Revista de Investigación Lingüística

Volumen 28 2025 Murcia (España) eISSN: 1989-4554 ISSN: 1139-1146

Departamento de Lengua Española y Lingüística General
FACULTAD DE LETRAS

Volumen 28 2025 Murcia (España) eISSN: 1989-4554 ISSN: 1139-1146

Revista de Investigación Lingüística (*RIL*)

Volumen 28 2025 Murcia (España) eISSN: 1989-4554 ISSN: 1139-1146

Revista de Investigación Lingüística (RIL)

La *Revista de Investigación Lingüística* es una publicación periódica científica dedicada al estudio de la lengua española y la lingüística general en todas sus variedades, desde cualquier modelo teórico o planteamiento metodológico oportunos para el enfoque que sugiera el autor. Los trabajos pueden adoptar una perspectiva tanto sincrónica como diacrónica. Además de acoger artículos de corte filológico tradicional, la revista pretende actuar como medio de difusión de los últimos enfoques teóricos y metodológicos desarrollados en los estudios de lengua española y lingüística general.

La línea editorial de *RIL* contempla muy diversos ámbitos de estudio: análisis del discurso, historia de la lengua, historia de la lingüística, lexicografía, lexicología, lingüística general, pragmática, semántica, sintaxis, sociolingüística, terminología, variedades del español. De acuerdo con los informes confidenciales de evaluadores externos, la revista decide sobre la publicación de los artículos recibidos, que deben ser originales inéditos.

Desde el año 2004 la *Revista de Investigación Lingüística* tiene una periodicidad anual. En la actualidad, cada número consta de una sección general, en la que se publican artículos sobre lengua española y lingüística general, y una sección dedicada a la recensión de libros. Además, la revista puede publicar monografías, cuyos artículos abordan un tema específico bajo la coordinación de uno o varios especialistas.

La *Revista de Investigación Lingüística* se publica desde 2020 exclusivamente en edición electrónica mediante el sistema OJS, disponible en la dirección https://revistas.um.es/ril, donde se ofrece el contenido de todos sus números en formato digital.

Dirección científica

Revista de Investigación Lingüística
Departamento de Lengua Española y Lingüística General
Facultad de Letras
Universidad de Murcia
C/ Santo Cristo, 1
30001 ~ Murcia

Dirección administrativa

Servicio de Publicaciones
Universidad de Murcia
Edificio Pleiades
Campus de Espinardo
30071 ~ Murcia

Indexación, bases de datos y catálogos

La *Revista de Investigación Lingüística* cuenta con el Sello FECYT desde 2021 (renovado en la convocatoria de 2025 para los próximos dos años), está indexada en ESCI (Clarivate) e incluida en el Catálogo Latindex 2.0 (36 de 38 características cumplidas). En Dialnet Métricas (2024) *RIL* se posiciona en el C1 de Filologías (48 de 312 revistas) y en el C2 de Lingüística (27 de 70 revistas). Según MIAR, en 2021 *RIL* tenía un ICDS de 10; en la actualidad presenta la siguiente difusión: c1+m6+e3+x6.

CARHUS Plus+ ~ Sistema de clasificación de revistas científicas de los ámbitos de Ciencias Sociales y Humanidades
CIRC ~ Clasificación Integrada de Revistas Científicas (de Ciencias Sociales y Humanas)
Dialnet ~ Portal de difusión de la producción científica hispana. Universidad de La Rioja
DOAJ ~ Directory of Open Access Journals
Dulcinea ~ Proyecto coordinado por el Ministerio de Educación y Ciencia para identificar y analizar las políticas editoriales de las revistas científicas españolas
ERIH Plus ~ Índice europeo de referencia para las disciplinas humanísticas y sociales
ESCI (Emerging Sources Citation Index) ~ Producto de la Web of Science (WoS) editado por Thomson Reuters
Fuente Académica Plus ~ Base de datos bibliográfica
Google Scholar ~ Buscador de Google especializado en documentos académicos con recuento de citas
IBZ On line ~ Bibliografía internacional de publicaciones periódicas de Humanidades y Ciencias Sociales
IDR (Índice Dialnet de Revistas) ~ Recurso que informa sobre el impacto científico de una revista, su evolución y su posición respecto del resto de las revistas de la especialidad. Universidad de La Rioja
ÍnDICEs-CSIC ~ Recurso bibliográfico multidisciplinar que recopila y difunde principalmente artículos de investigación publicados en revistas científicas españolas
Latindex ~ Sistema regional de información en línea para revistas científicas de América Latina, El Caribe, España y Portugal. Universidad Autónoma de México
LB (Linguistic Bibliography) ~ Catálogo en línea que abarca las diferentes disciplinas lingüísticas
LLBA (Linguistics & Language Behavior Abstracts) ~ Base de datos de revistas lingüísticas
MIAR (Matriz de Información para el Análisis de Revistas) ~ Base de datos que reúne información clave para la identificación y el análisis de revistas
OCLC WORLDCAT ~ Catálogo en línea que facilita el acceso a material bibliográfico
PIO (Periodicals Index Online) ~ Base de datos internacional de revistas de Artes, Humanidades y Ciencias Sociales
REDIB ~ Red Iberoamericana de Innovación y Conocimiento Científico
REGESTA IMPERII ~ Base de datos bibliográfica
SUDOC ~ Catálogo colectivo de referencias bibliográficas realizado por las bibliotecas y centros de documentación de educación superior e investigación franceses
ULRICH'S ~ Directorio de publicaciones periódicas
ZDB/EZB ~ Catálogo colectivo de revistas electrónicas

Derechos de autor

eISSN: 1989-4554
ISSN: 1139-1146
Depósito Legal: MU-646-1988
Dirección web *RIL*: https://doi.org/10.6018/ril
Archivo: https://revistas.um.es/ril/issue/archive
Envíos: https://revistas.um.es/ril/about/submissions

Volumen 28 2025 Murcia (España) eISSN: 1989-4554 ISSN: 1139-1146

ÍNDICE

Artículos

Reseñas

Revista de Investigación Lingüística, 28 (2025), pp. 9-10
eISSN: 1989-4554, ISSN: 1139-1146, https://doi.org/10.6018/ril

Revista de Investigación Lingüística, 28 (2025)
eISSN: 1989-4554, ISSN: 1139-1146
https://doi.org/10.6018/ril.662171
Universidad de Murcia

Diseño y metodología de un etiquetador semántico-ontológico multilingüe: ESMAS-ES+ *

Design and methodology of a multilingual semantic-ontological tagger: ESMAS-ES+

María José Domínguez Vázquez
Universidad de Santiago de Compostela
majo.dominguez@usc.es
ORCID ID: https://orcid.org/0000-0002-6060-9577

RECIBIDO: 10 de mayo de 2025
ACEPTADO: 24 de septiembre de 2025

RESUMEN: El etiquetador automático ESMAS-ES+ tiene como objetivo central la anotación semántico-ontológica de textos en español, francés, alemán y gallego. Junto con el estudio de la viabilidad de un nuevo método de análisis, el desarrollo del etiquetador requiere explorar nuevas vías para el procesamiento inteligente de la información y conocimiento, y, por ende, para la comprensión profunda del significado. Esta publicación presenta los principios metodológicos para su diseño, así como una panorámica de técnicas y estrategias aplicables para la generación de conocimiento lingüístico, multilingüe y tecnológico sostenible, lo que, a su vez, contribuirá al diseño de herramientas extrapolables a diferentes lenguas. La evolución de ESMAS-ES+ puede repercutir en algunas áreas del procesamiento del lenguaje natural, en especial, en aquellas ligadas a la comprensión y desambiguación del significado. De este modo, puede contribuir a favorecer la legibilidad y comprensión de datos lingüísticos por parte de máquinas.

PALABRAS CLAVE: ontologías, significado categorial, procesamiento del lenguaje natural, sostenibilidad.

ABSTRACT: The automatic tagger ESMAS-ES+ aims to annotate semantically and ontologically texts in Spanish, French, German and Galician. Besides examining the feasibility of a new method of analysis, the development of the tagger involves investigating new approaches to intelligent information and knowledge processing, and also to a deep comprehension of meaning. This paper outlines the methodological principles of the tagger's design and provides an overview of the techniques and strategies applicable for generating sustainable linguistic, multilingual and technological knowledge. These insights will support in turn the development of tools that are adaptable to various languages. The development of ESMAS-ES+ can have a positive impact on several areas of natural language processing, particularly those related to meaning comprehension and disambiguation. Consequently, it can enhance machine-driven readability and understanding of linguistic data.

KEYWORDS: ontologies, categorial meaning, natural language processing, sustainability.

* Proyecto PID2022-137170OB-I00, financiado por MICIU/AEI/10.13039/501100011033 y por FEDER/ UE.

1. INTRODUCCIÓN

Avanzar en una descripción y comprensión más profunda del significado resulta imprescindible, no solo para la generación de conocimiento lingüístico teórico y descriptivo, sino también para diversas aplicaciones de Procesamiento y Generación del Lenguaje Natural (PLN, GLN) y de Traducción Automática (TA). Es bien sabido que el procesamiento y modelado automatizado de la información semántica requiere fundamentos teórico-semánticos y recursos con datos lingüísticos (corpus, diccionarios, terminologías, etc.). Además, se necesitan analizadores y etiquetadores, reglas gramaticales y léxicos vinculados a ontologías estables, así como la aplicación de diferentes tecnologías de PLN.

En este contexto, el presente trabajo presenta la metodología propuesta para el desarrollo de un etiquetador semántico-ontológico multilingüe, ESMAS-ES^{+}. Dicho recurso anotará automáticamente el léxico nominal en español, francés, alemán y gallego[1]. A diferencia de otros recursos afines[2], ESMAS-ES^{+} detectará, procesará y etiquetará automáticamente textos a partir del vocabulario recogido en paquetes léxicos, siguiendo una ontología construida de forma *bottom-up*. Desde el punto de vista del usuario, será posible la consulta y recuperación conjunta de información lingüística atendiendo a parámetros ontológicos y relacionales (véanse ejemplos ilustrativos en la sección 3.1.).

El desarrollo del anotador, todavía en curso, se articula en dos ejes centrales: por un lado, la descripción ontológica del léxico basada en la Ontología 2.0 (Domínguez, 2025). Esta proporciona una caracterización detallada de los rasgos ontológicos atribuibles a las unidades léxicas nominales. Por otro lado, la vinculación sintáctico-semántica relacional, la cual ofrece una visión integrada del significado en contexto. Uno de los pilares de nuestra aproximación se fundamenta en el hecho de que los roles semánticos se expresan mediante lexemas con rasgos ontológico-categoriales específicos.

Aunque existen estudios y técnicas comparables, no conocemos ningún método que integre simultáneamente: (i) la sintaxis valencial y la semántica combinatoria de Engel (2009); (ii) las ontologías de WordNet; (iii) una ontología *bottom-up* de vocabulario prototipado; (iv) caudal léxico traducido automáticamente; (v) modelos del lenguaje (LLMs) y (vi) datos multilingües ya anotados con información sintáctica, semántica y ontológica. El proyecto, además, se fundamenta en principios de sostenibilidad[3], como la reutilización de datos y la interoperabilidad de recursos.

Este trabajo presenta, por tanto, una propuesta metodológica modular y sostenible para el desarrollo del etiquetador nominal. En la sección 2 se expone el contexto y los recursos relacionados con nuestra propuesta; la sección 3 presenta los fundamentos teóricos y metodológicos, incluyendo el modelo combinado y la arquitectura del etiquetador (sección 3.1), así como las herramientas aplicadas y los primeros resultados (sección 3.2). La Sección 4 describe el etiquetador previsto y el desarrollo potencial de herramientas fundamentadas en sus datos. La sección 5 recoge las conclusiones del estudio.

[1] La selección de lenguas permite contrastar una lengua germánica con lenguas románicas. Véase también la nota al pie número 2.

[2] Muchos recursos existentes (diccionarios, corpus y redes semánticas) presentan limitaciones para su explotación por procesadores lingüísticos: no están anotados semánticamente, no son de acceso abierto, carecen de interoperabilidad y suelen estar disponibles solo para lenguas mayoritarias, especialmente el inglés. Esta carencia es especialmente significativa en lenguas como el gallego. Esto contrasta con la creciente demanda de herramientas léxicas por parte de diferentes disciplinas.

[3] La sostenibilidad parte del empleo de datos procedentes de herramientas ya diseñadas por el equipo de investigación u otros equipos. Este enfoque permite aprovechar de forma óptima los avances previos y otorgar nuevo valor a esfuerzos realizados anteriormente. Para asegurar la continuidad y la reutilización del conocimiento generado, la interoperabilidad se revela como un principio clave.

Revista de Investigación Lingüística, 28 (2025), pp. 175-192
eISSN: 1989-4554, ISSN: 1139-1146, https://doi.org/10.6018/ril.662171

2. ESTADO DEL ARTE Y RECURSOS

A continuación, esbozaremos *per negationem* algunas de las características centrales de ESMAS-ES+ con la finalidad de contextualizar nuestro etiquetador (vid. sección 4) frente a recursos de su entorno más cercano:

(a) Los DICCIONARIOS son excelentes contenedores de información semántica (REDES, de Bosque, 2004; DICE), pero, hasta el momento, no han propiciado el desarrollo de una aplicación de etiquetado semántico automático, como tampoco ha sucedido desde la generación automática de contenido lexicográfico.

(b) La mayor parte de los CORPUS, RECURSOS LÉXICOS y GESTORES DE CORPUS permiten recuperar y medir estadísticamente las propiedades distributivas del vocabulario y codificar sus propiedades sintagmáticas. Sin embargo, dichas herramientas no suelen conjugar un filtrado simultáneo de datos semánticos, formales y estadísticos, dado que estas, por norma general, carecen de anotación semántica. No obstante, algunos recursos léxicos, como PDEV/CPA, Verbario, PropBank, NomBank, los corpus Ancora, OntoNotes 2.5. o SemLink sí ofrecen anotación semántica. Se dispone también de léxicos que describen solo determinadas clases de palabras atendiendo a su interfaz sintáctico-semántica, como *VerbNet*, por citar alguno.

(c) Entre las REDES SEMÁNTICAS y BASES DE DATOS LÉXICAS destacan WordNet y FrameNet, las cuales proporcionan datos semánticos de diferentes niveles: en el caso de WordNet, diferentes ontologías[4]; en el caso de FrameNet, también el significado relacional o roles semánticos[5]. Existen, además, diccionarios multilingües enciclopédicos (como BabelNet) y recursos colaborativos de código abierto (como Wikcionario).

(d) Los SOFTWARES DE ETIQUETADO analizados se centran mayoritariamente en etiquetar unidades léxicas o textuales desde un punto de vista formal o morfosintáctico (TreeTagger, FreeLing o XIADA) o atendiendo al análisis del discurso (CorefAnnotator). En el escenario actual solo conocemos dos aplicaciones que muestran concomitancias con nuestro proyecto:

- La aplicación UAM Corpus Tool. Ofrece anotación sintáctico-semántica; no obstante, el etiquetado en el plano semántico se ocupa únicamente del paradigma verbal haciendo hincapié en el aspecto o el modo verbal.
- El etiquetador semántico USAS (UCREL Semantic Analysis System), desarrollado inicialmente para el inglés (con problemas de desambiguación) y su categorización semántica, así como el etiquetador semántico en Python PyMusas.

(e) En las últimas décadas, las estrategias de aprendizaje automático basadas en REDES NEURONALES, el modelado estadístico del lenguaje (como los *n*-gramas) y otros enfoques lingüísticos neuronales (como el aprendizaje profundo) han revolucionado el campo del PLN (Peters *et al*., 2018; Li *et al*., 2021). Como resultado, contamos con métodos predictivos para el modelado y la representación del lenguaje, basados en *word embeddings* (re-

[4] Cabe señalar aquí la *Suggested Upper Merged Ontology* (Niles y Pease, 2001), la *Top Concept Ontology* (Álvez *et al*., 2008), los *WordNet Domains* (Bentivogli *et al*., 2004), el *Basic Level Concept* (Izquierdo *et al*., 2007), los *Epinónimos* (Gómez y Solla, 2018) y los *primitivos semánticos* de Miller *et al*. (1990).

[5] ESMAS-ES+ se diferencia de WordNet y FrameNet en varios aspectos clave, tanto en su enfoque como en su funcionalidad.

presentaciones vectoriales de formas léxicas generadas mediante técnicas de procesamiento de corpus)[6]. Estas técnicas permiten reproducir y codificar relaciones léxicas, y, por tanto, capturar el conocimiento semántico.

(f) La descripción del plano semántico exige también una organización y parametrización, por tanto, ONTOLOGÍAS y TAXONOMÍAS sólidas y granulares. Sobre ontologías y extracción automática de datos mediante PLN se viene trabajando de modo muy intenso en las últimas décadas. Así, por ejemplo, FunGramKB estudia la organización ontológica y el conocimiento léxico-conceptual y el grupo Tecling desarrolla taxonomías para la anotación automática de sustantivos y verbos. Entre el abanico de taxonomías y ontologías destacamos Kind, Louw & Nida Model, Semantic Domains, USAS y WordNet.[7]

(g) La irrupción de chatbots basados en inteligencia artificial (IA) —como ChatGPT, Copilot, Gemini o DeepSeek— ha transformado rápidamente los procesos de obtención y análisis de datos. Aunque, por el momento, no sustituyen la investigación sobre codificación semántica automática, sí la complementan. Como señala Trap-Jensen (2018: 34), el conocimiento humano sigue siendo esencial en este proceso, ya que la eficacia de estos sistemas varía según el tipo de tarea y la precisión del *prompt*:

> Language technology and artificial intelligence are moving into a phase where the word lists and morphological lexicons developed inside the NLP environment itself are insufficient to meet the demands for developing smarter and more sophisticated products. Automatic content summaries, domain classification and virtual assistants are but a few examples of applications that require 'knowledge' or some way of handling the semantics of human language.

3. FUNDAMENTOS TEÓRICOS Y METODOLÓGICOS

3.1. Modelo combinado y arquitectura del etiquetador

Con el objetivo de describir, codificar y procesar automáticamente la complejidad de los sistemas lingüísticos, particularmente la información semántica, se propone la aplicación de un método combinado multimodular (véase sección 1). A continuación, se ofrece una panorámica como punto de partida:

[6] El etiquetado de datos se beneficia de técnicas de PLN basadas en redes neuronales, como Word2vec (Mikolov *et al*., 2013). Emplea una RNN (*Recurrent Neural Network*) de dos niveles con dos implementaciones diferentes: i) el algoritmo *Skip-gram* intenta predecir el contexto más adecuado de una palabra analizando su vector y los vectores de las palabras vinculadas al vector de la palabra origen, ii) CBOW (*Continuous Bag of Words*) prevé la palabra más adecuada para un contexto específico, es decir, la palabra que más frecuentemente ocupa un espacio o aparece en concurrencia. Estos modelos permiten analizar tanto relaciones paradigmáticas como sintagmáticas. Así, siguiendo criterios cuantitativos (medición de cosenos y criterios de similitud) se pueden predecir las unidades léxicas más probables en un contexto lingüístico determinado. Esto facilita su análisis estadístico, la categorización y la elaboración de patrones de similitud semántica (cf. modelos preentrenados como Sketch Engine, Derekovecs o SemantiGal).

[7] Nuestra experiencia en proyectos de lexicografía electrónica multilingüe, GLN y combinatoria léxica nos ha permitido constatar que muchos recursos con información semántica suelen ser sistemas cerrados, poco escalables, no interoperables y dependientes de una clase de palabra o dominio específico, lo que limita su aplicabilidad a otros contextos. Aunque se han logrado avances en la anotación lingüística de corpus y recursos léxicos para lenguas como el español (cf. esta sección), el etiquetado semántico en francés o alemán aún está por detrás del inglés. La situación del gallego es incluso más precaria.

Revista de Investigación Lingüística, 28 (2025), pp. 175-192
eISSN: 1989-4554, ISSN: 1139-1146, https://doi.org/10.6018/ril.662171

a) La teoría valencial

Los enfoques semánticos aplicados en PLN (grafos conceptuales, redes semánticas, etc.) tienen en común la interpretación de las palabras individuales como predicados, con sus consiguientes argumentos. Estos, a su vez, pueden actuar nuevamente de predicados. De entre los numerosos enfoques lingüísticos, destacamos aquí la teoría de dependencias y valencias de la escuela de Ulrich Engel (1996, 2009). Esta describe la valencia sintáctica (diferenciando argumentos obligatorios o facultativos frente a circunstantes) y la valencia semántica (el significado combinatorio, conformado por el plano relacional y categorial). Engel (1988: 875) describe este último como sigue: «Kombinatorische Bedeutung, die für die Umgebung eines Wortes gilt (Gegensatz: inhärente Bedeutung); soviel wie Inhaltsvalenz des betreffenden Wortes; es sind kategorielle und relationale Bedeutung zu unterscheiden»[8].

Siguiendo esta teoría, se pueden analizar patrones argumentales y su potencial semántico combinatorio —los roles semánticos y rasgos ontológicos de los elementos implicados en una expresión— (Engel, 1996). Dicha aproximación permite establecer prototipos léxicos y clases semánticas actualizables en diferentes casillas funcionales. Las investigaciones realizadas en el desarrollo de recursos como Xera, XeraWord, Combinatoria y CombiContext[9] han confirmado la viabilidad de la teoría de valencias de Engel para describir y procesar el potencial combinatorio nominal, tanto en el eje sintagmático y paradigmático como en el oracional. Por tanto, la propuesta cuenta con un sólido fundamento semántico, el cual puede verse enriquecido con los modelos de Hanks (patrones semánticos; véase PDEV/CPA) o Mel'čuk (2013; funciones léxicas)[10]. Pese a las diferencias teóricas entre las aproximaciones citadas, estas posibilitan un modelado formal de la combinatoria léxica y la estructura nominal/oracional, como ejemplifican diferentes diccionarios y recursos basados en dichas teorías.

b) Los rasgos ontológicos y paquetes léxicos

El punto nuclear de nuestro método lo conforman los denominados *paquetes léxicos*. Actualmente disponemos de 3600 unidades. Estas resultan de la recurrencia de las APIs[11] a las relaciones semánticas de WordNet y a las ontologías vinculadas a los *synsets* en el modelo de EuroWordNet. Estos paquetes incluyen un identificador único, una descripción del tipo de objeto caracterizado, su clasificación combinada en las ontologías de WordNet y una lista de lexemas flexionados, etiquetados y caracterizados ontológicamente. Cada lexema está vinculado al Índice Interlingüístico (ILI), utilizado tanto en WordNet como en el *Multilingual Central Repository* (MCR[12]). Por tanto, estos paquetes compilan vocabulario prototipado ontológicamente y están asignados a una clase semántica concreta.

Para el etiquetado semántico manejamos la Ontología 2.0. (Domínguez, 2025). Esta aúna el inventario de rasgos categoriales de la teoría valencial (Engel, 2009) y las ontologías de WordNet. Un ejemplo del etiquetado ontológico granular a partir de los datos de los paquetes léxicos se recoge en la imagen 1:

[8] Traducción de la cita: Significado combinatorio, que se aplica al entorno de una palabra (en contraste con el significado inherente); equivale a la valencia de contenido de la palabra en cuestión. Se debe distinguir entre significado categorial y significado relacional.

[9] Su descripción se encuentra en la página http://portlex.usc.gal/combinatoria/.

[10] Mel'čuk (2013) representa un claro ejemplo de la posible formalización del significado para su posterior tratamiento informático.

[11] Las herramientas se encuentran en abierto en este enlace: http://portlex.usc.gal/combinatoria/.

[12] Proporciona coherencia ontológica a todos los Wordnets y elementos allí integrados. Además, el MCR 3.0 también proporciona grupos de sinónimos, definiciones y relaciones semánticas (hipónimos, hiperónimos).

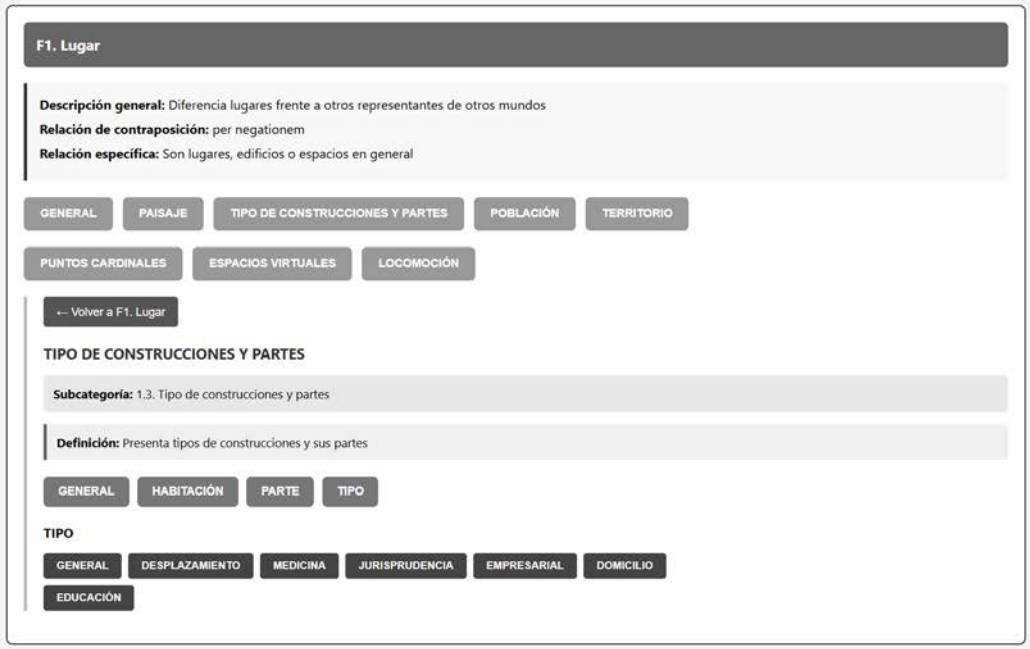

Imagen 1. Descripción ontológica. Ejemplo {*Lugar* | *Tipo de construcciones y partes* | *Tipos*}[13]

Dada la relación entre rasgos ontológico-categoriales y roles semánticos (imagen 2), esta descripción ontológica resulta esencial desde una perspectiva lingüística. Además, es clave para el procesamiento automático de datos, al servir de vínculo directo con las ontologías WordNet, a las que recurrimos en el diseño de nuestro etiquetador.

Una vez definido el aparato descriptivo y el inventario ontológico, es imprescindible disponer de un corpus para la evaluación del grado de acierto o desacierto en la anotación automática. Para tal fin, hemos seleccionado textos del repositorio OPUS (*Open parallel corpora*), en concreto, del subcorpus TED2020, el cual ofrece transcripciones de conferencias TED en varios idiomas[14]. La compilación textual se asienta, además, en criterios ligados a la representatividad del vocabulario, la cohesión de los textos y la distribución de las clases ontológicas[15]. Actualmente, el corpus *Gold Standard* está compuesto por una selección de ocho textos que suman un total de 19 182 *tokens* con una cobertura nominal de un 18,35 % sobre el total de *tokens*. Aunque el tamaño del corpus es reducido para entrenar y validar un etiquetador automático, en esta fase inicial se utiliza para verificar la viabilidad del método propuesto. En este contexto, resulta fundamental que el conjunto de datos sea manejable para su anotación manual, sin que ello impida su ampliación progresiva en futuras etapas del proyecto.

[13] Para el diseño de algunas de las imágenes se han utilizado herramientas de inteligencia artificial.

[14] Esta selección no supuso un trabajo menor teniendo en cuenta, por ejemplo, que la lengua gallega estaba menos representada (o no representada) en los diferentes subcorpus de OPUS.

[15] Los textos son coherentes y cohesionados desde un punto de vista lingüístico. No están vinculados a ningún dominio específico y muestran una representación equilibrada de las distintas categorías ontológicas. Estos textos incluyen unidades léxicas que requieren desambiguación, por ejemplo, RATÓN como 'animal' frente a RATÓN como 'dispositivo electrónico'. A su vez, cuentan con vocabulario polisémico interpretable en función del contexto, como es el caso de PERIÓDICO: puede referirse tanto a un objeto con contenido como a un lugar de trabajo. Proporcionan, además, expresiones complejas y compuestos, especialmente relevantes en lenguas como el alemán.

Revista de Investigación Lingüística, 28 (2025), pp. 175-192
eISSN: 1989-4554, ISSN: 1139-1146, https://doi.org/10.6018/ril.662171

c) COMBINATORIA Y CONTEXTO

Cualquier estudio sobre el significado del vocabulario tiene que contemplar el análisis de la polisemia y la actualización del significado en contexto, por tanto, prever estrategias de desambiguación (Arias-Arias, 2025; Raganato *et al.*, 2017; OntoNotes 2.5; Renau *et al.*, 2019).

Independientemente de lo que los autores entiendan por contexto, dado que se trata de un concepto multidimensional (Domínguez y Gouws, 2023), es un hecho que muchas investigaciones sobre el contexto tienen como objetivo último describir, representar o codificar significados o comprender el comportamiento de las lenguas. No solo desde la lingüística, sino también desde la ciencia cognitiva (McDonald y Ramscar, 2001) y los enfoques neurobiológicos (Pereira *et al.*, 2018) se destaca la importancia de los diferentes tipos de contextos. En esta línea, la concurrencia contextual es considerada uno de los factores más informativos en el estudio del significado. Veamos un ejemplo de contexto sintáctico-semántico: como bien es sabido, una frase preposicional en español puede expresar diferentes roles semánticos dotados de diferentes rasgos categoriales. Esto depende, por una parte, de la unidad regente y, por otra, de la presencia de diferentes argumentos y sus interacciones. La imagen 2 muestra para el sustantivo DEBATE combinaciones para diferentes roles[16]:

Ejemplos	
el debate {del tema} RoI2	el debate {del equipo de investigación} RoI1
el debate {del tema} RoI2 {con el equipo de investigación} RoI3	el debate {del equipo de investigación} RoI1 {con los participantes} RoI3
el debate {del tema} RoI2 {entre la delegación y los disidentes} RoI1	el debate {del tema} RoI2 {entre los participantes} RoI1

Leyenda: RoI1 : 'aquel que debate' (agente) | RoI2 : 'aquello que es debatido' (tema/objeto) | RoI3 : 'aquel con el que se debate' (co-agente)

Los elementos entre llaves {} indican los complementos que realizan cada rol semántico.

Imagen 2. Distribución de roles. Ejemplo DEBATE

En el análisis del contexto y de los esquemas argumentales también cabe contemplar la presencia o ausencia de diferentes complementos o circunstantes. Ya Harris (1954: 146) proponía: «Each language can be described in terms of a distributional structure, i.e., in terms of the occurrence of parts relative to other parts.». Dicho análisis también puede contribuir a la desambiguación de significados, como se observa en los siguientes ejemplos de Sketch Engine (deTenTen), donde el vocablo PRÄSENZ adopta diferentes significados según el contexto:

(1) Seine **Präsenz** ist überwältigend, und nur wenige können seinem Charisma widerstehen.
Traducción: Su presencia es abrumadora y pocos pueden resistirse a su carisma.
Acepción: Presencia = 'aspecto', 'carisma'

[16] Este tipo de datos son especialmente relevantes en todos aquellos diccionarios o recursos que tienen la finalidad de describir el potencial combinatorio de las lenguas, como, por ejemplo, los diccionarios de valencias.

(2) HELVETIAROCKT engagiert sich [...] für eine stärkere **Präsenz** der Musikerinnen auf der Bühne, in den Medien und in der Öffentlichkeit.
Traducción: HELVETIAROCKT se compromete [...] a una mayor presencia de las músicas en el escenario, en los medios y en el espacio público.
Acepción: 'Presencia = 'estar en un lugar', 'visibilidad física o mediática'

Para tareas de desambiguación en contextos nominales, y siguiendo un principio de retroalimentación entre datos y recursos, podemos manejar la herramienta CombiContext (Domínguez *et al*., 2021). Esta herramienta incluye un conjunto de 81 155 estructuras verbales, 29 726 adjetivos y un número similar de adverbios. Las estructuras argumentales están anotadas con información ontológica y valencial. Así, por ejemplo, el sustantivo COLOR aparece en nuestro recurso vinculado a 768 estructuras verbales. La imagen 3 presenta un patrón o estructura argumental, en este caso con COLOR como núcleo argumental de la frase nominal:

Imagen 3. Patrón valencial en CombiContext. Ejemplo COLOR

Como se desprende de la imagen 3, a partir de los datos compilados en CombiContext es posible realizar tareas de desambiguación. Retomemos el ejemplo de RATÓN: como se muestra en la imagen 4, tanto la estructura nominal-verbal como la anotación ontológica indican que la única interpretación posible de RATÓN en este contexto es la de {animado | animal}. Por tanto, mediante este procedimiento se puede distinguir la lectura de 'roedor' frente a la de 'dispositivo informático'.

Imagen 4. Patrón valencial en CombiContext. Ejemplo RATÓN

Revista de Investigación Lingüística, 28 (2025), pp. 175-192
eISSN: 1989-4554, ISSN: 1139-1146, https://doi.org/10.6018/ril.662171

Una visión de conjunto del método la proporciona la imagen 5:

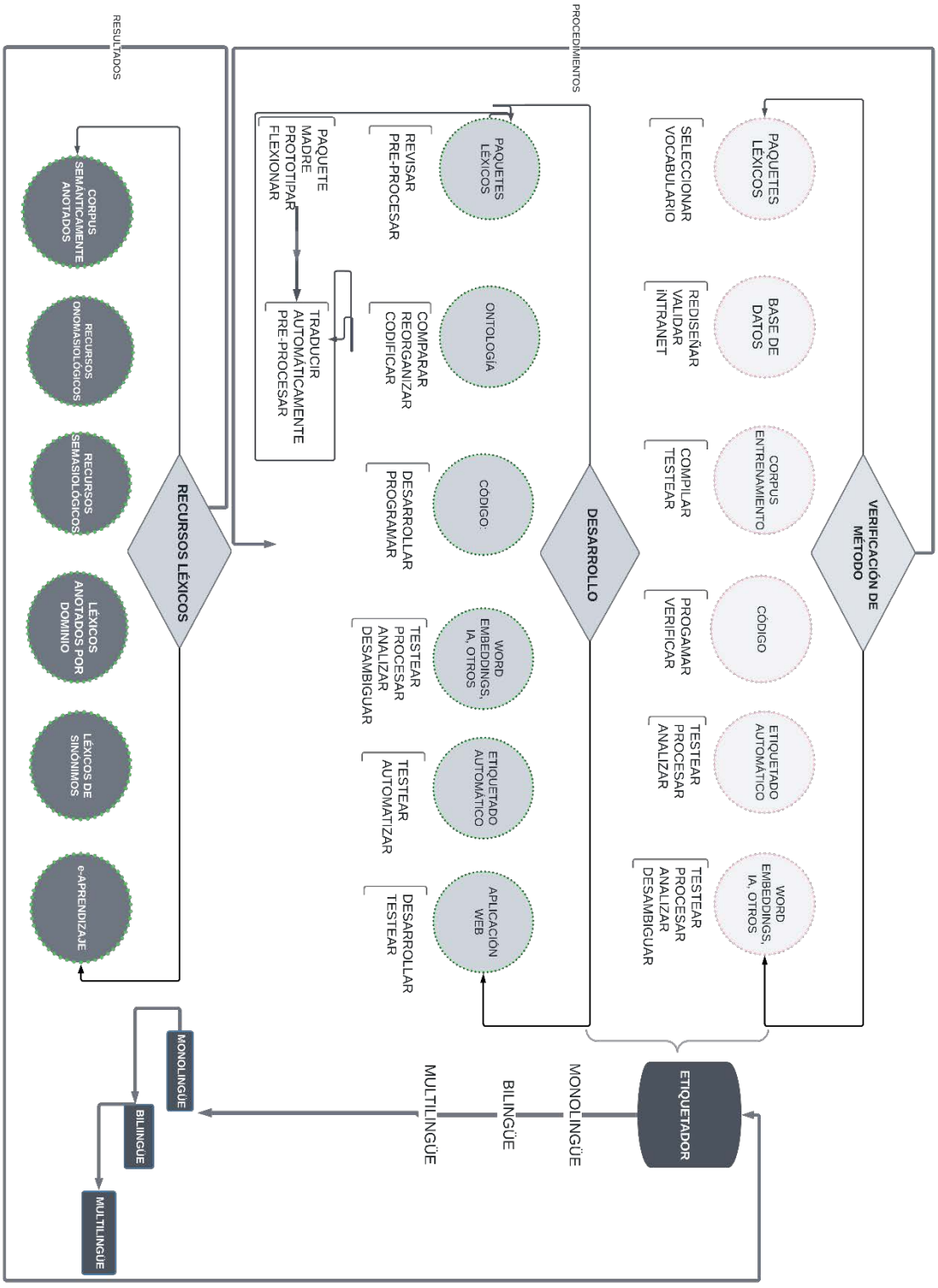

Imagen 5. Fases de trabajo, procedimientos y aplicaciones

Para la corrección del *output* se aplicarán diferentes modelos: corrección directa en la base de datos mediante la comparación multilingüe y/o elaboración de plantillas de corrección siguiendo modelos como el DQF-MQM, manejada en la revisión de datos procedentes de la traducción automática.

Por tanto, tenemos el punto de partida para poder desarrollar un nuevo tipo de herramienta de etiquetado semántico automático, esto es, nuestro etiquetador ESMAS-ES^{+}. En fases posteriores, el procedimiento descriptivo de rasgos ontológicos y patrones argumentales puede perfeccionarse tanto en favor de la desambiguación como de la ampliación de datos. En este contexto, se contemplan varias líneas de trabajo: i) el uso de Transformers como BERT, ii) la integración de recursos con anotaciones semánticas (como Kind) y/o sintáctico-semánticas (como OntoNotes 5.0; cfr. Weischedel *et al.*, 2022) o iii) el entrenamiento de LLMs[17], entre otros.

3.2. Recursos aplicados y primeros resultados

En esta sección, abordamos aquellos fenómenos que han dado lugar a inconsistencias en el proceso de anotación actual. Se proponen, además, diferentes opciones de optimización:

(a) Divergencias en la anotación humana: Se han observado discrepancias en la anotación del corpus *Gold Standard* realizada por los especialistas (*Inter-Annotator Agreement*, IAA). Las principales divergencias radican tanto en la complejidad descriptiva inherente a la realidad lingüística (por ejemplo, la dificultad para interpretar ciertas unidades incluso en un contexto determinado), como en la ausencia de algunas subclases en la Ontología léxica (Domínguez *et al*., 2021). En este sentido, se han establecido las siguientes líneas de actuación:

- Granularidad de la ontología: Dado que se trata de una ontologíao *bottom-up* —esto es, se construye a partir de la realidad lingüística y no a la inversa—, puede enriquecerse con nuevas subclases ontológicas. Tomando en consideración tanto los resultados del IAA como otros estudios preliminares (véase b)), la Ontología léxica (Domínguez *et al*., 2021; compárese Martín Gascueña, 2023) ha sido revisada y ampliada, dando lugar a la Ontología 2.0 (Domínguez, 2025). Esto conlleva, por tanto, un aumento de rutas ontológicas y de paquetes léxicos[18].
- Revisión de paquetes léxicos para cada una de las lenguas: Dado que los paquetes léxicos sirven para testar y entrenar el prototipo de etiquetador habrá que comprobar que el caudal léxico esté *de facto* adecuadamente compilado en un paquete léxico. Esta revisión posibilita dos vías de trabajo: permite optimizar el recurso y, a su vez, sienta las bases para el diseño de herramientas con aplicación en la enseñanza y aprendizaje de lenguas, como, por ejemplo, recursos para determinar la dificultad de aprendizaje de un texto o su adecuación para un cierto nivel de lengua.

[17] En este último caso, utilizando textos del corpus *Gold Standard, prompts* específicos y la Ontología 2.0. (Domínguez, 2025) como punto de partida.

[18] Con el fin de conseguir una mayor granularidad también es posible recurrir a otras ontologías o taxonomías, como USAS, la propia TOP de WordNet, *Kind,* o proceder como Martinelli *et al*. (2024).

Revista de Investigación Lingüística, 28 (2025), pp. 175-192
eISSN: 1989-4554, ISSN: 1139-1146, https://doi.org/10.6018/ril.662171

- Edición de paquetes madre: La elaboración de nuevos paquetes léxicos se fundamenta en el concepto de 'lengua madre' propuesto por Gouws (2014). Para su extracción, tratamiento y creación manejamos recursos descritos previamente, tales como WordNet, APIs, etc.; para la lematización, análisis de dependencias y etiquetado empleamos como soporte léxico diversas librerías de Python y para la flexión los diccionarios de Freeling.
- Traducción automática: Mediante la traducción de los paquetes léxicos se automatizan procedimientos y reutiliza la información. Una herramienta útil en este procedimiento es TraduWord. Dicho recurso posibilita la traducción automática de repertorios de clases semánticas, así como de sus ejemplares en relación paradigmática a partir de los datos extraídos automáticamente de WordNet, por tanto, a partir de los paquetes léxicos existentes en otras lenguas[19]. También es posible recurrir a la traducción automática de paquetes léxicos mediante otros recursos.

(b) Divergencias en la anotación automática: Actualmente se están llevando a cabo pruebas exploratorias para determinar si algún LLM específico resulta adecuado para los objetivos de nuestra investigación. Aunque pueda parecer una afirmación evidente, para poder anotar ontológicamente una unidad léxica es imprescindible identificarla previamente. En este sentido, nuestros estudios piloto (Domínguez, 2025) indican que Copilot y DeepSeek han demostrado ser más eficaces que ChatGPT y Gemini en el reconocimiento de sustantivos. En cuanto a la anotación semántica, también se ha observado que Copilot y DeepSeek son los más adecuados para tareas de anotación ontológica, tanto por su precisión como por su alineación con la jerarquía taxonómica. Gemini y ChatGPT, en cambio, presentan problemas de consistencia en la clasificación de datos. No obstante, estos resultados preliminares deben contrastarse con otros estudios que ofrecen perspectivas complementarias. Así, en una investigación posterior (Arias-Arias y Martín-Cancela, 2025) se evaluó la viabilidad de la anotación semántica mediante LLMs (ChatGPT, Gemini y Claude) utilizando técnicas de *zero-shot prompting*. Los resultados obtenidos muestran que Claude supera a ChatGPT y Gemini en todas las métricas evaluadas[20]. Esto sugiere que los LLMs no deberían excluirse de tareas de anotación ontológica (véase también Puraivan *et al.*, 2024). Estos estudios apuntan, además, a la posibilidad de desarrollar un entrenamiento específico orientado a explorar relaciones ontológicas en contexto. En general, conviene reflexionar sobre el uso de los LLMs como herramientas iniciales de anotación o como recursos de verificación en etapas finales.

(c) Junto con las optimizaciones señaladas previamente**,** el desarrollo del etiquetador conlleva una serie de tareas asociadas:

[19] Las traducciones automáticas en *TraduWord* se calculan utilizando dos enfoques diferentes: (a) dados los índices interlingüísticos presentes en los paquetes léxicos, que identifican unívocamente un concepto, se consulta la API y a partir de aquí se obtiene una lista de variantes para cada unidad presente en el paquete madre; (b) de no haber resultados disponibles procedentes de WordNet, o si la cantidad de variantes es demasiado restringida en comparación con el número de variantes en el paquete madre, se realiza una petición a la API de *MyMemory* para recoger su traducción. Se revisa el *output* de dichas herramientas y, de este modo, se obtiene el vocabulario de los paquetes léxicos.

[20] En términos de precisión, Claude alcanza un valor de 0.65 frente a ChatGPT con 0.0543 y Gemini con 0.487. En cuanto al *recall*, los resultados son 0.471 para Claude, 0.205 para ChatGPT y 0.444 para Gemini. Finalmente, el F1 *Score* se sitúa en 0.511 para Claude, 0.0862 para ChatGPT y 0.452 para Gemini. Estos datos evidencian una mayor consistencia y fiabilidad de Claude en la identificación y anotación de sustantivos.

- Es necesario continuar con el desarrollo y optimización de herramientas para el *matching* de los paquetes léxicos y los textos. Usando como punto de partida los resultados de los primeros experimentos realizados, se prevé el desarrollo de un motor capaz de procesar y anotar datos de manera automática.
- Optimización de herramientas, técnicas y estrategias: El análisis cualitativo y cuantitativo del etiquetado automático conducirá, en algunos casos, a la necesaria optimización de herramientas existentes, al desarrollo de nuevos recursos o al reajuste de la aproximación metodológica. Dichos análisis serán incorporados como nuevas capas de procesamiento dentro del motor de etiquetado original de los datos y vendrán a enriquecer y reajustar la aproximación metodológica inicial.

En los apartados anteriores, se han señalado los principales recursos, casi todos ellos de acceso abierto y gratuito, que se manejarán en el desarrollo del etiquetador. Los siguientes apartados sirven para su compilación:

(1) GENERADORES MULTILINGÜES[21] del lenguaje natural: Estos recursos permiten la consulta (semi)automática de las ontologías de WordNet o la extracción y reutilización de esquemas sintáctico-semánticos de la frase nominal y oración.

(2) WORDNET: Atendiendo a la estructura multilingüe de EuroWordNet[22] y MCR 3.0, es posible llevar a cabo una interconexión entre las lenguas implicadas en el análisis, así como el análisis de diferentes anotaciones semánticas y la extracción de vocabulario ligado a dichas clases semánticas.

(3) MULTITOOLS[23]: Bajo este epígrafe se compendia una colección de recursos de libre acceso, que dada su adecuación al proyecto se pueden reutilizar, lo que además contribuye a la sostenibilidad:
   - APIs: Permiten extraer datos léxicos de las consultas que recurren a las relaciones semánticas de WordNet y a las ontologías vinculadas a los *synsets* en el modelo de EuroWordNet.
   - Motores de búsqueda en WordNet: 1) Lematiza utiliza código derivado de diferentes proyectos del Seminario de Lingüística Informática de la Universidad de Vigo. Enlaza con la interfaz de Galnet para ilustrar la identificación del significado de formas léxicas y proporciona el lema de cada argumento con sus variantes de significado (*synsets*) ligadas a las diferentes ontologías de WordNet; 2) Combina permite combinar o cotejar los resultados de varias consultas de las APIs a WordNet, por tanto, la extracción semiautomática de datos compartidos o combinados de las ontologías de WordNet.
   - FLEXIONA (español, francés y alemán) y FLEXIONADOR (gallego y portugués): Aportan lemas flexionados.

(4) TREETAGGER y LINGUAKIT: Para la lematización y PoS-tagging del alemán, español y francés se maneja el TreeTagger. Con la misma finalidad, pero para el gallego, se usa Linguakit.

[21] Todos ellos accesibles en http://portlex.usc.gal/combinatoria/.

[22] En EuroWordNet, los wordnets están interconectados con enlaces interlingüísticos almacenados en el índice interlingual (Vossen, 1998).

[23] Estas herramientas están accesibles bajo dicho epígrafe en el siguiente enlace: http://portlex.usc.gal/combinatoria/.

Revista de Investigación Lingüística, 28 (2025), pp. 175-192
eISSN: 1989-4554, ISSN: 1139-1146, https://doi.org/10.6018/ril.662171

(5) ONTOLOGÍAS *bottom-up*.

(6) TRADUWORD: Herramienta de traducción de caudal léxico o a partir de datos extraídos de modo automático de WordNet en combinación con MyMemory.

(7) PAQUETES LÉXICOS MULTILINGÜES: 3600 paquetes léxicos, esto es, caudal léxico etiquetado y anotado semánticamente.

(8) BASE DE DATOS (mySQL) para el tratamiento y almacenamiento de datos.

(9) Corpus *Gold Standard* para tareas de validación.

(10) Herramientas de combinatoria aplicables a tareas de desambiguación.

(11) LLMs como recursos de anotación o verificación.

4. ETIQUETADOR Y NUEVOS RECURSOS LÉXICOS

El etiquetador se visualizará como una ventana de rastreo de texto para detectar, procesar y anotar automáticamente textos a partir del vocabulario recogido en los paquetes léxicos. El filtrado y etiquetado hace uso de las relaciones jerárquicas de la ontología, que podrán ampliarse para mejorar la granularidad.

En su fase final, junto con apartados como la descripción del recurso, el aparato teórico-aplicado, la ontología y el equipo, el etiquetador será una aplicación activa y supervisada: permitirá introducir un texto, seleccionar la lengua y obtenerlo anotado semánticamente. Contará con una interfaz interactiva con funciones avanzadas, como filtros por clase semántica y visualización de descripciones ontológicas mediante *mouse over*. El recurso será de acceso abierto y código libre, con posibilidad de descarga en varios formatos para facilitar su integración en otros sistemas. De este modo, los datos generados podrán integrarse en recursos externos, como diccionarios computacionales, bases terminológicas o sistemas de traducción automática, y servirán para crear plantillas semánticas en diccionarios multilingües (semasiológicos y onomasiológicos), léxicos anotados por dominio o herramientas de e-aprendizaje. Entre los posibles recursos desarrollados a partir del etiquetador se perfilan: i) una herramienta de sinónimos en contexto para evitar la pobreza léxica; ii) un clasificador automático de textos por tipología (p. ej., TFG vs. informe); iii) un sistema de evaluación de textos por nivel de lengua, útil en enseñanza de lenguas extranjeras y iv) un módulo de detección de errores combinatorios frecuentes, basado en patrones valenciales anotados. Esto tiene una clara aplicación en la enseñanza de lenguas extranjeras y segundas lenguas, por ejemplo, contribuyendo a evitar errores habituales ligados al potencial combinatorio de una unidad léxica (Müller-Spitzer *et al*., 2018). Estas propuestas resultan especialmente relevantes para lenguas con menos recursos, contribuyendo a su fortalecimiento digital y lingüístico.

5. CONCLUSIÓN

Este trabajo presenta un método combinado y sostenible para el desarrollo de un etiquetador semántico automático y multilingüe de textos.

Gracias a su arquitectura modular, el anotador puede adaptarse a distintos niveles de granularidad descriptiva y a diversas lenguas, incluidas aquellas con menos recursos digitales. Los avances en la automatización de la obtención eficaz de información sintáctica, léxica y combinatoria contribuirán tanto a la optimización de recursos existentes como al diseño de nuevos instrumentos lingüísticos. Para ello, será necesario continuar con la ampliación del corpus *Gold Standard*, incorporar nuevas subclases semánticas en la ontología, mantener una metodología circular entre datos y recursos y evaluar el entrenamiento de LLMs para tareas de desambiguación.

En definitiva, ESMAS-ES^{+} aportará datos anotados, técnicas de procesamiento, pruebas exploratorias y metodologías replicables. La sostenibilidad de la investigación se garantiza mediante la reutilización de datos, el acceso abierto a los recursos generados y su posible integración en aplicaciones externas. En el marco de la lexicografía y la investigación lingüística sostenible, esta propuesta contribuye a revalorizar esfuerzos previos y a generar valor social y científico.

Declaración de contribución de autoría

María José Domínguez Vázquez: conceptualización, curación de datos, obtención de fondos, investigación, metodología, administración del proyecto, supervisión, redacción - borrador original, redacción - revisión y edición.

Bibliografía[24]

Álvez, Javier, Jordi Atserias, Jordi Carrera, Salvador Climent, Egoitz Laparra, Antoni Oliver y German Rigau (2008): «Complete and Consistent Annotation of WordNet using the Top Concept Ontology», en Nicoletta Calzolari *et al.* (eds.), *Proceedings of the 6th Language Resources and Evaluation Conference (LREC'08)*, Marrakech, Morocco, pp. 1529-1534.

Arias-Arias, Iván (en prensa): «Nuevas vías para la desambiguación en frases nominales en alemán: fundamentos metodológico-lingüísticos para el desarrollo de una herramienta de anotación semántica (semi)automática», *Círculo de Lingüística Aplicada a la Comunicación*, 104.

Arias-Arias, Iván y Elena Martín-Cancela (en prensa): «Bridging Human and AI Perspectives: Semantic Annotation of Generic Nouns in German», *Proceedings of the eLex 2025 conference*.

Bentivogli, Luisa, Pamela Forner, Bernardo Magnini y Emanuele Pianta (2004): «Revising WordNet Domains Hierarchy: semantics, coverage and balancing», en Gilles Sérasset *et al.* (eds.), *Proceedings of Workshop on Multilingual Linguistic Resources*, Stroudsburg, Association for Computational Linguistics, pp. 101-108. En línea: <https://dl.acm.org/doi/10.5555/1706238.1706254>.

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

[24] Último acceso a todos los recursos electrónicos: 24/9/2025.

Revista de Investigación Lingüística, 28 (2025), pp. 175-192
eISSN: 1989-4554, ISSN: 1139-1146, https://doi.org/10.6018/ril.662171

DOMÍNGUEZ VÁZQUEZ, María José (2025): *Ontología 2.0. ESMAS-ES+*, Santiago de Compostela. En línea: <https://grupoportlex.github.io/ontologia/>.

DOMÍNGUEZ VÁZQUEZ, María José y Rufus H. GOUWS (2023): «The Definition, Presentation and Automatic Generation of Contextual Data in Lexicography», *International Journal of Lexicography*, 36(3), pp. 233-259. DOI: https://doi.org/10.1093/ijl/ecac020

DOMÍNGUEZ VÁZQUEZ, María José, Carlos VALCÁRCEL RIVEIRO y Daniel BARDANCA OUTEIRIÑO (2021): *Ontología léxica*, Santiago de Compostela. En línea: <http://portlex.usc.gal/ontologia/>.

DOMÍNGUEZ VÁZQUEZ, María José (dir.), Carlos VALCÁRCEL RIVEIRO, Daniel BARDANCA OUTEIRIÑO, José Antonio CALAÑAS CONTINENTE, Natalia CATALÁ TORRES, Rosa MARTÍN GASCUEÑA, Mónica MIRAZO BALSA, María Teresa SANMARCO BANDE y Laura PINO SERRANO (2021): *CombiContext. Prototipo online para la generación automática de contextos frasales y oraciones de la frase nominal en alemán, español y francés*, Santiago de Compostela. En línea: <http://portlex.usc.gal/combinatoria/verbal>.

ENGEL, Ulrich (1988): *Deutsche Grammatik,* Heidelberg, Julius Gross Verlag.

ENGEL, Ulrich (1996): «Semantische Relatoren. Ein Entwurf für künftige Valenzwörterbücher», en Nico Weber (ed.), *Semantik, Lexikographie und Computeranwendung*, Tubinga, Niemeyer, pp. 223-236. DOI: https://doi.org/10.1515/9783111555522.223

ENGEL, Ulrich (2009): *Syntax der deutschen Gegenwartssprache*, 4.ª ed., Berlín, Schmidt.

GÓMEZ GUINOVART, Xavier y Miguel SOLLA PORTELA (2018): «Construyendo el WordNet gallego: métodos y aplicaciones», *Recursos y evaluación de idiomas*, 52(1), pp. 317-339.

GOUWS, Rufus (2014): «Towards bilingual dictionaries with Afrikaans and German as language pair», en María José Domínguez Vázquez *et al.* (eds.), *Zweisprachige Lexicographie zwischen Translation und Didaktik*, Berlín, De Gruyter, pp. 249-262. DOI: https://doi.org/10.1515/9783110366631.249

HARRIS, Zellig (1954): «Distributional Structure», *Word*, 10(2-3), pp. 146-162. DOI: https://doi.org/10.1080/00437956.1954.11659520

IZQUIERDO, Rubén, Armando SUÁREZ y German RIGAU (2007): «Exploring the automatic selection of basic level concepts», en Ruslan Mitkov, Galia Angelova, y Kalina Bontcheva (eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Shoumen, INCOMA, pp. 298-302. En línea: <https://adimen.si.ehu.es/~rigau/publications/ranlp07-isr.pdf>.

LI, Belinda, Maxwell NYE y Jacob ANDREAS (2021): «Implicit Representations of Meaning in Neural Language Models», en Chengqing Zong *et al.* (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing,* vol. 1, Online: Association for Computational Linguistics, pp. 1813-1827. DOI: https://doi.org/10.18653/v1/2021.acl-long.143

MARTÍN GASCUEÑA, Rosa (2023): «Diseño de una ontología de semántica léxica para los proyectos MultiGenera y MultiComb», *RILEX. Revista Sobre Investigaciones léxicas*, 6(3), pp. 77-106. DOI: https://doi.org/10.17561/rilex.6.3.8083

MARTINELLI, Giuliano, Francesco Maria MOLFESE, Simone TEDESCHI, Alberte FERNÁNDEZ-CASTRO y Roberto NAVIGLI (2024): «CNER: Concept and Named Entity Recognition», en Kevin Duh, Helena Gomez y Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, México D.F., Association for Computational Linguistics, pp. 8336-8351. DOI: https://doi.org/10.18653/v1/2024.naacl-long.461

MCDONALD, Scott y Michael RAMSCAR (2001): «Testing the distributional hypothesis: The influence of context on judgments of semantic similarity», en Johanna Moore y Keith Stenning, (eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society,* Londres, LEA, pp. 611-616.

MEL'ČUK, Igor (2013): *Semantics. From meaning to text*, Ámsterdam/Filadelfia, John Benjamins.

MIKOLOV, Tomas, Kai CHEN, Greg CORRADO y Jeffrey DEAN (2013): «Efficient Estimation of Word Representations in Vector Space», en Yoshua Bengio y Yann Lecun (eds.), *Proceeding of the International Conference on Learning Representations Workshop Track*, Arizona, Conference Track Proceedings, pp. 1-12. DOI: https://doi.org/10.48550/arXiv.1301.3781

MILLER, George A., Richard BECKWITH, Christiane FELLBAUM, Derek GROSS y Katherine J. MILLER (1990): «Introduction to WordNet: An On-line Lexical Database», *International Journal of Lexicography*, 3, pp. 235-244. DOI: https://doi.org/10.1093/ijl/3.4.235

MÜLLER-SPITZER, Carolin, Martina Nied CURCIO, María José DOMÍNGUEZ VÁZQUEZ, Idalete Maria SILVA DIAS y Sascha WOLFER (2018): «Recherchepraxis bei der Verbesserung von Interferenzfehlern aus dem Italienischen, Portugiesischen und Spanischen: Eine explorative Beobachtungsstudie mit DaF-Lernenden», *Lexicographica*, 34(1), pp. 157-182. DOI: https://doi.org/10.1515/lex-2018-340108

NILES, Ian y Adam PEASE (2001): «Towards a Standard Upper Ontology», en Nicola Guarino, Barry Smith y Christopher Welty (eds.), *2nd International Conference on Formal Ontology in Information Systems (FOIS-2001),* Ogunquit, Main, ACM, pp. 17-19. DOI: https://doi.org/10.1145/505168.505170

PEREIRA, Francisco, Bin LOU, Brianna PRITCHETT, Samuel RITTER, Samuel J. GERSHMAN, Nancy KANWISHER, Matthew BOTVINICK y Evelina FEDORENKO (2018): «Toward a universal decoder of linguistic meaning from brain activation», *Nature communications,* 9, pp. 1-13. DOI: https://doi.org/10.1038/s41467-018-03068-4

PETERS, Matthew, Mark NEUMANN, Mohit IYYER, Matt GARDNER, Christopher CLARK, Kenton LEE y Luke ZETTLEMOYER (2018): «Deep Contextualized Word Representations», en Marilyn Walker (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1*, New Orleans, Louisiana, Association for Computational Linguistics, pp. 2227-2237. DOI: https://doi.org/10.18653/v1/N18-1202

PURAIVAN, Eduardo, Irene RENAU y Nicolás RIQUELME (2024): «Metaphor Identification and Interpretation in Corpora with ChatGPT», *SN Computer Science*, 5, art. n.º 976 (2024). DOI: https://doi.org/10.1007/s42979-024-03331-0

RAGANATO, Alessandro, Jose CAMACHO-COLLADOS y Roberto NAVIGLI (2017): «Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison», en Mirella Lapata, Phil Blunsom y Alexander Koller (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Association for Computational Linguistics, pp. 99-110. DOI: https://doi.org/10.18653/v1/E17-1010

RENAU, Irene, Rogelio NAZAR, Ana CASTRO, Benjamín LÓPEZ y Javier OBREQUE (2019): «Verbo y contexto de uso: Un análisis basado en corpus con métodos cualitativos y cuantitativos», *Revista Signos*, 52(101), pp. 878-901. DOI: http://dx.doi.org/10.4067/S0718-09342019000300878

TRAP-JENSEN, Lars (2018): «Lexicography beHtween NLP and Linguistics: Aspects of Theory and Practice», en Jaka Čibej *et al*. (eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, Ljubljana, Ljubljana University Press, pp. 25-37.

VOSSEN, Piek (1998): «EuroWordNet: Building a Multilingual Database with wordnets for European Languages», *ELRA Newsletter*, 3(1), pp. 7-10.

WEISCHEDEL, Ralph, Martha PALMER, Mitchell MARCUS, Hovy EDUARD, Sameer PRADHAN, Lance RAMSHAW, Nianwen XUE, Ann TAYLOR, Jeff KAUFMAN, Michelle FRANCHINI, Mohammed El-BACHOUTI, Robert BELVIN y Ann HOUSTON (2022): *OntoNotes Release 5.0* (Version V1), Borealis. DOI: https://doi.org/10.5683/SP2/KPKFPI

Recursos electrónicos[25]

AnCora = http://clic.ub.edu/corpus/es/ancora
BabelNet = https://babelnet.org/
ChatGPT = https://chat.chatbotapp.ai/
Combina = http://portlex.usc.gal/develop/combina.php
Copilot = https://www.microsoft.com/es/microsoft-copilot/organizations
CorefAnnotator = https://github.com/nilsreiter/CorefAnnotator
DeepSeek = https://chat.deepseek.com/
Derekovecs = https://corpora.ids-mannheim.de/openlab/derekovecs/
DICE = http://www.dicesp.com/paginas/index/2
DQF-MQM = https://www.taus.net/resources/blog/dqf-mqm-beyond-automatic-mt-quality-metrics
EuroWordNet = https://archive.illc.uva.nl/EuroWordNet/
EuroWordNet Top-Ontologie = https://archive.illc.uva.nl/EuroWordNet/corebcs/ewnTopOntology.html#_Toc419884299
Flexiona = http://portlex.usc.gal/develop/flexiona.php
Flexionador = https://ilg.usc.gal/flexionador
FrameNet = https://framenet.icsi.berkeley.edu/fndrupal/
FreeLing's dictionaries = http://nlp.lsi.upc.edu/freeling/node/1
FunGramKB = https://fungramkb.ucam.edu/
Gemini = https://gemini.google.com
GermaNet = https://uni-tuebingen.de/en/142806
Kind = http://www.tecling.com/kind
Lematiza = http://portlex.usc.gal/develop/lematiza/
Linguakit = https://linguakit.com/es/analisis-completo
Louw & Nide Model = https://ucrel.lancs.ac.uk/usas/Louw&Nida/Louw&Nida_frameset.htm
MyMemory = https://mymemory.translated.net/
Multilingual central repository = https://adimen.si.ehu.es/web/MCR
Multitools = http://portlex.usc.gal/combinatoria/
NomBank = https://nlp.cs.nyu.edu/meyers/NomBank.html
Odgen = http://ogden.basic-english.org/bewords.html
OntoNotes 5.0 = https://catalog.ldc.upenn.edu/LDC2013T19
OPUS = https://opus.nlpl.eu/
PDEV/ CPA = https://pdev.org.uk/

[25] Último acceso a todos los recursos electrónicos: 24/9/2025.

PropBank = http://verbs.colorado.edu/~mpalmer/projects/ace.html
PyMusas = https://pypi.org/project/pymusas/
Semantic Domains = https://semdom.org
SemantiGal = https://tec.citius.usc.es/demos-lingua/index
SemLink = https://verbs.colorado.edu/semlink/
SenSem = http://grial.edu.es/sensem/corpus/main
Sketch Engine = https://www.sketchengine.eu
Tecling = https://www.tecling.com/
TraduWord = https://ilg.usc.gal/gl/proxectos/interoperabilidade-de-recursos-e-producion-automatica-de-linguaxe-natural
TreeTagger = https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
UAM Corpus Tool = www.corpustool.com/index.html
USAS = http://ucrel-api.lancaster.ac.uk
Verbario = http://www.tecling.com/verbario/
VerbNet = https://verbs.colorado.edu/~mpalmer/projects/verbnet.html
Wikcionario = https://es.wikipedia.org/wiki/Wikcionario
WordNet = https://wordnet.princeton.edu
Xera = http://portlex.usc.gal/combinatoria/usuario
XeraWord = http://ilg.usc.es/xeraword/en/
XIADA = http://corpus.cirp.gal/xiada

Revista de Investigación Lingüística, 28 (2025), pp. 175-192
eISSN: 1989-4554, ISSN: 1139-1146, https://doi.org/10.6018/ril.662171