EDICIONES COMPLUTENSE

Círculo de Lingüística Aplicada a la Comunicación
ISSN-e: 1576-4737

ARTÍCULOS

Nuevas vías para la desambiguación en frases nominales en alemán: fundamentos metodológico-lingüísticos para el desarrollo de una herramienta de anotación semántica (semi)automática

Iván Arias-Arias
Instituto da Lingua Galega, Universidade de Santiago de Compostela

https://dx.doi.org/10.5209/clac.90641 Enviado: 24 de julio del 2023 • Aceptado: 2 de noviembre del 2024

ES **Resumen:** La presente investigación describe algunas propuestas iniciales para el diseño de una herramienta (semi)automática de anotación ontológica de corpus lingüísticos en alemán. Con el propósito de evaluar la viabilidad del modelo propuesto, el artículo se centra en el análisis de estructuras argumentales nominales dentro del campo semántico de la expresión. Para este fin, se proporciona una descripción detallada de la metodología aplicada, se presentan algunos de los resultados preliminares y se sugieren nuevas vías de trabajo para el desarrollo de una posible herramienta.
Palabras clave: anotación con ontologías, corpus, desambiguación categorial, estructura argumental, interfaz sintáctico-semántica.

ENG New avenues for disambiguation within noun phrases in German: methodological-linguistic foundations for the development of a (semi) automatic semantic annotation tool.

Abstract: The present research describes some initial proposals for a (semi)automatic tool for ontological annotation of German linguistic corpora. The focus is on the analysis of nominal argument structures within the semantic field of expression, in order to assess the feasibility of the proposed model. To this end, a detailed description of the methodology used will be provided, some preliminary results will be presented and further avenues for the development of a potential tool will be suggested.
Keywords: ontology-based annotation, corpus, categorial disambiguation, argument structure, syntax-semantics interface.

Cómo citar: Arias-Arias, I. (2025). 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 (2025): 169-181. https://dx.doi.org/10.5209/clac.90641

1. Introducción

En las tesis *Villa Vigoni* (Schierholz, 2021), elaboradas por especialistas en el campo de la lexicografía para abordar los desafíos actuales de la disciplina, se hace hincapié en que uno de los retos fundamentales consiste en mejorar la documentación sistemática del código lingüístico mediante el recurso a datos automáticamente extraídos de corpus y tratados computacionalmente. Se afirma, asimismo, que la información lingüística debe presentarse teniendo en cuenta al usuario y sus necesidades. Esta observación demuestra en qué medida la lexicografía y la lingüística aplicada se encuentran todavía frente a importantes desafíos en la actual era digital.

La mayoría de las aplicaciones para el tratamiento de texto (véase apartado 3) permite el estudio de patrones lingüísticos formales o morfosintácticos, pero no ofrece la posibilidad de aplicar filtros de corte puramente semántico de forma exhaustiva. La falta de anotación semántica en corpus representa una importante limitación, especialmente para la extracción y análisis de esquemas argumentales (López, 2020; Domínguez, 2022a; Valcárcel & Pino, 2023), ya que las consultas en *Corpus Query Language* (CQL; Christ, 1994) se restringen a análisis formales y estadísticos. Esto plantea un obstáculo para la desambiguación de ciertos significados al nivel de la estructura argumental debido a la inexistencia de filtros semántico-categoriales, aspecto del que nos ocuparemos en este trabajo. Estructuras como *Frage der Zeit* ('cuestión del tiempo'), *Frage der Teilnehmer* ('pregunta de los participantes') o *Frage der Sicherheit* ('cuestión de seguridad') solamente pueden ser desambiguadas mediante criterios puramente semánticos y no a través de patrones formales. En consecuencia, surge la pregunta de investigación que guiará este estudio: ¿en qué medida se puede automatizar el proceso de anotación semántica de corpus mediante un algoritmo diseñado *ad hoc* que permita una desambiguación efectiva de este tipo de estructuras?

En las últimas décadas, han surgido algunos enfoques con orientación marcadamente semántica, entre los cuales cabe destacar la inducción de significados (*word sense induction*, WSI), que permite predecir si una unidad léxica es polisémica y que presenta su información semántica agrupada en clústeres (véase *Mandinga*; Nazar, 2010); la aproximación *Corpus Pattern Analysis* (Hanks, 2004; Hanks, 2013), cuya finalidad consiste en identificar realizaciones morfosintácticas y comportamientos semánticos para una serie de unidades léxicas; o *FrameNet* (Fillmore, 1977; Fillmore, 1985), base de datos léxica para describir la organización cognitivo-conceptual del léxico. Sobresale, en esta línea, la existencia de sistemas computacionales basados en redes neuronales artificiales para el procesamiento o generación de datos lingüísticos (véanse modelos como BERT o GPT).

Sin embargo, los escasos recursos que nos permiten trabajar con la interfaz sintáctico-semántica constituyen el punto de partida para el diseño de una metodología que permita avanzar inicialmente en la anotación semántica y automática de corpus. Dado que el método introducido en este artículo (apartados 4, 5 y 6) debe ser entendido como una aproximación piloto, nos centraremos fundamentalmente en el análisis de frases y de la argumentación nominal. Partimos de la hipótesis de que en la frase ya se puede desambiguar el significado de algunas unidades léxicas con valor polisémico, si bien es cierto que a veces es necesario recurrir a un contexto más amplio (Gross, 2013). Teniendo en cuenta esto y la intención de desarrollar una herramienta que permita procesar datos lingüísticos de la interfaz sintáctico-semántica, se presentará en esta contribución una metodología exploratoria para la anotación semántica —en este caso del significado ontológico o categorial (Engel, 2004)— de frases nominales en alemán.

El presente artículo se asienta, desde un punto de vista teórico, en la gramática y lexicografía de valencias con aplicaciones derivadas de procesamiento de lenguaje natural (PLN) y se restringe el ámbito de aplicación a un campo semántico o escenario concreto: el de la expresión (apartados 2 y 3). Además, se presentan los procedimientos derivados de PLN que subyacen al método propuesto (apartado 4), el cual será evaluado en términos cuantitativos (apartado 5). Por su parte, se discutirán los resultados obtenidos desde un punto de vista lingüístico-cualitativo (apartado 6) y se concluirá si es viable (o no) desarrollar un sistema basado en el método aquí presentado (apartado 7).

2. Gramática de valencias: aproximación a la estructura argumental nominal

En la bibliografía sobre gramática y lexicografía de valencias se ha discutido ampliamente sobre la consideración de distintas clases de palabras como portadoras de valencia, dado que Tesnière (1959) describe exclusivamente los verbos como categoría gramatical con patrones valenciales. Domínguez (2011) pone énfasis en el desarrollo multimodal del concepto original de valencia, tanto por la relevancia del nivel semántico-pragmático para la descripción fidedigna de patrones argumentales como por la observación de que clases de palabras como el adjetivo o el nombre pueden ser portadores valenciales (planteamiento también presente en Mel'čuk, 2015). Herbst (2014) añade que las estructuras argumentales valenciales puede reinterpretarse desde la gramática de construcciones, ya que este enfoque aborda directamente la interfaz léxico-sintáctica. Al fusionar la identificación de los roles de los participantes con los esquemas sintáctico-semánticos que los estructuran, la gramática de construcciones captaría la interacción entre los aspectos formales y léxicos y podría proporcionar un marco para el análisis del potencial valencial de las unidades léxicas.

En el marco de esta investigación, se parte de la premisa de que los sustantivos pueden abrir casillas valenciales a su alrededor, funcionando como predicados y activando, de este modo, esquemas argumentales (Engel, 2004; Domínguez, 2011; Mel'čuk, 2012). Teubert (1979) define la valencia nominal como sistema *sui generis* al observar que no todos los sustantivos que activan esquemas argumentales son necesariamente deverbales o tienen una correspondencia clara en una estructura verbal, como en el caso de *Straße nach Rom* ('camino a Roma'). Así, para la presentación de la estructura argumental a nivel frasal interesan tres aspectos que tienen que ver con la valencia activa del predicado (Domínguez, 2022b, p. 736; Mel'čuk, 2015, p 110): el morfosintáctico, el argumental —en cuanto abstracción de lo expresado— y el semántico (categorial y relacional). Por tanto, las unidades léxicas seleccionadas —sustantivos— funcionan como portadores valenciales y abren casillas argumentales, tanto en el plano morfosintáctico como semántico.

Con la finalidad de alcanzar una descripción conjunta de mecanismos lingüísticos de la interfaz sintáctico-semántica, nos centraremos en el campo semántico o escenario de la expresión. En este caso, entendemos que dentro de este escenario se cumple, en general, la condición de que se transmite un mensaje de

una persona A (emisor) hacia una persona B (receptor). El mensaje suele desempeñar un rol central, puesto que su especificación lingüística —aunque no siempre presente de forma explícita— se realiza en numerosos eventos:

(1) *die rege Diskussion* ***über die heutige Ermordung***[contenido_mensaje] (‘la intensa discusión **sobre el asesinato de hoy**[contenido_mensaje]’)

(2) *die Frage* ***nach Computerlinguistik***[contenido_mensaje] (‘la pregunta **sobre lingüística computacional**[contenido_mensaje]’)

En lo que concierne a la codificación lingüística de la persona emisora del mensaje, Schumacher (1986, p. 885) establece una distinción clara entre expresión monológica —con intervención de una persona (véase el ejemplo 3)— y dialógica —con dos o más participantes en el evento (véase el ejemplo 5). Asimismo, se puede distinguir entre expresiones transaccionales (en el sentido de transmisión de información) unidireccionales y bidireccionales. Aunque ambas comparten una representación mental-cognitiva que implica la copresencia de varios participantes, en las unidireccionales (véase el ejemplo 4) el rol de emisor es fijo, mientras que en las bidireccionales (véase el ejemplo 5) los participantes comparten y se intercambian entre sí el rol de emisor.

(3) *der Bericht* **des Bundesrates**[emisor_agente] **über Motionen**[contenido_mensaje] (‘el informe del **Consejo Federal**[emisor_agente] **sobre las mociones**[contenido_mensaje]’)

(4) *die Frage* ***der neuen Abteilungsleiterin***[emisor_agente] ***an die Studentinnen***[receptor] (‘la pregunta **de la nueva directora de departamento**[emisor_agente] **a las estudiantes**[receptor])

(5) *das Gespräch* ***zwischen den Ärzten***[participante_agente] (‘la conversación **entre los médicos**[participante_agente]’)

El rol del receptor desempeñaría, de esta forma, un papel relevante en el caso de las expresiones transaccionales unidireccionales o monológicas (ejemplos 3 y 4), puesto que se observa claramente la transmisión de información de A a B. En lo relativo a los eventos transaccionales bidireccionales (ejemplo 5), se afirma que en su significado intrínseco ya se entrama la copresencia de varios participantes, los cuales intercambian entre sí los papeles de emisión y recepción.

De esta forma, “partimos de la premisa de que los hablantes nativos cuentan con una escena prototípica mental de cómo se desarrollan determinados eventos, acciones o procesos, esto es, que tienen una representación mental-cognitiva de un tipo de evento” (Domínguez, 2022b, p. 737). Esta constatación implica, por tanto, que el hecho de acotar el análisis a un determinado escenario (Fillmore, 1977) —que corresponde, a grandes rasgos, con un campo semántico— permitirá obtener una visión más exhaustiva de cómo se articulan algunos eventos o acciones tanto dentro de ese escenario (Boguslavsky, 2016, p. 52), como frente a otros (como podría ser el de movimiento o desplazamiento).

Como ya se señaló anteriormente, una de las principales dificultades es la obtención de ejemplos específicos para atestiguar los diferentes patrones argumentales. En este sentido, si bien es cierto que motores de tratamiento de corpus como *Sketch Engine* ya cuentan con potentes herramientas de análisis estadístico-distribucional, estas no suelen ser suficientes para el análisis de esquemas argumentales sintáctico-semánticos. Asimismo, la utilización de búsquedas CQL (Christ, 1994) no permite obtener una desambiguación en el plano semántico-categorial, ya que carecen de este tipo de información. Así, a través de una búsqueda CQL, obtenemos resultados que, desde una perspectiva valencial y de semántica combinatoria, nos transmiten poca —o ninguna— información, como por ejemplo *Diskussion der Zeit* (‘discusión del tiempo’) o *Frage des Tages* (‘cuestión del día’).

Por tanto, para describir la interfaz sintáctico-semántica desde una perspectiva valencial, se debe prestar atención a la dicotomía entre complemento/argumento y suplemento/circunstante ya introducida por Tesnière (1959) y presente en aproximaciones actuales de la valencia (véase Domínguez, 2011; Herbst, 2014; López, 2020 o Valcárcel & Pino, 2023). En términos generales, se podría establecer que una unidad es clasificada como complemento/argumento cuando su realización semántica y formal depende directamente del portador valencial al que acompaña, pues se presenta como elemento obligatorio (sea de forma explícita o implícita) que coocurre con la unidad léxica en función de predicado. Domínguez (2014) propone un inventario de los complementos que pueden aparecer en una frase nominal con una categorización sintáctico-formal: complemento sujeto (*die Klage des Gemeinderate*s, ‘la actuación del consejo municipal’), complemento objeto (*die Annahme der Gelder*, ‘la admisión de los fondos’), complemento prepositivo (*die Angst vor der Zukunft*, ‘el miedo al futuro’), complemento adverbial (*die Reise in die USA*, ‘el viaje a EEUU’), complemento nominal (*seine Annerkennung als ein erfahrener Lehrer*, ‘su reconocimiento como profesor experimentado’) y complemento verbal (*die Überlegung, ob es Sinn macht*, ‘la consideración de si tiene sentido’). Por su parte, una entidad será considerada como suplemento/circunstante cuando tanto su forma como su semántica no vengan determinadas por el portador valencial y, por tanto, cuando su realización no sea obligatoria para completar el significado del predicado (véanse los ejemplos mencionados anteriormente: *Diskussion der Zeit* o *Frage des Tages*).

No obstante, otras ocurrencias en corpus sí que se adecuan más para establecer una descripción fidedigna de esquemas argumentales, si bien los resultados obtenidos solo se pueden desambiguar semánticamente de forma manual, como sucede con estructuras como las que se presentan a continuación:

– *Diskussion der Ergebnisse* (‘discusión de los resultados’): dado que es posible una paráfrasis oracional del tipo “se discuten los resultados”, la frase en genitivo (*der Ergebnisse*) desempeña la función de

genitivus obiectivus (complemento objeto) y el rol semántico de afectado (Engel, 1996) o "aquel o aquello afectado: tema" (Domínguez, 2022a, p. 180).

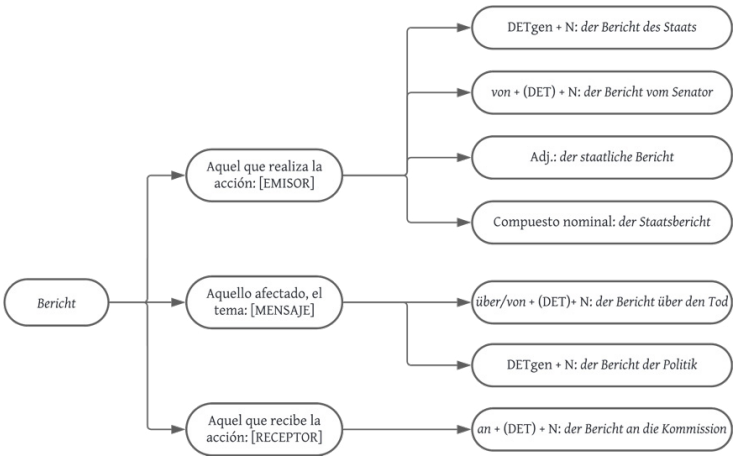
– *Diskussion der Kandidaten* ('la discusión de los candidatos'): la frase en genitivo (*der Kandidaten*) realiza —en contexto frasal— la función sintáctica *genitivus agentivus* (complemento sujeto) (Engel, 1996), dado que es posible una paráfrasis oracional como "los candidatos discuten". Desempeña un rol semántico agente, parafraseable como "aquel o aquello que realiza una acción" (Domínguez, 2022a, p. 180).

A partir de los ejemplos anteriores y del marco teórico presentado, se observa que, para extraer datos lingüísticos relevantes para el análisis de cuestiones relacionadas con la interfaz sintáctico-semántica y con la estructura argumental, es necesario establecer una serie de parámetros que nos permita proponer un método de anotación y extracción siguiendo criterios semánticos. De esta forma, al igual que una consulta CQL posibilita descartar o seleccionar determinados aspectos morfosintácticos, se propone en este estudio un método de filtrado semántico (en primer lugar, de tipo categorial; véase Engel, 2004) que nos permita afinar las consultas en corpus, y, por tanto, los resultados obtenidos.

Con este objetivo, tras consultar las categorías onomasiológicas del *Wörterbuch Deutsch als Fremdsprache* (Kempcke, 2000, p. 1298), decidimos trabajar con tres sustantivos del campo de la expresión, en concreto *Bericht*, *Diskussion* y *Frage*. Para la selección de estas unidades léxicas, se ha considerado su estructura argumental y las posibles realizaciones formales o morfosintácticas mediante las cuales se actualizan sus complementos:

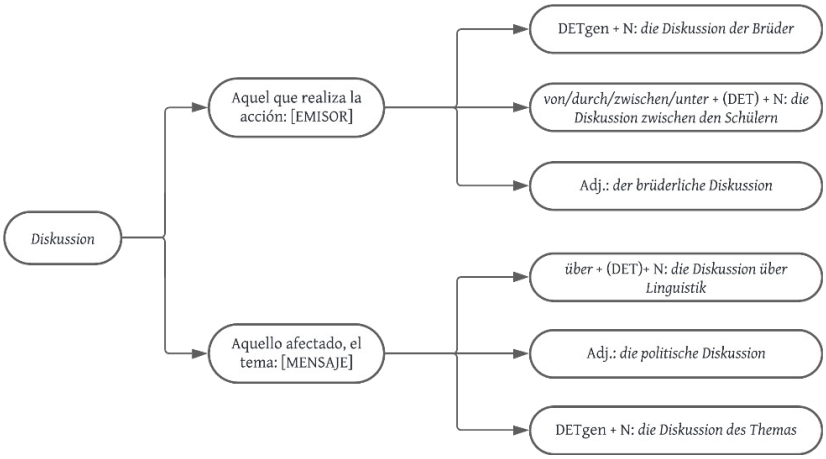
– *Bericht* ('informe'): desde un punto de vista del significado, este evento caracterizado como monológico implica la transmisión de una información detallada de tipo oficial y puramente objetivo de un emisor a un receptor (Hernández, 1993, p. 131).

Figura 1. Estructura argumental y realizaciones formales de *Bericht*



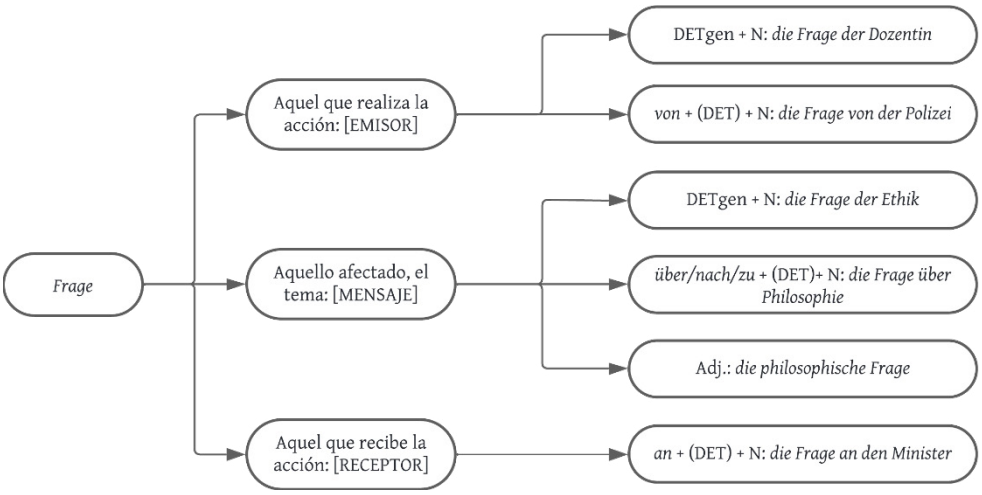
– *Diskussion* ('discusión'): se trata de un evento transaccional bidireccional en el que varios participantes intercambian mensajes y argumentan perspectivas habitualmente diferentes o controvertidas (Schumacher, 1986, p. 704).

Figura 2. Estructura argumental y realizaciones formales de *Diskussion*



– *Frage* (‘pregunta’): aunque a veces se clasifica y entiende como evento expresivo monológico o, en este caso como transaccional unidireccional, esta unidad léxica activa en el paradigma mental-cognitivo la existencia de una respuesta. En consecuencia, los sustantivos *Frage* (‘pregunta’) y *Antwort* (‘respuesta’) podrían entenderse como reacciones recíprocas entre sí (Hernández, 1993, p. 94). Además, su estructura argumental nos permite diferenciar de forma evidente dos acepciones: (i) las realizaciones del emisor y del receptor aparecen con el significado ‘interrogación que se plantea a un receptor en un esquema interactivo’ y (ii) la realización del mensaje tiene mayor tendencia a aparecer con el significado ‘cuestión o tema a debatir’.

Figura 3. Estructura argumental y realizaciones formales de *Frage*



Como se puede observar en las figuras 1-3, las unidades léxicas que pertenecen a un mismo escenario —en este caso, el de la expresión— activan roles semánticos y conceptuales parecidos que posibilitan un análisis lingüístico-cognitivo más detallado (Domínguez, 2022b, p. 737). Así, a partir de la presentación de las estructuras argumentales se observa que, para los tres predicados nominales seleccionados, el argumento “aquel que realiza la acción”, que corresponde con el complemento sujeto, presenta la posibilidad de realización morfosintáctica con una frase nominal en genitivo.

La realización en genitivo es frecuente para varios argumentos en alemán, entre los que se encuentran el complemento sujeto o el complemento objeto, y es ahí donde radica la poca eficacia de consultas CQL para el análisis de estructuras argumentales nominales, dado que la misma expresión morfosintáctica puede actualizar diferentes argumentos. En el marco de esta investigación se analizará la realización en genitivo en el complemento sujeto, puesto que esta línea de trabajo nos permitirá, por una parte, adquirir conocimientos sobre un escenario delimitado y, por otra parte, abrir nuevas vías para la desambiguación de esquemas argumentales a través de la aplicación (semi)automática de rasgos categoriales (véase Domínguez, 2022a o Boguslavsky, 2016).

De tal forma, para la actualización del rol agentivo propio del complemento sujeto en el escenario de la expresión se requieren, normalmente, lexemas de la clase semántica {humano}, aunque los niveles de especificación pueden ser más granulares si observamos la frecuencia de las unidades léxicas seleccionadas por los predicados nominales escogidos para la presente investigación. Esta granularidad y especificación se manifiesta expresamente en la ontología léxica compilada por Domínguez, Valcárcel y Bardanca (2021), en la que se recogen categorías específicas que aparecen asociadas a un determinado núcleo nominal, como {animado, humano, familia} o {animado, humano, personaje histórico}.

En esta aproximación se le concede, por tanto, una importancia especial al plano semántico, dado que este puede ser el único elemento que posibilite una desambiguación fidedigna de estructuras argumentales. La idea de que la semántica es crucial para comprender el funcionamiento de estructuras argumentales y eventos se relaciona directamente con el concepto de composicionalidad en el sentido de Pustejovsky y Batiukova (2019, p. 321), pues en su obra defienden que “the meaning of a complex expression is compositional if it is determined by the meanings of its component parts and their syntactic arrangement”.

En esta línea, cabe mencionar el mecanismo de selección (Pustejovsky, 2011) según el cual la realización que se pretende activar para un determinado predicado tiene que ser directamente satisfecha por la clase semántica del argumento que aparece. Dicho de otro modo, si se pretende expresar, por ejemplo, el rol relacional agentivo (“aquel/aquello que realiza una acción”) es necesario habitualmente recurrir a unidades léxicas pertenecientes a la clase {humano} o, al menos, a la clase {animado}. Estas relaciones afectan en primer lugar a la interacción entre el predicado y sus argumentos y se estructuran, por tanto, en un plano sintagmático.

El nivel semántico es, además, trascendental en lo que atañe al eje paradigmático, pues una casilla valencial puede estar cubierta por distintos candidatos léxicos que cumplen la condición de pertenecer a una

clase semántica específica. En este sentido, destaca el concepto de prototipo presentado por Domínguez (2021, p. 31), que difiere, al menos parcialmente, de la teoría de prototipos más tradicional:

> "Als Prototype fassen wir typische bzw. repräsentative Instanzen für eine konkrete Argumentstruktur oder Slotbesetzung auf. Dabei handelt es sich nicht um semantische abstrakte Konzepte, da der Prototypbegriff mit der syntaktischen Argumentstruktur eines konkreten nominalen Valenzträgers sowie mit der Aktualisierung einer konkreten Bedeutungslesart zusammenhängt. Ihre Typikalität sowie Repräsentativität lässt sich auf die Interaktion valenzfundierter und frequenzbezogener Parameter stützen [...].[1]"

No se trata, por tanto, de buscar el mejor candidato para representar una categoría léxica o de fijar prototipos segmentables en estancias mínimas de significado. Se trata más bien de encontrar lemas concretos que puedan aparecer en una casilla funcional específica y que realicen el papel semántico pretendido.

3. Herramientas digitales para el análisis lingüístico-valencial

En el campo del PLN, lenguajes de programación como *Python* disponen de poderosas y avanzadas bibliotecas para el tratamiento, anotación, extracción y edición de datos formales y lingüísticos (consúltense las bibliotecas *spaCy* o *NLTK*). Estas herramientas optimizan el manejo de grandes corpus lingüísticos y ofrecen soluciones para tareas como la tokenización, la lematización o el análisis sintáctico mediante la integración de modelos estadísticos y neuronales (véase Jurafsky & Martin, 2021). Desde una perspectiva lingüística, el uso de estas bibliotecas resulta crucial para la investigación, ya que permite llevar a cabo tareas como las que se detallan a continuación:

- Tokenización y lematización: estos sistemas permiten clasificar automáticamente las unidades que aparecen en un texto en *tokens* y *types*, facilitando así la segmentación y el análisis morfológico de textos.
- Identificación de la clase de palabra (*part-of-speech tagging*) y análisis morfológico: utilizando modelos estadísticos y neuronales, bibliotecas como *spaCy* predicen la categoría gramatical de cada unidad léxica. Estos enfoques han reemplazado los antiguos sistemas basados en reglas, proporcionando una mayor precisón.
- Análisis sintáctico y de dependencias: a través de técnicas de aprendizaje profundo, como las redes neuronales recurrentes y los modelos transformadores, contamos con sistemas que permiten identificar con precisión las relaciones de dependencia entre unidades léxicas y frases (véase Kondratyuk & Straka, 2019).
- Anotación semántica: herramientas como *pymusas*, en este caso basada en el *UCREL Semantic Analysis System*, permiten anotar texto categorialmente en diferentes lenguas según 21 clases semánticas. Sin embargo, esta biblioteca no resulta válida para esta investigación debido a la falta de compatibilidad con el alemán y a su diseño no adaptado específicamente para el análisis de estructuras argumentales nominales.
- Reconocimiento de entidades nombradas (*named entity recognition*, NER): los sistemas de NER identifican automáticamente entidades como nombres de persona, lugares o instituciones mediante el uso de modelos predictivos que han sido entrenados previamente con grandes corpus textuales.
- Vectores distribucionales: los modelos basados en la semántica distribucional (Harris, 1954) calculan la similitud semántica entre unidades léxicas en función de su contexto de aparición. El sistema *word2vec* (Mikolov et al., 2013) es uno de los primeros modelos en implementar este enfoque, pero herramientas más recientes como *FastText* (Bojanowski et al., 2017) han mejorado la precisión para abordar cuestiones semánticas.

Es importante señalar que los corpus lingüísticos empleados como *input* pueden variar considerablemente en cuanto a tamaño, tipo de discurso y variedad lingüística. No obstante, para una adecuada descripción de usos lingüísticos, es esencial que los corpus textuales sean representativos y fiables. Esto implica que deben abarcar el mayor número de variaciones posibles para que las conclusiones sean extensibles y generalizables al análisis de la lengua objeto de estudio. De esta forma, recurrir a textos procedentes de fuentes digitales ha ganado en importancia, dado que estos suelen cubrir un amplio abanico de géneros y registros (véase Paquot & Gries, 2020). Siguiendo esta tendencia, en este trabajo utilizaremos el corpus *deTenTen20* (disponible en *Sketch Engine*).

Aunque las herramientas citadas contribuyen significativamente a la descripción del código lingüístico, todavía carecemos de recursos que faciliten el análisis sintáctico-semántico de patrones argumentales nominales mediante un método (semi)automático. Por ello, recurrimos frecuentemente a diccionarios especializados en la valencia del sustantivo, entre los que cabe destacar, para el alemán, el *Wörterbuch zur Valenz und Distribution der Substantive* de Sommerfeldt y Schreiber, (1983). Una perspectiva innovadora, fundamentalmente debido a su enfoque multilingüe, la ofrece el diccionario en línea *Portlex* (Domínguez & Valcárcel, 2020), que otorga especial importancia al plano semántico. El presente trabajo surge, en este sentido, como

[1] Traducción propia: "Definimos los prototipos como instancias típicas o representativas de una estructura argumental concreta o de ocupación de una casilla. No se trata de conceptos semánticos abstractos, ya que la noción de prototipo está relacionada con la estructura argumental sintáctica de un portador de valencia nominal concreto y con la actualización de una acepción concreta. Su tipicidad y representatividad pueden basarse en la interacción de parámetros relativos a la valencia y a la frecuencia."

una respuesta a la falta de anotación semántica de corpus señalada por Domínguez y Valcárcel (2020) para la búsqueda de ejemplos que satisfagan la descripción de esquemas valenciales durante la fase de compilación del diccionario.

A su vez, cabe resaltar la contribución de los proyectos de generación de lenguaje natural (GNL) derivados del diccionario *Portlex* y, más directamente relacionado con los objetivos de este trabajo, el generador de frases monoargumentales *Xera* (Domínguez, 2022b). Se trata de un generador de estructuras monoargumentales disponible para español, francés y alemán en el que es posible seleccionar, además del predicado nominal, las posibles realizaciones morfosintácticas y los criterios semánticos que coocurrirán en los ejemplos creados automáticamente. Así, si se selecciona la realización morfosintáctica del complemento sujeto en genitivo, *Xera* vuelca ejemplos que cumplen ese criterio y se obtienen frases como *die Fragen des Klassenlehrers* ('la pregunta del profesor') o *die Frage der Kollegin* ('la pregunta de la compañera'). El análisis semántico constituye, por tanto, uno de los puntos clave de *Xera*, ya que permite distinguir, a través de filtros semánticos, entre las estructuras mencionadas y otra como *die Frage der Übungen* ('la pregunta de los ejercicios'), una diferenciación que solo se puede explicar con elementos categoriales y relacionales.

4. Aproximación metodológica: anotación semántica en estructuras argumentales

Como se ha señalado en los apartados previos, la falta de anotación semántico-categorial y semántico-relacional en corpus dificulta el proceso de extracción de datos lingüísticos relativos a la interfaz sintáctico-semántica en muchos de los corpus actuales (López, 2020 o Domínguez, 2022b). Consecuentemente, esta investigación pretende contribuir a la evaluación de un método piloto que pueda servir como base para el desarrollo de herramientas (semi)automáticas de anotación semántica. En esta línea, resulta pertinente retomar la pregunta de investigación planteada (véase el apartado 1), la cual explora la viabilidad de automatizar el proceso de anotación semántica en corpus mediante un algoritmo diseñado *ad hoc*. Esta cuestión se fundamenta en la hipótesis de que, al igual que una búsqueda CQL facilita el procesamiento de datos a nivel formal y morfosintáctico (Christ, 1994; Kilgarriff et al., 2014), una consulta con filtro semántico debería proporcionar resultados favorables para la extracción de información del plano categorial o relacional. A continuación, se presenta una secuencia metodológica que resume los pasos seguidos en este estudio, los cuales se explicarán en detalle más adelante:

1. Selección de la categoría semántico-ontológica {humano} y generación de una lista de vocabulario.
2. Extracción de datos lingüísticos del corpus *deTenTen20* mediante consultas CQL en formato XML.
3. Anotación de las concordancias en formato XML gracias a la aplicación de un algoritmo diseñado *ad hoc*. En primer lugar, se hace una prueba de funcionamiento del algoritmo con datos extraídos de *Xera*.
4. Evaluación cuantitativa y cualitativa del método (apartados 5 y 6).

De este modo, atendiendo a la secuenciación metodológica, se evalúa la posibilidad de alcanzar una metodología de desambiguación de estructuras argumentales nominales gracias a la aplicación de rasgos ontológicos exclusivamente. Atendiendo al objetivo del presente estudio, una delimitación en sentido más amplio de la categoría {humano} debería considerarse como suficiente (véase paso 1). Con esta finalidad, se recurre a la herramienta *Lematiza*, en la que, a partir de la introducción de datos procedentes de corpus para las estructuras argumentales ya delimitadas, se obtienen datos sobre cada argumento que se realiza a nivel frasal. Este recurso nos devuelve los lemas argumentales con un hipervínculo a *WordNet* con datos que posteriormente se pueden utilizar para consultar información ontológico-categorial en *Combina* (figura 4) gracias a la conjunción de consultas API en el repositorio multilingüe que subyace a *WordNet* (Domínguez Vázquez *et al*., 2019).

Figura 4. Interfaz de búsqueda en *Combina*

El recurso a distintas ontologías del repositorio multilingüe *EuroWordNet* mediante búsquedas API predeterminadas nos devuelve una serie de lemas almacenados en cada una de ellas. Como resultado final, se obtiene una lista de vocabulario prototipado que podría cubrir la casilla funcional deseada. Así, se puede configurar una clase semántica para la anotación semántica pretendida en el marco de este estudio en el que se encuentran un total de 2 500 sustantivos como los siguientes: *Ausländer* ('extranjero'), *Benutzer* ('usuario'), *Chefin* ('jefa'), *Europäer* ('europeo') o *Großvater* ('abuelo'). En conclusión, se crea una clase semántica unificada con el rasgo {humano} a partir de sustantivos que, en origen, pertenecen a grupos diferentes, como {animado, humano, origen}, {animado, humano, cargo} o {animado, humano, familia}.

Una vez definida esta lista de vocabulario, extraemos, en primer lugar, datos lingüísticos de corpus a través de la aplicación de filtros morfosintácticos (véase paso 2). La realización de consultas CQL en *Sketch Engine* facilita, en este sentido, el volcado de información lingüístico-formal. En el marco de esta investigación, se maneja, como ya se ha mencionado, el corpus *deTenTen20*, constituido por diferentes tipologías textuales procedentes de páginas web. En la búsqueda por concordancias de *Sketch Engine* se fijan, para el corpus seleccionado, las consultas CQL que se realizarán para la retirada de información relativa al complemento sujeto con su realización en genitivo con los predicados nominales *Bericht*, *Diskussion* y *Frage*:

- [lemma="Bericht"][tag="(ART\.(Def|Indef)|PRO.(Dem|Poss).Attr).Gen.*"][tag="N.*"]
- [lemma="Diskussion"][tag="(ART\.(Def|Indef)|PRO.(Dem|Poss).Attr).Gen.*"][tag="N.*"]
- [lemma="Frage"][tag="(ART\.(Def|Indef)|PRO.(Dem|Poss).Attr).Gen.*"][tag="N.*"]

Gracias a estas búsquedas delimitadas, se extraen datos lingüísticos anotados formalmente en *eXtensible Markup Language* (XML) con la etiqueta *Key Word in Context* (<kwic>...</kwic>), de forma que se facilita el procesamiento posterior. Se extraen, para cada una de las estructuras seleccionadas, hasta 10 000 concordancias, sobre las cuales se aplicará la anotación semántica posterior. Así, para la adecuada implementación del algoritmo diseñado *ad hoc* atendiendo al objetivo de este estudio, es indispensable introducir dos elementos como *input* al sistema:

a) La lista de vocabulario creada con *Combina* y definida como {animado, humano};
b) Las concordancias anotadas en XML con demarcación formal o morfosintáctica extraídas previamente de *Sketch Engine*.

No obstante, antes de comenzar el procesamiento de datos y con el fin de comprobar la eficacia (o no) del método propuesto, se generan estructuras monoargumentales en *Xera* (véase apartado 3) con complementos sujetos realizados en caso genitivo (véase paso 3). Si los resultados obtenidos resultan plausibles, se podrán introducir en el modelo piloto todas las realizaciones en genitivo para clasificarlas y diferenciarlas de otras funciones sintáctico-semánticas. Al ejecutar el algoritmo desarrollado, se obtiene una anotación en lenguaje XML como la siguiente:

- <kwic>die Diskussion der <sem_tag type="human"> Abteilungsleiterinnen </sem_tag> </kwic>
- <kwic>die Diskussion der <sem_tag type="human">Amtsinhaber </sem_tag> </kwic>

La lista de vocabulario creada posibilitará, por tanto, anotar aquellas unidades léxicas incluidas como {animado, humano}. El adecuado funcionamiento del modelo a pequeña escala permite que se introduzcan posteriormente como *input* los datos extraídos de *Sketch Engine* en XML (véase paso 3). No obstante, cabe destacar que el algoritmo propuesto actúa como un anotador simbólico, comparando estrictamente si las unidades léxicas están o no presentes en la lista de vocabulario para proceder a su anotación. En este sentido, es importante subrayar que se trata de una aproximación inicial, concebida para prever y sistematizar posibles casos imprevistos (véase apartado 6) en futuras fases de desarrollo.

5. Evaluación cuantitativa de la anotación semántica

Para una evaluación cuantitativa efectiva, se recurre en este caso a una matriz de confusión, técnica estadística frecuentemente utilizada para la evaluación de algoritmos en aprendizaje computacional y en estudios de corte lingüístico-computacional (Gries, 2021). Para el análisis estadístico se establecen cuatro grupos diferenciados según los cuales se puede analizar la adecuación de la anotación en las distintas concordancias de corpus:

- Verdadero positivo (VP): se trata de las ocurrencias que el algoritmo anota como {animado, humano} y que, desde un punto de vista humano, pertenecen efectivamente a esta clase semántica.
- Verdadero negativo (VN): hace referencia a las concordancias que no son anotadas por no ser reconocidas por el algortimo como {animado, humano} al no estar incluidas en la lista de vocabulario porque efectivamente no pertenecen a ese grupo.
- Falso positivo (FP): alude a las concordancias que el programa anota como {animado, humano}, pero que siguiendo un análisis humano no pertenecen a esta clase ontológica. En este caso, se puede proceder a una (re)consideración de las unidades léxicas que han sido incluidas en la lista de vocabulario, pero que no activan necesariamente la casilla funcional analizada (se trata, por ejemplo, de casos de polisemia regular; Apresjan, 1974).

– Falso negativo (FN): remite a los datos lingüísticos que no son reconocidos como {animado, humano} por el programa por no figurar en la lista de vocabulario, a pesar de que pertenecen a esta categoría. Al contrario que en el caso anterior, esto nos puede ayudar a (re)evaluar su status como posibles candidatos para ser incluidos en la lista de vocabulario.

A partir del análisis del *output*, se pueden extraer conclusiones en términos matemáticos que contribuyen a una valoración objetiva de los resultados. Interesa prestar atención especial a las categorías de falsos positivos y falsos negativos, dado que es en los puntos débiles donde se debe realizar una mejora del método. En esta línea, se ofrecerá a continuación una visión estadística-cuantitativa acerca del funcionamiento de nuestra propuesta para, posteriormente, discutir cuáles son los aspectos que desde una perspectiva lingüístico-cualitativa deben ser abordados e integrados en el futuro desarrollo de una herramienta de anotación (apartado 6).

Resulta, así, interesante destacar las dos posibilidades sobre las que se asienta el análisis cuantitativo aquí presentado: el método propuesto puede fallar o acertar en su clasificación simbólica y es exactamente esa información en la que se basan los cálculos derivados de las métricas estadísticas. Nos interesa, especialmente, ver cuál es la distribución de cada una de las clases predeterminadas considerando este método de clasificación.

Tabla 1. Matriz de confusión con los resultados de la anotación semántica

		Predicción	
		Positivo	**Negativo**
Actual	**Positivo**	2 956	79
	Negativo	3 481	23 483

A partir de la tabla 1 se puede explicar detalladamente cuál es la distribución de las categorías dentro de la matriz de confusión. Cabe mencionar que los verdaderos positivos representan una frecuencia normalizada de 9.9 % con respecto a las ocurrencias totales. Por su parte, el grupo más representativo es el de las concordancias etiquetadas como verdaderos negativos, con una frecuencia normalizada del 78.3 %. En lo que atañe a los valores clasificados como falsos, 0.3 % como falsos positivos y 11.6 % como falsos negativos, se confirma que estos representan un porcentaje inferior en términos globales, lo que parece apuntar a un adecuado funcionamiento preliminar del método, aunque estos resultados serán analizados más detalladamente con métricas concretas a continuación.

A continuación, se presentan algunas de las medidas que nos permiten extraer conclusiones sobre la viabilidad (o no) de la metodología de anotación propuesta:

a) En primer lugar, dada la relación de los valores que han sido predichos adecuadamente con el número total de ocurrencias, se confirma la precisión del método. En este caso, el resultado de la precisión es de 88.1 %, lo cual implica que tenemos un resultado bastante elevado para esta métrica, aunque existe margen de mejora.
b) La evaluación del método requiere considerar los parámetros de sensibilidad y especificidad. En el primer caso, se atiende a la relación entre VP fraccionada por la suma de VP y FN. El resultado obtenido en este caso es de 45.9 %, cifra derivada de la frecuencia relativizada del grupo de FN, constituyéndose, así, un punto que debe ser afinado en futuros estadios del modelo. Por su parte, el valor de la especificidad, derivado de la ratio de VN entre la suma de VN y FP, es claramente superior y el resultado alcanzado es de 99.7 %. Esto se debe a que los FP representan un porcentaje muy reducido en el conjunto.
c) Por último, y debido a la diferencia entre las dos métricas expuestas en b), parece interesante hacer referencia a la fórmula F1, que ayuda a comprender, atendiendo a una mayor armonización, la relación existente entre la especificidad y la sensibilidad. En este caso, el resultado estadístico es de 62.4 %, lo cual implica que el método está funcionando, pero se requiere una identificación y optimización de diferentes aspectos (véase apartado 6).
d) No obstante, si se estudian los datos relativos a cada uno de los predicados nominales seleccionados para el estudio por separado, se deduce que las concordancias con *Bericht* son las que conducen a valores más bajos en lo que respecta a la sensibilidad, puesto que la categoría de FN es elevada. En el siguiente apartado se explicará, desde un punto de vista lingüístico, el posible porqué de estos valores. De todas formas, cabe hacer hincapié en el hecho de que *Bericht* es el único sustantivo de los escogidos que no estaba incluido en los datos de *Xera* con los que se probó el algoritmo a partir de los cuales se creó la lista de vocabulario utilizada (véase apartado 4).

En conclusión, si bien esta aproximación metodológica presenta algunos aspectos optimizables (fundamentalmente relacionados con la reducción del número de FN que darían lugar al consecuente aumento del valor derivado de la fórmula F1), su funcionamiento parece ser adecuado en términos estadísticos a pesar de su carácter simbólico y estricto. Para alcanzar un mayor conocimiento sobre

dichos aspectos y las consiguientes fases de trabajo para su mejora, conviene sistematizar los factores lingüístico-cualitativos que aún no detecta (véase apartado 6), que serán los que nos permitan integrar en fases futuras de desarrollo aspectos cuantitativos para la predicción de las clases semántico-ontológicas deseadas.

6. Evaluación cualitativa de la anotación semántica

El método de anotación semántica propuesto apunta a resultados favorables, puesto que los valores métricos derivados de su aplicación se sitúan en una escala positiva (véase apartado 5). En esta línea, interesa hacer un recorrido por los aspectos que el algoritmo identifica y anota adecuadamente, así como por los motivos que hacen que no reconozca algunas estructuras monoargumentales realizadas en genitivo con el rasgo categorial {animado, humano}. Para tal fin, se recurre a la aplicación de expresiones regulares, dado que nos pueden ayudar en la identificación de unidades léxicas que han sido clasificadas como falsos positivos o falsos negativos principalmente. De esta forma se persigue avanzar en la descripción semiautomática de la interfaz sintáctico-semántica, para lo cual una sistematización de expresiones regulares podrá posibilitar la identificación de puntos vulnerables del algoritmo y permitirá establecer perspectivas de optimización.

En primer lugar, se constata que el programa desarrollado funciona adecuadamente o de acuerdo con los objetivos para los que fue originalmente concebido debido a que no anota unidades léxicas que no fueron previamente recogidas en la lista de vocabulario predefinida. Esto garantiza que se puedan excluir directamente sustantivos que cubren otras casillas funcionales o argumentales. El caso más destacado es el del complemento objeto, pues la realización formal en genitivo también puede actualizar este argumento. De hecho, estructuras como *Bericht des* [contenido_mensaje] ('informe de$_{\text{[frase nominal en genitivo]}}$...') o *Frage der/des* [contenido_mensaje] ('cuestión/pregunta de$_{\text{[frase nominal en genitivo]}}$...') son altamente frecuentes en el corpus, pero no son anotadas por el algoritmo diseñado, puesto que los rasgos categoriales que aparecen pertenecen a categorías léxicas como {proceso}, {estado} o {mundo intelectual, contenido} (Domínguez, Valcárcel & Bardanca, 2021). Podría afirmarse el mismo comportamiento para el caso de *Diskussion*, pues estructuras argumentales como *Diskussion des Themas* ('discusión del tema') o *Diskussion der Ergebnisse* ('discusión de los resultados') representan una frecuencia del 8 % en el corpus empleado.

También se confirma que el programa no reconoce las unidades léxicas que actualizan la casilla funcional de circunstante, dado que los rasgos ontológicos que las definen, muy frecuentemente de la clase léxica {lugar} o {unidad, tiempo}, no han sido incluidas ni reflejadas en la lista de vocabulario que funciona como *input*. En consecuencia, estructuras como *Bericht des Jahres* ('informe del año') o *Diskussion der Zeit* ('discusión del tiempo') quedan excluidas de la anotación y son clasificadas en términos cuantitativos como verdaderos negativos, ya que no corresponden con el patrón predefinido.

Los aspectos aquí mencionados implican que el algoritmo no reconoce aquellas unidades que no han sido incluidas en la lista de vocabulario y demuestran, por tanto, que cumple con la función para la que ha sido diseñado. No obstante, en la clase de falsos negativos se encuentran estructuras argumentales cuyo actante cuenta con el rasgo ontológico {animado, humano}, pero que no han sido etiquetadas principalmente por dos motivos:

a) Las entidades no figuran en la lista de vocabulario creada y no son reconocidas.
b) Las unidades léxicas que aparecen en la casilla funcional del complemento sujeto presentan algunas alteraciones morfosintácticas o gráficas con respecto a las variantes lingüísticas que constan en la lista definida y, por tanto, no son anotadas.

Al primer grupo pertenecen, por ejemplo, las siglas o acrónimos, que pueden actualizar la realización de complemento sujeto por contar con el rasgo ontológico {animado, humano}, al referirse a menudo a nombres de instituciones u organizaciones. Este factor resulta especialmente llamativo en las estructuras argumentales con el predicado nominal *Bericht*, dado que el significado inherente de este sustantivo hace alusión a actos formales, objetivos o incluso jurídicos. Con bastante frecuencia se trata de siglas o acrónimos que remiten a nombres de periódicos nacionales o regionales, por ejemplo, en estructuras como *Bericht der HNA* ('informe del HNA').

En esta línea, y aunque relativamente menos frecuentes, cabe destacar la existencia de nombres propios de persona como realización morfosintáctica que desencadena la actualización del complemento sujeto en determinados patrones: por ejemplo, aparece la frase *Bericht des Lukas* ('informe de Lukas'). Algunos estudios (véase Nazar & Renau, 2016) señalan la dificultad de sistematizar taxonómicamente nombres propios para fines de anotación o extracción de información por su poca uniformidad y alta variabilidad. Una solución posible a este problema podría ser la utilización de sistemas de reconocimiento de entidades nombradas ya entrenados en algunas bibliotecas de *Python*, como *spaCy* o *Flair*.

Por su parte, la aplicación de expresiones regulares puede favorecer, a propósito de una aproximación semiautomática, el reconocimiento y consecuente etiquetado de algunas unidades léxicas. Destaca, en primer lugar, la elevada frecuencia de compuestos morfosintácticos y léxicos que activan la casilla funcional del argumento aquí analizado. Para su procesamiento y anotación, el recurso a expresiones regulares que incluyan el segundo miembro del compuesto, esto es, la palabra base, pueden ayudar a extraer información sobre la interfaz sintáctico-semántica. Un ejemplo son las expresiones regulares .*leute ('gente') o .*vorsitzende ('presidencia'), pues el hecho de que las palabras base del compuesto cuenten con el rasgo

{animado, humano} garantiza que la nueva formación morfosintáctica mantendrá el significado ontológico de partida.

Esta misma perspectiva puede ser aplicada con la frecuencia relativamente alta de participios de presente —cuya terminación es *-nde(n)*— como actualizadores de la casilla funcional del complemento sujeto. Esta tendencia, cada vez más habitual en el lenguaje inclusivo en alemán, puede abordarse desde un punto de vista computacional con la aplicación de expresiones regulares como .*nden</kwic>$, para etiquetar todas las concordancias que incluyan esta variante lingüística desempeñando la función de complemento sujeto. Así, unidades como *Teilnehmenden* ('participantes'), sin marca de género en alemán, pueden ser anotadas gracias a esta revisión manual, puesto que no habían sido incluidas originalmente en la lista de vocabulario.

Además, se debe mencionar que, en el corpus utilizado, *deTenTen20*, se incluyen concordancias con interferencias —fundamentalmente gráficas— o con (meta)datos no relevantes para el análisis lingüístico. Se trata fundamentalmente de errores en la fase de preprocesamiento textual que dan lugar a problemas en la fase de extracción de información sintáctico-semántica. Un ejemplo son los elementos gráficos que impiden su adecuado reconocimiento. Este problema se podría solucionar, sin embargo, mediante la realización de un nuevo preprocesamiento textual que nos permitiese eliminar, por ejemplo, signos de puntuación inadecuadamente colocados.

Por último, es necesario recordar que el contexto frasal dificulta la desambiguación de algunas estructuras, específicamente de aquellas en las cuales uno de los elementos —núcleo o argumento— cuenta con polisemia regular o lógica (véase Apresjan, 1974 o Pustejovsky, 1995). Entendemos que una unidad léxica presenta polisemia de este tipo cuando activa varios significados relacionados entre sí por relaciones semánticas como la metáfora o la metonimia. Así, aparecen más de 100 concordancias con la estructura *Bericht der Zeitung* ('informe del periódico') en las cuales el escaso contexto impide desambiguar de qué sentido se trata:

1. '{institución} responsable de publicar un periódico'. Se desencadena el rol relacional agentivo con la paráfrasis "el informe [realizado por personal] del periódico".
2. '{texto} que resulta del proceso de publicación periodística y que presenta noticias de actualidad'. Se activa el rol temático locativo que se puede parafrasear de este modo: "el informe [que se encuentra físicamente en una versión] del periódico".

En síntesis, el método propuesto funciona y es viable en términos cualitativos, si bien requiere ser afinado para una anotación más exacta en futuras ampliaciones, como en el complejo caso de la polisemia regular, quizás solucionable mediante la integración de aspectos cuantitativos de inducción de significado. La creación de listas de vocabulario específicas para el análisis de las realizaciones formales y funcionales manejando herramientas como *Combina* o la ontología léxica diseñada por Domínguez, Valcárcel y Bardanca (2021) ayuda a la desambiguación de estructuras argumentales. Asimismo, y aunque actualmente sea necesario recurrir a aproximaciones semiautomáticas, continuar y profundizar en esta línea supondría que mediante fases de aprendizaje profundo basadas en criterios estadístico-predictivos el sistema podría ser capaz de reconocer las unidades léxicas sin necesidad de predeterminar y preestablecer listas de vocabulario cerradas.

7. Conclusión y proyecciones futuras

El presente estudio constituye una primera aproximación al diseño de una herramienta piloto para la anotación (semi)automática y semántico-categorial de corpus textuales en lengua alemana. Se parte de la premisa de que el carente etiquetado de la interfaz sintáctico-semántica en la mayor parte de los corpus dificulta la extracción y análisis de estructuras argumentales sintáctico-semánticas. El método de análisis propuesto conjuga postulados de la teoría de valencias con aproximaciones de la lingüística computacional y PLN. Gross (2013) afirma que para alcanzar un análisis automático y eficaz de textos se debe haber realizado anteriormente una descripción muy detallada de los distintos niveles implicados en la expresión lingüística, en este caso, la sintaxis y la semántica.

De acuerdo con esa premisa se decide delimitar en esta aproximación piloto el ámbito de aplicación de nuestro estudio, de modo que se prueba el método de análisis exclusivamente con estructuras argumentales de sustantivos del escenario de la expresión. En concreto, se analizan frases nominales con complementos sujetos realizados en caso genitivo y con el rasgo categorial {animado, humano}. Los resultados demuestran la viabilidad del modelo propuesto para la anotación semántica a pequeña escala. Se requiere ahora aplicarlo en un mayor volumen de datos y optimizar las debilidades ya detectadas. En primer lugar, se debe ampliar el tamaño de la lista de vocabulario para garantizar una mayor cobertura de elementos lingüísticos con un complemento sujeto en genitivo. Asimismo, sería interesante crear listas para otras clases semánticas para poder anotar y reconocer ontológicamente unidades léxicas que activan otras casillas funcionales o valenciales. Se identifican, además, problemas como la polisemia regular, que representan un desafío importante para la consolidación de un modelo como el presentado en este artículo.

Si bien en esta investigación nos hemos centrado exclusivamente en el plano categorial, es necesario ampliar el ámbito de análisis incluyendo la identificación y anotación de significados relacionales (Engel, 2004). Para lograrlo, resulta indispensable profundizar en el estudio del plano sintagmático y considerar aspectos distribucionales y contextuales. El estudio piloto confirma, además, que el proceso de extracción

y anotación de información lingüística de la interfaz sintáctico-semántica puede ser automatizado de forma efectiva en términos estadístico-cuantitativos.

En resumen, se ha demostrado la viabilidad de un modelo semiautomático piloto basado en listas de vocabulario anotadas con rasgos ontológico-categoriales que permite extraer información lingüística para un análisis más profundo de la interfaz sintáctico-semántica. Se plantean ahora nuevos desafíos y vías de trabajo futuro, por lo que la confluencia de aproximaciones procedentes de la inteligencia artificial con otras propias de la lingüística aplicada parece ponernos en el camino adecuado para el análisis semiautomático y semántico de textos en corpus.

Agradecimientos

Este trabajo ha sido realizado en el marco del proyecto de investigación *Etiquetador semántico multilingüe automático y sostenible (ESMAS-ES+)* (PID2022-137170OB-I00), financiado por MCIN/AEI/10.13039/501100011033 y FEDER/UE "Una manera de hacer Europa". Agradezco, además, el apoyo del Ministerio de Ciencia, Innovación y Universidades a través de una ayuda para la Formación de Profesorado Universitario (FPU21/00188). Asimismo, expreso mi gratitud a los revisores anónimos por sus valiosas sugerencias y comentarios, los cuales han contribuido a mejorar y enriquecer este artículo.

Referencias

Apresjan, Juri D. (1974). Regular Polysemy. *Linguistics*, *12*(142), 5-32. https://doi.org/10.1515/ling.1974.12.142.5.

Boguslawsky, Igor (2016). On the Non-canonical Valency Filling. *Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces*, 51-60.

Bojanowski, Piotr, Grave, Edouard, Joulin, Armand, & Mikolov, Tomas (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics, 5*, 135-146.

Christ, Oliver (1994). A modular and flexible architecture for an integrated corpus query system. *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research*, 1-10.

Domínguez Vázquez, María José (2011). *Kontrastive Grammatik und Lexikographie: Spanisch-deutsches Wörterbuch zur Valenz des Nomens.* Iudicum.

Dominguez Vázquez, María José (2014). Nomenergänzungen aus grammatischer Sicht. Forschungsstand und Bestandsaufnahme. *Neuphilologische Mitteilungen*, *115*(1), 3-32.

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

Domínguez Vázquez, María José (2022a). Contribución de la semántica combinatoria al desarrollo de herramientas digitales multilingües. *Círculo de lingüística aplicada a la comunicación*, *90*, 171-188.

Domínguez Vázquez, María José (2022b). Estructura argumental del nombre: Generación automática. *Revista signos: estudios de lingüística*, *55*(110), 732-761.

Domínguez Vázquez, María José, Solla Portela, Miguel Anxo, & Valcárcel Riveiro, Carlos (2019). Resource Interoperability: Exploiting Lexicographic Data to Automatically Generate Dictionary Examples. *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal, 2019, págs. 51-71*, 51-71.

Domínguez Vázquez, María José, & Valcárcel Riveiro, Carlos (2020). PORTLEX as a multilingual and cross-lingual online dictionary. En *Studies on multilingual lexicography* (pp. 135-158). De Gruyter Mouton.

Domínguez Vázquez, María José, Valcárcel Riveiro, Carlos & Bardanca Outeiriño, Daniel (2021). Ontología léxica. Santiago de Compostela. http://portlex.usc.gal/ontologia/.

Engel, Ulrich (1996). Semantische Relatoren: Ein Entwurf für künftige Valenzwörterbücher. En *Semantik, Lexikographie und Computeranwendungen* (pp. 223-236). https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/3031.

Engel, Ulrich (2004). *Deutsche Grammatik: Neubearbeitung*. Iudicum.

Fillmore, Charles J. (1977). Scenes-and-frames semantics. En *Linguistic Structures Processing* (pp. 55-79). North-Holland Publishing.

Fillmore, Charles J. (1985). Frames and the semantics of understanding. *Quaderni Di Semantica*, *6*(2), 222-254.

Gries, Stefan Th. (2021). Statistics for linguistics with R: A practical introduction (3ª edición). De Gruyter.

Gross, Gaston (2013). *Manual de análisis lingüístico*. Editorial UOC.

Hanks, Patrick (2004). Corpus pattern analysis. *EURALEX 2004 Proceedings*, 87-97. https://euralex.org/publications/corpus-pattern-analysis/.

Hanks, Patrick (2013). *Lexical Analysis: Norms and Exploitations*. The MIT Press. https://doi.org/10.7551/mitpress/9780262018579.001.0001.

Harris, Zellig S. (1954). Distributional Structure. *Word, 10*(2-3), 146-162.

Herbst, Thomas (2014). The valency approach to argument structure constructions. En *Constructions Collocations Patterns* (Vol. 282, pp. 167-216). De Gruyter. https://doi.org/10.1515/9783110356854.167.

Hernández, Eduardo Jorge (1993). *Verba dicendi. Kontrastive Untersuchungen Deutsch-Spanisch*. Peter Lang.

Jurafsky, Daniel, & Martin, James H. (2021). *Speech and Language Processing* (3ª edición). Pearson.

Kempcke, Günter (2000). *Wörterbuch Deutsch als Fremdsprache*. De Gruyter.

Kilgarriff, Adam, Baisa, Vít, Bušta, Jan, Jakubíček, Miloš, Kovář, Vojtech, Michelfeit, Jan, Rychlỳ, Pavel, & Suchomel, Vít (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7-36.

Kondratyuk, Dan, & Straka, Milan (2019): 75 Languages, 1 Model: Parsing Universal Dependencies Universally. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2779-2795.
López Iglesias, Nerea (2020). Analysing nominal phrase contexts for the automatic extraction of linguistic and lexicographic data [Trabajo Fin de Máster]. Universidade do Minho. *RepositóriUM*, https://hdl.handle.net/1822/68562.
Mel'čuk, Igor (2012). *Semantics. From meaning to text*. John Benjamins.
Mel'čuk, Igor (2015). *Semantics. From meaning to text* (vol. 3). John Benjamins.
Mikolov, Tomas, Chen, Kai, Corrado, Greg, & Dean, Jeffrey (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781.*
Nazar, Rogelio (2010). A quantitative approach to concept analysis [Tesis doctoral]. Universitat Pompeu Fabra. *TDX (Tesis Doctorals en Xarxa)*. https://www.tdx.cat/handle/10803/7516.
Nazar, Rogelio, & Renau, Irene (2016). A Taxonomy of Spanish Nouns, a Statistical Algorithm to Generate it and its Implementation in Open Source Code. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1485-1492.
Paquot, Magali, & Gries, Stefan Th. (Eds.). (2020). *A Practical Handbook of Corpus Linguistics*. Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1.
Pustejovsky, James (1995). *The Generative Lexicon*. The MIT Press.
Pustejovsky, James (2011). Coercion in a general theory of argument selection. *Linguistics*, *49*(6). https://doi.org/10.1515/ling.2011.039.
Pustejovsky, James, & Batiukova, Olga (2019). *The Lexicon*. Cambridge University Press.
Schierholz, Stefan (2008). Corpusbasierte Operationalisierungsstrategien zur Bestimmung von Valenzpartnern. En *Akten des XI. Internationalen Germanistenkongresses Paris 2005. «Germanistik im Konflikt der Kulturen»* (Vol. 80, pp. 37-48). Peter Lang.
Schierholz, Stefan (2021). *Die Villa Vigoni Thesen zur Lexikographie*. Lexicographica, 37(1), 303-305. https://doi.org/10.1515/lex-2021-0017.
Schumacher, Helmut (1986). *Verben in Feldern: Valenzwörterbuch zur Syntax und Semantik deutscher Verben*. De Gruyter. https://doi.org/10.1515/9783110861853.
Sommerfeldt, Kar-Erns, & Schreiber, Herbert (1983). *Wörterbuch zur Valenz und Distribution der Substantive*. Max Niemeyer Verlag. https://doi.org/10.1515/9783111549491.
Tesnière, Lucien (1959). *Eléments de syntaxe structurale*. Klincksieck.
Teubert, Wolfgang (1979). *Valenz des Substantivs. Attributive Ergänzungen und Angaben.* Pädagogischer Verlag Schwann.
Valcárcel Riveiro, Carlos, & Pino Serrano, Laura (2023). Application d'une méthodologie d'analyse des prédicats nominaus: l'exemple du lexème MORT$_1$. *Çédille, Revista de Estudios Franceses, 24*, 557-579. https://doi.org/10.25145/j.cedille.2023.24.27.

Otros recursos electrónicos

Combina = http://portlex.usc.gal/develop/combina.php
CPA = https://pdev.org.uk/
Flair = https://huggingface.co/flair/ner-english
FrameNet = https://framenet.icsi.berkeley.edu/
Lematiza = http://portlex.usc.gal/develop/lematiza/
Mandinga = http://www.tecling.com/cgi-bin/mandinga/index.pl
NLTK = https://www.nltk.org/
pymusas = https://github.com/UCREL/pymusas
Portlex = http://portlex.usc.gal/diccionario/
Sketch Engine = https://www.sketchengine.eu/
spaCy = https://spacy.io/
UCREL Semantic Analysis System = https://ucrel.lancs.ac.uk/usas/
WordNet = https://wordnet.princeton.edu/
Xera = http://portlex.usc.gal/combinatoria/usuario