RILEX

REVISTA SOBRE INVESTIGACIONES LÉXICAS

VOLUMEN MONOGRÁFICO

Coordinado por

María José Domínguez Vázquez y Carlos Valcárcel Riveiro

DESARROLLO DE APLICACIONES PARA LA GENERACIÓN AUTOMÁTICA DEL LENGUAJE: LOS RECURSOS DEL PORTAL LEXICOGRÁFICO PORTLEX

DICIEMBRE, 2023

María José Domínguez Vázquez
Natália Català Torres
Daniel Bardanca Outeiriño
Rosa María Martín Gascueña
Carlos Valcárcel Riveiro
Laura Pino Serrano
Nerea López Iglesias

REVISTAS CIENTÍFICAS DE LA UNIVERSIDAD DE JAÉN

https://doi.org/10.17561/rilex.6.3

RILEX. Revista sobre investigaciones léxicas
ISSNe: 2605-3136

Índice

DESARROLLO DE APLICACIONES PARA LA GENERACIÓN AUTOMÁTICA DEL LENGUAJE: LOS RECURSOS DEL PORTAL LEXICOGRÁFICO PORTLEX

Tal y como indica el título, el presente número monográfico presenta diferentes recursos alojados en el portal lexicográfico *PORTLEX*. A lo largo del volumen se pueden encontrar referencias al diccionario multilingüe de la valencia del nombre *Portlex*, el cual sirvió de catalizador para el desarrollo de nuevos proyectos. Los trabajos aquí compilados, no obstante, ponen el foco en la descripción de los fundamentos teóricos y metodológicos, así como en las técnicas y herramientas aplicadas para el desarrollo de recursos plurilingües de análisis y generación automática del lenguaje natural con aplicación lexicográfica. Se trata, en concreto, de los prototipos *Xera*, *XeraWord*, *Combinatoria* y *CombiContext*, los cuales generan automáticamente ejemplos a partir de parámetros de consulta sintácticos-semánticos. Por tanto, estos simuladores ofrecen información sobre el potencial combinatorio de sustantivos valenciales creando automáticamente ejemplos dinámicos con diferentes finalidades, entre ellas la ejemplificación lexicográfica. Estos ejemplos generados por los propios usuarios muestran el vocabulario que puede ocupar determinadas casillas funcionales. De este modo, dichos recursos aportan combinaciones tanto en el eje sintagmático como paradigmático. El acceso en la interfaz de usuario es, dependiendo del recurso, semasiológico u onomasiológico.

El público encontrará además detalladas descripciones de herramientas esenciales para el análisis y la extracción de datos de WordNet, el método seguido para la anotación semántico-ontológica de los datos recogidos, así como la propia ontología diseñada para la finalidad de los proyectos en los que se enmarcan los recursos. Destacan aquí, junto con los generadores, herramientas como *Combina*, *Lematiza*, *Ontología léxica*, *TraduWord* y el etiquetador en desarrollo ESMAS-ES^{+}, todos ellos detalladamente explicados en la monografía. En el volumen también se presta

atención a resultados ligados a la aplicación de redes neuronales y métodos predictivos para el análisis de la similitud semántica y las coocurrencias.

Junto con los temas señalados, la monografía aporta aproximaciones teóricas al estudio valencial de la frase nominal, discute dificultades en el análisis lingüístico y computacional para desarrollar herramientas de procesamiento del lenguaje natural y propone el uso de los generadores del lenguaje en la enseñanza de lenguas junto con modelos concretos para su aplicación.

El volumen se articula en los siguientes artículos y temas:

El estudio *LA AVENTURA DE LOS GENERADORES AUTOMÁTICOS DEL LENGUAJE NATURAL: DEL ANÁLISIS LINGÜÍSTICO AL PROCESAMIENTO AUTOMÁTICO DE DATOS* de María José Domínguez Vázquez dota al volumen de una descripción general de los generadores desarrollados en el portal *PORTLEX* y sus fundamentos. De este modo, se describe el marco teórico, la gramática y lexicografía de valencias, así como las características generales de los cuatro generadores automáticos del lenguaje y la tipología de datos que ofrecen.

Natalia Catalá Torres aporta en *SOBRE LA ESTRUCTURA DE LOS SINTAGMAS NOMINALES* una aproximación teórica amplia sobre el sintagma nominal y la estructura argumental, abordando los tipos y propiedades de nominalizaciones, los sustantivos no derivados y los sintagmas nominales en los proyectos *MultiGenera* y *MultiComb*.

Una visión de conjunto de diferentes recursos y herramientas manejadas en diferentes estadios de desarrollo de los simuladores de generación, así como las técnicas y estrategias aplicadas se encuentra en *GUÍA DE TÉCNICAS, ESTRATEGIAS Y HERRAMIENTAS EN EL DISEÑO Y DESARROLLO DE GENERADORES AUTOMÁTICOS DEL LENGUAJE* de Daniel Bardanca Outeiriño y María José Domínguez Vázquez. El estudio también refleja diferentes fases metodológicas hasta alcanzar el objetivo final, la generación de ejemplos dinámicos y esquemas argumentales del nombre anotados semánticamente.

Uno de los recursos centrales, diseñado *ad hoc* para el desarrollo de los simuladores de generación automática del lenguaje, es la ontología léxica *bottom up*. Rosa María Martín Gascueña ofrece en su investigación *DISEÑO DE UNA ONTOLOGÍA DE SEMÁNTICA LÉXICA EN LOS PROYECTOS MULTIGENERA Y MULTICOMB* un estudio detallado de las ontologías de WordNet, elemento clave en el desarrollo de los prototipos de generación. La autora presenta las diferentes fases en el desarrollo de la ontología léxica propia aplicada en todos los simuladores para el etiquetado semántico.

El estudio de Carlos Valcárcel Riveiro y Laura Pino *HERRAMIENTAS Y DIFICULTADES EN EL ANÁLISIS DEL GRUPO NOMINAL EN FRANCÉS PARA SU PROCESAMIENTO COMPUTACIONAL* revisa el trabajo desarrollado para la lengua francesa en los diferentes proyectos del portal *PORTLEX*, como ejemplo extrapolable a las otras lenguas incluidas en los recursos. Se presentan los resultados obtenidos y se hace un análisis detallado de las herramientas utilizadas por los equipos de trabajo para el análisis, la extracción y el procesamiento de datos en francés atendiendo tanto a sus funcionalidades como a sus limitaciones.

Cierra el volumen el trabajo de Nerea López Iglesias *MUCHO MÁS QUE EJEMPLOS: APLICACIONES DIDÁCTICAS DE LOS GENERADORES AUTOMÁTICOS*, el cual incide en la importancia del contexto en el desarrollo de la competencia léxica. La autora describe cómo los ejemplos generados automáticamente por los generadores pueden ser de aplicación directa en el aula de lenguas extranjeras, pero también propone actividades concretas en línea diseñadas a partir de los datos que ofrecen los prototipos de generación automática.

Un hilo conductor de todos los trabajos es la importancia concedida a la interoperabilidad y sostenibilidad. De este modo, los prototipos de generación automática del lenguaje se retroalimentan entre sí y sus datos son integrables en otros recursos. Otro ejemplo de ello es el desarrollo del nuevo recurso compilado en el portal lexicográfico, el etiquetador plurilingüe semántico y

automático ESMAS-ES+, actualmente en elaboración. Este bebe de las fuentes de los datos lingüísticos anotados semánticamente, de las herramientas diseñadas para la generación automática del lenguaje y de la traducción automática del caudal léxico. Algunas de las optimizaciones de los recursos de generación automática presentados resultan de la investigación al abrigo de ESMAS-ES+.

Las diferentes herramientas, recursos y simuladores se han desarrollado u optimizado al abrigo de diferentes proyectos competitivos:

- *MultiGenera. Generación multilingüe de estructuras argumentales del sustantivo y automatización de extracción de datos sintáctico-semánticos*. Fundación BBVA. Ayudas Fundación BBVA a Equipos de Investigación Científica - Humanidades Digitales. 2017-2020. http://portlex.usc.gal/multigenera/
- *MultiComb. Generador multilingüe de estructuras argumentales del sustantivo con aplicación en la producción en lenguas extranjeras*. FI2017-82454-P: Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia, Generación de Conocimiento. MCIN/AEI/ FEDER "Una manera de hacer Europa" (EXCELENCIA 2017, 2017-PN091). 2018-2021. http://portlex.usc.gal/multicomb/
- *Ferramentas TraduWord e XeraWord: tradución de caudal léxico e xeración automática da linguaxe natural en galego e portugués*. 2020-PU004. Convocatoria proyectos de colaboración. Universidade de Santiago de Compostela. https://ilg.usc.gal/xeraword/
- *Etiquetador semántico multilingüe automático y sostenible*. ESMAS-ES+. PID2022-137170OB-I00: Programa Estatal para Impulsar la Investigación Científico-Técnica y su Transferencia, del Plan Estatal de Investigación Científica, Técnica y de Innovación 2021-2023. Generación de Conocimiento. MCIN/AEI//FEDER "Una manera de hacer Europa". 2023-2027.

A su vez, los resultados de investigación han sido propiciados, son objeto de estudio o son aplicados por el grupo de investigación Humboldt (Grupo GI 1920, Universidad de Santiago de Compostela), el grupo de innovación docente MeReLing (Universidad de Vigo) y el Instituto da Lingua Galega (Universidad de Santiago de Compostela), entre otros.

Nuestro más sincero agradecimiento a las instituciones que apoyan nuestro trabajo, a los evaluadores y las evaluadoras por sus contribuciones, así como a la revista por permitirnos presentar los resultados de nuestra investigación en este foro.

Los editores

LA AVENTURA DE LOS GENERADORES AUTOMÁTICOS DEL LENGUAJE NATURAL: DEL ANÁLISIS LINGÜÍSTICO AL PROCESAMIENTO AUTOMÁTICO DE DATOS

THE ADVENTURE OF GENERATORS OF NATURAL LANGUAGE: FROM LINGUISTIC ANALYSIS TO AUTOMATIC DATA PROCESSING

María José Domínguez Vázquez
Universidade de Santiago de Compostela
majo.dominguez@usc.es

Resumen

Este capítulo ofrece una visión panorámica de los generadores *Xera*, *XeraWord*, *Combinatoria* y *CombiContext*. En el apartado 2 se explican los motivos que condujeron al equipo de investigación del diccionario *Portlex* a explorar una vía de trabajo hasta el momento desconocida para nosotros: la generación automática de datos lingüístico-valenciales anotados semánticamente junto con sus ejemplos. El apartado 3 sirve de descripción general de los fundamentos de los generadores en su conjunto: aspectos como su tipología, las estructuras de acceso a la información o los niveles informativos. Una breve sinopsis de su estructura desde la perspectiva de su uso se encuentra en el apartado 4.

Palabras clave: generadores automáticos del lenguaje natural, lexicografía, anotación valencial, interfaz sintáctico-semántica.

Abstract

This chapter aims to give an overview of the generators *Xera*, *XeraWord*, *Combinatoria*, and *CombiContext*. The reasons that led the *Portlex* dictionary research team to explore an approach hitherto unknown to us -the automatic generation of semantically annotated linguistic-valuative data and its examples- are explained in Section 2. Section 3 provides the fundamentals of the generators as a whole: aspects such as their typology, information access structures, or information levels. A brief synopsis of their structure and the perspective of their use can be found in section 4.

Keywords: automatic natural language generators, lexicography, valency annotation, syntax-semantics interface.

DOI: https://doi.org/10.17561/rilex.6.3.8105

1. INTRODUCCIÓN

Los prototipos de generación automática del lenguaje natural *Xera*, *XeraWord*, *Combinatoria* y *CombiContext* se caracterizan en líneas generales como sigue:

- Se trata de prototipos de generación automática que cumplen al completo la finalidad para la que han sido concebidos: servir de experimentos piloto para verificar un nuevo método combinado de análisis.
- Describen en su conjunto el español, gallego, francés, alemán y portugués.
- Reutilizan datos, técnicas y modelos verificados previamente, así como recursos en abierto. Por tanto, siguen principios de retroalimentación e interoperabilidad en favor de la sostenibilidad.
- Describen y generan automáticamente ejemplos de valencia activa y pasiva del nombre siguiendo diferentes patrones argumentales. Desde un punto de vista tipológico son especialmente novedosos frente a otros generadores de la lengua (Domínguez Vázquez, 2022b): teniendo en cuenta al usuario, aportan opciones de consulta sintáctico-semántica no contempladas en otros recursos de su entorno más cercano.

Los prototipos presentados en este volumen se desarrollan al amparo de diferentes proyectos competitivos:

- *MultiGenera. Generación multilingüe de estructuras argumentales del sustantivo y automatización de extracción de datos sintáctico-semánticos*. Fundación BBVA. Ayudas Fundación BBVA a Equipos de Investigación Científica - Humanidades Digitales. 2017-2020. http://portlex.usc.gal/multigenera/
- *MultiComb. Generador multilingüe de estructuras argumentales del sustantivo con aplicación en la producción en lenguas extranjeras*. FI2017-82454-P: Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia, Generación de Conocimiento. MCIN/

AEI/ FEDER "Una manera de hacer Europa" (EXCELENCIA 2017, 2017-PN091). 2018-2021. http://portlex.usc.gal/multicomb/

- *Ferramentas TraduWord e XeraWord: tradución de caudal léxico e xeración automática da linguaxe natural en galego e portugués*. 2020-PU004. Convocatoria proyectos de colaboración. Universidade de Santiago de Compostela. https://ilg.usc.gal/xeraword/

Tal y como recoge la bibliografía, este capítulo compendia resultados de las publicaciones más recientes sobre dichos generadores (Domínguez Vázquez, 2022a, 2022b; Domínguez Vázquez, 2021; Domínguez Vázquez, Bardanca Outeiriño & Simões, 2021; Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019, por citar algunas). Para cuestiones teóricas y metodológicas más detalladas remitimos al lector a estos trabajos y a la página web de los proyectos (http://portlex.usc.gal/), así como a los diferentes capítulos de este volumen.

2. EN EL PUNTO DE PARTIDA

La idea de diseñar generadores automáticos del lenguaje natural nace al abrigo de un proyecto competitivo previo, el diccionario multilingüe *cross-lingual* de la valencia del nombre en alemán, español, francés, italiano y gallego, *Portlex*[1] (Domínguez Vázquez & Valcárcel Riveiro, 2020)[2]. Las investigaciones realizadas en este contexto nos permitieron constatar, por una parte, la complejidad de verificar estructuras sintáctico-semánticas valenciales en corpus para todas y cada una de las realizaciones de superficie, así como para todas las lenguas contempladas en el recurso. A esto se suma, por otra parte, la costosa tarea de compilar ejemplos de corpus adecuados a los propósitos de nuestro diccionario para las cinco lenguas del proyecto.

[1] *Portlex*: Ref.FFI2012-32456. *Portal Lexicográfico: Diccionario online modular multilingüe y corpus informatizado anotado de la frase nominal*. Ministerio de Economía y Competitividad. 2013-2015.
[2] http://portlex.usc.gal/portlex/

Junto con los desequilibrios observados en cuanto al volumen y representatividad de los datos extraídos de los diferentes corpus manejados (CREA para el español, CORGA para el gallego, DeReKo para el alemán, FRANTEXT para el francés y PAISÀ para el italiano)[3], comprobamos la inadecuación de muchos ejemplos debido a factores como su sobresaturación informativa (resultando difícil mostrar aquello para lo que estaba concebido el diccionario), el papel de los pronombres o las anáforas en los mismos, y, en definitiva, la necesidad de encontrar ejemplos con un vocabulario representativo para todas y cada una de las combinaciones posibles en las cinco lenguas (Valcárcel Riveiro & Pino Serrano, 2023).

Asimismo, el tipo de diccionario –un diccionario de valencias– requiere un análisis y descripción de la valencia no sólo sintáctica, sino también semántica (roles semánticos y rasgos ontológicos). De este modo, han de contemplarse fenómenos relacionados con la obligatoriedad o facultatividad de los argumentos del esquema valencial y su relación con la acepción de significado actualizada en los diferentes casos. En esta línea, la aplicación de filtros comunes de extracción de datos de corpus, tales como criterios de frecuencia o co-ocurrencias, permiten obtener lógicamente datos cuantitativos, pero estos no son necesariamente determinantes para un diccionario de estas características: la frecuencia de un elemento no se encuentra en necesaria correlación con su función de complemento específico, cuya descripción es el fin último de todo diccionario de valencias. Por tanto, el hecho de no contar con

[3] CREA = *Corpus de referencia del español actual*. Real Academia Española. http://corpus.rae.es/creanet.html
CORGA = *Corpus de referencia do galego actual*. Centro Ramón Piñeiro para a Investigación en Humanidades. http://corpus.cirp.es/corga
DeReKo = *Das Deutsche Referenzkorpus*. Institut für Deutsche Sprache. http://www1.ids-mannheim.de/kl/projekte/korpora
FRANTEXT = *Base textuelle* FRANTEXT. ATILF - CNRS & Université de Lorraine. http://www.frantext.fr
PAISÀ = *Piattaforma per l'Apprendimento dell'Italiano Su corpora Annotati*. Università di Bologna/CNR Pisa/Accademia Europea di Bolzano/Università di Trento. http://www.corpusitaliano.it/

corpus anotados semánticamente para las lenguas objeto de análisis dificulta y aumenta exponencialmente el trabajo manual de extracción, compilación y depurado de los datos, así como la obtención de ejemplos representativos. Así, por ejemplo, en el motor Sketch Engine (español) para una búsqueda CQL

> [[tag="D.*"][lemma="discusión"][word="de|del"][tag="DA.*|DD.*|DI.*|DP*"][tag="N.C.*"]]

aparece en el top 20, por ejemplo, *la discusión de esta mañana*. Ninguna de las pruebas que se suelen aplicar (reconversión a oración, test de la anáfora, test de la pregunta; vid. Domínguez, 2011) permite categorizar esta frase preposicional como posible argumento del sustantivo DISCUSIÓN. Por el contrario, *la discusión de los docentes* sí que permite su clasificación como tal: *los docentes discuten*, en donde la frase nominal se reconvierte en complemento sujeto de la oración. Por tanto, el manejo de corpus no evita un elevado trabajo manual para documentar ejemplos de patrones argumentales, complejidad que se acrecienta a medida que dichos patrones argumentales se vuelven más complejos.

Desde la perspectiva del usuario aprendiente de lenguas la situación no es mucho más satisfactoria. Cuando comenzamos con la elaboración del diccionario *Portlex*, observamos también la escasa o nula posibilidad de consultar datos aplicando criterios sintáctico-semánticos para las lenguas que describe el recurso. Así, por ejemplo, gramáticas, libros de textos y diccionarios no permiten al usuario plantear una consulta individualizada: no solo el número de ejemplos aportados, sino las opciones de filtrado son restringidas. La limitación de espacio, que parece haber sido superada con los recursos en línea, tampoco revierte esta situación: si bien los diccionarios y portales lexicográficos aportan más ejemplos que los recursos impresos –muchos sistemas lexicográficos y plataformas ofrecen más ejemplos tanto dentro de su propia estructura como mediante hipervinculación a recursos externos al de acceso primario de consulta–, estos no dejan de ser en su mayoría nuevamente ejemplos de corpus sin opciones de filtrado semántico. En resumen, el usuario no puede seleccionar ejemplos o estructuras

específicas según filtros sintáctico-semánticos concretos atendiendo a sus necesidades de consulta.

Todos estos factores nos condujeron a la idea de generar automáticamente patrones argumentales sintáctico-semánticos y ejemplos dinámicos, en lugar de extraerlos de diferentes corpus. Este es el origen de los proyectos *MultiGenera*, *MultiComb* y *XeraWord* (vid. 1). Al amparo de dichos proyectos se diseñan diferentes herramientas para el análisis lingüístico y automatización de procedimientos (vid. Bardanca Outeiriño y Domínguez Vázquez en este volumen), pero, en especial, se desarrollan los generadores *Xera*, *Combinatoria* y *CombiContext* para el español, francés y alemán, y *XeraWord* para el gallego y portugués.

3. TEORÍA VALENCIAL

Un estudio detenido de la literatura científica permite constatar no solo diferentes aproximaciones al sustantivo y a su capacidad para ser portador valencial, sino que además evidencia las notables diferencias terminológicas y la asimetría en los inventarios de complementos del nombre y sus realizaciones formales. En el primer caso contamos con aproximaciones que entienden que el sustantivo no puede abrir casillas valenciales (Ágel, 2000) hasta aquellas que consideran la valencia nominal como un sistema *sui generis* (Teubert, 1979; Kubczak & Schumacher, 1998). A su vez, también es conocido que la escasa atención dedicada al sustantivo se debe, entre otras, a los postulados sobre la herencia de su potencial combinatorio a partir de sus bases derivativas, si bien sustantivo y palabra base pueden diferir cuantitativa y cualitativamente desde un punto de vista valencial (Díaz Hormigo, 2002; Domínguez Vázquez, 2011)[4].

Siguiendo el modelo de Engel (2004), Domínguez Vázquez (2011) propone un modelo contrastivo para la valencia del nombre, que recurre al concepto

[4] No me detendré más en estos aspectos, véase para una descripción detallada Domínguez Vázquez (2011).

de escenario y escenas como *tertium comparationis* inter- e intralingüístico. Según esta aproximación, el número y tipo de roles centrales, así como el rol focalizado son conceptos clave para poder adjudicar un sustantivo a una escena concreta, y, en su nivel jerárquico superior, a un escenario. Esto se ejemplifica sencillamente con sustantivos como HUIDA frente a otros como ESTANCIA. El primero cuenta con un argumento focalizado, expresado explícitamente o no, que expresa movimiento en su esquema argumental, el cual no está presente en el segundo de los sustantivos citados. La Figura 1 presenta la relación entre escenas y subescenas en el escenario MOVIMIENTO y, por tanto, ejemplifica la delimitación de HUIDA frente a otros sustantivos de su mismo escenario.

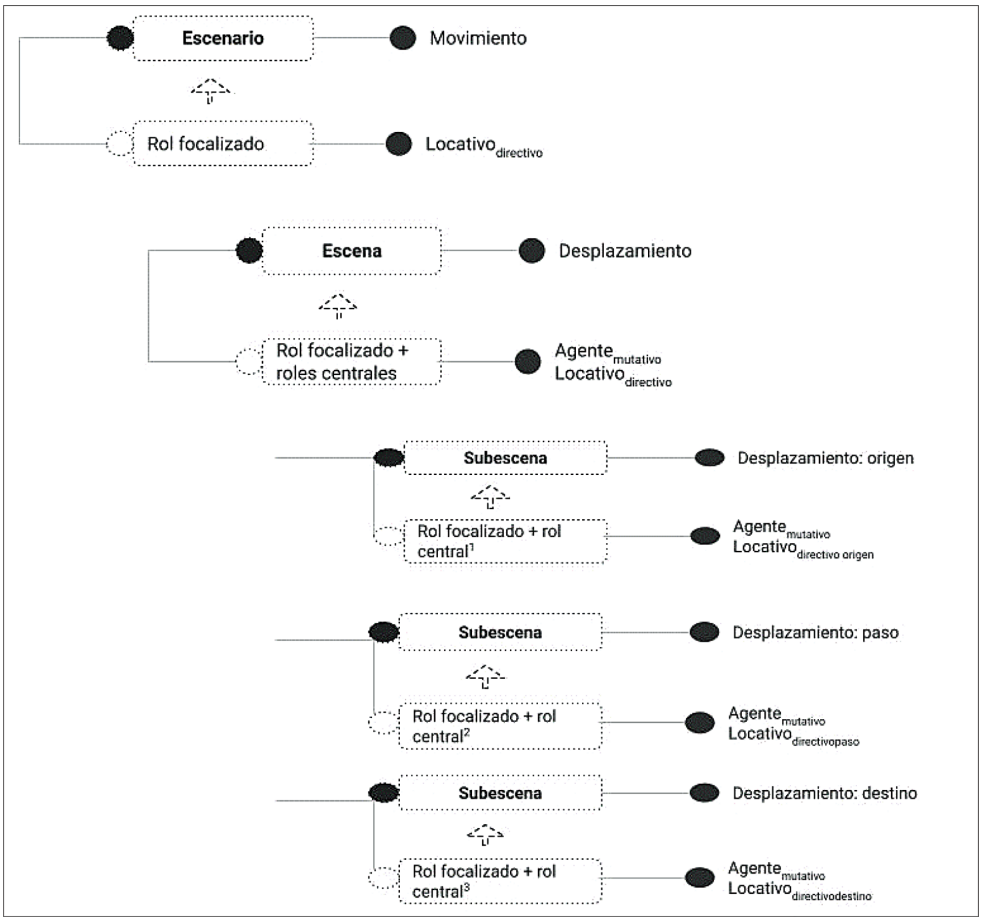


Figura 1: *Escenario MOVIMIENTO con escenas y subescenas*

Como se desprende de la Figura 1 existe una relación de mapeo entre escenas, subescenas y escenario, de modo que aquellos sustantivos de un escenario o escena comparten el marco conceptual-semántico y, por tanto, los roles semánticos y argumentos centrales.

El modelo descriptivo comprende diferentes niveles de análisis:

i) el plano semántico-combinatorio (significado relacional y categorial-ontológico; vid. Engel, 1996). La descripción del significado relacional se asienta en el siguiente inventario de roles semánticos ya aplicado en el diccionario *Portlex* (Figura 2):

Aquel/aquello que realiza una acción
Aquel/aquello afectado
Aquel/aquello no afectado
Aquel/aquello no afectado: Tema
Aquel/aquello que existe o es
Clasificativo
Aquel/aquello que tiene o dispone de algo
Objetivo no espacial
Extensión
Localización abstracta
Aquel/aquello que experimenta un cambio/ que experimenta un estado o situación
Aquel/Aquello que es el origen o la causa
Origen
Paso
Destino
Locación

Figura 2: *Inventario de roles semánticos*

Los rasgos categoriales parten de los inventarios de la gramática y lexicografía valencial (Engel, 2004, E-Valbu) y van evolucionando hasta una ontología léxica *bottom up*, que se retroalimenta de las ontologías de WordNet (vid. capítulo Bardanca Outeiriño y Domínguez Vázquez en este volumen, así como Martín Gascueña).

ii) el plano sintáctico argumental o el patrón argumental: según Domínguez Vázquez (2011) los tipos sintácticos complementos del nombre son *Genitivus subiectivus*, *Genitivus obiectivus*, Complemento/Suplemento prepositivo, Complemento adverbial, Complemento verbativo y Complemento nominal.

iii) el plano morfosintáctico: en este nivel se describen las diferentes realizaciones formales de los argumentos nominales. Junto con las frases preposicionales y el genitivo (para el alemán) el inventario de realizaciones también contempla adjetivos, compuestos (para el alemán) y las aposiciones N+N (Valcárcel Riveiro, 2017). La inclusión de dichas realizaciones no es común en la gramática y lexicografía valencial. En nuestros recursos entendemos que en ejemplos como *la huida apresurada* frente *a la huída marítima* la función sintáctica realizada por uno y otro adjetivo es diferente. Así, el segundo ejemplo explicita la vía de huida, del mismo modo que sucede en *la huida por el Mediterráneo*. Dado que *por el Mediterráneo* se considera un argumento nominal locativo de paso al cumplir los criterios de pregunta y anáfora, nada debería de impedir, por tanto, categorizar esta realización adjetival del mismo modo.

4. GENERADORES AUTOMÁTICOS DEL LENGUAJE NATURAL

4.1. *Descripción general*

El principal objetivo de los simuladores es ofrecer información sobre el potencial combinatorio de sustantivos valenciales junto con ejemplos dinámicos y, por tanto, el vocabulario que puede cubrir el eje paradigmático y sintagmático de diferentes complementos específicos. Para tal fin se atiende a

- su aparición aislada, como en una frase preposicional simple del tipo *el olor a rancio* (datos de los generadores *Xera* y *Xeraword*)
- su realización combinada a nivel frasal, como en *el desagradable olor a rancio* frente a **el agradable olor a rancio* (resultados ofrecidos por *Combinatoria*)

- su combinatoria en el plano oracional, como en *El desagradable olor a rancio se extendía por la habitación* (datos aportados por *CombiContext*).

Los generadores tienen en común la aplicación de una metodología combinada que permite procesar datos con información semántica. Se fundamentan en la interoperabilidad y retroalimentación de recursos, así como en diferentes aproximaciones lingüístico-computacionales: i) la gramática de valencias, la teoría de los prototipos léxicos y clases semánticas, ii) el análisis de corpus, ontologías, bases de datos de coocurrencias y redes semánticas, como WordNet, iii) el procesamiento del lenguaje natural (PLN; recuperación y extracción de información), iv) modelos neuronales y métodos predictivos, como *word2vec* (Mikolov, Chen, Corrado & Dean, 2013) y *fastText* (Bojanowski, Grave, Joulin & Mikolov, 2017), v) la generación automática (GLN) y vi) la traducción automática (en el caso del cuarto generador, *XeraWord*).

Un sinóptico de la interrelación entre el flujo de trabajo y las herramientas aplicadas se presenta en la Figura 3, actualizada a partir de Domínguez Vázquez (2022a).

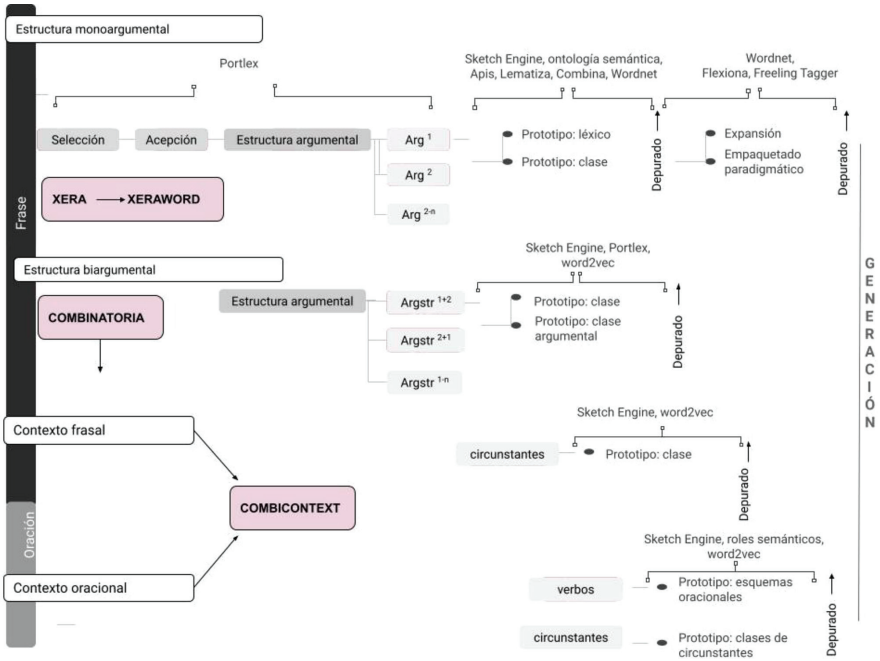


Figura 3: *Interrelación de flujo de trabajo y herramientas*

Desde una perspectiva tipológica los generadores representan un nuevo modelo de sistemas de información digitales dinámicos e individualizados, en concreto, de diccionarios de valencias plurilingües. Han sido diseñados para destinatarios humanos, pero también para su aprovechamiento por máquinas, siendo estos integrables y exportables como léxicos computacionales. Comparten con otros recursos de su espectro más cercano características mediales como su accesibilidad en red y su uso simultáneo por más de un usuario. Son gratuitos y de libre acceso.

Frente a otros recursos, una de sus principales novedades es su dinamismo personalizado: en portales lexicográficos, diccionarios y plataformas suele ser posible obtener más ejemplos consultando el propio portal o a través de un enlace a ejemplos externos al recurso. De este modo, el usuario puede inducir mediante la observación de un conjunto amplio de ejemplos determinadas reglas (o también se puede perder en el volumen de datos). No obstante, dichas reglas no son explícitas y una búsqueda concreta e individualizada de datos no es siempre posible. A diferencia de otros recursos, los prototipos proponen un enfoque intermedio: sus ejemplos no son ni extraídos directamente de corpus para su integración directa en los generadores, ni son elaborados *ad hoc* por el equipo lexicográfico, sino que son generados automáticamente. Esta vía intermedia permite evitar la sobresaturación informativa del ejemplo, así como ligar el vocabulario a clases semánticas y rasgos ontológicos –los cuales sirven además como filtro de consulta. Por tanto, en los generadores es posible una consulta y selección de datos y ejemplos concretos siguiendo filtros sintáctico-semánticos aplicados por el usuario. Esto los diferencia de otros recursos con ejemplos y patrones estáticos. De este modo, se permite al usuario corroborar (o no) su hipótesis de consulta inicial (Müller-Spitzer, Domínguez Vázquez, Nied Curcio, Silva Dias, & Wolfer, 2018) y se responde a la pregunta de si ciertas combinaciones son posibles (o no) en determinadas situaciones de producción. A su vez, se posibilita descubrir (o confirmar) el uso de determinadas unidades léxicas y sus entornos sintácticos o contextos (Domínguez Vázquez &

Gouws, 2023). Dicha consulta selectiva es posible porque los generadores integran una descripción de la interfaz sintáctico-semántica, y, por tanto, permiten extraer y consultar los datos atendiendo a dicho aspecto.

Dado que las propias herramientas también han sido concebidas para facilitar información de diferente calado, el tipo de ejemplos que aportan muestra similitudes, pero también diferencias entre sí. Así, *Xera* y *XeraWord* presentan ejemplos de frases simples (monoargumentales) que incluyen las realizaciones de superficie y los rasgos ontológicos vinculados a un argumento valencial. *Combinatoria* aporta datos semejantes, pero para estructuras complejas o biargumentales. Un ejemplo de los argumentos y esquemas de *Xera* y *Combinatoria*, así como el tipo de información se ofrece en la Tabla 1:

ARGUMENTOS				
Plano formal	Plano semántico		Ejemplo	Recurso
	Rol semántico	Características ontológicas		
determinante + adjetivo + ESTANCIA + de + determinante+ ARG 1	N1: Aquel que realiza la estancia	{Humano condición negativa}: *paciente* | *enfermo*	*La estancia del paciente*	*Xera*
determinante + adjetivo + ESTANCIA + en + determinante + ARG 2	N2: lugar en donde se realiza la estancia	{Lugar edificio: tipo: medicina}: *hospital* | *clínica*	*La estancia en el hospital*	
PATRONES SINTÁCTICO-SEMÁNTICOS				
determinante + adjetivo + ESTANCIA + de + determinante+ ARG 1 + en + determinante + ARG 2	N1 {Humano condición negativa} + N2 {Lugar edificio Tipo: medicina}		*La estancia del paciente en el hospital*	*Combinatoria*

Tabla 1: *Ejemplo de argumentos y patrones*

El último de los generadores diseñados para alemán, francés y español, *CombiContext*, ofrece información sobre el marco oracional en el que se incrustan las frases nominales simples y complejas. Desde un punto de vista cuantitativo, cabe señalar que *CombiContext*, se retroalimenta de los a) 3.625 argumentos

específicos y b) 20.600 esquemas sintáctico-semánticos que generan *Xera* y *Combinatoria*. También se nutre de 29.700 modificadores adjetivales y más de 820 verbos, ofreciendo en su estado actual 90.000 esquemas argumentales. La Figura 4 recoge estos datos y ejemplos concretos de los diferentes bloques de información.

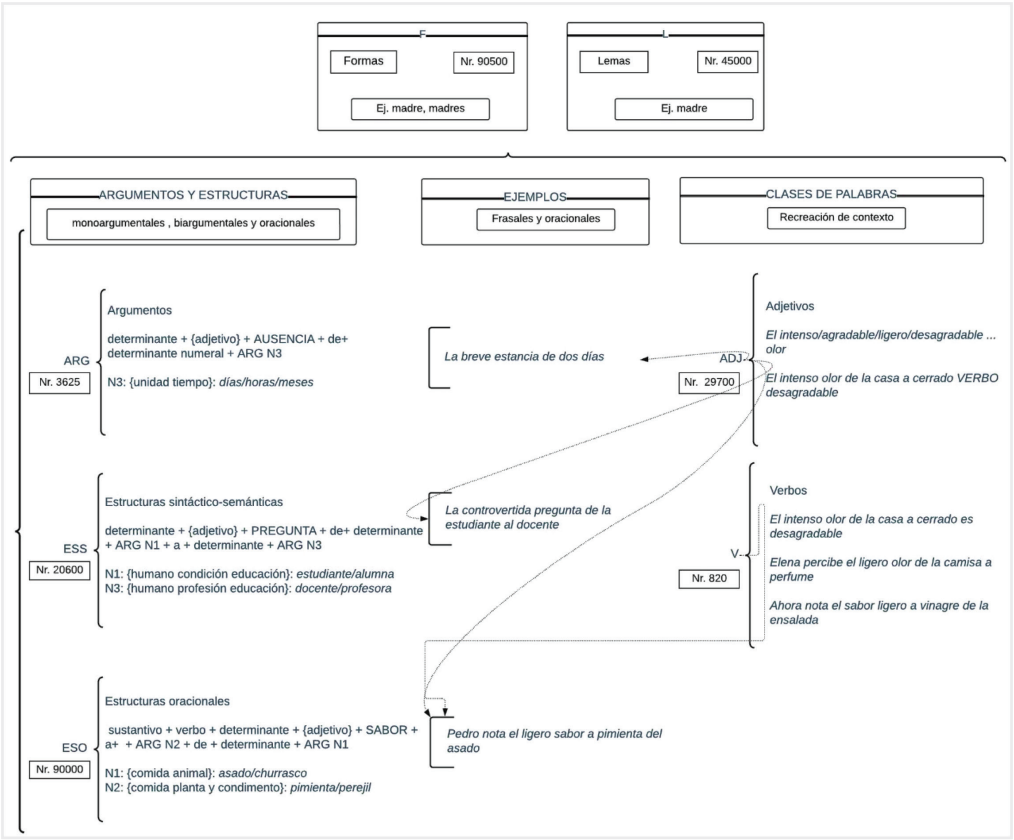


Figura 4: *Niveles informativos de los generadores*

Otro de los aspectos a los que se le ha dedicado especial atención en el diseño de los recursos son las rutas de acceso a la información en la interfaz de usuario. Así, en *Xera* y *XeraWord*[5], los dos recursos monoargumentales, la aproximación a la consulta es formal. En la herramienta *Combinatoria* la aproximación es ontológico-conceptual. La herramienta *CombiContext* parte

[5] *XeraWord*, la herramienta piloto para el gallego y portugués,cuenta con estructuras monoargumentales para dichas lenguas. Su diseño perseguía comprobar la viabilidad de nuevas vías de automatización para analizar, extraer y generar ejemplos de patrones argumentales. En XeraWord, por tanto, se incorpora como metodología de trabajo la traducción automática del caudal léxico de *WordNet*.

de una aproximación distribucional –la posición de la frase con núcleo valencial en relación con el verbo (vid. 4.2.).

A modo de resumen, la tabla 2 ofrece una comparativa de las características generales de los generadores:

	Xera	XeraWord	Combinatoria	Combi-Context
Lenguas: alemán, español, francés	✓		✓	✓
Lenguas: gallego, portugués		✓		
Patrones monoargumentales	✓	✓		
Patrones biargumentales			✓	
Marco frasal y oracional				✓
Descripción formal	✓	✓	✓	✓
Descripción ontológica	✓	✓	✓	✓
word embeddings			✓	✓
Traducción automática	✓	✓[1]		
Acceso libre y gratuito	✓	✓	✓	✓
Interfaz de usuario: acceso formal primario	✓	✓		
Interfaz de usuario: acceso semántico primario			✓	
Interfaz de usuario: acceso distribucional primario				✓
Generación *ad libitum*	✓	✓	✓	✓
Exportación de datos	✓	✓	✓	✓

Tabla 2: *Información de los generadores en contraste*[6]

4.2. *Generadores automáticos de la lengua paso a paso*

El primero de los generadores diseñados es *Xera*. Proporciona esquemas y ejemplos monoargumentales de la frase nominal, como, por ejemplo, *el ancho de los muebles, el viaje a Beirut* o *el aumento de la inflación*.

Su acceso es formal (con unas abreviaturas poco intuitivas que requieren cierto hábito) y posteriormente semántico-conceptual. Esto supone que una vez seleccionada la realización formal objeto de consulta (en la Figura 5 [determinante + adjetivo + *viaje* + adjetivo +de + determinante + actante

[6] Leyenda: [1] = exclusivamente.

N1]), el usuario puede seleccionar una o diferentes clases semánticas, las cuales están acompañadas de ejemplos estándar a modo de guía. Seleccionada una de las clases semánticas, por ejemplo, {animado humano grupo o colectivo militar} y clicando en GENERAR, se visualizan ejemplos concretos que cumplen los requisitos formales y semánticos aplicados (Figura 6). De este modo, se pueden observar las restricciones de coaparición sintácti-co-semántica y, tras su selección, generar ejemplos *ad libitum*.

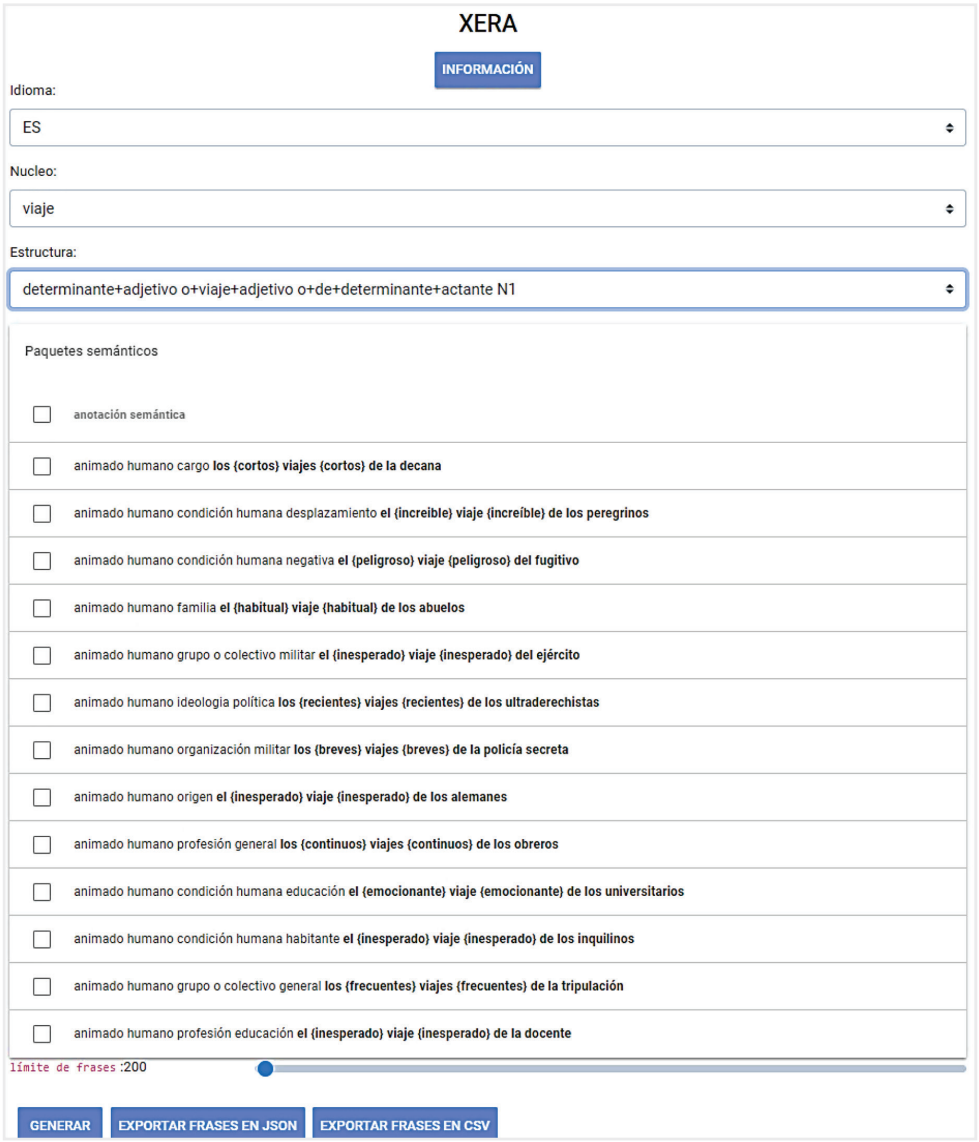


Figura 5: *Acceso en Xera*

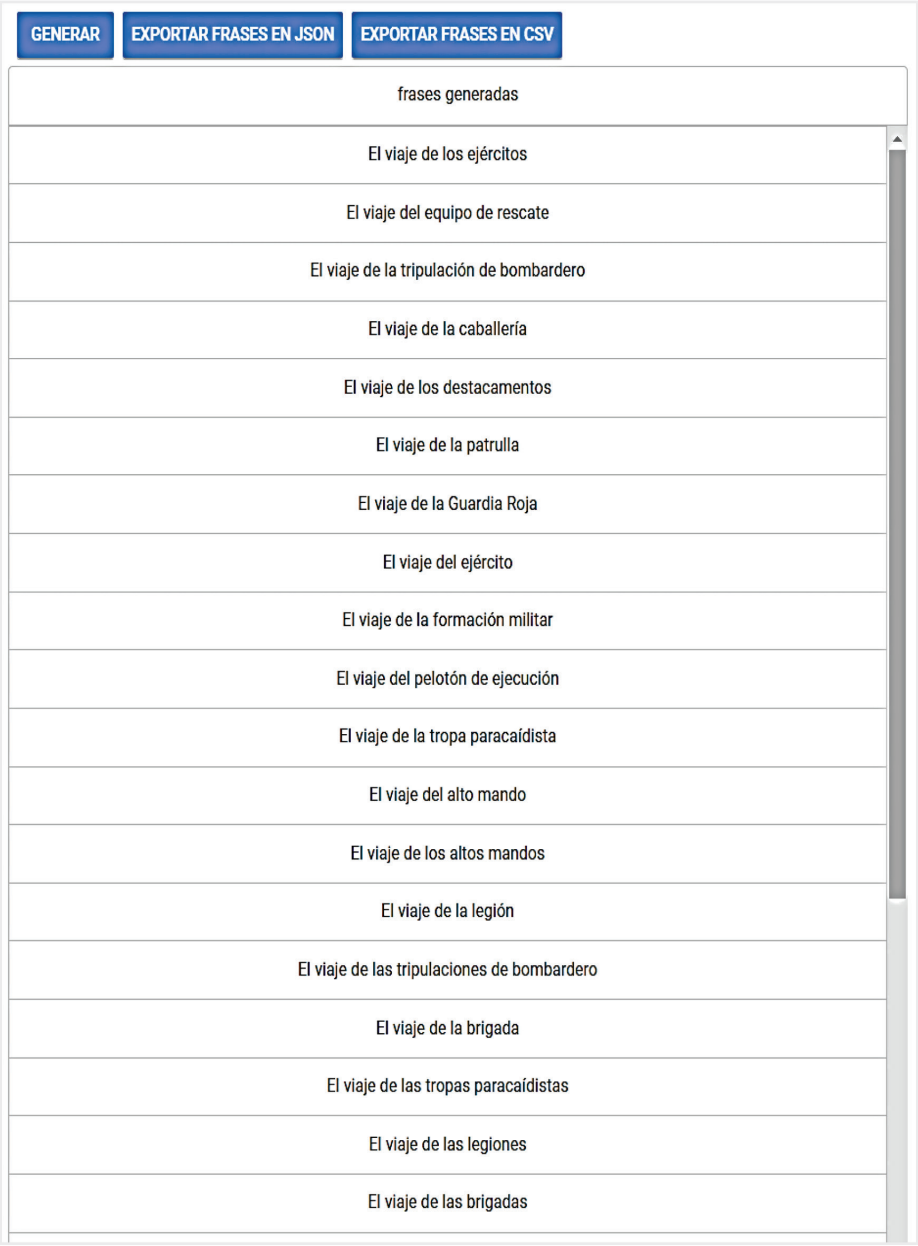


Figura 6: *Volcado de datos en Xera*

El segundo de los prototipos, *Combinatoria*, ofrece esquemas y ejemplos biargumentales, tales como *el viaje del explorador a Tierra Santa*, *el sabor del pastel a canela* o *la estancia formativa del investigador.* Este prototipo propone una perspectiva de consulta inversa a la de *Xera*: de lo conceptual a lo formal. Con la finalidad de favorecer la selección de las clases semánticas que

se quieren combinar o consultar, la herramienta incorpora ventanas emergentes con un ejemplo concreto de la clase semántica en cuestión (Figura 7).

La selección del argumento que aparece en primer lugar desglosa automáticamente posibles combinatorias de argumentos que pueden aparecer en segunda posición acompañando a la clase semántica seleccionada (por ejemplo {condición humana desplazamiento} en la Figura 7). La Figura 8 presenta todos los paquetes léxicos combinables con el primer actante seleccionado.

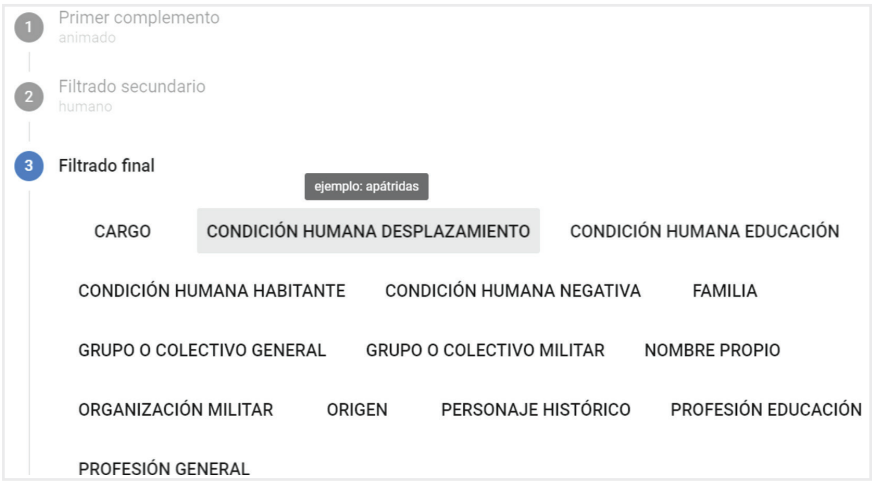
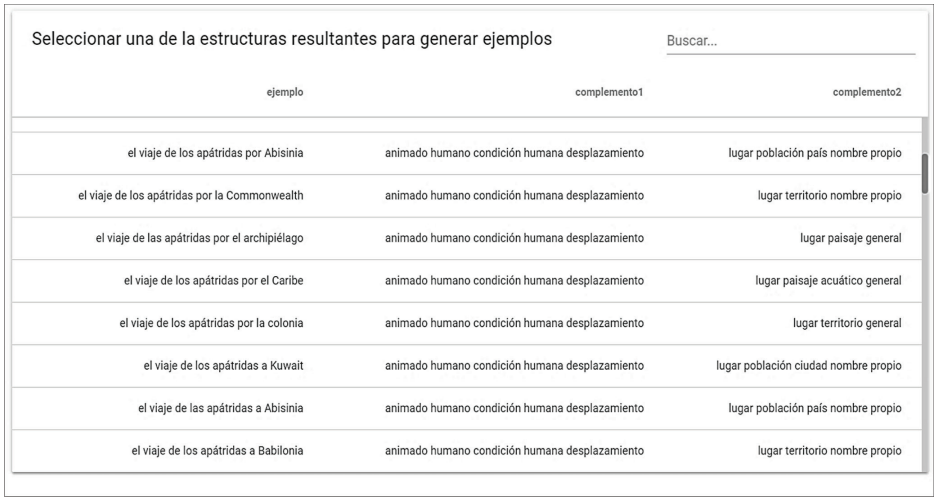


Figura 7: *Selección ontológica del argumento que aparece en primera posición*



ejemplo	complemento1	complemento2
el viaje de los apátridas por Abisinia	animado humano condición humana desplazamiento	lugar población país nombre propio
el viaje de los apátridas por la Commonwealth	animado humano condición humana desplazamiento	lugar territorio nombre propio
el viaje de las apátridas por el archipiélago	animado humano condición humana desplazamiento	lugar paisaje general
el viaje de los apátridas por el Caribe	animado humano condición humana desplazamiento	lugar paisaje acuático general
el viaje de los apátridas por la colonia	animado humano condición humana desplazamiento	lugar territorio general
el viaje de los apátridas a Kuwait	animado humano condición humana desplazamiento	lugar población ciudad nombre propio
el viaje de las apátridas a Abisinia	animado humano condición humana desplazamiento	lugar población país nombre propio
el viaje de los apátridas a Babilonia	animado humano condición humana desplazamiento	lugar territorio nombre propio

Figura 8: *Combinatoria con un segundo argumento no filtrado [vista parcial]*

De este modo, se obtiene una visión de conjunto de todas las combinatorias posibles con el primer elemento. Existe también la posibilidad de predeterminar el segundo argumento mediante la selección de una clase semántica concreta, como, por ejemplo, {país nombre propio} en la Figura 9 y de una estructura concreta de esa clase, como, por ejemplo, *el viaje de los apátridas desde Abisinia/por Abisinia/a Abisinia*, etc. (Figura 10):

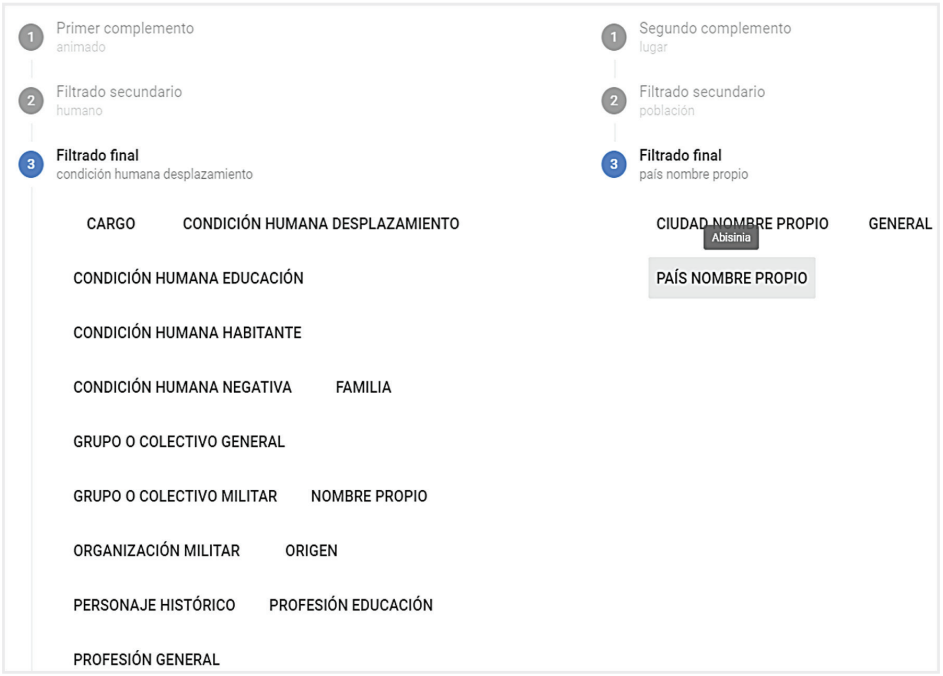


Figura 9: *Panel de combinatoria para el argumento que aparece en segunda posición*

Seleccionar una de la estructuras resultantes para generar ejemplos

Buscar...

ejemplo	complemento1	complemento2
el viaje de las apátridas desde Abisinia	animado humano condición humana desplazamiento	lugar población país nombre propio
el viaje de los apátridas por Abisinia	animado humano condición humana desplazamiento	lugar población país nombre propio
el viaje de las apátridas a Abisinia	animado humano condición humana desplazamiento	lugar población país nombre propio
el viaje de las apátridas hasta Abisinia	animado humano condición humana desplazamiento	lugar población país nombre propio
el viaje de los apátridas desde Abisinia	animado humano condición humana desplazamiento	lugar población país nombre propio
el viaje de los apátridas por Abisinia	animado humano condición humana desplazamiento	lugar población país nombre propio
el viaje de las apátridas a Abisinia	animado humano condición humana desplazamiento	lugar población país nombre propio
el viaje de los apátridas hasta Abisinia	animado humano condición humana desplazamiento	lugar población país nombre propio

Figura 10: *Combinatoria biargumental con filtro de selección para ambos argumentos*

Para visualizar los ejemplos generados automáticamente se clica encima de la combinatoria biargumental seleccionada. La Figura 11 muestra ejemplos para el patrón argumental [determinante + VIAJE + de +determinante +N:{condición humana desplazamiento} a + N: {país nombre propio}], seleccionadas previamente (Figuras 9 y 10):

frases generadas
El viaje de las refugiadas a Bolivia
El viaje de las expatriadas a América
El viaje de los apátridas a Argelia
El viaje de los personas desplazadas a Armenia
El viaje de las exiliadas a Argelia
El viaje de los desplazados a Balcanes
El viaje de las exiliadas a Antigua
El viaje de las asiladas a Bután
El viaje de los refugiados a Canadá
El viaje de las peregrinas a Canadá
El viaje de los deportados a Bulgaria
El viaje de las peregrinas a Bielorrusia
El viaje de los desplazados a Armenia
El viaje de las exiliadas a Afganistán

Figura 11: *Ejemplos de combinatoria biargumental con filtro de selección [vista parcial]*

CombiContext es el simulador que proporciona contexto frasal y oracional a las frases generadas automáticamente por *Xera* y *Combinatoria*. Este nuevo generador aplica en primer lugar un filtro distribucional: la posición de la frase nominal con respecto al verbo (antes o después del mismo). Tras aplicar este filtro, la herramienta presenta diferentes estructuras formales, esto es, [*viaje* + a], [*viaje* + hacia], etc., a las que acompañan ejemplos estándar (Figura 12):

Figura 12: *Realizaciones de viaje y sus argumentos en posición preverbal*

Si se selecciona, por ejemplo [*viaje* +de] (Figura 13), se obtiene un desplegable con esta realización y todas las posibles combinaciones oracionales, siguiendo, por tanto, el mismo tipo de acceso a la información que el generador *Combinatoria*. De este modo, se enmarca la frase nominal compleja en el contexto oracional y frasal (Figura 14), pudiéndose obtener ejemplos *ad libitum*. En este estadio, se puede indicar si se desea obtener datos de consulta filtrados con *word2vec* (Figura 13; vid. Bardanca Outeiriño y Domínguez Vázquez en este volumen).

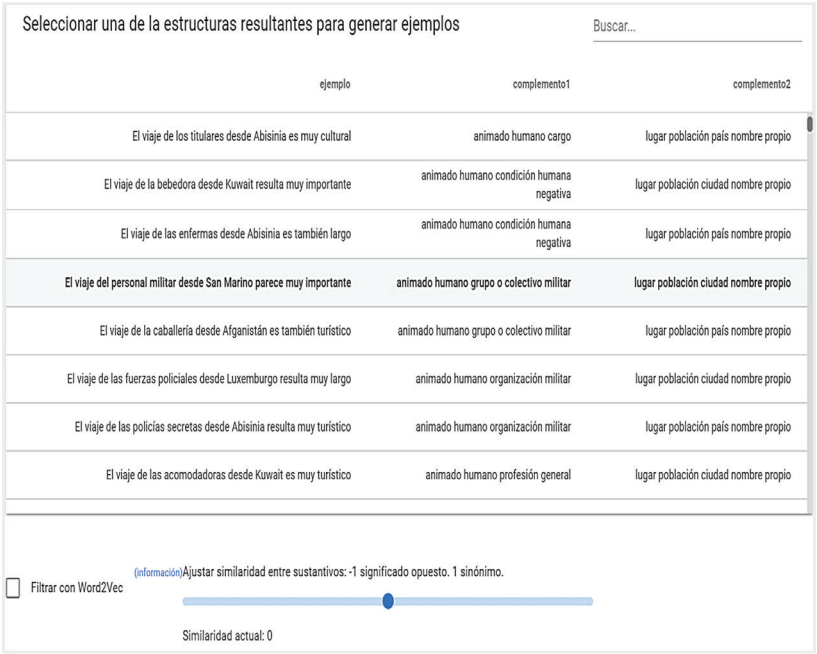


Figura 13: *Combinatoria oracional*

El viaje de los cuerpos desde Berlín es muy largo
El viaje de la tropa desde Jerusalén parece muy largo
El viaje del cuerpo desde Argel es muy turístico
El viaje de la guardia pretoriana desde Maracaibo parece muy largo
El viaje de las formaciones militares desde Tartu resulta también cultural
El viaje de los equipos de salvamento desde Buenos Aires es también turístico
El viaje de los destacamentos desde Maracaibo es muy importante
El viaje de los cuerpos desde Tokio parece también turístico
El viaje de los altos mandos desde Cork parece muy importante
El viaje del cuerpo desde Roterdam resulta también turístico

Figura 14: *Desplegable de ejemplos generados de combinatoria oracional [vista parcial]*

5. A MODO DE RESUMEN

Como se señalaba previamente, los generadores cumplen el objetivo para el que han sido diseñados. De este modo, ha sido posible verificar la validez del método aplicado y sus principales potencialidades, al mismo tiempo que se han detectado las limitaciones del propio modelo y las posibles optimizaciones. Sin lugar a duda, abren una puerta a un buen número de aplicaciones, algunas de las cuales se da cuenta en esta monografía, como son sus aplicaciones didácticas (vid. López Iglesias en este volumen) y contrastivas (Domínguez Vázquez & Caíña Hurtado, 2021; vid. también Pino Serrano y Valcárcel Riveiro en este tomo), por citar algunas.

REFERENCIAS BIBLIOGRÁFICAS

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

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. https://doi.org/10.1162/tacl_a_00051

Díaz Hormigo, M.ª T. (2002). Sintaxis y semántica de la construcción con sustantivo en posición nuclear. *Estudios de Lingüística Española*, *16*. http://elies.rediris.es/elies16/

Domínguez Vázquez, M.ª J. (2011). *Kontrastive Grammatik und Lexikographie: spanisch-deutsches Wörterbuch zur Valenz des Nomens*. Iudicium.

Domínguez Vázquez, M.ª J. (2021). Zur Darstellung eines mehrstufigen Prototypbegriffs in der multilingualen automatischen Sprachgenerierung: vom Korpus über word embeddings bis zum automatischen Wörterbuch. *Lexikos*, *31*, 20-50. https://doi.org/10.5788/31-1-1623

Domínguez Vázquez, M.ª J. (2022a). Estructura argumental del nombre: generación automática. *Revista Signos. Estudios de Lingüística*, *55*(110), 732-761. https://doi.org/10.4067/S0718-09342022000300732

Domínguez Vázquez, M.ª J. (2022b). Contribución de la semántica combinatoria al desarrollo de herramientas digitales multilingües. *Círculo de Lingüística Aplicada a la Comunicación*, *90*, 171-18.

Domínguez Vázquez, M.ª 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 conference* (pp. 269-287). Lexical Computing CZ. https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_16_pp269-287.pdf

Domínguez Vázquez, M.ª J. & Caíña Hurtado, M. (2021). Aplicación de recursos de xeración automática da lingua para estudos comparativos. *Estudos de Lingüística Galega, 130*, 139-172. https://doi.org/10.15304/elg.13.7409

Domínguez Vázquez, M.ª J. & Gouws, R. (2023). The definition, presentation, and automatic generation of contextual data in lexicography. *International Journal of Lexicography*, 1-27. https://doi.org/10.1093/ijl/ecac020

Domínguez Vázquez, M.ª J., Solla Portela, M. A. & Valcárcel Riveiro, C. (2019). Resources interoperability: Exploiting lexicographic data to automatically generate dictionary examples. En I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek & C. Tiberius (eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference* (pp. 51-71). Lexical Computing CZ. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_4.pdf

Domínguez Vázquez, M.ª J. & Valcárcel Riveiro, C. (2020). PORTLEX as a multilingual and cross-lingual online dictionary. En M.ª J. Domínguez Vázquez, M. Mirazo Balsa & C. Valcárcel Rivero (eds.), *Studies on Multilingual Lexicography* (pp. 135-158). De Gruyter. https://doi.org/10.1515/9783110607659-008

Engel, U. (1996). Semantische Relatoren. Ein Entwurf für künftige Valenzwörterbücher. En N. Weber (ed.), *Semantik, Lexikographie und Computeranwendungen* (pp. 223-236). Niemeyer. https://doi.org/10.1515/9783111555522.223

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

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). Gunter Narr Verlag.

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. 1-12). Conference Track Proceedings. https://arxiv.org/pdf/1301.3781.pdf

Müller-Spitzer, C., Domínguez Vázquez, M.ª J., Nied Curcio, M., Silva Dias, M. & Wolfer, S. (2018). Correct Hypotheses and Careful Reading Are Essential: Results of an Observational Study on Learners Using Online Language Resources. *Lexikos, 28*, 287-315. https://doi.org/10.5788/28-1-1466

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

Valcárcel Riveiro, C. (2017). Las construcciones N1N2 como realizaciones actanciales del sustantivo en francés y su tratamiento en el diccionario multilingüe PORTLEX. En M.ª J. Domínguez Vázquez & S. Kutscher (eds.), *Estudios contrastivos y multicontrastivos: Interacción entre gramática, didáctica y lexicografía* (pp. 193-207). De Gruyter. https://doi.org/10.1515/9783110420784-015

Valcárcel Riveiro, C. & Pino Serrano, L. (2023). Application d'une méthodologie d'analyse des prédicats nominaux: l'exemple du lexème MORT1. *Çédille. Revista de estudios franceses*, 24 (en prensa).

Recursos

CombiContext = Domínguez Vázquez, M.ª J. (dir.), Valcárcel Riveiro, C., Bardanca Outeiriño, D., Calañas Continente, J. A., Catalá Torres, N., Martín Gascueña, R., Mirazo Balsa, M., Sanmarco Bande, M. T. & Pino Serrano, L. (2021). *CombiContext. Prototipo online para la generación automática de contextos frasales y oraciones de la frase nominal en alemán, español y francés*. Universidade de Santiago de Compostela. Instituto da Lingua Galega. Consultado el 30 de mayo de 2023. http://portlex.usc.gal/combinatoria/verbal

Combinatoria = Domínguez Vázquez, M.ª J. (dir.), Valcárcel Riveiro, C., Bardanca Outeiriño, D., Calañas Continente, J. A., Catalá Torres, N., López Iglesias, N., Martín Gascueña, R., Mirazo Balsa, M., Sanmarco Bande, M. T. & Pino Serrano, L. (2020). *Combinatoria. Prototipo online para la generación biargumental de la frase nominal en alemán, español y francés*. Universidade de Santiago de Compostela. Consultado el 30 de mayo de 2023. http://portlex.usc.gal/combinatoria/usuario

E-Valbu = *Elektronisches Valenzwörterbuch deutscher Verben*. Consultado el 30 de mayo de 2023. https://grammis.ids-mannheim.de/verbvalenz

Ontología léxica = Domínguez Vázquez, M. J., Valcárcel Riveiro, C. & Bardanca Outeiriño, D. (2021). *Ontología léxica*. Universidade de Santiago de Compostela. Consultado el 30 de mayo de 2023. http://portlex.usc.gal/ontologia/

Portlex = Domínguez Vázquez, M.ª J. (dir.), Valcárcel Riveiro, C., Mirazo Balsa, M., Sanmarco Bande, M.ª T., Simões, A. & Vale, M. J. (2018). *Portlex. Diccionario multilingüe de la*

valencia del nombre. Universidade de Santiago de Compostela. Consultado el 30 de mayo de 2023. http://portlex.usc.gal/portlex/

Traduword = Consultado el 30 de mayo de 2023. https://ilg.usc.gal/es/proxectos/interoperabilidad-de-recursos-y-produccion-automatica-de-lenguaje-natural-0

WordNet = *WordNet*. Princeton University. Consultado el 30 de mayo de 2023. https://wordnet.princeton.edu/

Xera = Domínguez Vázquez, M.ª J. (dir.), Valcárcel Riveiro, C., Bardanca Outeiriño, D., Calañas Continente, J. A., Catalá Torres, N., López Iglesias, N., Martín Gascueña, R., Mirazo Balsa, M., Sanmarco Bande, M.T. & Pino Serrano, L. (2020). *Xera. Prototipo online para la generación automática monoargumental de la frase nominal en alemán, español y francés*. Universidade de Santiago de Compostela. Consultado el 30 de mayo de 2023. http://portlex.usc.gal/combinatoria/usuario

XeraWord = Domínguez Vázquez, M.ª J. (dir.), Bardanca Outeiriño, D., Caíña Hurtado, M., Gómez Guinovart, X., Iglesias Allones, J. J., Simões, A., Valcárcel Riveiro, C., Álvarez de la Granja, M. & Cidrás Escaneo, F. A. (2020). *XeraWord. Prototipo de xeración automática da argumentación da frase nominal en galego e portugués*. Santiago de Compostela: Instituto da Lingua Galega. Consultado el 30 de mayo de 2023. http://ilg.usc.gal/xeraword/

SOBRE LA ESTRUCTURA DE LOS SINTAGMAS NOMINALES

ON THE STRUCTURE OF NOMINAL PHRASES

Natália Català Torres
Universitat Rovira i Virgili
natalia.catala@urv.cat

Resumen

Con el objetivo de situar los proyectos *MultiGenera*[1] y *MultiComb*[2] en un contexto teórico algo distinto del que nos ofrece la gramática valencial, nos aproximamos a los debates que han suscitado las nociones de estructura argumental y estructura eventiva en el ámbito del sintagma nominal.

En este trabajo asumimos, por una parte, que el aspecto léxico, que inicialmente se consideró una propiedad de los verbos y las nominalizaciones, resulta también pertinente para evaluar el comportamiento de los adjetivos y de los nombres que no tienen relación morfológica con raíces verbales. Por otra, contemplamos la hipótesis de que los análisis de los sustantivos deverbales y de otros tipos de sustantivos pueden unificarse recurriendo, en todos los casos, al examen de tres propiedades relevantes: delimitación, duratividad y causación.

Ilustramos estas cuestiones con el examen de algunos sintagmas nominales, incluidos en los proyectos mencionados, que confirman las hipótesis iniciales respecto a la clasificación de los sustantivos, aunque reconocemos la necesidad de profundizar en algunos aspectos relacionadas con las diferencias entre los cambios de estado y los cambios de posición.

Palabras clave: estructura argumental, estructura eventiva, sintagma nominal, sustantivos deverbales, sustantivos no deverbales, aspecto léxico.

Abstract

In order to situate the *MultiGenera* and *MultiComb* projects in a theoretical context somewhat different from the one offered by valency grammar, we approach the debates that the notions of argument structure and event structure have raised in the field of noun phrases.

In this paper we assume, on the one hand, that the lexical aspect, which was initially considered a property of verbs and nominalizations, is also relevant for evaluating the behavior of adjectives and nouns that are not morphologically related to verbal roots. On the other hand, we consider the hypothesis that the analysis of deverbal nouns and other types of nouns can be unified by examining three relevant properties in all cases: delimitation, durativity and causation.

We illustrate these issues by examining some noun phrases, included in the above-mentioned projects, which confirm the initial hypotheses regarding the classification of nouns, although we recognize the need to go deeper into some aspects related to the differences between changes of state and changes of position.

Keywords: argument structure, event structure, noun phrase, deverbal nouns, non deverbal nouns, lexical aspect.

[1] *MultiGenera. Generación multilingüe de estructuras argumentales del sustantivo y automatización de extracción de datos sintáctico-semánticos*. Fundación BBVA. Ayudas Fundación BBVA a Equipos de Investigación Científica - Humanidades Digitales. 2017-2020. http://portlex.usc.gal/multigenera/

[2] *MultiComb. Generador multilingüe de estructuras argumentales del sustantivo con aplicación en la producción en lenguas extranjeras*. FI2017-82454-P: Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia, Generación de Conocimiento. MCIN/ AEI/ FEDER "Una manera de hacer Europa" (EXCELENCIA 2017, 2017-PN091). 2018-2021. http://portlex.usc.gal/multicomb/

DOI: https://doi.org/10.17561/rilex.6.3.8192

1. APROXIMACIÓN TEÓRICA

En la actualidad parece haberse llegado a cierto consenso en torno a la idea de que el concepto de predicación no debe restringirse a los verbos, sino que debe extenderse a otras categorías como los nombres y los adjetivos, lo que implica la existencia de una estructura argumental en los sintagmas nominales y adjetivales y, por tanto, temática, con condiciones de asignación de papeles temáticos similares a las de los sintagmas verbales.

Pero, a pesar de ese consenso, dado que la mayor parte de los estudios se han centrado en las nominalizaciones deverbales, especialmente en las nominalizaciones que expresan eventos, participantes, estados (Fábregas Alfaro & Marín Gálvez, 2012) o cualidades (Pena Seijas, 2004; Arche & Marín Gálvez, 2015; Zato, 2020), el análisis del potencial combinatorio de otros sustantivos no está adecuadamente representado en la literatura científica (Domínguez Vázquez & Mirazo Balsa, 2017), aunque contamos con algunos trabajos que, como el de Escandell Vidal (1995), parten del análisis de las diferentes realizaciones sintácticas de los argumentos y adjuntos del sintagma nominal o que, como el de Fábregas Alfaro (2014), se centran en la realización de los argumentos nominales como sintagmas preposicionales, y, en concreto, sin negar la heterogeneidad del conjunto, en aquellas realizaciones que recurren al SP *de* y que suelen asociarse con el caso genitivo[3].

1.1. *Sobre los tipos de nominalización*

Grimshaw (1990) distinguía tres tipos de nombres en función de su denotación: sustantivos que denotan un evento complejo, sustantivos que denotan un evento simple, y sustantivos que denotan el resultado de una acción.

(1) a. La destrucción de la ciudad por el ejército fue muy rápida

b. La llegada del ejército tuvo lugar ayer

c. El examen está sobre la mesa

[3] Badia Cardús (2002) propone las etiquetas *genitivo subjetivo* para referirse a los argumentos externos de los nombres y *genitivo objetivo* para referirse a los argumentos internos.

En (1a) aparece un nombre procedente de un verbo transitivo y los complementos de ese nombre son el argumento interno y el argumento externo de ese verbo; en (1b) aparece un nombre procedente de un verbo inacusativo y el complemento de ese nombre es el argumento externo del verbo; y en (1c) aparece un nombre relacionado con un verbo transitivo sin complementos.

Para la autora, esta diferencia está relacionada con la capacidad de selección de argumentos: tan solo los nominales eventivos complejos seleccionan argumentos. Picallo i Soler (1999), Alexiadou (2001), o Badia Cardús (2002), entre otros, argumentan, en cambio, que tanto los sustantivos eventivos como los resultativos tienen la capacidad de seleccionar argumentos.

1.2. *Sobre las propiedades de las nominalizaciones*

Para establecer las características de los tipos de nominalización, se ha recurrido a criterios[4] de distinta naturaleza. Algunos hacen referencia a cuestiones morfosintácticas; otros, en cambio, a cuestiones sintáctico-semánticas:

a) La clase de verbo de la que deriva el sustantivo:

Picallo i Soler (1999) y Alexiadou (2001) mantienen que los verbos inergativos[5] dan lugar siempre a sustantivos resultativos (2a), mientras que los inacusativos[6] originan sustantivos ambiguos (2b):

(2) a. El viaje de Juan

b. La llegada de Pedro

Respecto a los verbos transitivos, Picallo i Soler (1999) sostiene que de ellos pueden derivar tanto nominales resultativos (3a) como eventivos (3b) o tener una interpretación ambigua:

(3) a. La declaración del testigo se extravió

[4] Aunque estos criterios permiten identificar algunas características de los distintos tipos de nominales, es relativamente sencillo encontrar contraejemplos que cuestionan su pertinencia.

[5] Verbo intransitivo que denota una acción controlada por un agente.

[6] Verbo intransitivo que denota un evento que afecta a una entidad.

b. La declaración de intenciones por parte del testigo fue confusa

b) La presencia del equivalente nominal al argumento interno en las nominalizaciones eventivas:

(4) La destrucción de las pruebas se consideró probada

c) La presencia de un complemento preposicional equivalente al agente en las nominalizaciones eventivas:

(5) La destrucción de las pruebas por parte de la policía se consideró probada

Picallo i Soler (1999) afirma que un complemento agentivo introducido por la preposición *por* o la locución prepositiva *por parte de* implica una lectura eventiva, mientras que un complemento introducido por la preposición *de* implica una lectura resultativa del nominal:

(6) La traducción de Kafka por Borges/La traducción de Kafka de Borges

Los nombres derivados de verbos volitivos y psicológicos no pueden ser considerados eventivos, ya que su argumento externo es un experimentador y no un agente:

(7) La preocupación de los ciudadanos por la economía/*La preocupación por parte de los ciudadanos por la economía

d) La compatibilidad con los predicados del tipo *tener lugar/ocurrir* con los nominales eventivos (Grimshaw, 1990; Picallo i Soler, 1999):

(8) La invasión tuvo lugar en febrero

e) La compatibilidad de los nominales eventivos con modificadores de tiempo y aspecto (sintagmas preposicionales) o adjetivos adverbiales de tiempo y aspecto:

(9) a. La construcción del puente durante la guerra

b. La probable construcción del puente

f) La compatibilidad de los nombres resultativos con la marca de plural:

(10) Los bombardeos de Ucrania/*Las destrucciones de la ciudad

g) La compatibilidad de los nombres resultativos con todo tipo de determinantes:

(11) Estas/algunas traducciones son impresentables

h) La presencia de un modificador intencional como indicador de una interpretación eventiva (Grimshaw, 1990):

(12) La destrucción deliberada de las pruebas

j) La compatibilidad de las estructuras de control en oraciones finales de infinitivo con los nominales eventivos (Grimshaw, 1990; Picallo i Soler, 1991, 1999):

(13) La publicación de los documentos para exculpar al acusado/*La publicación semanal para exculpar al acusado

a) La derivación de los sustantivos resultativos de verbos transitivos atélicos[7] frente a la derivación de los sustantivos eventivos de verbos transitivos télicos (Alexiadou, 2001):

(14) Hizo posible la construcción de la casa de sus sueños en un mes/*Ordenó la destrucción de la presa durante un mes

Algunos autores (Fábregas Alfaro & Marín Gálvez, 2012) señalan, además, la diferencia que parece existir entre la clase de los nombres procedentes de verbos volitivos y psicológicos y otras clases:

(15) a. La construcción del puente tuvo lugar durante la guerra

b. La constante construcción de puentes

(16) a. *La preocupación de John por la economía tuvo lugar el verano pasado

b. La constante preocupación de John por la economía

(17) a. *La construcción de piedra tuvo lugar en el siglo XVI

[7] Si la noción expresada tiene un límite temporal es *télica*; si no lo tiene, es *atélica*.

(18) b. *El constante examen de matemáticas

El rango de nominalizaciones que permite un verbo y las propiedades de cada clase de nominalización parecen determinadas en parte por la estructura aspectual del verbo base.

1.3. *Sobre los sustantivos no derivados*

Bosque (1999) incluye en la clase de los sustantivos eventivos nombres no deverbales como *conferencia* y señala, como propiedades de esta clase de sustantivos, algunas de las propuestas para las nominalizaciones: pueden ser sujetos de verbos del tipo *tener lugar* o de otros que se refieren a los límites del evento (*empezar*, *durar*, *concluir*) pueden ser complementos de la preposición *durante* y de los adverbios *antes* y *después*:

(19) a. La conferencia empezó a las tres

b. Llegó después de la conferencia

También el trabajo de Resnik (2010), centrado en el estudio de los nombres eventivos no deverbales, defiende la idea de que la eventividad no es una propiedad exclusiva de las nominalizaciones. La autora asume la existencia de distintos tipos aspectuales entre los nombres que no tienen relación morfológica con el verbo, basándose en tres valores que parecen sintácticamente relevantes: delimitación, duratividad y causación.

En español, para identificar eventos delimitados, suelen utilizarse tests que evalúan la compatibilidad de algunas unidades léxicas con modificadores del tipo *en una hora/un mes* y su incompatibilidad con modificadores durativos como *durante*. Pero esa compatibilidad con modificadores durativos no es, para Resnik (2010), señal de telicidad, sino de duratividad, por lo que nos permite distinguir las realizaciones de los logros, pero no de las actividades o los estados. La duración del evento es independiente de la telicidad: construir una casa y descubrir la clave son eventos télicos, pero sólo el primero es durativo. Nombres como *huelga* y *motín* son télicos (realizaciones) y admiten modificadores como *una hora/un mes* porque indican duratividad

y no atelicidad: las realizaciones, los estados y las actividades admiten estos modificadores, pero los rechazan los logros:

(20) a. Dos horas de conferencia/Un año de dolor/Un minuto de silencio

b. *Dos minutos de accidente

En el análisis de la estructura eventiva denotada por los verbos, se suele considerar también la causación un parámetro relevante y también lo es en la determinación de tipos aspectuales de nombres eventivos no deverbales, aunque, en el contexto de la sintaxis nominal, Resnik (2010) se limita a constatar este valor en los casos en que hay agentividad, es decir, en los eventos controlados[8]. Por ejemplo, el predicado *planear* es compatible con los eventos causados, pero no con los eventos no causados:

(21) a. Planearon una excursión/una cita

b. *Planearon una crisis/un terremoto

Por el contrario, predicados como *ocurrir* o *suceder* seleccionan como argumento interno eventos no causados:

(22) a. Ocurrió un accidente

b. *Ocurrió una excursión

Los nombres eventivos simples que tienen causación externa pueden controlar cláusulas finales, mientras que los eventos sin causación no pueden controlar este tipo de cláusulas:

(23) a. La fiesta del pueblo para celebrar el fin de la sequía

b. *La rabia del pueblo para recuperar su poder adquisitivo

Resnik (2010) propone así cuatro clases de nombres no deverbales se corresponden con los cuatro tipos aspectuales de la clasificación de Vendler (1967):

[8] En el concepto amplio de causación, se incluyen también los casos en que hay una causa interna que no corresponde a un agente.

1. Actividades: presentan -telicidad, +durativiдad y +causación: *concierto, guerra, conferencia…*
2. Realizaciones: presentan +telicidad, +durativiдad y +causación: *motín*, *boicot*, *huelga…*
3. Estados: presentan -telicidad, +durativiдad y -causación: *rabia*, *pánico*, *crisis…*
4. Logros: presentan +telicidad, -durativiдad y -causación: *accidente*, *terremoto*, *desastre…*

2. SOBRE LOS SINTAGMAS NOMINALES DE *MULTIGENERA* Y *MULTICOMB*

Las estructuras nominales que aparecen en nuestros proyectos incluyen los siguientes sustantivos: *color, dolor, sabor, olor, ausencia, presencia, discusión, conversación, aumento, texto, muerte, huída, pregunta, respuesta, amor, estancia, mudanza, viaje, video, ancho*. En este trabajo nos centraremos en el examen de solo algunas de ellas:

MUERTE

(24) a. La lamentable muerte de la princesa Diana de Gales

b. La muerte de un soldado de/por envenenamiento

c. La dolorosa muerte paterna

Estas secuencias con un sustantivo derivado de un verbo inacusativo admiten sintagmas preposicionales con *de* que apuntan al único argumento del verbo. También admiten sintagmas preposicionales con *de/por* que nos remiten a una causa no controlada. La presencia de un adjetivo complementando al sustantivo también nos remite al único argumento del verbo del que procede.

De acuerdo con la clasificación de Resnik (2010), nos encontraríamos frente a un logro, ya que el sustantivo tiene los rasgos +telicidad, -durativiдad y -causación[9].

[9] Como ya hemos dicho, el concepto de causa que se contempla aquí es una causa controlada.

VIAJE

(25) a. El próximo viaje del presidente a Berlín

b. El viaje paterno al pueblo

c. El pesado viaje de Carlos desde Australia

Este sustantivo admite sintagmas preposicionales con *de* que apuntan a un argumento externo[10] y sintagmas preposicionales con *a* y con *de/desde* que nos remiten, respectivamente, al lugar de destino (Meta) y de procedencia (Origen).

De acuerdo con la clasificación de Resnik (2010), nos encontraríamos frente a una actividad, ya que el sustantivo tiene los rasgos -telicidad, +duratividad y +causación.

El sustantivo HUÍDA tendría un comportamiento similar.

CONVERSACIÓN

(26) a. Las frecuentes conversaciones de los hermanos

b. Las conversaciones entre las familias reales

c. La inevitable conversación conyugal

d. Las triviales conversaciones con el mecánico

e. La interminable conversación sobre los estatutos

f. La conversación antropológica

Los sintagmas preposicionales con *de* que aparecen con este sustantivo apuntan al argumento externo del verbo, y, en este caso, alternan con el sintagma preposicional encabezado por la preposición *entre*. También admite un adjetivo que hace referencia a ese argumento externo.

Además, pueden aparecer otros sintagmas preposicionales que remiten a los argumentos internos del sustantivo: sintagmas con la preposición *sobre*

[10] La presencia de un adjetivo complementando al sustantivo también nos remite al argumento externo.

que apuntan al Tema de la predicación y sintagmas con la preposición *con* que nos remiten al rol semántico de Compañía.

Puede aparecer también un adjetivo relacionado con el Tema.

De acuerdo con la clasificación de Resnik (2010), nos encontraríamos frente a un sustantivo que tiene los rasgos +/-telicidad, +durativIdad y +causación. Como puede verse, el rasgo telicidad no parece, en principio, determinante, puesto que el verbo del que deriva puede considerarse tanto télico como atélico, es decir, puede manifestarse como actividad o como realización.

El sustantivo DISCUSIÓN comparte estas características.

AUSENCIA

(27) a. La injustificable ausencia de los trabajadores

b. La necesaria ausencia paterna

c. La inesperada ausencia de María del baile

Los sintagmas preposicionales con *de* que aparecen con este sustantivo apuntan o al argumento externo que está en la situación denotada[11] o al lugar de esta situación (Ubicación).

De acuerdo con la clasificación de Resnik (2010), nos encontraríamos frente a un sustantivo que tiene los rasgos -telicidad, +durativIdad y -causación. Se trataría, por tanto, de un estado.

El sustantivo PRESENCIA comparte estas características, aunque el sintagma preposicional que denota el lugar no requeriría la preposición *de*, sino la preposición *en*.

DOLOR

(28) a. Un insoportable dolor de cabeza

b. El persistente dolor menstrual

c. El dolor de Clara

Los sintagmas preposicionales con *de* que aparecen con este sustantivo apuntan al argumento externo (Tema) del verbo *doler*, que también admite un

[11] También admite un adjetivo que hace referencia a ese argumento.

adjetivo que hace referencia a ese argumento externo. El argumento interno (Experimentador) puede aparecer en la estructura.

De acuerdo con la clasificación de Resnik (2010), nos encontraríamos frente a un sustantivo que tiene los rasgos -telicidad, +durativiad y -causación, por tanto, frente a un estado. En este caso, la peculiaridad estructural procedería de la presencia de un argumento con el papel temático de Experimentador que compite y convive con el papel temático de Tema[12]:

(29) El insoportable dolor de cabeza de Clara

3. A MODO DE CONCLUSIÓN

La mayor parte de las investigaciones parecen haber prestado más atención al potencial combinatorio de las nominalizaciones deverbales que al potencial combinatorio de otros sustantivos. Los datos, sin embargo, parecen avalar la idea de que la eventividad no es exclusiva de las nominalizaciones, por lo que una clasificación basada en aspectos como delimitación, durativiad y causación podría ser adecuada en otros contextos.

Propuestas como la de Resnik (2010) nos permiten establecer generalizaciones sintácticas avaladas por las estructuras nominales generadas y nos garantizan un tratamiento bastante uniforme de la diversidad nominal, aunque, por supuesto, habría que resolver algunas cuestiones: a pesar de la importancia de la distinción entre los eventos que suponen un cambio y los que no, las posibles implicaciones sintácticas de las diferencias entre cambios de estado y cambios de posición todavía no se han examinado en profundidad

Cabe preguntarse también si una aproximación de estas características podría facilitar una explicación más sistemática de la variación morfológica.

REFERENCIAS BIBLIOGRÁFICAS

Alexiadou, A. (2001). *Functional Structure in Nominals: Nominalization and Ergativity*. John Benjamins. https://doi.org/10.1075/la.42

[12] Entre ambos existiría una relación de pertenencia/parte-todo.

Arche, M.ª J. & Marín Gálvez, R. (2015). On the edge: Nominalizations from Evaluative Adjectives in Spanish. En J. Smith & T. Isane (eds.), *Romance Linguistics 2012: Selected papers from the 42nd Linguistic Symposium on Romance Languages* (pp. 261–274). John Benjamins. https://doi.org/10.1075/rllt.7.17arc

Badia Cardús, T. (2002). Els complements nominals. En J. Solà Cortassa, M.ª R. Lloret, J. Mascaró & M. Pérez Saldanya (dirs.), *Gramàtica del català contemporani*, vol. 2 (pp. 1591-1640). Empúries.

Bosque, I. (1999). Sustantivos eventivos. En I. Bosque & V. Demonte (eds.), *Gramática descriptiva de la lengua española*. Tomo1 (pp. 51-53). Espasa Calpe.

Domínguez Vázquez, M.ª José & Mirazo Balsa, M. (2017). Aproximación multilingüe a los argumentos oracionales del sustantivo. En M.ª J. Domínguez & S. Kutscher (eds.), *Interacción entre gramática, didáctica y lexicografía: Estudios contrastivos y multicontrastivos* (pp. 353-368). De Gruyter. https://doi.org/10.1515/9783110420784-026

Escandell Vidal, M.ª V. (1995). *Los complementos del nombre*. Arco Libros.

Fábregas Alfaro, A. (2014). Los genitivos múltiples en español: restricciones léxicas y sintácticas. *Lexis*, *XXXVIII*(2), 269-306. https://doi.org/10.18800/lexis.201402.002

Fábregas Alfaro, A. & Marín Gálvez, R. (2012). The role of Aktionsart in deverbal nouns: State nominalizations across languages. *Journal of Linguistics*, *48*, 35-70. https://doi.org/10.1017/S0022226711000351

Grimshaw, J. (1990). *Argument Structure*. The MIT Press.

Pena Seijas, J. (2004). Morfología de los nombres de cualidad derivados. *Verba*, *31*, 7-42.

Picallo i Soler, C. (1991). Nominals and Nominalizations in Catalan. *Probus*, *3*(3), 279-316. https://doi.org/10.1515/prbs.1991.3.3.279

Picallo i Soler, C. (1999). La estructura del sintagma nominal: las nominalizaciones y otros sustantivos con complementos argumentales. En I. Bosque & V. Demonte (eds.), *Gramática descriptiva de la lengua española*. Tomo 3 (pp. 4367- 4422). Espasa Calpe.

Resnik, G. (2010). *Los nombres eventivos no deverbales en español* [Tesis doctoral. Universitat Pompeu Fabra]. Repositorio Digital de la UPF. http://hdl.handle.net/10803/22647

Vendler, Z. (1967). *Linguistics in Philosophy*. Cornell University Press. https://doi.org/10.7591/9781501743726

Zato, Z. (2020). *The role of state-kinds in the morphosemantics of Spanish deadjectival nominalizations* [Tesis doctoral. Universidad del País Vasco]. Archivo digital docencia investigación. ADDI. http://hdl.handle.net/10810/51257

GUÍA DE TÉCNICAS, ESTRATEGIAS Y HERRAMIENTAS EN EL DISEÑO Y DESARROLLO DE GENERADORES AUTOMÁTICOS DEL LENGUAJE

GUIDE OF TECHNIQUES, STRATEGIES AND TOOLS INVOLVED IN THE DESIGN AND DEVELOPMENT OF AUTOMATIC LANGUAGE GENERATORS

Daniel Bardanca Outeiriño
Universidade de Santiago de Compostela
danielbardanca.outeirino@usc.es

María José Domínguez Vázquez
Universidade de Santiago de Compostela
majo.dominguez@usc.es

Resumen

Este capítulo aborda la descripción de diferentes técnicas, métodos y herramientas desarrolladas y aplicadas para el diseño de la cadena de generadores descrita en el capítulo 1 de este volumen. Tanto el método combinado como los generadores y el conjunto de recursos que los soportan se asientan en principios de sostenibilidad, interoperabilidad y retroalimentación de datos.

Palabras clave: generadores automáticos del lenguaje natural, WordNet, ontología, significado relacional y categorial.

Abstract

This chapter focuses on the explanation of the different techniques, methods, and tools applied during the development of the generators described in chapter 1 of this volume. The combining methodology, generators and resources that support them are based on principles of sustainability, interoperability, and data feedback.

Keywords: automatic generators, WordNet, ontology, relational and categorial meaning.

DOI: https://doi.org/10.17561/rilex.6.3.8087

1. INTRODUCCIÓN

Las herramientas diseñadas al abrigo de los proyectos *MultiGenera*[1], *MultiComb*[2] y *XeraWord*[3] persiguen finalidades diversas ligadas a diferentes fases de trabajo o *workflow*. Se pueden agrupar, por tanto, atendiendo al objetivo final para el que han sido concebidas (Figura 1)[4]:

a) Herramientas para la investigación lingüística,
b) Herramientas de revisión y corrección en la Intranet,
c) Herramientas de generación automática del lenguaje o generadores,
d) Herramientas de difusión de los propios generadores y actividades relacionadas con los proyectos,
e) Herramientas de aplicación didáctica.

El estudio se articula como sigue: en el capítulo 2 se contextualiza el conjunto de recursos aplicados en el desarrollo de los prototipos de generación automática. El capítulo 3 describe las herramientas y métodos de investigación y generación en consonancia con las diferentes fases de trabajo. Una evaluación de la cadena de generadores se aporta en el capítulo 4. El capítulo 5 sirve a modo de conclusión.

---

[1] *MultiGenera. Generación multilingüe de estructuras argumentales del sustantivo y automatización de extracción de datos sintáctico-semánticos*. Fundación BBVA. Ayudas Fundación BBVA a Equipos de Investigación Científica - Humanidades Digitales. 2017-2020. http://portlex.usc.gal/multigenera/

[2] MultiComb.*Generador multilingüe de estructuras argumentales del sustantivo con aplicación en la producción en lenguas extranjeras*. FI2017-82454-P: Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia, Generación de Conocimiento. MCIN/AEI/ FEDER "Una manera de hacer Europa" (EXCELENCIA 2017, 2017-PN091). 2018-2021. http://portlex.usc.gal/multicomb/

[3] *Ferramentas TraduWord e XeraWord: tradución de caudal léxico e xeración automática da linguaxe natural en galego e portugués*. 2020-PU004. Convocatoria proyectos de colaboración. Universidade de Santiago de Compostela. https://ilg.usc.gal/xeraword/

[4] Este capítulo compendia y detalla algunas de las herramientas recogidas en Domínguez Vázquez, Solla Portela & Valcárcel Riveiro (2019) y Domínguez Vázquez, Bardanca Outeiriño & Simões, (2021), pero también herramientas manejadas evolucionadas en el proyecto Proyecto ESMAS-ES+ (PID2022-137170OB-I00) financiado por MCIN/AEI//FEDER "Una manera de hacer Europa".

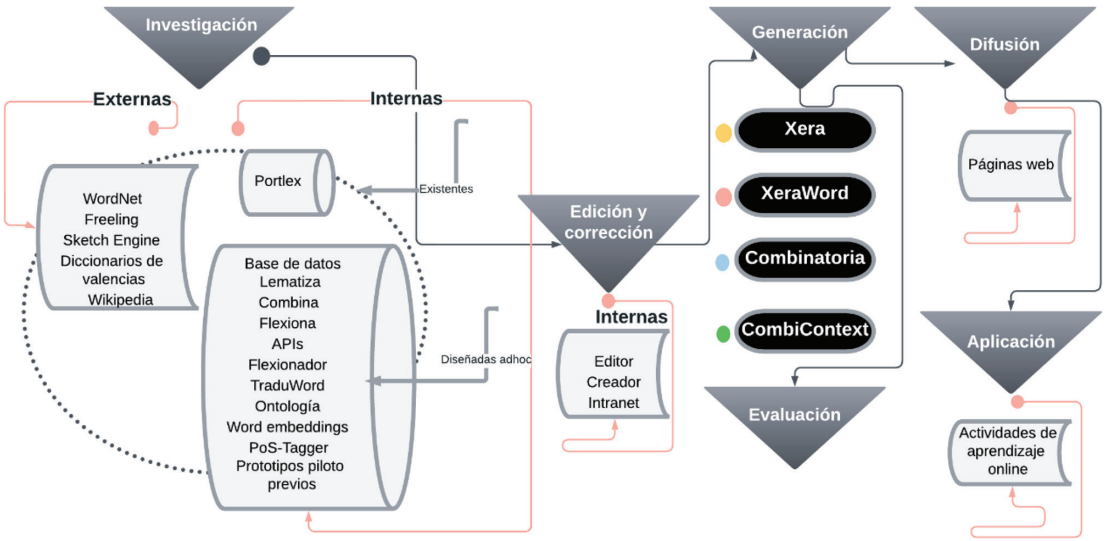


FIGURA 1: *Herramientas y etapas implicadas en el flujo de trabajo*

2. MÉTODO Y TIPOLOGÍA DE HERRAMIENTAS: UNA VISIÓN DE CONJUNTO

Atendiendo a los principios de sostenibilidad, interoperabilidad y retroalimentación y al propio objeto de estudio se ha diseñado un método combinado que la propia evolución de la investigación ha verificado como adecuado. Dicho método aúna principios de la gramática de valencias, la teoría de los prototipos léxicos y las clases semánticas y el procesamiento del lenguaje natural (recuperación y extracción de la información y generación del lenguaje natural). Este conjunto de teorías y aproximaciones permite describir, procesar y generar sintáctica y semánticamente el potencial combinatorio de la frase nominal en el eje sintagmático y paradigmático frasal (la valencia activa) y en el oracional (la valencia pasiva). Para tal fin, se analizaron patrones argumentales –frasales y oracionales, así como las selecciones léxicas de sujetos, verbos y objetos (Domínguez Vázquez & Valcárcel Riveiro, 2020; Valcárcel Riveiro, 2017)–, el significado combinatorio –roles semánticos y rasgos ontológicos de los elementos implicados en una expresión—, así como los prototipos léxicos y clases semánticas actualizables en las diferentes casillas funcionales (Engel, 2004; Domínguez Vázquez, 2022).

La Figura 1 recoge cinco fases centrales de trabajo, así como el abanico de herramientas y técnicas aplicadas en dichos estadios. Como se puede observar, los datos compilados en esta figura son de diferente cariz, si bien en su conjunto contribuyen al desarrollo global de los generadores y de los proyectos de investigación en los que se enmarcan.

A continuación, se describen pautas generales sobre el conjunto de herramientas:

a) El epígrafe "Investigación" engloba herramientas y recursos de investigación para la compilación y análisis de datos lingüísticos. Diferenciamos aquí tres tipos de herramientas: las externas, las internas existentes y las internas diseñadas *ad hoc*. El concepto de interno o externo refiere aquí a si las herramientas han sido diseñadas por grupos o investigadores externos al equipo de los proyectos de investigación, como es el caso de *WordNet* o *Wikipedia*. Dado que nuestros generadores también se retroalimentan de recursos desarrollados previamente por el equipo de investigación, denominamos a este tipo "internas existentes". Este es el caso del diccionario *Portlex*. Un tercer bloque lo conforman aquellas herramientas concebidas para la investigación y diseñadas por el equipo *ad hoc*. Dichas herramientas van ligadas a diferentes fases de trabajo y contribuyeron a avanzar en el diseño y optimización de los generadores, así como a agilizar determinadas fases de trabajo automatizando procedimientos.
b) Bajo la etiqueta "Edición y Corrección" se enmarcan recursos imprescindibles en tareas de corrección y edición de los datos compilados y generados. Estos son accesibles para los equipos de trabajo en la intranet.
c) Las herramientas de generación son nuestros prototipos *online* de generación automática de la lengua, en concreto, *Xera* (generación automática monoargumental de la frase nominal en alemán, español y francés), *XeraWord* (generación automática monoargumental de la frase nominal en gallego y portugués), *Combinatoria* (generación

biargumental de la frase nominal en alemán, español y francés) y *CombiContext* (generación de la combinatoria en contexto en alemán, español y francés). Para todos ellos se despliegan diferentes interfaces de usuario con diferentes estructuras de acceso.

d) Las herramientas de difusión (páginas web propias, pero también post y vídeos en redes sociales) son centrales en los proyectos de investigación, ya no sólo por su valor a la hora de trasladar los resultados científicos a la sociedad, sino por las posibles vías de colaboración que de ahí puedan surgir. Asimismo, su actualización y mantenimiento resultan relevantes en la planificación de la carga de trabajo de los equipos.

e) Una aplicación directa de los generadores es el desarrollo de actividades *online* de aprendizaje automatizadas. Estas nos permiten explorar las posibilidades de uso didáctico de los ejemplos generados con los prototipos *Xera*, *Combinatoria* y *Combicontext*. La figura 1. las recoge bajo "Aplicación".

A su vez, resulta relevante destacar que el desarrollo de este conjunto de herramientas para diferentes lenguas, junto con los diferentes equipos de investigación implicados en los diferentes proyectos supone una dificultad añadida en cuanto al tiempo invertido en su propio diseño y en la coordinación de los miembros participantes.

3. HERRAMIENTAS, MÉTODO Y FASES DE TRABAJO: INVESTIGACIÓN Y GENERACIÓN

Como se ha señalado, para el correcto funcionamiento de los prototipos de generación se aplicaron y/o diseñaron diferentes herramientas y técnicas que sostienen diversos procedimientos y fases de trabajo. A continuación, presentaremos recursos de investigación y de generación automática de datos. Esta aproximación aporta, a su vez, una visión de conjunto de la interrelación entre a) dichas herramientas, técnicas de análisis y métodos y b) las diferentes fases de trabajo en las que se detectó la necesidad de manejarlas o diseñarlas.

3.1. *Estableciendo los patrones argumentales*

Los generadores describen la estructura argumental sintáctico-semántica de sustantivos valenciales y aportan ejemplos seleccionados según el filtro de selección del usuario (Domínguez Vázquez, 2022)[5]. El establecimiento de dichos patrones no es solo una cuestión cuantitativa, sino también cualitativa: es necesario determinar el tipo de casillas funcionales, así como qué unidades léxicas suelen o pueden cubrir esos espacios funcionales. Observemos los siguientes ejemplos de la figura 2:

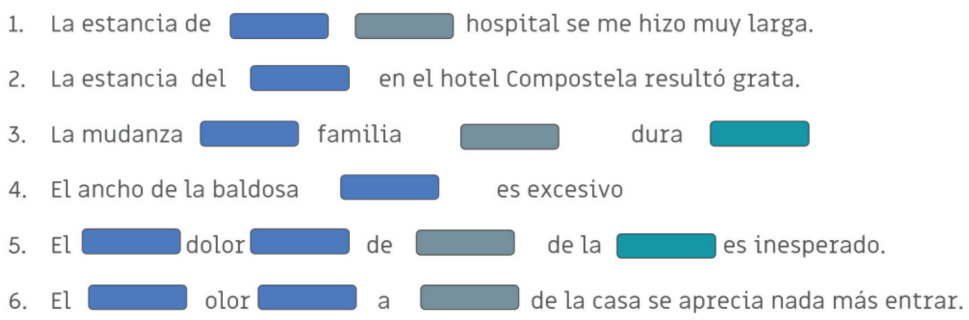


Figura 2: *Ejemplos de patrones argumentales*

En el ejemplo 1. la primera casilla valencial puede ser ocupada por un elemento expansivo, por ejemplo, *La estancia de dos meses*, si bien nada impide la realización de un elemento humano, *La estancia de Pedro*. Por tanto, en primer lugar, hay que establecer la interfaz sintáctico-semántica. Dichas estructuras o patrones argumentales del nombre los extraemos del diccionario multilingüe Portlex[6], otro de los recursos recogidos en el portal lexicográfico con el mismo nombre. Portlex nos permite determinar el patrón argumental atendiendo a criterios formales, pero también contemplando los

[5] Finalmente, cabe subrayar aún que, si bien para el establecimiento de los patrones argumentales aplicamos criterios valenciales sintáctico-semánticos, en las diferentes interfaces de consulta de los generadores no se explicitan ni las funciones sintácticas ni los roles semánticos de cada casilla funcional de cada sustantivo, pero estos subyacen al análisis. El usuario sí observa en dicha interfaz realizaciones formales y los rasgos ontológicos.

[6] Se trata de un diccionario online multilingüe, multilateral y modular de la frase nominal en francés, alemán, español, gallego e italiano, concebido como diccionario colaborativo (Domínguez Vázquez & Valcárcel Riveiro, 2020). Teóricamente se fundamenta en Domínguez Vázquez (2011) y Engel (2004). El sistema de gestión de bases de datos es MySQL.

roles semánticos (o significado relacional) y las entidades ontológicas (o significado categorial) en un nivel general (vid. 3.2.). De este modo, determinamos que un espacio funcional-valencial como el señalado previamente para el primer ejemplo puede ser actualizado por

- un completo sujeto expresado mediante la estructura [preposición *de* (+ determinante) + Nombre: {humano}]: *La estancia de Pedro/del profesor.*
- un complemento dilativo o expansivo con el patrón [preposición *de* (+ determinante) + Nombre: {unidad de tiempo}]: *La estancia de dos meses.*

Como muestran los ejemplos anteriores, una misma realización formal [*la estancia* + *de*] puede realizar en superficie diferentes funciones sintáctico-semánticas. Por tanto, en esta fase de trabajo aplicamos una aproximación cuantitativa y una cualitativa, ambas imprescindibles atendiendo a nuestros propósitos:

- Aproximación cuantitativa: Con el fin de obtener el caudal léxico que puede cubrir un espacio funcional compilamos, en primer lugar, datos cuantitativos mediante consultas CQL (*corpus query language*) en Sketch Engine. La Tabla 1 muestra los datos del sustantivo ESTANCIA en la estructura [estancia + en + determinante][7].
- Aproximación semántico-cualitativa: Es a todas luces evidente que los corpus, como Sketch Engine, posibilitan la agrupación de los datos extraídos mediante criterios de frecuencia y criterios formales, como ejemplifica la Tabla 1. También es sabido que para las lenguas objeto de estudio no contamos con corpus anotados sintáctica-semánticamente. Dicha anotación es imprescindible para los generadores, no solo desde un punto de vista lingüístico, sino también computacional. Para superar este primer obstáculo, el equipo de investigación lleva a cabo una depuración de los datos extraídos de Sketch Engine siguiendo criterios valenciales. A continuación, dichos datos se prototipan semánticamente (vid. 3.2).

[7] CQL-Query: [lemma=""estancia""][lemma=""en""][tag=""D.*""][tag=""A.*""]?[tag=""N.*""].

Lema	Frecuencia
1. estancia en el ciudad	2831
2. estancia en el extranjero	2723
3. estancia en el país	2637
4. estancia en el hospital	2390
5. estancia en el hotel	2339
6. estancia en el capital	1587
7. estancia en el isla	1319
8. estancia en el cárcel	1128
9. estancia en el Universidad	1051
10. estancia en el centro	1030
11. estancia en uno hotel	750
12. estancia en el casa	721
13. estancia en el Hotel	604
14. estancia en nuestro hotel	532
15. estancia en nuestro país	487
16. estancia en el universidad	474
17. estancia en el zona	470
18. estancia en el lugar	458
19. estancia en este hotel	434
20. estancia en este ciudad	407

Tabla 1: *Consulta CQL en Sketch Engine para [estancia + en + determinante]*

3.2. *Prototipando: características ontológicas y clases semánticas*

Una vez determinadas las características centrales de los patrones argumentales se requiere prototipar y agrupar los candidatos léxicos susceptibles de realización en un determinado *slot* valencial. Nuestro modelo descriptivo se fundamenta aquí en un concepto propio de prototipo léxico –léxico más frecuente que ocupa un determinado espacio funcional– y en las clases semánticas prototípicas –caudal léxico agrupado en clases semánticas tras un proceso de prototipado ontológico (Domínguez Vázquez, 2021). Por tanto, tras haber depurado los listados de frecuencia que obtenemos de Sketch Engine (vid. 3.1.), anotamos el vocabulario según sus rasgos ontológicos y lo agrupamos semánticamente. Arranca aquí la fase de prototipado léxico (vid. tabla 2).

El inventario de rasgos que aplicamos en el proceso de prototipado conforma lo que denominamos *ontología léxica bottom-up* (Domínguez Vázquez, Bardanca Outeiriño & Simões, 2021; vid. Martín Gascueña, en este volumen). Para su elaboración partimos del inventario de rasgos categoriales de la gramática y lexicografía valencial (Engel, 2004; Domínguez Vázquez, 2011) y de las ontologías de WordNet, cuyos *synsets* están asociados a rasgos semántico-cognitivos. La conjunción de diferentes recursos se debe al hecho de que para una descripción detallada del material lingüístico los rasgos categoriales del inventario valencial no son lo suficientemente granulares atendiendo a nuestros propósitos, dado que solo nos permiten identificar categorías o clases generales[8] –por ejemplo, {situación}, {material} en la Figura 3-.

Combinaciones

Realización formal	Rasgo categorial	Tipo complemento
de	Material	Complemento sujeto
a	Material, Situación	Complemento prepositivo

Frase tipo:
El olor del agua a gasolina

Ejemplos y notas:
Me encanta el **olor** de la casa a galletas recién hechas y esa desconexión que sólo consigo cuando las estoy haciendo y decorando.

WEB

Realización formal	Rasgo categorial	Tipo complemento
a	Material, Situación	Complemento prepositivo
de	Material	Complemento sujeto

Frase tipo:
El olor a matarratas de una fábrica cercana

Ejemplos y notas:
Y había el **olor** a gasolina de los libros de Fausto. Y los cuadernos de Fausto siempre estaban ajados, gastados, con el sudor de las manos. Y también, siempre a medio usar, sus lápices.

CREA: García Vega, Lorenzo: Los años de Orígenes, Monte Avila Editores: Caracas, 1978.

Realización formal	Rasgo categorial	Tipo complemento
de	Material, Situación	Complemento prepositivo
de	Material	Complemento sujeto

Frase tipo:
El olor de limpieza de su ropa

Ejemplos y notas:
¿Cómo debes deshacerte del mal **olor** de pies de los zapatos?

WEB

Realización formal	Rasgo categorial	Tipo complemento
Adjetivo	Material, Situación	Complemento prepositivo
de	Material	Complemento sujeto

Frase tipo:
El olor fecal de las alcantarillas

Ejemplos y notas:
El virus de la gripe aviar se puede detectar por el **olor** fecal de las aves infectadas.

WEB

FIGURA 3: *Captura del diccionario multilingüe Portlex*

En favor de una mayor regularidad decidimos recurrir también a las ontologías y recursos manejados en WordNet: la Suggested Upper Merged Ontology (SUMO) (Niles & Pease, 2001), la Top Concept Ontology (TOP) (Álvez, Atserias, Carrera, Climent, Laparra, Oliver & Rigau, 2008), los WordNet Domains

[8] Estos rasgos generales, sin embargo, cumplen una función central porque funcionan de vínculo entre los datos lingüísticos y las ontologías de WordNet.

(Bentivogli, Forner, Magnini & Pianta, 2004), el Basic Level Concept (Izquierdo Beviá, Suárez Cueto & Rigau, 2007), los Epinónimos (Gómez Guinovart & Solla Portela, 2018), así como a los primitivos semánticos (Miller, Beckwith, Fellbaum, Gross & Miller, 1990).

Un ejemplo concreto del resultado de prototipado se muestra en la Tabla 2. El vocabulario recogido para ESTANCIA en la estructura [estancia + en + determinante] (Tabla 1) refiere a lugares, pero de diferentes características. El elemento más frecuente es *ciudad* (obsérvese la posición 1 y la 20), con lo cual *ciudad* será el prototipo léxico de la clase {lugar población general} frente a *hospital*, que, siendo también muy frecuente, refiere a un {lugar construcción tipo medicina}. La aplicación de la ontología permite ir agrupando el léxico como sigue:

	1. nivel	2. nivel	3. nivel	4. nivel
ciudad	lugar	población	general	
capital	lugar	población	general	
país	lugar	territorio	general	
hospital	lugar	construcción	tipo	medicina
cárcel	lugar	construcción	tipo	jurisprudencia
universidad	lugar	construcción	tipo	educación
isla	lugar	paisaje	acuático	general

TABLA 2: *Ejemplo de prototipado*

La figura 4 muestra un ejemplo de algunos de los rasgos categoriales aplicados para la descripción de lugares (véase tabla 2, por ejemplo {paisaje} y {construcción}).

Cabe señalar que dicha clasificación ontológica ha sido desarrollada expresamente para la finalidad de los proyectos y se va enriqueciendo a medida que se añaden nuevas unidades nominales de análisis. En este sentido, es parcial e incompleta: su granularidad depende de la necesidad de describir con mayor o menor detalle las unidades ontológicas que pueden ocupar determinadas casillas funcionales, para, de este modo, plasmar las restricciones de combinatoria semántico-categorial de cada argumento valencial y sus diferentes realizaciones de superficie.

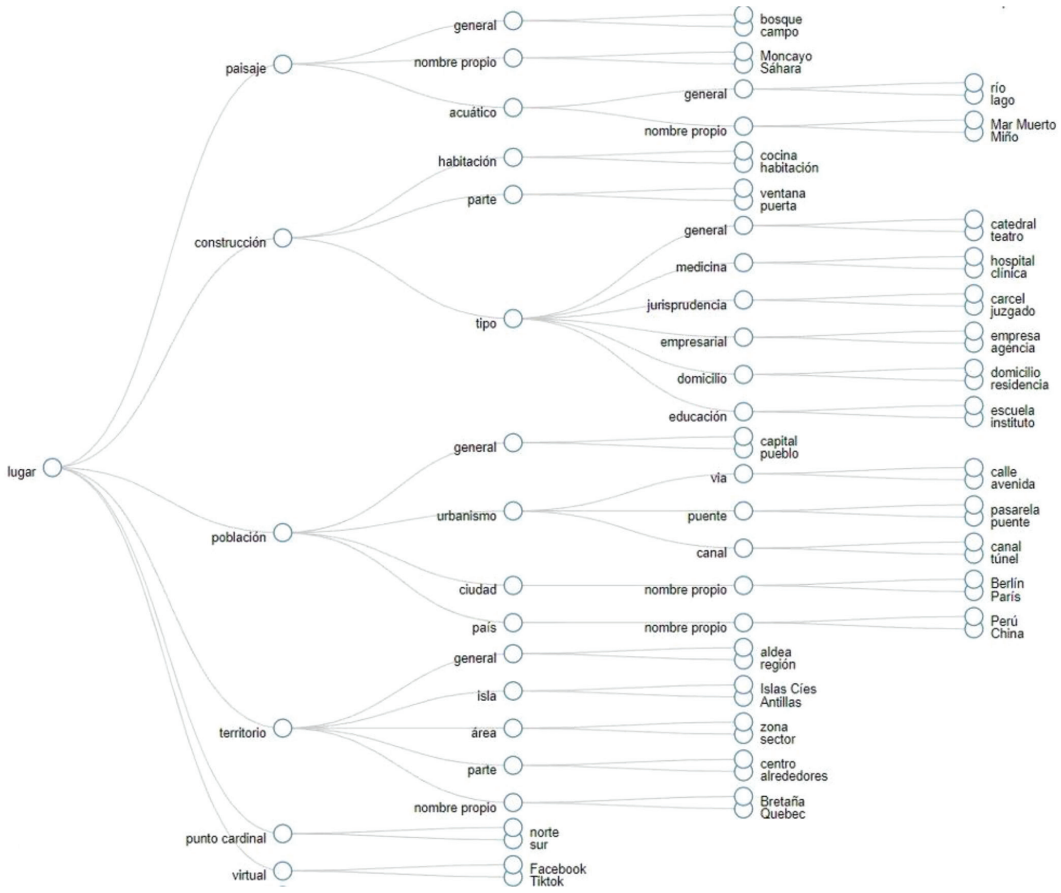


Figura 4: *Vista parcial de la ontología léxica bottom-up*

3.3. *Expandiendo*

Habiéndose analizado los candidatos léxicos y las clases semánticas estándar, se plantea la pregunta de cómo obtener para *slots* concretos más caudal léxico ya agrupado semánticamente. Empieza la fase de expansión léxica en el eje paradigmático. Pongamos un ejemplo: la clase semántica {construcción tipo jurisprudencia}, representada con el prototipo léxico *cárcel* (Tabla 2), tiene que tener un abanico de lexemas similares y compatibles en el eje paradigmático con el propio prototipo, como, por ejemplo, *La estancia en la penitenciaría* | *el talego* | *la celda* | *el calabozo* | *la prisión* | *los reformatorios* | *la institución correccional*. Alguno de estos lexemas puede no aparecer en el listado inicial de los aproximadamente 500 lexemas de

media extraídos de Sketch Engine, pero son igualmente relevantes. Dada la imposibilidad de filtrar y extraer información semántica agrupada de los corpus actuales para las lenguas objeto de estudio, decidimos desarrollar las herramientas *Lematiza* y *Combina*:

- *Lematiza* es un lematizador de actantes de corpus que funciona subiendo ficheros en formato xml o csv extraídos de Sketch Engine (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019, véase Figura 5). Con esta herramienta obtenemos la información recogida en las diferentes ontologías de WordNet para los lemas extraídos del corpus Sketch Engine.

FIGURA 5: *Interfaz de usuario de Lematiza*

La Figura 6 permite visualizar, para el ejemplo de "hospital", el tipo de información a la que accedemos para todo el vocabulario contenido en el archivo manejado. Ofrece la posibilidad de consultar diferentes acepciones de significado y navegar directamente por las ontologías de WordNet, pero, a su vez, nos permite ir afinando nuestra ontología léxica *bottom-up*.

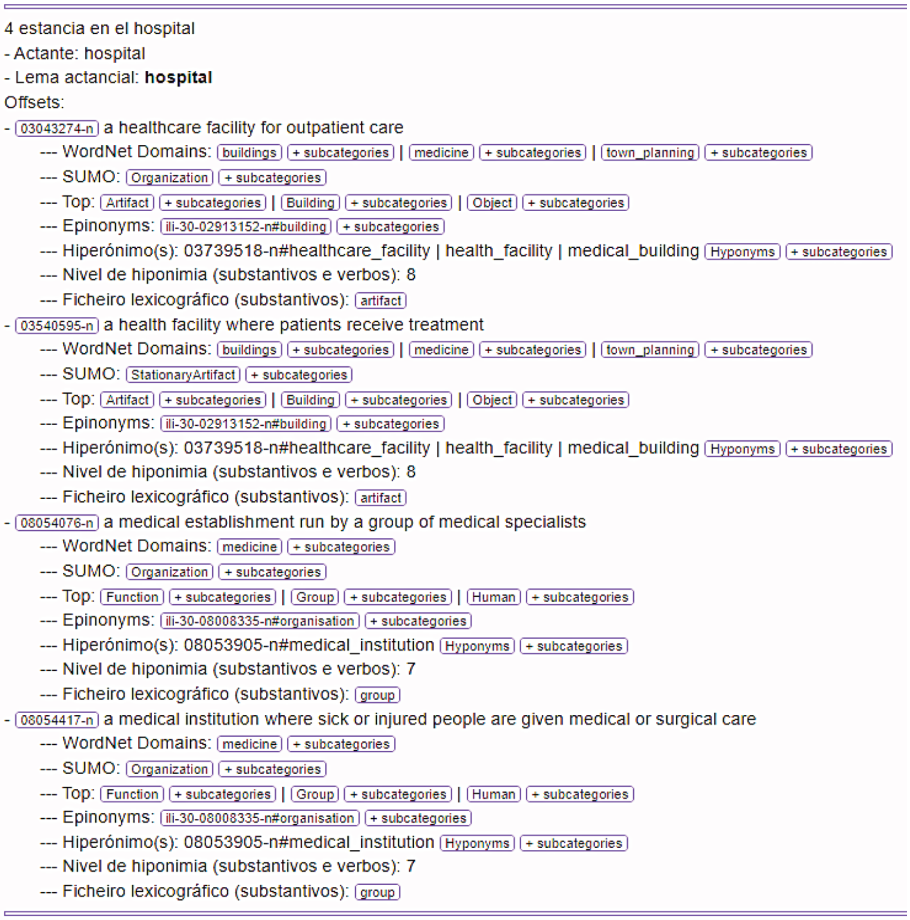


Figura 6: *Información provista por Lematiza*

- A continuación, desarrollamos *Combina,* la cual nos permite combinar y cotejar los resultados de varias consultas sobre el caudal léxico recogido en las ontologías de Wordnet. De este modo extraemos una selección léxica en el eje paradigmático, la cual comparte las características semánticas del prototipo léxico-semántico tomado como punto de partida (Domínguez Vázquez, 2021; Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019). Siguiendo con el ejemplo de {lugar construcción tipo medicina} obtenemos, por tanto, ejemplos como los que recoge la Tabla 3 para el español:

1 02820798-n casa de locos	13 03540595-n enfermería
2 02820798-n crazy house	14 03540595-n hospital
3 02820798-n gallinero	15 03540595-n hospitales
4 02820798-n loquera	16 03650803-n lazareto
5 02820798-n manicomio	17 03650803-n leprosería
6 03043274-n clínica	18 03746574-n hospital psiquiátrico
7 03043274-n hospital	19 03746574-n manicomio
8 03129471-n inclusa	20 03746574-n psiquiátrico
9 03210552-n ambulatorio	21 03746574-n siquiátrico
10 03210552-n dispensario	22 03762982-n hospital militar
11 03333349-n hospital de campaña	23 04133497-n sanatorio
12 03540595-n clínica	

Tabla 3: *Resultados de* Combina *para {lugar construcción tipo medicina} en español*

En algunos casos es necesario depurar los resultados obtenidos: el acceso directo gracias a dicha herramienta a los *synset* y a las diferentes ontologías es de especial ayuda en esta tarea. Finalmente, los datos se pueden descargar en formato Json y txt y pueden ser depurados manualmente, de ser necesario, y reutilizados.

Para el funcionamiento de *Lematiza*, y posteriormente de *Combina*, se han desarrollado en primer lugar 4 APIs (para el alemán, español, francés y gallego). Dichas APIs permiten extraer datos léxicos de las consultas recurriendo a las relaciones semánticas de WordNet y a las ontologías vinculadas a los *synsets* en el modelo de EuroWordNet. Estas, así como la propia herramienta *Lematiza*, utilizan código derivado de diferentes proyectos del Seminario de Lingüística Informática de la Universidad de Vigo. También enlazan

con la interfaz de Galnet para ilustrar la identificación del significado de formas léxicas (Gómez Guinovart & Solla Portela, 2018). Los datos lingüísticos para el español y los enlaces con las ontologías provienen de Galnet, que integra el repositorio central multilingüe, además de los Epinonyms (comp. Gonzalez-Agirre et al., 2012). En el caso del francés los datos se tuvieron que adaptar desde WOLF (comp. Sagot & Fišer, 2008). Los datos del alemán proceden del Open Multilingual Wordnet (Bond & Foster, 2013) y parcialmente también del UWN/MENTA (Melo & Weikum, 2010). Para *XeraWord*, los datos lingüísticos del gallego y portugués y los enlaces con las ontologías procede de Galnet y Pulo, los cuales también integran el Repositorio Central Multilingüe (MCR; González, Aguirre Laparra & Rigau, 2012).

Como ya se ha señalado previamente, la granularidad en el establecimiento de las clases semánticas es importante no solo para el proceso de generación, sino también para configurar los paquetes léxicos (3.4.) también desde un punto de vista formal. Así, el sustantivo alemán UMZUG (`mudanza´) selecciona una u otra preposición directiva dependiendo del lugar al que uno se mude, como muestran las figuras 7 y 8 (para otras cuestiones contrastivas, Pino y Valcárcel Riveiro en este tomo):

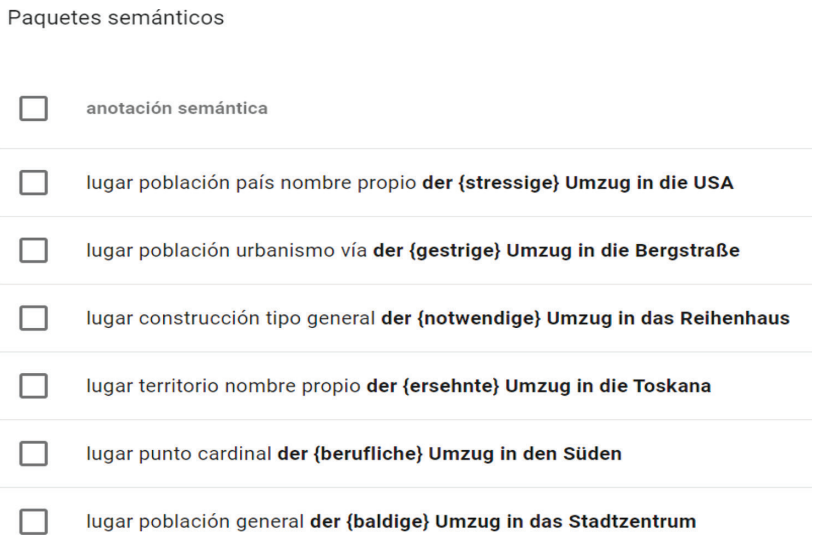


Figura 7: *Paquetes léxicos combinables en la expresión de la dirección del sustantivo Umzug con la preposición in*

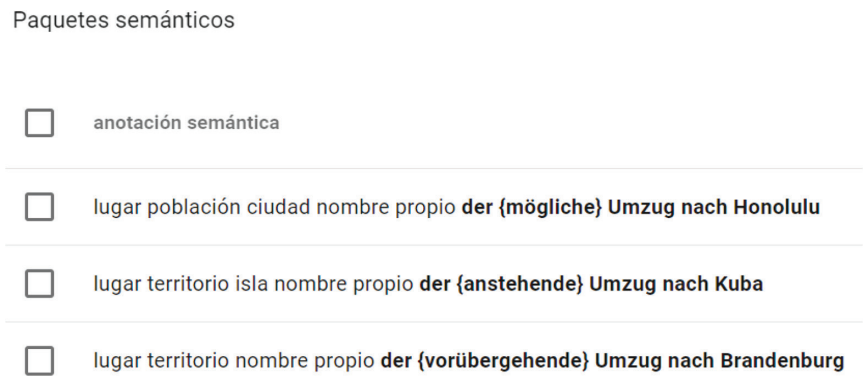


Figura 8: *Paquetes léxicos combinables en la expresión de la dirección del sustantivo Umzug con la preposición nach*

3.4. *Flexionando y empaquetando*

En esta fase, obtenemos los paquetes léxicos, elementos nucleares flexionados para la generación automática (Domínguez Vázquez, Bardanca Outeiriño & Simões, 2021). Una vez establecidos los lemas que conforman cada clase semántica necesitamos flexionarlos, para lo cual recurrimos en el caso del español, alemán y francés a la herramienta *Flexiona* y en el caso del gallego y portugués a *Flexionador*. Los flexionadores utilizan los diccionarios del analizador lingüístico FreeLing. Para generar la sintaxis y la morfología (conjugadores, flexionadores nominales y adjetivales, etc.) recurrimos al lenguaje de programación Python, en el que desarrollamos nuestra propia librería de generación frasal y verbal.

Finalmente, cada uno de estos paquetes léxicos incluye, para cada casilla valencial, un identificador único, una descripción del tipo de objeto que se está caracterizando, su clasificación en la ontología y una lista de lemas. Cada lema está enlazado con el respectivo Índice Interlingüístico (ILI), utilizado tanto en WordNet, como en el Repositorio Central Multilingüe (MCR).

3.5. *Traduciendo semi-automáticamente paquetes léxicos*

Si bien nuestros generadores para el español, alemán y francés no se sustentan inicialmente en principios de traducción automática, sí que se han

explorado diferentes vías para la optimización de resultados y la agilización de diferentes fases de trabajo recurriendo a un traductor automático de datos extraídos de WordNet. Es así como nace *XeraWord*, una herramienta piloto de generación automática de la frase nominal simple en gallego y portugués, basada en la traducción automática de léxico extraído de WordNet.

Para su desarrollo el Instituto da Lingua Galega (ILG) de la Universidad de Santiago de Compostela diseñó *ad hoc* una herramienta que permite la traducción automática de paquetes léxicos, en concreto, de los ejemplares en relación paradigmática compilados a partir de los datos extraídos automáticamente de Wordnet. Dicho traductor recurre a Mymemory y al WordNet del portugués —Pulo (Simões & Gómez Guinovart, 2014)— y del gallego —Galnet (Gómez Guinovart, 2011; Gómez Guinovart & Solla Portela, 2018).

3.6. *Combinando estructuras biargumentales*

El prototipo *Combinatoria* genera frases nominales complejas biargumentales, por tanto, patrones biargumentales con aleatoriedad restringida en el eje sintagmático y paradigmático. La generación automática de los ejemplares léxicos sigue un principio de aleatoriedad restringida, el cual no afecta ni a las clases semánticas ni a los roles, puesto que a) las clases semánticas están preestablecidas para cada sustantivo y sus representantes léxicos y b) el usuario selecciona previamente la estructura argumental y la clase semántica. *Combinatoria* constituye un claro ejemplo de la necesidad de contar con clases semánticas granulares. Así, para el sustantivo ESTANCIA es necesario delimitar paquetes semánticos como {lugar construcción tipo medicina} o {lugar construcción tipo jurisprudencia}, porque, entre sus combinatorias previsibles, se encuentran los paquetes léxicos de lexemas como *paciente* o *prisionero*, respectivamente: *La estancia del paciente en el hospital* y *la estancia del prisionero en la cárcel.* Dichas combinatorias se recogen en la herramienta, tal y como se observa en la figura 9:

Figura 9: *Interfaz de usuario en Combinatoria*

3.7. *Creando el contexto frasal y oracional*

La herramienta *CombiContext* aporta el marco frasal y oracional en el que se pueden incrustar las secuencias generadas automáticamente por los generadores *Xera* y *Combinatoria*. Contribuye, además, a humanizar los resultados de los generadores. *CombiContext* se retroalimenta de los datos de los generadores previos, si bien requiere la aplicación de nuevos métodos y recursos, así como el tratamiento de nuevos datos lingüísticos.

Desde un punto de vista lingüístico, se decide en un primer estadio de trabajo predeterminar estructuras básicas oracionales. De este modo, los diferentes equipos lingüísticos comienzan con el análisis de estructuras copulativas e intransitivas y, paulatinamente, van avanzando hacia las transitivas y preposicionales:

(i) sujeto$^{\text{frase nominal}}$ + verbo + adverbio: *El viaje de Mario a Berlín termina así*;

(ii) sujeto$^{\text{frase nominal}}$ + verbo + atributo: *La huida de los refugiados desde la frontera es peligrosa*;

(iii) adverbio + sujeto + verbo + objeto directo$^{\text{frase nominal}}$: *Ahora Carlos nota el olor a humedad de la casa*;

(iv) sujeto + verbo + adverbio + suplemento$^{\text{frase nominal}}$: *Nerea escribe brevemente sobre el amor de Marco Antonio por Cleopatra*.

Dichas estructuras formales muestran variabilidad paradigmática relativa a las clases semánticas combinables entre sí, así como a los representantes léxicos que las conforman. Los ejemplos cuentan también con variabilidad sintagmática, dado que los verbos, adverbios y adjetivos se generan con aleatoria restringida. Por tanto, un ejemplo como

la discusión de los alumnos con el profesor es intensa

puede mostrar diferentes realizaciones como

(i) posibles adjetivos en relación paradigmática para las posiciones prenominal y posnominal: *la reciente* | *acalorada* | *intensa* | *etc. discusión*; *la discusión posterior* | *previa* | *final* | etc.

(ii) diferentes adverbios: *probablemente* | *seguramente* | *normalmente, etc.*; *ahora* | *después* | etc.

(iii) estructuras diversas con diferentes verbos en relación paradigmática: *ser* | *resultar* | *parecer* + atributo (que nuevamente muestran variabilidad paradigmática); *mantener* | *escuchar* + complemento directo; *participar en* + suplemento, etc.

De este modo se pueden generar automáticamente oraciones *ad libitum*[9].

Para asegurar el funcionamiento de *CombiContext* ha sido necesario alimentarla de datos lingüísticos nuevos: la selección de los verbos que aportan el marco en el que se incrustan los eductos generados automáticamente, así como la de los adjetivos –valenciales o no– y la de los diferentes complementos circunstanciales se determinan a partir de un PoS-Tagger, que recurre a Wikimedia. El WikiExtractor de clases de palabra organiza todos los textos tomados de los wikidumps, esto es, una colección de todos los datos textuales disponibles en

[9] Para datos cuantitativos véase el capítulo 1. en esta monografía.

la base de datos de Wikimedia, separados por lengua. Para poder manejar estos datos ha sido necesario procesarlos centrándose en la extracción de concordancias en las que se incluye alguno de los núcleos presentes en la herramienta. Una vez extraída la lista de concordancias, se almacenan en la base de datos para su posterior consulta a través de la interfaz del recurso.

La consulta en tiempo real de los datos es posible, puesto que la herramienta se apoya en la integración de un PoS-tagger, desarrollado con Spacy (Honniba & Montani, 2017), para las distintas lenguas del proyecto. La integración de estos etiquetadores permite la interoperabilidad y consulta de los datos producidos durante las fases previas del proyecto[10] y los nuevos datos tomados de Wikimedia, que están almacenados en bruto. Es decir, el procesamiento de los datos es el resultado de una petición hecha a la carta por parte del usuario y que se realiza sobre la marcha. De este modo, la herramienta procesa todos los datos disponibles desde Wikimedia para el núcleo y devuelve aquellos sustantivos, adverbios, adjetivos y verbos que se encuentren en construcciones sintácticamente relevantes. Así, por ejemplo, para la extracción de adjetivos

> determinante-adjetivo_o-nucleo-adjetivo_o-de-actante N1-en-actante N3

la herramienta generará una nueva estructura, resultado de la abstracción de la estructura básica, que contiene el esqueleto indispensable para la validación de los datos de búsqueda:

> (comienzo de frase)-(hasta dos elementos desconocidos)-nucleo-(hasta un elemento desconocido)-de-sustantivo-en-sustantivo

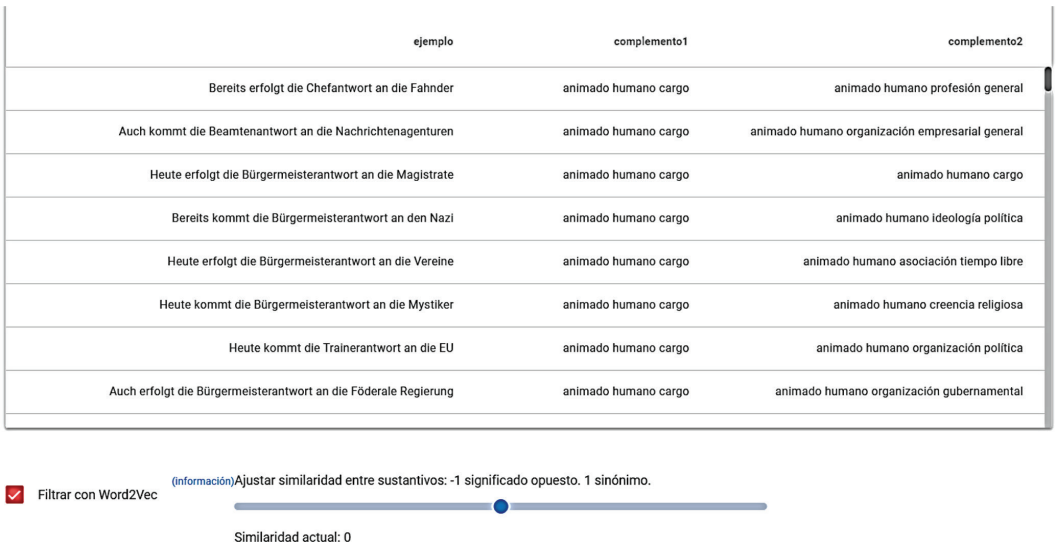
La extracción de los verbos se realiza de manera independiente de la estructura seleccionada, pero ligada al núcleo. Así, se obtiene una lista ordenada

[10] Con el objetivo de rentabilizar por completo esta aproximación, los investigadores parten exclusivamente de las estructuras formales ya documentadas durante el desarrollo de la combinatoria nominal para cada núcleo, a las que pueden añadir, mediante la herramienta, nuevos elementos para combinarlas como estructuras verbales. Estos nuevos elementos se corresponden principalmente con las etiquetas de adjetivo, adverbio, sustantivo y verbo.

por frecuencia de los verbos que aparecen con el sustantivo nuclear en los datos extraídos de Wikimedia.

Para facilitar la automatización de los resultados generados, poder aprovechar el trabajo original realizado durante la fase de desarrollo de los generadores y analizar la admisibilidad de las combinatorias generadas hemos integrado en nuestra herramienta *word embeddings* –el método *Word2vec* de Mikolov, Chen, Corrado y Dean (2013)–. Este se basa en el uso de una red recurrente neuronal. Por *word embeddings* se entiende, en pocas palabras, la representación en un espacio vectorial de una forma léxica resultado de distintas técnicas de procesamiento de corpus. En este caso, también hemos aplicado *Glove* (Pennington, Socher & Manning, 2014), un algoritmo que se basa en la agrupación de coocurrencias dentro del corpus de entrenamiento.

La principal motivación para la integración de *word embeddings* en las herramientas es evitar conflictos semánticos en contextos semánticamente válidos. Atiéndase al siguiente ejemplo: *el dolor de ovarios del abuelo*. Este tipo de errores derivan de la combinación ciega generada a partir de los paquetes originales usados en *Xera*. *Word2vec*, por lo tanto, permite filtrar este tipo de problemas basándose en la métrica de similitud contextual producida al comparar los dos vectores de los lemas *abuelo* y *ovario*. La integración de *Word2vec* en la interfaz de usuario permite controlar a través de un deslizador (vid. Figura 10) la mayor cercana o distancia entre los paquetes léxicos seleccionados a la hora de ser combinados, para a continuación generar los ejemplos. La selección de -1 en la herramienta se corresponde con la ausencia de filtro, es decir, todas las opciones, por pequeña que sea la correspondencia entre los lemas, se consideran válidas. En el otro extremo, la selección en el deslizador de una igual a 1 indica que solo aquellas frases cuyo vector contextual se corresponda por completo con la otra palabra, que está siendo comparada, serán mostradas en la interfaz de consulta. Por defecto, el deslizador se sitúa en 0. Este parámetro se corresponde con una similitud entre los dos vectores contextuales de 50% o más.



ejemplo	complemento1	complemento2
Bereits erfolgt die Chefantwort an die Fahnder	animado humano cargo	animado humano profesión general
Auch kommt die Beamtenantwort an die Nachrichtenagenturen	animado humano cargo	animado humano organización empresarial general
Heute erfolgt die Bürgermeisterantwort an die Magistrate	animado humano cargo	animado humano cargo
Bereits kommt die Bürgermeisterantwort an den Nazi	animado humano cargo	animado humano ideología política
Heute erfolgt die Bürgermeisterantwort an die Vereine	animado humano cargo	animado humano asociación tiempo libre
Heute kommt die Bürgermeisterantwort an die Mystiker	animado humano cargo	animado humano creencia religiosa
Heute kommt die Trainerantwort an die EU	animado humano cargo	animado humano organización política
Auch erfolgt die Bürgermeisterantwort an die Föderale Regierung	animado humano cargo	animado humano organización gubernamental

Figura 10: *Aplicación de Word2vec en CombiContext*

4. EVALUACIÓN DE LOS GENERADORES

La principal limitación de los generadores reside en la necesidad de integrar reglas simbólicas capaces de capturar de manera eficiente la generación automática de todas las frases y oraciones analizadas lingüísticamente por el equipo humano. Este obstáculo es compensado mediante la relación de dependencia de los paquetes léxicos en la marcación manual con las distintas estructuras esperadas. La generación, por lo tanto, está restringida a la lista de etiquetas que han sido diseñadas de manera conjunta entre el equipo lingüístico e informático. Consecuentemente, si se quisiese incluir una lengua nueva que necesite hacer uso de etiquetas no registradas todavía, como puede ser la inclusión de la flexión nominal de una lengua eslava como el ucraniano o ruso, se produciría un error en la generación de los datos. Esto se debería a que no está codificado de manera informática el tipo sustantivo instrumental.

Esta restricción en las estructuras es análoga a la necesidad de procesar primero la flexión de todos los lexemas con los que se desea producir nuevas combinaciones. Es decir, el sistema sólo puede producir oraciones con léxico que ha sido vinculado a un paquete léxico con anterioridad. La herramienta

de extracción de sustantivos y verbos permite esquivar en parte esta limitación al posibilitar la inclusión de nuevos sustantivos derivados del procesamiento de los datos de Wikimedia en la generación verbal. Esto también está presente en la dependencia por parte de los generadores de los datos revisados por los especialistas para la construcción de paquetes semánticos. Dicho obstáculo es en parte evitado con la construcción de herramientas de traducción semi-automática (3.5).

Una tercera limitación deriva de la capacidad física del servidor para almacenar todas las formas, lemas, combinaciones, ejemplos, modelos de *embeddings*, datos para el procesamiento con PoS-tagger, por nombrar algunos de los principales conjuntos de agrupamiento de datos resultantes.

Desde un punto de lingüístico, diferentes estudios exploratorios permiten observar la corrección formal de los ejemplos generados automáticamente. Se constatan también ciertas incongruencias semánticas, que bien pueden ser de tipo cultural –*La estancia del equipo de fútbol en Constantinopla*– o bien estar relacionadas con la generación aleatoria de todos y cada uno de los elementos partícipes en la oración. Así, concluimos que uno de los principales elementos de calidad de los generadores —la variabilidad de los datos generados (Hashimoto, Zhang & Liang, 2019) conseguida a través de la fase de expansión léxica (vid. 3.3.)— resulta ser uno de los factores que más dificultades supone.

Cabe la pena subrayar que los ejemplos generados por *CombiContext* pueden ser manejados para diferentes finalidades:

- ejemplos estándar^{plus}: Están filtrados mediante *Word2vec*. Siguen, por tanto, los criterios de Atkins y Rundell (2008) de naturalidad y tipicidad, son además informativos e inteligibles. Dichos ejemplos pueden servir de base para la elaboración de manuales o unidades didácticas y actividades de práctica controlada (conocidos también como ejercicios estructurales o "drills"), ya sea para el aula, ya para aplicaciones de aprendizaje de lenguas asistido por ordenador.

- ejemplos estándar^{minus}: Ejemplos sin el filtro de *Word2vec*, que pueden ser manejados en el aula de modo guiado (por ejemplo, para detectar restricciones de combinatoria), así como con propósitos de investigación.

5. CONCLUSIÓN

Los generadores diseñados verifican la viabilidad de la propuesta metodológica que los sustentan, pero a su vez la necesidad de ciertas optimizaciones (vid. 4). Para tal fin, actualmente se están explorando diferentes vías para automatizar el análisis y la evaluación de los datos.

En su fase actual los generadores son de libre acceso y gratuitos y cuentan con diferentes aplicaciones y finalidades (López en este volumen). Por una parte, no conocemos ninguna aplicación informática de WordNet que se aplique en la generación automática de combinatorias argumentales, con excepción de estos generadores piloto; por otra parte, configuran un nuevo modelo de recursos plurilingües valenciales con ejemplos dinámicos e interacción por parte del usuario. A su vez, la aplicación de diferentes técnicas, estrategias y recursos, algunos de ellos ya existentes, los significan como un buen ejemplo de una lexicografía más sostenible. Profundizando en esta idea: Muchos de los datos lingüísticos obtenidos, así como diferentes aplicaciones, técnicas y herramientas están siendo incorporados y manejados en el proyecto de etiquetado semántico ESMAS-ES^{+}[11].

[11] Proyecto ESMAS-ES+ (PID2022-137170OB-I00) financiado por MCIN/AEI//FEDER "Una manera de hacer Europa".

REFERENCIAS BIBLIOGRÁFICAS

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 Sixth International Conference on Language Resources and Evaluation* (LREC'08) (pp. 1529-1534). European Language Resources Association (ELRA). https://adimen.si.ehu.es/~rigau/publications/gwc08-tco.pdf

Atkins, B. T. S. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Bentivogli, L., Forner, P., Magnini, B. & Pianta, E. (2004). Revising WordNet Domains Hierarchy: semantics, coverage, and balancing. *Proceedings of the Workshop on Multilingual Linguistic Ressources. MLR '04* (pp. 101-108). Association for Computational Linguistics. https://doi.org/10.3115/1706238.1706254

Domínguez Vázquez, M.ª J. (2011). *Kontrastive Grammatik und Lexikographie: spanisch-deutsches Wörterbuch zur Valenz des Nomens*. Iudicium.

Domínguez Vázquez, M.ª J. (2021). Zur Darstellung eines mehrstufigen Prototypbegriffs in der multilingualen automatischen Sprachgenerierung: vom Korpus über word embeddings bis zum automatischen Wörterbuch. *Lexikos, 31*, 20-50. https://doi.org/10.5788/31-1-1623

Domínguez Vázquez, M.ª J. (2022). Contribución de la semántica combinatoria al desarrollo de herramientas digitales multilingües. *Círculo de Lingüística Aplicada a la Comunicación*, *90*, 171-18.

Domínguez Vázquez, M.ª 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 conference* (pp. 269-287). Lexical Computing CZ. https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_16_pp269-287.pdf

Domínguez Vázquez, M.ª 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, S. & C. Tiberius (eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference* (pp. 51-71). Lexical Computing CZ. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_4.pdf

Domínguez Vázquez, M.ª J. & Valcárcel Riveiro, C. (2020). PORTLEX as a multilingual and cross-lingual online dictionary. En M.ª J. Domínguez Vázquez, M. Mirazo Balsa & C. Valcárcel Rivero (eds.), *Studies on multilingual lexicography* (pp. 135-158). De Gruyter. https://doi.org/10.1515/9783110607659-008

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

Gómez Guinovart, X. (2011). Galnet: WordNet 3.0 do galego. *Linguamática*, *3*(1), 61-67.

Gómez Guinovart, X. & Solla Portela, M. A. (2018). Construyendo el WordNet gallego: métodos y aplicaciones. *Recursos y evaluación de idiomas*, *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)*. Japón. https://adimen.si.ehu.es/~rigau/publications/gwc12-glr.pdf

Hashimoto, T., 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*. Vol I. (pp. 1689-1701). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1169

Izquierdo Beviá, R. Suárez Cueto, A. & Rigau, G. (2007). Exploring the automatic selection of basic level concepts. En R. Mitkov, G. Angelova & K. Bontcheva (eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (pp. 298-302). Shoumen. https://adimen.si.ehu.es/~rigau/publications/ranlp07-isr.pdf

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. 1-12). Conference Track Proceedings. https://arxiv.org/pdf/1301.3781.pdf

Miller, G. A, Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. J. (1990). Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, *3*(4), 235-244. https://doi.org/10.1093/ijl/3.4.235

Niles, I. & Pease, A. (2001). Towards a standard upper ontology. *FOIS '01. Proceedings of the International Conference on Formal Ontology in Information Systems* (pp. 2-9). ACM. https://doi.org/10.1145/505168.505170

Pennington, J., Socher, R. & Manning, C. (2014). GloVe: Global Vectors for Word Representation. En A. Moschitti, B. Pang & W. Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532-1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Simões, A. & Gómez Guinovart, X. (2014). Bootstrapping a Portuguese WordNet from Galician, Spanish and English Wordnets. En J. L Navarro Mesa, A. Ortega, A. Teixeira, E. Hernández Pérez, P. Quintana Morales, A. Ravelo García, I. Guerra Moreno, D. T. Toledano (eds.), *Advances in Speech and Language Technologies for Iberian Languages* (pp. 239-248). Springer. https://doi.org/10.1007/978-3-319-13623-3_25

Valcárcel Riveiro, C. (2017). Las construcciones N1N2 como realizaciones actanciales del sustantivo en francés y su tratamiento en el diccionario multilingüe PORTLEX. En M.ª J. Domínguez Vázquez & S. Kutscher (eds.), *Interacción entre gramática, didáctica y lexicografía* (pp. 193-207). De Gruyter. https://doi.org/10.1515/9783110420784-015

Recursos propios

CombiContext = Domínguez Vázquez, M.ª J. (dir.), Valcárcel Riveiro, C., Bardanca Outeiriño, D., Calañas Continente, J. A., Catalá Torres, N., Martín Gascueña, R., Mirazo Balsa, M., Sanmarco Bande, M. T. & Pino Serrano, L. (2021). *CombiContext. Prototipo online para la generación automática de contextos frasales y oraciones de la frase nominal en alemán, español y francés*. Universidade de Santiago de Compostela. Consultado el 28 de noviembre de 2022. http://portlex.usc.gal/combinatoria/verbal

Combina = Recuperado el 28 de noviembre de 2022, de http://portlex.usc.gal/develop/combina.php

Combinatoria = Domínguez Vázquez, M.ª J. (dir.), Valcárcel Riveiro, C., Bardanca Outeiriño, D., Calañas Continente, J. A., Catalá Torres, N., López Iglesias, N., Martín Gascueña, R., Mirazo Balsa, M., Sanmarco Bande, M. T. & Pino Serrano, L. (2020). *Combinatoria. Prototipo online para la generación biargumental de la frase nominal en alemán, español y francés*. Universidade de Santiago de Compostela. Instituto da Lingua Galega. Consultado el 28 de noviembre de 2022. http://portlex.usc.gal/combinatoria/usuario

Flexiona = Consultado el 28 de noviembre de 2022. http://portlex.usc.gal/develop/flexiona.php

Flexionador = Consultado el 28 de noviembre de 2022. https://ilg.usc.gal/flexionador/

Lematiza = Consultado el 28 de noviembre de 2022. http://portlex.usc.gal/develop/lematiza/

Ontología léxica = Domínguez Vázquez, M.ª J., Valcárcel Riveiro, C. & Bardanca Outeiriño, D. (2021). *Ontología léxica*. Universidade de Santiago de Compostela. Consultado el 28 de noviembre de 2022. http://portlex.usc.gal/ontologia/

Portlex = M.ª J. Domínguez Vázquez (dir.), Valcárcel Riveiro, C., Mirazo Balsa, M., Sanmarco Bande, M. T., Simões, A. & Vale, M. J. (2018). Portlex. Diccionario multilingüe de la valencia del nombre. Universidade de Santiago de Compostela. Consultado el 28 de noviembre de 2022. http://portlex.usc.gal/portlex/

TraduWord = Consultado el 28 de noviembre de 2022. https://ilg.usc.gal/es/proxectos/interoperabilidad-de-recursos-y-produccion-automatica-de-lenguaje-natural-0

Xera = Domínguez Vázquez, M.ª J. (dir.), Valcárcel Riveiro, C., Bardanca Outeiriño, D., Calañas Continente, J. A., Catalá Torres, N., López Iglesias, N., Martín Gascueña, R., Mirazo Balsa, M., Sanmarco Bande, M. T. & Pino Serrano, L. (2020). *Xera. Prototipo online para la generación automática monoargumental de la frase nominal en alemán, español y francés*. Universidade de Santiago de Compostela. Consultado el 28 de noviembre de 2022. http://portlex.usc.gal/combinatoria/usuario

XeraWord = Domínguez Vázquez, M.ª J. (dir.), Bardanca Outeiriño, D., Caíña Hurtado, M., Gómez Guinovart, X., Iglesias Allones, J. J., Simões, A., Valcárcel Riveiro, C., Álvarez de la Granja, M. & Cidrás Escaneo, F. A. (2020). *XeraWord. Prototipo de xeración automática da argumentación da frase nominal en galego e portugués*. Santiago de Compostela: Instituto da Lingua Galega. Consultado el 28 de noviembre de 2022. http://ilg.usc.gal/xeraword/

Recursos externos

Open Multilingual Wordnet = Bond, F. & Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. En H. Schuetze, P. Fung, M. Poesio (eds.), *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013* (pp. 1352-1362). Association for Computational Linguistics.

FreeLing = Consultado el 28 de noviembre de 2022. http://nlp.lsi.upc.edu/freeling/

Galnet = Consultado el 28 de noviembre de 2022 http://sli.uvigo.gal/galnet/

MCR = Multilingual Central Repository. Consultado el 28 de noviembre de 2022. https://adimen.si.ehu.es/web/MCR

MyMemmory = Consultado el 28 de noviembre de 2022. https://mymemory.translated.net/

PULO = Consultado el 28 de noviembre de 2022. http://wordnet.pt/

Spacy = Consultado el 28 de noviembre de 2022. https://spacy.io/; Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*, 411-420.

Sketch Engine = Consultado el 28 de noviembre de 2022. https://www.sketchengine.eu

PULO = Consultado el 28 de noviembre de 2022. http://wordnet.pt/

UWN/MENTA = de Melo, G. & Weikum, G. (2010). Towards Universal Multilingual Knowledge *Bases*. En P. Bhattacharyya, C. Fellbaum & P. T. J. M. Vossen (eds.) (2010), *Principles, construction and application of Multilingual Wordnets. Proceedings of the Fifth Global WordNet Conference* (pp. 149-156). Narosa Publishing House. https://doi.org/10.1145/1871437.1871577

Wikimedia = Consultado el 28 de noviembre de 2022. https://commons.wikimedia.org/wiki/Main_Page

WOLF = Wordnet Libre du Français – Sagot, B. & Fišer, D. (2008). Building a free French wordnet from multilingual resources. En N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, D. Tapias (eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC'08). European Language Resources Association (ELRA).

WordNet = https://wordnet.princeton.edu/

DISEÑO DE UNA ONTOLOGÍA DE SEMÁNTICA LÉXICA EN LOS PROYECTOS MULTIGENERA Y MULTICOMB

DESIGN OF A LEXICAL SEMANTICS ONTOLOGY IN THE MULTIGENERA AND MULTICOMB PROJECTS

Rosa María Martín Gascueña
Universidad Complutense de Madrid
rosamartingascuena@filol.ucm.es

Resumen

El trabajo se enfoca en la presentación del diseño, desarrollo y funcionalidad de una ontología en los proyectos *MultiGenera*[1] y *MultiComb*[2], que se llevan a cabo en múltiples idiomas (español, alemán y francés) con el propósito de generar de manera automática frases nominales (FN) y sus contextos oracionales. La ontología juega un papel fundamental en estos proyectos al formalizar las propiedades esenciales de los elementos léxicos siguiendo principios lingüísticos teóricos de la gramática y de la lexicografía valencial. Esta ontología proporciona etiquetas de marcado semántico para las clases semánticas que conforman el entorno sintagmático de veinte sustantivos de diversas áreas de conocimiento. Su objetivo es organizar los datos léxicos para su almacenamiento en una base de datos y su posterior recuperación a través de las aplicaciones prototipo *Xera* y *Combinatoria*. En las que se generan automáticamente las estructuras nominales de los sustantivos y su combinación en el contexto oracional.

El proceso de elaboración de la ontología se divide en dos fases: en la primera, se utilizan las mismas clases semánticas que en *Portlex* para clasificar los actantes de cada nombre empleados para la generación automática de FN en las tres lenguas en *Xera* 1.0. En la segunda fase, se reconfigura la ontología al conectar automáticamente con *synsets* de WordNet y otras ontologías relacionadas para aumentar la base de datos léxica de los proyectos. Esta versión posterior de la ontología presenta una mayor granularidad, lo que facilita la programación para generar los entornos de los nombres y sus combinaciones en contextos oracionales. La ontología se aplica para aprender idiomas gracias a su interfaz amable en las diversas aplicaciones. Los usuarios pueden utilizarla para seleccionar posibles combinaciones léxicas en FN. Es importante señalar que la ontología está disponible en línea y sigue en constante desarrollo.

Palabras clave: lexicografía, ontologías, categorías semánticas, frase nominal.

Abstract

The aim of this paper is to present the design, development, and functionality of the ontology in the *MultiGenera* and *MultiComb* multilingual projects, (Spanish, German, and French) for the automatically generation of the noun phrases (FN) and its sentence contexts. It is important to highlight the primordial role of this ontology for the development of these, as it involved a formalisation of the fundamental properties of lexical items in accordance with valential grammar and lexicography. The ontology provides semantic tagging labels for the semantic classes that make up the syntagmatic environment of the twenty nouns, belonging to different areas of knowledge. Its purpose is to organise the lexical data for storage in the database and subsequent retrieval through the prototype applications *Xera* and *Combinatoria*, which automatically generate the nominal structures of the nouns and their combination in sentence contexts.

The ontology development process is divided into two phases: in the first phase, the same semantic classes as in *Portlex* to classify the actants of each noun used for the automatic generation of FN in the *Xera* 1. 0. in the three languages. In the second phase, the ontology is reconfigured by automatically connecting to the WordNet database *synsets* and other related ontologies to increase the lexical database of the projects. The latter ontology presents greater granularity, which favours programming for the generation of name environments and their sentence combination. The application of the ontology in language learning is shown by the user-friendly interface in various applications to select the possible lexical items combinations in FN. It should be noted that the ontology is still under development and is available online.

Keywords: lexicography, ontologies, semantic categories, noun phrase.

[1] *MultiGenera. Generación multilingüe de estructuras argumentales del sustantivo y automatización de extracción de datos sintáctico-semánticos*. Fundación BBVA. Ayudas Fundación BBVA a Equipos de Investigación Científica - Humanidades Digitales. 2017-2020. http://portlex.usc.gal/multigenera/

[2] *MultiComb. Generador multilingüe de estructuras argumentales del sustantivo con aplicación en la producción en lenguas extranjeras*. FI2017-82454-P: Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia, Generación de Conocimiento. MCIN/ AEI/ FEDER "Una manera de hacer Europa" (EXCELENCIA 2017, 2017-PN091). 2018-2021. http://portlex.usc.gal/multicomb/

DOI: https://doi.org/10.17561/rilex.6.3.8083

1. INTRODUCCIÓN

El objetivo de este trabajo es presentar la ontología de los proyectos lexicográficos, multilingües (español, alemán y francés), consecutivos y complementarios *MultiGenera* y *MultiComb*, disponibles en la red. Al comenzar el estudio, las primeras cuestiones que se plantearon fueron definir el concepto ontología en esta investigación, cómo sería su elaboración y cuál sería su utilidad en el desarrollo de los proyectos.

Se revisaron diferentes definiciones: etimológicamente, el término *ontología* significa el estudio del Ser, aunque a finales del siglo XX, adquirió otra significación en el ámbito informático y en la documentación. En el primero, una ontología designa una colección de objetos diferentes relacionados (Lenat & Guha, 1990)[3] y se define como "An *ontology* is an explicit specification of a conceptualization" (Gruber, 1993, p. 1), para otros autores "una ontología es una especificación explícita y formal de una conceptualización compartida" (Borst, 1997; Studer, Benjamins & Fensel, 1998, p. 25); una ontología es un tipo especial es un tipo especial de objeto de información o artefacto computacional (Guarino, Oberle & Staab, 2009). En el Procesamiento del Lenguaje Natural (PLN) y la Inteligencia Artificial (IA), las ontologías se emplean para modelizar formalmente la estructura de un sistema, las entidades y sus relaciones. Y para algunos autores, la base de datos WordNet (Miller, Beckwith, Fellbaum, Gross & Miller, 1990) es una ontología léxica utilizada como modelo para crear otras ontologías como EuroWordNet[4] (Vossen, 1998) (vid. 3). Por otro lado, en el ámbito documental, una ontología es una tecnología

[3] Lenat y Guha (1990) en el proyecto CYC, sobre inteligencia artificial que combina bases de datos y ontología: "the ontology of Cyc is organized around the concept of categories. We also refer to them as classes or collections. The categories are organized in a generalization-specialization hierarchy" (1990, p. 42).

[4] La ontología en EuroWordnet (Vosse, 1998) se define como un modelo abstracto, una estructura interpretable por el computador, que define las relaciones e implicaciones necesarias para modelar conceptualmente un área de conocimiento.

para los sistemas de información. Algunos conciben la ontología como un tesauro: "una lista controlada y estructurada de términos para el análisis temático y la búsqueda de documentos"[5] o como una jerarquía taxonómica cuyos componentes principales son las clases (p. e. *mamíferos*), las subclases (p. e. *gatos*) y los individuos o las instancias (p. e., *siamés*) más sus relaciones, que deben estar codificadas en un lenguaje informático con una base lógica para especificar las relaciones e inferencias entre ellas, y además, pueden centrarse en un área de conocimiento concreto (Codina & Pedraza-Jiménez, 2011).

La definición adoptada aquí considera la ontología como un modelo conceptual abstracto que describe las relaciones e implicaciones léxicas necesarias para modelar áreas de conocimiento y, es interpretable computacionalmente. En otras palabras, una ontología es un constructo de categorías semánticas relacionadas dentro en un dominio.

La motivación para elaborar y desarrollar una ontología propia estaba justificada, porque, en principio, ninguna otra ontología de libre acceso podía dar cuenta de las restricciones combinatorias de los sustantivos en la frase nominal (FN), tal como se plantea en la gramática y lexicografía valencial (Domínguez Vázquez, 2011; Domínguez Vázquez, Valcárcel Riveiro & Lindemann, 2018) base de la investigación. Así pues, se decidió crear una ontología a medida para el tratamiento computacional de los datos léxicos, que, además, pudiera emplearse con fines didácticos en las aplicaciones lexicográficas de los proyectos y que en un futuro sirva como índice interlingüístico entre las tres lenguas.

La ontología inicial es diseñada *ad hoc*; a partir de las categorías semánticas definidas en el proyecto anterior[6]. Se emplea para organizar el léxico que

[5] Thesauro de la Unesco http://vocabularies.unesco.org/browser/thesaurus/en/?clang=es [consulta 20/05/2021]
[6] Proyecto anterior: *Portlex*.

nutre las aplicaciones de generación automática de las frases nominales. De este modo, se concibe una estructura de clases o categorías semánticas que contribuyen a formalizar las relaciones sintagmáticas en la FN y generar automáticamente la combinatoria del sustantivo a partir de esas categorías semánticas. Posteriormente, la ontología se enlaza automáticamente con WordNet y otras ontologías relacionadas con ella para incrementar los elementos léxicos de las clases semánticas, como se explicará en el apartado 4. Esto conlleva un cambio en la configuración inicial de la ontología: favorecer la formación del léxico en las aplicaciones *Xera* y *Combinatoria* como se explicará en el apartado 4.2.

El empleo de ontologías en aplicaciones informáticas es frecuente, aunque no tanto el hecho de que las clases semánticas de la ontología se muestren en la interfaz del usuario, como en los prototipos *Xera* y *Combinatoria*, para que este elija entre las posibles opciones y así se generen automáticamente frases nominales y su contexto oracional en varias lenguas. Esta propuesta novedosa es de gran utilidad para la enseñanza de lenguas y la traducción automática.

Este trabajo se divide en varios apartados, el 2 presenta el marco teórico del que parte la ontología; el 3 indica las ontologías léxicas con las que se enlazará automáticamente y presenta trabajos relacionados; en el 4 se expone la metodología empleada para el diseño de la ontología y las aplicaciones; el 5 expone las conclusiones y, para finalizar; en el 6 están las referencias bibliográficas.

2. FUNDAMENTOS LINGÜÍSTICOS DE LA ONTOLOGÍA

En este apartado se resumen los fundamentos lingüísticos teóricos de esta investigación de los que se parte para la creación de la ontología. Estos residen en la gramática y lexicografía valencial (Domínguez Vázquez, 2011; Domínguez Vázquez, Valcárcel Riveiro & Lindemann, 2018), los principios

de categorización, la teoría del prototipo (Rosch, 1978), los corpus de Sketch Engine[7] y el concepto de *synsets* de WordNet (1987).

2.1. *El análisis valencial del sustantivo*

La valencia del sustantivo se identifica con sus características sintáctico-semántico particulares. Esto implica que cada sustantivo selecciona unos elementos léxicos, denominados *actantes*, y cada actante desempeña una función semántica (agente, experimentante…) en consonancia con unas propiedades semánticas determinadas; por ejemplo, el sustantivo *dolor* tiene un significado que se relaciona con un actante (experimentante) con el rasgo semántico [animado] como *el dolor de su madre*. La ontología se concibe para formalizar semánticamente el análisis valencial de los sustantivos (Domínguez Vázquez, 2011).

En estos proyectos se ha trabajado con veinte sustantivos de cinco áreas léxico-conceptuales diferentes (Tabla 1): *Expresión, Clasificación Afección, Locación-situación, Locación-dirección*, según la clasificación de un proyecto anterior para un diccionario multilingüe sobre la valencia del nombre (*Portlex*).

Expresión	Clasificación	Afección	Locación-situación	Locación-dirección
1. Pregunta 2. Respuesta 3. Conversación 4. Discusión 5. Texto 6. Video	1. Olor 2. Sabor 3. Color 4. Ancho	1. Amor 2. Aumento 3. Dolor 4. Muerte	1. Presencia 2. Ausencia 3. Estancia	1. Huida 2. Viaje 3. Mudanza

Tabla 1: *Áreas léxico-conceptuales para los sustantivos en español*

La valencia de cada sustantivo en algunos casos coincide en las tres lenguas, pero en otros puede presentar peculiaridades formales distintas, aunque

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

compartan la misma perspectiva onomasiológica. El comportamiento actancial de cada sustantivo se analiza indicando los roles semánticos que desempeñan sus argumentos. En el esquema actancial de los sustantivos se indican los actantes o argumentos que se completan con unidades léxicas con unos rasgos categoriales determinados: [Animado] [Humano]... Así, por ejemplo, el sustantivo *amor,* que pertenece al *área de la afección*, presenta una valencia de tres posibles actantes A1 (1), A2 (2), A3 (3), cuyos roles semánticos son para A1: aquel / aquello que experimenta un nuevo estado o situación, agente; A2: aquel/aquello que tiene o dispone de algo; A3: aquel/aquello no afectado, tema.

(1) El intenso amor de Pepe. (A1) [Animado] [Nombre Propio]
(2) Un sincero amor por la humanidad. (A2) [Animado] [Humano]
(3) Un profundo amor a los libros. (A3) [Inanimado][Objeto]

Después de analizar el comportamiento actancial de todos los sustantivos se hicieron búsquedas en los corpus de Sketch Engine para encontrar las palabras más frecuentes que cubrían los diferentes actantes y a partir de ahí categorizar los prototipos léxicos. Este proceso fue muy importante para la construcción de las categorías y relaciones semánticas de la ontología (vid. 4.1.1).

2.2. *El proceso de categorización semántica*

Lakoff (1987, p. 5) indica que: "There is nothing more basic than categorization to our thought, perception, action, and speech. Every time we see something as a kind of thing, for example, a tree, we are categorizing". La categorización es una de las operaciones lingüísticas de conceptualización que intervienen en la construcción del significado. Es una actividad mental que implica organizar, estructurar y agrupar elementos que comparten rasgos de significado conceptual mediante propiedades y funciones. Al categorizar se produce un ajuste de nuestro foco de atención hacia ciertas características ignorando otras y así, creamos categorías conceptuales que se definen como

construcciones teóricas abstractas formadas por unidades léxicas con propiedades comunes. Así, *discusión;* y *texto*; se incluyen dentro del área conceptual de la *expresión* (Tabla 1). Estas clasificaciones representan la forma en que articulamos nuestra experiencia del mundo para hacerlo manejable. Una categoría conceptual incluye conceptos y a su vez, en sí misma, es un concepto, desde la perspectiva onomasiológica (Martín-Gascueña, 2013, p. 90).

La categorización es un fenómeno cognitivo relacionado desde la lógica, con la extensión e intensión del significado. El punto de vista extensional indica continuidad del significado, relacionado con la categorización, y parte de presupuestos lógicos para definir las relaciones de inclusión de significados (hiponimia y meronimia). La relación por excelencia en la creación de taxonomías es la hiponimia, caracterizada por ser implicativa, transitiva y asimétrica (Lyons, 1977, p. 274; Cruse, 1986, 2002, 2004; Brown, 2002). Por ejemplo, si *texto* está incluido en la categoría[8] de *expresión*, entonces todos los *textos* son *expresión,* pero no toda *expresión* es un *texto*; esto indica que ser un *texto* implica ser una expresión, aunque no al revés, los *textos* son un grupo dentro de la expresión, luego es una relación implicativa y asimétrica. En cuanto a la propiedad transitiva, está basada en la capacidad de contención de los significados y está condicionada por el contexto (Croft & Cruse, 2004). Por ejemplo, si *texto* es un hipónimo de *expresión* y *expresión* es un hipónimo de *comunicación*, entonces *texto* es un hipónimo de software *comunicación*. Asimismo, la intensión está relacionada con la composicionalidad del significado. La intensión implica una inclusión de significados de las unidades léxicas de las propiedades de niveles superiores o hiperónimos en las de niveles inferiores o hipónimos, caracterizados por heredar sus características y distinguirse por tener algún rasgo más. La combinación de extensión e intensión permite generar inferencias e implicaturas de significación.

[8] En este trabajo *categoría* y *clase semántica* son equiparables y ese utilizarán indistintamente.

Las categorías conceptuales o semánticas en la teoría clásica se definen por una lista de condiciones necesarias y suficientes que deben cumplir cualquiera de sus miembros para pertenecer a ella. Sus límites están definidos y la relación entre sus miembros es lineal. En la teoría del prototipo (Rosch, 1978), las categorías interesan por su organización interna en torno al prototipo, que es el mejor representante, cuya referencia sociocultural puede corresponderse con un ser real o ideal y se identifica con el Modelo Cognitivo Ideal (MCI) que es una estructura abstracta que interviene en los procesos de construcción del significado (Lakoff, 1987). Esta organización categorial determina las relaciones intercategorial e intracategorial de las unidades léxicas. Las clases semánticas organizan el conocimiento léxico en la ontología.

2.3. *La base de datos WordNet*

La base de datos del inglés WordNet se creó en 1985 por un grupo de psicolingüistas de la universidad de Princeton dirigidos por Miller y es considerada por algunos investigadores como una ontología léxica a pesar de que no fuera concebida como tal. Esta base de datos léxica está fundamentada en los modelos de memoria léxica basados en la semántica de Lyons (1977). La idea original era probar que el uso de diccionarios conceptuales era mejor que el simple orden alfabético. La característica más ambiciosa de WordNet es el intento de organizar la información léxica, desde la perspectiva onomasiológica, por los significados de las palabras, más que por su forma o visión semasiológica.

En esta base de datos, las palabras se diferencian por categorías gramaticales: sustantivos, adjetivos, verbos; su significante, forma física, y significado, referido a un concepto que puede estar representado por una o varias palabras. De este modo, el significado de una palabra se representa por un grupo de sinónimos denominados *synsets* que simbolizan conceptos, son *nodos* en la red léxica unidos por relaciones semánticas horizontales, sinonimia y verticales, hiperonimia y meronimia. El modelo de red de significados agrupados

en *synsets* por relaciones de inclusión y sinonimia también está presente en nuestra ontología.

WordNet representa un modelo de organización léxica, donde los sustantivos se clasifican en jerarquías, lo que implica un sistema de herencia de significados. Así pues, se establecen varios niveles de categorización: en el nivel más alto de la jerarquía está el significado más abstracto, la {entidad}, en el nivel inmediatamente inferior se incluye la categoría {objeto, cosa} e {idea}, a continuación, la clasificación entre *animados* e *inanimados* y posteriormente, se identifican con alguno de los veinticinco primitivos semánticos que funcionan como categorías conceptuales o semánticas cerradas (Tabla 2).

Primitivos Semánticos		
{act, action, activity}	{food}	{process}
{animal, fauna}	{group, collection}	{quantity, amount}
{artifact}	{location, place}	{relation}
{attribute, property}	{motive}	{shape}
{body, corpus}	{natural object}	{state, condition}
{cognition, knowledge}	{natural phenomenon}	{substance}
{communication}	{person, human being}	{time}
{event, happening}	{plant, flora}	
{feeling, emotion}	{possession}	

Tabla 2: *Los primitivos semánticos en WordNet* (Miller, Beckwith, Fellbaum, Gross & Miller, 1990)

La posibilidad de descargar y utilizar la base de datos WordNet con todo su entramado de relaciones semánticas ha permitido que muchos proyectos la utilizasen como una ontología léxica, aunque presentase algunas inconsistencias en las categorizaciones (Gangemi, Navigli & Velardi, 2003), que se fueron corrigiendo en sucesivas versiones como WordNet 1.7 y posteriores. En consecuencia, se ha transformado en una ontología léxica para representar el conocimiento. Distingue entre relaciones de subclases e individuos o instancias, y asocia identificadores únicos a cada categoría o clase (Martin,

2003). WordNet es considerada por algunos una ontología superior, porque las categorías propuestas en sus grupos de *synsets* y sus relaciones son modelos compartidos con otras ontologías que además han compartido datos. En consecuencia, WordNet funciona como un Índice Interlingüístico (ILI) entre otras ontologías de diferentes idiomas, facilitando la interoperabilidad semántica en la definición de términos.

Muchos proyectos utilizan WordNet como fuente para recuperar información y también para desambiguar significados. En muchos casos, sus categorías se han convertido en una especificación formal para extraer automáticamente asociaciones de información léxica de WordNets en distintas lenguas e interpretarlas como un conjunto de relaciones conceptuales (inclusión y sinonimia), formalmente definidas en la ontología. Este es el caso del proyecto del Istituto di Scienze e Tecnologie della Cognizione DOLCE[9] (Descriptive Ontology for Linguistic and Cognitive Engineering), creado por Masolo, Borgo, Gangemi, Guarino y Oltramari (2003).

WordNet ha sido un elemento clave para el desarrollo computacional de algunas ontologías de semántica léxica que se verán a continuación y con las que se ha enlazado automáticamente la ontología. Ninguna se consideró totalmente apropiada para basar los proyectos en ellas porque en principio, no respondía a las necesidades del análisis valencial (vid. 2.1) y se optó por diseñar y elaborar una propia, que partía de la investigación de *Portlex*.

3. ONTOLOGÍAS DE SEMÁNTICA LÉXICA

WordNet ha sido un elemento clave para el desarrollo de algunas ontologías de semántica léxica de código abierto con las que se ha enlazado la ontología de esta esta investigación, a través de la WordNet del gallego: Galnet (Solla Portela & Gómez Guinovart, 2015). Así las clases semánticas

[9] DOLCE https://www.istc.cnr.it/it/content/dolce-descriptive-ontology-linguistic-and-cognitive-engineering

han incrementado su léxico al conectar automáticamente con las ontologías que presentamos a continuación por orden cronológico de creación (vid. 4.1.2).

3.1. *Top Ontology*

Top Ontology (TOP) es una jerarquía independiente de clases, diseñadas para agrupar, comparar e intercambiar conceptos entre idiomas en EuroWordNet[10] (Rodríguez, Climent, Vossen, Bloksma, Peters, Alonge, Bertagna & Roventini, 1998). Está basada en relaciones semánticas de inclusión y sinonimia como en WordNet 1.3, aunque la categorización semántica del léxico se basa en el lexicón generativo de Pustejovsky (1995, p. 76). En concreto, en la estructura de *qualia* o modos de explicación, formada por diferentes roles o *qualia: formal*, representa la información que distingue el contenido de una palabra de otras relacionadas con ella, dentro de su dominio correspondiente; *constitutivo*, recoge la información sobre la entidad y sus partes; *télico*, especifica la finalidad de la entidad y *agentivo*, detalla los factores que originan la existencia de la entidad.

La TOP es un repositorio de información de semántica léxica que ha aumentado con más rasgos los *synsets* de WordNet y el resultado es WordNet 1.6. La TOP está programada con el lenguaje de marcado OWL (Web Ontology Language), y es una jerarquía que consta de 63 características organizadas en tres tipos diferentes de entidades: 1stOrderEntity: cosas físicas (imagen); 2ndOrderEntity: eventos, estados y propiedades (imagen); 3rdOrderEntity: entidades no observables (Vossen, 1998).

Esta ontología no respondía a la división en área conceptuales que se plantea en la ontología de los proyectos para facilitar el tratamiento computacional.

[10] EuroWordNet (1996) es una base de datos léxica, multilingüe para el neerlandés, el italiano y el español, cada lengua posee sus propios WordNets enlazados con la base de datos WordNet 1.5, unidos por un índice interlingüístico ILI de significados que es la ontología superior, denominada Top Ontology.

3.2. *Suggested Upper Merged Ontology*

Suggested Upper Merged Ontology (SUMO), diseñada por Teknowledge Corporation (Pease, Niles & Li, 2002), es una ontología de comunicación computacional básica para distintos sistemas informáticos de procesamiento de información. SUMO es un esquema jerárquico de clases o categorías, reglas de relación y relaciones que establece enlaces con los *synsets* de WordNet 1.3. En principio, era una ontología de nivel general, que incluía niveles medios de especificación. Se fue ampliando con otras muchas ontologías de dominio específico. Un hecho que hay que resaltar es que la nomenclatura empleada en la ontología sigue estándares que le permiten emplear los mismos significados y maximizar su compatibilidad. Por este motivo la SUMO permite la interoperatividad con otras aplicaciones de razonamiento automatizados.

Esta ontología de carácter general fue concebida para la comunicación entre máquinas, no para establecer clases o categorías con mayor especificidad que permitieran la programación con mayor granularidad.

3.3. *WordNet Domains*

WordNet Domains, desarrollado en el Instituto per la Ricerca Scientifica e Tecnológica (Bentivogli, Forner, Magnini & Pianta, 2004), es un recurso léxico creado de forma semiautomática. Posee un conjunto de unas doscientas etiquetas de *dominios básicos* con un grado de granularidad adecuado para su procesamiento computacional aplicado a la categorización de textos y a la desambiguación del sentido de las palabras. Asimismo, WordNet se incrementa al anotar sus *synsets* con las etiquetas de dominio semántico de WordNet Domains Hierarchy (WDH). Estos dominios pueden incluir *synsets* de distintas categorías gramaticales, de diferentes niveles jerárquicos de WordNet y reunir varios sentidos de la misma palabra en clusters homogéneos, para reducir la polisemia de palabras en WordNet (Bentivogli, Forner, Magnini & Pianta, 2004).

WordNet Domains realiza el mapeo entre sus dominios básicos, los temas de WordNet 1.5 y las categorías emergentes de Wikipedia. De este modo, se

consigue una alineación aproximada entre WordNet y Wikipedia, útil para producir corpus multilingües y específicos de dominio. El multilingüismo se logra a través de enlaces entre las categorías de Wikipedia en diferentes idiomas. WordNet Domains se ha integrado en la base de datos léxica multilingüe MultiWordNet. En un principio enlazaba la WordNet del italiano con la versión 1.3 de WordNet, posteriormente se han incluido más WordNets del español, portugués, hebreo rumano y latín (Pianta, Bentivogli & Girardi 2002) y se distribuye bajo una licencia Creative Commons Attribution.

Enlazar con los dominios de WordNet excedía las dimensiones de *MultiGenera* y *MultiComb*, ya que no se centraba en ningún dominio conceptual específico, sino que tenía un carácter de clasificación de léxico general en varias lenguas.

3.4. *Galnet*

Galnet es una WordNet para el gallego que, mediante un índice interlingüístico enlaza con WordNet 3.0 en inglés y con otras ontologías como la TOP, la SUMO y WordNet Domains. Hay una transferencia de conocimiento de WordNet a Galnet, se estructura igual mediante relaciones de inclusión (hiponimia y meronimia) y sinonimia. A los *synsets* nominales se les denomina *epinónimos* y representan la categoría de un área semántica, a la que se asignarán automáticamente otros *synsets* mediante algoritmos que evaluarán su proximidad a través del tratamiento terminológico de las relaciones léxico-semánticas (Solla Portela & Gómez Guinovart, 2015). Galnet facilita la interoperatividad semántica entre las distintas ontologías indicadas, mediante la conversión de los formatos con el lenguaje de marcado LMF (Lexical Markup Framework) y la compatibilidad con los modelos OWL existentes, con lo que se consigue la conexión entre diferentes capas de información, permitiendo reutilizar datos.

La ontología de esta investigación ha expandido el léxico de sus clases semánticas gracias a Galnet, que ha permitido mediante su índice interlingüístico enlazar con todas las ontologías antes mencionadas.

En los siguientes apartados presentamos cómo se han utilizado estas ontologías y la influencia que han ejercido en la evolución de la ontología semántica de los proyectos lexicográficos cuyas aplicaciones *Xera* y *Combinatoria* crean ejemplos a partir de patrones argumentales valenciales y son prototipos para nuevos modelos de diccionarios plurilingües automáticos y dinámicos.

4. METODOLOGÍA

La cuestión que se planteó era cómo diseñar una ontología para estos proyectos que diera cuenta, desde la perspectiva semasiológica y onomasiológica, de las características semánticas de los prototipos léxicos para los diferentes sustantivos, en alemán, francés y español. La dificultad radicaba en establecer las etiquetas de las clases semánticas para almacenar y recuperar los datos léxicos, y de este modo, crear paradigmas semánticamente coherentes para la generación automática de la FN con cierta autonomía del contexto. Y, además, que, en un futuro, sirvieran como índice interlingüístico entre las distintas lenguas de los proyectos, de tal forma, que a través de la ontología se pudiera establecer equivalencias entre las lenguas para una aproximación comparativa y contrastiva del funcionamiento de estos sustantivos en alemán, español y francés.

Gangemi, Navigli y Velardi (2003) proponen distintos modelos para la construcción de una ontología, dependiendo de su utilidad; el enfoque *top down*, descendente, es el más adecuado para elaborar ontologías generales. En una estructura jerárquica se parte de los nodos superiores independientes de un dominio. El enfoque *bottom up*, ascendente, es para ontologías terminológicas, se intenta llegar a un nodo más general a partir de nodos locales. Y el modelo *híbrido* que intenta aprovechar los dos enfoques anteriores, descendente y ascendente. Asimismo, la construcción de la ontología puede completarse con otras ya existentes, según sea su finalidad conectando con ellas automáticamente. Hay ontologías construidas con modelos de funciones

léxicas para unir diferentes bases de datos y ontologías para comunicarse computacionalmente[11].

La ontología del proyecto es de carácter general y se pensó como una estructura conceptual que sirviera para formalizar el tratamiento computacional, tanto para el almacenaje como la recuperación y la programación de los sustantivos y sus actantes.

En el proceso de elaboración de la ontología, se siguió un modelo *híbrido*; en la primera fase, se siguió un modelo ascendente, *bottom up,* a partir de los datos léxicos se crearon y denominaron las clases semánticas en la ontología 0.1. En la segunda fase, el hecho de conectarse con otras ontologías para aumentar el léxico de los paquetes semánticos de los sustantivos obligó a revisar las clases de la ontología 0.1 para favorecer la conexión computacional. El modelo de referencia fue WordNet para incrementar los niveles categoriales y las clases semánticas, por este motivo, lo denominamos enfoque descendente, *top down*.

4.1. *Enfoque ascendente*

La ontología se concibió como una estructura jerárquica de categorías semánticas para organizar el léxico, facilitar el almacenaje, la recuperación y la programación de las aplicaciones. Siguiendo este enfoque *bottom up*, se decidió construir las clases semánticas a partir de los datos obtenidos en los corpus

[11] Actualmente, se trabaja en la creación de estándares internacionales para la representación de lexicones que posibiliten la comunicación entre máquinas y así permitan, unir datos y compartir recursos. Este es el caso del proyecto LEMON, Lexicon Model for Ontologies, que propone un modelo para modelar el léxico y hacer diccionarios legibles para la máquina, no contiene datos lingüísticos, sino que establece definiciones formales para la creación de diccionarios monolingües, inicialmente, en formato LMF, Lexical Markup Framework. A su vez, el proyecto Ontolex-Lemón (Bosque-Gil & García, 2019) propone un modelo de estandarización para la representación de lexicones computacionales, unido a las tecnologías OWL (Web Ontology Language) para obtener significado y RDF (Resource Description Framework), que relaciona datos por ejemplo triples <books> <haswriter> <writer> Así se consigue compartir los recursos léxicos fácilmente y construir otros recursos léxico-semánticos multilingües de acuerdo con estas tecnologías de Web Semántica y la nube de datos abiertos vinculados, para establecer equivalencias entre ellos para que puedan ser compartidos y reutilizados en las comunidades científicas internacionales (Bosque-Gil, 2019).

de Sketch Engine y así, establecer los niveles categoriales generales y las etiquetas semánticas, de acuerdo con el esquema básico de catorce rasgos (Domínguez Vázquez, 2011) utilizados en *Portlex* como puede verse en la Tabla 3.

Abreviaturas	Significado categorial
mat	material (concretos)
anim	animado
hum	humano
zool	animal
inst	instituciones
inanim	inanimado
mas	masa (concretos no contables, como por ejemplo el *agua*)
obj	objeto (inanimado contable)
plant	planta
inmat	inmaterial (abstractos)
intel	conceptos delimitables y contables (por ejemplo, *idea*) y sistemas reglados jerárquicos (por ejemplo, los *días, meses, comunismo*, etc.)
situ	situación
situest	situación estática
situdin	situación dinámica

Tabla 3: *Categorías semánticas propuestas en Portlex*

Los veinte sustantivos de las tres lenguas se analizaron valencialmente para obtener su esquema actancial (vid. 2) con el que se realizaron las búsquedas en los corpus de Sketch Engine (Figura 1), para extraer información mediante las *querys* o consultas sobre los actantes de cada sustantivo y así obtener las unidades léxicas más frecuentes y representativas de cada casilla funcional, o sea, sus prototipos semánticos. Los resultados obtenidos se descargaban en un fichero CSV (Comma-Separated Values), o XML y se etiquetaban semánticamente según las catorce categorías de la Tabla 3, de este modo se configuran los niveles categoriales de la ontología.

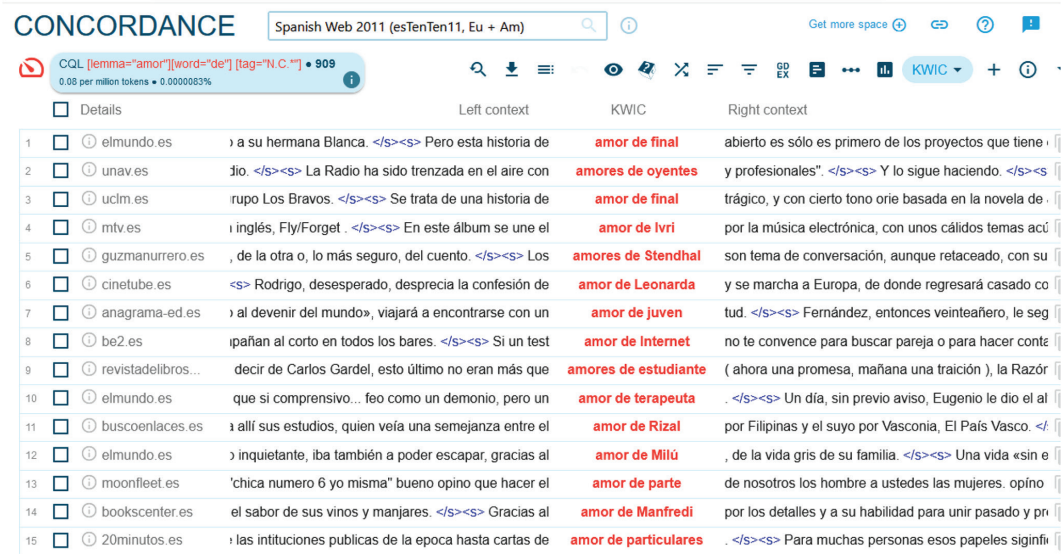


FIGURA 1: *Búsqueda del* actante 1 *(A1) del sustantivo* amor

Los nombres y adjetivos se ordenaron por frecuencias y etiquetaron dentro de tres niveles categoriales, el más abstracto, la dicotomía *material* e *inmaterial*. Así se obtuvieron los prototipos, la Tabla 4 muestra un ejemplo de los prototipos léxicos de otro sustantivo (*texto*) para su actante tema (A3) *El texto de la reforma*.

COLOCACIÓN	Género	1 NIVEL	2 NIVEL	3 NIVEL	FRE	LL SCORE
tema	m.	material	inanimado	objeto	4.397	8.36
reforma	m.	material	inanimado	objeto	842	8.62
ley	f.	material	inanimado	objeto	746	8.35
proyecto	m.	material	inanimado	objeto	573	7.55
problema	f.	material	inanimado	intelectual	486	7.40
asunto	m.	inmaterial	inanimado	intelectual	470	7.30

TABLA 4: *Categorías semánticas de prototipos del A3 de texto*

El siguiente paso, tras destacar los prototipos léxicos, fue agrupar las palabras que compartían los mismos rasgos en ficheros, denominados aquí *paquetes semánticos paradigmáticos*, para almacenarlos en la base de datos. Esta

modelización de los datos en sinónimos y cohipónimos se relaciona con el concepto de *synsets* de WordNet (vid. 3). Estos paquetes contenían sustantivos o adjetivos con características semánticas afines e identificados con dos niveles categoriales.

En definitiva, el enfoque es ascendente, porque las clases semánticas de la ontología se proponen a partir de las características semánticas de los prototipos en las tres lenguas, alemán, español y francés. Esta ontología 0.1 se aplicó en *MultiGenera* en la aplicación *Xera* versión 0.1 del generador monoargumental de la FN (vid. 4.1.1).

4.1.1. Proceso de desarrollo de la ontología 0.1

Los datos léxicos etiquetados y almacenados con la ontología 0.1 resultaban escasos. Entonces se enlazó con otras ontologías de código abierto (vid. 3) para incrementar el vocabulario de los paquetes semánticos de cada uno de los veinte sustantivos, a través de diferentes APIs (Application Programming Interfaces) (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019), API1 (Lematiza)[12] y API2 (Combina)[13] creadas para estos proyectos y disponibles en la web (Gómez Guinovart & Solla Portela, 2018).

En el proceso, primero, se subieron los paquetes semánticos de los sustantivos en las tres lenguas a la API 1, para aumentar su número de palabras. Esta API conectaba con otras ontologías: Word Domains, SUMO, TOP, Basic Level Concept, conceptos frecuentes de WordNet y Epinonyms, a través de la WordNet del gallego: Galnet (vid. 3). La Figura 2 muestra la interfaz desde donde se asociaban los paquetes semánticos con estas ontologías para la expansión léxica.

[12] (API1) Conecta las palabras de un fichero a través de Galnet con otras ontologías relacionadas con WordNet. Disponible en http://portlex.usc.gal/develop/lematiza/
[13] (AP2) Disponible en http://portlex.usc.gal/develop/combina.php

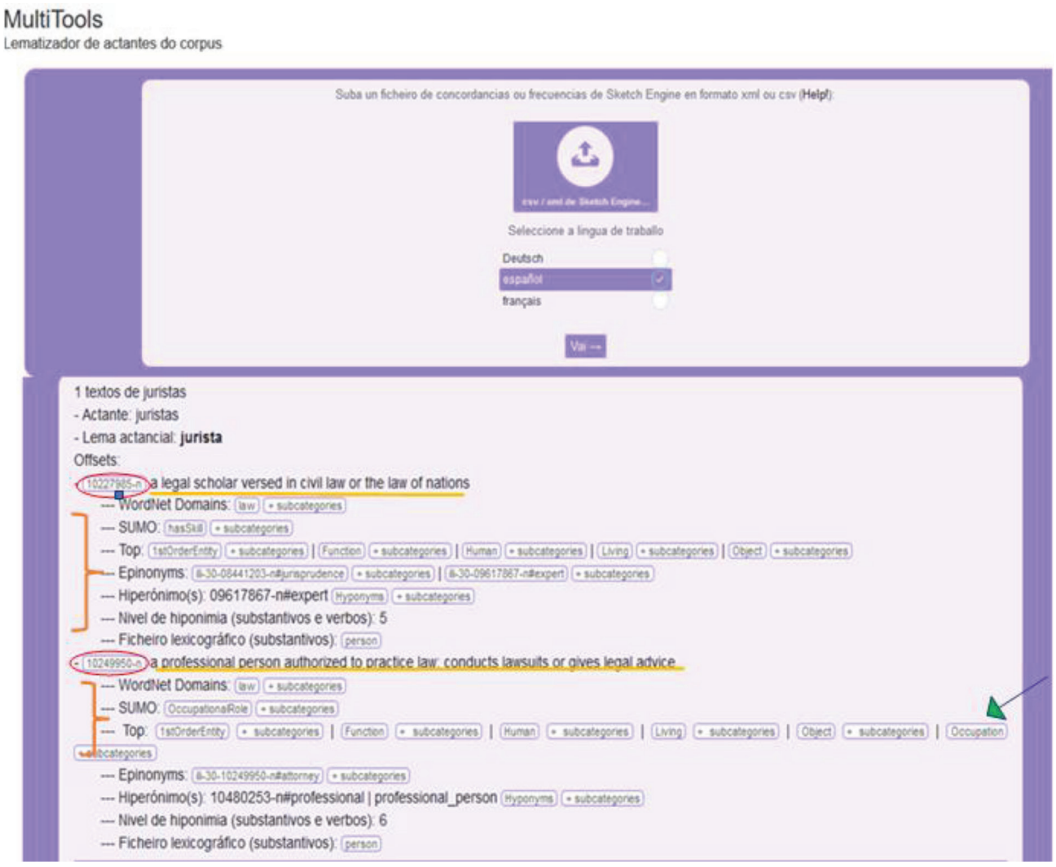


FIGURA 2: *API 1 para expandir el paquete de* texto

Los actantes lematizados podían enlazar con los índices u *offsets* de los distintos *synsets* o nodos léxicos de un dominio o área conceptual con los que se correspondía el lema y el nivel categorial al que se asocia en las diferentes ontologías relacionadas con WordNet. Por ejemplo, la Figura 2 muestra un ejemplo del fichero del actante 1 de *texto,* el primer lema obtenido(*jurista*) se puede asociar al offset=10227985-n y al offset=10249950-n de los *synsets* asociados a distintas áreas de conocimiento y las categorías de diferentes ontologías.

Una vez seleccionados los *offsets*, en las ontologías, se eligieron las categorías más cercanas a los prototipos, se anotaron sus direcciones en la AP2 (Figura 3) para mapear los datos entre las distintas ontologías y se

incrementó el número de palabras con las mismas características semánticas en los paquetes semánticos, correspondientes a los actantes de un sustantivo, por ejemplo [humano-cargo] del sustantivo *texto*.

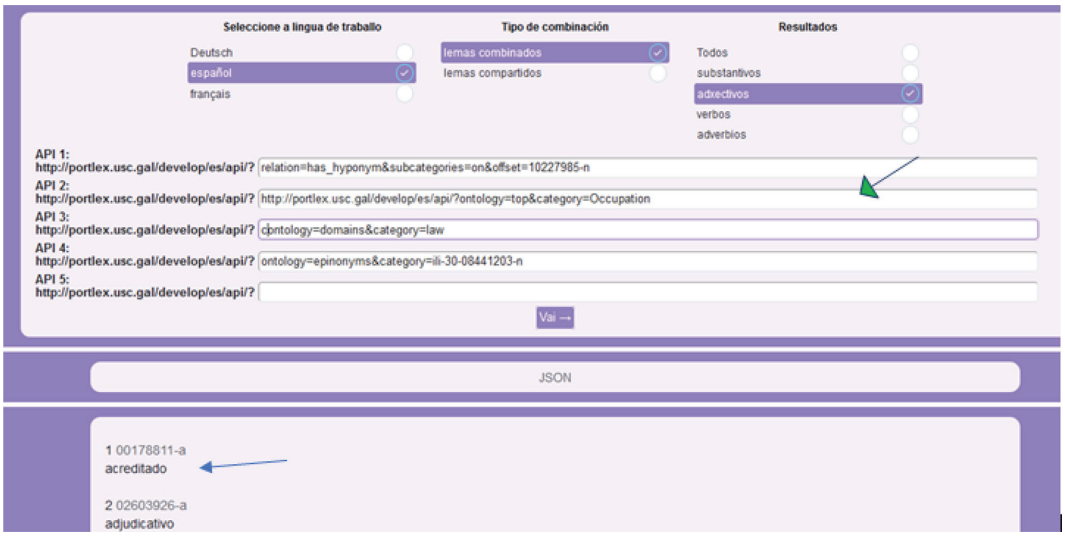


FIGURA 3: *Pantalla de API2 con mapeo a las ontologías*

Entonces, los documentos .txt obtenidos en (API2) se subieron a la API3 (Flexiona)[14], un flexionador morfológico, y los resultados se depuraron de agramaticalidades e incoherencias originadas por la regularización morfológica. Todos los paquetes semánticos expandidos, es decir, incrementados con más datos léxicos y etiquetados con la ontología 0.1 se emplearon para programar en la *Xera* 0.1 (Figura 4).

4.1.2. Aplicación de la ontología 0.1 a *Xera* (0.1)

La ontología 0.1 se concibió como una estructura con varios niveles de clases semánticas que podían ser argumentos de un sustantivo concreto. Se muestra la interfaz de la primera versión del generador monoargumental de

[14] (API3) http://portlex.usc.gal/develop/flexiona.php

Figura 4: *O: Sustantivo* texto *en* Xera *0.1*

la FN en *Xera1*.0. En la Figura 4, vemos su funcionamiento, al seleccionar *texto* se despliega un menú con su estructura actancial. Aquí aparece seleccionado [Determinante] + texto + [Actante N1] y debajo las clases semánticas que pueden cubrirlo, la elección [humano-cargo] genera las frases nominales mostradas.

4.2. *Enfoque descendente*

Cuando la ontología 0.1 se enlazó automáticamente con otras ontologías (vid. 3) se planteó la necesidad de ampliar y reajustar los niveles categoriales para

establecer equivalencias y favorecer la conexión de datos con otras ontologías. Se vio la necesidad de establecer categorías equiparables para facilitar la conexión de datos léxicos y la interoperabilidad de recursos. Al reelaborar la ontología original se tuvo como referencia WordNet y las ontologías relacionadas, lo que supuso un enriquecimiento de la estructura de esta ontología léxica, se ampliaron las categorías superordinadas y la granularidad en los niveles más específicos, para favorecer la programación en *Xera* y *Combinatoria*. Los niveles superiores aumentaron en tres, desde la categoría más abstracta: *material/inmaterial* de la ontología 0.1 (Figura 1, vid. 4.1.1). El resultado fue la ontología 0.2 con ocho niveles de inclusión, y dividida en ocho áreas de conocimiento o dominios: *Mundo animado; Mundo material; Mundo intelectual: Comunicación, pensamiento y cognición; Eventos dinámicos: procesos y actividades; Situaciones estáticas y condiciones/estados; Locación; Tiempo y Cantidad* (Figura 5).

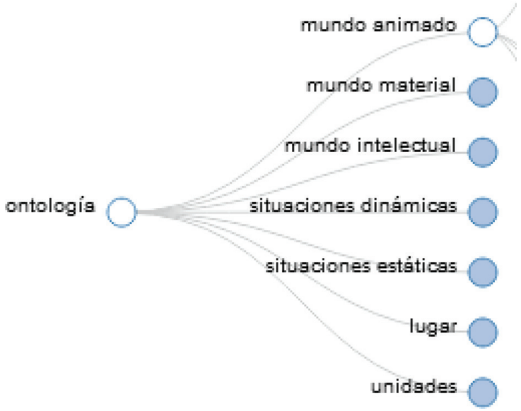


Figura 5: *Dominios de la ontología de* MultiGenera *y* MultiComb

Los dominios o áreas conceptuales son los hiperónimos que representan los niveles más abstractos e incluyen dos niveles intermedios hasta llegar al marcado como nivel 1, que coincide con el nivel superior de la ontología 0.1, desde el cual aumenta la granularidad hasta en cuatro niveles de subordinación (Figura 6).

			Nivel 1	Nivel2	Nivel 3	Nivel 4	Nivel 5
Mundo animado	Ser humano	Ser humano y condición/relación					
			animado	humano	general		
			animado	humano	origen		
			animado	humano	familia		
			animado	humano	personaje histórico		
			animado	humano	condición humana	negativa	
			animado	humano	condición humana	positiva	
			animado	humano	humana	acto negativo	

FIGURA 6: *Detalle de la ontología*[15]

En definitiva, la ontología 0.2 es el resultado de la evolución de la ontología 0.1 con más niveles categoriales, que facilitan la conexión con otras aplicaciones y restringen la selección de los argumentos semánticos en el procesamiento computacional. De esta manera, se optimizó la modelización y formalización del significado del léxico en la versión *Xera* versión 0.2 y *Combinatoria*.

4.2.1. La ontología 0.2 en *Xera* Versión 2 y *Combinatoria* 2

La ontología 0.2 se emplea en la base de datos de los proyectos, en la programación de las aplicaciones, y sus categorías semánticas se utilizan en la comunicación con el usuario de las aplicaciones *Xera* 0.2 y *Combinatoria* para seleccionar los argumentos que generan automáticamente la FN.

La aplicación *Xera* 0.2 de libre acceso presenta una interfaz más amigable para el usuario que la versión 0.1 y genera FN de un solo argumento. El funcionamiento es similar a la anterior, se elige el idioma, uno de los 20 sustantivos y para la consulta, primero se selecciona la estructura sintáctica formal y después las clases semánticas. Esto generará los ejemplos siguiendo

[15] Disponible en http://portlex.usc.gal/ontologia/

el filtrado realizado por los prototipos léxicos, representado por categorías semánticas, adjudicables a cada argumento. La Figura 7 muestra el ejemplo de *texto* con el actante A1 y las posibles combinaciones semánticas con las propiedades [animado humano familiar].

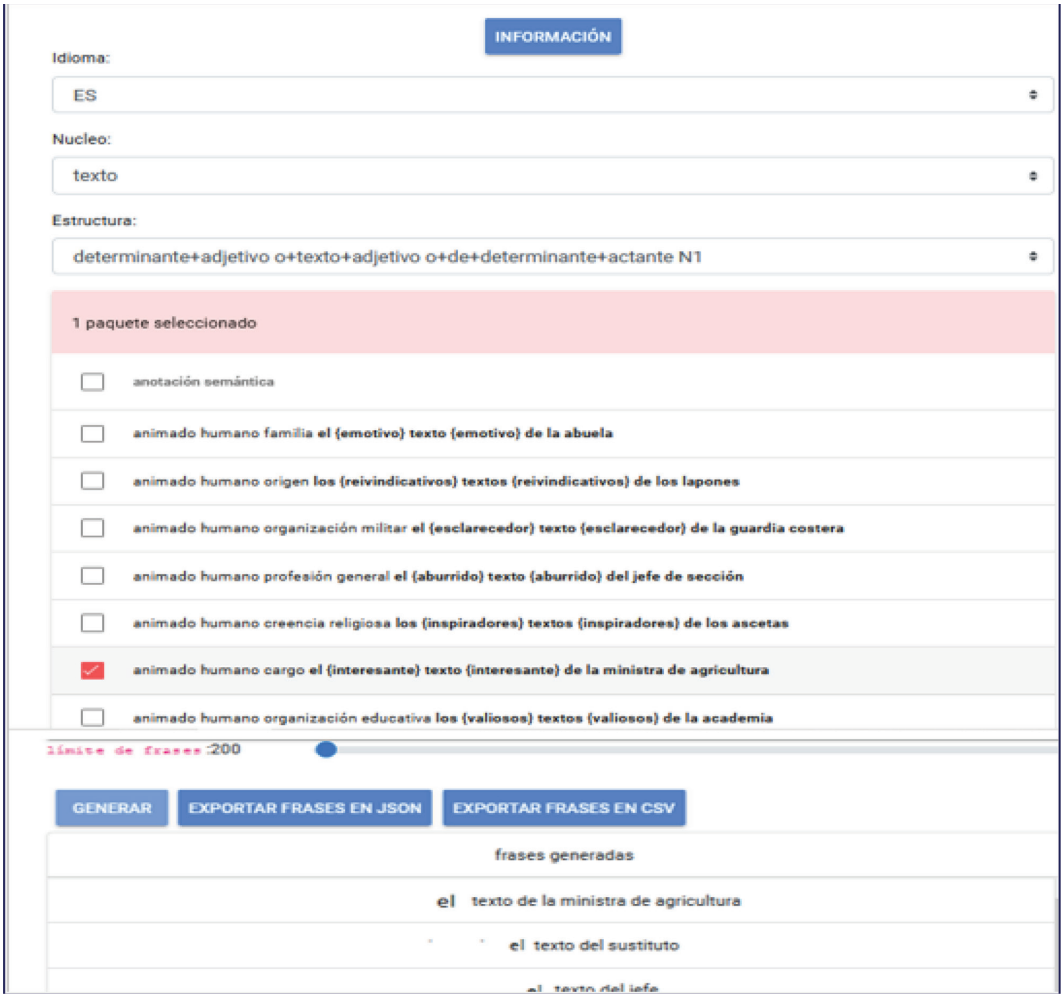


FIGURA 7: Xera*: ejemplo de* texto
Fuente: http://portlex.usc.gal/combinatoria/usuario.

4.2.2. Combinatoria

La aplicación *Combinatoria* es un generador automático biargumental de la FN en los tres idiomas, disponible online. En la interfaz con el usuario se pueden hacer consultas sobre los sustantivos y sus combinaciones en tres idiomas. Inicialmente, se elige la categoría semántica que completará el argumento 1.

Existen tres niveles de selección para cada uno. En el primero, se selecciona una las ocho categorías de esta ontología: *animado, material, intelectual, lugar, estado, situación, unidad* y *proceso*. Esta elección implica una herencia de significado limitando el número de categorías semánticas del nivel 2. A su vez, la elección de una de ellas restringe más las posibilidades semánticas del nivel 3. En consecuencia, las características sintáctico-semánticas del complemento1 limitan las posibles combinaciones semánticas del argumento, complemento 2. En el nivel 1, el complemento 2 solo muestra las clases semánticas que pueden combinarse con el complemento 1, siguiendo así el mismo proceso restrictivo en los dos niveles siguientes para la selección de la categoría semántica más específica del complemento 2. La Figura 8 muestra un ejemplo de la combinatoria de texto y las frases generadas.

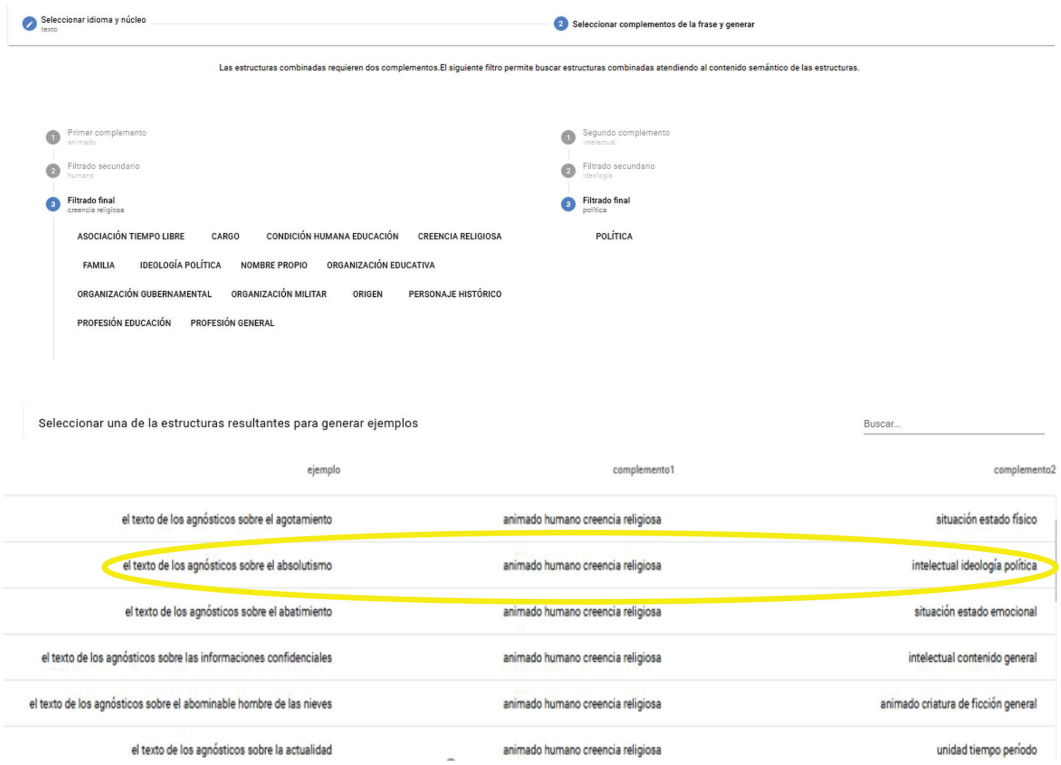


Figura 8: *Ejemplo de combinatoria biargumental para texto en* Combinatoria
Fuente: http://portlex.usc.gal/combinatoria/.

La mayor granularidad de la ontología permite programar con mayor nivel de restricción y generar frases nominales más adecuadas para su aplicación en la enseñanza de lenguas y la traducción. En la actualidad se está trabajando para generar el contexto oracional en un fututo.

5. CONCLUSIONES

En este trabajo, hemos demostrado como la ontología es un recurso léxico-conceptual que facilita la representación del mundo mediante categorías conceptuales relacionadas entre sí. Se ha concebido como un esquema conceptual, una estructura jerárquica cercana a una taxonomía, que organiza las propiedades semánticas del léxico bien a través de relaciones verticales de inclusión y horizontales de identidad. Las relaciones jerárquicas, en general son la base de la ontología, la herencia es una propiedad destacada en la configuración de las categorías semánticas que la conforman. En los proyectos *MultiGenera* y *MultiComb* su utilidad es indiscutible para el almacenamiento y recuperación del léxico de la base de datos y para el desarrollo del software, los niveles categoriales y las clases propuestas en la ontología funcionan como un esquema conceptual básico para la formalización de las restricciones semánticas en la programación de la generación automática de frases nominales con un argumento en *Xera* y con dos en *Combinatoria*.

La motivación para diseñar una ontología propia para los proyectos era dar cuenta de las restricciones semánticas en la combinatoria de los sustantivos en la FN tal como se planteaba en la gramática y lexicografía valencial (Domínguez Vázquez, 2011; Domínguez Vázquez, Valcárcel Riveiro & Lindemann, 2018). Cuando se comenzó la modelización del *MultiGenera*, se pensó que ninguna ontología de libre acceso se adecuaba a la finalidad del proyecto, por lo que se decidió profundizar en las categorías semánticas empleadas en el *Portlex* constituyendo la ontología 0.1. No obstante, al avanzar la investigación, esta ontología se enlazó automáticamente con WordNet y otras

ontologías, para aumentar los datos léxicos. Entonces se vio la necesidad de ampliar las categorías semánticas, sus relaciones y reajustar el etiquetado de las mismas para facilitar la conexión, por esta razón, la ontología evoluciona a la versión 0.2, tomando como referencia WordNet. La ontología 0.2 sigue en periodo de prueba, se espera que pueda ser portable, que permita la reutilización de datos y en un futuro la interoperatividad entre las distintas lenguas de los proyectos[16].

En definitiva, la ontología es una herramienta lexicográfica para la enseñanza de lenguas, con la didactización de las aplicaciones *Xera* y *Combinatoria* que, desde un enfoque onomasiológico, cognitivo organiza el léxico disponible para el aprendiz de lenguas, las categorías semánticas de la ontología que se muestran al usuario funcionan como principio selectivo en las relaciones sintagmáticas de la FN y permiten explorar el sentido dinámico del significado (Croft & Cruse, 2004).

Para concluir, la ontología tiene un valor en sí misma y como herramienta lexicográfica tanto para la lectura humana y como para la lectura computacional. Hay que destacar que de momento sólo se ha probado en estos proyectos. En un futuro, la optimización de esta ontología nos permitirá la interoperabilidad semántica con diferentes aplicaciones y funcionará como índice interlingüístico entre las tres lenguas para su traducción y el análisis contrastivo. La ontología y las demás aplicaciones mencionadas en este trabajo están disponible online en la web de los proyectos[17].

REFERENCIAS BIBLIOGRÁFICAS

Bentivogli, L., Forner, P., Magnini, B. & Pianta, E. (2004). Revising the Wordnet Domains Hierarchy: Semantics, Coverage and Balancing. En G. Sérasset, S. Armstrong, C. Boitet, A. Popescu-Belis & D. Tufis (eds.), *Proceedings of the Workshop on Multilingual Linguistic Ressources MLR2004.* Universidad de Ginebra. https://doi.org/10.3115/1706238.1706254

[16] Esta ontología está disponible en la web http://portlex.usc.gal/ontologia/
[17] La ontología y todas las herramientas están disponibles en http://portlex.usc.gal/

Borst, W. N. (1997). *Construction of Engineering Ontologies for Knowledge Sharing and Reuse* [Tesis doctoral. Universidad de Twente]. Centre for Telematics and Information Technology (CTIT). https://research.utwente.nl/en/publications/construction-of-engineering-ontologies-for-knowledge-sharing-and-

Bosque-Gil, J. (2019). *Linguistic Linked Data for Lexicography* [Tesis doctoral. Universidad Politécnica de Madrid]. Archivo Digital UPM. https://doi.org/10.20868/UPM.thesis.57887.

Bosque-Gil, J. & García, J. (eds.) (2019). *The Ontolex Lemon Lexicography Module Specificatio*n. Ontology Lexica under the W3C Community Final Specification Agreement (FSA). https://www.w3.org/2019/09/lexicog/

Brown, C. (2002). Paradigmatic relations of inclusion and identity I: Meronymy. En A. Cruse, F. Hundsnurscher, J. Michael & P. R. Lutzeier (eds.), *Lexikologie: Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen 1. Halbband. Lexicology: An International Handbook on the Nature and Structure of Words and Vocabularies,* vol. I. (pp. 480-485). de Gruyter.

Codina, L. & Pedraza-Jiménez, R. (2011). Tesauros y ontologías en sistemas de información documental. *El profesional de la información*, *20*(5), 555-563. https://doi.org/10.3145/epi.2011.sep.10

Combinatoria = Domínguez Vázquez, M.ª J. (dir.), Valcárcel Riveiro, C., Bardanca Outeiriño, D., Calañas Continente, J. A., Catalá Torres, N., López Iglesias, N., Martín Gascueña, R., Mirazo Balsa, M., Sanmarco Bande, M. T. & Pino Serrano, L. (2020). *Combinatoria. Prototipo online para la generación biargumental de la frase nominal en alemán, español y francés*. Universidade de Santiago de Compostela. Consultado el 30 de mayo de 2023. http://portlex.usc.gal/combinatoria/usuario

Croft, W. & Cruse, D. A. (2004). *Cognitive Linguistic*. Cambridge University Press. https://doi.org/10.1017/CBO9780511803864

Cruse, A. (1986). *Lexical Semantics*. Cambridge University Press.

Cruse, A. (2002). Descriptive models for sense relations II: Cognitive semantics. En A. Cruse, F. Hundsnurscher, J. Michael & P. R. Lutzeier (eds.), *Lexikologie: Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen 1. Halbband. Lexicology: An International Handbook on the Nature and Structure of Words and Vocabularies,* vol. I. (pp. 542-549). de Gruyter. https://doi.org/10.1515/9783110113082.1.15.542

Cruse, D. A. (2004). *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford University Press.

Domínguez Vázquez, M.ª J. (2011). *Kontrastive Grammatik und Lexikographie: spanisch-deutsches Wörterbuch zur Valenz des Nomens*. IUDICIUM Verlag GmbH München.

Domínguez Vázquez, M.ª J., Solla Portela, M. A. & Valcárcel Riveiro, C. (2019). Resource Interoperability: Exploiting Lexicographic Data to Automatically Generate Dictionary Examples. En I. Kosem, T. Z. Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.), *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019* (pp. 51-71). Lexical Computing.

Domínguez Vázquez, M.ª J., Valcárcel Riveiro, C. & Lindemann, D. (2018). Multilingual Generation of Noun Valency Patterns for Extracting Syntactic-Semantical Knowledge from Corpora (*MultiGenera*). En J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.), *XVIII EURALEX International Congress. Lexicography in Global Contexts* (pp. 847-854). Ljubljana University Press.

Gangemi, A., Navigli, R. & Velardi, P. (2003). The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet. En R. Meersman, Z. Tari & C. Schmidt (eds.), *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE* (pp. 820-838). Springer. https://doi.org/10.1007/978-3-540-39964-3_52

Gómez Guinovart, X. & Solla Portela, M. A. (2018). Building the Galician wordnet: methods and applications. *Language Resources and Evaluation*, *52*(1), 317-339. https://doi.org/10.1007/s10579-017-9408-5

Gruber, T. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, *5*(2), 199-220. https://doi.org/10.1006/knac.1993.1008

Guarino, N., Oberle, D. & Staab S. (2009). What Is an Ontology? En S. Steffen & R. Studer (eds.), *Handbook on Ontologies. International Handbooks on Information Systems* (pp. 1-17). Springer. https://doi.org/10.1007/978-3-540-92673-3_0

Lakoff, G. (1987). *Women, Fire and Dangerous Things: what Categories Reveal about the Mind*. The University of Chicago Press. https://doi.org/10.7208/chicago/9780226471013.001.0001

Lenat, D. & Guha, R. V. (1990). CYC: A Mid-Term Report. *AI Magazine, 11*(3), 32-59.

Lyons, J. (1977). *Semántica*. Editorial Teide. https://doi.org/10.1017/CBO9781139165693

Martin-Gascueña, R. (2013). La hiponimia en un área conceptual. *Revista Pragmalingüística*, *21*, 86–106 http://revistas.uca.es/index.php/pragma/issue/view/139/showToc

Martin, P. (2003). Correction and Extension of WordNet 1.7 for Knowledge-based Applications. En B. Ganter, A. Moor & W. Lex (eds.), *Conceptual Structures for Knowledge Creation and Communication* (pp. 160-173). Springer. https://doi.org/10.1007/978-3-540-45091-7_11

Masolo, C., Borgo, S., Gangemi, A., Guarino, N. & Oltramari, A. (2003). *DOLCE: Descriptive Ontology for Linguistic and Cognitive Engineering* https://www.istc.cnr.it/it/content/dolce-descriptive-ontology-linguistic-and-cognitive-engineering

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, J. (1990). Introduction to WordNet: An on-line Lexical Database. *International Journal of Lexicography, 3*(4), 235-244. https://doi.org/10.1093/ijl/3.4.235

Ontología léxica = Domínguez Vázquez, M.ª J., Valcárcel Riveiro, C. & Bardanca Outeiriño, D. (2021). *Ontología léxica*. Universidade de Santiago de Compostela. Consultado el 30 de mayo de 2023 http://portlex.usc.gal/ontologia/

Pease, A., Niles, I. & Li, J. (2002). The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Application. *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web. July-August*. Edmonton.

Pianta, E., Bentivogli, L. & Girardi, C. (2002). MultiWordNet Developing an aligned multilingual database. *Proceedings of the 1st International WordNet Conference* (pp. 293-302).

Portlex = Domínguez Vázquez, M.ª J. (dir.), Valcárcel Riveiro, C., Mirazo Balsa, M., Sanmarco Bande, M.ª T., Simões, A. & Vale, M. J. (2018). *Portlex. Diccionario multilingüe de la valencia del nombre*. Universidade de Santiago de Compostela. Consultado el 30 de mayo de 2023. http://portlex.usc.gal/portlex/

Pustejovsky, J. (1995) *The Generative Lexicon*. The MIT Press Cambridge.

Rodríguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Alonge, A., Bertagna, F. & Roventini, A. (1998). The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. P. Vossen (ed.) *EuroWordNet: A multilingual database with lexical semantic networks* (pp. 45-80). Springer. https://doi.org/10.1007/978-94-017-1491-4_3

Rosch, E. (1978). Principles of Categorization. En B. Lloyd & E. Rosch (eds.), *Cognition and categorization* (pp. 27-48). Lawrence Erlbaum.

Solla Portela, M. A. & Gómez Guinovart, X. (2015). Galnet o WordNet do galego. Aplicacións lexicolóxicas e terminolóxicas. *Revista galega de filoloxía, 16*, 169-201. https://doi.org/10.17979/rgf.2015.16.0.1383

Studer, R., Benhjamins, V. R. & Fensel, D. (1998). Knowledge Engineering: Principles and methods. *Data & Knowledge Engineering*, *25*(1-2), 161-197. https://doi.org/10.1016/S0169-023X(97)00056-6

Vossen, P. J. T. M. (1998). EuroWordnet: Building a Multilingual Database with Word Nets for European Languages. *ELRA Newsletter*, *3*(1), s.p. https://doi.org/10.1007/978-94-017-1491-4_1

WordNet = The Trustees of Princeton University (2023). *WordNet*. Princeton University. Consultado el 30 de mayo de 2023. https://wordnet.princeton.edu/

Xera = Domínguez Vázquez, M.ª J. (dir.), Valcárcel Riveiro, C., Bardanca Outeiriño, D., Calañas Continente, J. A., Catalá Torres, N., López Iglesias, N., Martín Gascueña, R., Mirazo Balsa, M., Sanmarco Bande, M.T. & Pino Serrano, L. (2020). *Xera. Prototipo online para la generación automática monoargumental de la frase nominal en alemán, español y francés*. Universidade de Santiago de Compostela. Consultado el 30 de mayo de 2023. http://portlex.usc.gal/combinatoria/usuario

HERRAMIENTAS Y DIFICULTADES EN EL ANÁLISIS DEL GRUPO NOMINAL EN FRANCÉS PARA SU PROCESAMIENTO COMPUTACIONAL

TOOLS AND DIFFICULTIES IN THE ANALYSIS OF THE FRENCH NOUN PHRASE FOR COMPUTATIONAL PROCESSING

Carlos Valcárcel Riveiro
Universidade de Vigo
carlos.valcarcel@uvigo.gal

Laura Pino Serrano
Universidade de Santiago de Compostela
laura.pino@usc.es

Resumen

La presente contribución presenta el trabajo desarrollado sobre la lengua francesa en el portal lexicográfico multilingüe PORTLEX. Concretamente, el artículo se centra en las herramientas utilizadas por los equipos de francés de tres proyectos de investigación: el diccionario PORTLEX y el desarrollo de prototipos para la generación automática de sintagmas nominales en los proyectos *MultiGenera* y *MultiComb*. Se detallan los recursos utilizados distinguiendo las herramientas de consulta (diccionarios, gramáticas) de las computacionales (corpus, bases de datos). Asimismo, se describen las principales dificultades asociadas al uso de cada herramienta o recurso y se explican las decisiones adoptadas para solventarlas o minimizarlas.

Palabras clave: francés, herramientas, recursos, lexicografía, PLN.

Abstract

This paper presents the work carried out on the French language within the multilingual lexicographic portal PORTLEX. In particular, the article focuses on the tools used by the teams working on French in three research projects: the PORTLEX dictionary and the development of prototypes for the automatic generation of noun phrases within *MultiGenera* and *MultiComb* projects. The resources used are described in detail, distinguishing between reference tools (dictionaries, grammars) and computational tools (corpora, databases). In addition, the main difficulties associated with the use of each tool or resource are described and the choices made to resolve or minimise them are explained.

Keywords: French, tools, resources, lexicography, NLP.

DOI: https://doi.org/10.17561/rilex.6.3.8392

1. INTRODUCCIÓN

Desde sus inicios, hace ya diez años, el francés ha sido una de las lenguas de trabajo en el portal lexicográfico multilingüe PORTLEX[1]. Este portal se creó en la Universidad de Santiago bajo la dirección de Domínguez Vázquez durante el desarrollo del diccionario del mismo nombre. Más allá de un espacio web centrado en la lexicografía, PORTLEX constituye un lugar de encuentro y colaboración de especialistas en los ámbitos de la lingüística y de la computación para desarrollar proyectos de investigación, herramientas lingüísticas y diferentes acciones de divulgación científica (congresos, cursos, talleres, etc.). Por lo tanto, la inclusión del francés en las actividades de este portal, y toda la investigación paralela desarrollada en diferentes áreas de la lingüística, ha implicado la colaboración de diferentes especialistas en esta lengua. Así, a lo largo de estos años grupos de francesistas, en colaboración con otros equipos especializados en las otras lenguas de trabajo, han realizado diferentes tareas de tipo teórico y aplicado: el análisis y anotación sintáctico-semántica de frases nominales, la identificación y anotación semántica de prototipos léxicos, la elaboración de paquetes léxicos, la revisión de textos generados automáticamente o incluso el diseño de bases de datos lingüísticos, entre otras.

En la presente contribución trataremos, pues, de la labor desarrollada por estos equipos de francés en PORTLEX centrándonos en las herramientas utilizadas y las dificultades encontradas a la hora de afrontar todas estas tareas. Como era de esperar, estas últimas fueron numerosas y de diverso tipo. Si bien una parte de estos desafíos eran comunes a las otras lenguas meta de los proyectos, varios se relacionaban directamente con particularidades de la lengua francesa. En primer lugar, describiremos las tareas realizadas y los resultados obtenidos en los tres proyectos desarrollados en el portal hasta el

[1] http://portlex.usc.gal/

momento: el diccionario PORTLEX (FFI2012-32456)[2] (Domínguez Vázquez & Valcárcel Riveiro, 2020) y los prototipos de generación automática de frases nominales de los proyectos *MultiGenera*[3] y *MultiComb* (FFI2017-82454-P)[4] (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019; Domínguez Vázquez, Bardanca Outeiriño & Simões, 2021). Haremos también referencia al etiquetador semántico desarrollado en el proyecto en curso ESMAS-ES+ (PID2022-137170OB-I00)[5]. Pasaremos después a hablar de las principales herramientas de trabajo del equipo de francés de PORTLEX, haciendo una distinción entre herramientas de consulta (diccionarios y gramáticas) y herramientas de procesamiento del lenguaje natural (corpus, bases de datos y bibliotecas). Describiremos sus funcionalidades y limitaciones más destacadas, así como las dificultades que estas últimas han supuesto para el trabajo con la lengua francesa. En las conclusiones extraeremos las principales lecciones aprendidas y mencionaremos algunas líneas de trabajo futuras en el marco del portal PORTLEX.

2. LOS PROYECTOS PORTLEX Y SUS APORTACIONES AL FRANCÉS

Los cuatro proyectos desarrollados en el marco del portal lexicográfico PORTLEX tienen en común tres aspectos esenciales: su carácter multilingüe, su orientación dependencial y su interés por el grupo nominal. En primer

[2] El proyecto *Portal Lexicográfico: Diccionario online modular multilingüe y corpus informatizado anotado de la frase nominal* fue financiado por el Ministerio de Economía y Competitividad, por la Unión Europea a través del Fondo Europeo de Desarrollo Regional (FEDER) 2007-2013 y por la Red de Lexicografía RELEX (R2014/042), 2013-2015.

[3] Este proyecto fue financiado por la Fundación BBVA a través del programa de Ayudas Fundación BBVA a Equipos de Investigación Científica - Humanidades Digitales, 2017-2020.

[4] Este proyecto fue financiado por la Agencia Estatal de Investigación (Ministerio de Ciencia e Innovación) en el marco del Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia, Generación de Conocimiento (EXCELENCIA 2017, 2017-PN091) y por la Unión Europea a través del Fondo Europeo de Desarrollo Regional (FEDER) "Una manera de hacer Europa", 2018-2021.

[5] Este proyecto está financiado por la Agencia Estatal de Investigación (Ministerio de Ciencia e Innovación) y por la Unión Europea a través del Fondo Europeo de Desarrollo Regional (FEDER) "Una manera de hacer Europa".

lugar, la inclusión de múltiples idiomas en el análisis lingüístico permite una comprensión más amplia y comparativa de las estructuras lingüísticas. Así, al abordar diferentes idiomas, los proyectos PORTLEX pueden identificar similitudes y diferencias en la construcción de los grupos nominales y sus combinatorias en diversas lenguas. De hecho, el desarrollo de interfaces que simplifiquen la consulta contrastiva de la información de varias lenguas ha sido una preocupación de los equipos de PORTLEX desde sus inicios. Así, en los recursos disponibles en este portal lexicográfico, los datos referentes al francés pueden ser contrastados de diferentes formas con los de las otras lenguas de trabajo, principalmente el castellano y el alemán.

En segundo lugar, el enfoque valencial o dependencial aplicado al estudio del grupo nominal tampoco deja de constituir un aspecto novedoso en los estudios franceses. A pesar de que la gramática dependencial moderna fue formulada inicialmente por un lingüista francés, Tesnière (1959), los análisis de corte valencial todavía son menos frecuentes para el francés que para otras lenguas como el alemán. Menos aún lo son las investigaciones sobre los predicados nominales, ya que el análisis del predicado verbal y de sus argumentos ha ocupado siempre un lugar prioritario en el campo de la gramática de dependencias. Sin embargo, este verbocentrismo no ha frenado el desarrollo de investigaciones sobre los argumentos del sustantivo, su combinatoria y las restricciones formales y semánticas asociadas. Para el francés destacan en esta línea los estudios desarrollados en el marco de la Teoría Sentido-Texto (Mel'čuk, 1997) y del *Laboratoire d'Informatique Documentaire et Linguistique* (LIDL) (Gross, 2012). Como veremos más adelante, estos y otros estudios resultan, pues, esenciales para el trabajo desarrollado en el portal PORTLEX sobre la lengua francesa.

2.1. *El diccionario PORTLEX*

La elaboración de este diccionario electrónico constituyó el primer proyecto del portal y se desarrolló entre los años 2013 y 2017. Esencialmente, PORTLEX es un diccionario multilingüe y valencial sobre la estructura del

grupo nominal en cinco idiomas, entre ellos el francés. Por lo tanto, las entradas están constituidas por lexemas nominales y en ellas se analizan sus diferentes argumentos, así como la combinatoria que se da entre ellos. Probablemente, los aspectos más novedosos de este diccionario son la estructura de la base de datos que lo sustenta y la interfaz de consulta contrastiva, que permite visualizar en pantalla entradas equivalentes en dos lenguas (Domínguez Vázquez & Valcárcel Riveiro, 2020).

Para la lengua francesa se elaboraron un total de 37 entradas, lo que supuso la descripción de 197 realizaciones actanciales[6] y 152 combinatorias. En ambos casos se proporciona un análisis pormenorizado en el que se detalla la función sintáctica de las realizaciones, su rol semántico, los rasgos semánticos asociados y una descripción de su estructura formal. Además, tanto las realizaciones como las combinatorias se ilustran con un mínimo de tres ejemplos buscados y seleccionados manualmente en corpus. Para el francés se extrajeron un total 1047 ejemplos y, sin duda, esta fue la parte más laboriosa del proyecto. Esto se debió principalmente a las limitaciones del corpus utilizado que se comentarán más adelante. Prácticamente todo el trabajo sobre el francés en el diccionario PORTLEX corrió a cargo de Valcárcel Riveiro, de la Universidad de Vigo[7].

2.2. *MultiGenera y MultiComb*

Como ya se ha comentado, una de las mayores dificultades experimentadas por los equipos de trabajo durante la elaboración del diccionario PORTLEX fue la búsqueda y selección de ejemplos válidos. Para intentar solventar este problema se desarrollaron entre 2018 y 2021 dos proyectos paralelos de

[6] En lo referente a las realizaciones actanciales analizadas para el francés, cabe destacar como novedad que el diccionario incluya las estructuras apositivas N_1N_2 (p. ej., *la consommation poisson*, *une question santé*) (Valcárcel Riveiro, 2017).

[7] Como parte de su programa de formación, varios estudiantes del Máster europeo de lexicografía EMLex editaron y revisaron algunas entradas. Se trató en concreto de Océane Meyan, Nikolay Chepurnykh y Polina Mikhel.

investigación aplicada que tenían como objetivo el desarrollo de dos prototipos para generar, en una primera fase, sintagmas nominales (*Combinatoria*)[8] y, en una segunda, contextos oracionales para estos (*CombiContext*)[9]. Aun tratándose de prototipos, nos encontramos ante herramientas complejas, ya no solo por su carácter multilingüe, sino sobre todo porque son los propios usuarios quienes establecen los parámetros de los sintagmas y oraciones generados en función de sus necesidades. Las estructuras creadas por estos prototipos pueden servir para diversos fines, pero principalmente para nutrir con ejemplos personalizados herramientas lexicográficas (diccionarios, glosarios) y actividades de aprendizaje de lenguas (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019; Domínguez Vázquez, Bardanca Outeiriño & Simões, 2021).

Los prototipos desarrollados en *MultiGenera* y *MultiComb* cubren tres lenguas (alemán, castellano y francés) y veinte lexemas o núcleos nominales. Además de las complicaciones computacionales, la construcción de estos prototipos requirió un ingente trabajo a nivel lingüístico por parte de los equipos investigadores involucrados: identificación de prototipos léxicos para cada realización argumental, etiquetado semántico de estos prototipos[10], creación y depuración de paquetes semánticos para generar argumentos nominales y validación tanto de combinatorias de realizaciones como de estructuras oracionales (Domínguez Vázquez, Valcárcel Riveiro & Lindemann, 2018; Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019; Domínguez Vázquez, Bardanca Outeiriño & Simões, 2021). Asumir toda esta carga de trabajo requirió conformar un equipo más consolidado para la lengua francesa. Además de Valcárcel Riveiro, que asumió también tareas computacionales, trabajaron en estos proyectos Pino Serrano (Universidade de Santiago de Compostela),

[8] http://portlex.usc.gal/combinatoria/usuario
[9] http://portlex.usc.gal/combinatoria/verbal
[10] Para tal fin se elaboró una ontología léxica propia (Domínguez Vázquez, Valcárcel Riveiro & Bardanca Outeirño, 2021).

centrada en las tareas de análisis gramatical y validación de estructuras, y Malingret (Universidade de Santiago de Compostela), encargada de la revisión y evaluación de las estructuras generadas por los prototipos.

En total, trabajamos con 20 lexemas nominales[11] de diferentes campos semánticos (*absence*, *amour*, *augmentation*, *conversation*, *couleur*, *déménagement*, *discussion*, *douleur*, *fuite*, *largeur*, *mort*, *odeur*, *présence*, *question*, *réponse*, *saveur*, *séjour*, *texte*, *vidéo*, *voyage*), para los que se procesaron 152 realizaciones. Por parte del equipo esto implicó la elaboración y revisión de 624 paquetes semánticos para generar automáticamente estas realizaciones, la validación de 1491 estructuras oracionales y la revisión sistemática de los sintagmas y oraciones generados. Si bien los prototipos desarrollados ofrecen una cobertura muy limitada, la gran cantidad de datos analizados sin duda ha permitido un mejor conocimiento sobre el funcionamiento del grupo nominal en francés. Además, se han desarrollado herramientas específicas (API, Combina, Lematiza) para facilitar la extracción y procesamiento de datos léxicos en francés y otros idiomas (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019). Finalmente, todo esto nos ha permitido implementar una metodología para analizar en detalle las características semánticas de los argumentos nominales (Valcárcel Riveiro & Pino Serrano, 2023).

2.3. *ESMAS-ES+*

Actualmente, dentro del portal PORTLEX se está desarrollando ya un nuevo proyecto en el que también trabaja un equipo de francesistas: ESMAS-ES+. Además de Valcárcel Riveiro, componen el equipo las profesoras Vázquez Rodríguez y Castro Lorenzo, ambas de la Universidad de Vigo. Con este nuevo proyecto se pretende construir un prototipo de etiquetador semántico multilingüe y sostenible, es decir, que aproveche los recursos elaborados

[11] Es decir, solo se trabajó con una determinada acepción de estos vocablos, no con todas. Por ejemplo, para *question* los prototipos cubren la acepción 'pregunta dirigida a alguien' y no otras como 'tema' o 'problema'.

en el marco de PORTLEX en proyectos anteriores. Principalmente, se parte de la ontología y los paquetes semánticos desarrollados para *MultiGenera* y *MultiComb* para alimentar el etiquetador semántico. Esto conlleva la revisión y traducción[12] de los paquetes de datos existentes, lo que por ahora limita la cobertura del etiquetador a los sustantivos. Además, se prevé el desarrollo de una base de datos y una interfaz específicas.

3. LAS HERRAMIENTAS DE TRABAJO: VENTAJAS, LIMITACIONES Y DIFICULTADES

Tanto en los proyectos *MultiGenera* y *MultiComb*, como en la elaboración del diccionario PORTLEX, el equipo de francés tuvo que emplear diversas herramientas a lo largo de las distintas fases de trabajo. Por un lado, el análisis semántico y sintáctico de las frases nominales, esencial en los tres proyectos realizados, demandaba el uso de herramientas de consulta, como diccionarios y gramáticas. Por otro lado, la extracción y procesamiento computacional de la gran cantidad de datos manejados exigía la búsqueda y manejo de corpus, repositorios y bases de datos disponibles para la lengua francesa, así como el desarrollo de herramientas específicas. En este apartado, describiremos las principales herramientas utilizadas para ambos fines, deteniéndonos en las dificultades encontradas por nuestro equipo.

3.1. *Herramientas de consulta*

En todos los proyectos desarrollados en el marco de PORTLEX el equipo de lingüistas se encontró ante el desafío de analizar numerosas estructuras argumentales de sustantivos en las diferentes lenguas de trabajo. En el caso del francés, los predicados nominales estudiados a diferentes niveles se acercan ya a la cincuentena. Para llevar a cabo este análisis previo se tornó esencial contar con obras de consulta fiables y orientadas al tipo de marco

[12] En los proyectos anteriores no se elaboraron los mismos paquetes semánticos para todas las lenguas. Para traducir los paquetes de una lengua a otra se desarrolló otra herramienta, TraduWord, en colaboración con el Instituto da Lingua Galega (Domínguez Vázquez, Bardanca Outeiriño & Simões, 2021, p. 278).

conceptual en el que se desarrollaban los proyectos: la gramática valencial o de dependencias. Estas fuentes no solo brindaron respuestas a ciertos interrogantes que fueron surgiendo durante el desarrollo de los diferentes proyectos, sino que permitieron sobre todo visualizar modelos para analizar esquemas argumentales de sustantivos, proporcionando así una base sólida para la comprensión y generación automática de las estructuras lingüísticas estudiadas.

3.1.1. Diccionarios y léxicos

Todo procesamiento automático del lenguaje implica ineludiblemente trabajar con el léxico y, por lo tanto, la disponibilidad de recursos lexicográficos es indispensable en la mayoría de los proyectos. Las necesidades lexicográficas de cada proyecto varían considerablemente en función de su orientación monolingüe o multilingüe, pudiendo incluir diccionarios generales y especializados, así como léxicos técnicos y bases terminológicas. En el caso de los proyectos desarrollados en el marco de PORTLEX se requerían recursos lexicográficos con amplia información sobre predicados verbales y, sobre todo, nominales. Desgraciadamente, la disponibilidad de diccionarios o léxicos de orientación valencial es escasa y a menudo los recursos existentes no tienen una gran cobertura. El francés no constituye una excepción, pero cuenta con un diccionario electrónico de amplia cobertura y de libre acceso que responde a las necesidades de los proyectos desarrollados: el *Trésor de la langue française informatisé* o *TLFi* (Dendien, 2020).

Si bien el *TLFi* no podría definirse como un diccionario valencial en sentido estricto (Domínguez Vázquez, 2018), su amplia cobertura y sobre todo la información específica que ofrecen sus entradas sobre cuestiones gramaticales lo convierten en una obra de referencia indispensable. Sin duda, se trata de una obra monumental que contiene más de 100.000 entradas y 270.000 definiciones. Incluye numerosos elementos metatextuales como códigos gramaticales, etiquetas semánticas y estilísticas, así como indicadores de dominio. Además de esto, se pueden encontrar 430.000 ejemplos extraídos de

obras literarias. Finalmente, *TLFi* también ofrece información detallada de carácter etimológico y bibliográfico (Pierrel, 2003, pp. 159-161).

Sin embargo, a pesar de toda la información compilada en el *TFLi*, este diccionario no explicita en las entradas de muchos sustantivos los complementos esenciales o argumentos. En muchos casos, y no de manera sistemática, los argumentos vienen expresados en la propia definición de los lemas. Por ejemplo, en el caso de la primera acepción de *séjour*, con la que trabajamos en MultiComb, el *TFLi* la define como "Fait de demeurer un certain temps dans un lieu, un endroit". Esta definición evoca el argumento 'agente' (*le séjour de Marie*) y el argumento locativo (*le séjour en Italie*) como parte del significado, es decir, del predicado de *séjour* (*le séjour de Marie en Italie*). En otros casos, los argumentos se expresan mediante ejemplos en la sección SYNT. (*syntagme, syntaxe*) mezclados con complementos no esenciales. Así, en el caso de la primera acepción de *odeur*[13] argumentos como *l'odeur du bois* o *odeur d'herbe* aparecen en la misma sección SYNT. que *odeur forte* o *odeur délicieuse*. Esta falta de coherencia en la presentación de la información sintáctica de los sustantivos hace a veces difícil la consulta de este recurso tan importante.

Otro recurso lexicográfico esencial para el equipo de francés en los proyectos PORTLEX es el *Dictionnaire explicatif et combinatoire du français contemporain* (*DECFC*) (Mel′čuk, Arbatchewsky-Jumarie, Iordanskaja, Mantha & Polguère 1984-1999). Los cuatro volúmenes publicados de esta obra tienen como objetivo proporcionar una descripción completa y sistemática del léxico francés siguiendo los principios de la lexicología explicativa y combinatoria (Mel′čuk, Clas & Polguère, 1995). Desarrollado en el marco de la Teoría del Sentido-Texto (TST), el DECFC pretende presentar, de forma coherente y sistemática, toda la información necesaria para que los hablantes

[13] La definición ofrecida por el TLFi es la siguiente: "Émanation propre à un corps pouvant être perçue par l'homme ou par un être animé grâce à des organes particuliers et avec des impressions diverses (agréable, désagréable, indifférente)".

expresen de forma lingüísticamente correcta cualquier significado que deseen comunicar. Para ello, incluye detalles exhaustivos para cada unidad léxica, como pronunciación, categoría gramatical y patrones sintácticos, así como las propiedades y restricciones de las coocurrencias léxicas y la combinatoria argumental (Mel′čuk & Milićević, 2014, pp. 269-290). La marcada orientación valencial de este recurso, que incluye el análisis minucioso de los esquemas argumentales de numerosos sustantivos, hacen del *DECFC* una obra de consulta obligada para el equipo de francés del portal PORTLEX. Sin embargo, la cobertura de este diccionario es muy limitada ya que sólo describe 510 vocablos en francés y, salvo *conseil, apprentissage*, *maladie* y *risque*, analizados parcialmente en el diccionario PORTLEX[14], el resto de sustantivos cubiertos por los diferentes proyectos no se encuentran en el *DECFC*. De todas formas, este recurso constituye un referente para el análisis de predicados nominales y de las combinatorias que se dan en ellos.

Además del *DECFC*, dos recursos desarrollados en el marco de la TST resultaron de gran utilidad en momentos puntuales del trabajo. Por un lado, consultamos el *Lexique actif du français* (*LAF*) (Mel′čuk & Polguère, 2007), una versión accesible del *DECFC* y con una marcada orientación didáctica. De hecho, este recurso está dirigido tanto a profesorado y aprendientes del francés como a traductores y otros profesionales del lenguaje. De nuevo, la cobertura, aunque es más amplia que la del *DECFC*, sigue siendo insuficiente, ya que se analiza un número limitado de lexemas (781), agrupados en 386 vocablos. Por otro lado, y aunque se trata de un diccionario de lengua española centrado en las colocaciones, el *Diccionario de colocaciones del español* (*DiCE*) (Alonso-Ramos, 2004) se centra en el análisis de sustantivos, para los que se describen de manera sistemática los esquemas argumentales de sus

[14] Cabe señalar aquí que, mientras el *DECFC* describe vocablos y sus diferentes lexemas o acepciones, el diccionario PORTLEX se limita normalmente a presentar información sobre un lexema o acepción específicos.

predicados (Alonso-Ramos, 2017, pp. 185-192). Dada la proximidad tipológica del francés y el castellano, resultó de interés consultar el análisis que presenta de lexemas equivalentes a los analizados en los proyectos PORTLEX (p. ej., *texto*, *discusión*) o pertenecientes a su mismo campo semántico.

3.1.2. Gramáticas e investigaciones gramaticales

Dada la fuerte orientación gramatical de los proyectos desarrollados en el portal PORTLEX, la consulta de investigaciones en este campo resultó imprescindible. No sólo fue necesario verificar la gramaticalidad de numerosas construcciones, sino también clarificar cuestiones conceptuales más básicas como, por ejemplo, determinar el carácter valencial de las realizaciones adjetivales de ciertos argumentos[15]. Como ya se ha indicado, el enfoque adoptado para el análisis gramatical en los proyectos PORTLEX es el dependencial o valencial, más concretamente el elaborado por Domínguez Vázquez (2011) para los grupos nominales del castellano y el alemán a partir de la gramática de Engel (2004). Sin embargo, el trabajo con la lengua francesa suscita cuestiones particulares a esta lengua que requieren el manejo de obras específicas.

Naturalmente, se consultaron gramáticas de referencia para el francés, entre las que cabe destacar el clásico *Le bon usage* (Grevisse & Goosse, 2008). Resultó de especial interés la reciente *Grande grammaire du français* (Abeillé & Godard, 2021) por su tratamiento del grupo nominal más próximo al enfoque de los proyectos PORTLEX. Por otro lado, también se consultaron con frecuencia algunas gramáticas de autor como la *Grammaire méthodique du français* (Riegel, Pellat & Rioul, 2009) y, más concretamente, la *Grammaire critique du français de* Wilmet (1997) debido a su estudio exhaustivo

[15] Así, si en el grupo nominal *le séjour italien de tes amis* el carácter valencial del adjetivo *italien* (= *en Italie*) parece claro, ya no lo parece tanto en el syntagma *un texte italien sur les émotions*. Las investigaciones revelan que un mismo adjetivo puede ocupar una casilla valencial o circunstancial dependiendo del predicado, es decir, del contexto (Rigau, 1999).

de los sintagmas nominales, especialmente de la determinación. Finalmente, la marcada orientación semántica de la *Grammaire du sens et de l'expression* de Charaudeau (1992) resultó particularmente útil en el análisis de los roles semánticos de los argumentos nominales.

Más allá de todas estas obras de referencia, dos escuelas gramaticales de corte dependencial constituyeron los principales marcos de referencia para los equipos de francés en el portal PORTLEX: por un lado, la ya mencionada Teoría Sentido-Texto formulada por Mel′čuk y, por otro, la Teoría de las clases de objeto desarrollada por Gross en el *Laboratoire d'Informatique Documentaire et Linguistique* (LIDL). Ambos enfoques se han venido desarrollando, además, para impulsar avances en el ámbito del procesamiento del lenguaje natural desde hace décadas (Gross, 2004; Iordanskaja, Kim & Polguère, 1996). Resultaron de especial importancia los trabajos de Gross sobre el funcionamiento del grupo nominal en francés (Gross, 1991, 2012) y, más concretamente, sobre la noción de clase de objeto (Gross, 2002, 2008)[16]. Siguiendo esta línea, también se consultaron contribuciones de Blanco (1997, 1999), quien trabaja desde una perspectiva contrastiva francés-español. Finalmente, los trabajos de Lazard (1988, 1994), aunque centrados en los predicados verbales, se consultaron para evaluar el carácter actancial de algunos complementos nominales, algo para lo que también se tuvo en cuenta el estudio de Stage (1994).

Asimismo, se consultaron diferentes investigaciones más particulares sobre la estructura de los predicados, normalmente centradas en los predicados constituidos por sustantivos deverbales (Condette, Marín & Merlo, 2012; Stage, 1997). Finalmente, aunque se realizasen ya en el ámbito de la

[16] En concreto, este concepto de clase de objeto, entendido como un "ensemble de substantifs, sémantiquement homogènes, qui détermine une rupture d'interprétation d'un prédicat donné, en délimitant un emploi spécifique" (Gross, 2008, p. 11), se tuvo muy presente en la creación y anotación semántica de los paquetes léxicos en los proyectos *MultiGenera* y *MultiComb*.

lingüística española, resultaron de gran utilidad los capítulos sobre el sintagma nominal de Rigau (1999) y Picallo (1999) en la *Gramática descriptiva de la lengua española*, así como la tesis de doctorado de Barrios Rodríguez (2010).

3.2. Herramientas de procesamiento del lenguaje natural para la lengua francesa

Además de las herramientas diseñadas específicamente para los proyectos en cuestión, el equipo de francés utilizó diferentes herramientas específicas para esta lengua disponibles en línea. Afortunadamente, para el francés hay disponibles numerosos recursos y herramientas útiles que permitieron avanzar de manera eficiente en todas las fases de trabajo. Así, contamos con corpus para la extracción de datos y la definición de prototipos léxico-semánticos, bases de datos como WordNet para desarrollar los paquetes léxicos o librerías de datos morfológicos como la facilitada en *Freeling*, esenciales para generar la flexión verbal en los contextos oracionales producidos con *CombiContext*. En diferentes casos estos recursos se usaron para alimentar herramientas de elaboración propia (Lematiza, Combina, Flexiona, etc.), también compatibles con el francés. En este apartado nos centraremos, por lo tanto, en comentar los principales recursos utilizados y las limitaciones o dificultades que presentan.

3.2.1. El trabajo con corpus y sus dificultades

Los corpus constituyeron una herramienta fundamental, tanto para la elaboración de entradas en el diccionario PORTLEX como en los proyectos subsiguientes. Permitieron obtener a gran escala datos relevantes y representaron un recurso esencial para el análisis lingüístico. En el caso del diccionario se usó FRANTEXT[17] (Montémont, 2020; Pierrel, 2003) para identificar tanto argumentos nominales como sus combinaciones e ilustrar todo esto

[17] https://www.frantext.fr/

con ejemplos. FRANTEXT es una base de datos de textos franceses[18] que contiene una gran cantidad de textos literarios, científicos y técnicos desde el siglo XVIII hasta la actualidad. El corpus consta de más de 3.600 textos y 215 millones de palabras. Sin duda, es uno de los mayores corpus de textos en francés disponibles en línea y es utilizado por numerosos investigadores y estudiantes para realizar estudios lingüísticos y literarios. Este potente recurso permite localizar palabras específicas, lemas y expresiones regulares en una obra concreta o en un conjunto de fuentes.

Sin embargo, dado el carácter eminentemente literario de sus textos, en muchos casos resultaba difícil encontrar ejemplos de ciertas realizaciones argumentales. Concretamente, durante el análisis de la combinatoria argumental, FRANTEXT no permitía verificar, por falta de ejemplos, numerosas realizaciones existentes en el uso de la lengua. Por ejemplo, en el caso del lexema *consommation* ('uso de bienes o productos para la alimentación'), esta base de datos no proporcionaba ejemplos para un argumento tan habitual como el agentivo *par* + *determinante* + *nombre*, p. ej. *La consommation* [*de vin*] *par les Français*. Tampoco lo hacía para muchas combinaciones argumentales del sustantivo *apprentissage* ('acción de aprender un oficio o profesión') y ejemplos relativamente comunes como *l'apprentissage par l'enfant des structures syntaxiques* o *un apprentissage de deux ans comme consultant* tuvieron que obtenerse en corpora basados en la web.

Estas limitaciones llevaron al equipo de francés a contar exclusivamente con el corpus FrTenTen (Jakubíček, Kilgarriff, Kovář, Rychlý & Suchomel, 2013) y el interfaz de búsqueda Sketch Engine para los proyectos *MultiGenera* y *MultiComb*. Este corpus se basa en webs francesas y su versión de 2017, la utilizada en los proyectos que nos ocupan, contiene más de 10 mil millones de palabras. El uso de este potente recurso hizo posible el reconocimiento de

[18] La base de datos es mantenida por el Centre National de la Recherche Scientifique (CNRS) y la Universidad de Chicago, dentro del proyecto ARTFL.

prototipos léxicos en las realizaciones argumentales mediante la extracción de datos de frecuencias. Sin embargo, el uso de corpus presenta limitaciones para obtener datos significativos sobre el comportamiento de realizaciones argumentales menos frecuentes, tanto en solitario como combinadas con otros argumentos. A este respecto cabe señalar las dificultades encontradas para obtener ejemplos válidos y suficientes para muchas realizaciones adjetivas de argumentos como, por ejemplo, en *déménagement* (*démenagement familial* = *de la famille*, 'agente animado') o en *mort* (*mort cancérigène* = *par cancer*, 'causa'). Pero, sin duda, la principal limitación que presentan los corpus para el tipo de investigación que realizamos es el hecho de que ninguno está anotado semánticamente. Como veremos, esto imposibilita, entre otras cosas, la desambiguación automática de realizaciones similares de argumentos diferentes de un mismo sustantivo como, por ejemplo, *le séjour de Pierre* ('agente') y *le séjour de plaisance* ('clase').

Asimismo, surgieron dificultades en el empleo del lenguaje CQL (Corpus Query Language) (Lexical Computing, 2023a), que permite extraer datos precisos del corpus siguiendo diversos criterios. Para simplificar nuestras búsquedas y evitar obtener resultados ambiguos o carentes de relevancia, en los proyectos *MultiGenera* y *MultiComb* se optó por extraer únicamente datos relacionados con la primera posición argumental de los predicados nominales, es decir, la posición contigua a su núcleo[19]. La inclusión de otras posiciones de los argumentos en nuestra investigación habría requerido la definición de expresiones regulares mucho más complejas y el procesamiento de un volumen de datos difícil de asumir. La Tabla 1 muestra las estructuras para las que se extrajeron datos.

[19] Cabe recordar aquí que en estos proyectos la finalidad de nuestras búsquedas en corpus no era un estudio exhaustivo de los predicados nominales de los sustantivos escogidos, sino la identificación de prototipos léxicos en cada argumento nominal para proceder a la elaboración semiautomática de paquetes semánticos (Valcárcel Riveiro & Pino Serrano, 2023).

Estructuras consultadas	Ejemplos
Det. + Núcleo + Prep. + Nombre	*Le voyage de Marie* *Le voyage en Italie*
Det. + Núcleo + Prep. + Det. + Nombre	*Le voyage de la professeure* *Le voyage depuis le Japon*
Det. + Núcleo + Adjetivo	*Le voyage présidentiel* *Le voyage asiatique*

Tabla 1: *Estructuras consultadas en Sketch Engine para los proyectos* MultiGenera *y* MultiComb

Esta decisión de analizar sólo las estructuras contiguas al núcleo implica que los KWIC[20] extraídos para el análisis presenten problemas de interpretación semántica. La frecuencia de estos casos hace inviable la consulta directa de los ejemplos en el corpus para su desambiguación, por lo que se adoptaron varias soluciones de compromiso. Así, los KWIC no muestran los núcleos argumentales cuando en esta posición se encuentran adjetivos o sustantivos compuestos sin guiones. Por ejemplo, cuando en el núcleo del argumento aparecen *jeune fille* o *petit ami* en el KWIC solo vemos *la présence de la jeune*, *la vidéo de son petit*. En estos casos se analizan los elementos que aparecen en el KWIC (*jeune*, *petit*) como sustantivos. Cuando resulta obvio que se trata de un adjetivo (p.ej. *la conversation avec ce beau* [*garçon*]), se descarta el KWIC. Un problema similar también hizo inviable el procesamiento de sintagmas nominales con determinantes cuantificadores adverbiales (p.ej., *beaucoup de*, *peu de*) o nominales (*la moitié de*, *un tas de*, *la majorité de*) (Gross, 2012, pp. 177-178)[21] ya que los KWIC no mostraban el

[20] KWIC es un acrónimo en inglés que significa *Key Word In Context*. En Sketch Engine, un KWIC se obtiene mediante una expresión regular en CQL y permite visualizar la palabra o expresión buscada en un contexto más amplio. Esto facilita el análisis contextual de los resultados de las búsquedas en el corpus. La palabra clave se muestra en el centro de la pantalla, rodeada de las palabras que la acompañan en las oraciones donde aparece (Lexical Computing, 2023b).

[21] Wilmet (1997, pp. 168-171, 227-229) analizó en detalle estas estructuras clasificándolos como cuantificadores estrictos compuestos (p. ej. *un morceau de*, *beaucoup de*, *assez de*) o como cuantificadores-caracterizantes preposicionales (p. ej. *une sorte de*, *une espèce de*).

núcleo del argumento, p.ej.: *la présence de la majorité*, *la largeur de beaucoup de*, *l'odeur de la plupart*.

Asimismo, esta limitación de contexto en las búsquedas dificultó la desambiguación de palabras polisémicas. Por ejemplo, en el KWIC *le voyage de la femme* es imposible saber, sin verificarlo en el ejemplo fuente, si aquí el sustantivo *femme* es sinónimo de *épouse (la femme de Carla*) o si se refiere al género de una persona (*la femme de l'épicerie*). Sin embargo, en KWIC *le voyage de sa femme* el posesivo indica claramente que el significado es 'esposa'. Para zanjar el problema, estos casos ambiguos se anotaron semánticamente con la etiqueta más general: en el caso de *femme* como 'ser humano femenino'.

Las dificultades en el análisis semántico pueden aparecer también en algunos núcleos nominales como *présence* o *mort*. En estos casos se tuvieron que descartar las búsquedas con determinantes en plural (por ejemplo, *les présences, les morts*) porque esto podría interpretarse de diferentes maneras. Así, en el caso de *présence*, la forma plural *les présences* indica más bien 'persona o entidad presente' (p.ej., *les présences dans la salle de réunion étaient nombreuses*) o más raramente 'momento en el que alguien está presente' (p.ej., *ses présences dans la salle de classe étaient toujours appréciées par les élèves*), y no ya el significado estudiado en los proyectos referente al 'hecho de encontrarse en un lugar'.

3.2.2. WordNet y las herramientas PORTLEX

Los proyectos MultiGenera y MultiComb demandaban la extracción de una gran cantidad de datos léxicos etiquetados semánticamente. Con ellos se elaboraron paquetes semánticos para alimentar los prototipos que generaban frases nominales personalizadas de manera automática (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019; Domínguez Vázquez, Bardanca Outeiriño & Simões, 2021). WordNet respondía en gran medida a las necesidades de los proyectos y resultó esencial para su desarrollo desde un primer momento. Esto se debe, sin duda, al particular tratamiento de la

información semántica en WordNet. Esta sólida base de datos léxica organiza semánticamente palabras en conjuntos de sinónimos conocidos como *synsets* (Miller, Beckwith, Fellbaum, Gross & Miller, 1990). Sin embargo, este recurso lingüístico va más allá de un mero diccionario, ya que facilita el análisis automático de textos desde un punto de vista semántico y, por tanto, el desarrollo de herramientas como las que se crearon en los proyectos *MultiGenera* y *MultiComb*.

Otro factor esencial es la amplia cobertura multilingüe de WordNet, lo que garantiza la disponibilidad de una gran cantidad de datos léxicos anotados semánticamente para todas las lenguas de trabajo en los proyectos desarrollados en PORTLEX: castellano, gallego, francés, italiano y alemán. Más concretamente, en el caso del francés se utilizó la versión 1.0b4 del *WordNet Libre du Français* (WOLF) (Sagot & Fišer, 2008)[22]. Esencialmente, WOLF se construyó a partir del WordNet original desarrollado en Princeton. A lo largo del tiempo, se ha ido ampliando y mejorando en diferentes versiones en las que se han procesado polisemias y nominalizaciones, recurriendo incluso a la depuración manual de datos inconsistentes. Este recurso se encuentra en formato XML y se puede descargar libremente en la web del proyecto[23].

Sin embargo, aunque la cobertura léxica de esta herramienta respondía a las necesidades de los proyectos *MultiGenera* y *MultiComb*, la información semántica asociada en el WOLF a las unidades de significado o *synsets* es limitada. Esta carencia se suplió mediante un emparejamiento con los datos de EuroWordNet[24] del Multilingual Central Repository (MCR), ya que

[22] El WOLF fue desarrollado en Francia por el equipo ALMAnaCH, que trabaja en el seno del Institut National de Recherche en Informatique et en Automatique (INRIA).
[23] https://almanach.inria.fr/software_and_resources/WOLF-fr.html
[24] EuroWordNet es una base de datos multilingüe que incluye WordNets en varios idiomas europeos. Cada idioma estructura su propio WordNet en *synsets* y relaciones semánticas básicas entre ellos siguiendo una organización similar al WordNet original de Princeton.

estos están asociados a información semántica categorizada en diferentes ontologías como Suggested Upper Merged Ontology (SUMO), Top Concept Ontology (Top), WordNet Domains, Basic Level Concept y Epinónimos. Este paso tan importante se realizó en colaboración con el equipo de GalNet[25], la versión de WordNet para el gallego (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019; Gómez Guinovart & Solla Portela, 2020). Gracias a esto, resultó posible hacer búsquedas semánticas de datos léxicos y refinarlas, además, por tipos de palabra. De esta manera se extrajeron de manera masiva los datos necesarios para crear paquetes anotados semánticamente siguiendo un sistema de etiquetado concebido *ad hoc* para los proyectos del portal PORTLEX (Domínguez Vázquez, Valcárcel Riveiro & Bardanca Outeiriño, 2021). Sin embargo, para poder realizar la extracción de datos en Galnet de manera efectiva se desarrollaron diferentes herramientas específicas.

En primer lugar, se diseñaron diferentes APIs de consulta a Galnet, siendo una de ellas específica para el francés. Su función es la de obtener datos léxicos filtrados por categorías semánticas de las diferentes ontologías (SUMO, TOP, etc.). Esta API proporciona resultados con gran rapidez en formato JSON, lo que facilita la utilización de los datos obtenidos en otros recursos. Sin embargo, los datos proporcionados por la API se limitaban a una única categoría de una ontología concreta. Esto suponía un problema, ya que estas categorías ontológicas estaban organizadas por criterios semánticos generales y, por lo tanto, no atendían al contexto de uso de cada palabra o lexema, esto es, a su predicado y a las combinatorias que se dan en él. Así,

---

Estos WordNets están conectados a través de un índice interlingüe (ILI), lo que permite consultar lexemas similares en otros idiomas y acceder a una ontología compartida con 63 distinciones semánticas comunes a todos los idiomas. El proyecto EuroWordNet comenzó en 1994 y se completó en 1999, estableciendo una base sólida para la expansión de recursos lingüísticos en diferentes idiomas europeos (Vossen, 1999).

[25] GalNet es un recurso desarrollado por el Seminario de Lingüística Informática de la Universidad de Vigo. Este proyecto se enmarca en un esfuerzo más amplio de integrar de forma coordinada las versiones del WordNet 3.0 en español, catalán, gallego, vasco y portugués. (Gómez Guinovart & Solla Portela, 2018, 2020).

dos substantivos como *pain* (07679356-n)[26] y *farine* (07567707-n)[27] aparecen incluidos en la misma categoría *Food* de la ontología SUMO, también en *Comestible* de TOP o en *Gastronomy* de WordNet Domains. Sin embargo, desde un punto de vista sintáctico-semántico *pain* y *farine* pertenecen a clases o categorías diferentes puesto que sus rasgos semánticos no permiten la combinación con verbos o adjetivos similares[28]. En términos de Gross (2008, 2012), *pain* y *farine* pertenecen a dos clases de objetos diferentes y, aunque ambos lexemas comparten rasgos semánticos como +material +comestible, *pain* es una comida y *farine* un ingrediente. Así, en nuestros prototipos de *Combinatoria* y *Xera* (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019; Domínguez Vázquez, Bardanca Outeiriño & Simões, 2021) dejar estos dos substantivos en un mismo paquete semántico llevaría a la generación de frases nominales semánticamente inaceptables como **le goût de la farine* o **la largeur de la farine*. Para evitar este tipo de resultados y mantener la automatización del proceso de importación de datos se necesitó más granularidad semántica en su extracción desde WordNet.

Para dar respuesta a esta necesidad, se desarrolló la herramienta *Combina*, que permite combinar múltiples consultas semánticas en Galnet para el mismo idioma, ya sea mediante la adición de datos de una consulta inicial a otra o la intersección de resultados de diferentes consultas (ver Figura 1). Los datos resultantes se generan tanto en formato de texto como en JSON. Por ejemplo, con esta herramienta podemos obtener los sustantivos compartidos en las categorías *Gastronomy* de WordNet Domains, *Food* de SUMO y *Artifact* de TOP con el fin de elaborar un paquete semántico de platos o

[26] La glosa o definición proporcionada en WordNet para este *synset* es "food made from dough of flour or meal and usually raised with yeast or baking powder and then baked".
[27] La glosa en inglés para este *synset* es "coarsely ground foodstuff; especially seeds of various cereal grasses or pulse".
[28] Así, por ejemplo, *pain* no puede ser objeto directo de verbos como *saupoudrer* o *tamiser*, ni *farine* puede ser objeto directo de los verbos *couper*, *griller* o *manger*. Por otro lado, *farine* no puede adjetivarse con *rassise* o *croustillante*, ni *pain* con *moulu*.

productos gastronómicos en el que se encuentre *pain*, pero no *farine*. Esta búsqueda cruzada en *Combina* nos proporciona 411 lemas entre los que encontramos *bouillon*, *confiture*, *fondue*, *gâteau, quiche*, *pâté* o *ragoût*, pero no *farine*, *oeuf* o *cannelle*[29].

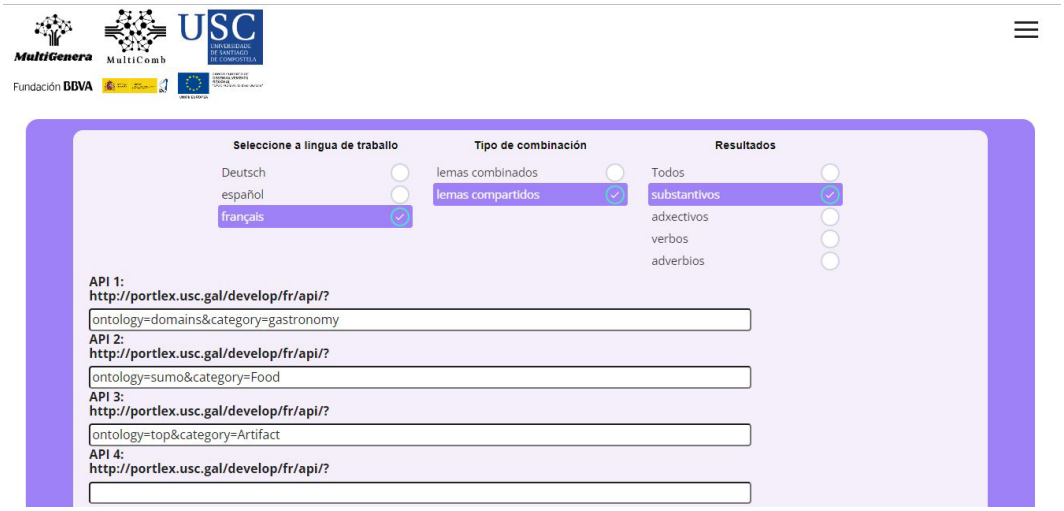


Figura 1: *Vista de la interfaz de Combina en una búsqueda de datos compartidos por tres categorías ontológicas en WordNet*

Naturalmente, extraer datos precisos con Combina requiere tener un buen conocimiento de la estructura de las diferentes ontologías asociadas a WordNet. Dado que este no era el caso de los integrantes de los equipos de lingüistas, entre ellos el de francés, se requirió el desarrollo de una nueva herramienta que facilitase rápidamente la localización de un determinado lexema en las categorías de las diferentes ontologías. Esta nueva herramienta se denominó Lematiza y acepta concordancias de los corpus disponibles en

[29] De todas formas, aunque con *Combina* se obtienen datos mucho más precisos, su revisión y depuración humana es todavía necesaria. Por ejemplo, entre los resultados obtenidos en la búsqueda mencionada antes se incluyen ítems repetidos como *purée*, que se asocian a varios *synsets* o significados en WordNet, o sustantivos como *assiette* ou *viennoiserie*, que no responden completamente a la delimitación semántica del paquete que pretendemos crear.

Sketch Engine o listas de frecuencias en formatos csv y xml (ver Figura 2). Al ejecutarla, se obtienen los lemas de estas consultas, con sus diferentes *synsets* y las categorías ontológicas a las que se asocian en WordNet. Además, la herramienta proporciona enlaces de la API correspondientes a consultas de ontologías externas relacionadas, así como a otros datos de tipo semántico (hipónimo, hiperónimos, etc.) (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019).

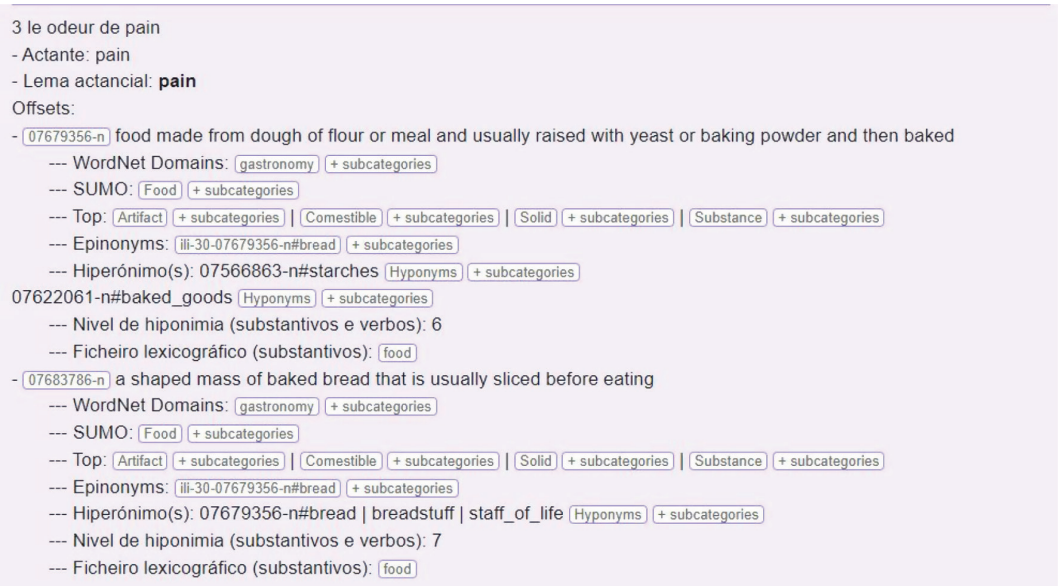


Figura 2: *Vista de los resultados proporcionados por* Lematiza *para los tres* synsets *del lema* pain

3.2.3. FreeLing y el tratamiento de los paradigmas morfológicos

Además de la extracción masiva de datos léxicos y su organización semántica, tanto *Xera* como los prototipos de combinatoria frasal y verbal requerían información morfológica detallada para generar tanto sintagmas nominales como sus contextos oracionales. En el prototipo Combinatoria se necesitaba flexionar el género y número de sustantivos y las correspondientes concordancias en determinantes y adjetivos, así como conjugar los verbos en *CombiContext*. Entre las diferentes opciones disponibles para cada idioma, se priorizó la búsqueda de un recurso multilingüe. De esta manera, se optó

finalmente por FreeLing, una biblioteca desarrollada en C++ que proporciona un conjunto de recursos multilingües para el procesamiento del lenguaje natural. Entre sus diferentes funcionalidades se encuentra el análisis morfológico de numerosas lenguas, incluido el francés. De esta forma FreeLing proporciona información morfológica detallada sobre la flexión de sustantivos, adjetivos, determinantes y verbos (Padró & Stanilovsky, 2012).

De la biblioteca de Freeling no se extrajeron todos los datos disponibles para el francés y las otras lenguas de trabajo, sino solamente las formas flexionadas de los sustantivos y adjetivos presentes en los diferentes paquetes semánticos. Para automatizar la extracción de la información morfológica de FreeLing y su incorporación en los paquetes semánticos obtenidos con *Combina* se diseñó otra herramienta: Flexiona (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019). Gracias a ella, seleccionando la lengua de consulta y cargando el paquete semántico correspondiente, se realizaba rápidamente esta operación. En los paquetes, las formas flexionadas importadas se asocian al lema y a su correspondiente *synset*. Se indica primero la forma flexionada, después el lema, los códigos de tipo de palabra, género y número, y finalmente el *synset*. Los ejemplos (1) y (2) muestran la disposición de los datos en columnas correspondientes al sustantivo *maladie* y al adjetivo *violent*.

(1)	maladie	maladie	N	F	S	14061805-n
	maladies	maladie	N	F	P	14061805-n
(2)	violent	violent	A	M	S	02510879-a
	violents	violent	A	M	P	02510879-a
	violente	violent	A	F	S	02510879-a
	violentes	violent	A	F	P	02510879-a

Aunque esta automatización de la anotación morfológica de los paquetes léxicos fue satisfactoria para la mayoría de los lemas, surgieron problemas en diferentes casos:

a. Los lemas multipalabra (p. ej. *peste noire*, *fruit de la passion*) no se incluyen en la biblioteca de Freeling y, por lo tanto, no se pudo importar para ellos la información morfológica. Esta se tuvo que introducir manualmente durante la elaboración de los paquetes semánticos.

b. Las formas femeninas de singular y plural tampoco constaban en Freeling para muchas profesiones (p. ej, *écrivain*, *professeur*, *transporteur*), por lo que también hubo que incorporarlas de forma manual a los paquetes de datos. Para verificar las formas femeninas de estos sustantivos (p. ej., *écrivaine*, *professeure*, *transporteuse*) se consultó la guía *Femme, j'écris ton nom*, publicado por el CNRS y el Institut de la Langue Fraçaise (Cerquiglini, 1999).

c. En los argumentos locativos (p. ej. *le voyage en Italie*, *le séjour au Portugal*) surge el problema de la variación de la preposición *en* ~ *à*, vinculada generalmente al tipo de lugar y, en el caso de los países, a su género. Aunque Freeling proporcionaba los datos de género para los países, hubo que crear paquetes específicos[30] ya que en realidad la regla contempla no pocas excepciones. Así, entre los países que rigen *en* podemos encontrar nombres masculinos (p. ej., *Israël*, *Iran*) y entre los países que rigen *à* + determinante nombres femeninos (p. ej., *Philippines*, *Seychelles*). Esto llevó, por un lado, a la revisión manual de los paquetes y a la programación de estructuras diferentes para cada preposición, es decir, procesarlas como realizaciones diferentes del mismo argumento.

d. Las palabras con la llamada "h aspirada" no se identifican en las bibliotecas de Freeling y esto complicó la programación de los prototipos. Hubo que crear código específico para evitar que en las frases con

[30] Se elaboraron separadamente tres paquetes de nombres de países que regían la preposición *à* (p. ej. *Cuba*, *Chypre*), nombres de países que regían à + determinante (p. ej. *Yémen*, *Portugal*) y países o regiones que regían la preposición *en* (p. ej. *Italie, Bretagne*).

palabras con h aspirado no se generasen elisiones ni se colocasen apóstrofos (p. ej. *le hamburger* y no **l'hamburguer*, *la Hongrie* y no **L'Hongrie*).

Como se puede ver, todas estas cuestiones hicieron necesaria una revisión manual y exhaustiva de todos los paquetes semánticos básicos obtenidos de *Combina* (92), lo que implicó una importante inversión de tiempo por parte del equipo de trabajo[31]. La necesidad de concentrar tiempo y esfuerzos en la revisión de los paquetes llevó a limitar la morfología de los determinantes y de los verbos con el fin de evitar más complicaciones en la generación automática. En el caso de los determinantes, se decidió limitar al artículo determinado la generación de estos en los sintagmas nominales. Esto se debe a las numerosas restricciones a las que está sujeto el uso de determinantes en estas estructuras (Gross, 1991, pp. 269-270). Esta decisión ahorró eventuales revisiones de las frases generadas, pero llevó a generar grupos nominales poco frecuentes en el uso, véase:

(3) *Les voyages de la mère sont fréquents.*

(4) *La largeur de la tête est surprenante.*

En los dos ejemplos *mère* y *tête* son dos sustantivos asociados a la propiedad inalienable de alguien, que se expresa mediante un complemento preposicional o un determinante posesivo (Rigau, 1999, pp. 345-346). La decisión de limitar la generación de determinantes a los artículos definidos excluía obtener en estos casos resultados más comunes:

(5) *Les voyages de ta mère sont fréquents.*

(6) *La largeur de sa tête est surprenante.*

[31] Cabe recordar aquí que la elaboración de los paquetes conllevó dos revisiones manuales además de esta: una anterior, en la que se depuraban los lexemas no pertinentes extraídos de WordNet, y otra posterior en la que se depuraban los lexemas inapropiados para la combinatoria con un determinado sustantivo. Así, el paquete *fr_animado_planta_arbol* que se combina con el argumento 'tipo' del sustantivo *odeur* no contiene lexemas como *saule* o *hêtre* (p. ej., **l'odeur de saule*) pero sí el que se combina con el argumento 'tema' de *conversation* (p. ej. *la conversation sur le saule*).

En el caso de los verbos, se limitó la generación de la conjugación a las terceras personas del presente del indicativo. Se excluyó, por un lado, el uso de pronombres personales sujeto (*je*, *tu*, *elle*, etc.), y por el otro la generación de verbos en pasado, en futuro o en imperativo. Así, el número de formas importadas de Freeling para cada verbo se limita a tres (infinitivo, tercera persona del singular y tercera persona del plural), p. ej. *penser*, *pense*, *pensent*. Esto permitió reducir considerablemente el número de datos almacenados en la base de datos y simplificar eventuales revisiones.

4. CONCLUSIONES

En la presente contribución se ha hecho un repaso al trabajo desarrollado sobre el francés en el portal lexicográfico PORTLEX. Los equipos que trabajaron con esta lengua en los tres proyectos realizados (el diccionario PORTLEX, *MultiGenera* y *MultiComb*) se enfrentaron a importantes desafíos relacionados, sobre todo, con la obtención y procesamiento de grandes cantidades de datos léxicos. Para ello se usaron una serie de herramientas y recursos procedentes de diferentes ámbitos como la lingüística computacional (corpus, bases de datos, repositorios), la lexicografía (diccionarios, léxicos) o la gramática (gramáticas, artículos de investigación). Además, dentro del portal se desarrollaron nuevas aplicaciones para simplificar el trabajo de los equipos de lingüistas.

Sin embargo, a pesar de todas las herramientas disponibles, los proyectos PORTLEX requirieron mucho tiempo de trabajo por parte de estos equipos. La particular naturaleza valencial y multilingüe del diccionario y los prototipos de generación desarrollados exigió un minucioso trabajo previo de análisis lingüístico y depuración de datos léxicos imposible de automatizar completamente. A esto se le sumaron las diferentes limitaciones que presentan las herramientas y recursos empleados y que se detallan en la parte central del artículo. Afortunadamente, todos estos problemas no supusieron una interrupción del trabajo ya que se fueron encontrando soluciones que los

resolvían o los paliaban. De esta forma, se consiguieron alcanzar los principales objetivos en cada proyecto y, finalmente, tanto el diccionario PORTLEX como los prototipos de generación *Combinatoria* y *CombiContext* son una realidad. De hecho, los resultados alcanzados para el francés se pueden consultar en acceso libre y de manera contrastada con los datos de otras lenguas de trabajo.

Todo lo aprendido en estos tres proyectos redundará, sin duda, en una mayor eficacia de los equipos a la hora de afrontar el nuevo desafío que supone ESMAS-ES+. De hecho, este proyecto, además de presentar la orientación multilingüe y valencial que caracteriza el trabajo en el portal PORTLEX desde sus inicios, se define por una vocación de sostenibilidad. Esto implicará la reutilización de datos y herramientas de los proyectos anteriores para desarrollar un etiquetador semántico evitando el desperdicio de tiempo, material y dinero. En última instancia, se pretende que estos esfuerzos contribuyan a impulsar un futuro más eficiente y sostenible en la investigación lingüística y el procesamiento del lenguaje natural.

REFERENCIAS BIBLIOGRÁFICAS

Abeillé, A. & Godard, D. (dirs.) (2021). *La grande grammaire du français*. Actes Sud.

Alonso-Ramos, M. (2004). *Diccionario de colocaciones del español (DICE)*. http://www.dicesp.com/paginas

Alonso-Ramos, M. (2017). Diccionarios combinatorios. *Estudios de Lingüística del Español*, *38*, 173-201. https://doi.org/10.36950/elies.2017.38.8651

Barrios Rodríguez, M. A. (2010). El dominio de las funciones léxicas en el marco de la Teoría Sentido-Texto. *Estudios de Lingüística del español (ELiEs)*, *30*. http://elies.rediris.es/elies30/index30.html

Blanco, X. (1997). De las clases de objetos a las clases de predicados. *Verba*, *24*, 371-385.

Blanco, X. (1999). *Lexicographie bilingue français-espagnol et classes d'objets*. Universitat Autònoma de Barcelona.

Cerquiglini, B. (dir.) (1999). *Femme, j'écris ton nom… Guide d'aide à la féminisation des noms de métiers, titres, grades et fonctions*. Centre National de la Recherche Scientifique & Institut de la Langue Française. https://medias.vie-publique.fr/data_storage_s3/rapport/pdf/994001174.pdf

Charaudeau, P. (1992). *Grammaire du sens et de l'expression*. Hachette-Éducation.

Condette, M. H., Marín, R. & Merlo, A. (2012). La structure argumentale des noms déverbaux: du corpus au lexique et du lexique au corpus. En F. Neveu, V. M. Toke, P. Blumenthal, T. Klingler, P. Ligas, S. Prévost & S. Teston-Bonnard (eds.), *Actes du 3ème Congrès Mondial de Linguistique Française, Lyon, France, 4-7 juillet 2012* (pp. 845-858). SHS Web of conferences. https://doi.org/10.1051/shsconf/20120100271.

Dendien, J. (2020). *Le TLFi Trésor de la langue française informatisé*. Analyse et traitement informatique de la langue française. http://atilf.atilf.fr/tlfv3.htm

Domínguez Vázquez, M.ª J. (2011). *Kontrastive Grammatik und Lexikographie: spanisch-deutsches Wörterbuch zur Valenz des Nomens*. Iudicium

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

Domínguez Vázquez, M.ª 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.), *Post-editing Lexicography – Elex 2021. Proceedings of the eLex 2021 conference* (pp. 269-287). European Lexicographic Infrastructure. https://elex.link/elex2021/wp-content/uploads/eLex_2021-proceedings.pdf

Domínguez Vázquez, M.ª 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. Ferreira, 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. 1-3 October 2019, Sintra, Portugal* (pp. 51-71). Lexical Computing CZ, s.r.o.

Domínguez Vázquez, M.ª J. & Valcárcel Riveiro, C. (2020). PORTLEX as a multilingual and cross-lingual online dictionary. En M.ª J. Domínguez Vázquez, M. Mirazo Balsa & C. Valcárcel Riveiro (eds.), *Studies on multilingual lexicography* (pp. 135-158). De Gruyter. https://doi.org/10.1515/9783110607659-008

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

Domínguez Vázquez, M.ª J., Valcárcel Riveiro, C. & Lindemann, D. (2018). Multilingual Generation of Noun Valency Patterns for Extracting Syntactic-Semantical Knowledge from Corpora (*MultiGenera*). En J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana, Slovenia* (pp. 847-854). Ljubljana University Press.

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

Gómez Guinovart, X. & Solla Portela, M. A. (2018). Building the Galician WordNet: methods and applications. *Language Resources & Evaluation 52*(1), 317-339. https://doi.org/10.1515/9783110607659-010

Gómez Guinovart, X. & Solla Portela, M. A. (2020). Construction of a WordNet-based multilingual lexical ontology for Galician. En M.ª J. Domínguez Vázquez, M. Mirazo Balsa

& C. Valcárcel Riveiro (eds.), *Studies on multilingual lexicography* (pp. 179-196). De Gruyter.

Grevisse, M. & Goosse, A. (2008). *Le bon usage. Grammaire française*. De Boeck-Duculot.

Gross, G. (1991). Syntaxe du complément de nom. *Linguisticae Investigationes, 15*, 255-284. https://doi.org/10.1075/li.15.2.02gro

Gross, G. (2002). Analyse de compléments du nom en termes de classes d'objets. *Le français moderne*, *70*(2), 187-209.

Gross, G. (2004). Réflexions sur le traitement automatique des langues. En G. Purnelle, C. Fairon & A. Dister (eds.), *Le Poids des mots. Actes des 7es Journées internationales d'Analyse Statistique des Données Textuelles (JADT 2004)* (545-556). Presses Universitaires de Louvain.

Gross, G. (2008). Les classes d'objets. *Lalies*, *28*, 111-165.

Gross, G. (2012). *Manuel d'analyse linguistique*. Presses Universitaires du Septentrion. https://doi.org/10.4000/books.septentrion.115128

Iordanskaja, L., Kim, M. & Polguère, A. (1996). Some Procedural Problems in the Implementation of Lexical Functions for Text Generation. En L. Wanner (ed.), *Lexical functions in lexicography and natural language processing* (pp. 279-297). John Benjamins.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The Tenten Corpus Family. En A. Hardie & R. Love (eds.), *Proceedings of the 7th International Corpus Linguistics Conference CL* (pp. 125127). Lancaster University. https://www.sketchengine.eu/wp-content/uploads/The_TenTen_Corpus_2013.pdf

Lazard, G. (1988). Définition des actants dans les langues européennes. En J. Feuillet (ed.), *Actance et valence dans les langues de l'Europe* (pp. 11-146). Mouton de Gruyter. https://doi.org/10.1515/9783110804485.11

Lazard, G. (1994). *L'actance*. Presses Universitaires de France.

Lexical Computing (2023a). CQL - Corpus Query Language. *Sketch Engine*. https://www.sketchengine.eu/documentation/corpus-querying/

Lexical Computing (2023b). Concordance - a tool to search corpus. *Sketch Engine*. https://www.sketchengine.eu/guide/concordance-a-tool-to-search-a-corpus/

Mel'čuk, I. (1997). *Vers une linguistique Sens-Texte. Leçon inaugurale*. Collège de France. http://olst.ling.umontreal.ca/pdf/melcukColldeFr.pdf

Mel′čuk, I., Arbatchewsky-Jumarie, N., Iordanskaja, L., Mantha, S. & Polguère, A. (eds.) (1984-1999). *Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques I-IV*. Presses de l'Université de Montréal. https://doi.org/10.2307/j.ctv69t5n2

Mel′čuk, I., Clas, A. & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Éditions Duculot.

Mel′čuk, I. & Milićević, J. (2014). *Introduction à la linguistique. Volume 1*. Éditions Hermann.

Mel'čuk, I. & Polguère, A. (2007). *Lexique actif du français: l'apprentissage du vocabulaire fondé sur 20000 dérivations sémantiques et collocations du français*. De Boeck.

Miller, G. A., Beckwith, R., Fellbaum, C, Gross, D & Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, *3*(4), 235-244. https://doi.org/10.1093/ijl/3.4.235

Montémont, V. (2020). De Frantext 1 à Frantext 2: la cure de jouvence d'une vieille dame. En D. Aquino-Weber & Y. Greub (eds.), *La lexicographie informatisée: les vocabulaires nationaux dans un contexte européen* (pp. 41-66). Académie suisse des sciences humaines et sociales.

Padró, L. & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. En N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.), *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. European Language Resources Association. https://nlp.lsi.upc.edu/publications/papers/padro12.pdf

Picallo, M. C. (1999). La estructura del sintagma nominal: las nominalizaciones y otros sustantivos con complementos argumentales. En I. Bosque & V. Demonte (dirs.), *Gramática descriptiva de la lengua española* (pp. 363-393). Espasa Calpe.

Pierrel, J. M. (2003). Un ensemble de ressources de référence pour l'étude du français: TLFI, FRANTEXT et le logiciel STELLA. *Revue québécoise de linguistique*, *32*, 155-176. https://doi.org/10.7202/012248ar

Riegel M., Pellat, J.C. & Rioul, R. (2009). *Grammaire méthodique du français*. Presses Universitaires de France.

Rigau, G. (1999). La estructura del sintagma nominal: los modificadores del nombre. En I. Bosque & V. Demonte (dirs.), *Gramática descriptiva de la lengua española* (pp. 311-362). Espasa Calpe.

Sagot, B. & Fišer, D. (2008). Building a free French wordnet from multilingual resources. En N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (eds.), *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. European Language Resources Association. https://inria.hal.science/inria-00614708/document

Stage, L. (1994). La valence des noms en français, En M. Herslund (ed.), *Noun Phrase Structures* (pp. 93-131). Samfundslitteratur.

Stage, L. (1997). La transposition des actants dans le syntagme nominal. Étude sur la nominalisation nucléaire et l'emploi des prépositions. *Revue Romane*, *32*(1), 51-86.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Klincksieck.

Valcárcel Riveiro, C. (2017). Las construcciones N_1N_2 como realizaciones actanciales del sustantivo en francés y su tratamiento en el diccionario multilingüe PORTLEX. En M.ª J. Domínguez Vázquez & S. Kutscher (eds.), *Interacción entre gramática, didáctica y lexicografía* (pp. 193-207). De Gruyter. https://doi.org/10.1515/9783110420784-015

Valcárcel Riveiro, C. & Pino Serrano, L. (2023). Application d'une méthodologie d'analyse des prédicats nominaux: l'exemple du lexème MORT$_1$. *Çédille. Revista de estudios franceses*, *24* (en prensa).

Vossen, P. (ed.) (1999). *EuroWordNet. General Document 1, Final*. University of Amsterdam. https://archive.illc.uva.nl/EuroWordNet/docs/GeneralDocDOC.zip

Wilmet, M. (1997). *Grammaire critique du français*. Duculot.

MUCHO MÁS QUE EJEMPLOS: APLICACIONES DIDÁCTICAS DE LOS GENERADORES AUTOMÁTICOS

BEYOND EXAMPLES: DIDACTIC APPLICATIONS OF AUTOMATIC LANGUAGE GENERATORS

Nerea López Iglesias
Saxony International School
nerea.iglesias@trias.lernsax.de

Resumen

En este capítulo se aborda la aplicación de las herramientas de generación automática del lenguaje natural en el aula de lenguas extranjeras. Se discute la utilidad de los ejemplos generados automáticamente en el contexto del aula, se describe una tipología de actividades diseñadas con diversas herramientas digitales y se analiza su potencial en el desarrollo de la competencia léxica y lingüística del alumnado.

Palabras clave: aplicaciones didácticas, enseñanza de lenguas, generadores automáticos del lenguaje natural, herramientas digitales.

Abstract

This chapter focuses on the implementation of natural language generation tools in the teaching of foreign languages. It examines the value of automatically generated examples within the teaching-learning context setting, presents a range of activities devised with diverse digital tools, and evaluates their potential in fostering students' lexical and linguistic proficiency.

Keywords: didactic applications, language teaching, natural language generators, digital tools.

DOI: https://doi.org/10.17561/rilex.6.3.8191

1. INTRODUCCIÓN

El desarrollo de herramientas de generación automática permite solventar algunas de las dificultades que acarrea el trabajo lexicográfico: específicamente, aquellas que tienen que ver con la búsqueda y selección de ejemplos representativos de la lengua real para ser incluidos en las herramientas lexicográficas (Kilgarriff, Husák, McAdam, Rudnell & Rychlý, 2008; Kosem, Koppel, Zingano, Michelfeit & Tiberius, 2019). Asimismo, sabemos también que en la literatura científica se demandan cada vez más recursos en los que se aporte información de combinatoria argumental (Domínguez Vázquez & Caíña Hurtado, 2021), como los desarrollados en el marco de *MultiGenera*[1] y *MultiComb*[2]. La información que se puede extraer de estas herramientas[3], que parte a su vez de la lengua real gracias al análisis y a la extracción de datos en corpus, es de utilidad, por tanto, a la hora de elaborar diccionarios plurilingües de valencias con un enfoque didáctico (Fuertes-Olivera, Niño Amo & Sastre Ruano, 2019, p. 79) y otras herramientas lexicográficas, pero también puede complementar la información de gramáticas y manuales de lenguas. De esto se infiere que los generadores automáticos pueden tener una aplicación indirecta en el ámbito de la enseñanza y aprendizaje de lenguas. Sin embargo, estas herramientas nos ofrecen mucho más que ejemplos, y prueba de ello es la posibilidad de elaborar recursos didácticos a partir de la información que estas aportan, poniéndose así de manifiesto sus aplicaciones didácticas más directas. A lo largo del presente capítulo llevaremos a

[1] *MultiGenera. Generación multilingüe de estructuras argumentales del sustantivo y automatización de extracción de datos sintáctico-semánticos*. Fundación BBVA. Ayudas Fundación BBVA a Equipos de Investigación Científica - Humanidades Digitales. 2017-2020. http://portlex.usc.gal/multigenera/

[2] *MultiComb. Generador multilingüe de estructuras argumentales del sustantivo con aplicación en la producción en lenguas extranjeras*. FI2017-82454-P: Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia, Generación de Conocimiento. MCIN/ AEI/ FEDER "Una manera de hacer Europa" (EXCELENCIA 2017, 2017-PN091). 2018-2021. http://portlex.usc.gal/multicomb/

[3] Las herramientas concretas son *Xera*, *Combinatoria* y *CombiContext*.

cabo una presentación de los diferentes recursos didácticos digitales que se han diseñado con los prototipos de los proyectos *MultiGenera* y *MultiComb*, explorando así el potencial pedagógico de estas herramientas.

2. EL USO DE EJEMPLOS GENERADOS AUTOMÁTICAMENTE EN EL AULA DE LENGUAS EXTRANJERAS

En el aula de lenguas extranjeras no siempre es sencillo propiciar situaciones en las que el alumnado pueda tener un contacto directo con la lengua meta, aunque sabemos que una mayor exposición a esta tiene una repercusión directa sobre la calidad del aprendizaje y el progreso del alumnado en la adquisición de la lengua (Juan-Garau, 2008). El uso de ejemplos de la lengua real es un buena estrategia para aportar un contexto al aprendizaje de contenidos en la lengua meta. Así, por ejemplo, la adquisición de conocimientos léxicos no se limita a la memorización de palabras aisladas, sino que para la correcta comprensión del significado y uso de estas debemos conocer el contexto sintáctico-semántico (Laufer & Nation, 2012) en el que se insertan. Dicho de otro modo, el contexto juega un papel fundamental en el aprendizaje de vocabulario, ya que proporciona la información necesaria para comprender y recordar palabras de manera efectiva. Asimismo, ayuda a identificar matices y connotaciones y permite, por lo tanto, captar el sentido preciso de una palabra en diferentes situaciones. La adquisición del léxico, en definitiva, va más allá del reconocimiento de palabras aisladas y su significado: "the knowledge of a word not only implies a definition, but also implies how that word fits into the world" (Steven, 2005, p. 95).

El contexto es una parte esencial del desarrollo de la competencia léxica, que es un componente clave en el *Marco Común Europeo de Referencia para las Lenguas* (MCER, 2002). El MCER la define como la habilidad de un individuo para comprender, utilizar y ampliar su vocabulario en la lengua meta. El dominio de un amplio repertorio léxico es esencial para la comunicación efectiva y, por ende, para la consecución de la competencia lingüística en

general. La competencia léxica implica la capacidad de reconocer y comprender palabras en diferentes contextos, así como de seleccionar y utilizar términos apropiados en diversas situaciones comunicativas.

Los ejemplos de lengua real son, como hemos indicado, un método efectivo para proporcionar este contexto en el aula de lenguas extranjeras y, por lo tanto, para mejorar la competencia léxica en particular y, en general, la competencia lingüística del alumnado. Ahora bien, no siempre es sencillo encontrar estos ejemplos en corpus o crearlos *ad hoc*. Además, los docentes no siempre disponen del tiempo y las herramientas necesarias para buscar estos ejemplos y, además, los corpus no siempre ofrecen los ejemplos deseados para cada contexto de enseñanza-aprendizaje. Es por ello por lo que los ejemplos generados automáticamente resultan de utilidad, específicamente aquellos que aportan información de combinatoria argumental, ya sea en el nivel de la frase o de la oración. Los generadores automáticos de los proyectos *MultiGenera* y *MultiComb* ofrecen la posibilidad de generar automáticamente un gran número de ejemplos frasales (monoargumentales y biargumentales) (Domínguez Vázquez, Solla Portela, & Valcárcel Riveiro, 2019; Domínguez Vázquez, Bardanca Outeiriño & Simões, 2021) y, a partir de estas frases nominales, también es posible generar oraciones de cuatro tipos:

a) oraciones copulativas,
b) oraciones en las que la frase nominal tiene la función de SUJETO,
c) oraciones en las que la frase nominal tiene la función de COMPLEMENTO DIRECTO,
d) oraciones en las que la frase nominal tiene la función de SUPLEMENTO.

Esto se hace, además, desde una perspectiva multilingüe, lo que permite también al usuario contrastar la información con otras lenguas y puede ser de utilidad en el aula de lenguas extranjeras, donde a veces resulta útil hacer comparaciones entre la lengua meta y la(s) lengua(s) inicial(es) del alumnado. En definitiva, se generan ejemplos para distintas lenguas con un contexto

enriquecido, variable en cuanto a la selección léxica y filtrados semánticamente. Son ejemplos automatizados (aunque con base en datos de corpus, lo que los acerca a la lengua real) y personalizados. Esto se traduce en una mayor adaptabilidad a las necesidades del usuario y, por consiguiente, del profesorado y alumnado.

3. DISEÑO DE ACTIVIDADES PARA EL AULA DE LENGUAS

Teniendo en cuenta las posibilidades que ofrecen los generadores automáticos, una de las tareas del equipo de investigación del proyecto *MultiComb* ha sido la de poner a prueba su aplicación en el diseño de actividades para el aula de lenguas extranjeras. A continuación se presentarán las herramientas empleadas para el desarrollo de estas actividades, así como la tipología de materiales didácticos creados a partir de la información que ofrecen los generadores.

3.1. Herramientas

La selección de herramientas digitales en línea para la creación de actividades se ha realizado con base en tres criterios principales: que se trate de herramientas gratuitas (tanto para la creación como para la realización de actividades), de acceso libre e integrables en webs y plataformas de aprendizaje. El objetivo principal es, por tanto, que las actividades resulten accesibles tanto para el profesorado como para el alumnado y, por otro lado, que contribuyan también a trabajar la competencia digital en el aula. Teniendo en cuenta estos criterios, se han seleccionado los siguientes recursos digitales para la creación de las actividades:

- *Kahoot!,* una plataforma de aprendizaje en línea que ofrece juegos interactivos basados en preguntas y respuestas.
- *Quizlet*, una plataforma centrada en la práctica y aprendizaje de vocabulario, que permite crear tarjetas de estudio y actividades interactivas, juegos y tests.

- *Flippity,* que permite convertir datos en actividades interactivas como *flashcards*, juegos de memoria, crucigramas y cuestionarios, entre otras.
- *Learningapps.org,* que permite crear y compartir actividades interactivas y recursos educativos. Ofrece una amplia variedad de herramientas, como cuestionarios y juegos, que pueden adaptarse a diferentes niveles.
- *Padlet,* que permite a los usuarios crear y colaborar en tableros virtuales donde se pueden agregar textos, imágenes, vídeos y enlaces, de manera que los usuarios pueden organizar y compartir ideas, proyectos y recursos de un modo visual y colaborativo.

Estas son algunas de las herramientas que cumplen los criterios anteriormente mencionados y que pueden ser utilizadas para el desarrollo de actividades en línea a partir de los datos que ofrecen los generadores. No obstante, podrían emplearse otras muchas herramientas digitales que no se mencionan en la lista anterior y con las que también se podrían explorar las posibilidades didácticas de los ejemplos generados automáticamente. En todo caso, el trabajo con las herramientas mencionadas ha permitido obtener un primer panorama de cuál es la tipología de actividades que se pueden diseñar con estos ejemplos.

3.2. Tipología de actividades

Siguiendo a Nation (2007), las actividades o tareas que se llevan a cabo en el aula de lenguas extranjeras pueden clasificarse en cuatro tipos o *strands*, según en el desarrollo de la competencia en el que pongan el foco:

a) Actividades basadas en la lengua (*language focused*), centradas en la estructura lingüística. Se emplean normalmente ejercicios estructurales o de *drill*.
b) Actividades de recepción oral y escrita (*meaning-focused input).*
c) Actividades de producción oral y escrita *(meaning-focused output*).
d) Actividades del desarrollo de la fluidez (*fluicy development*), como aquellas de lectura rápida de textos conocidos.

A partir de los ejemplos de los generadores automáticos, se han diseñado una serie de actividades-modelo que pueden ser descritas en función de la clasificación anterior. El trabajo en el desarrollo de este tipo de actividades ha permitido observar que las herramientas digitales ofrecen diversas posibilidades de creación de actividades basadas en la lengua, especialmente de aquellas tipo *drill*. Este es el caso de la herramienta *Kahoot!* (Figura 1), con la que se pueden diseñar fácilmente este tipo de tareas en línea partiendo de los datos que ofrecen los generadores:

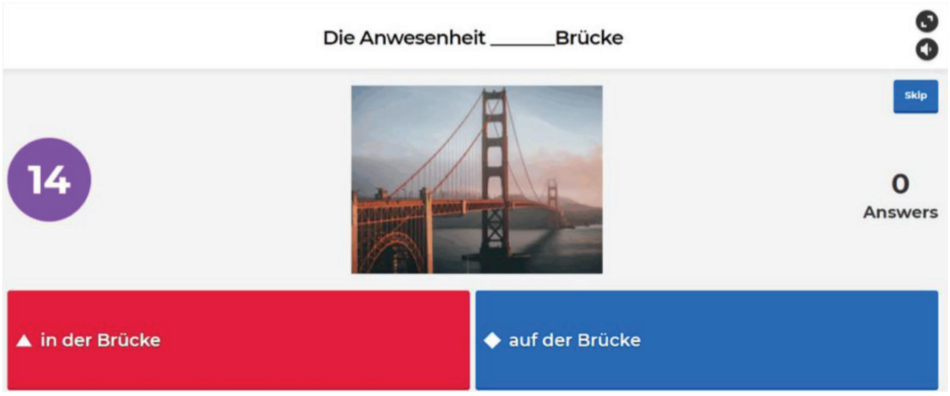


Figura 1: *Ejemplo de pregunta en la actividad de* Kahoot!*: Preposiciones en la frase nominal, para el aprendizaje de alemán*

Como vemos, la actividad consiste en la selección de la preposición y el artículo adecuados para la frase. Los ejemplos que se han tomado como base para el desarrollo de la actividad son aquellos creados por el generador automático de frases nominales y el objetivo de la actividad es el trabajo en los ejes semántico y paradigmático.

También *Quizlet* permite el diseño de este tipo de actividades (Figura 2), como podemos ver en el siguiente ejemplo, donde lo que se ha desarrollado en este caso es un *drill* de asociación en el que el usuario debe seleccionar la frase nominal que puede preceder a la frase dada:

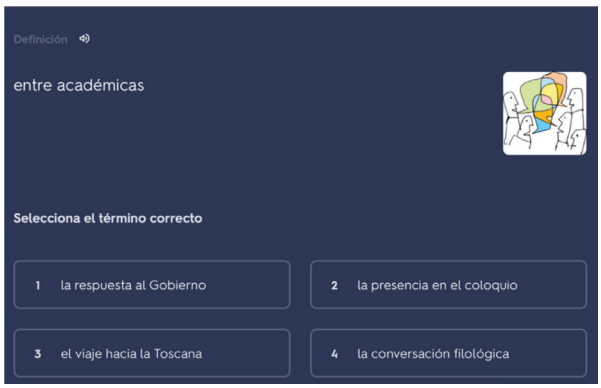


Figura 2: *Actividad de asociación de grupos de frases nominales creada con la herramienta* Quizlet

Tal y como se observa en la imagen, esta actividad se centra en la asociación de frases nominales con sus complementos. El objetivo en este caso es que el alumnado practique por un lado la estructura de la frase nominal y, a la vez, la asociación de significados, lo que puede ser de utilidad para la práctica de la recepción y de la producción escrita y oral.

Por otra parte, para la práctica de la producción escrita y oral se pueden diseñar actividades en las que al alumnado se le da un input léxico y se les pide que construyan frases u oraciones con este input. En este caso, a partir de los ejemplos de los generadores automáticos, también se pueden diseñar actividades lúdicas en línea empleando las herramientas digitales:

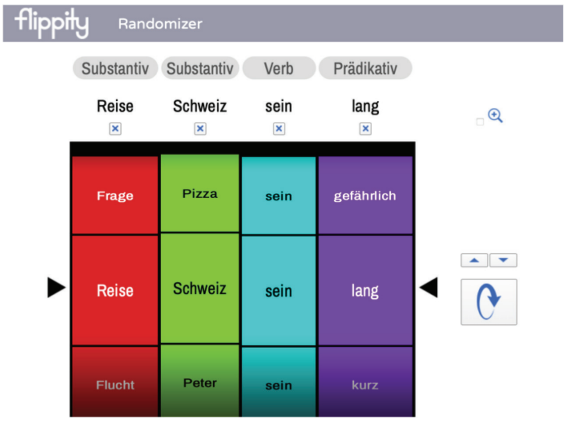


Figura 3: *Actividad de construcción de oraciones a partir del* randomizer *de la herramienta* Flippity

Como vemos en la Figura 3, *Flippity* permite crear ruletas en las que se le ofrece un input léxico al alumnado, con el que deberá producir oraciones con base en una estructura dada. Se trabaja, de este modo, la producción de unidades relativamente simples, aunque la combinación de estas tareas podría emplearse para requerir también la producción de textos más complejos, según las necesidades del aula.

Por otro lado, también se han realizado actividades basadas en la lengua y centradas en el aprendizaje de aspectos sintáctico-semánticos empleando la herramienta *Learningapps* (Figura 4). En este caso, se han realizado actividades en las que el foco está más centrado en el plano semántico y otras en las que prima el plano sintáctico. Todas estas actividades están presentadas en forma de secuencia, por lo que es necesaria la consecución de cada una de ellas para que se vayan desbloqueando las siguientes. Además, todas ellas van precedidas de un enunciado que explica la actividad (Figura 5), lo que permite un mayor grado de autonomía a la hora de llevarlas a cabo.

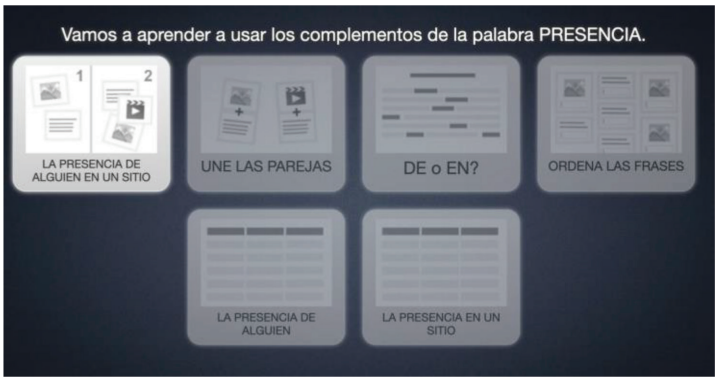


Figura 4: *Actividades para el aprendizaje de los complementos de presencia en español empleando la herramienta* Learningapps

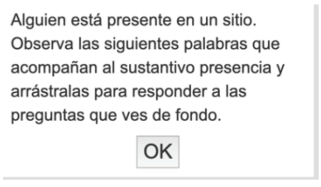


Figura 5: *Enunciado de la primera actividad de la secuencia sobre los complementos de* presencia *en* Learningapps

En la primera actividad (Figura 6) de la secuencia se trabaja fundamentalmente el plano semántico por medio de la asociación de elementos. En ella se pide al usuario que clasifique los complementos del sustantivo *presencia* en función de si pertenecen a la clase semántica de 'animado' o 'lugar', formulando estas dos clases semánticas como preguntas, de forma que sean más fácilmente comprensibles para el usuario:

Figura 6: *Primera actividad de la secuencia sobre los complementos de* presencia *en* Learningapps

En la siguiente actividad (Figura 7) se trabaja con estructuras monoargumentales del sustantivo *presencia* en las que, en función de la selección léxica, la preposición (y, si procede, el artículo) que introduce el complemento variará. Esto permite trabajar simultáneamente el plano sintáctico y el semántico, y a la vez propiciar la reflexión sobre cuestiones gramaticales de modo inductivo:

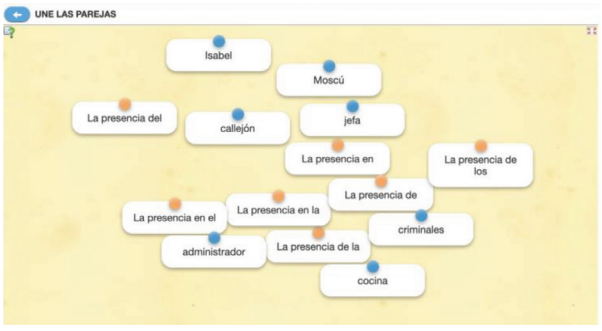


Figura 7: *Segunda actividad de la secuencia sobre los complementos de* presencia *en* Learningapps

Asimismo, esta herramienta también permite diseñar actividades de *drill* (Figura 8) en un sentido más clásico, como es el caso de la tercera actividad de la secuencia, donde el alumnado deberá escoger la preposición correcta ('de' o 'en') en cada una de las estructuras biargumentales:

Figura 8: *Tercera actividad de la secuencia sobre los complementos de* presencia *en* Learningapps

En la siguiente actividad de la secuencia (Figura 9) se pide al usuario que ordene una serie de frases, de modo que en este caso no solo se ofrece una tarea del tipo *language-focused,* sino que al mismo tiempo se realiza un ejercicio de producción escrita controlada que puede servir como primer paso hacia la realización de otras tareas de producción más complejas:

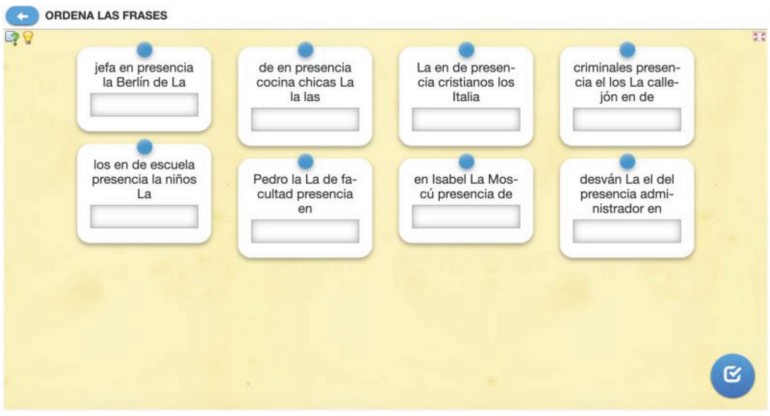


Figura 9: *Cuarta actividad de la secuencia sobre los complementos de* presencia *en* Learningapps

A continuación, se propone una quinta actividad (Figura 10) en la que se trabajan con más detalle los tipos de clases semánticas en las que se pueden agrupar todos los argumentos del sustantivo. Se trata, por lo tanto, de una actividad práctica que promueve la reflexión semántica y que puede servir no solo para el aprendizaje de cuestiones lingüísticas, sino también para la familiarización con las categorías de la ontología empleada en las herramientas de los proyectos *MultiGenera* y *MultiComb*:

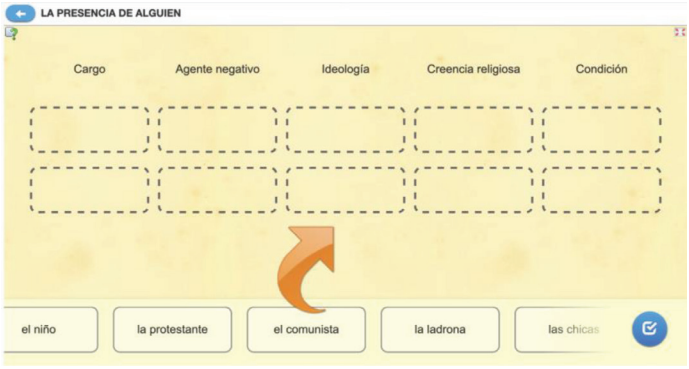


Figura 10: *Quinta actividad de la secuencia sobre los complementos de* presencia *en* Learningapps

Paralelamente, también se ha desarrollado una actividad de asociación semántica para los argumentos del tipo 'lugar' (Figura 11), con una presentación idéntica a la actividad anterior, en la que se clasificaban los argumentos de tipo 'animado'. Estas dos actividades sirven como conclusión a la secuencia, pues tienen una dificultad mayor y promueven la reflexión sobre aspectos semánticos y gramaticales por parte del alumnado.

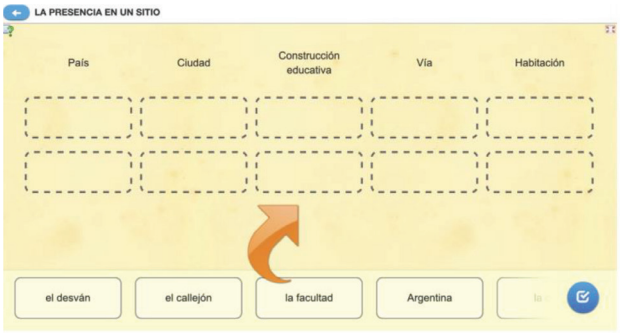


Figura 11: *Sexta actividad de la secuencia sobre los complementos de* presencia *en* Learningapps

Por último, el trabajo con la herramienta *Padlet* ha permitido desarrollar un modelo de actividad de producción escrita y/u oral más compleja que las presentadas hasta ahora (Figura 12). En este caso, se parte de una oración dada y el alumnado debe escribir un pequeño texto o grabar un vídeo o un audio breve. Esta se trata de una actividad interactiva y de gran potencial, no solo para trabajar la producción, sino también para fomentar el trabajo en equipo, la reflexión sobre la producción de textos propios y/o de otros aprendientes, así como la corrección conjunta de la tarea.

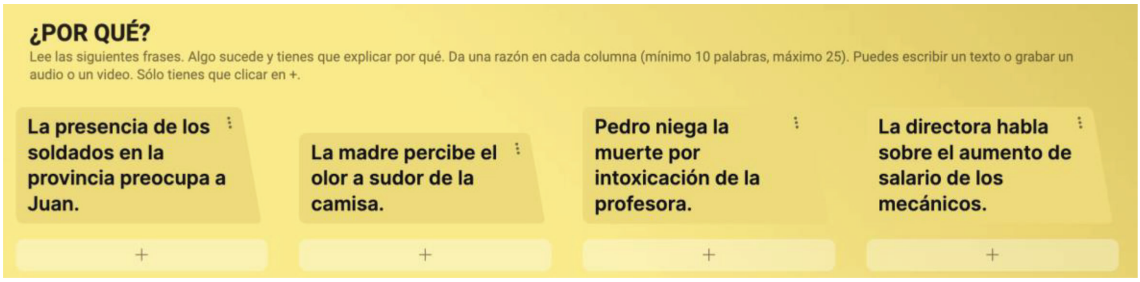


Figura 12: *Actividad de producción diseñada con la herramienta* Padlet *a partir de los ejemplos obtenidos con los generadores automáticos*

Como se ha podido observar en el análisis de las actividades-modelo presentadas, las posibilidades que ofrecen los generadores automáticos permiten enriquecer el diseño de actividades didácticas para la enseñanza-aprendizaje de lenguas extranjeras: evitan la necesidad de buscar ejemplos en corpus o de crearlos *ad hoc* y ofrecen información sintáctico-semántica multilingüe a docentes y alumnado. Mediante la aplicación de herramientas digitales gratuitas y accesibles, es posible diseñar diferentes tipos de actividades: ejercicios de drill, de asociación de frases, construcción de oraciones, clasificación de elementos sintáctico-semánticos, actividades producción escrita y oral, entre otras. Estas actividades son útiles para trabajar la competencia lingüística y fomentar el desarrollo de habilidades comunicativas en el aula de lenguas. El uso de herramientas digitales ha facilitado su implementación, permitiendo un enfoque interactivo, colaborativo y accesible tanto para profesorado como para alumnado.

4. CONCLUSIÓN

La generación automática de ejemplos, además de complementar y suplir algunas carencias de los corpus lingüísticos y herramientas lexicográficas, facilitando así el trabajo lexicográfico, puede tener aplicaciones directas e indirectas en el aula de lenguas. Como ya se ha mencionado, las aplicaciones indirectas tienen que ver con la inclusión de los ejemplos en diccionarios (especialmente en diccionarios plurilingües de valencias), manuales de lengua, gramáticas y otros recursos. En el aula de lenguas extranjeras, los ejemplos generados automáticamente son una estrategia efectiva para proporcionar contexto y mejorar la competencia léxica del alumnado. Los generadores automáticos de los proyectos *MultiGenera MultiComb* ofrecen la posibilidad de generar ejemplos frasales y oraciones en diferentes lenguas con un contexto enriquecido, lo que los hace adaptables a las necesidades del profesorado y alumnado. Por esta razón, este capítulo se ha centrado en las aplicaciones directas, es decir, en el empleo de los ejemplos para la realización de actividades didácticas y didáctico-lúdicas que pueden llevarse a cabo en el aula.

Tal y como se ha podido observar en el análisis de la tipología de actividades, los ejemplos generados automáticamente tienen una gran utilidad a la hora de diseñar actividades de tipo *language focused,* empleando para ello herramientas digitales diversas. Se han diseñado, empleando herramientas digitales en línea, ejercicios de *drill* y de asociación con interfaces y diseños diversos y con grados de dificultad diferentes. Sin embargo, también se ha podido observar que es posible desarrollar otro tipo de tareas, como por ejemplo aquellas que trabajan la producción oral y escrita, así como otras centradas en la reflexión sobre aspectos sintáctico-semánticos. Se trata, no obstante, de una batería de actividades que puede (y debe) ser ampliada y adaptada a niveles diversos (adaptados a las exigencias del MCER), pero que demuestra las posibilidades que ofrecen los generadores automáticos y las aplicaciones directas que tienen los ejemplos en el aula de lenguas. Además,

el diseño de actividades digitales abiertas y en línea ofrece una forma accesible y efectiva de trabajar la competencia lingüística en el aula de lenguas extranjeras, tanto para docentes como para el alumnado que desea aprender o practicar de manera autónoma.

REFERENCIAS BIBLIOGRÁFICAS

Consejo de Europa. (2002). *El Marco común europeo de referencia para las lenguas aprendizaje, enseñanza, evaluación*. Anaya y CVC. https://bit.ly/2IkNKo4

Domínguez Vázquez, M.ª 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 conference* (pp. 269-287). Lexical Computing CZ. https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_16_pp269-287.pdf

Domínguez Vázquez, M.ª J. & Caíña Hurtado, M. (2021). Aplicación de recursos de xeración automática da lingua para estudos comparativos. *Estudos De Lingüística Galega*, *13*, 139-172. https://doi.org/10.15304/elg.13.7409

Domínguez Vázquez, M.ª J., Solla Portela, M. A. & Valcárcel Riveiro, C. (2019). Resources interoperability: Exploiting lexicographic data to automatically generate dictionary examples. En I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek & C. Tiberius (eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference* (pp. 51-71). Lexical Computing CZ. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_4.pdf

Fuertes-Olivera, P. A., Niño Amo, M. & Sastre Ruano, A. (2019). Tecnología con fines lexicográficos: Su aplicación a los *Diccionarios Valladolid-UVa*. *RILE. Revista Internacional de Lenguas Extranjeras, 10*, 75-100. https://doi.org/10.17345/rile10.75-100

Juan-Garau, M. (2008). Contexto y contacto en el aprendizaje de lenguas extranjeras. *IN. Investigació i Innovació Educativa i Socioeducativa*, *1*, 47-66.

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

Kosem, I., Koppel, K., Zingano, T., Michelfeit, J. & Tiberius, C. (2019). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, *32*, 119-137. https://doi.org/10.1093/ijl/ecy014

Laufer, B. & Nation, P. (2012). Vocabulary. S. M. Gass & A. Mackey (eds.), *The Routledge Handbook of Second Language Acquisition* (pp. 163-176). Routledge.

Nation, I. S. P. (2007). The four strands. *Innovation in Language Learning and Teaching, 1*(1), 1-12. https://doi.org/10.2167/illt039.0

Steven, A. S. (2005). *Four Problems with Teaching Word Meanings (and What to Do to Make Vocabulary an Integral Part of Instruction)*. Mahwah.

Recursos propios

CombiContext = Domínguez Vázquez, M.ª J. (dir.), Valcárcel Riveiro, C., Bardanca Outeiriño, D., Calañas Continente, J. A., Catalá Torres, N., Martín Gascueña, R., Mirazo Balsa, M., Sanmarco Bande, M.T. & Pino Serrano, L. (2021). *CombiContext. Prototipo online para la generación automática de contextos frasales y oraciones de la frase nominal en alemán, español y francés*. Universidade de Santiago de Compostela. Instituto da Lingua Galega. Consultado el 28 de noviembre de 2023, de http://portlex.usc.gal/combinatoria/verbal

Combinatoria = Domínguez Vázquez, M.ª J. (dir.), Valcárcel Riveiro, C., Bardanca Outeiriño, D., Calañas Continente, J. A., Catalá Torres, N., López Iglesias, N., Martín Gascueña, R., Mirazo Balsa, M., Sanmarco Bande, M. T. & Pino Serrano, L. (2020). *Combinatoria. Prototipo online para la generación biargumental de la frase nominal en alemán, español y francés*. Universidade de Santiago de Compostela. Consultado el 28 de noviembre de 2022, de http://portlex.usc.gal/combinatoria/usuario

Xera = Domínguez Vázquez, M.ª J. (dir.), Valcárcel Riveiro, C., Bardanca Outeiriño, D., Calañas Continente, J. A., Catalá Torres, N., López Iglesias, N., Martín Gascueña, R., Mirazo Balsa, M., Sanmarco Bande, M.T. & Pino Serrano, L. (2020). *Xera. Prototipo online para la generación automática monoargumental de la frase nominal en alemán, español y francés*. Universidade de Santiago de Compostela. Consultado el 28 de noviembre de 2023, de http://portlex.usc.gal/combinatoria/usuario

Recursos externos

Domínguez Vázquez, M.ª J. (2021). *Kahoot! Preposiciones en la frase nominal*. https://create.kahoot.it/share/preposiciones-en-la-frase-nominal/ff001680-34a3-46f8-bbf6-8c6c3833aee8

Domínguez Vázquez, M.ª J. (2021). *Randomizer*. https://www.flippity.net/ra.asp?k=1FMDTYJNePlaX0Cybg2spfg64JBxW7RUVhM8L_aQ-S7Y

López Iglesias, N. (2021). *Grupos de frases nominales*. https://quizlet.com/578008083/view-screen?redir=%2F578008083%2Flearn

Padlet (2023). *Padlet*. https://padlet.com

Verein LearningApps interaktive Bausteine (s.f.). *LearningApps*. https://learningapps.org