Núñez-Torres, Fredy (2025) Modelo para la desambiguación léxica automática basado en una medida híbrida. Journal of Computer-Assisted Linguistic Research 9: 43-62.

Modelo para la desambiguación léxica automática basado en una medida híbrida

Fredy Núñez-Torres*
Pontificia Universidad Católica de Chile, Chile
*Autor de correspondencia: frnunez@uc.cl
Recibido: 27 Junio 2025 / Aceptado: 21 Septiembre 2025 / Publicado: 20 Noviembre 2025

Abstract
This research presents the development of a more robust model for measuring semantic similarity than those currently available for solving the problem of word sense disambiguation applied to natural language processing. The model is based both linguistically and statistically on the interaction of two approaches to taxonomic exploration: path-based and information content, through the incorporation of FunGramKB as a sense inventory. In terms of evaluation, the proposed similarity measure consistently generated efficient results from a linguistic perspective in the automatic lexical disambiguation process.

Keywords: lexical ambiguity, word sense disambiguation, natural language processing, computer-based linguistic resources

Resumen
Esta investigación presenta el desarrollo de un modelo más robusto de medida para la similitud y relación semántica que los disponibles actualmente para resolver el problema de la desambiguación léxica automática, aplicado al procesamiento del lenguaje natural, y fundamentado tanto lingüística como estadísticamente en la interacción de dos enfoques de exploración taxonómica: distancia entre rutas y contenido de información, a través de la incorporación de FunGramKB como inventario de sentidos. En cuanto a la evaluación, la medida de similitud propuesta logró resultados consistentemente eficientes desde un punto de vista lingüístico en el proceso de desambiguación léxica automática.

Palabras clave: ambigüedad léxica, desambiguación léxica automática, procesamiento del lenguaje natural, recursos lingüísticos informatizados

1. Introducción

En este artículo se presenta la propuesta de un modelo de desambiguación léxica automática basado en una medida híbrida. Su implementación se fundamenta en la interacción de dos

DOI: https://doi.org/10.4995/jclr.2025.24934
Publisher: Universitat Politècnica de València.

enfoques de exploración taxonómica: distancia entre rutas (Leacock & Chodorow, 1998) y contenido de información (Zhou *et al*., 2008). Además, se utiliza la base de conocimiento léxico-conceptual-gramatical FunGramKB [1] (Periñán-Pascual & Arcas-Túnez, 2007; 2010; Periñán-Pascual, 2012) como un inventario de sentidos que permite explotar dicha interacción. En primer lugar, se exponen los fundamentos para la propuesta de una medida híbrida y su posicionamiento en el modelo de desambiguación. En segundo lugar, se describe la naturaleza de la información conceptual disponible en la base de conocimiento FunGramKB, a partir de ejemplos específicos de las unidades léxicas que fueron seleccionadas para los experimentos de aprendizaje automático, junto con sus correlatos conceptuales. Finalmente, se presenta la evaluación del modelo de desambiguación léxica automática propuesto: primero, se exponen las variables necesarias para el cálculo de la medida híbrida, cuyos valores han sido extraídos desde la ontología de FunGramKB; segundo, se proponen tres casos de evaluación junto con sus respectivos resultados, considerando los pasos necesarios para la aplicación de la medida de similitud $SIM_{\text{híbrida}}(c_i, c_j)$.

La desambiguación léxica automática constituye uno de los desafíos más relevantes en el ámbito del procesamiento del lenguaje natural (PLN). La selección precisa del sentido de una palabra en un contexto determinado es una condición necesaria para diversas tareas, como la traducción automática, la recuperación de información o la generación de texto. En la actualidad, la aparición de grandes modelos de lenguaje, como los modelos generativos, ha reconfigurado el panorama del PLN. Sin embargo, estos sistemas, aunque altamente eficientes, proporcionan resultados difíciles de explicar y de controlar desde el punto de vista semántico y cognitivo. En contraste, los modelos basados en conocimiento siguen siendo esenciales para comprender los mecanismos subyacentes de interpretación semántica, pues permiten explicitar el proceso de desambiguación mediante relaciones conceptuales.

Así, desde una perspectiva lingüística, la relevancia del fenómeno radica en que la desambiguación léxica refleja procesos cognitivos fundamentales implicados en la comprensión del lenguaje y que intentan ser reproducidos artificialmente. Según lo anterior, la manera en que un sistema computacional resuelve la ambigüedad puede compararse, en cierto sentido, con las estrategias inferenciales humanas, que integran información léxica, conceptual y contextual para acceder al significado. Por ello, un modelo de desambiguación léxica automática que articule principios lingüísticos con estructuras de conocimiento explícitas no solo aporta al desarrollo de sistemas más precisos, sino también más coherentes con la arquitectura cognitiva que subyace al procesamiento humano del lenguaje.

2. El problema computacional de la ambigüedad léxica

Desde la perspectiva del PLN, la desambiguación léxica se define como una habilidad computacional, y por lo tanto automática, para activar una interpretación posible de una unidad léxica en un contexto oracional determinado. Este problema es de suma importancia, ya que no solamente implica una adecuada descripción del fenómeno en análisis, sino que requiere

[1] FunGramKB, *Functional Grammar Knowledge Base*, es una base de conocimiento léxico-conceptual-gramatical multipropósito, cuyo objetivo es el procesamiento de las lenguas naturales.

del desarrollo de un algoritmo que sea capaz de asignar sentidos a unidades léxicas que funcionan en contextos lingüísticos auténticos. El procedimiento tradicional supone etiquetar un corpus y proveer un número consistente de reglas para la selección de sentidos de una unidad léxica en un contexto oracional determinado. En este sentido, el desarrollo de medidas para hacer más eficientes distintos procedimientos de extracción de información desde bases de datos léxicas es el componente fundamental de los métodos estocásticos. Así, la ambigüedad léxica se puede manifestar en dos casos: homonimia o polisemia, como ejemplifican Núñez-Torres y Pérez-Cabello (2023).

a. Homonimia: coincidencia de dos lemas en su escritura o pronunciación, aún cuando difieren en sus sentidos. Presenta, a su vez, dos tipos:
    i. Homógrafo: coincidencia gráfica en la escritura.
    p. ej. *can* (poder - lata)
    ii. Homófono: coincidencia en la pronunciación, diferencia en la escritura.
    p. ej. *write* (escribir) | *right* (derecho).
b. Polisemia: coincidencia de dos lemas en su escritura y pronunciación, cuyos sentidos se encuentran relacionados con el mismo significado:
    i. p. ej. *bank* (institución) | *bank* (edificio de la institución).

El tratamiento de la polisemia presente en estos ejemplos considera la dificultad de seleccionar adecuadamente los parámetros del *output* que una búsqueda pueda arrojar a partir de un determinado *input*. Estos sistemas deben comprender una alta eficiencia computacional para procesar grandes volúmenes de información. En efecto, se sitúa la polisemia como uno de los problemas más importantes que enfrenta actualmente el desarrollo de sistemas para la recuperación de información, sobre todo aquellos que se focalizan en sistemas que permiten desambiguar términos en contextos de conocimiento especializado, como PubMed[2], o generales como FrameNet[3]. Desde la década de 1950 ha existido interés, en el contexto del tratamiento computacional del lenguaje, por la desambiguación del sentido de las palabras como una labor necesaria e intermedia para lograr determinadas tareas de PLN. Las máquinas no poseen la habilidad inherente para procesar lenguas naturales y, por tanto, son incapaces de reconocer casos de polisemia a menos que se les provea de mecanismos específicos para lograr esta tarea. Es fundamental comprender el problema de la desambiguación léxica como un aspecto del desarrollo de la inteligencia artificial que solo puede ser solucionado en la medida que sean abordados otros dos problemas: (a) la representación del conocimiento enciclopédico y (b) la representación del conocimiento que proviene del sentido común. A este respecto, el proceso de desambiguación léxica se define, según Nevzorova *et al*. (2015), como la habilidad de la máquina para identificar el significado de un ítem léxico en determinados contextos a partir de procedimientos computacionales.

En general, se asume que la desambiguación léxica constituye un procedimiento de

[2] https://www.ncbi.nlm.nih.gov/pubmed

[3] https://framenet.icsi.berkeley.edu/fndrupal

asignación de significado a una unidad léxica, que está basado en el contexto en el que esa unidad ocurre (Patwardhan *et al*., 2003). Esto es coherente con la idea de que un significado adecuado para un determinado ítem léxico será seleccionado a partir de un inventario de sentidos, que definen a su vez el rango de posibilidades para esa unidad léxica en un contexto particular. Existen ámbitos del PLN en los que la resolución del problema de la desambiguación es crítica para establecer de manera eficiente la comunicación hombre-máquina.

Asimismo, la desambiguación léxica requiere de un proceso que lleve a cabo la asociación entre cierto ítem léxico en un texto (o discurso) y una definición o significado (sentido) entre varios significados potencialmente atribuibles a esa palabra. La traducción mecánica, por su parte, tiene por objetivo determinar la opción correcta de significado para un ítem léxico, a partir de un número finito de significados posibles. Este método se utiliza principalmente en la traducción de textos técnicos, es decir, aquellos que pertenecen a dominios reducidos y/o especializados. Esta distinción podría facilitar la desambiguación en tanto se trata de información contextual relevante para el proceso.

Como otras tareas en el ámbito del PLN, la desambiguación léxica automática está sometida al llamado problema del cuello de botella en la adquisición del conocimiento, abordado por Buchanan & Wilkins (1993), Wagner (2006), Aussenac-Gilles & Gandon 2013, y Pasini (2020). Para comprenderlo, es fundamental reconocer que la representación del conocimiento (en este caso, entendido como conocimiento de mundo proveniente desde el sentido común) se realiza mediante una base de conocimiento que opera a través de un sistema de inteligencia artificial. Entonces, cualquier sistema de PLN basado en conocimiento debería funcionar mediante la relación entre la representación del conocimiento, una base de conocimiento y un motor de inferencia. Así, según Wagner (2006), el problema del cuello de botella en la adquisición del conocimiento consiste en el límite que se plantea para la representación del conocimiento, dado que el desarrollo de una ontología generalmente requiere de expertos que puedan contribuir a poblar la base de conocimiento. Estos expertos, a su vez, emprenden manualmente una tarea que, evidentemente, es costosa en términos de tiempo y recursos. Además, no serían capaces de dar cuenta de todo el conocimiento disponible en las fuentes de información, pues en la medida que una base de conocimiento crece, también lo hace el requisito de mantenimiento y actualización. Así, la eficiencia de un sistema de desambiguación dependerá de la calidad y cantidad de información que contenga un recurso lingüístico informatizado que constituirá el corpus de entrenamiento en el caso del aprendizaje supervisado, o bien las instancias en análisis en el caso de las métricas de desambiguación.

3. UNA MEDIDA HÍBRIDA: FUNDAMENTOS DE LA PROPUESTA

Los sistemas actuales de desambiguación léxica automática utilizan recursos lingüísticos informatizados de diversa naturaleza técnica, ya sea como fuente de conocimiento endógena o exógena. No obstante, una pregunta crítica aún permanece sin respuesta en el ejercicio de comparar estos sistemas: si bien los algoritmos de aprendizaje automático más tradicionales, como *Naïve Bayes*, junto con las métricas de similitud semántica, logran resultados de desambiguación léxica similares, ¿por qué no alcanzan una concordancia perfecta o suficiente

como para declarar la resolución del problema de la desambiguación léxica automática de manera definitiva? Esto se debe, probablemente, a que todos los métodos computacionales diseñados para tareas de desambiguación automática realizan predicciones basándose en conjuntos de características previamente manipuladas.

Según lo anterior, los sistemas de desambiguación léxica automática serían altamente dependientes de las fuentes de conocimientos que utilizan, en el caso de los métodos de aprendizaje automático, y de los inventarios de sentidos, en el caso de los métodos de similitud semántica. En este contexto, la propuesta de una medida híbrida debe ser capaz de explotar el potencial de un método que integre la medida de la distancia entre rutas, por un lado, y la medida del contenido de información, por otro, con una fuente de conocimiento o inventario de sentidos. Esta perspectiva estaría basada en la disponibilidad de recursos léxicos y de conocimiento tanto lingüístico como conceptual, como lo es FunGramKB, que contiene la información necesaria para eliminar la ambigüedad a partir de sus representaciones conceptuales. Esta aproximación sería menos restrictiva que los enfoques de aprendizaje automático, por ejemplo, pues no dependería de la generalización de patrones para grandes corpus, o de la consistencia que alcancen quienes etiqueten un corpus de entrenamiento.

En cuanto a la fundamentación desde una perspectiva psicolingüística, Moreno-Sandoval (1998), basado en los aportes de Bob & Scha (1996), propone tres argumentos a favor de los métodos basados en datos y su relación con los procesos cognitivos de procesamiento lingüístico:

a. Los hablantes registran frecuencias de uso y diferencias entre frecuencias.
b. Los hablantes prefieren los análisis que ya han experimentado a análisis que tienen que construir por primera vez.
c. La preferencia está influida por la frecuencia de aparición de los análisis, de manera que prefieren los análisis más frecuentes a los menos frecuentes.

Luego, las características que debe tener esta propuesta de medida son las siguientes:

a. Debe ser una medida aplicada a tareas de procesamiento de datos textuales, coherente con la definición de Rada *et al*. (1989), en la que se establece como una función matemática, basada en la medición de la distancia conceptual. Específicamente, se trata de la medición de la distancia geométrica de puntos que representan conceptos o información conceptual.
b. Las propiedades que debe satisfacer una función $f(x, y)$ son las siguientes:
    i. $f(x, x) = 0$; la distancia entre los puntos es cero, solo si los puntos son iguales.
    ii. $f(x, y) = f(y, x)$; la función es simétrica.
    iii. $f(x, y) \geq 0$; la distancia entre los puntos es positiva (también llamada no-negatividad).
    iv. $f(x, y) + f(y, z) \geq f(x, z)$; se aplica la desigualdad triangular para espacios vectoriales.

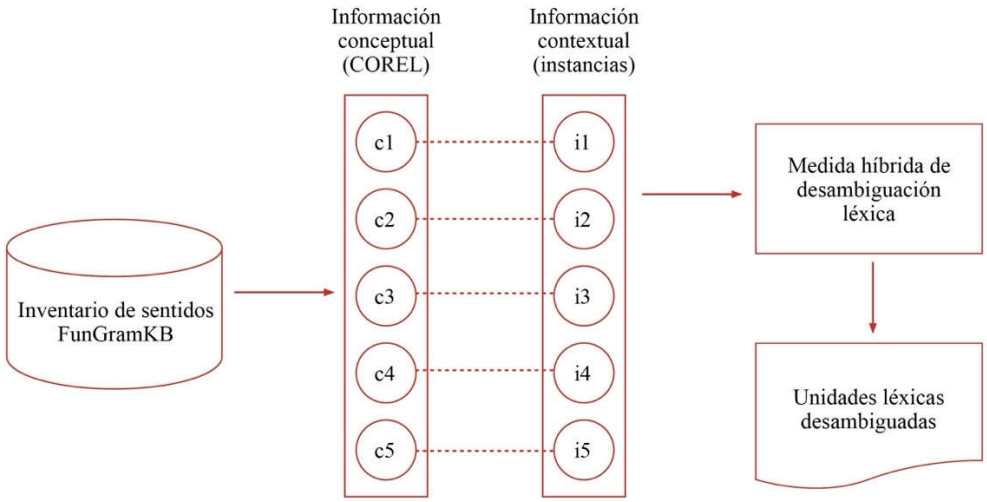
c. Debe ser evaluable matemáticamente, y devolver un valor cuyo rango se encuentre entre [0, 1], como valor normalizado.

d. Debe facilitar la comparación entre dos conceptos.

e. Debe poder comparar unidades léxicas que pertenezcan a diferentes partes de la oración, como sustantivos, verbos y adjetivos, cuyos correlatos conceptuales corresponden a entidades, eventos y cualidades en FunGramKB.

En consecuencia, el modelo de desambiguación propuesto considera, en primer lugar, la base de conocimiento FunGramKB como un inventario de sentidos aplicable a la evaluación de un sistema de desambiguación léxica automática basado en una medida de similitud semántica. En segundo lugar, la desambiguación se basa en una comparación entre la información contextual, que contiene instancias en las que aparece una palabra objetivo, y la información conceptual provista por relaciones taxonómicas en FunGramKB y los postulados de significado que caracterizan cada sentido potencial de la palabra por medio de información lingüística. En tercer lugar, se aplica una medida híbrida de desambiguación, que integra el enfoque basado en la distancia entre rutas y el enfoque basado en el contenido de información. Finalmente, es posible obtener un resultado numérico, dada una función f (c_i, c_j), que establecerá el puntaje para la similitud de dos palabras-objetivo puestas en un contexto oracional específico. Este modelo de desambiguación automática se ilustra como diagrama de flujo en la Figura 1.

FIGURA 1. MODELO DE DESAMBIGUACIÓN AUTOMÁTICA



Los postulados de significado en FunGramKB constituyen representaciones del conocimiento semántico independientes de cualquier lengua natural. Específicamente, cada metaconcepto, concepto básico o concepto terminal de la ontología se asocia a un postulado de significado, que a su vez se encuentra construido en el lenguaje de representación conceptual

COREL (Conceptual Representation Language), cuyo propósito es formalizar las relaciones semánticas y cognitivas entre unidades conceptuales (Periñán-Pascual y Mairal-Usón, 2010).

Nuestro modelo de desambiguación léxica automática integra como componente central una medida híbrida, cuyo objetivo es asignar un valor numérico para la función $f(c_i, c_j)$, equivalente a la interacción entre la información contextual y aquella que se puede extraer desde la exploración taxonómica, considerando a la base de conocimiento FunGramKB como inventario de sentidos.

El primer componente que considerar es la medida de la distancia entre rutas (*path base*, en adelante PB), basado en la propuesta de Leacock & Chodorow (1998). En la exploración taxonómica es relevante integrar la perspectiva de la relación semántica basada en PB, entendida como una función entre la longitud de la ruta que relaciona dos conceptos (*IS-A*), y la posición de esos conceptos en la taxonomía. Esta integración será particularmente importante considerando la jerarquía conceptual específica para cada una de las subontologías de la base de conocimiento. La ecuación es la siguiente:

$$PB_{Leacock\ \&\ Chodorow}(c_i, c_j) = -\log\left(\frac{len(c_i, c_j)}{2 \times depth\_max}\right)$$

la cual devuelve un valor que establece la cercanía de los nodos entre dos conceptos en la taxonomía. Para incluirla como parte de nuestra medida, es necesario determinar dos valores:

i. $len(c_i, c_j)$, equivale a la longitud del camino más corto entre los conceptos c_i y c_j de FunGramKB.

ii. *depth_max = 14*, equivale al valor máximo de profundidad de la taxonomía. En el caso de la subontología #ENTITY, este número es un valor constante para los niveles taxonómicos en FunGramKB.

El segundo componente para la medida híbrida corresponde a la medida del contenido de información. Uno de los problemas más relevantes que considera esta propuesta es que las métricas basadas en IC, dado que se establecen a partir WordNet como inventario de sentidos, no consideran únicamente la relación taxonómica *IS-A*. Por tanto, en esos casos no siempre se podrá identificar un ancestro común más cercano. En efecto, una de las críticas que plantean Gangemi *et al*. (2003), es que los enlaces léxicos de WordNet son de naturaleza heterogénea, aunque la mayoría de ellos corresponden a la subsunción habitual *IS-A*. Por el contrario, en las subontologías de FunGramKB, dado que utiliza exclusivamente la relación taxonómica *IS-A*, no siempre se podrá establecer un ancestro común más cercano que sea relevante semánticamente. Esto se debe a que se entiende la ponderación de la relación semántica entre dos unidades léxicas como un proceso de comprensión, en el que interviene un proceso inferencial. Sin embargo, aunque esta relación de subsunción provoca que no se puedan conectar directamente las cualidades con las entidades o con los eventos, sí se pueden relacionar a través de los postulados de significado.

Específicamente, la máquina debe encontrar una inferencia necesaria o vínculo perdido[4]. Según Croft y Cruse (2008), se puede afirmar que "[…] para que *X es un tipo de Y* sea aceptable, basta con que los *Xs* prototípicos se hallen dentro de los límites categoriales de *Y*. Esto supone asumir que *X* e *Y* son elementos léxicos y que denotan categorías conceptuales fijas" (p. 194). Adicionalmente, se establece que un enunciado hiponímico se juzgaría como aceptable si existen conceptualizaciones de *X* y de *Y* fácilmente aceptables y basadas en interpretaciones contextuales no anómalas. Esta inferencia necesaria es fundamental para poner en relación las dos unidades léxicas en análisis, puesto que establece un tránsito desde la palabra explícita (oída o leída), hasta el conocimiento tácito. Por ejemplo, si se consideran las unidades léxicas «manzana» y «pera», el vínculo perdido es que 'manzanas y peras pertenecen a la categoría fruta'. es decir, se trata de una relación *IS-A*. Por otra parte, si se consideran las unidades léxicas «automóvil» y «viaje», el vínculo perdido corresponde al hecho de que 'el automóvil es uno de los medios de transporte que se puede utilizar para viajar'; es decir, se establece una relación semántica más cercana a un tipo de clase (llamada también taxonimia). Según lo anterior, en principio se consideró la propuesta adaptada de Seco *et al*. (2004), en la que se establece una ecuación capaz de determinar un valor para el IC en tanto que, cuanto más abajo se encuentre un concepto en la taxonomía, mayor será el valor para el IC con respecto a sus conceptos superordinados en la ruta *IS-A*. Este valor demostraría un potencial favorable para establecer el vínculo perdido entre dos conceptos:

$$IC_{Seco\ et\ al.}(c) = 1 - \frac{\log\left(hypo(c) + 1\right)}{\log\left(\mathrm{C}\right)}$$

donde *hypo*(*c*) establece el número de hiperónimos de *c* y *C* corresponde al número total de conceptos que se encuentran en la dimensión metacognitiva a la que pertenece *c*. No obstante, en coherencia con el potencial del contenido conceptual alojado en la ontología de la base de conocimiento FunGramKB, y sus diferencias con la taxonomía de WordNet, en la que se basan propuestas como la de Seco *et al*. (2004), es necesario seleccionar una ecuación para establecer el valor de IC, que considere que la estructura taxonómica está organizada y regulada por determinados principios a partir de los cuales los hipónimos proporcionan información conceptual más específica que los hiperónimos, al mismo tiempo que es posible explotar las relaciones de herencia e inferencia para la información conceptual. Según esto, el trabajo de Zhou *et al*. (2008), que toma como antecedente la propuesta Seco *et al*. (2004), propone una medida que supera el método establecido como convencional para obtener el IC, en el que se establece una función entre el conocimiento de la estructura taxonómica y las instancias derivadas de un corpus. La nueva ecuación considera la información que heredan los hipónimos, junto con la profundidad de la estructura taxonómica. Así, el IC se establece como un valor más dependiente del inventario de sentidos que del contexto oracional en el que aparece la palabra objetivo. Los resultados de esta medida presentan, entonces, la ventaja de que no está basados en el análisis de las características del corpus y, por tanto, es posible evitar el problema de la escasez o la imprecisión de los datos textuales disponibles. En definitiva, la

[4] También llamado *enlace omitido* desde el marco conceptual de la semántica cognitiva.

medida para establecer el IC, adaptada de Zhou *et al*. (2008), es la siguiente:

$$IC_{Zhou\ et\ al.}(c) = k(1 - \frac{\log(hypo(c) + 1)}{\log(node\_max)}) + (1 - k)(\frac{\log(depth(c) + 1)}{\log(depth\_max)})$$

Por una parte, se condiera el logaritmo del número de hipónimos de *c* (es decir, de cada uno de los conceptos en análisis) dividido por el número máximo de conceptos en la taxonomía. Por otra parte, se considera el logaritmo de la profundidad del concepto *c* en relación con el nodo raíz, dividido por el logaritmo de la profundidad máxima de la taxonomía, como valor constante. En cuanto a los componentes, se integra el parámetro *k*, de adaptación manual, mediante el cual es posible ajustar el peso para cada uno de los dos componentes de la ecuación.

En definitiva, la ventaja de una propuesta híbrida basada en las funciones matemáticas propuestas por Leacock & Chodorow (1998) y por Zhou *et al*. (2008) es que facilitan la interacción de dos componentes para el cálculo de un valor numérico establecido como una función entre PB e IC. Para integrar cada uno de estos componentes de la medida es necesario normalizar sus valores. Según Han *et al*. (2012), la normalización de valores tiene el objetivo de entregar a todas las variables de la medida el mismo peso estadístico. De esta manera, se espera evitar la dependencia derivada de la elección de una u otra unidad de medida, lo que provocaría un mayor peso o efecto de una determinada variable. En términos matemáticos, el procedimiento consiste en la transformación de los datos para que se encuentren dentro de un rango común, en este caso [0, 1].

Según lo anterior, para el caso de la medida híbrida propuesta, y siguiendo la metodología de Periñán-Pascual (2015), utilizaremos una técnica de normalización de valores en la que se establece una transformación lineal de los datos originales, pero conservando sus relaciones de magnitud. Esta normalización, entonces, se expresará como *norm*(*PB*(c_i, c_j)) para el componente de distancia entre rutas y *norm*(*IC*(*c*)) para el componente de contenido de información, respectivamente:

$$norm\left(PB\left(c_i, c_j\right)\right) = 1 - \frac{1}{log_2(2 + PB\left(c_i, c_j\right))}$$

$$norm\left(IC(c)\right) = 1 - \frac{1}{log_2(2 + \left(IC(c)\right))}$$

Posteriormente, se integran a la medida los parámetros de optimización α y β, con el objetivo de evaluar el impacto que tendría cada componente. De acuerdo con Daelemans & Hoste (2002), la optimización de parámetros se define como el proceso mediante el cual se ajustan los valores de los atributos de una ecuación en un sistema de PLN, con el objetivo de configurar las condiciones óptimas para el desempeño en una tarea particular. En el caso de nuestra propuesta de medida híbrida, se integrarían dos parámetros de optimización, donde $\alpha + \beta = 1$, de tal

manera que el valor resultante todavía sería un valor normalizado. Finalmente, la propuesta de similitud semántica basada en una medida híbrida es la que se presenta a continuación:

$$SIM_{híbrida}(c_i, c_j) = \alpha\ (norm\left(PB(c_i, c_j)\right)) + \beta\ ((norm(IC(c_i)) + \ (norm\left(IC(c_j)\right))$$

En definitiva, se trata de una propuesta de medida híbrida que está basada en la medición del IC como una dimensión que permite una comprensión eficiente de la semántica de un concepto y su relación con su ámbito metaconceptual, a partir de la estimación de su grado ya sea de generalidad o de concreción. Esta aproximación considera que la única relación taxonómica válida en FunGramKB, y que permite establecer el valor de PB, es la subsunción (a diferencia de otros recursos similares).

4. Evaluación

4.1. Experimento

A continuación, exponemos tres casos para el cálculo de la similitud semántica, correspondientes a cada una de las tres unidades léxicas en análisis. Estos casos permitirán evaluar el potencial de la medida híbrida de similitud. El criterio para la selección, tanto de unidades léxicas como de sus correlatos conceptuales en la base de conocimiento FunGramKB, fue la verificación de que su contenido conceptual estuviese incluido en la ruta taxonómica para cada uno de los sentidos de las unidades léxicas «cabeza», «cara» y «carta» junto con sus correlatos conceptuales en la base de conocimiento, respectivamente, y que además se pudiese clasificar objetivamente a partir de un sentido en particular.

Así, en primer lugar, para «cabeza» seleccionamos como hiperónimo la unidad léxica «órgano», cuyo significado está capturado por el postulado de significado correspondiente al concepto básico +EXTERNAL_ORGAN_00, definido a su vez como 'un órgano que está situado sobre o cerca de la superficie del cuerpo':

+(e1: +BE_00 (x1: +EXTERNAL_ORGAN_00)Theme (x2: +BODY_PART_00)Referent)
*(e2: +BE_02 (x1)Theme (x3: +BODY_00)Location (f1: +OUT_00)Position)

En segundo lugar, para el caso de «cara» hemos escogido como hiperónimo la unidad léxica «superficie», que se define como 'el límite exterior bidimensional extendido de un objeto tridimensional'. El significado correspondiente está capturado por el postulado de significado del concepto básico +SURFACE_00:

+(e1: +BE_00 (x1: +SURFACE_00)Theme (x2: +PARASITIC_PLACE_00)Referent)
+(e2: +BE_02 (x1)Theme (x3: +PLACE_00 ^ +NATURAL_OBJECT_00 ^
+ARTIFICIAL_OBJECT_00)Location)

En tercer lugar, para el caso de «carta» seleccionamos como hiperónimo la unidad léxica «documento», cuyo significado está capturado por el postulado de significado del concepto +DOCUMENT_00, correspondiente a un 'tipo de escrito que contiene información':

+(e1: +BE_00 (x1: +DOCUMENT_00)Theme (x2: +WRITING_00)Referent)
+(e2: +CONTAIN_00 (x3: +INFORMATION_00)Theme (x1)Location)

Finalmente, el objetivo de la evaluación es determinar el valor más alto de similitud semántica entre los pares (*cabeza*$_{\text{head}}$, *órgano*), (*cabeza*$_{\text{intelligence}}$, *órgano*), (*cabeza*$_{\text{chief}}$, *órgano*), y (*cabeza*$_{\text{leader}}$, *órgano*), correspondientes a los sentidos de la unidad léxica «cabeza»; entre los pares (*cara*$_{\text{face}}$, *superficie*) y (*cara*$_{\text{side}}$, *superficie*), que constituyen los sentidos de la unidad léxica «cara»; y entre los pares (*carta*$_{\text{letter}}$, *documento*), (*carta*$_{\text{card}}$, *documento*), y (*carta*$_{\text{menu}}$, *documento*), que conforman los sentidos de la unidad léxica «carta», aplicando la medida de similitud híbrida que hemos propuesto. Para cada una de las evaluaciones de los casos en análisis, se incorporaron los siguientes valores de α y β como parámetros de optimización:

TABLA 1. VALORES PARA LOS PARÁMETROS DE OPTIMIZACIÓN α Y β.

Comb.	**Valores α**	**Valores β**
1	0	1
2	0,1	0,9
3	0,2	0,8
4	0,3	0,7
5	0,4	0,6
6	0,5	0,5
7	0,6	0,4
8	0,7	0,3
9	0,8	0,2
10	0,9	0,1
11	1	0

Los parámetros de optimización se definen como una selección de valores que mejoran el desempeño de un sistema de PLN en un aspecto específico. Cada parámetro representa un determinado peso de una variable en la medida, lo que en definitiva provocará un sesgo tanto

en la interpretación de los datos como en el funcionamiento de la medida. Este sesgo permitirá determinar cuáles son las variables que más influyen en la eficiencia del sistema de desambiguación. En este sentido, los trabajos de Hoste *et al.* (2002), Van Gompel & Van Den Bosch (2013) y Lu *et al*. (2019), entre otros, han demostrado que la inclusión de parámetros de optimización para modificar las variables disponibles en algoritmos de aprendizaje automático produce mejoras significativas en la precisión de las clasificaciones que puede realizar un sistema de desambiguación léxica automática.

En el caso de nuestra medida híbrida, los parámetros de optimización le otorgarán, para cada una de las combinaciones expuestas en la Tabla 1, un peso específico ya sea al valor del contenido de información o al valor de la distancia entre rutas. Esto conducirá a determinar los valores a partir de los cuales la medida obtiene resultados más eficientes.

4.2. Evaluación de similitud semántica para el caso «órgano» y los sentidos de «cabeza»

El objetivo de este caso es determinar la similitud semántica más alta entre los pares (*cabeza*$_{\text{head}}$, *órgano*), (*cabeza*$_{\text{intelligence}}$, *órgano*), (*cabeza*$_{\text{chief}}$, *órgano*), y (*cabeza*$_{\text{leader}}$, *órgano*) mediante la comparación de los valores de $SIM_{\text{híbrida}}(c_i, c_j)$, considerando las variables que constituyen tanto la distancia entre rutas como el contenido de información para los conceptos +HEAD_00 ('como parte superior del cuerpo en humanos, o anterior en algunas especies animales, en la que típicamente se encuentra el cerebro, y que podría considerar además la parte superior de una cosa'), +INTELLIGENCE_00 ('como la habilidad para elaborar pensamientos o para ejercer el juicio de manera acertada'), +CHIEF_00 ('como persona que preside, gobierna o está formalmente a cargo de un grupo, comunidad o corporación'), +LEADER_00 ('Como persona que dirige, inspira o es el punto de referencia para otras personas') y +EXTERNAL_ORGAN_00. Según lo anterior, la hipótesis que se intentará comprobar es que, dado el campo semántico de la unidad léxica «cabeza», el sentido (*cabeza*$_{\text{head}}$) sería el que alcance el puntaje de similitud más alto tras la aplicación del modelo de desambiguación léxica automática basado en nuestra medida híbrida.

Se aplicó para los pares en análisis la medida $SIM_{\text{híbrida}}(c_i, c_j)$, considerando cada una de las combinaciones para los parámetros de optimización (α y β). Los resultados se exponen en la Tabla 2:

TABLA 2. RESULTADOS DE APLICACIÓN DE LA MEDIDA HÍBRIDA PARA $SIM_{\text{HÍBRIDA}}$(*CABEZA, ÓRGANO*)

Comb.	(cabeza$_{\text{head}}$, órgano)	(cabeza$_{\text{intelligence}}$, órgano)	(cabeza$_{\text{chief}}$, órgano)	(cabeza$_{\text{leader}}$, órgano)
1	0,899	0,889	0,887	0,876
2	0,853	0,848	0,817	0,807
3	0,807	0,806	0,746	0,737
4	0,761	0,765	0,676	0,668
5	0,715	0,723	0,605	0,598
6	0,670	0,682	0,535	0,529
7	0,714	0,729	0,553	0,547
8	0,578	0,599	0,394	0,390
9	0,532	0,557	0,323	0,321

10	0,486	0,516	0,253	0,251
11	0,440	0,474	0,182	0,182
Prom.	0,678	0,690	0,543	0,537

Según los resultados expuestos en la Tabla 2, el puntaje de $SIM_{\text{híbrida}}(cabeza_{\text{head}}, órgano)$ promedio [X = 0,678; DE = 0,152] presenta sus valores más altos en las combinaciones uno [= 0,899], dos [=0,853] y tres [=0,807]. Estas combinaciones incorporan un peso nulo o mínimo para la variable de la distancia entre rutas. Por otra parte, el puntaje promedio para $SIM_{\text{híbrida}}(cabeza_{\text{intelligence}}, órgano)$ se convirtió en el resultado más alto para esta evaluación [X = 0,690; DE = 0,138] y presentó, en la mayoría de sus combinaciones, valores en los que el peso de la distancia entre rutas fue el más preponderante.

4.3. Evaluación de similitud semántica para el caso «superficie» y los sentidos de «cara»

El objetivo de este caso en evaluación es determinar la similitud semántica más alta entre los pares ($cara_{\text{face}}$, *superficie*) y ($cara_{\text{side}}$, *superficie*) mediante la comparación de los valores de $SIM_{\text{híbrida}}(c_i, c_j)$, considerando las variables que constituyen tanto la distancia entre rutas como el contenido de información para los conceptos +FACE_00 (´como la parte frontal de la cabeza humana, en la que se perciben órganos como la boca, nariz, ojos, etc.'), +SIDE_00 ('como la superficie de una cosa o de un lugar´) y +SURFACE_00. Así, la hipótesis a demostrar será que, dado el campo semántico de la unidad léxica «cara», el sentido ($cara_{\text{side}}$), alcanzará el puntaje de similitud más alto tras la aplicación del modelo de desambiguación léxica automática basado en nuestra medida híbrida. Los resultados, considerando las combinaciones para los parámetros de optimización α y β, fueron los siguientes:

TABLA 3. RESULTADOS DE APLICACIÓN DE LA MEDIDA HÍBRIDA PARA $SIM_{\text{HÍBRIDA}}(CARA, SUPERFICIE)$

Comb.	**($cara_{\text{face}}$, *superficie*)**	**($cara_{\text{side}}$, *superficie*)**
1	0,896	0,870
2	0,846	0,827
3	0,796	0,784
4	0,746	0,741
5	0,696	0,698
6	0,646	0,655
7	0,685	0,699
8	0,545	0,569
9	0,495	0,526
10	0,445	0,483
11	0,395	0,440
Prom.	0,654	0,663

La Tabla 3 muestra que los valores de similitud semántica para (*cara*$_{\text{side}}$, *superficie*) presentan un promedio de [X = 0,663; DE = 0,141]. El grupo de las combinaciones cinco a once fue consistentemente más alto, alcanzando su valor máximo en la combinación cinco [= 0,698], correspondiente a los valores de [α = 0,4] para la distancia entre rutas y [β = 0,6] para el contenido de información. No obstante, se observa que los valores seis a once corresponden a la asignación del parámetro de optimización cuyo mayor peso estuvo en la variable de la distancia entre rutas. Este caso en particular resultó problemático debido a que tanto el concepto +FACE_00 como el concepto +SIDE_00 presentan al concepto +SURFACE_00 como hiperónimo común más cercano. Esto provoca que, al asignarle mayor peso al parámetro de PB siempre se seleccione el sentido +FACE_00, y al establecer un mayor peso al parámetro de IC, se seleccione el sentido +SIDE_00. Según lo anterior, se presenta la dificultad de que el puntaje final de la medida se vea afectado significativamente por la presencia o ausencia del hiperónimo común más cercano, dado que esta información se obtiene desde el conocimiento intrínseco en FunGramKB.

4.4. Evaluación de similitud semántica para el caso «documento» y los sentidos de «carta»

El objetivo de este ejemplo es determinar la similitud semántica más alta entre los pares (*carta*$_{\text{letter}}$, *documento*), (*carta*$_{\text{card}}$, *documento*) y (*carta*$_{\text{menu}}$, *documento*) mediante la comparación de los valores de $SIM_{\text{híbrida}}(c_i, c_j)$, considerando las variables que constituyen tanto la distancia entre rutas como el contenido de información para los conceptos +LETTER_00 (´como una misiva o papel escrito que una persona envía a otras personas'), +CARD_00 (´como una cartulina rectangular con números y dibujos que se utiliza para juegos´), $MENU_00 (´como una cartulina rectangular con números y dibujos que se utiliza para juegos´) y +DOCUMENT_00. La hipótesis que se intentará comprobar en este ejemplo es que, dado el campo semántico de la unidad léxica «carta» en su sentido (*carta*$_{\text{letter}}$), será este último el sentido que alcance el puntaje de similitud más alto tras la aplicación del modelo de desambiguación automática basado en nuestra medida híbrida. Los resultados se muestran a continuación:

TABLA 4. RESULTADOS DE APLICACIÓN DE LA MEDIDA HÍBRIDA PARA $SIM_{\text{HÍBRIDA}}$(*CARTA*, *DOCUMENTO*)

Comb.	**(*carta*$_{\text{letter}}$, *documento*)**	**(*carta*$_{\text{card}}$, *documento*)**	**(*carta*$_{\text{menu}}$, *documento*)**
1	0,933	0,941	0,929
2	0,884	0,871	0,862
3	0,834	0,801	0,795
4	0,785	0,731	0,728
5	0,736	0,661	0,661
6	0,687	0,591	0,594
7	0,731	0,615	0,619
8	0,588	0,451	0,459

9	0,539	0,381	0,392
10	0,489	0,311	0,325
11	0,440	0,241	0,258
Prom.	0,695	0,600	0,602

En cuanto a los resultados de la Tabla 4, se observa que el valor promedio más alto para los valores de similitud corresponde a (*carta*$_{\text{letter}}$, *documento*), con un [X = 0,695; DE = 0,163]. En este caso, los valores de similitud para cada combinación fueron consistentemente más altos desde la combinación dos hasta la once, alcanzando su resultado más eficiente en la combinación dos [= 0,884], correspondiente a los valores de [α = 0,1] para la distancia entre rutas, y [β = 0,9] para el contenido de información. Cabe señalar que los valores más altos se encontraron precisamente en las combinaciones dos a cinco, que presentaban un mayor peso en el parámetro correspondiente al contenido de información.

4.5. Discusión de los resultados

Como se estableció en la sección anterior, la propuesta de medida híbrida para el cálculo de la similitud semántica $SIM_{\text{híbrida}}(c_i, c_j)$ considera las relaciones de subsunción en la taxonomía de la base de conocimiento FunGramKB para determinar el valor de la distancia entre rutas y el contenido de información. Se trata de una medida que, si bien se fundamenta en la complementariedad de los valores de IC y PB, establece predominantemente el valor de IC como una dimensión que permite una comprensión eficiente de la semántica de un concepto y la relación con su ámbito metaconceptual, a partir de la estimación de su grado ya sea de generalidad o de concreción. En cuanto a los resultados, los valores más altos para los casos de las unidades en análisis en los que se alcanza la desambiguación según los sentidos esperados para cada una de las hipótesis propuestas, basados en la aplicación de la medida híbrida, corresponden a aquellos en los que se evidenció un mayor peso en el parámetro de optimización correspondiente a la variable IC. Esta variable, al ser complementada con el valor de PB como la medición de la especificidad de un concepto en cuanto a su significado, es capaz de describir de manera eficiente la naturaleza de la similitud entre dos conceptos. En efecto, la organización taxonómica de FunGramKB representa las unidades conceptuales como un sistema de relaciones que es consistente con la manera en la que los hablantes organizan su lexicón mental; es decir, es capaz de incorporar el conocimiento del mundo al conocimiento léxico. Esto es coherente con la preponderancia del valor del contenido de información en los resultados, dado que la similitud semántica como método aplicado a la desambiguación léxica se basa en la información léxico-semántica y en la relación taxonómica entre los sentidos disponibles de las unidades léxicas.

Según lo anterior, los resultados permiten determinar que las combinatorias más eficientes para la medida híbrida de similitud se encuentran, típicamente, en el intervalo de los valores uno a cinco, en los que el valor de PB α aumenta progresivamente desde 0, mientras que el valor de IC en β disminuye desde 1. Específicamente, es la combinatoria tres, correspondiente a los valores de *[α = 0,2]* para la distancia entre rutas y [β = 0,8] para el contenido de información,

la que obtiene un desempeño eficiente y más consistente para los casos en evaluación. Lo anterior se muestra en la Tabla 4 para los pares (*cabeza*$_{\text{head}}$, *órgano*), (*cara*$_{\text{side}}$, *superficie*) y (*carta*$_{\text{letter}}$, *documento*):

TABLA 5. RESULTADOS DE APLICACIÓN DE $SIM_{\text{HÍBRIDA}}(C_I, C_J)$ PARA LA COMBINACIÓN TRES.

Comb. Tres	**(*cabeza*$_{\text{head}}$, *órgano*)**	**(*cara*$_{\text{side}}$, *superficie*)**	**(*carta*$_{\text{letter}}$, *documento*)**
$[\alpha = 0{,}2; \beta = 0{,}8]$	0,807	0,784	0,834

Finalmente, la medida híbrida propuesta ha mostrado ser eficiente desde el punto de vista de los métodos de similitud semántica, considerando el conocimiento lingüístico que provee la organización taxonómica de la base de conocimiento FunGramKB.

5. CONCLUSIONES

El objetivo general que hemos planteado para esta propuesta es desarrollar un modelo más robusto de medida para la similitud y relación semántica que los disponibles actualmente para la desambiguación léxica automática aplicado al PLN. Así, hemos propuesto una medida híbrida para la evaluación de la similitud semántica, basada en la interacción de dos enfoques de exploración taxonómica: distancia entre rutas (Leacock y Chodorow, 1998) y contenido de información (Zhou *et al*., 2008), considerando la arquitectura cognitiva de la base de conocimiento FunGramKB (Periñán-Pascual & Mairal-Usón, 2010) como un inventario de sentidos que permite explotar dicha interacción.

No obstante, es relevante considerar que la descripción del fenómeno de la ambigüedad léxica implica la interacción de distintos niveles de análisis lingüístico en el proceso de desambiguación: léxico, sintáctico, semántico y/o fonológico. Según esto, y desde una perspectiva formal, el proceso en el que el hablante es capaz de determinar el sentido específico de una unidad léxica en un contexto oracional particular puede tratarse como una integración modular de información, en la que el procesamiento de un nivel afectará el procesamiento de otros niveles adyacentes.

A partir de lo anterior, hemos abordado el objetivo específico de representar formalmente un procedimiento computacional para poder resolver la desambiguación léxica automática, basado en la propuesta de una medida de desambiguación híbrida. Nuestro modelo de desambiguación léxica automática, cuyo componente central es esta medida híbrida, ha cumplido con tres propiedades fundamentales: (1) considera la base de conocimiento FunGramKB como inventario de sentidos; (2) se basa en una comparación entre la información contextual, que contiene instancias en las que aparece una palabra objetivo, junto con la información conceptual provista por relaciones taxonómicas en FunGramKB y los postulados de significado que caracterizan cada sentido potencial de la palabra por medio de información lingüística; y (3) integra el enfoque basado en la distancia entre rutas y el enfoque basado en el contenido de información, a partir de los que es posible obtener un resultado numérico, dada

una función f (c_i, c_j), que establece el puntaje para la similitud de dos palabras objetivo puestas en un contexto oracional específico.

Posteriormente, hemos realizado la evaluación del modelo de desambiguación léxica automática, basado en la medida híbrida propuesta, para la unidad léxica «cabeza» y sus sentidos capturados en los postulados de significado de los conceptos +INTELLIGENCE_00, +HEAD_00, +LEADER_00, y +CHIEF_00; la unidad léxica «cara» y sus correspondientes sentidos capturados en los postulados de significado de los conceptos +FACE_00 y +SIDE_00; y la unidad léxica «carta», junto con sus sentidos capturados en los postulados de significado de los conceptos +LETTER_00, +CARD_00, y $MENU_00. Además, se integraron parámetros de optimización con el objetivo de configurar las condiciones óptimas para el desempeño de la tarea de desambiguación considerando el peso específico de cada una de las variables. En el caso de nuestra propuesta de medida híbrida, se integraron dos parámetros de optimización, donde $\alpha + \beta = 1$. Luego de este proceso, y considerando los resultados, la combinatoria correspondiente a los valores de [α = 0,2] para la distancia entre rutas y [β = 0,8] para el contenido de información, es la que obtuvo un desempeño eficiente y más consistente para los casos en análisis, donde la similitud semántica para c_i y c_j será igual a la suma entre el parámetro 0,2 multiplicado por el valor normalizado de la distancia entre rutas del par (c_i, c_j) y el parámetro 0,8 multiplicado por el valor normalizado del contenido de información de c_i más el valor normalizado del contenido de información de c_j. Esta medida híbrida ha mostrado ser eficiente desde el punto de vista de los métodos de similitud semántica, considerando el conocimiento lingüístico que provee la organización taxonómica de la base de conocimiento FunGramKB.

En este estudio se han presentado tres casos experimentales para la aplicación del modelo híbrido de desambiguación léxica automática. Aunque el número de casos analizados podría considerarse limitado y, por tanto, insuficiente para generalizar conclusiones desde el punto de vista estadístico, cada unidad léxica abordada ejemplifica un tipo de relación polisémica y de complejidad taxonómica en FunGramKB. Además, los resultados obtenidos en estos casos son consistentes con los alcanzados en experimentos adicionales desarrollados en el marco de la tesis doctoral del autor (Núñez-Torres, 2021), donde se aplicó la misma medida híbrida a un conjunto más amplio de unidades léxicas y sus correlatos conceptuales. Así, la coherencia entre ambos conjuntos de resultados refuerza la validez del modelo propuesto.

Finalmente, el modelo que hemos presentado ha puesto en evidencia que una medida basada en la información léxico-semántica y en relaciones taxonómica disponibles puede funcionar de manera eficiente cuando se proporciona una fuente de conocimiento apropiada y coherente con criterios de análisis lingüísticos. En definitiva, esta propuesta contribuye en el diseño de un modelo coherente tanto lingüística como matemáticamente, que es capaz de determinar el valor numérico que representa un puntaje de similitud semántica para la comparación entre la información tanto contextual como conceptual entre dos unidades léxicas, considerando la caracterización de cada sentido potencial de la palabra en análisis por medio de información lingüística.

AGRADECIMIENTOS

El desarrollo de esta investigación fue patrocinado por la Agencia Nacional de Investigación y Desarrollo (ANID) del Ministerio de Ciencia, Tecnología, Conocimiento e Innovación del Gobierno de Chile, en el marco del Programa de Formación de Capital Humano Avanzado (Beca de Doctorado Nacional 2016, folio Nº 21160361). Igualmente, esta publicación es parte del proyecto de I+D+i PID2023-147137NB-I00 financiado por MICIU/AEI/10.13039/501100011033 y por por FEDER, UE.

REFERENCIAS

Aussenac-Gilles, Nathalie, and Fabien Gandon. 2013. "From the knowledge acquisition bottleneck to the knowledge acquisition overflow: A brief French history of knowledge acquisition." *International Journal of Human-Computer Studies* 71 (2): 157-65. https://doi.org/10.1016/j.ijhcs.2012.10.009.

Buchanan, Bruce, and David Wilkins. 1993. *Readings in Knowledge Acquisition and Learning: Automating the Construction and Improvement of Expert Systems.* San Mateo, CA: Morgan Kaufmann.

Croft, William, and David Alan Cruse. 2008. *Lingüística cognitiva.* Madrid: Akal.
Daelemans, Walter, and Véronique Hoste. 2002. "Evaluation of machine learning methods for natural language processing tasks." In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria. http://www.lrec-conf.org/proceedings/lrec2002/pdf/94.pdf.

Gangemi, Aldo, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. 2003. "Sweetening WordNet with DOLCE." *AI Magazine* 24 (3): 13-24. https://doi.org/10.1609/aimag.v24i3.1715.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2012. *Data Mining: Concepts and Techniques.* 3rd ed. Waltham, MA: Morgan Kaufmann.

Hoste, Véronique, Iris Hendrickx, Walter Daelemans, and Antal van den Bosch. 2002. "Parameter optimization for machine learning of word sense disambiguation." *Natural Language Engineering* 8 (4): 311-25.

Leacock, Claudia, and Martin Chodorow. 1998. "Combining local context and WordNet similarity for word sense identification." In *WordNet: An Electronic Lexical Database*, edited by Christiane Fellbaum, 265-83. Cambridge, MA: MIT Press.

Lu, Wenpeng, Fanqing Meng, Shoujin Wang, Guoqiang Zhang, Xu Zhang, Antai Ouyang, and Xiaodong Zhang. 2019. "Graph-based Chinese word sense disambiguation with multi-knowledge integration." *Computers, Materials & Continua*61 (1): 197-212.

Moreno-Sandoval, Antonio. 1996. *Lingüística computacional.* Madrid: Síntesis.

Nevzorova, Olga, Alfiya Galieva, and Vladimir Nevzorov. 2015. "Sentence context and resolving lexical ambiguity for special groups of words on the base of corpus data." *Procedia – Social and Behavioral Sciences* 198: 359-66.

Núñez-Torres, Fredy. 2021. *Diseño y desarrollo de un modelo de desambiguación léxica automática para el procesamiento del lenguaje natural.* Tesis doctoral, Pontificia Universidad Católica de Chile.

Núñez Torres, Fredy, and María Beatriz Pérez Cabello de Alba. 2023. "Desarrollo de un sistema de aprendizaje automático supervisado para la desambiguación léxica automática utilizando DAMIEN (Data Mining Encountered)." *Revista Electrónica de Lingüística Aplicada* 21 (1): 150-78. https://doi.org/10.58859/rael.v21i1.504.

Pasini, Tommaso. 2020. "The knowledge acquisition bottleneck problem in multilingual word sense disambiguation." In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20).*

Patwardhan, Siddharth, Satanjeev Banerjee, and Ted Pedersen. 2003. "Using measures of semantic relatedness for word sense disambiguation." In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics.*

Periñán-Pascual, Carlos. 2012. "En defensa del procesamiento del lenguaje natural fundamentado en la lingüística teórica." *Onomázein* 26: 13-48.

Periñán-Pascual, Carlos. 2015. "The underpinnings of a composite measure for automatic term extraction: The case of SRC." *Terminology* 21 (2): 151-79.

Periñán-Pascual, Carlos, and Francisco Arcas-Túnez. 2007. "Cognitive modules of an NLP knowledge base for language understanding." *Procesamiento del Lenguaje Natural* 39: 197-204.

Periñán-Pascual, Carlos, and Francisco Arcas-Túnez. 2010. "The architecture of FunGramKB." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 2667-74. Valletta, Malta: European Language Resources Association (ELRA).

Periñán-Pascual, Carlos, and Ricardo Mairal-Usón. 2010. "La gramática de COREL: un lenguaje de representación conceptual." *Onomázein* 21: 11-45.

Rada, Roy, Hafedh Mili, and María Blettner. 1989. "Development and application of a metric on semantic nets." *IEEE Transactions on Systems, Man and Cybernetics* 19 (1): 17-30.

Seco, Nuno, Tony Veale, and Jer Hayes. 2004. "An intrinsic information content metric for semantic similarity in WordNet." In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*. https://doi.org/10.5555/3000001.3000272.

Van Gompel, Maarten, and Antal van den Bosch. 2013. "WSD2: Parameter optimisation for memory-based cross-lingual word-sense disambiguation." In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (SEM 2013)*, 183-87.

Wagner, Christian. 2006. "Breaking the knowledge acquisition bottleneck through conversational knowledge management." *Information Resources Management Journal* 19 (1): 70-83.

Zhou, Zili, Yanna Wang, and Junzhong Gu. 2008. "New model of semantic similarity measuring in WordNet." In *Proceedings of the Second International Conference on Future Generation Communication and Networking Symposia*.https://doi.org/10.1109/FGCNS.2008.4730937.