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".

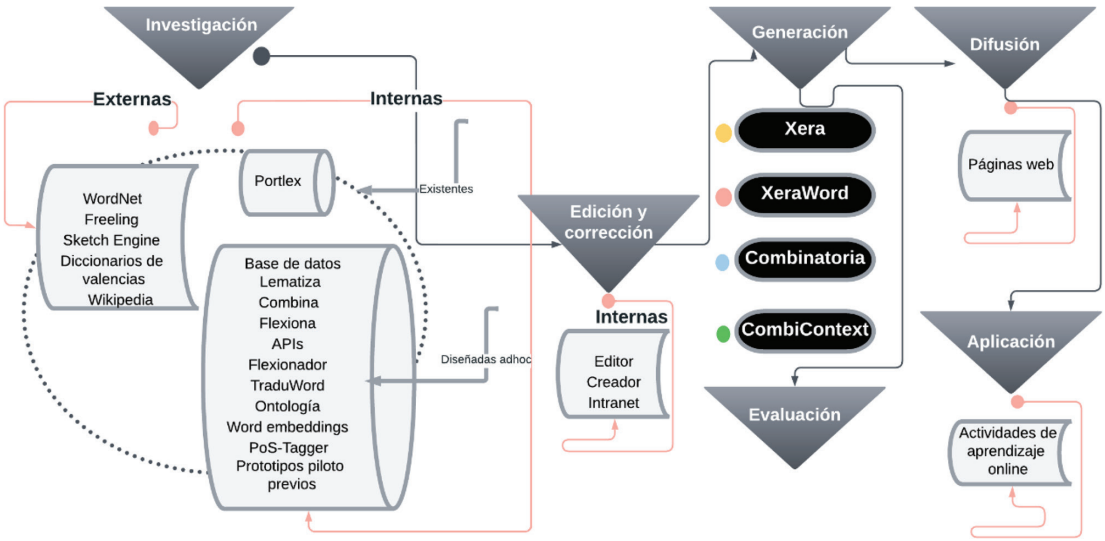


FIGURA 1: *Herramientas y etapas implicadas en el flujo de trabajo*

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:

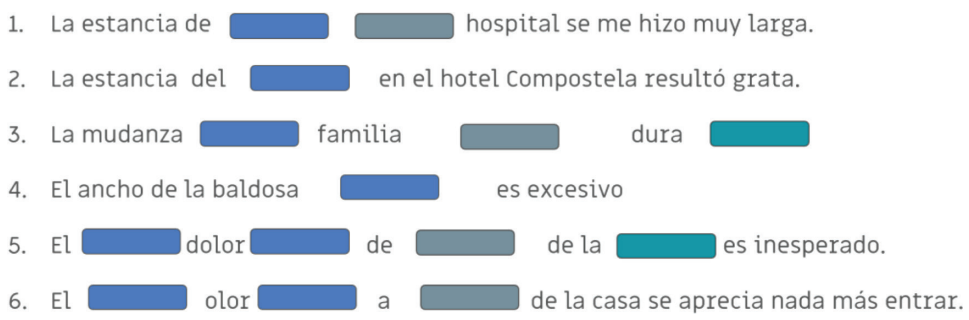


Figura 2: *Ejemplos de patrones argumentales*

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.

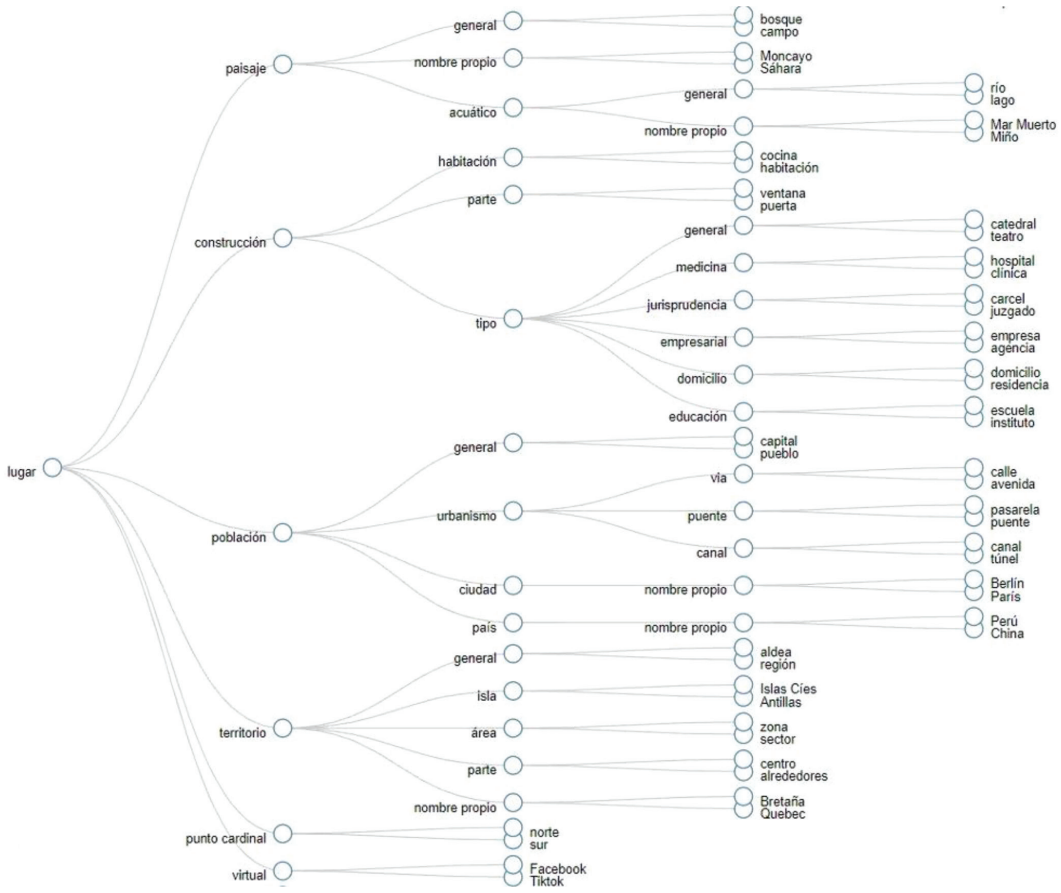


Figura 4: *Vista parcial de la ontología léxica bottom-up*

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*.

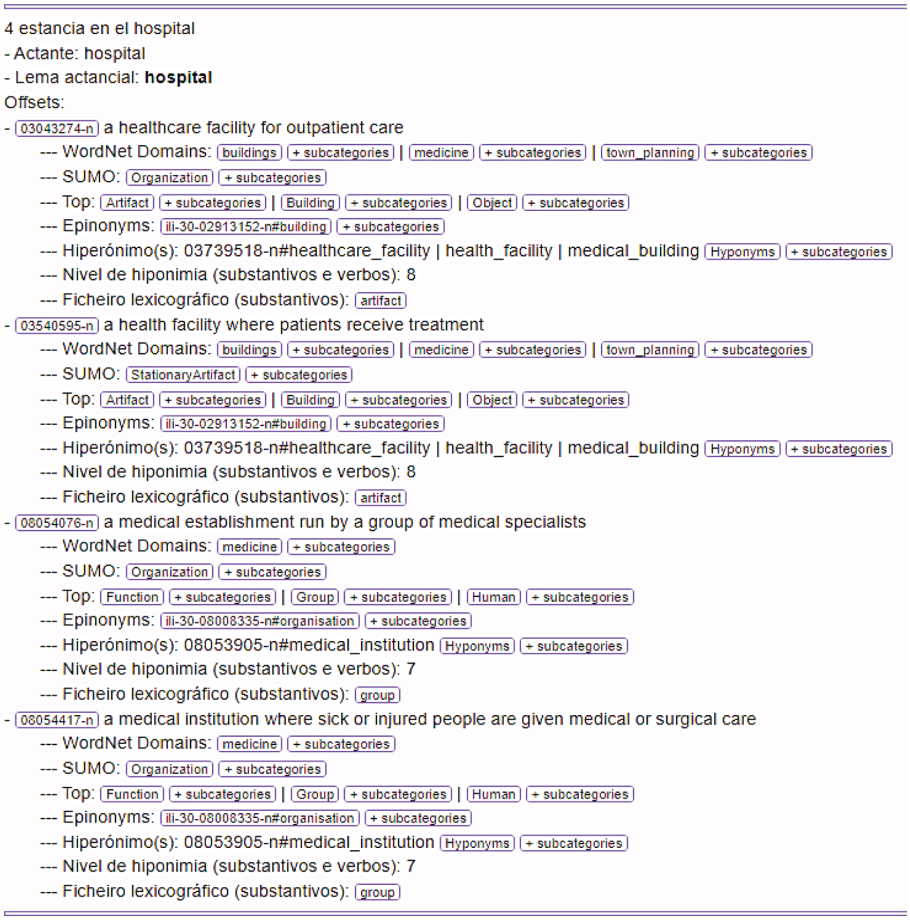


Figura 6: *Información provista por Lematiza*

- 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):

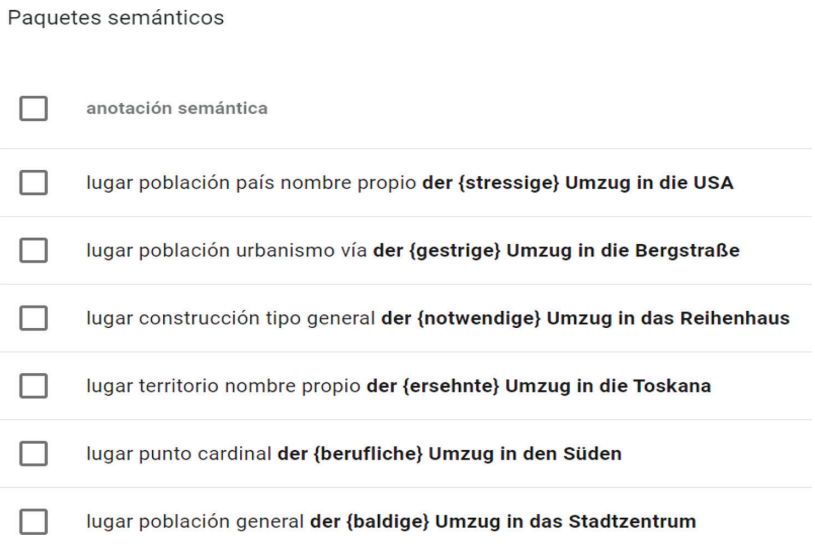


Figura 7: *Paquetes léxicos combinables en la expresión de la dirección del sustantivo Umzug con la preposición in*

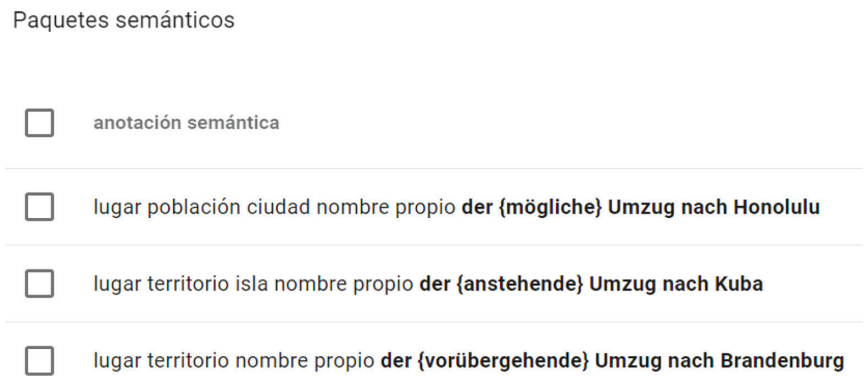


Figura 8: *Paquetes léxicos combinables en la expresión de la dirección del sustantivo Umzug con la preposición nach*

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

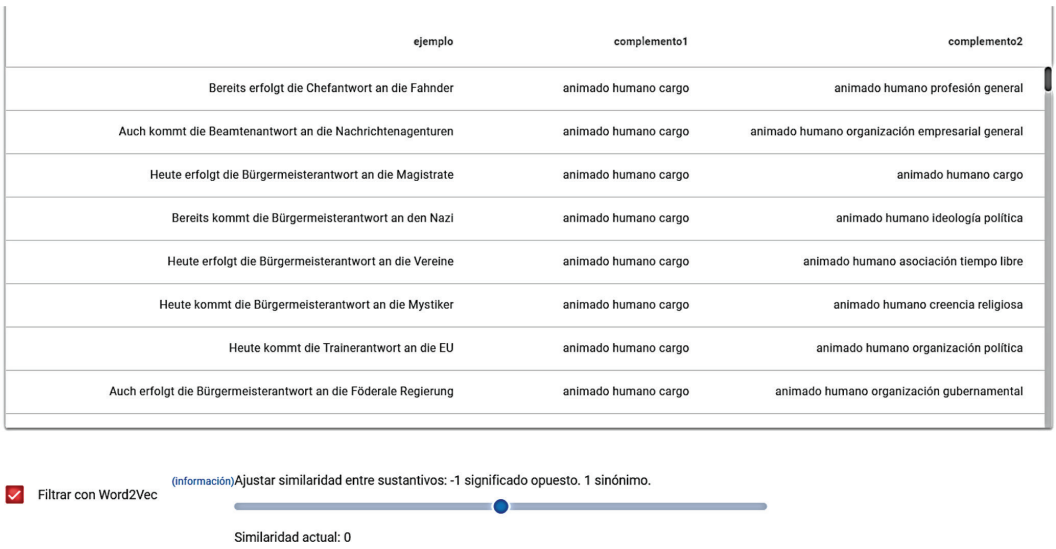
La extracción de los verbos se realiza de manera independiente de la estructura seleccionada, pero ligada al núcleo. Así, se obtiene una lista ordenada

[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/