Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 277-289 recibido 05-12-2024 revisado 19-01-2025 aceptado 02-02-2025

Detección de terminologización en sustantivos

Detecting terminologization in nouns

Javiera Ahumada, Rogelio Nazar
Pontificia Universidad Católica de Valparaíso
javiera.ahumada.i@mail.pucv.cl, rogelio.nazar@pucv.cl

Resumen: Este artículo presenta una propuesta metodológica para detectar automáticamente la terminologización de sustantivos en un dominio especializado. La metodología se basa en la detección de alteraciones en los perfiles de coocurrencia de los sustantivos y es evaluada aplicándola a un corpus especializado compuesto por artículos de investigación en el ámbito del procesamiento del lenguaje natural. Si bien se trata de un trabajo aun preliminar, los resultados muestran que el método propuesto puede ser de utilidad, ya que permite identificar sustantivos con alta probabilidad de terminologización. La implementación es, por tanto, de interés para terminólogos e investigadores trabajando en el tema del cambio semántico. Además de una valoración del método, se señalan las limitaciones del enfoque y algunas ideas para su ulterior desarrollo.
Palabras clave: terminologización, cambio semántico, neología semántica, coocurrencia léxica, extracción terminológica.

Abstract: This paper presents a methodological proposal for the automatic detection of noun terminologization in specialized domains. The methodology is based on the detection of disruptions in the co-occurrence profiles of a target noun, and it is evaluated in a specialized corpus of research articles of the domain of natural language processing. Although this is still work in progress, results show that the proposed methodology can be useful, as it allows users to identify nouns that are likely to have been terminologized. The implementation may be of help for terminologists and researchers interested in semantic change. In addition to the assessment of the method, the paper presents its limitations and some ideas for future development.
Keywords: terminologization, semantic change, semantic neology, lexical co-occurrence, terminology extraction.

1 Introducción

El cambio lingüístico es un fenómeno inherente a las lenguas vivas, manifestándose en distintos niveles, incluido el semántico (Bloomfield, 1984; Ullmann, 1965). Dentro de este ámbito, la neología semántica ha atraído el interés de una gran cantidad de investigadores en las últimas décadas. Por ejemplo, la extracción de neologismos semánticos de forma automática o semiautomática (Santamaría, 2013; Torres Rivera, 2020), la clasificación de neologismos semánticos (Freixa, Luna, y Suárez de la Torre, 2010; Díaz Hormigo, 2020) y la neología semántica en ámbitos especializados (Estornell Pons y Soto, 2016; Berri, 2013; Adelstein, 2022, entre otros).

Un caso particular del fenómeno de la neología semántica es la terminologización, entendida como una forma de cambio semántico caracterizado por la aparición de un nuevo significado terminológico o especializado en un significante ya conocido (Martín, 2004; Sánchez, 2009; Roelcke, 2018). Por ejemplo, la palabra *semilla*, que en la lengua general refiere a un grano vegetal (Real Academia Española, 2024), se utiliza en informática para denotar un número empleado para inicializar un generador de números pseudoaleatorios en los lenguajes de programación (Kernighan y Ritchie, 1991).

En la actualidad, existen distintas herramientas para la extracción automática de terminología, aunque distintos estudios han señalado limitaciones tales como una variable tasa de error (Llanos, 2023; Arce y Seghiri, 2021), la diferente calidad del resultado según la disciplina e idioma con el que se trabaja (Santamaría y Krallinger, 2018) o

ISSN 1135-5948 DOI 10.26342/2025-74-19

la dificultad de la segmentación de unidades poliléxicas (Cabré, Estopà, y Vivaldi, 2001). Actualmente se está empleando para la extracción de unidades terminológicas herramientas como Sketch Engine (Kilgarriff et al., 2004), AntConc (Anthony, 2012) o Termostat (Drouin, 2003). Sin embargo, estas presentan limitaciones ya que son sistemas que se centran en el significante o forma de las palabras. Basan su método en la comparación de frecuencias de las palabras con un corpus de referencia, que se supone es representativo del uso no especializado de la lengua. Esto tiene como consecuencia que presenten dificultad para la detección de términos cuya forma coincide con una palabra de uso general.

Con la actual proliferación de corpus digitales y el desarrollo de herramientas computacionales, surgen nuevas oportunidades para abordar estas limitaciones, ya que es posible aplicar metodologías que integren análisis semántico. Con base en un trabajo previo (Ahumada, 2024), en el presente artículo proponemos un método automatizado para detectar la terminologización de sustantivos en un dominio de especialidad, utilizando como indicador de terminologización la comparación de coocurrencias entre un corpus especializado y uno de referencia, representativo del vocabulario general.

Para el desarrollo de esta investigación, el dominio de especialidad con el que se trabaja es el procesamiento del lenguaje natural (PLN), dado que es un campo abundante en neologismos, por efecto de la cantidad de innovaciones que se producen en la disciplina, y con abundantes ejemplos de neologismos por terminologización.

El desarrollo de este estudio representa contribuciones tanto teóricas como prácticas. En el plano teórico, la metodología propuesta puede ser útil para ampliar la comprensión del fenómeno de la terminologización al proponer una manera de formalizar la descripción. En el plano práctico, en tanto, el sistema descrito puede servir para facilitar la detección de terminologización en beneficio de investigadores interesados en el estudio del cambio semántico.

Los datos de los resultados y el código de la implemantación se encuentran libremente disponibles en el sitio web del proyecto[1].

[1]https://www.tecling.com/neuter

2 Trabajo relacionado

2.1 La investigación en neología

Las lenguas, aunque se presentan como una estructura estable de hábitos léxicos y gramaticales, experimentan un constante proceso de cambio lingüístico que se produce de manera gradual (Bloomfield, 1984). Entre los distintos niveles donde puede ocurrir el cambio lingüístico, el léxico representa uno de los ámbitos más dinámicos, siendo el área donde las transformaciones ocurren con mayor rapidez (Álvarez de Miranda, 2009; Nazar, 2023).

Estos cambios léxicos se manifiestan principalmente a través de dos fenómenos: la neología, que implica la incorporación de una unidad léxica nueva a una lengua (Cabré, 2008; Álvarez de Miranda, 2009; Díaz Hormigo, 2020) y la pérdida léxica –también llamada obsolescencia o muerte léxica– que consiste en la desaparición de algunos signos en una lengua (Seco, 1989; Álvarez de Miranda, 2009).

Múltiples investigaciones desarrolladas en torno a la neología han establecido diversas taxonomías para clasificar los neologismos. La distinción aparentemente más aceptada es la que establece las dos categorías de neología formal y neología semántica (Bastuji, 1974; Guilbert, 1974; Rey, 1976; Pottier Navarro, 1979; Guerrero Ramos, 1995; Cabré et al., 2002; Díaz Hormigo, 2007). Existen propuestas adicionales que complementan esta clasificación con la incorporación de otros tipos de neologismos, tales como el préstamo lingüístico (Auger y Rousseau, 1977), la neología funcional (Cabré, 1993) y la neología sintáctica (Cabré, 2006; Domènech, 2008).

Para el estudio de la terminologización, resultan relevantes la neología semántica y la especializada. La primera se define como el fenómeno mediante el cual aparece un nuevo significado en palabras formalmente conocidas (Guilbert, 1975; Pottier Navarro, 1979; Nazar, 2011; Díaz Hormigo, 2020; Renau, 2023), y se manifiesta principalmente a través de cuatro mecanismos: la metáfora, la metonimia, la especialización y la generalización (Bloomfield, 1984; Ullmann, 1965; Sablayrolles, 1997; Cabré, 2006; Geeraerts, 2010; Renau, 2023). Por su parte, la neología especializada –denominada también neonimia por Rondeau (1984)– se emplea para referirse a los elementos que representan una nueva noción en las lenguas de especialidad (Estopà,

2016), por tanto, es un tipo de neología propia de los discursos especializados (Fuentes et al., 2009).

2.2 Terminologización

La terminologización se encuentra en el proceso dinámico de cambio lingüístico, particularmente donde se manifiesta la interacción entre el léxico general y el de dominio especializado. El proceso de terminologización es un fenómeno mediante el cual una palabra del vocabulario general pasa a un léxico especializado. De este modo, se convierte en un término y, por tanto, adquiere un nuevo significado (Martín, 2004; Sánchez, 2009; Roelcke, 2018). Este cambio implica variación a un nivel semántico y pragmático del término, ya que se precisa o modifica su significado, aunque sea sutilmente (Sanz, 2008).

En el caso de la terminologización, los principales mecanismos lingüísticos que la producen son la metáfora y la metonimia. La metáfora opera estableciendo relaciones de semejanza entre conceptos del léxico común y el especializado (Holeš y Honová, 2023), destacándose por su carácter alusivo y evocador, así como también por facilitar el logro de mayor claridad en la exposición de un razonamiento (Barán, 1999). La metonimia, que ha sido menos estudiada que la metáfora en textos especializados (Berri y Bregant, 2015), opera por una relación de contigüidad (Holeš y Honová, 2023) y restricción denotativa (Martí, 2009).

Mantener una separación entre lo especializado y lo general en ocasiones suele representar una dificultad, puesto que ciertas palabras poseen significados especializados claros que pueden ser comprendidos por un amplio público, como se observa en el término *coronavirus* (Estopà, 2022). Esta difusión entre los límites de lo general y especializado se explica mediante procesos como la divulgación y popularización de la ciencia, ya que el conocimiento científico o disciplinar no es hermético, sino que se difunde hacia esferas más amplias de la sociedad (Cassany, 2007). Esta dinámica resalta la importancia del estudio de los términos en sus contextos comunicativos específicos y considerando su naturaleza evolutiva (Freixa, 2016) en la labor de su recopilación y atesoramiento (Sager, 1990).

2.3 Estudios sobre la detección del cambio semántico

En los últimos años, se ha percibido un aumento en el interés de parte de la comunidad académica por los métodos y herramientas computacionales que apoyan la investigación sobre el cambio semántico (Tahmasebi, Borin, y Jatowt, 2021). Este interés es impulsado por la disponibilidad de datos lingüísticos diacrónicos y el avance de tecnologías para identificar el cambio de significado de las palabras (Tang, 2018). Entre los principales enfoques metodológicos, Heyer et al. (2017) distinguen tres categorías: análisis de patrones y pistas lingüísticas, exploración del espacio semántico latente, y análisis de pertinencia temática.

Un método destacado es el basado en la coocurrencia, en donde se emplea esta información sobre las palabras que coocurren con una unidad léxica objetivo para determinar la fuerza de asociación entre ellas. Por ejemplo, en la investigación de Sagi, Kaufmann, y Clark (2009), se aplicaron vectores de contexto para detectar la generalización y especialización del significado en un corpus diacrónico mediante el análisis de densidad semántica. Tenemos también antecedentes de detección de neología semántica basada en la coocurrencia léxica y mediante técnicas de *clustering* basadas en grafos (Nazar y Vidal, 2010). Por su parte, Gulordava y Baroni (2011) desarrollaron un enfoque automático utilizando modelos de similitud distribucional con el corpus Google Books Ngram. Más recientemente, Gonen et al. (2021) propusieron una metodología que se caracteriza por su simplicidad, estabilidad e interpretabilidad, basada en la observación de la coocurrencia de palabras entre diferentes corpus para detectar posibles cambios semánticos en las unidades léxicas.

Las *word embeddings* representan otro enfoque interesante. Por ejemplo, Hamilton, Leskovec, y Jurafsky (2016) proponen una metodología que emplea tres técnicas: *positive point-wise mutual information* (PPMI), SVD y *skip-gram with negative sampling* (SGNS), y sus hallazgos establecieron dos tendencias: la ley de conformidad, donde las palabras más frecuentes tienden a cambiar más lentamente, y la ley de innovación, donde las palabras más polisémicas cambian más rápido. Sin embargo, este método presenta desafíos en la alineación de espacios vecto-

riales entre diferentes períodos. Para abordar esta limitación, Yao et al. (2018) desarrollaron un modelo de optimización conjunta que aprende simultáneamente las representaciones de palabras para distintos períodos temporales.

En cuanto al uso de técnicas como *topic modeling* y *clustering*, Montariol, Martinc, y Pivovarova (2021) emplearon técnicas de clustering mediante *embeddings* contextuales utilizando BERT, aunque señalan limitaciones en términos de escalabilidad y consumo de memoria y tiempo. Por otro lado, Frermann y Lapata (2016) propusieron SCAN, un modelo bayesiano dinámico que infiere tanto los sentidos como la prevalencia de las palabras en diferentes períodos, destacándose por su versatilidad en la detección de nuevos sentidos y la clasificación de cambios de significado.

Si bien estos métodos han demostrado ser efectivos para la detección del cambio semántico en general, no están exentos de dificultades y, además, son pocos los que están centrados en el fenómeno particular de la terminologización. Esto representa una oportunidad para desarrollar métodos automatizados que no solo detecten el fenómeno, sino que también proporcionen un análisis y comprensión más profundo de los cambios y asociaciones que experimentan las palabras al migrar a dominios especializados. De esta forma, se logra predecir el fenómeno mediante una clasificación basada en patrones de coocurrencia y esto, naturalmente, puede tener a su vez consecuencias prácticas para el mantenimiento de diccionarios o bases de datos terminológicas.

3 Metodología

Como ya se mencionó en la introducción, esta investigación aborda la detección de la terminologización proponiendo un método para identificar automáticamente cómo los sustantivos experimentan un cambio semántico al integrarse en un dominio específico de conocimiento. Para el desarrollo metodológico, se utilizaron tres recursos principales:

- **Corpus especializado (ESP_PLN)**: Corpus compuesto por un total de 902 textos completos (artículos de investigación) de la revista Procesamiento del Lenguaje Natural, entre los años 2003 y 2023. El corpus es representativo de la disciplina en lengua castellana, ya que la revista ha publicado desde 1983 y goza de alto prestigio académico.
- **Corpus de referencia (REF_PRE)**: Corpus de referencia de la lengua general conformado por 31.497 textos de prensa del periódico *El País* entre los años 2000 y 2002.
- **Lemario del Diccionario de la Lengua Española (DLE)**: Listado de 49.254 sustantivos monoléxicos que aparecen en este diccionario.

A continuación, se detallan las tres fases en las que se divide la metodología de investigación.

3.1 Definición del conjunto de unidades léxicas para el análisis

El propósito de esta fase consiste en establecer un conjunto de análisis que permita identificar los sustantivos potencialmente susceptibles de terminologización. Para ello, el proceso inició con el etiquetado morfosintáctico de los corpus ESP_PLN y REF_PRE utilizando UDPipe (Straka y Straková, 2017). Este paso permitió eliminar las variantes flexivas mediante lematización y seleccionar los sustantivos gracias al etiquetado de categorías gramaticales, lo que ofrece además la posibilidad de distinguir los tipos de palabra con los que coocurre la unidad léxica de interés, comparar resultados en función de dicha categoría o bien incluir o excluir categorías en el estudio.

El conjunto de unidades léxicas de interés se denota con el símbolo C y se define como la intersección entre V, que es el vocabulario del corpus ESP_PLN (específicamente, sustantivos monoléxicos en castellano), y el DLE (1).

$$C = V \cap DLE \tag{1}$$

A su vez, el conjunto C está compuesto por dos subconjuntos. Por un lado, un subconjunto hipotético T, que reúne las unidades que han experimentado terminologización (como, por ejemplo, *semilla*); y el subconjunto N, con aquellas unidades que no han sido terminologizadas (como, por ejemplo, el sustantivo *discrepancia*). En esencia, el problema que aquí se plantea es cómo separar el conjunto C en los dos subconjuntos N y T.

3.2 Criterios de filtrado de sustantivos

La metodología implicó la aplicación de diferentes filtros que permitan mejorar la calidad de los resultados obtenidos. El primer filtro se basa en un criterio de longitud del sustantivo de interés, ya que se descartan aquellos con tres caracteres o menos, dado que estas unidades cortas tienen baja probabilidad de ser útiles para el análisis. Este criterio aplica también para las palabras que cooccurren con un determinado candidato, hecho que se vincula con el criterio siguiente.

El segundo criterio de filtrado está asociado con la frecuencia del candidato. Como el método está basado en el análisis de la coocurrencia de cada sustantivo en los respectivos corpus, es condición necesaria que el candidato tenga una frecuencia de coocurrencia con otras palabras. Este filtro se aplica indirectamente, ya que mide la frecuencia mínima del elemento que con mayor frecuencia coocurre con el candidato evaluado: elimina cualquier candidato que no coocurra al menos 15 veces con otra palabra. Este criterio tiene una doble ventaja. La más evidente es que de este modo se filtran los candidatos de baja frecuencia, y la otra es que elimina también sustantivos que, aun siendo frecuentes, no están sintagmáticamente relacionados con otras palabras y, por ende, son poco informativos. Sería el caso de unidades funcionales de la lengua, como por ejemplo elementos que forman parte de marcadores discursivos y que no tienen interés terminológico.

Otro criterio de eliminación es el de los sustantivos que coinciden formalmente con palabras de alta frecuencia en inglés (como *once, quite, pose*, etc.). Existe, naturalmente, gran afluencia de palabras en inglés dado que la revista publica también en inglés y cada artículo, aunque esté escrito en castellano, incorpora títulos, resúmenes, palabras clave, fragmentos de texto de citas y títulos bibliográficos en inglés. Para obtener el listado de vocabulario de alta frecuencia del inglés se utilizó el corpus de inglés del *Projekt Deutscher Wortschatz* de la Universidad de Leipzig (Goldhahn, Eckart, y Quasthoff, 2012).

3.3 Obtención de un perfil de coocurrencia de las unidades léxicas analizadas

En esta segunda fase se diseñó un método para extraer las coocurrencias de las unidades que conforman el conjunto C. El perfil de coocurrencia de una unidad es lo que aquí se considera como indicador del posible cambio semántico, ya que la diferencia entre las palabras con las que suele coincidir una unidad léxica sería un indicio de cambio semántico y, para este caso, de terminologización.

El procedimiento consistió en extraer las palabras que acompañan a cada unidad i de C, tanto en el corpus ESP_PLN (E) como para REF_PRE (R). Esto permitió comparar las coocurrencias de las unidades léxicas (C_i) de cada corpus y, como se verá a continuación, evaluar si existe un cambio según el dominio en el que se encuentran.

Para refinar el análisis, se aplicó una lista de exclusión (*stoplist*) a estos listados de coocurrencias. La función de la *stoplist* consiste en descartar como coocurrentes palabras funcionales (artículos, preposiciones, conjunciones, entre otros) y así generar una comparación de las unidades que aporten un valor semántico para los dominios de interés.

3.4 Detección de alteraciones en el perfil de coocurrencia léxica

En esta última fase se realizó la comparación de las coocurrencias obtenidas en la fase anterior y se estableció una función que operacionaliza la terminologización indicando si una unidad léxica ha experimentado un cambio en su perfil de coocurrencia. Una vez generado el perfil de coocurrencia, se calculó la intersección entre n unidades que coocurren más frecuentemente con C_i, tanto para el corpus ESP_PLN (E) como el REF_PRE (R). En ambos casos, el análisis se realizó considerando las primeras n coocurrencias en orden decreciente de frecuencia ($E_i = \{e_{i,1}, e_{i,2}, e_{i,3}, \ldots, e_{i,n}\}$), con n fijado en 100. El resultado de este cálculo es el valor de coocurrencia r (2).

$$r(C_i) = |E_i \cap R_i| \tag{2}$$

El valor r obtenido es la clave para determinar si un candidato C_i ha sufrido un proceso de terminologización (es decir, $C_i \in T$) o no ($C_i \notin T \wedge C_i \in N$). Esta decisión puede tomarse con un umbral de corte k, con el cual es posible definir una función que permita aceptar o rechazar a C_i (3).

$$f(C_i) = \begin{cases} C_i \in T & r(C_i) \leq k \\ C_i \in N & \text{otherwise} \end{cases} \tag{3}$$

Más allá del umbral k, el valor r permite obtener como resultado el listado de sustantivos analizados ordenado de forma descendente con respecto a ese, resaltando de esa manera las unidades que tienen menor intersección. Esto permite, además, definir el umbral k de manera empírica.

4 Resultados

Los resultados obtenidos a partir de la aplicación del método automatizado se sintetizan en el diagrama de línea de la Figura 1, que representa la precisión acumulada en el listado ordenado de un total de 1.474 sustantivos monoléxicos que aparecen en el corpus especializado, dispuestos en el eje horizontal. Este es un subconjunto de C, que originalmente contiene 5.639 unidades, reducido después de la aplicación de los criterios de filtrado.

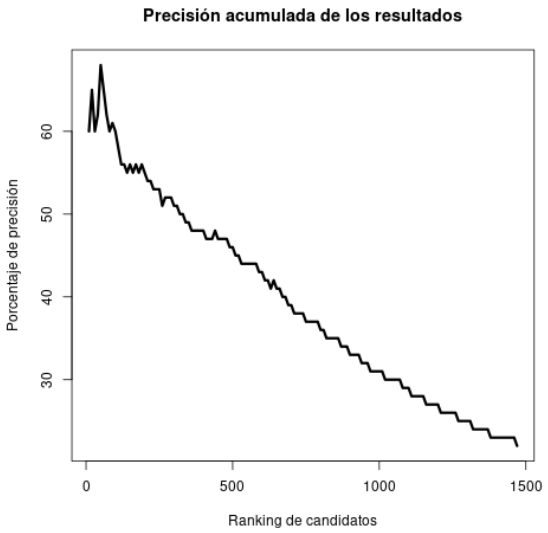

Figura 1: Precisión acumulada del ranking de sustantivos terminologizados.

El algoritmo diseñado para esta propuesta proporcionó como resultado el listado de sustantivos de análisis ordenado de forma descendente a partir del valor r. Esta disposición de los datos refleja que, mientras menor es el número de coincidencias entre las coocurrencias de un sustantivo, mayor es la probabilidad de que sea candidato a terminologización. En este sentido, el resultado expuesto en la Figura 1 permite observar una tendencia decreciente de los candidatos positivos a terminologización, puesto que el porcentaje de sustantivos verdaderos positivos a este proceso va disminuyendo en la medida que se avanza en el eje horizontal.

Cabe destacar la concentración de candidatos correctos a terminologización ubicados en las primeras posiciones del ranking. Los primeros sustantivos analizados se caracterizan por una precisión que supera el 60 %, lo que indica que el método propuesto logra detectar con relativa eficacia los sustantivos que tienen una mayor probabilidad de ser terminologizados. El porcentaje de verdaderos positivos a terminologización disminuye gradualmente en la medida que incrementa el valor r en el listado de sustantivos, lo que indica que la coocurrencia de un mismo sustantivo en diferentes contextos (corpus especializado y corpus de referencia) es efectivamente un indicador de cambio semántico, en este caso, de terminologización.

Este sistema de ranking de los resultados tiene implicancias para el estudio de palabras del léxico general que se desplazan hacia un lenguaje de especialidad. Por una parte, el orden descendente aporta a la comprensión de la dinámica del proceso de terminologización, en donde no solo se obtiene el número de coincidencias que tiene un sustantivo en distintas áreas de uso, sino que también se puede observar aquellas palabras con las que coocurre una unidad. Por otra parte, una presentación de mayor a menor potencial de terminologización de unidades léxicas representa un apoyo para los investigadores del fenómeno, quienes pueden enfocar su esfuerzo e interés en la revisión y validación manual de los elementos concentrados en los primeros puestos para identificar a los sustantivos con mayor potencial de terminologización.

4.1 Análisis de resultados

El método propuesto permitió identificar distintos sustantivos que han experimentado un proceso de terminologización en el dominio del PLN. En la Tabla 1 se presentan los primeros 30 sustantivos del ranking, junto con la cantidad de coincidencias entre sus coocurrencias (r) y su clasificación como candidatos a terminologización. Por una parte, el valor de r representa el número de coincidencias entre los listados de coocurrencias extraídos del corpus especializado y el corpus de referencia, es decir, indica el grado de correspondencia entre el perfil de coocurrencia de cada sustantivo analizado. Por otra parte, la clasificación de los candidatos a terminologización se manifiesta mediante un 1 si es correcta y 0 si es incorrecta.

	Candidato	r	Correcto
1	anotador	1	1
2	basa	1	0
3	disparador	1	1
4	indexación	1	1
5	interdisciplinariedad	1	0
6	pare	1	0
7	pulsación	1	0
8	solapamiento	1	1
9	categorización	2	1
10	clasificador	2	1
11	clic	2	0
12	colon	2	0
13	cope	2	0
14	herencia	2	1
15	maximización	2	1
16	nodo	2	1
17	notación	2	1
18	parada	2	1
19	polaridad	2	1
20	portabilidad	2	1
21	quechua	2	0
22	quita	2	0
23	regalo	2	0
24	reutilización	2	0
25	separador	2	1
26	tilde	2	0
27	transformador	2	1
28	tripleta	2	1
29	varianza	2	1
30	agregación	3	1

Tabla 1: Primeros 30 resultados del ranking de candidatos a terminologización.

Un ejemplo de terminologización es *anotador*. Este sustantivo en el lenguaje general se emplea para aludir a alguien que anota (Real Academia Española, 2024), mientras que, para el área del PLN, un anotador consiste en una herramienta que asigna etiquetas lingüísticas a diversos elementos de un texto. En el caso del corpus especializado empleado, se asocia con la ejecución de tareas de análisis morfosintáctico, como se observa en el ejemplo 1.

(1) "La anotación morfosintáctica en TEITOK se lleva a cabo con el **anotador** automático NeoTag [...], un analizador probabilístico del tipo HMM que adjudica a cada palabra la etiqueta correspondiente según la función gramatical que cumple en el texto."

Un caso similar ocurre con el sustantivo *disparador*, el cual está originalmente asociado a la pieza de las armas portátiles que permite dispararlas o con la persona que dispara (Real Academia Española, 2024). Sin embargo, en el lenguaje de especialidad con el que se trabajó, un disparador –empleado como equivalente de *trigger*– refiere a un elemento que lleva a cabo una acción determinada ante un evento en específico. En el caso del ejemplo 2, se asocia con el análisis de sentimiento:

(2) "Para ello, se introduce un nuevo concepto denominado **disparador** de emoción. Inicialmente, se construye de forma incremental una base de datos léxica de disparadores de emoción asociados a la cultura con la que se quiere trabajar".

Otro proceso de terminologización ocurre con *solapamiento*, que en su uso en el léxico general se relaciona con la acción y efecto de superponer elementos (Real Academia Española, 2024). No obstante, para el lenguaje de especialidad con el que se trabajó, el sustantivo, equivalente al inglés *overlapping*, se utiliza para denotar la intersección o coincidencia entre distintas unidades de texto, tal como se presenta en el ejemplo 3.

(3) "el procedimiento determina los sentidos de las palabras que ocurren en un contexto particular basándose en una medida de **solapamiento** entre las definiciones de un diccionario y dicho contexto".

Otro ejemplo interesante de terminologización es el de *herencia*, sustantivo relacionado con el conjunto de bienes transmisibles a familiares o herederos (Real Academia Española, 2024). Para el caso de la disciplina del PLN, *herencia* describe el mecanismo capaz de transferir propiedades o características a otros elementos en modelos computacionales. El ejemplo 4 pone de manifiesto esta acepción.

(4) "Además, el formalismo ofrece un mecanismo de **herencia** de paradigmas para crear paradigmas similares, como el paradigma de verbos defectivos en 'ar' para 'tronar' a partir del paradigma de verbos en 'ar' ".

Si bien la Tabla 1 y los distintos ejemplos expuestos con anterioridad reflejan la eficacia del método desarrollado para la detección de

la terminologización, se reconocen algunas limitaciones que provocan una clasificación incorrecta de candidatos. Una de ellas consiste en el amplio número de sustantivos que fueron falsos positivos en los primeros tramos del ranking generado por el método propuesto. Ejemplos de sustantivos considerados erróneamente como terminologizados al estar en las primeras posiciones del listado jerarquizado según el coeficiente r son, por ejemplo, *basa, interdisciplinariedad, quechua, quita, regalo*, entre otros.

Es posible aventurar distintos motivos por los que puede ocurrir la categorización de falsos positivos. Una de las razones está asociada con que son sustantivos de baja frecuencia en los corpus y que, si bien no han sido descartados mediante el filtro de frecuencia aplicado en la metodología, aun así no tienen un alto nivel de uso en el corpus especializado ni en el de referencia. La Tabla 2 muestra la cantidad de coincidencias (hits) que tienen las palabras para cada corpus.

Palabras	ESP_PLN	REF_PRE
basa	309	5
adición	42	22
pare	18	6
pulsación	65	25
colon	27	19
quita	9	18

Tabla 2: Ejemplos de error con su frecuencia.

A partir de los datos de la Tabla 2, es posible observar que algunos sustantivos presentan un bajo nivel de uso para el corpus ESP_PLN y para el REF_PRE, como el caso de *adición, pare, colon* y *quita*. Esta situación explica su errónea posición en los primeros puestos de los resultados de acuerdo con su valor r, dado que las coocurrencias obtenidas probablemente no son representativas para reflejar el proceso de terminologización. Una problemática similar ocurre con aquellos sustantivos que tienen una elevada frecuencia en el corpus REF_PRE y baja en el ESP_PLN, como es el caso de un falso positivo como *regalo*. Este sustantivo está asociado en el corpus REF_PRE con su acepción tradicional de obsequio o presente que se hace de manera voluntaria, y de esta misma forma está expresado en ESP_PLN. Sin embargo, en este corpus se identifica una menor incidencia de esta unidad léxica, puesto que está siendo empleada mayoritariamente en ejemplos de estudio de flexión verbal (5).

(5) "Pedro se llevó el **regalo**".

Otro motivo que generó dificultades para que el método categorizara al sustantivo como terminologizado fueron los problemas asociados con el etiquetado de los corpus. Esta situación se manifiesta en el caso de unidades como *basa* y *quita*. La Tabla 3 expone un extracto del corpus ESP_PLN en donde el etiquetado morfosintáctico detectó erróneamente la forma léxica *basa* con la categoría de sustantivo (NOUN), en lugar de la forma verbal *basar*.

Algo similar ocurre con *quita*, donde el etiquetador nuevamente considera una forma de sustantivo en lugar de un verbo. En este sentido, el problema del etiquetado, si bien afecta a la precisión del método mediante el establecimiento del ranking de resultados, también es un indicador de la relevancia que tiene el preprocesamiento de los datos para el desarrollo de un método automatizado. Por ejemplo, el realizar pruebas con diferentes etiquetadores morfosintácticos, descartar las palabras en inglés y el proceso de limpieza de corpus son esenciales para obtener un resultado de calidad.

Tratándose de una propuesta metodológica, lo ideal sería ofrecer una evaluación comparativa con otros sistemas. Sin embargo, en este caso es difícil al tratarse de un trabajo sin antecedentes metodológicos claros ya que, que sepamos, no se han producido hasta ahora trabajos de investigación que aborden esta misma temática. Como una referencia general, llevamos a cabo una prueba a baja escala utilizando el modelo Llama3 (Meta, 2024), en su versión reducida (8B), utilizando la implementación disponible a través del Proyecto Ollama (Ollama, 2024).

Dada la imposibilidad de analizar la totalidad del corpus con este tipo de sistemas, como consecuencia de su complejidad inherente, nos vimos obligados a pensar en una operativa diferente, sin corpus. De esta forma, sometemos sustantivos al arbitrio del modelo, para que decida si son o no casos de terminologización en el ámbito de la lingüística computacional. El modelo tarda alrededor de dos minutos en dar un dictamen por cada palabra, con lo cual se podría procesar los cerca de 50.000 sustantivos del diccionario en alre-

1	Nuestro	Nuestro	ADP	IN	_	5
2	modelo	modelo	ADJ	JJ	Degree=Pos	5
3	computacional	computacional	ADJ	JJ	Degree=Pos	5
4	se	se	NOUN	NN	Number=Sing	5
5	basa	basa	NOUN	NN	Number=Sing	0
6	en	en	ADP	IN	_	7

Tabla 3: Ejemplo de error de etiquetado morfosintáctico de la forma *basa*.

dedor de 70 días [2]. Como pilotaje, sometimos a análisis un conjunto de 143 palabras, compuesto por sustantivos que comienzan con la letra "a" y que aparecen tanto en el diccionario de la RAE como en el corpus especializado. Llama3 eligió 19 palabras (Tabla 4), entre las cuales encontramos 7 casos correctos: *alerta, analizador, anotación, aplicación, asignación, asistente y alineamiento*. Otros casos, como *algoritmo* o *autómata*, pueden tener un significado técnico en el corpus, pero no han sido terminologizados.

#	Caso	Ok
1	abreviatura	
2	acrónimo	
3	acta	
4	algoritmo	
5	alias	
6	alerta	1
7	analizador	1
8	anotación	1
9	aplicabilidad	
10	aplicación	1
11	aporte	
12	asignación	1
13	asistente	1
14	autómata	
15	automático	
16	abstracción	
17	adaptación	
18	alineamiento	1
19	ampliación	

Tabla 4: Terminologización según *Llama3*.

La tasa de acierto Llama3 en esta tarea es de 37 % (7 / 19), inferior a la del algoritmo aquí propuesto. Si bien se trata de una prueba a baja escala, permite poner en perspectiva el valor de precisión de nuestro sistema. Además de baja precisión, este sistema también omite muchos de los sustantivos terminologizados que nuestro sistema sí fue capaz de detectar, tales como *alineación* (Llama 3 solo detecta la forma *alineamiento*), o *anotador* (solo detecta *anotación*), etc.

[2]Cabría experimentar con distintas instrucciones para determinar hasta qué punto es posible acelerar el proceso sin perder calidad.

5 Conclusiones y planes de trabajo futuro

En esta investigación hemos propuesto un método automatizado para la detección de la terminologización de los sustantivos en un dominio de especialidad, y hemos puesto a prueba el método con un análisis de un corpus del dominio del PLN. Los hallazgos obtenidos representan una contribución para la comprensión y estudio de los fenómenos asociados al cambio semántico y, en particular, al proceso de terminologización.

Los principales resultados del estudio exponen la eficacia del método propuesto, especialmente en la distinción de la probabilidad de sustantivos con alta y baja capacidad de experimentar terminologización. El algoritmo diseñado demostró una precisión relativamente alta al colocar en las primeras posiciones del *ranking* a los sustantivos con mayor potencial. Este hecho demuestra que el enfoque basado en la comparación de coocurrencias entre un corpus especializado y uno de referencia permite detectar la terminologización y, además, comprender la dinámica de migración de unidades léxicas entre el lenguaje general y los lenguajes de especialidad.

Es preciso destacar, además, la simplicidad del método, que en este caso se traduce directamente en eficiencia computacional. Esto es importante porque permitirá escalar el volumen de datos procesados. Además, la simplicidad conceptual del método también va a facilitar futuras implementaciones en diferentes escenarios.

La evaluación cualitativa de los resultados refleja algunos aspectos por mejorar en

la implementación del método propuesto, que sugieren posibilidades de trabajo futuro. En primer lugar, un aspecto evidente es la revisión y selección de un etiquetador morfosintáctico más preciso en la detección de categorías gramaticales, principalmente con las unidades léxicas de interés para el estudio (sustantivos, adjetivos, verbos y nombres propios). Corresponde, asimismo, mejorar el análisis cuantitativo de los resultados. Una posibilidad sería una medición de la cobertura del sistema, lo que requerirá la identificación previa, independiente y manual de un conjunto de casos de terminologización.

Por último, una vía de trabajo futuro necesaria será intentar reproducir el estudio en otros dominios de especialidad y también en otras lenguas. Cabría determinar, por ejemplo, cuál es el efecto de la selección del dominio de especialidad en la calidad del resultado. Ciertos dominios, como el de la tecnología, se encuentran también en la prensa, y esto podría interferir con el método. Con todo, se debe considerar que la naturaleza de los términos del discurso especializado es distinta a la del vocabulario general. No se utilizan las mismas unidades léxicas ni de la misma manera.

Agradecimientos

Esta investigación ha sido posible gracias a la financiación de la Beca de Magíster Nacional/2023 de la Agencia Nacional de Investigación y Desarrollo (ANID) del Gobierno de Chile. Agradecemos también a los revisores por su valioso trabajo, que ha contribuido a la mejora de este artículo.

Bibliografía

Adelstein, A. 2022. Neología y semántica: grados de neologicidad en el ámbito nominal. En E. Bernal J. Freixa, y S. Torner, editores, *La neología del español: del uso al diccionario*. Iberoamericana Vervuert, páginas 327–346.

Ahumada, J. 2024. Propuesta metodológica para detectar la terminologización de sustantivos. Tesis de magíster, Pontificia Universidad Católica de Valparaíso.

Álvarez de Miranda, P. 2009. Neología y pérdida léxica. En E. de Miguel, editor, *Panorama de la lexicología*. Ariel, páginas 133–157.

Anthony, L. 2012. AntConc. Waseda University, Tokyo, Japan.

Arce, L. y M. Seghiri. 2021. Generation of a glossary for the translation of housing purchase and sale agreements in Spain, Argentina, the United Kingdom, and the United States. *Hermēneus. Revista de Traducción e Interpretación*, 24:87–118.

Auger, P. y L. Rousseau. 1977. *Méthodologie de la recherche terminologique*. Régie de la langue française.

Barán, M. 1999. El pulso ondulante de la metaforización en terminologías de la lengua española. *Moenia*, 5:247–255.

Bastuji, J. 1974. Aspects de la néologie sémantique. *Langages*, 36:6–19.

Berri, M. 2013. Neología semántica nominal y metáfora. En G. Ciapuscio, editor, *Variedades del español de la Argentina: estudios textuales y de semántica léxica*. Eudeba, páginas 131–150.

Berri, M. y L. Bregant. 2015. Identificación de metonimias y metáforas: cuestiones metodológicas. *Lenguaje*, 43(2):219–245.

Bloomfield, L. 1984. *Language*. University of Chicago Press.

Cabré, M. T. 1993. *La terminología: teoría, metodología, aplicaciones*. Antártida-Empúries.

Cabré, M. T. 2006. La clasificación de neologismos: una tarea compleja. *ALFA*, 50(2):229–250.

Cabré, M. T. 2008. La neología efímera. En M. T. Cabré J. Freixa, y E. Solé, editores, *Lèxic i Neologia*. Barcelona: Observatori de Neologia, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, páginas 13–28.

Cabré, M. T., R. Bayà, E. Bernal, J. Freixa, E. Solé, y T. Vallés. 2002. Evaluación de la vitalidad de una lengua a través de la neología: a propósito de la neología espontánea y de la neología planificada. En M. T. Cabré J. Freixa, y E. Solé, editores, *Lèxic i Neologia*. Barcelona: Observatori de Neologia, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, páginas 159–202.

Cabré, M. T., R. Estopà, y J. Vivaldi. 2001. Automatic term detection: A review of current systems. En D. Bourigault C. Jacquemin, y M.-C. L'Homme, editores, *Recent Advances in Computational Terminology*. John Benjamins, páginas 53–88.

Cassany, D. 2007. La lengua entre las humanidades y las ciencias. *Página y Signos*, (2):15–27.

Díaz Hormigo, M. T. 2007. Aproximación lingüística a la neología léxica. En Unknown, editor, *Morfología: investigación, docencia, aplicaciones: Actas del II Encuentro de Morfología: investigación, docencia*, páginas 33–54. Universidad de Extremadura.

Díaz Hormigo, M. T. 2020. Precisiones para una caracterización lingüística de la neología semántica. *ELUA*, 34:73–94.

Domènech, O. 2008. Metodología de trabajo del observatorio de neología del instituto de lingüística aplicada de la universidad pompeu fabra. En A. Pérez R. Montoro del Arco, y E. Tomás, editores, *Neologismo y morfología*, páginas 11–38. Universidad de Murcia.

Drouin, P. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Estopà, R. 2016. La neología especializada: Términos médicos en la prensa española. En C. Sánchez y D. Azorín, editores, *Estudios de neología del español*, páginas 109–129.

Estopà, R. 2022. Neología general y especializada. En E. Bernal J. Freixa, y S. Torner, editores, *La neología del español: del uso al diccionario*, páginas 151–174. Iberoamericana Vervuert.

Estornell Pons, M. y A. Soto. 2016. La metáfora y la metonimia como procedimientos de creación neológica en el discurso gastronómico actual. *Tonos Digital*, páginas 1–24.

Freixa, J. 2016. Terminología. En G. Gutiérrez, editor, *Enciclopedia de Lingüística Hispánica*. Routledge, páginas 326–333.

Freixa, J., R. Luna, y M. Suárez de la Torre. 2010. La neología semántica en las antenas neológicas: descripción de algunos fenómenos relevantes. En M. Cabré O. Domènech R. Estopà J. Freixa, y M. Lorente, editores, *Actes del Congrés Internacional de Neología en les Llengües Romàniques (CINEO)*, páginas 837–852. Institut de Lingüística Aplicada/Universitat Pompeu Fabra.

Frermann, L. y M. Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Fuentes, M., C. Gerding, A. Pecchi, G. Kotz, y P. Cañete. 2009. Neología léxica: Reflejo de la vitalidad del español de chile. *Revista de Lingüística Teórica y Aplicada*, 47(1):103–124.

Geeraerts, D. 2010. *Theories of Lexical Semantics*. Oxford University Press.

Goldhahn, D., T. Eckart, y U. Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. En N. Calzolari K. Choukri T. Declerck M. Doğan B. Maegaard J. Mariani A. Moreno J. Odijk, y S. Piperidis, editores, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, páginas 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Gonen, H., G. Jawahar, D. Seddah, y Y. Goldberg. 2021. Simple, interpretable and stable method for detecting words with usage change across corpora. *arXiv preprint arXiv:2112.14330*.

Guerrero Ramos, G. 1995. *Neologismos en el español actual*. Arco/Libros.

Guilbert, L. 1974. Grammaire générative et néologie lexicale. *Langages*, 36:34–44.

Guilbert, L. 1975. *La creativité lexicale*. Larousse Université.

Gulordava, K. y M. Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. En *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, páginas 67–71.

Hamilton, W. L., J. Leskovec, y D. Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Heyer, G., C. Kantner, A. Niekler, M. Overbeck, y G. Wiedemann. 2017. Modeling the dynamics of domain specific terminology in diachronic corpora. *arXiv preprint arXiv:1707.03255*, páginas 1–16.

Holeš, J. y Z. Honová. 2023. La métaphore terminologique sur l'exemple des termes tchèques et français du domaine d'astronomie et d'astrophysique. *Linguistica Silesiana*, 44(2):109–120.

Kernighan, B. W. y D. M. Ritchie. 1991. *El lenguaje de programación C*. Pearson Educación.

Kilgarriff, A. R., P. Rychly, P. Smrz, y D. Tugwell. 2004. The Sketch Engine. En G. Williams y S. Vessier, editores, *Proceedings of the 11th EURALEX International Congress*, páginas 105–116. Université de Bretagne Sud.

Llanos, J. 2023. Aplicación de herramientas de extracción de términos para la definición de palabras clave especializadas en SARS-COV-2. *Debate Terminológico. Nueva Época*, 1(1):5–13.

Martí, M. 2009. La discriminación de los sintagmas terminológicos en los glosarios especializados. *Lingüística Española Actual*, 31:61–88.

Martín, J. 2004. Los procesos neológicos del léxico científico. Esbozo de clasificación. *Anuario de Estudios Filológicos*, 17:157–174.

Meta. 2024. Llama. https://www.llama.com. Con acceso: 17-02-2025.

Montariol, S., M. Martinc, y L. Pivovarova. 2021. Scalable and interpretable semantic change detection. En *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, páginas 4642–4652.

Nazar, R. 2011. Neología semántica: un enfoque desde la lingüística cuantitativa. Seminario IULAterm. Disponible en: http://tecling.com/nazar/111214nazar.pdf.

Nazar, R. 2023. Extensión, variación y extensión del léxico español. En S. Torner P. Battaner, y I. Renau, editores, *The Routledge Handbook of Spanish Lexicography*. Routledge, páginas 204–218.

Nazar, R. y V. Vidal. 2010. Aproximación Cuantitativa a la Neología. En M. Cabré O. Domènech R. Estopà J. Freixa, y M. Lorente, editores, *Actes del Congrés Internacional de Neología en les Llengües Romàniques (CINEO)*. Institut de Lingüística Aplicada/Universitat Pompeu Fabra, páginas 867–880.

Ollama. 2024. Ollama. https://ollama.com. Con acceso: 17-02-2025.

Pottier Navarro, H. 1979. La néologie en espagnol contemporaine. *Les Langues Néolatines*, 229:148–172.

Real Academia Española. 2024. Diccionario de la lengua española. 23.7 en línea.

Renau, I. 2023. A corpus-based study of semantic neology of the Covid-19 pandemic. *Quaderns de Filologia: Estudis Lingüístics*, 28:55–76.

Rey, A. 1976. Le néologisme: un pseudoconcept? *Cahiers de Lexicologie*, (28):3–7.

Roelcke, T. 2018. Technical Terminology. En J. Humbley G. Budin, y C. Laurén, editores, *Languages for Special Purposes: An International Handbook*. De Gruyter Mouton, páginas 489–508.

Rondeau, G. 1984. *Introduction à la terminologie*. Gaëtan Morin.

Sablayrolles, J. F. 1997. Néologismes: Une typologie des typologies. *Cahier du CIEL*, páginas 11–48.

Sager, J. 1990. *Practical Course in Terminology Processing*. John Benjamins.

Sagi, E., E. Kaufmann, y B. Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. En *GEMS 2009*, páginas 104–111.

Santamaría, I. 2013. La representación de la neología semántica en los diccionarios del español. *Revista de Lexicografía*, 19:139–166.

Santamaría, J. y M. Krallinger. 2018. Construcción de recursos terminológicos médicos para el español: el sistema de extracción de términos CUTEXT y los repositorios de términos biomédicos. *Procesamiento del Lenguaje Natural*, 61:49–56.

Sanz, G. 2008. Traducción de textos de Ciencias Humanas: problemas terminológicos.

En L. Pegenaute J. Decesaris M. Tricás, y E. Bernal, editores, *Actas del III Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación*, páginas 1–11.

Seco, M. 1989. *Gramática esencial del español. Introducción al estudio de la lengua.* Espasa-Calpe.

Straka, M. y J. Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. En *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, páginas 88–99. Association for Computational Linguistics.

Sánchez, M. C. 2009. Procedimientos trópicos en la neología semántica: sistematicidad y creatividad. *Revista de investigación lingüística*, 12:1139–1146.

Tahmasebi, N., L. Borin, y A. Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change*, 6(1):1–56.

Tang, X. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

Torres Rivera, A. 2020. *Detección y extracción de neologismos semánticos especializados: un acercamiento mediante clasificación automática de documentos y estrategias de aprendizaje profundo.* Tesis doctoral, Universitat Pompeu Fabra.

Ullmann, S. 1965. *Semántica: Introducción a la ciencia del significado.* Ediciones Aguilar S.A.

Yao, Z., Y. Sun, W. Ding, N. Rao, y H. Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. En *Proceedings of the eleventh acm international conference on web search and data mining*, páginas 673–681.