

Palabras que conectan saberes

Semántica aplicada
e interdisciplinariedad



Iván Arias-Arias
Elena Martín-Cancela
(eds.)



EDITORIAL COMARES

Iván Arias-Arias
Elena Martín-Cancela
(eds.)

Palabras que
conectan saberes
Semántica aplicada
e interdisciplinariedad

Granada, 2026

Colección indexada en la MLA International Bibliography desde 2005

EDITORIAL COMARES

INTERLINGUA

431

Colección fundada por:
EMILIO ORTEGA ARJONILLA y PEDRO SAN GINÉS AGUILAR

Comité Científico (Asesor):

ESPERANZA ALARCÓN NAVÍO Universidad de Granada	CATALINA JIMÉNEZ HURTADO Universidad de Granada
JESÚS BAIGORRI JALÓN Universidad de Salamanca	ÁNGELA LARREA ESPIRAL Universidad de Córdoba
CHRISTIAN BALLIU ISTI, Bruxelles	HELENA LOZANO Università di Trieste
LORENZO BLINI LUSPIO, Roma	MARIA JOAO MARÇALO Universidade de Évora
ANABEL BORJA ALBÍ Universitat Jaume I de Castellón	FRANCISCO MATTE BON UNINT, Roma
NICOLÁS A. CAMPOS PLAZA Universidad de Murcia	JAVIER MARTÍN PÁRRAGA Universidad de Córdoba
MIGUEL Á. CANDEL-MORA Universidad Politécnica de Valencia	ANTONIO RAIGÓN RODRÍGUEZ Universidad de Córdoba
ÁNGELA COLLADOS AÍS Universidad de Granada	CHELO VARGAS-SIERRA Universidad de Alicante
MIGUEL DURO MORENO Universidad de Málaga	MERCEDES VELLA RAMÍREZ Universidad de Córdoba
FRANCISCO J. GARCÍA MARCOS Universidad de Almería	ÁFRICA VIDAL CLARAMONTE Universidad de Salamanca
GLORIA GUERRERO RAMOS Universidad de Málaga	GERD WOTJAK Universidad de Leipzig

ENVÍO DE PROPUESTAS DE PUBLICACIÓN:

Las propuestas de publicación han de ser remitidas (en archivo adjunto, con formato PDF) a alguna de las siguientes direcciones electrónicas: anabelen.martinez@uco.es, psgines@ugr.es

Antes de aceptar una obra para su publicación en la colección INTERLINGUA, ésta habrá de ser sometida a una revisión anónima por pares. Para llevarla a cabo se contará, inicialmente, con los miembros del comité científico asesor. En casos justificados, se acudirán a otros especialistas de reconocido prestigio en la materia objeto de consideración.

Los autores conocerán el resultado de la evaluación previa en un plazo no superior a 60 días. Una vez aceptada la obra para su publicación en INTERLINGUA (o integradas las modificaciones que se hiciesen constar en el resultado de la evaluación), habrán de dirigirse a la Editorial Comares para iniciar el proceso de edición.

Financiación: Proyecto Etiquetador Semántico Multilingüe, Automático y Sostenible (ESMAS-ES*) (PID2022-137170OB-100), financiado por MICIU/AEI/10.13039/501100011033 y por FEDER/UE.



© Los autores

© Editorial Comares, 2026

Polígono Juncaril • C/ Baza, parcela 208 • 18220 Albolote (Granada) España • Tlf.: 958 465 382

<https://www.comares.com> • E-mail: libreriacomares@comares.com

<https://www.facebook.com/Comares> • <https://twitter.com/comareseditor>

<https://www.instagram.com/editorialcomares>

ISBN: 979-13-7033-144-3 • Depósito legal: Gr. 988/2026

Impresión y encuadernación: COMARES

Sumario

Introducción. ¿PARA QUÉ QUEREMOS UN ANOTADOR SEMÁNTICO EN LA ERA DE LA INTELIGENCIA ARTIFICIAL?.....	1
<i>Iván Arias-Arias</i>	
<i>Elena Martín-Cancela</i>	
1. UN CAMBIO DE PARADIGMA.....	1
2. UN ANOTADOR SEMÁNTICO EN LA ERA DE LA IA: ¿PARA QUÉ?.....	4
2.1. Fiabilidad y verificabilidad de los datos	4
2.2. Transparencia metodológica	4
2.3. Equidad multilingüe	5
2.4. Mitigación de sesgos.....	6
2.5. Sostenibilidad ambiental y responsabilidad ecológica.....	6
2.6. Reutilización responsable de datos	7
2.7. Democratización real	7
2.8. Reflexión final	8
3. ÉSTE VOLUMEN: PROPUESTAS Y NUEVAS PERSPECTIVAS	8
4. REFERENCIAS BIBLIOGRÁFICAS	10
Capítulo I.—¿QUÉ ES UNA ONTOLOGÍA Y POR QUÉ ES IMPORTANTE PARA LA ANOTACIÓN SEMÁNTICA?.....	13
<i>María José Domínguez Vázquez</i>	
1. INTRODUCCIÓN.....	13
2. ONTOLOGÍA Y RED DE CONCEPTOS AFINES... O NO TANTO	13
2.1. ¿De qué estamos hablando?	13
2.2. Estructuras prototípicas de eventos y <i>scripts</i> : lo conceptual	14
2.3. Campo semántico y familia léxica	15
2.4. Taxonomías	16
2.5. Redes semánticas.....	17
3. ONTOLOGÍAS	19
3.1. Introducción	19
3.2. Ontologías computacionales.....	20
3.3. Ontologías en lingüística	23
4. ONTOLOGÍAS Y SU IMPORTANCIA PARA LA ANOTACIÓN SEMÁNTICA.....	26
5. CONCLUSIÓN	29
6. DECLARACIÓN DE USO DE IA GENERATIVA	30
7. REFERENCIAS BIBLIOGRÁFICAS	30

Capítulo II.—VIAJE AL CENTRO DEL ANOTADOR ESMAS-ES+: ¿CÓMO SE SUSTENTA UN ANOTADOR SEMÁNTICO DESDE EL PUNTO DE VISTA TECNOLÓGICO? ...	33
<i>Daniel Bardanca Outeiriño</i>	
1. INTRODUCCIÓN.	33
2. ANOTACIÓN SEMÁNTICA DESDE EL PUNTO DE VISTA COMPUTACIONAL.	34
2.1. Anotación, almacenaje y revisión de los datos.	34
2.2. Interactuando con los datos.	36
3. DISEÑO DE LOS DATASETS, MODELOS Y FINE-TUNING.	37
3.1. Escogiendo un modelo.	38
3.2. Datasets y plantilla de entrenamiento.	39
4. INSTRUCCIÓN DEL MODELO.	41
4.1. Gold standards y evaluación.	42
5. IMPLEMENTACIÓN DEL MODELO EN UN SERVIDOR.	43
5.1. Desplegar el modelo con una API.	43
5.2. La interfaz del usuario.	43
6. CONCLUSIÓN.	44
7. REFERENCIAS BIBLIOGRÁFICAS.	45
Capítulo III.—¿PARA QUÉ SIRVE ETIQUETAR PALABRAS SI YA TENEMOS DEEPL?	47
<i>Lucía Esperanza del Castillo Vázquez</i>	
<i>Sofía García González</i>	
1. INTRODUCCIÓN.	47
2. EVOLUCIÓN DE LA TRADUCCIÓN AUTOMÁTICA.	49
2.1. Traducción basada en reglas.	49
2.2. Traducción basada en ejemplos.	51
2.3. Traducción estadística.	52
2.4. Traducción automática neuronal.	52
2.5. Los grandes modelos de lenguaje (LLM).	54
3. LA ANOTACIÓN SEMÁNTICA EN EL PLN.	55
3.1. Desambiguación léxica.	56
3.2. Etiquetado de roles semánticos.	57
3.3. Anotación semántica categorial u ontológica.	58
3.4. Anotación semántica para el reconocimiento de entidades.	59
3.5. Representación abstracta del significado.	60
4. ANÁLISIS CRÍTICO.	61
5. REFERENCIAS BIBLIOGRÁFICAS.	62
Capítulo IV.—¿PODEMOS APRENDER Y ENSEÑAR UNA LENGUA CON ETIQUETAS SEMÁNTICAS?	67
<i>Carlos Valcárcel Riveiro</i>	
1. INTRODUCCIÓN.	67
2. BASES TEÓRICAS PARA ENSEÑAR UNA L2 CON ETIQUETADO SEMÁNTICO.	68
2.1. El lexicon mental y las redes semánticas.	68
2.2. El papel de la semántica en el aprendizaje gramatical.	69
2.3. Lingüística de corpus y anotación lingüística.	70

SUMARIO

3.	ONTOLOGÍAS, TAXONOMÍAS Y ETIQUETAS	72
3.1.	Ontologías semánticas	72
3.2.	Taxonomías semánticas	72
3.3.	Etiquetado de roles semánticos	73
3.4.	Las etiquetas semánticas	73
4.	EL USO DE ETIQUETAS SEMÁNTICAS EN EL AULA DE L2	74
4.1.	Comprensión escrita	74
4.2.	Comprensión oral	75
4.3.	Producción escrita	76
4.4.	Producción oral	77
4.5.	Etiquetas y agrupamiento semántico	78
5.	CONCLUSIONES	79
6.	REFERENCIAS BIBLIOGRÁFICAS	80
Capítulo V.—A VUELTAS CON LOS NÚMEROS: ¿CÓMO NOS PUEDE AYUDAR LA ESTADÍSTICA EN LA ANOTACIÓN SEMÁNTICA?		85
<i>Vítor Míguez-Rego</i>		
<i>Jose Ameijeiras-Alonso</i>		
1.	INTRODUCCIÓN	85
2.	CONCEPTOS BÁSICOS: ENTENDIENDO EL ACUERDO	87
3.	MIDIENDO EL ACUERDO ENTRE ANOTADORES HUMANOS	89
4.	MIDIENDO EL RENDIMIENTO DE LA ANOTACIÓN AUTOMÁTICA	94
5.	ELIGIENDO LA MÉTRICA ADECUADA SEGÚN EL TIPO DE TAREA	98
6.	CONCLUSIONES	100
7.	REFERENCIAS BIBLIOGRÁFICAS	101
Capítulo VI.—PALABRAS, DATOS Y DICCIONARIOS. ¿POR QUÉ (NO) VAMOS A DEJAR DE USAR DICCIONARIOS?		103
<i>María Concepción Maldonado González</i>		
1.	ALGUNAS CUESTIONES PREVIAS	103
2.	VARIAS PREGUNTAS ACTUALIZADAS	105
2.1.	¿Qué es un diccionario?	105
2.2.	Diccionario físico y diccionario mental	107
2.3.	El uso del diccionario en el aprendizaje de segundas lenguas	107
2.4.	El uso del diccionario en la lengua materna	108
3.	ALGUNAS REFLEXIONES A MODO DE RESPUESTAS	109
4.	REFERENCIAS BIBLIOGRÁFICAS	110
Capítulo VII.—HACIA UNA RED DINÁMICA DE RECURSOS SEMÁNTICOS: SOSTENIBILIDAD TECNOLÓGICA E INTEROPERABILIDAD		111
<i>María José Domínguez Vázquez</i>		
<i>Iván Arias-Arias</i>		
1.	INTRODUCCIÓN	111
2.	¿QUÉ ENTENDEMOS POR «SOSTENIBILIDAD TECNOLÓGICA»?	112
3.	RED DINÁMICA DE RECURSOS	113

4.	VISUALIZACIÓN Y NAVEGACIÓN EN EL MACRO-RECURSO	118
4.1.	Estructura de acceso y navegación facetada.	118
4.2.	Microestructura dinámica y jerarquización de la información.	120
4.3.	Interactividad y enriquecimiento ontológico	121
4.4.	Combinatoria sintagmático-ontológica e interoperabilidad externa	123
4.5.	Otros elementos ligados a la visualización.	124
5.	CONSIDERACIONES FINALES	125
6.	REFERENCIAS BIBLIOGRÁFICAS	126
Capítulo VIII.—TECNOLOGÍA LINGÜÍSTICA PARA UNA SOCIEDAD MULTILINGÜE: APLICACIONES, RETOS Y BUENAS PRÁCTICAS		129
<i>María Teresa Sanmarco Bande</i>		
1.	QUÉ ES LA TECNOLOGÍA LINGÜÍSTICA Y QUÉ DEMANDA LA SOCIEDAD	129
1.1.	De propuestas heterogéneas a infraestructuras normalizadas (2020-2025)	130
1.2.	La TL como infraestructura en capas e interoperable	132
2.	LOS GRANDES MODELOS DE LENGUAJE (LLM): CARACTERÍSTICAS Y FUNCIONAMIENTO	133
2.1.	El contraste semántico: lo explícito vs. lo implícito	133
2.2.	El impacto de los LLM en las humanidades digitales y en la lexicografía	135
3.	APLICACIONES ESPECÍFICAS DE LA TL: DEL PANORAMA GENERAL A UN CASO EJEMPLAR	137
3.1.	ESMAS-ES ⁺ : de la teoría a la práctica en tecnologías del lenguaje	137
4.	INTEGRAR LLM EN LA INVESTIGACIÓN LINGÜÍSTICA: RETOS Y ESTRATEGIAS	138
4.1.	Construyendo puentes: hacia una integración responsable	139
4.2.	Una arquitectura coherente, no un catálogo de buenas intenciones	140
5.	CONCLUSIÓN. HACIA UN DIÁLOGO INTERDISCIPLINAR IMPRESCINDIBLE.	141
6.	DECLARACIÓN DE USO DE IA GENERATIVA	141
7.	REFERENCIAS BIBLIOGRÁFICAS	142

Introducción

¿Para qué queremos un anotador semántico en la era de la inteligencia artificial?¹

Iván Arias-Arias

<https://orcid.org/0000-0003-2673-0899>

Instituto da Lingua Galega, Universidade de Santiago de Compostela

Elena Martín-Cancela

<https://orcid.org/0000-0002-9391-6418>

Universidade da Coruña

1. UN CAMBIO DE PARADIGMA

La presunta democratización de los grandes modelos de lenguaje (LLM por sus siglas en inglés) ha transformado radicalmente el panorama de la investigación lingüística, generando un intenso debate sobre la capacidad de adaptación de las disciplinas tradicionales a los desafíos de la era digital. Con esta transformación ha surgido, además, un temor palpable entre algunos especialistas respecto a la utilidad y supervivencia misma de áreas históricamente consolidadas como la lexicografía, hasta el punto de que algunas voces han proclamado su «fin» (de Schryver & Joffe, 2023).

Esta renovación se refleja en numerosos artículos científicos publicados en esta línea (entre otros: Lew, 2023; Rundell, 2023; Lew *et al.*, 2024; Arias-Arias *et al.*, 2024; Fuentes Morán *et al.*, 2024), así como en congresos internacionales que articulan sus temáticas centrales en torno a la relación entre lexicografía e inteligencia artificial. Así, AFRILEX 2025 presenta como hilo conductor *The Role of Artificial Intelligence and Digitalisation in the Future of Lexicography*, eLex 2025 aborda el tema *Intelligent Lexicography* y EURALEX 2026 se centra en un asunto bastante similar, *Lexicography in the Age of AI*. Estos espacios de debate evidencian un interés creciente por potenciar una mayor interdisciplinariedad y colaboración entre dos disciplinas cuya vinculación se ha hecho patente ya durante los últimos años: la lexicografía (o, más generalmente, la lingüística) y la inteligencia artificial (o, de forma más integral, la informática).

¹ Esta investigación se enmarca en el proyecto ESMAS-ES⁺ (PID2022-137170OB-I00), financiado por MICIU/AEI/10.13039/501100011033 y por FEDER/UE. Además, Iván Arias-Arias es beneficiario de una ayuda Formación de Profesorado Universitario (FPU21/00188) concedida por el Ministerio de Ciencia, Innovación y Universidades del Estado español.

Sin embargo, lejos de constituir una amenaza terminal, debemos entender este contexto como una oportunidad histórica para la revitalización y la reinención de las disciplinas lingüística y lexicográfica. Este cambio de paradigma no anuncia el ocaso de estas especialidades, sino que exige una profunda readaptación capaz de conducir la práctica investigadora y lexicográfica hacia un nuevo estadio de relevancia social y aplicabilidad. Klosa-Kückelhaus y Tiberius (2024, p. 9) proponen, en este sentido, una reconceptualización del proceso lexicográfico considerando que algunas ramas de la lingüística (fundamentalmente, la semántica, la lexicología y la lexicografía) deben orientarse hacia la creación de nuevas herramientas entendidas primordialmente como bases de datos estructuradas, más que como diccionarios tradicionales. Estos datos, adecuadamente procesados, podrán servir de fundamento para diferentes aplicaciones lingüísticas destinadas al desarrollo de modelos de lengua y recursos de amplio alcance.

En el caso de la lexicografía, el futuro podría residir, como sugieren algunos expertos (cf. Tarp & Gouws, 2023), en herramientas «invisibles» e integradas de manera natural en nuestro ecosistema digital. Para construir este ecosistema integrado resulta imprescindible el trabajo colaborativo entre instituciones, investigadores y desarrolladores. La integración de información lingüística y lexicográfica de forma sutil ya es una realidad actualmente: el acceso a la definición de cualquier palabra en un editor de texto está a tan solo un clic del ratón, y muchas plataformas digitales incorporan consultas léxicas sin que el usuario sea plenamente consciente de ello. En esta línea, las bases de datos de proyectos lexicográficos previos deben reutilizarse estratégicamente en la reinención de la disciplina, evitando duplicaciones innecesarias y aprovechando inversiones previas en conocimiento estructurado (cf. Colman, 2016).

No obstante, conviene reivindicar con firmeza el valor documental, histórico y cultural de las herramientas lingüísticas fundamentadas en criterios rigurosos. Este valor es precisamente lo que los modelos de lengua generativos, por su propia naturaleza estadístico-probabilística, no pueden garantizar de manera sistemática (cf. Arias-Arias *et al.*, 2025). Las herramientas lexicográficas tradicionales ofrecen trazabilidad, verificabilidad y responsabilidad sobre el conocimiento que vehiculan, cualidades que resultan esenciales para la investigación científica y la confianza de los usuarios.

En este contexto, la figura del lingüista adquiere mayor relevancia que nunca. La construcción de herramientas de inteligencia artificial fiables y éticamente responsables depende de datos de entrenamiento de alta calidad, cuya supervisión y diseño no pueden seguir el criterio exclusivo de algoritmos, sino que han de apoyarse en criterios humanos. Son especialmente esclarecedoras las palabras de Maldonado González a este respecto (2023, p. 33):

De nosotros depende que el desarrollo de herramientas lingüísticas de IA [...] esté en manos de equipos interdisciplinarios que cuenten con buenos lingüistas entre sus filas, o que quede al arbitrio de algoritmos y desarrollos técnicos que desestimen qué es el lenguaje lo que nos constituye como seres humanos.

Precisamente por estas consideraciones, este nuevo paradigma exige una redefinición clara de propósitos y funciones. Recuperando el concepto de «propósito genuino» de Wiegand (1998), cabe preguntarse por el fin último de estas tecnologías y de las herramientas lexicográficas. Un modelo de lenguaje generativo pretende crear lengua a través de procesos estadísticos complejos, su objetivo es puramente productivo, sin que medien una capacidad cognitiva real ni una comprensión genuina de los contenidos que ofrece. Por el contrario, una herramienta lexicográfica, como puede ser un sistema de anotación semántica, persigue un fin científico definido: facilitar el trabajo de investigadores, profesores, traductores o lexicógrafos, entre otros profesionales de la lengua, y servir como base para un conocimiento más estructurado y verificable. La diferencia fundamental radica en que una inteligencia artificial generativa, desprovista de sensibilidad y rigor humanos, opera en el plano de la generación de lengua *ad libitum*, mientras que un anotador semántico constituye un instrumento cognitivo diseñado para mejorar la comprensión y representación del conocimiento lingüístico.

Así, además, se hace necesario abogar decididamente por una red integrada de elementos y sistemas con financiación pública que construya infraestructuras como el proyecto ELEXIS (Tiberius *et al.*, 2023), capaces de perseguir objetivos como el que se enuncia a continuación: «Lexicography will encounter the shift towards open access structured data enabling re-use and linking of dictionary data along with aggregating stand-alone lexicographic (and terminological) resources into numerous dictionary portals».

Para responder a estas exigencias, emerge la necesidad urgente de repensar las disciplinas lingüísticas desde una perspectiva transversal e integradora. El plan de acción de la Comisión Europea para la Década Digital 2030² subraya esta visión estratégica al potenciar el desarrollo de tecnologías que permitan entrenar modelos fiables a partir de corpus elaborados con criterios éticos y una perspectiva centrada en lo humano. Las tesis de Villa Vigoni (cf. Schierholz, 2021) ya abogaban por una apertura interdisciplinar de la lexicografía mediante colaboraciones efectivas entre empresas tecnológicas, centros de investigación e instituciones públicas, un modelo que se revela cada vez más imprescindible.

Estas cuestiones nucleares serán debatidas a lo largo del presente volumen, cuyos contenidos se sintetizan ya en este capítulo introductorio. El objetivo fundamental es explorar los avances y debates surgidos en el seno del proyecto de investigación *Etiquetador Semántico Multilingüe Automático y Sostenible* (ESMAS-ES⁺), partiendo de una premisa ineludible: la necesidad de desarrollar herramientas semánticas sostenibles, automatizadas y multilingües que ofrezcan una alternativa robusta, verificable y fundamentada al uso indiscriminado de chatbots y modelos generativos. Argumentamos que un anotador semántico multilingüe no solo contribuye al desarrollo de un nuevo

² <https://eur-lex.europa.eu/ES/legal-content/summary/2030-digital-decade-policy-programme.html>

paradigma de ciencia sostenible y socialmente responsable (Calañas Continente & Domínguez Vázquez, 2023), sino que también puede enriquecer significativamente la web semántica estructurada y fomenta el desarrollo de una inteligencia artificial más transparente y sujeta a principios éticos verificables.

2. UN ANOTADOR SEMÁNTICO EN LA ERA DE LA IA: ¿PARA QUÉ?

Una vez establecida la necesidad de repensar las disciplinas lingüísticas en el contexto de la revolución tecnológica actual, resulta pertinente preguntarse por la justificación específica de desarrollar un anotador semántico automático. La proliferación de grandes modelos de lenguaje plantea interrogantes fundamentales que merecen respuesta: ¿por qué invertir recursos en herramientas lingüísticas especializadas cuando existen chatbots aparentemente capaces de responder a cualquier consulta y satisfacer las necesidades potenciales de los usuarios? La respuesta exige un examen crítico y riguroso de las limitaciones estructurales de los LLM, así como el reconocimiento del valor irremplazable de herramientas desarrolladas conforme a criterios científicos sólidos. Existen razones de peso (de naturaleza técnica, ética, científica y social) que justifican plenamente esta apuesta por herramientas lingüísticas sostenibles y multilingües como alternativa fundamentada a los sistemas generativos comerciales.

2.1. Fiabilidad y verificabilidad de los datos

La fiabilidad de los datos constituye una preocupación central en la construcción de cualquier recurso científico, especialmente ante las alucinaciones bien documentadas de los chatbots actuales (cf. Rundell, 2023; Trap-Jensen, 2024). Mientras que los sistemas de inteligencia generativa pueden producir respuestas aparentemente coherentes, pero factualmente incorrectas, un anotador semántico desarrollado por especialistas garantiza información contrastada, trazable, replicable y verificable (Chomsky *et al.*, 2023). Esta diferencia adquiere relevancia crucial cuando consideramos la centralidad del usuario: quien consulta un recurso lexicográfico (especialmente si no posee formación especializada) espera información fiable y validada académicamente, no pseudo-información generada algorítmicamente sin posibilidad de verificación directa.

La dificultad de confirmar afirmaciones extraídas de LLM radica precisamente en la imposibilidad de contrastar sus fuentes originales o validar sus procesos inferenciales. Para los usuarios no es posible verificar críticamente sus hipótesis iniciales ante un sistema que produce «alucinaciones de confianza»: formulaciones que suenan persuasivas y transmiten apariencia de solidez académica, pero carecen de fundamento verificable y frecuentemente resultan factual o lógicamente incorrectas (cf. Arias-Arias *et al.*, 2025).

2.2. Transparencia metodológica

Por otro lado, la opacidad de los grandes modelos de lenguaje comerciales agrava significativamente la problemática de la fiabilidad de los datos. Funcionando como

auténticas «cajas negras», estos sistemas impiden la trazabilidad completa de sus procesos de decisión, dificultando la identificación y corrección de errores. Los usuarios quedan, por tanto, a merced de *outputs* generados mediante cálculos probabilísticos cuya lógica permanece inaccesible.

Es importante precisar que esta crítica se aplica específicamente a los grandes modelos de lenguaje comerciales. Existen modelos de lenguaje alternativos, más sostenibles por su menor explotación de recursos, desarrollados específicamente para propósitos de investigación con un control riguroso de los datos de entrenamiento (cf. Gamallo *et al.*, 2024). Estos enfoques representan una dirección responsable en el desarrollo de tecnologías lingüísticas.

De forma análoga a estos últimos, anotadores semánticos automáticos contruidos a partir de ontologías contrastadas y recursos validados con un recorrido histórico en la disciplina (como puede ser el caso de *WordNet*) permiten supervisar todas las fases del desarrollo con pleno conocimiento metalingüístico. Esta supervisión garantiza transparencia metodológica y calidad científica en cada etapa del proceso, desde la concepción inicial hasta la implementación final.

2.3. Equidad multilingüe

El respeto al multilingüismo representa un desafío fundamental, especialmente para lenguas en situación de desigualdad o con escasos recursos digitales. Los grandes modelos de lenguaje, entrenados predominantemente con datos en lenguas hegemónicas como el inglés, perpetúan desequilibrios lingüísticos estructurales que pueden acelerar procesos de extinción de lenguas minorizadas. Esta perspectiva eurocéntrica y limitada al mundo occidental ignora a comunidades enteras que carecen de acceso efectivo a estas tecnologías o cuyas lenguas no están representadas en los corpus de entrenamiento.

Esta problemática es de doble naturaleza: por una parte, muchas comunidades no tienen la capacidad técnica de entrenar modelos propios o se encuentran con lenguas sin registros escritos suficientes para entrenar LLM de forma efectiva, lo que amplifica intrínsecamente la brecha digital. Por otra parte, incluso para iniciativas estatales como ALIA, es necesario incluir un volumen desproporcionadamente mayor de datos en inglés frente a lenguas cooficiales del Estado español³. Lenguas como el asturiano o el aragonés, que no disfrutan de estatus oficial, quedan sistemáticamente excluidas de esta integración tecnológica.

Las consecuencias trascienden lo meramente lingüístico. La invisibilización digital de lenguas minoritarias y minorizadas contribuye directamente a la erosión de identidades culturales e impide la transmisión de sistemas de conocimiento tradicional. Lo que observamos a pequeña escala con lenguas minorizadas (su relegación sistemática

³ https://www.newtral.es/el-gobierno-destinara-10-millones-a-alia-un-sistema-de-ia-entrenado-en-espanol-y-lenguas-cooficiales/20240524/#google_vignette

a un segundo plano en los recursos digitales) anticipa procesos de mayor envergadura que amenazan la pervivencia de culturas completas, perpetuando una dinámica que privilegia culturas occidentales en detrimento de la diversidad global.

Las herramientas lingüísticas desarrolladas conforme a criterios explícitos de equidad, representatividad y rigor científico (Gamallo *et al.*, 2024) pueden convertirse en recursos efectivos para la normalización lingüística, ofreciendo una vía concreta para contrarrestar la homogeneización cultural inherente a los modelos hegemónicos.

2.4. Mitigación de sesgos

Los grandes modelos de lenguaje comerciales, por su proceso de construcción, reproducen y amplifican los sesgos presentes en sus datos de entrenamiento. Los sesgos de género perpetúan estereotipos y roles tradicionales, los sesgos políticos reflejan las ideologías dominantes en las fuentes utilizadas y los sesgos lingüísticos consolidan la supremacía de lenguas coloniales. Esta problemática no es meramente técnica, sino profundamente política: al naturalizar desigualdades preexistentes, estos sistemas pueden contribuir a su legitimación y propagación exponencial. La amplificación de estos sesgos agrava, además, la difusión de información falsa en un entorno donde resulta cada vez más difícil contrastar la precisión de los contenidos.

Herramientas lingüísticas diseñadas *ad hoc*, como puede ser el caso de un anotador semántico, al estar fundamentadas en recursos previamente validados, permiten identificar, documentar y mitigar sesgos de forma sistemática y transparente. La supervisión humana especializada en cada fase del desarrollo proporciona una salvaguarda contra la reproducción acrítica de prejuicios.

2.5. Sostenibilidad ambiental y responsabilidad ecológica

La sostenibilidad ambiental añade una dimensión ética ineludible a esta discusión. El entrenamiento y utilización de grandes modelos de lenguaje comerciales conllevan un consumo energético e hídrico desproporcionado, con estimaciones que señalan el consumo de millones de litros de agua para refrigerar centros de datos, además de emisiones de carbono equiparables a las industrias pesadas (cf. Ding & Shi, 2024). Este impacto ambiental resulta difícilmente justificable en el contexto de la crisis climática global y de los compromisos internacionales para la reducción de emisiones.

Un anotador semántico automático, por el contrario, requiere recursos computacionales significativamente menores tanto en su fase de desarrollo como en su funcionamiento operativo cotidiano. Una vez contruidos, estos sistemas operan con una fracción mínima de la energía necesaria para ejecutar consultas equivalentes en LLM, representando una alternativa tecnológica demostradamente más sostenible y responsable con la preservación del medio ambiente.

2.6. Reutilización responsable de datos

La reutilización indiscriminada de recursos plantea interrogantes cruciales sobre la integridad de la calidad científica. Los modelos de lenguaje, impulsados por una lógica de consumo acelerado de datos de entrenamiento, incorporan información no contrastada extraída de la web sin validación previa, generando un ciclo de degradación informativa. Esta práctica resulta contraproducente para la producción científica sostenible, pues se genera información a partir de datos previamente no validados, lo que habitualmente conlleva a la aparición de imprecisiones epistémicas.

El desarrollo de herramientas lingüísticas construidas con criterios específicos garantiza que todas las fases del proceso estén supervisadas por especialistas y fundamentadas exclusivamente en recursos previamente validados. Este enfoque asegura que cada componente del sistema (desde las ontologías base hasta el corpus anotado) cumpla criterios rigurosos y verificables de calidad científica. Esta metodología permite una auténtica reutilización responsable del conocimiento existente, en lugar de la degradación informativa característica de los enfoques basados en datos masivos sin curación.

2.7. Democratización real

Finalmente, es imprescindible cuestionar críticamente la democratización ficticia que prometen los modelos comerciales. Aunque se presentan públicamente como herramientas de acceso universal al conocimiento, la realidad operativa evidencia lo contrario: numerosas empresas tecnológicas monetizan activamente el acceso a versiones avanzadas de estos modelos, estableciendo barreras económicas explícitas que reproducen y amplifican desigualdades preexistentes, perpetuando la conocida como brecha digital. Además, muchos de los datos utilizados para entrenar estos sistemas provienen de recursos en abierto y disponibles para todo el público general, sin embargo, tras ser incorporados en estos modelos, se transforman en productos privatizados que solo son accesibles mediante transacciones comerciales.

Las herramientas lexicográficas, como los anotadores semánticos, desarrollados frecuentemente con financiación pública, permanecen abiertos y accesibles para toda la sociedad sin restricciones económicas. Esta apertura genuina contribuye a una democratización real del conocimiento. El uso de fondos públicos para desarrollar herramientas que permanecerán en abierto de forma permanente representa un modelo de responsabilidad social que antepone explícitamente el bien común al beneficio privado, permitiendo que comunidades de todo el mundo accedan sin restricción a recursos lingüísticos de calidad certificada. Este compromiso con la accesibilidad universal conecta directamente con la visión articulada anteriormente respecto a la construcción de una inteligencia artificial más transparente, fiable y verdaderamente al servicio de la sociedad.

2.8. Reflexión final

Los argumentos expuestos demuestran que la apuesta por herramientas lexicográficas y semánticas sostenibles, multilingües y de código abierto no constituye una resistencia al cambio tecnológico, sino un modelo ético y científicamente riguroso para construir herramientas que sirvan genuinamente al interés público y a la comunidad científica internacional. El presente volumen aspira a contribuir a esta democratización en sentido pleno, abordando desde una perspectiva transversal temas de vital importancia para la generación de anotadores semánticos de calidad, presentándolos de manera que resulten accesibles tanto para especialistas como para un público más amplio interesado en comprender los desafíos y oportunidades de la tecnología lingüística contemporánea.

Solo mediante este compromiso colectivo (integrando excelencia científica, responsabilidad ética, sostenibilidad ambiental y equidad social) podremos contribuir efectivamente a la construcción de un futuro digital más justo, transparente y sostenible que respete la diversidad lingüística y cultural de nuestras sociedades y las generaciones futuras.

3. ESTE VOLUMEN: PROPUESTAS Y NUEVAS PERSPECTIVAS

La presente obra nace de los debates, avances y reflexiones surgidos en el marco del proyecto *Etiquetador Semántico Multilingüe Automático y Sostenible* (ESMAS-ES⁺), pero aspira a trascender los límites de un volumen monográfico academicista. Este volumen adopta deliberadamente una perspectiva transversal que busca tender puentes entre diversos ámbitos del conocimiento y públicos lectores.

Este carácter democratizador no es retórico, sino que responde a una convicción profunda: las decisiones sobre el desarrollo de herramientas lingüísticas y su integración en nuestro ecosistema digital no pueden permanecer confinadas a círculos especializados. Afectan a cuestiones de justicia lingüística, sostenibilidad ambiental, preservación cultural y acceso democrático al conocimiento que conciernen al conjunto de la sociedad. Con esta premisa en mente, cada capítulo se ha diseñado para resultar accesible tanto a investigadores especializados como a profesionales de campos adyacentes, estudiantes avanzados y lectores interesados en comprender las implicaciones de las tecnologías del lenguaje en nuestro presente y futuro inmediato.

La estructura del volumen refleja esta voluntad integradora, aunando aproximaciones teóricas, propuestas aplicadas, consideraciones metodológicas y reflexiones éticas. Los capítulos que siguen abordan aspectos fundamentales para la construcción de anotadores semánticos multilingües (desde cuestiones ontológicas y computacionales hasta desafíos de representación y evaluación), pero siempre contextualizados en el marco más amplio del debate sobre qué tipo de infraestructuras digitales queremos construir y a quién han de servir.

En un momento en el que se cuestiona abiertamente qué lugar ocupan hoy la lexicografía y la lingüística en un ecosistema dominado por la inteligencia artificial, **María**

José Domínguez Vázquez, en *¿Qué es una ontología y por qué es importante para la anotación semántica?* abre el volumen volviendo a una pregunta básica, pero nada trivial: cómo organizamos el conocimiento léxico. A partir de modelos bien conocidos —desde los *scripts* procedimentales o las estructuras prototípicas de eventos hasta las familias léxicas, los campos semánticos o las redes semánticas—, el capítulo sitúa las ontologías como una pieza clave para representar el significado de forma estructurada.

La pregunta por la fiabilidad y la transparencia de las herramientas lingüísticas cobra especial peso en *Viaje al centro del anotador: ¿Cómo se sustenta un anotador semántico desde una perspectiva tecnológica?* Aquí, **Daniel Bardanca Outeiriño** describe, paso a paso, cómo se construye un anotador semántico de sustantivos mediante el ajuste fino de un modelo de lenguaje. El proyecto se fundamenta en la reutilización de datos ya existentes y en un mantenimiento accesible. Además, el capítulo pone el acento en decisiones que hoy resultan cruciales: optar por soluciones técnicamente sencillas, reducir costes computacionales y, sobre todo, partir de datos ontológicos cuidadosamente curados por especialistas.

La sensación de que «los traductores automáticos ya lo hacen todo» sirve de telón de fondo para *¿Para qué sirve etiquetar palabras si ya tenemos DeepL?* **Lucía Esperanza del Castillo Vázquez** y **Sofía García González** examinan el papel que desempeña la anotación semántica en tareas concretas como la desambiguación léxica, el etiquetado de roles semánticos o la representación del significado abstracto. Nos explican, a través de un recorrido por la evolución de la traducción automática, que, cuando hablamos de fiabilidad, verificabilidad o transparencia metodológica, no todo puede dejarse en manos de modelos opacos que «aciertan» sin explicar cómo ni por qué.

En el capítulo *¿Podemos aprender y enseñar una lengua con etiquetas semánticas?*, **Carlos Valcárcel Riveiro** dedica unas páginas a examinar el valor pedagógico en el terreno de la enseñanza-aprendizaje de lenguas. El punto de partida es sencillo y muy humano: aprendemos palabras no como listas aisladas, sino como redes de relaciones. A partir de ahí, examina cómo el etiquetado semántico puede convertirse en un apoyo real —y no meramente técnico— para la comprensión, la producción lingüística y la organización del vocabulario en contextos educativos.

Se aborda también la cuestión de cómo evaluar con rigor la calidad de los datos anotados y el rendimiento real de los sistemas automáticos. Para ello, partiendo de la problemática del inevitable desacuerdo entre anotadores humanos ante casos ambiguos, **Vítor Míguez-Rego** y **Jose Ameijeiras-Alonso**, en su capítulo *A vueltas con los números: cómo nos puede ayudar la estadística en la anotación semántica*, nos presentan ejemplos prácticos y conceptos como el *gold standard*, la precisión o el coeficiente kappa, para introducir las principales métricas utilizadas para medir el acuerdo interanotador y validar anotadores automáticos.

Como es natural, no se puede hablar de lexicografía y semántica sin tener presentes los diccionarios. Así, **María Concepción Maldonado González**, en *¿Por qué (no) vamos a dejar de usar los diccionarios?* reflexiona sobre la vigencia de los diccionarios en un

entorno digital en el que las consultas son inmediatas y muchas veces invisibles para el usuario y, a partir de ahí, plantea una redefinición de la lexicografía que va más allá de los repertorios tradicionales, apostando por los diccionarios como bases de datos léxicas estructuradas, pensadas para integrarse —casi sin que nos demos cuenta— en aplicaciones y servicios lingüísticos contemporáneos.

Se ha hablado mucho a lo largo de estas líneas sobre la necesidad de crear redes y sistemas interconectados con los recursos lexicográficos digitales existentes para asegurar una sostenibilidad tecnológica real. Este es precisamente el tema del siguiente capítulo, *Hacia una red dinámica de recursos semánticos: sostenibilidad tecnológica y navegabilidad integrada*. Aquí, **María José Domínguez Vázquez** e **Iván Arias-Arias** presentan una propuesta concreta para articular distintos recursos lexicográficos y semánticos en una red navegable, pensada para facilitar el acceso, la interoperabilidad y el mantenimiento a largo plazo. Además de crear buenos datos, hay que pensar cómo se conectan, cómo se reutilizan y cómo se sostienen en el tiempo sin reinventar la rueda en cada nuevo proyecto.

Por último, **María Teresa Sanmarco Bande**, en *Tecnología lingüística para una sociedad multilingüe: aplicaciones, retos y buenas prácticas*, amplía el foco y mira el panorama en su conjunto. Aborda los retos que plantea el desarrollo de tecnologías lingüísticas en contextos marcados por la diversidad lingüística, desde la trazabilidad semántica y la mitigación de sesgos hasta la sostenibilidad computacional. En este marco más amplio, el proyecto ESMAS-ES+ aparece como un punto de referencia para pensar una inteligencia artificial que no solo sea eficaz, sino también transparente, equitativa y, en última instancia, más alineada con las necesidades reales de la sociedad.

En conjunto, la variedad de enfoques que se dan cita en este volumen permite apreciar hasta qué punto la lexicografía puede aportar cuando se orienta hacia la sostenibilidad y la responsabilidad social. Cada capítulo ofrece una visión del panorama innovador de la lexicografía sostenible y socialmente responsable, desde la reflexión teórica y la formalización semántica hasta la aplicación práctica en contextos educativos, traductológicos y tecnológicos. El resultado es una visión más abierta de la lexicografía, más integrada y atenta a su papel en una sociedad tan diversa y en constante transformación.

4. REFERENCIAS BIBLIOGRÁFICAS

- ARIAS-ARIAS, Iván et al. (2024). «Der Effizienz- und Intelligenzbegriff in der Lexikographie und künstlichen Intelligenz: kann ChatGPT die lexikographische Textsorte nachbilden?». *Lexikos*, 34(1), 51-76. <https://doi.org/10.5788/34-1-1879>.
- (2025). «People, machines and dictionaries: Is artificial intelligence killing dictionaries?». En Dias, I. et al. (Eds.), *Regarding the Significance of Lexicography* (pp. 235-258). De Gruyter. <https://doi.org/10.1515/9783112216644-012>
- CALAÑAS CONTINENTE, José Antonio y DOMÍNGUEZ VÁZQUEZ, María José (2023). «About Sustainable and Socially Responsible Lexicography: State of the Art». *Quaderns de Filologia: Estudis Lingüístics*, 28, 9-19. <https://doi.org/10.7203/qf.0.27659>.
- CHOMSKY, Noam et al. (2023). «The false promise of ChatGPT». *The New York Times*. <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>

- COLMAN, Lut (2016). «Sustainable Lexicography: Where to go from here with the ANW (Algemeen Nederlands Woordenboek, an Online General Language Dictionary of Contemporary Dutch)». *International Journal of Lexicography*, 29(2), 139-155. <https://doi.org/10.1093/ijl/ecw008>.
- DE SCHRYVER, Gilles-Maurice y JOFFE, David (2023). «The end of lexicography, welcome to the machine: On how ChatGPT can already take over all of the dictionary maker's tasks». Comunicación presentada en el 20º Seminario CODH, Tokyo. <http://codh.rois.ac.jp/seminar/lexicography-chatgpt-20230227/>
- DING, Yi y SHI, Tianyao (2024). «Sustainable LLM Serving: Environmental Implications, Challenges, and Opportunities». En *2024 IEEE 15th International Green and Sustainable Computing Conference* (pp. 37-38). <https://doi.org/10.1109/IGSC64514.2024.00016>.
- FUENTES MORÁN, M.^a Teresa et al. (2024). «La inteligencia artificial como herramienta lexicográfica: estudio analítico sobre el rendimiento de ChatGPT, Copilot y Gemini en unidades léxicas del español». *RLA. Revista de lingüística teórica y aplicada*, 62(1), 13-38. <http://dx.doi.org/10.29393/rla62-1iamf30001>
- GAMALLO, Pablo et al. (2024). «Open Generative Large Language Models for Galician». *Procesamiento del Lenguaje Natural*, 73, 259-270. <https://doi.org/10.48550/arXiv.2406.13893>.
- KLOSA-KÜCKELHAUS, Annette y TIBERIUS, Carole (2024). «The lexicographic process revisited». *International Journal of Lexicography*, 38(1), 1-12. <https://doi.org/10.1093/ijl/ecae016>.
- LEW, Robert (2023). «ChatGPT as a COBUILD lexicographer». *Humanities and Social Sciences Communications*, 10, 1-10. <https://doi.org/10.1057/s41599-023-02119-6>.
- LEW, Robert et al. (2024). «The effectiveness of ChatGPT as a lexical tool for English, compared with a bilingual dictionary and a monolingual learner's dictionary». *Humanities and Social Sciences Communications*, 11, 1-10. <https://doi.org/10.1057/s41599-024-03775-y>.
- MALDONADO GONZÁLEZ, M.^a Concepción (2023). «Prehistoria y ciencia ficción en Lexicografía». *Quaderns de Filologia: Estudis Lingüístics*, 28, 21-36. <https://doi.org/10.7203/qf.0.26609>.
- RUNDELL, Michael (2023). «Automating the creation of dictionaries: are we nearly there?». En *Proceedings of the 2023 ASIALEX Conference* (pp. 9-17).
- SCHIERHOLZ, Stefan (2021). «Die Villa Vigoni Thesen zur Lexikographie». *Lexicographica*, 37(1), 303-305. <https://doi.org/10.1515/lex-2021-0017>.
- TARP, Sven y GOUWS, Rufus H. (2023). «A Necessary Redefinition of Lexicography in the Digital Age: Glossography, Dictionography and Implications for the Future». *Lexikos*, 33(1), 425-447. <https://doi.org/10.5788/33-1-1826>
- TIBERIUS, Carole et al. (2023). «A Lexicographic Practice Map of Europe». *International Journal of Lexicography*, 37(1), 1-28. <https://doi.org/10.1093/ijl/ecad023>.
- TRAP-JENSEN, Lars (2024). «The Best of Two Worlds: Exploring the Synergy between Human Expertise and AI in Lexicography». En *Proceedings of the International Conference Lexicography in the XXI Century* (pp. 49-62). https://lexicography21.iliauni.edu.ge/wp-content/uploads/2024/06/03_Lars-Trap-Jensen.pdf
- WIEGAND, Herbert Ernst (1998). *Wörterbuchforschung: Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. De Gruyter.

colección:
INTERLINGUA

431

Dirigida por:
Ana Belén Martínez López y Pedro San Ginés Aguilar

¿Qué ocurre cuando la inteligencia artificial transforma la manera en la que producimos, organizamos y consultamos el conocimiento lingüístico? ¿Tienen sentido todavía los diccionarios, las ontologías o la propia investigación lingüística en la era de *ChatGPT*? *Palabras que conectan saberes. Semántica aplicada e interdisciplinariedad* reúne voces especializadas de la lingüística, la lexicografía, la informática, la estadística y la didáctica para abordar el papel de las herramientas lingüísticas construidas con criterios científicos en un ecosistema digital dominado por modelos de lenguaje cada vez más presentes en nuestra vida cotidiana.

Con un enfoque deliberadamente transversal, los capítulos abordan cuestiones que van desde la arquitectura de las ontologías hasta la evaluación estadística de anotadores, pasando por la didáctica de lenguas, la traducción automática y la sostenibilidad tecnológica, y muestran cómo los datos lingüísticos estructurados siguen siendo fundamentales para construir tecnologías más transparentes, inclusivas y fiables. Una obra dirigida tanto a especialistas como a cualquier lector interesado en comprender qué está en juego cuando las palabras se convierten en datos.



COMARES
editorial

ISBN 979-13-7033-144-3



9 791370 331443