El etiquetado gramatical automático en el procesamiento del habla coloquial

MARTA GARROTE SALAZAR
Universidad Autónoma de Madrid
marta.garrote@uam.es

Resumen: El etiquetado gramatical (*part-of-speech tagging*) es una técnica esencial del Procesamiento del Lenguaje Natural (PLN) que consiste en asignar etiquetas gramaticales a cada palabra en un determinado texto. Mientras que el etiquetado gramatical se ha estudiado de manera extensa en datos lingüísticos formales y bien estructurados, el etiquetado preciso de textos de habla coloquial plantea dificultades específicas. Este artículo tiene como finalidad explorar las dificultades a las que se enfrentan las técnicas de etiquetado gramatical a la hora de analizar e interpretar el habla coloquial. Se discuten fenómenos como el impacto en la precisión del etiquetado gramatical de los dialectos, la jerga, el contexto cultural y social y las disfluencias discursivas, entre otros. Además, tras una revisión del estado de la cuestión, se identifican soluciones potenciales y se establece una prospectiva de estudio que mejore el desempeño del etiquetado gramatical en contextos de habla coloquial.

Palabras clave: etiquetado gramatical automático, procesamiento del lenguaje natural (PLN), habla coloquial, dificultades de etiquetado, prospectiva de investigación.

Automatic POS tagging in colloquial speech processing

Abstract: Part-of-speech (POS) tagging is a fundamental Natural Language Processing (NLP) task that involves assigning grammatical labels to each word in a given text. While POS tagging has been extensively studied in formal, well-structured language data, the accurate tagging of colloquial speech corpora presents unique challenges. This paper aims to explore the difficulties faced when employing POS tagging techniques on colloquial speech texts. We discuss the impact of dialects, slang, cultural and social context, and speech disfluencies on the accuracy of POS tagging. Furthermore, after a review of the state of

the art, we identify potential solutions and future research directions to improve the performance of POS tagging in colloquial speech contexts.

Keywords: automatic POS tagging, natural language processing (NLP), colloquial speech, tagging difficulties, research prospective.

1. Introducción

El procesamiento del lenguaje natural (PLN) es un área de estudio interdisciplinar que se sitúa entre la inteligencia artificial y la lingüística computacional, y que se centra en el análisis de la interacción entre la máquina y el ser humano. El objetivo principal del PLN es permitir que las máquinas comprendan, interpreten y respondan al lenguaje de manera similar a como lo hace el ser humano (Bolaños 2015). El PLN es esencial para muchas aplicaciones, como los asistentes virtuales (como Siri o Alexa), los traductores automáticos (como Google Translate), los sistemas de recomendación (como los de Amazon o Youtube), los chatbots (como ChatGPT), la extracción de información y muchos otros sistemas que interactúan a través del lenguaje humano. Para lograr esa interacción, el PLN abarca una gran variedad de técnicas, entre las cuales se encuentra el etiquetado gramatical o *part-of-speech* (POS) *tagging*.

El etiquetado gramatical es una técnica del PLN que consiste en asignar categorías gramaticales (como nombre, verbo, adjetivo, preposición, etc.) a las palabras de un texto con la finalidad de analizar y comprender la estructura gramatical de las oraciones mediante la clasificación de sus componentes en categorías sintácticas y, así, interpretar su significado correcto, resolviendo la ambigüedad en tareas de análisis sintáctico estructural (Chiche y Yitagesu 2022).

El etiquetado gramatical juega un papel esencial en el PLN, como en la traducción automática, en la minería de datos o en la extracción de información. En la mayoría de los estudios y desarrollos, los modelos de etiquetado gramatical han sido entrenados y evaluados a partir de corpus escritos de lengua formal (Taulé et al. 2015). No obstante, el creciente interés por analizar y procesar lenguaje coloquial genera nuevos retos debido a su naturaleza informal. Si el lenguaje informal supone mayor dificultad que el formal en la precisión del análisis del etiquetador, esta dificultad incrementa cuando se trata de analizar el habla coloquial, es decir, oral, ya que, además de las características propias del lenguaje informal, la lengua oral incluye elementos comunicativos, como la prosodia o la discontinuidad de la cadena hablada, que

hacen que la precisión de los etiquetadores automáticos disminuya. Profundizaremos en estos fenómenos más adelante.

Este estudio plantea las dificultades que, a pesar de los enormes avances en el PLN, aún existen en el etiquetado gramatical del habla coloquial y explora estrategias potenciales para solventar estas dificultades de manera eficiente. Para ello, se explicará en profundidad qué es el etiquetado gramatical, cuáles son los principales métodos existentes, qué carencias presentan las distintas técnicas a la hora de analizar e interpretar el habla coloquial y qué soluciones se deben adoptar en futuros trabajos.

2. ¿Qué es el etiquetado gramatical?

Como se ha mencionado, el etiquetado gramatical es una técnica o tarea muy habitual dentro del PLN que consiste en asignar etiquetas gramaticales a cada palabra en un texto (Chiche y Yitagesu 2022). Un ejemplo es el etiquetario utilizado en el corpus CHIEDE (Garrote 2010: 86) que se reproduce en la Tabla 1:

Categoría	Etiqueta	Ejemplo
Sustantivo	N	mesa
Nombre propio	NPR	Natalia
Adjetivo	ADJ	rojo
Artículo	ART	el, la, lo
Posesivo	POSS	mi, tu, su
Demostrativo	DEM	este, ese
Cuantificador	Q	uno, primero, muchos
Pronombre	P	yo, tú, lo_que
Relativo	REL	que
Verbo	V	comer
Auxiliar	AUX	he comido
Preposición	PREP	ante, bajo, con
Adverbio	ADV	aquí, así, allí
Conjunción	C	y, pero, ni
Marcador discursivo	MD	oye, o_sea, es_decir
Interjección	INTJ	madre_mía, uy

Tabla 1. Etiquetario del corpus CHIEDE (Garrote 2010: 86).

Mediante el etiquetado gramatical, la máquina no solo puede interpretar la oración de manera adecuada, atendiendo a las relaciones sintácticas de los elementos que la conforman, sino que también puede evitar la ambigüedad semántica, entre otras funciones. Por ejemplo, si el etiquetado es preciso –siguiendo los ejemplos de la Tabla 1–, la máquina interpretará que *madre_mía* es una interjección, no una referencia a la madre del hablante.

2.1. Usos del etiquetado gramatical en el PLN

La información que proporciona el etiquetado gramatical (Ghosh y Mishra 2020) es crucial para tareas del PLN como:

— **Análisis sintáctico** (*parser*): el etiquetado gramatical es un paso preliminar a tareas más complejas como el análisis sintáctico (Castillo Velásquez *et al.* 2020), mediante el cual se establecen las relaciones entre palabras y su papel sintáctico. El análisis sintáctico automático es una herramienta fundamental para analizar y procesar la estructura de un texto, de acuerdo con las reglas definidas, para asegurar que se ajusta a la gramática esperada y permite una interpretación posterior adecuada.

— **Reconocimiento de entidades nombradas** (*Named-entity recognition*): la identificación de la categoría gramatical de cada palabra ayuda en el reconocimiento de entidades, como nombres propios de personas, organizaciones, localizaciones, etc. (Landolsi *et al.* 2024). El reconocimiento de entidades nombradas es una herramienta poderosa en el PLN que permite la extracción automática y estructurada de información clave a partir de textos. No obstante, uno de los desafíos a los que se enfrentan los sistemas de reconocimiento de entidades nombradas es la ambigüedad, ya que una misma palabra (o grupo de palabras) puede referirse a diferentes entidades según el contexto. Para ello, un etiquetador gramatical puede ser de ayuda para identificar la categoría gramatical y facilitar la correcta interpretación. Por ejemplo, la palabra *chile* puede etiquetarse como N o como NPR (ver Tabla 1) y, dependiendo de la categoría gramatical que se le asigne, el significado del enunciado oracional variará.

— **Traducción automática**: entender la estructura gramatical de una oración es esencial para acometer una traducción rigurosa (Sánchez-Cartagena 2024). Este proceso de traducción automática implica varios niveles de análisis lingüístico, desde el nivel léxico hasta el sintáctico y semántico, pasando por el etiquetado gramatical.

— **Recuperación de información**: el etiquetado gramatical puede ser de utilidad en los sistemas de recuperación de información para mejorar los motores de búsqueda al considerar las categorías gramaticales y los papeles sintácticos de las palabras (Cherradi y Haddadi 2024).

— **Reconocimiento y síntesis de voz** (*Speech-to-Text and Text-to-Speech systems*): conocer la categoría gramatical de cada palabra ayuda a interpretar y generar un discurso más natural y adecuado al contexto (Ying *et al.* 2024).

Respecto a los métodos, existen distintos enfoques para abordar el etiquetado gramatical de un texto. En la siguiente sección haremos un breve repaso de la información más relevante.

2.2. Enfoques y métodos

El etiquetado gramatical nace de la necesidad de desambiguar palabras cuya categoría gramatical varía de un contexto a otro. El etiquetado manual es una tarea ardua; de ahí la búsqueda de la automatización. Para ello, se han desarrollado distintos enfoques. Los más generalizados son los enfoques basados en reglas (*rule-based approach*), los enfoques probabilísticos (*probabilistic approach*) y los basados en transformaciones (*transformational-based approach*) (Martínez 2012).

Las técnicas basadas en reglas asignan categorías gramaticales a las unidades léxicas a partir de reglas lingüísticas creadas manualmente; por ejemplo, se establece que ante una palabra ambigua como *bajo*, esta será etiquetada como nombre, y no como preposición, si va precedida de un determinante. Uno de los desarrollos más conocidos dentro de los modelos basados en reglas es el *Brill tagger* (Brill 1993).

Los enfoques probabilísticos parten de la frecuencia de las etiquetas presentes en un corpus ya etiquetado manualmente y asignan la etiqueta basándose en una mayor probabilidad (modelos estocásticos) de la categoría de una determinada palabra en un determinado contexto. El modelo oculto de Markov (*Hidden Markov Model*) es uno de los más comunes entre los estocásticos.

El tercer enfoque, el basado en transformaciones, aplica un sistema híbrido, combinando los dos anteriores y generando, así, reglas a partir de un corpus¹.

Recientemente, se están utilizando técnicas de la inteligencia artificial (IA) para realizar el etiquetado gramatical de manera automática. Dos ejemplos son el uso del aprendizaje automático (*machine learning*) y del aprendizaje profundo (*deep learning*). Chiche y Yitagesu (2022) realizan un análisis exhaustivo del estado de la cuestión, comparando distintos métodos basados en la IA para etiquetar automáticamente las partes del discurso de un texto y concluyen que, hasta el momento, los más eficientes son los basados en el aprendizaje profundo, seguidos de aquellos basados en el aprendizaje automático y, por último, los métodos híbridos.

De acuerdo con la mayoría de los estudios (Bonilla 2024), el porcentaje de precisión de los etiquetadores gramaticales automáticos se sitúa

¹ Véase Chiche y Yitagesu (2022), Kumawat y Jain (2015) y Martínez (2012) para una mayor información sobre los distintos métodos de etiquetado gramatical.

alrededor del 97 %. No obstante, esta cifra hace referencia al lenguaje formal. Este porcentaje disminuye considerablemente (dependiendo de la lengua y del dominio lingüístico) cuando nos referimos al lenguaje oral. Es decir, que incluso en textos formales y editados el porcentaje de acierto no alcanza el 100 %. Si bien dichos porcentajes de precisión son un éxito para el PLN, el reto consiste en conseguir una precisión del 100 % para cualquier tipo de registro lingüístico. Las peculiaridades idiosincrásicas de la lengua informal, especialmente del habla coloquial, hacen que su procesamiento automático aún esté lejos de ser una realidad. En la siguiente sección se abordará la definición de habla coloquial, sus características y, derivadas de estas, la dificultad de su procesamiento automático.

3. Dificultades para etiquetar gramaticalmente el habla coloquial

3.1. *Definición de habla coloquial*

El evento de habla coloquial se caracteriza por su espontaneidad y por ciertos elementos intrínsecos a ella, derivados de la presencialidad y la contextualidad (Garrote 2010: 66):

- Interacciones cara a cara multimodales: importancia del contexto.
- Referencia intersubjetiva a un espacio deíctico.
- Programación simultánea de la producción.
- Comportamiento lingüístico impredecible sujeto al contexto.
- Discontinuidad de la cadena hablada.
- Lenguaje corporal.
- Aspectos suprasegmentales.
- Aspectos pragmáticos.
- Características sociales y contextuales que influyen en el evento del habla.

El contexto, el espacio, la espontaneidad, la impredecibilidad, la ruptura del flujo discursivo (disfluencias), los elementos paralingüísticos o el lenguaje no verbal son elementos comunicativos que el ser humano, en circunstancias normales, procesa con naturalidad. No obstante, todos estos fenómenos comunicativos son difíciles de formalizar en un lenguaje que pueda procesar la máquina (Rojo 2021).

Según Briz (2016: 463), el habla coloquial es «[u]n registro o uso lingüístico [...] empleado en situaciones de inmediatez comunicativa, a la vez que favorecido por estas». Añade el autor que dichas situaciones se definen por las relaciones entre los hablantes, de igualdad social, proximidad, cotidianidad temática y espacial, el fin interpersonal, la falta de planificación y el tono informal, lo que hace que el discurso se caracterice por rasgos dialectales y sociolectales (propios de las características sociolingüísticas de los hablantes). Además, Briz (2016: 464) matiza que «a veces se escribe como si se hablara». Esto se aprecia especialmente en interacciones a través de redes sociales. Cuando hablamos con un amigo por Whatsapp, escribimos como si estuviéramos hablando, intentando reproducir, de hecho, fenómenos característicos de la lengua oral, como rasgos prosódicos (utilización de mayúsculas como si estuviéramos gritando), pérdida de sonidos (*pa mí*), léxico específico (sociolecto), marcadores discursivos prototípicamente orales (*bueno, mira...*) o una sintaxis que sigue un orden menos neutro que el estándar sujeto-verbo-objeto, etc. Si tomamos en consideración alguno de los ejemplos citados, el lexicón (diccionario) de un sistema de etiquetado gramatical no recogerá la palabra *pa* como apócope de la preposición *para*; y, probablemente, tampoco incluya neologismos y palabras propias de un sociolecto; además, el sistema de etiquetado podría categorizar de manera errónea la palabra *bueno* como adjetivo cuando esté funcionando como marcador discursivo.

Todas estas características del habla coloquial hacen que la descodificación e interpretación del discurso por parte de la máquina sean más complejas y costosas. El procesamiento del habla coloquial añade retos al PLN que detallaremos a continuación.

3.2. ¿Por qué es difícil etiquetar gramaticalmente el habla coloquial?

El etiquetado gramatical automático de textos de lenguaje coloquial (concretamente, del habla coloquial, aunque, como se ha mencionado, los textos escritos en ciertos contextos presentan características similares a las de la oralidad) plantea varios desafíos debido a la particularidad y a la variabilidad de este registro lingüístico. A pesar de que, en primer lugar, detallaremos aquellos propios de la oralidad, también se mencionarán rasgos propios de la escritura coloquial que afectan a la eficacia y al rendimiento del proceso de etiquetado gramatical automático. Respecto a la oralidad, los principales retos son la variedad léxica, la ambigüedad, las disfluencias, el contexto social y cultural o la innovación y creatividad lingüísticas (entre otros).

La variedad léxica, debido a cuestiones como el uso de dialectos, de jergas o de préstamos lingüísticos, es difícil de procesar mediante

un programa de etiquetado gramatical automático porque determinadas unidades léxicas no aparecen en los lexicones, ni son habituales en los corpus etiquetados utilizados como modelo de entrenamiento. La imposibilidad de que un corpus contenga todos los elementos de una lengua impide que se pueda recopilar un conjunto de datos completo con el que entrenar a la máquina (Rojo 2021). Así, se alcanza una mayor precisión cuando se trabaja con corpus de documentos de dominios lingüísticos específicos (Neunerdt *et al.* 2013) para los que se ha entrenado a la máquina.

La ambigüedad (que puede ser léxica o contextual), junto con la polisemia, plantea un gran reto a los etiquetadores gramaticales. De hecho, la desambiguación es la principal función de un etiquetador. En la lengua coloquial es más habitual que las palabras puedan tener diferentes significados según el contexto; es más, la falta de contexto claro en oraciones cortas y fragmentadas puede hacer difícil la desambiguación de términos. La ambigüedad y la presencia de palabras desconocidas reducen el índice de acierto o precisión de los etiquetadores gramaticales automáticos (Bonilla 2024).

Las disfluencias, entendidas como fenómenos orales que interrumpen el flujo del habla (Crible *et al.* 2019), como las pausas, las autointerrupciones, las repeticiones, los falsos inicios, los alargamientos vocálicos o las pausas sonoras (o *fillers*, como *eh*), complican el trabajo de etiquetado automático. Las disfluencias son un fenómeno característico del habla, especialmente de la coloquial o informal (aunque también de la formal). Tradicionalmente, se han considerado un obstáculo para la comunicación; sin embargo, en la actualidad, los expertos consideran que contribuyen al proceso de comunicación, facilitando la comprensión del discurso (Barr y Seyfeddinipur 2009). Las disfluencias ayudan a desambiguar el significado, evitando interpretaciones erróneas, permiten al interlocutor predecir el tema y apreciar las intenciones del hablante (si existe duda, si se pretende matizar un significado, etc.). Es decir, tienen una función pragmática que el cerebro humano sabe descodificar e interpretar. No obstante, estos fenómenos son un inconveniente para un etiquetador gramatical automático y para el PLN en general.

Los contextos social y cultural influyen enormemente en la generación del habla coloquial. Las referencias a eventos, personajes o chistes específicos de la cultura o del grupo social pueden resultar difíciles de interpretar sin el contexto adecuado. Si a esto añadimos las diferencias regionales (dialectos) y/o sociales (sociolectos) en el uso del lenguaje, estos usos pueden impedir que los modelos generalicen adecuadamente todas las formas de habla (Farasyn *et al.* 2022; Jamatia *et al.* 2015; Jørgensen *et al.* 2015; Rozovskaya *et al.* 2006).

La innovación y creatividad lingüísticas, tan propias de una lengua viva, implican la creación constante de nuevas palabras y expresiones en el lenguaje coloquial que pueden dejar obsoletos rápidamente a los modelos entrenados (Taulé *et al.* 2015). Los modelos necesitan actualizarse constantemente para mantenerse al día con los cambios en el uso del lenguaje.

Por último, a pesar de no ser fenómenos propios de la oralidad, la escritura coloquial también presenta retos para el etiquetado gramatical automático que merecen ser mencionados. Así, en la escritura también se producen alteraciones gráficas (dejando de lado los errores involuntarios) que obedecen a la innovación y la creatividad, como el uso intencionado de errores ortográficos y tipográficos con efectos estilísticos o humorísticos, el uso de formas abreviadas y siglas o el uso de símbolos y emoticonos, que pueden alterar el significado del texto y complicar el análisis gramatical (Neunerdt *et al.* 2013).

4. Propuestas y prospectiva de investigación

Abordar estos desafíos implica desarrollar modelos de PLN que sean robustos y adaptables. El uso de técnicas avanzadas de aprendizaje profundo, entrenadas con grandes y variados conjuntos de datos que incluyan ejemplos de habla coloquial, así como la implementación de algoritmos que puedan manejar la corrección ortográfica y la contextualización, son enfoques necesarios para mejorar la precisión del etiquetado gramatical en textos coloquiales en cualquier lengua. Además, la combinación de enfoques basados en reglas con modelos estadísticos y de aprendizaje automático puede ayudar a mejorar la precisión del etiquetado gramatical en textos coloquiales.

La investigación sobre el etiquetado gramatical ha identificado varios retos, incluyendo la necesidad de reducir la ambigüedad espuria (percepción incorrecta de que una oración tiene múltiples interpretaciones cuando, en realidad, solo hay una interpretación lógica o contextual correcta) en la gramática de dependencias categoriales (Alfared 2012), el impacto de las palabras desconocidas y la ambigüedad en la precisión (Anbananthen 2017) y el problema de etiquetar palabras desconocidas de manera eficiente (Gupta 2014). Para abordar estos desafíos, se han propuesto varios enfoques. Alfared (2012) sugiere usar un etiquetador gramatical automático destinado a mejorar la tasa de ambigüedad espuria, mientras que Anbananthen (2017) compara metodologías estocásticas y basadas en reglas para incrementar la precisión del etiquetado automático. Gupta (2014) propone un modelo para etiquetar palabras desconocidas y Tsai (2003) introduce un modelo de reglas de contexto para lograr una alta precisión y reducir la

revisión manual. Estos estudios, en conjunto, resaltan la importancia de abordar la ambigüedad, las palabras desconocidas y la escasez de datos para entrenar a los etiquetadores gramaticales.

A grandes rasgos, se podría hablar de cuatro perspectivas de investigación principales: modelos de etiquetado específicos para el habla coloquial, incrustaciones contextuales y redes neuronales, aprendizaje activo y colaboración abierta y métricas de evaluación y puntos de referencia.

En primer lugar, es necesario contar con modelos de etiquetado específicos para el habla coloquial. Desarrollar modelos de etiquetado gramatical automático especializados que se entrenen con corpus de habla coloquial puede ayudar a abordar muchos de los desafíos discutidos. La incorporación de características lingüísticas específicas del habla coloquial y el aprovechamiento de técnicas de adaptación del dominio lingüístico pueden mejorar la precisión de dichos modelos. Evidentemente, este es un trabajo arduo, ya que no solo supone recopilar y anotar corpus de habla coloquial, sino que, además, estos modelos deben tener en cuenta los distintos y variados dominios lingüísticos y deben ser específicos para cada lengua (Rozovskaya *et al.* 2006).

En segundo lugar, el uso de incrustaciones contextuales (*contextual embeddings*) y de redes neuronales, fruto del reciente avance en métodos de aprendizaje profundo, se ha mostrado prometedor en la solución de varios desafíos del PLN. Las incrustaciones contextuales consisten en otorgar significados dinámicos a las palabras según el contexto en el que aparecen, en contraposición a la tradicional asignación de significados estáticos que no tienen en cuenta el contexto (Kanade *et al.* 2020). En cuanto a las redes neuronales, son útiles para tareas de etiquetado de secuencias, predicción de la siguiente palabra, reconocimiento de patrones, etc. Aplicar estas técnicas al etiquetado gramatical automático de conjuntos de datos de habla coloquial puede capturar matices de significado contextuales y mejorar el rendimiento general de los etiquetadores.

En tercer lugar, aprovechar los métodos de aprendizaje activo y la colaboración abierta (*crowdsourcing*) puede ayudar a construir corpus anotados de habla coloquial a gran escala. Utilizar anotadores humanos especializados en lenguaje coloquial puede proporcionar datos valiosos para el entrenamiento y la evaluación de modelos de etiquetado gramatical automático en este registro lingüístico.

Por último, desarrollar métricas de evaluación y crear puntos de referencia (*benchmarks*) estandarizados específicamente adaptados al etiquetado gramatical del habla coloquial es vital para medir su rendimiento (Tintinago *et al.* 2018). Las métricas de evaluación comparan

valores como la precisión y la exhaustividad en el resultado de los etiquetadores. Los puntos de referencia son conjuntos de datos y estándares utilizados para evaluar y comparar diferentes modelos de etiquetado gramatical. Estos dos procedimientos pueden usarse para valorar objetivamente diferentes modelos de etiquetado e identificar las áreas de mejora.

5. Conclusiones

El etiquetado gramatical automático de corpus de habla coloquial presenta numerosos desafíos debido, entre otros fenómenos propios del registro, a la variedad léxica, a la ambigüedad, a las disfluencias, a los contextos social y cultural o a la innovación y la creatividad lingüísticas. Abordar estos desafíos requiere más investigación y desarrollo en la creación de modelos especializados en los distintos dominios lingüísticos, el aprovechamiento de la información contextual, la adopción de técnicas de aprendizaje activo y la comparación y evaluación de etiquetadores. A estos retos, se añade la dificultad de crear etiquetadores estándar para una lengua, ya que es imposible contar con un corpus que recoja por completo todos los dominios lingüísticos para entrenar a la máquina. A medida que aumente el interés en procesar datos de habla informal, los avances en el etiquetado gramatical para el lenguaje coloquial permitirán un PLN más preciso y aplicaciones en escenarios del mundo real.

BIBLIOGRAFÍA

- Alfred, Ramadan, y Denis Béchet (2012). «POS taggers and dependency parsing», *International Journal of Computational Linguistics and Applications*, 3 (2): 107-122.
- Anbananthen, Kalaiarasi S. M., Jaya K. Krishnan, M. Shohel Sayeed y Praviny Muniapan (2017), «Comparison of stochastic and rule-based POS tagging on Malay online text», *American Journal of Applied Sciences*: 14 (9), 843-851.
- Barr, Dale J., y Mandana Seyfeddinipur (2009), «The role of fillers in listener attributions for speaker disfluency», *Language and Cognitive Processes*, 25 (4): 441-455. DOI: 10.1080/01690960903047122.
- Bolaños, Sergio (2015), «La lingüística de corpus: perspectivas para la investigación lingüística contemporánea», *Forma y Función*, 28 (1): 31-54. DOI: 10.15446/fyf.v28n1.51970.

- Bonilla, Johnatan. E. (2024), «Spoken Spanish POS tagging: gold standard dataset», *Language Resources and Evaluation*: 1-30. DOI: 10.1007/s10579-024-09751-x.
- Brill, Eric D. (1993), *A corpus-based approach to language learning*, Philadelphia, University of Pennsylvania.
- Briz, Antonio (2016), «Español coloquial», en Javier Gutiérrez-Rexach (ed.) *Enciclopedia de lingüística hispánica*, vol. 2, Londres/Nueva York, Routledge: 463-476.
- Castillo Velásquez, Francisco A., José Luis Martínez Godoy, María del Consuelo P. Torres Falcón, Jonny P. Zavala De Paz, Adela Becerra Chávez, y José A. Rizzo Sierra (2020), «Atribución de autoría de mensajes de Twitter a través del análisis sintáctico automático», *Research in Computer Science*, 149 (11): 91-101.
- Cherradi, Mohamed, y Anass Haddadi (2024), «Exploration of scientific documents through unsupervised learning-based segmentation techniques», *Seminars in Medical Writing and Education*, 3: 1-9. DOI: 10.56294/mw202468.
- Chiche, Alebachew, y Betselot Yitagesu (2022), «Part of speech tagging: a systematic review of deep learning and machine learning approaches», *Journal of Big Data*, 9: 1-25. DOI: 10.1186/s40537-022-00561-y.
- Crible, Ludivine, Amandine Dumont, Lulia Grosman y Ingrid Notarrigo (2019), «(Dis)fluency across spoken and signed languages: application of an interoperable annotation scheme», en Liesbeth Degand, Gaëtanelle Gilquin, Laurence Meurant y Catherine Simon (eds.) *Fluency and disfluency across languages and language varieties*, Lovaina, Presses universitaires de Louvain: 17-40.
- Farasyn, Melisa, Anne-Sophie Ghyselen, Jacques Van Keymeulen y Anne Breitbarth (2022), «Challenges in tagging and parsing spoken dialects of Dutch», *Journal of Historical Syntax*, 6 (4-11): 1-36.
- Garrote, Marta (2010), *Los corpus de habla infantil: metodología y análisis*, Madrid, UAM Ediciones.
- Ghosh, Soumitra, y Brojo Kishore Mishra (2020), «Parts-of-speech tagging in NLP: utility, types, and some popular POS taggers», en Brojo Kishore Mishra y Raghvendra Kumar (eds.), *Natural language processing in artificial intelligence*, Palm Bay, Apple Academic Press: 131-165.

- Gupta, Aastha, Rachna Rajput, Richa Gupta, y Monika Arora (2014), «Improved POS tagging for unknown words», *International Journal of Soft Computing and Engineering*, 4: 47-50.
- Jamatia, Anupam, Jjörn Gambäck y Amitava Das (2015), «Part-of-speech tagging for code-mixed English-Hindi Twitter and Facebook chat messages», *Proceedings of Recent Advances in Natural Language Processing*: 239–248.
- Jørgensen, Anna K., Dirk Hovy y Anders Søgaard (2015), «Challenges of studying and processing dialects in social media», *Proceedings of the Workshop on Noisy User-Generated Text*: 9-18.
- Kanade, Aditya, Petros Maniatis, Gogul Balakrishnan y Kensen Shi (2020), «Learning and evaluating contextual embedding of source code», *International Conference on Machine Learning*: 5110-5121.
- Kumawat, Deepika, y Vinesh Jain (2015), «POS tagging approaches: a comparison», *International Journal of Computer Applications*, 118 (6): 32-38.
- Landolsi, Mohamed Y., Lotfi Ben Romdhane, y Lobna Hlaoua (2024), «Hybrid medical named entity recognition using document structure and surrounding context», *The Journal of Supercomputing*, 80 (4): 5011-5041.
- Martínez, Ángel R. (2012), «Part-of-speech tagging», *Wiley Interdisciplinary Reviews: Computational Statistics*, 4 (1): 107-113. DOI: 10.1002/wics.195.
- Neunerdt, Melanie, Bianca Trevisan, Michael Reyer y Rudolf Mathar (2013), «Part-of-Speech tagging for social media texts», en Iryna Gurevych, Chris Biemann y Torsten Zesch (eds.) *Language processing and knowledge in the web: lecture notes in computer science*, Berlín/Heidelberg, Springer: 139-150. DOI: 10.1007/978-3-642-40722-2_15.
- Rojo, Guillermo (2021), *Introducción a la lingüística de corpus en español*, Londres/Nueva York, Routledge.
- Rozovskaya, Alla, Richard Sproat, y Elabbas Benmamoun (2006) «Challenges in processing colloquial Arabic», *Proceedings of the International Conference on the Challenge of Arabic for NLP/MT*: 4-14.
- Sánchez-Cartagena, Víctor M., Juan Antonio Pérez-Ortiz, y Felipe Sánchez-Martínez (2024), «Understanding the effects of word-level linguistic annotations in under-resourced neural machine translation», *arXiv*: 2401.16078. DOI: 10.48550/arXiv.2401.16078.

- Taulé, Mariona, M. Antonia Martí, Ann Bies, Montserrat Nofre, Aina Garí, Zhiyi Song, Stephanie Strassel, y Joe Ellis, J. (2015), «Spanish treebank annotation of informal non-standard web text», en *Current Trends in Web Engineering: 15th International Conference, ICWE 2015 Workshops*, Rotterdam, Springer International Publishing: 15-27.
- Tintinago, Alfonso, Yordan Muñoz, Gustavo A. Uribe, y Pedro H. Álvarez (2018), «Etiquetado asistido de documentos de investigación mediante procesamiento de lenguaje natural y tecnologías de la web semántica», *Scientia et Technica*, 23 (4), 528-537.
- Tsai, Yu-Fang, y Keh-Jiann Chen (2003), «Context-rule model for POS tagging», *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation*, ACL Anthology: 146-151.
- Ying, Zelin, Chen Li, Yu Dong, Qiuqiang Kong, Qiao Tian, Yuanyuan Huo, y Yuxuan Wang (2024), «A unified front-end framework for English text-to-speech synthesis», *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*: 10181-10185.