



# 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](mailto: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.



## 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*<sup>1</sup> (Domínguez Vázquez & Valcárcel Riveiro, 2020)<sup>2</sup>. 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.

---

<sup>1</sup> *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.

<sup>2</sup> <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)<sup>3</sup>, 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

---

<sup>3</sup> 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 *Multi-Genera*, *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

---

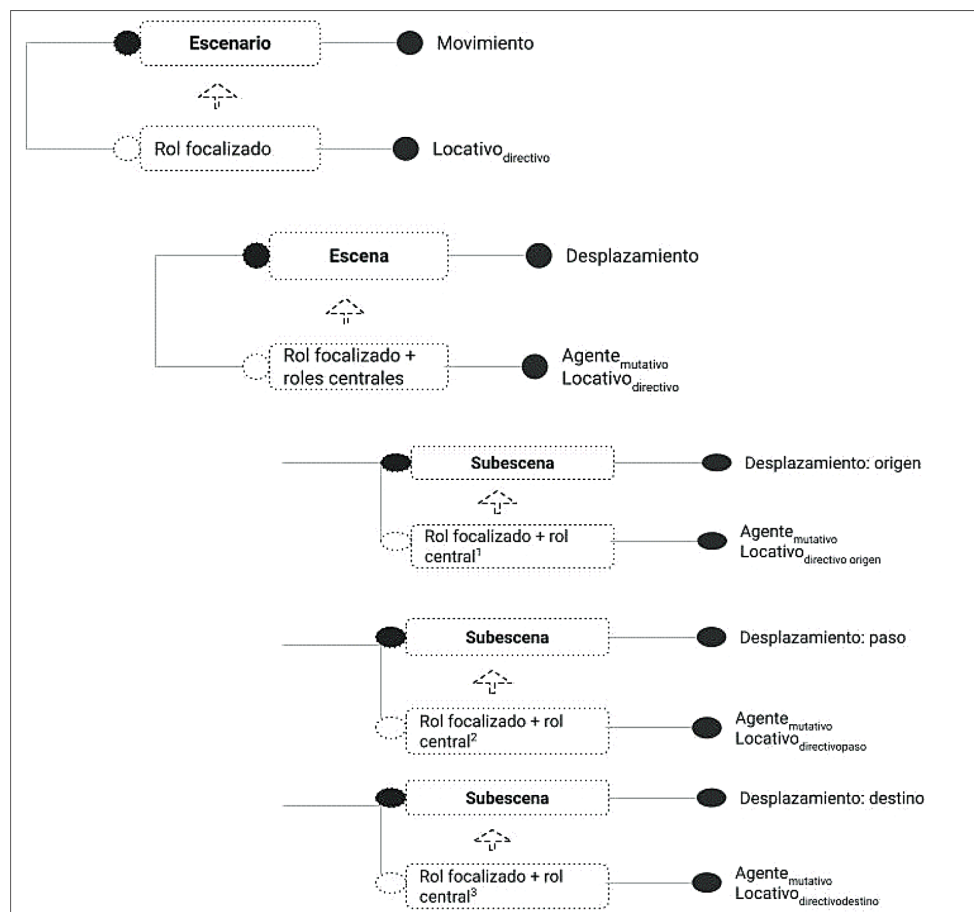
Un estudio detenido de la literatura científica permite constatar no solo diferentes aproximaciones al sustantivo y a su capacidad para ser portador valencial, sino que además evidencia las notables diferencias terminológicas y la asimetría en los inventarios de complementos del nombre y sus realizaciones 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)<sup>4</sup>.

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

---

<sup>4</sup>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):

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

**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

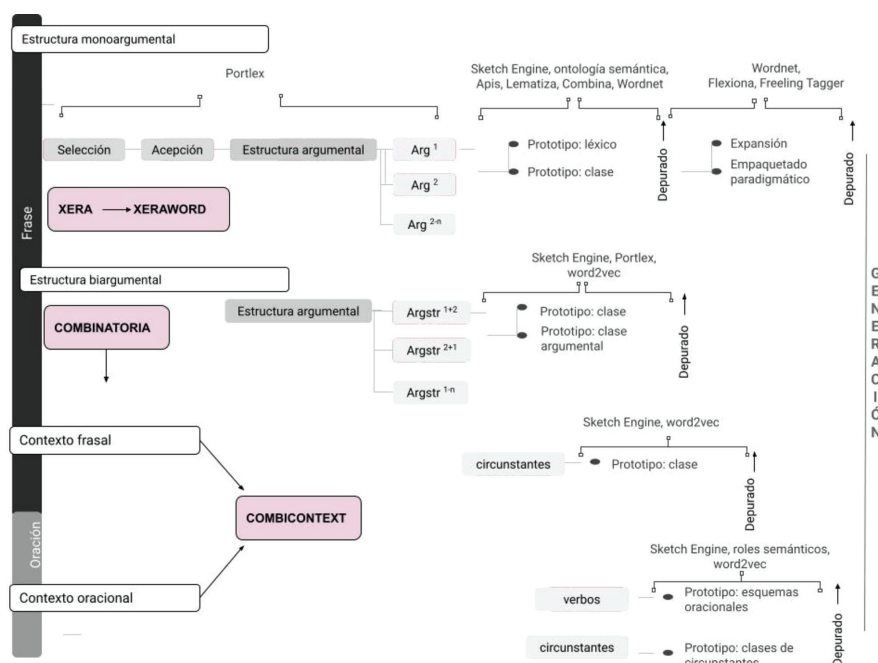
El principal objetivo de los simuladores es ofrecer información sobre el potencial combinatorio de sustantivos valenciales junto con ejemplos dinámicos y, por tanto, el vocabulario que puede cubrir el eje paradigmático y sintagmático de diferentes complementos específicos. Para tal fin se atiende a

- 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}: <i>paciente</i>   <i>enfermo</i>	<i>La estancia del paciente</i>	<i>Xera</i>
determinante + adjetivo + ESTANCIA + en + determinante + ARG 2	N2: lugar en donde se realiza la estancia	{Lugar edificio: tipo: medicina}: <i>hospital</i>   <i>clínica</i>	<i>La estancia en el hospital</i>	
PATRONES SINTÁCTICO-SEMÁNTICOS				
determinante + adjetivo + ESTANCIA + de + determinante+ ARG 1 + en + determinante + ARG 2	N1 {Humano condición negativa} + N2 {Lugar edificio Tipo: medicina}		<i>La estancia del paciente en el hospital</i>	<i>Combinatoria</i>

**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.

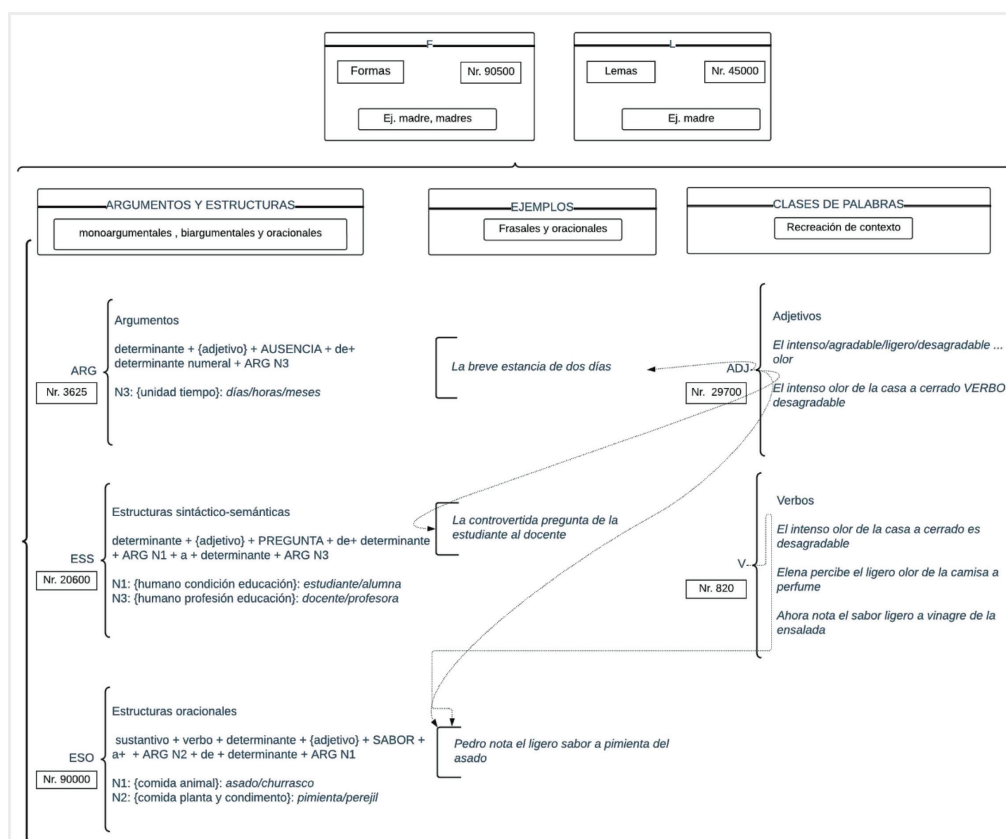


FIGURA 4: Niveles informativos de los generadores

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*<sup>5</sup>, los dos recursos monoargumentales, la aproximación a la consulta es formal. En la herramienta *Combinatoria* la aproximación es ontológico-conceptual. La herramienta *CombiContext* parte

<sup>5</sup> *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	✓	✓	✓	✓
<i>word embeddings</i>			✓	✓
Traducción automática	✓	✓ <sup>1</sup>		
Acceso libre y gratuito	✓	✓	✓	✓
Interfaz de usuario: acceso formal primario	✓	✓		
Interfaz de usuario: acceso semántico primario			✓	
Interfaz de usuario: acceso distribucional primario				✓
Generación <i>ad libitum</i>	✓	✓	✓	✓
Exportación de datos	✓	✓	✓	✓

**TABLA 2:** Información de los generadores en contraste<sup>6</sup>

#### 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

<sup>6</sup> Leyenda: <sup>1</sup> = 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 sintáctico-semántica y, tras su selección, generar ejemplos *ad libitum*.

## XERA

INFORMACIÓN

Idioma:

ES
⌵

Núcleo:

viaje
⌵

Estructura:

determinante+adjetivo o+viaje+adjetivo o+de+determinante+actante N1
⌵

Paquetes semánticos

<input type="checkbox"/>	anotación semántica
<input type="checkbox"/>	animado humano cargo los (cortos) viajes (cortos) de la decana
<input type="checkbox"/>	animado humano condición humana desplazamiento el (increíble) viaje (increíble) de los peregrinos
<input type="checkbox"/>	animado humano condición humana negativa el (peligroso) viaje (peligroso) del fugitivo
<input type="checkbox"/>	animado humano familia el (habitual) viaje (habitual) de los abuelos
<input type="checkbox"/>	animado humano grupo o colectivo militar el (inesperado) viaje (inesperado) del ejército
<input type="checkbox"/>	animado humano ideología política los (recientes) viajes (recientes) de los ultraderechistas
<input type="checkbox"/>	animado humano organización militar los (breves) viajes (breves) de la policía secreta
<input type="checkbox"/>	animado humano origen el (inesperado) viaje (inesperado) de los alemanes
<input type="checkbox"/>	animado humano profesión general los (continuos) viajes (continuos) de los obreros
<input type="checkbox"/>	animado humano condición humana educación el (emocionante) viaje (emocionante) de los universitarios
<input type="checkbox"/>	animado humano condición humana habitante el (inesperado) viaje (inesperado) de los inquilinos
<input type="checkbox"/>	animado humano grupo o colectivo general los (frecuentes) viajes (frecuentes) de la tripulación
<input type="checkbox"/>	animado humano profesión educación el (inesperado) viaje (inesperado) de la docente

Límite de frases: 200

GENERAR

EXPORTAR FRASES EN JSON

EXPORTAR FRASES EN CSV

FIGURA 5: Acceso en Xera

<a href="#">GENERAR</a> <a href="#">EXPORTAR FRASES EN JSON</a> <a href="#">EXPORTAR FRASES EN CSV</a>
frases generadas
El viaje de los ejércitos
El viaje del equipo de rescate
El viaje de la tripulación de bombardero
El viaje de la caballería
El viaje de los destacamentos
El viaje de la patrulla
El viaje de la Guardia Roja
El viaje del ejército
El viaje de la formación militar
El viaje del pelotón de ejecución
El viaje de la tropa paracaidista
El viaje del alto mando
El viaje de los altos mandos
El viaje de la legión
El viaje de las tripulaciones de bombardero
El viaje de la brigada
El viaje de las tropas paracaidistas
El viaje de las legiones
El viaje de las brigadas

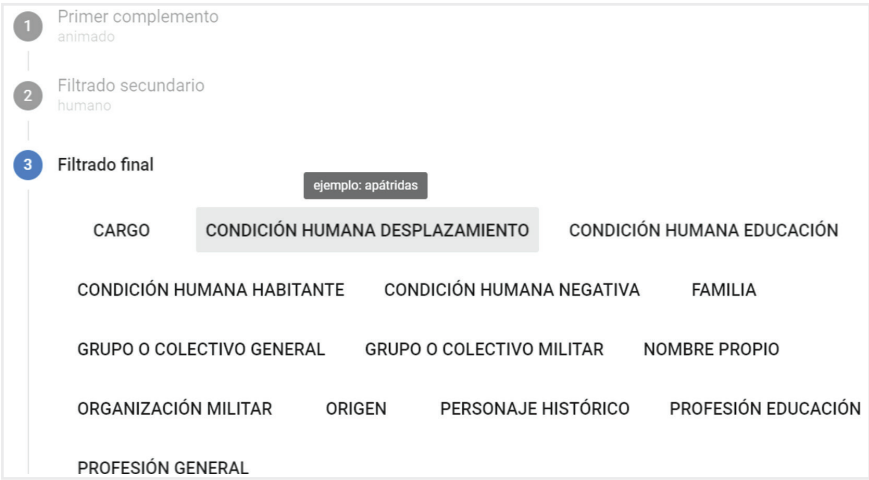
**FIGURA 6:** Volcado de datos en Xera

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

Seleccionar una de la estructuras resultantes para generar ejemplos			Buscar...
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):

1 Primer complemento  
animado

2 Filtrado secundario  
humano

3 Filtrado final  
condición humana desplazamiento

1 Segundo complemento  
lugar

2 Filtrado secundario  
población

3 Filtrado final  
país nombre propio

CARGO

CONDICIÓN HUMANA DESPLAZAMIENTO

CONDICIÓN HUMANA EDUCACIÓN

CONDICIÓN HUMANA HABITANTE

CONDICIÓN HUMANA NEGATIVA

FAMILIA

GRUPO O COLECTIVO GENERAL

GRUPO O COLECTIVO MILITAR

NOMBRE PROPIO

ORGANIZACIÓN MILITAR

ORIGEN

PERSONAJE HISTÓRICO

PROFESIÓN EDUCACIÓN

PROFESIÓN GENERAL

CIUDAD NOMBRE PROPIO

Abisinia

PAÍS NOMBRE PROPIO

GENERAL

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 clicca 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):

1
Seleccionar idioma y núcleo  
viaje

2
seleccionar posición del verbo  
escoger si el verbo aparece antes o después del núcleo non

3
estructura formal  
definir la estructura formal de las oraciones

4
contenido generado  
escoger el contenido semántico de las oraciones generadas

viaje + de  
ejemplos

El viaje de Pedro desde Kuwait es principalmente solitario  
El viaje de Jose Javier desde Abisinia es muy solitario  
El viaje de Euclides desde Kuwait es muy solitario  
El viaje de Armand Jean du Plessis desde Abisinia es principalmente solitario

ELEGIR

viaje + adjetivo  
ejemplos

El viaje posdoctoral de Carlos es largo  
El intento viaje vacacional de Inés es muy peligroso  
El viaje predoctoral de los progenitores es económico  
El viaje posdoctoral de la universitaria es cercano

ELEGIR

viaje + hacia  
ejemplos

El viaje ligero hacia Delfos desde la Commonwealth es turístico  
El viaje económico hacia Buenos Aires desde el área urbana es largo  
El viaje económico hacia Cusco desde el Caribe es largo  
El viaje económico hacia Berlín desde la colonia es largo

ELEGIR

viaje + a  
ejemplos

El viaje económico a Jordania por el área urbana es largo  
El viaje ligero a Costa de Marfil por el Caribe es turístico  
El ligero viaje a Mauritania por el Caribe es largo  
El económico viaje a América por el área urbana es turístico

ELEGIR

viaje + por  
ejemplos

El viaje ligero por el golfo hacia el Caribe es largo  
El viaje ligero por el Lago Michigan hacia el área urbana parece largo  
El importante viaje por el Mar Negro hacia Kuwait principalmente es turístico  
El viaje importante por el Lago Michigan hacia Abisinia especialmente es turístico

ELEGIR

viaje + desde  
ejemplos

El largo viaje desde Bangalore hasta Kuwait comienza  
El viaje largo desde Abu Dabi hasta Abisinia sigue  
El ligero viaje desde Augusta hasta Babilonia comienza  
El largo viaje desde Cali hasta Babilonia sigue

ELEGIR

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).

Seleccionar una de la estructuras resultantes para generar ejemplos

Buscar...

ejemplo	complemento1	complemento2
El viaje de los titulares desde Abisinia es muy cultural	animado humano cargo	lugar población país nombre propio
El viaje de la bebedora desde Kuwait resulta muy importante	animado humano condición humana negativa	lugar población ciudad nombre propio
El viaje de las enfermas desde Abisinia es también largo	animado humano condición humana negativa	lugar población país nombre propio
El viaje del personal militar desde San Marino parece muy importante	animado humano grupo o colectivo militar	lugar población ciudad nombre propio
El viaje de la caballería desde Afganistán es también turístico	animado humano grupo o colectivo militar	lugar población país nombre propio
El viaje de las fuerzas policiales desde Luxemburgo resulta muy largo	animado humano organización militar	lugar población ciudad nombre propio
El viaje de las policías secretas desde Abisinia resulta muy turístico	animado humano organización militar	lugar población país nombre propio
El viaje de las acomodadoras desde Kuwait es muy turístico	animado humano profesión general	lugar población ciudad nombre propio

☐ Filtrar con Word2Vec

información

Ajustar similitud entre sustantivos: -1 significado opuesto. 1 sinónimo.

Similitud actual: 0

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 Rotterdam 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](https://doi.org/10.1162/tacl_a_00051)

- Díaz Hormigo, M.<sup>a</sup> 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.<sup>a</sup> J. (2011). *Kontrastive Grammatik und Lexikographie: spanisch-deutsches Wörterbuch zur Valenz des Nomens*. Iudicium.
- Domínguez Vázquez, M.<sup>a</sup> 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.<sup>a</sup> 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.<sup>a</sup> 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.<sup>a</sup> J., Bardanca Outeiriño, D. & Simões, A. (2021). Automatic Lexicographic Content Creation: Automating Multilingual Resources Development for Lexicographers. En I. Kosem, M. Cukr, M. Jakubiček, 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](https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_16_pp269-287.pdf)
- Domínguez Vázquez, M.<sup>a</sup> 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.<sup>a</sup> 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.<sup>a</sup> 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. Jakubiček, 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](https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_4.pdf)
- Domínguez Vázquez, M.<sup>a</sup> J. & Valcárcel Riveiro, C. (2020). PORTLEX as a multilingual and cross-lingual online dictionary. En M.<sup>a</sup> 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.<sup>a</sup> 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.<sup>a</sup> 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.<sup>a</sup> J. (dir.), Valcárcel Riveiro, C., Bardanca Outeiriño, D., Calañas Continente, J. A., Catalá Torres, N., Martín Gascuña, 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.<sup>a</sup> J. (dir.), Valcárcel Riveiro, C., Bardanca Outeiriño, D., Calañas Continente, J. A., Catalá Torres, N., López Iglesias, N., Martín Gascuña, 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.<sup>a</sup> J. (dir.), Valcárcel Riveiro, C., Mirazo Balsa, M., Sanmarco Bande, M.<sup>a</sup> T., Simões, A. & Vale, M. J. (2018). *Portlex. Dicionario 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.<sup>a</sup> 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.<sup>a</sup> 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/>