Propuesta recibida en noviembre de 2018 y aceptada para publicación en octubre de 2019.

# Formalización de reglas para la detección del plural en castellano en el caso de unidades no diccionarizadas

**Formalization of rules for the detection of plurals in Spanish in the case of out-of-vocabulary units**

Rogelio Nazar
Instituto de Literatura y Ciencias del Lenguaje
Pontificia Universidad Católica de Valparaíso
rogelio.nazar@pucv.cl

Amparo Galdames
Instituto de Literatura y Ciencias del Lenguaje
Pontificia Universidad Católica de Valparaíso
amparo.galdames@pucv.cl

**Resumen**

En este artículo ofrecemos una formalización de reglas de pluralización en castellano para ser utilizada concretamente en el procesamiento de términos especializados, ya que con frecuencia estos no se encuentran registrados en los diccionarios de lengua general y, por tanto, no son reconocidos su categoría y lema. Esto tiene consecuencias negativas en tareas como la extracción de terminología, especialmente en el caso de lenguas con riqueza morfológica. Enfrentamos el problema con un diseño en forma de cascada de reglas de sustitución, expresiones regulares y adquisición léxica a partir de corpus de grandes dimensiones. Los resultados experimentales muestran una reducción significativa de la tasa de error de dos etiquetadores ampliamente utilizados: TreeTagger y UDPipe. Ofrecemos una implementación en código abierto que funciona como posproceso del etiquetado.

**Palabras clave**

etiquetado morfosintáctico, lematización, reglas de pluralización del castellano, unidades no diccionarizadas

**Abstract**

This paper presents a formalization of rules on plural formation in Spanish to be used in the processing of specialized terminology, as it is frequently the case that terms are not found in dictionaries of general language and therefore they cannot be lemmatized or POS-tagged. The absence of terms in general dictionaries has negative effects in tasks such as terminology extraction, particularly in the case of morphologically rich languages. We attack the problem by cascading through multiple trasnfser rules, regular expressions and lexical aquisition from large corpora. Results show significant reduction of the error rate of two POS-taggers: TreeTagger and UDPipe. We offer an open-source implementation which works as a post-process, cleaning up after the tagger.

**Keywords**

part-of-speech tagging, lemmatization, rules for plural in Spanish, out-of-vocabulary units

## 1 Introducción

Una de las tareas más elementales del área del procesamiento del lenguaje natural (PLN) es el etiquetado morfosintáctico o POS-tagging. Esto incluye la lematización de las formas que aparecen en el texto y la asignación de una categoría gramatical, con o sin análisis morfológico específico, que indique persona, género, número, etc. (Manning & Schütze, 1999). En otras palabras, se trata de reconocer la asociación entre una palabra flexionada y, por un lado, su lema o lexema —la forma elegida como entrada en los diccionarios— y, por otro lado, la categoría gramatical.

Los etiquetadores morfosintácticos fueron diseñados inicialmente como sistemas basados en reglas (Greene & Rubin, 1971), pero fueron gradualmente reemplazados por familias de algoritmos probabilísticos (Church, 1989; Schmid, 1994) y, más recientemente, neuronales (Ling et al., 2015; Straka & Straková, 2017; Qi et al., 2018). Estos últimos han mejorado sustancialmente la calidad del resultado y se han aplicado también al análisis sintáctico. El problema general de los métodos cuantitativos, sin embargo, es que si bien reducen el esfuerzo de la creación de reglas, exigen el de la producción de material de entrenamiento en forma de un corpus ya etiquetado o –al menos– revisado manualmente. El inconveniente radica en que entre el tamaño de este corpus de entrenamiento y la calidad del resultado existe una relación que parece describir un modelo logarítmico: se requiere cada vez más corpus de entrenamiento para obtener cada vez menos mejora en la calidad del resultado. Esta circunstancia

DOI: 10.21814/lm.11.2.285

representa una motivación para explorar formas híbridas de etiquetadores probabilísticos con reglas desarrolladas de manera manual o, al menos, un proceso posterior al etiquetado que corrija algunos de los errores. En el pasado se exploraron alternativas en este sentido, con algoritmos que complementaban sistemas de reglas con análisis probabilístico (Brill, 1992, 1995), pero no representan una tendencia actual, al menos no en este campo.

Está claro, en cualquier caso, que la tasa de error de los etiquetadores se ha ido reduciendo progresivamente. Sin embargo, y a pesar de estas mejoras, al día de hoy los problemas de los etiquetadores persisten. La desambiguación sigue siendo un desafío, ya que en muchos casos depende del sentido de la oración o incluso de la intención del emisor, con todos los matices y sutilezas que puedan darse (la ironía, el humor, etc.). Otro desafío en el análisis morfosintáctico es el de las unidades no diccionarizadas (out-of-vocabulary units; en adelante, UND), es decir, formas cuyo lema no es posible reconocer porque no está en el diccionario o corpus de entrenamiento del etiquetador. Esta es una dificultad especialmente frecuente cuando se trabaja en terminología, en que existe gran innovación léxica. Todo ello, además, tiene como consecuencia que no es posible proporcionar ni la etiqueta ni el lema de la forma analizada. Aplicar al procesamiento automático de textos especializados analizadores morfosintácticos que no cuentan con información morfológica específica de los términos genera una mayor tasa de error, lo que afecta tareas posteriores como la extracción de terminología. El problema se agudiza en particular en el caso de lenguas morfológicamente ricas como las romances.

Los Cuadros 1 y 2 ilustran la problemática. Se trata de fragmentos del resultado del análisis morfosintáctico de un corpus especializado (cf. Sección 3.4) después de haber sido sometido al analizador morfosintáctico TreeTagger (Schmid, 1994), que produce lema y categoría gramatical de las palabras ingresadas, y al analizador UDPipe (Straka & Straková, 2017), que además de esto ofrece un análisis sintáctico completo (nivel de análisis que para los fines del presente estudio no hemos tenido en cuenta). En general, se aprecia una mejora en la calidad de la lematización del segundo respecto al primero, pero en ambos casos es fácil advertir también la persistencia de errores cuando se trabaja en corpus especializados. En ambos casos se observa la lematización errónea de unidades léxicas desconocidas para el etiquetador, como *dopaminérgicos* en un caso y *micro-diálisis* y *demuestra* en el otro.

| **Forma** | **POS-tag** | **Lema** |
|---|---|---|
| Existen | VLfin | existir |
| grupos | NC | grupo |
| de | PREP | de |
| fármacos | NC | fármaco |
| que | CQUE | que |
| tienen | VLfin | tener |
| afinidad | NC | afinidad |
| hacia | PREP | hacia |
| los | ART | el |
| diversos | QU | diversos |
| receptores | NC | receptor |
| dopaminérgicos | ADJ | UNKNOWN |
| . | FS | . |

Cuadro 1: Resultado con TreeTagger. Se advierte error en la lematización de la forma *dopaminérgicos*.

| **Forma** | **POS-tag** | **Lema** |
|---|---|---|
| Es | AUX | ser |
| así | ADV | así |
| como | SCONJ | como |
| Gingrich | PROPN | Gingrich |
| , | PUNCT | , |
| in | NOUN | in |
| vivo | ADJ | vivo |
| con | ADP | con |
| micro-diálisis | NOUN | micro-diálisi |
| demuestra | VERB | demuestro |
| liberación | NOUN | liberación |
| de | ADP | de |
| DA | PROPN | Da |
| en | ADP | en |
| el | DET | el |
| N | PROPN | N |
| Acc | PROPN | Acc |

Cuadro 2: Resultado con el etiquetador UDPipe. Se advierte error en la lematización de las formas *micro-diálisis* y *demuestra*

En este contexto, el presente trabajo tiene por objetivo proponer la formalización de las reglas específicamente para la pluralización en castellano. La propuesta no pretende resolver todos los problemas; sin embargo, solo reconocer la flexión de número en sustantivos y adjetivos, como un proceso posterior al etiquetado, ya implicaría una mejora sustancial de la calidad del resultado. El método tiene la forma de una cascada de reglas de sustitución y expresiones regulares para restituir el lema de formas del plural en sustantivos y adjetivos. La propuesta se centra en el caso específico de los términos especializados porque es

el campo en el que el problema de las UND tiene mayor incidencia. En una serie de ensayos en un corpus especializado compuesto por artículos de investigación de una revista de neuropsiquiatría, obtuvimos mejoras significativas en la calidad del etiquetado de TreeTagger y de UDPipe para el caso específico del reconocimiento del plural.

La siguiente sección proporciona los antecedentes teóricos asociados al problema de las UND en el PLN en general, el POS-tagging en castellano en particular y las normas de pluralización en esta lengua, foco de interés de la presente investigación. Posteriormente, la Sección 3 presenta la propuesta metodológica que incluye la descripción del algoritmo diseñado y las reglas construidas para la lematización de las UND. La Sección **??** expone los resultados de una evaluación en el corpus de neuropsiquiatría. Finalmente, la Sección **??** presenta nuestras conclusiones y sugerencias para trabajo futuro. Hemos implementado este algoritmo en un prototipo en lenguaje Perl para que funcione como posproceso del etiquetado. Un demostrador del prototipo y su código fuente están disponibles en el sitio web del proyecto[1].

## 2 Marco Teórico

### 2.1 El problema de las unidades no diccionarizadas

El problema de las UND ha sido explorado en distintas áreas del PLN, como la adquisición léxica y el reconocimiento de habla (automatic speech recognition), donde las UND se presentan normalmente en forma de nombres propios de persona o lugar, pero también en unidades del vocabulario general, ya que este se encuentra en permanente cambio (Manning & Schütze, 1999; Bazzi & Glass, 2002; Bazzi, 2002; Parada et al., 2011; Qin, 2013). En el caso del reconocimiento de habla, el problema de las UND es particularmente agudo porque multiplica los errores en el reconocimiento de palabras vecinas: "When encountering an OOV word, the recognizer will incorrectly recognize the OOV word with one or more similar sounding in-vocabulary (IV) words. In addition, OOV words also affect the recognition performance of their surrounding IV words" (Qin, 2013, p. 3).

En el caso del POS-tagging, las UND han sido una de las primeras dificultades a superar: "Unknown words are a major problem for taggers, and in practice, the differing accuracy of different taggers over different corpora is often mainly determined by the proportion of unknown words" (Manning & Schütze, 1999, p. 351). Algunas propuestas para resolver el problema incluyen recurrir a pistas morfológicas calculando las probabilidades de combinación de raíces, sufijos y categorías morfológicas: "If we have never seen the word 'rakishly', then knowledge that 'ly' typically ends an adverb will improve our accuracy on this word –similarly, for 'randomizing' [etc.]" (Charniak et al., 1993, p. 787).

Manning (2011) calculó en 4.5 % el porcentaje de error que las UND representan respecto de la totalidad de errores cometidos en inglés, pero hay que suponer que la tasa de error siempre será más alta en el caso de los corpus especializados y en lenguas morfológicamente ricas. En castellano, no conocemos datos del porcentaje que representan las UND frente al total de los errores cometidos por los etiquetadores, pero esto variará de la misma manera en función de la naturaleza del texto analizado.

El mismo Manning (2011) y otros autores (*cf.* Biemann, 2006) propusieron nuevas formas de pensar el problema, entre las que se reconoce la aplicación de medidas de similitud distribucional para comparar las UND con las unidades que sí están en el vocabulario. Asimismo, la idea del modelamiento morfológico en la línea de Charniak et al. (1993) también ha sido explorada en distintas investigaciones posteriores para el tratamiento de las UND (Adams et al., 1994; Creutz et al., 2007). Una de las propuestas que se reconoce como la más reciente es la aplicación de las representaciones vectoriales de las secuencias de caracteres al interior de las palabras, secuencias transportadoras de información morfosintáctica (Santos & Zadrozny, 2014; Ling et al., 2015).

### 2.2 El POS-tagging en castellano

Al igual que en el caso del inglés y de otras lenguas, los primeros etiquetadores morfosintácticos para el castellano eran sistemas basados en reglas codificadas de manera manual, como sería el caso de Moreno & Goni (1995). Pero eventualmente fueron también los algoritmos probabilísticos los que fueron ganando popularidad. Entre estos destaca el ya mencionado TreeTagger (Schmid, 1994), basado en aprendizaje automático que está entre los más utilizados actualmente, al menos en el caso de las lenguas romances (Allauzen & Bonneau-Maynard, 2008; Parra Escartín & Martínez Alonso, 2015). Otras propuestas incluyen sistemas híbridos que combinan análisis probabilístico con sistemas de reglas manualmente codificadas, como Freeling (Carreras et al., 2004).

[1] http://www.tecling.com/pullpos

La lematización y el etiquetado morfosintáctico han cobrado nuevo impulso en el último tiempo, tanto en castellano como en otras lenguas. La aparición de nuevos productos parece sugerir que se producirán cambios en el escenario de los etiquetadores y que el análisis de dependencias sintácticas puede tener un rol en la desambiguación de categorías gramaticales. El POS-tagger que propuso Manning (2011) está basado en un clasificador que funciona con modelos de máxima entropía (Maximum Entropy models) y está disponible para el castellano entre otras lenguas, aunque no lematiza. El proyecto de las 'IXA pipes' (Agerri et al., 2014) contiene un POS-tagger basado en el mismo tipo de modelo, pero a diferencia del anterior sí incluye lematización, para lo cual utiliza un diccionario de 600.000 entradas que a su vez es expandido mediante la herramienta de análisis morfológico Morfologik (Miłkowski, 2010). Han aparecido recientemente distintas propuestas de etiquetadores que pueden aplicarse al castellano, como MateTools (Bohnet & Nivre, 2012) o spaCy (Honnibal, 2016; Honnibal & Johnson, 2015). En particular destacamos el sistema Lemming (Müller et al., 2015) porque incorpora algunas de las ideas que se presentan también en este artículo, tales como observar la frecuencia de aparición en corpus (Wikipedia en el caso de Lemming) para ponderar la validez de un candidato a lema. Sin embargo, su enfoque es distinto al nuestro, entre otras razones porque utilizan recursos externos, tales como el diccionario ASPELL.

Al margen de lo anterior, la razón fundamental de la renovación del interés por los etiquetadores viene dada por los nuevos avances en el campo de las redes neuronales, ya que dada su naturaleza también pueden aplicarse al castellano. En particular, el modelo Bi-LSTM (*Bidirectional long short-term memory*, un subtipo de redes neuronales recurrentes) es el que actualmente está dando mejores resultados (Ling et al., 2015; Plank et al., 2016) al aplicarse a la representación vectorial no ya de palabras (*word embeddings*), sino de fragmentos inferiores a la palabra (*subword* o *subtoken representations*). Trabajar al nivel inferior a la palabra resulta ser la clave para la lematización de las UND porque se puede generalizar a partir de la información morfológica de la palabras que sí son conocidas.

Dentro de la familia de algoritmos a los que se ha hecho referencia, destacamos el ya mencionado UDPipe (Straka et al., 2016; Straka & Straková, 2017), uno de los sistemas más recientes y con mejor desempeño tanto en etiquetado morfológico como en análisis sintáctico, que resultó el mejor entre 26 participantes de la *CoNLL 2018 UD Shared Task* (Straka, 2018). Cabe destacar, además, que ya es posible aplicarlo a una amplia variedad de lenguas.

Una evaluación extensiva de todos los sistemas existentes escapa a los límites del presente artículo. Sin embargo, ya un examen superficial de los distintos etiquetadores revela que aún existe amplio margen de mejora. Cuando los autores informan un 99 % de precisión, es necesario tener en cuenta, como ya se advirtió, que en corpus especializado el desempeño puede ser muy inferior. Además, por regla general, cuando se informan estos porcentajes de precisión, esto se hace teniendo en cuenta la precisión por token ("token ratio"), es decir, que se incluyen en el conteo aquellos tokens que solo tienen una etiqueta y una lematización. Incluso si se confirmaran estos diagnósticos optimistas sobre el desempeño de los etiquetadores, dos o tres errores cada cien palabras resultarán en una proporción mucho mayor de oraciones mal analizadas ("per-sentence ratio").

### 2.3 La pluralización en castellano

El plural en castellano es relativamente sencillo en comparación con otras lenguas que cuentan con una variación numérica como el dual, para dos elementos, o el paucal, para un grupo reducido de elementos (Dixon, 2009). Sin embargo, y a pesar de la aparente simplicidad del sistema castellano, la flexión de número ha sido un fenómeno que, en conjunto con la flexión de género, ha suscitado la atención de muchos gramáticos (Sánchez Corrales, 1994).

En castellano, los adjetivos tienen flexión de género y número y los sustantivos solo aparecen en su forma singular o plural, más allá de casos específicos como los nombres de cargos o profesiones, que también presentan flexión de género. En el plural, la flexión responde a una serie de restricciones que considera, entre otros aspectos, la estructura fonológica de la forma singular (Cedeño et al., 2014). De esta manera, el sonido, la acentuación y la estructura silábica han sido aspectos comunes en gran parte de las propuestas que explicitan una variedad de especificaciones en las normas para la formación de los plurales.

Dentro de las primeras formalizaciones que podemos reconocer para la flexión de número se encuentra la *Gramática de la lengua castellana* (De Nebrija, 1492). En esta publicación se utilizan ya los términos de singular y plural y se proponen las normas que permiten formar los plurales, básicamente el uso de las terminaciones *-s* y *-es*. Posteriormente, la *Ortografía española* (Real Academia Española, 1741), la *Gramática*

de Andrés Bello (1847), la *Gramática de la lengua castellana* (Real Academia Española, 1920) y, más recientemente, la Nueva gramática de la lengua española (Real Academia Española, 2009) han mantenido estas normas en lo esencial.

Alemany (1920) también mantiene en esencia las mismas normas, pero las complejiza al advertir que las marcas de la pluralización dependen de ciertas características del lema principal. Estas suelen asociarse a la combinatoria de letras en el término de la palabra, su sílaba átona o tónica y su dependencia gramatical en la sintaxis. Sin embargo, este autor aborda también el tema de la pluralización de los compuestos, a los que distingue entre perfectos e imperfectos. El primero corresponde a aquellos sustantivos que combinan dos elementos nominales (adjetivos+sustantivos, sustantivos+sustantivos, adjetivo+adjetivo, etc.) conformando una nueva forma nominal (*ferrocarril, portafusil*). Esta nueva construcción aceptaría el plural en su segundo término. Por su parte, los compuestos imperfectos son las formas que se componen por más de una palabra ortográfica, conformando así unidades poliléxicas. Dentro de estos casos, la pluralización interna se encarga de marcar solo el primer componente, que además es el núcleo del término (*hombres rana, cartas suicida, palabras clave*). El tema ha sido desarrollado por autores contemporáneos como Moyna (2011) o de León (2015). En el presente artículo, sin embargo, nos restringimos a la construcción de los plurales de unidades monoléxicas y fuera de contexto ("types").

Otros fenómenos dignos de atención respecto a la pluralización en español son 1) el caso del morfema ∅ (morfema cero) en sustantivos como *lunes, martes* o *paréntesis*, ya que en su forma singular acaban en *-s* y que no son agudos y reconocen el mismo uso tanto en su forma singular como plural (Real Academia Española, 1920); y 2) el de los *pluralia tantum*, que remiten generalmente a un solo objeto pero compuesto (*tijeras, pinzas, pantalones*).

La revisión de la bibliografía permite identificar que existen numerosos trabajos enfocados en la descripción de la pluralización español (Stockwell et al., 1965; Saporta, 1965; Foley, 1967; Alcina & Blecua, 1975; Hernández Alonso, 1984; Ambadiang, 1999, entre otros), y todos coinciden en que las marcas que afectan a esta flexión se identifican con *-s* y *-es*, dependiendo de los aspectos estructurales del lema y de otras características fonológicas de la palabra. Como tendencia general, se puede identificar que para la forma del singular terminada en vocal, suele añadirse *-s* a su forma plural; mientras que para la forma singular que acaba con consonante, es más recurrente añadir *-es* para su forma plural.

Entre las obras consultadas, nos hemos decantado por la sistematización que ofrece el *Diccionario panhispánico de dudas* (Asociación de Academias de la Lengua Española, 2005) para la implementación de las reglas. La obra no es del todo reciente y algunas de las reglas son de naturaleza normativa, lo que nos obligaría en el futuro a hacer algunos ajustes. Sin embargo, nos pareció una buena síntesis al menos para esta etapa inicial del proyecto. A continuación se presentan los 10 casos principales entre los 17 que regularizan la flexión del plural, ya que los restantes corresponden a particularidades asociadas a latinismos, notas musicales, plural de nombres de las letras del abecedario, abreviaturas y símbolos, entre otros.

(a) Sustantivos y adjetivos terminados en vocal átona o en *-e* tónica. Forman el plural con *-s* (*casas, estudiantes, taxis, planos, tribus, comités*).

(b) Sustantivos y adjetivos terminados en *-a* o en *-o* tónicas. forman el plural únicamente con *-s*. (*sofás, rococós, dominós*).

(c) Sustantivos y adjetivos terminados en *-i* o en *-u* tónicas. Admiten generalmente dos formas de plural, una con *-es* y otra con *-s*, aunque en la lengua culta suele preferirse *-es*. (*bisturíes/bisturís, carmesíes/carmesís, tabúes/tabús*).

(d) Sustantivos y adjetivos terminados en *-y* precedida de vocal. Forman tradicionalmente su plural con *-es*. (*rey/reyes; ley/leyes; buey/bueyes*). Los sustantivos y adjetivos con esta misma terminación que se han incorporado provienen por lo general de otras lenguas. Aplica para este caso que la *y* del singular pasar a escribirse *i* (*espray/espráis; yóquey/yoqueis*).

(e) Voces extranjeras terminadas en *-y* precedida de consonante. Deben adaptarse gráficamente al español sustituyendo la *-y* por *-i*. Su plural se forma añadiendo una *-s*. (*dandis, pantis, ferris*).

(f) Sustantivos y adjetivos terminados en *-s* o en *-x*. Si son monosílabos o polisílabos agudos, forman el plural añadiendo *-es* (*tos/toses; vals/valses, fax/faxes; compás/compases; francés/franceses*). En el resto de los casos, permanecen invariables (*crisis/crisis; tórax/tórax; fórceps/fórceps*).

(g) Sustantivos y adjetivos terminados en *-l, -r, -n, -d, -z, -j*. Si no van precedidas de otra consonante, forman el plural con *-es*. Los extranjerismos deben aplicar la misma regla (*dócil/dóciles; color/colores; pan/panes; césped/céspedes; cáliz/cálices; reloj/relojes*).

(h) Sustantivos y adjetivos terminados en consonantes distintas de *-l, -r, -n, -d, -z, -j, -s, -x, -ch* pluralizan en *-s*. Esta norma incluye onomatopeyas o voces procedentes de otros idiomas (*crac/cracs; zigzag/zigzags; esnob/esnobs; chip/chips; mamut/mamuts; cómic/cómics*).

(i) Sustantivos y adjetivos terminados en *-ch*. Procedentes todos ellos de otras lenguas, o bien se mantienen invariables en plural (*(los) crómlech, (los) zarévich, (los) pech*), o bien hacen el plural en *-es* (*sándwich/sándwiches; maquech/maqueches*).

(j) Sustantivos y adjetivos terminados en grupo consonántico. Procedentes todos ellos de otras lenguas, forman el plural con *-s* salvo aquellos que terminan ya en *-s*, y que siguen la regla (f) (*gong/gongs; iceberg/icebergs; récord/récords*). Se exceptúan de esta norma las voces *compost, karst, test, trust y kibutz*, que permanecen invariables en plural, porque añadir *-s* en estos casos daría lugar a una secuencia de difícil articulación en castellano.

Creemos que una sistematización e implementación computacional de estas reglas puede representar una ayuda para identificar el plural de lemas no incluidos en los diccionarios, mejorando así la calidad del trabajo en el procesamiento de terminología especializada.

## 3 Metodología

La propuesta metodológica está basada en un sistema que se puede aplicar con posterioridad al POS-tagger para detectar los casos de error con los plurales y corregirlos asignando el lema correspondiente del plural encontrado, si es que se trata realmente de una forma plural, porque si la unidad encontrada no es un plural, se asignará la misma forma como lema. Describimos a continuación los pasos del algoritmo que hemos diseñado para esta tarea. En primer lugar, la Sección 3.1 describe la adquisición léxica de un corpus de referencia de gran tamaño. La Sección 3.2 describe la implementación de las reglas de pluralización en castellano descritas ya en la Sección 2.3, pero utilizando la frecuencia de aparición de los elementos observada en el corpus de referencia como fundamento para la toma de decisión. La Sección 3.3 describe el proceso de extracción de parejas singular-plural por medio de la aplicación de las reglas de pluralización al corpus de referencia.

El proceso de adquisición léxica debe realizarse solamente una vez, ya que al completarse esta primera acción, el algoritmo pasa a la segunda fase, que es la que puede realizarse $n$ veces. Esta segunda fase corresponde al análisis de un corpus especializado en particular (Sección 3.4), de ahora en adelante designado como el corpus-objetivo, que ha sido previamente etiquetado con un POS-tagger. Sobre este corpus se procederá con la detección y corrección de errores.

### 3.1 Adquisición de un formario amplio

El punto de partida es la generación automática de un amplio formario del castellano, entendiendo por ello un listado de formas léxicas distintas. Para nuestros experimentos utilizamos un fragmento del corpus de castellano EsTenTen (Kilgarriff & Renau, 2013), compuesto por páginas web descargadas de manera aleatoria. El tamaño de la muestra utilizada es de aproximadamente dos mil millones de palabras sobre un total de $10^{10}$ que tiene el corpus.

El EsTenTen se ofrece ya etiquetado, pero nuestra metodología no utiliza esas etiquetas y por eso hemos procedido a eliminarlas dejando solo el texto plano (*cf.* comentarios al respecto en la Sección 5 al proyectar las posibilidades de trabajo futuro). En esta etapa del proceso, el algoritmo produce una tabla como la que se muestra en el Cuadro 3, con todo el vocabulario del corpus asociado a su frecuencia de aparición, reteniendo solamente aquellos elementos con frecuencia mínima $>= u$ ($u = 5$). El umbral es arbitrario pero obedece a cuestiones prácticas: un umbral muy bajo haría que el tamaño del formario sea demasiado grande como para manejarlo con facilidad, mientras que uno muy alto reduciría la cobertura del sistema. El tamaño del formario obtenido mediante este procedimiento es de 1.054.411 registros.

### 3.2 Reglas de pluralización

A partir de lo revisado sobre las normas de pluralización en castellano derivamos una serie de reglas para reconocer la relación entre un singular y un plural. En los casos de las formas más frecuentes, utilizamos simples reglas de sustitución porque consideramos que son un apuesta segura.

| **Forma** | **Frecuencia** |
|---|---|
| ... | ... |
| zigzagueaba | 13 |
| zigzagueaban | 7 |
| zigzagueado | 8 |
| zigzagueamos | 5 |
| zigzaguean | 42 |
| zigzagueando | 140 |
| zigzagueante | 258 |
| zigzagueantes | 108 |
| zigzaguear | 63 |
| zigzagueo | 54 |
| zigzagueos | 40 |
| zigzagues | 5 |
| zigzagueó | 6 |
| ... | ... |

Cuadro 3: Fragmento del formario del EsTenTen

Si se trata de terminaciones típicamente adjetivales o de nombres de profesiones, la forma singular remite directamente al masculino. En el resto de los casos, conservamos la forma terminada en *a*, ya que la decisión final se tomará en un momento posterior del análisis (Sección 3.4.3). En total, desarrollamos 55 reglas de este tipo. Los siguientes son algunos ejemplos:

```
-áceos     =>   -áceo
-ares      =>   -ar
-ari[ao]s  =>   -ario
-ciones    =>   -ción
-eces      =>   -ez
-eones     =>   -eón
-ices      =>   -iz
-idades    =>   -idad
-ines      =>   -ín
-siones    =>   -sión
```

Para los casos de palabras que no cumplen con este tipo de morfología, utilizamos cascadas de expresiones regulares que nos permiten establecer generalizaciones más amplias según la normativa de la pluralización del español. Para ello, tomamos como base la sistematización del *Diccionario panhispánico de dudas* expuesta en la Sección 2.3.

### 3.3 Derivación de parejas singular-plural a partir del formario extraído del corpus

A partir del formario obtenido en la Sección 3.1 y aplicando las reglas de la Sección 3.2, generamos un recurso léxico en forma de parejas singular-plural. Esta tabla utilizará luego el algoritmo para el análisis de cada corpus-objetivo.

Solo se admite una pareja en este diccionario si la proporción entre la frecuencia del singular y el plural se encuentran dentro de unos límites que se asignaron de manera empírica. Fundamentamos esta decisión en la expectativa de encontrar determinado equilibrio entre las formas singular y plural. Por ejemplo, si una palabra no aparece casi nunca en plural, se asume que en realidad no pluraliza. Definimos en (1) la razón $r$ entre la frecuencia observada de una forma plural $c(p)$ sobre la frecuencia de la forma singular $c(s)$, y en (2) una función binaria $a(p,s)$ que decide si el algoritmo admite o no la pareja $(p,s)$.

$$r(p,s) = \frac{c(p)}{c(s)+1} \tag{1}$$

$$a(p,s) = \begin{cases} 1 & x > r(p,s) < z \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

De esta forma, la función $a(p,s)$ controla que la razón $r(p,s)$ se encuentre dentro de la banda de tolerancia definida por los parámetros $x$ y $z$ ($x = 0{,}001 \wedge z = 120$). Esto impide que se produzca una disparidad excesiva entre la frecuencia del singular y del plural. En el primer caso se evitaría, por ejemplo, la unión de *algos*, con frecuencia 132, vs. *algo*, con frecuencia 1.186.957. En el caso opuesto, se evita por ejemplo la unión de *vívere*, con frecuencia 8, y *víveres*, con frecuencia 2.427.

A modo de orientación para la estimación de los parámetros $x$ y $z$, incluimos la Figura 1, en la que se puede apreciar el porcentaje de error que se obtiene tomando muestras aleatorias de 50 parejas en distintos intervalos. Tal como se puede apreciar en esa Figura, los valores de precisión disminuyen hacia los extremos mayor y menor.

Mediante este procedimiento se generó un total de 143.159 parejas (un fragmento se muestra en el Cuadro 4). Las parejas se componen principalmente de sustantivos y adjetivos, pero no porque el algoritmo utilice la información proporcionada por el etiquetador en el corpus de referencia. La razón, en cambio, es que son estas categorías gramaticales las que cumplen con las reglas de pluralización, y para una forma singular se ha encontrado efectivamente una coincidencia en el formario con una forma plural. Esto no es posible en el caso de otras categorías gramaticales, aunque sí se introducen errores que corresponden a formas en otras lenguas o nombre propios, pero la mayor parte de estos errores se corrigen en la etapas del análisis de un corpus-objetivo.

| **Plural** | **Singular** | **Frec 1** | **Frec 2** | **r(p,s)** | **a(p,s)** |
|---|---|---|---|---|---|
| ... | ... | ... | ... | ... | |
| luís | luí | 12880 | 54 | 238.5185185 | 0 |
| extremis | extremi | 2124 | 9 | 236 | 0 |
| holmes | holme | 7073 | 30 | 235.7666667 | 0 |
| escalopines | escalopín | 244 | 11 | 22.1818182 | 1 |
| fotomecánicas | fotomecánico | 24 | 16 | 1.5 | 1 |
| claroscuristas | claroscurista | 11 | 13 | 0.8461538 | 1 |
| antieconómicas | antieconómico | 68 | 161 | 0.4223602 | 1 |
| linfocíticas | linfocítico | 8 | 20 | 0.4 | 1 |
| fototérmicos | fototérmico | 9 | 24 | 0.375 | 1 |
| autoproclamaciones | autoproclamación | 7 | 69 | 0.1014493 | 1 |
| esquizofrénicas | esquizofrénico | 92 | 1067 | 0.0862231 | 1 |
| jurisprudencias | jurisprudencia | 62 | 18092 | 0.0034269 | 1 |
| moderaciones | moderación | 36 | 10547 | 0.0034133 | 1 |
| nazismos | nazismo | 11 | 4523 | 0.002432 | 1 |
| comos | como | 651 | 10574252 | 0.0000616 | 0 |
| madrids | madrid | 17 | 1239084 | 0.0000137 | 0 |
| relacionares | relacionar | 5 | 425566 | 0.0000117 | 0 |
| ... | ... | ... | ... | ... | |

Cuadro 4: Fragmento del listado de parejas singular-plural. El valor de la columna *a(p,s)* define si el algoritmo aceptará o no la pareja propuesta

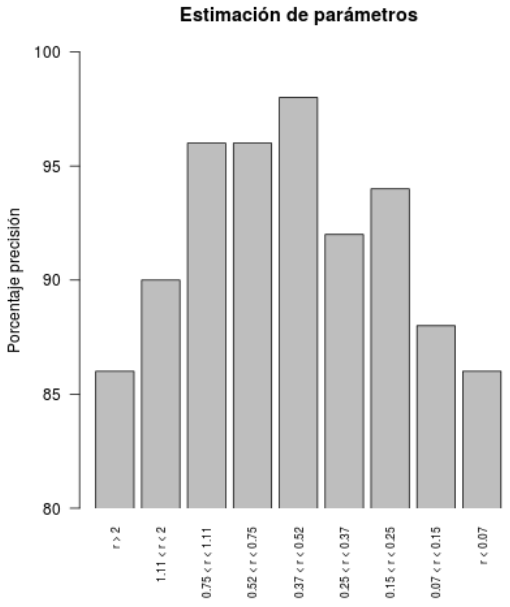


Figura 1: Diagrama de barras con valores orientativos para estimar los parámetros $x$ y $z$

## 3.4 Análisis de un corpus-objetivo

Ya con los recursos generados en la etapa anterior (tabla de parejas singular-plural + reglas de pluralización), pasamos ahora a describir cómo es la metodología de análisis de un corpus especializado en particular, el que denominamos corpus-objetivo.

### *3.4.1 Etiquetado y posprocesamiento del corpus-objetivo*

Tal como hemos descrito en las secciones anteriores, proponemos una secuencia de pasos en los que, en primer lugar, se somete el corpus-objetivo a un etiquetado con la herramienta habitual. A continuación, se reenvía el resultado de esta operación al script que implementa nuestro algoritmo de corrección del plural.

Por cada forma del corpus, este sistema puede detectar casos en los que la palabra termina en *-s*. Llamaremos $W_p$ a una palabra que cumpla con esta condición, como por ejemplo *anticoagulantes*. $W_p$ será sometida a una función binaria $f(W_p)$ (3) que determinará si $W_p$ tiene o no lematización.

$$f(W_p) = \begin{cases} 1 & W_p \rightarrow W_s \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Para cada $W_p$ existen solo dos resultados posibles: la lematización correspondiente (1) o su rechazo como forma plural (0). El valor 1 se devuelve en el caso de cumplirse $W_p \rightarrow W_s$, es decir, que para la supuesta forma plural $W_p$ se ha encontrado una forma singular satisfactoria $W_s$ (por ejemplo, *anticoagulante*). Por el contrario, si $f(W_p) = 0$, se considerará un caso de no plural y no presentarán flexión de número (sería el caso, por ejemplo, de una forma como *apoptosis*).

A continuación se explica en detalle este proceso de decisión.

### *3.4.2 Primeras operaciones de descarte*

Antes de someter a una determinada forma a la batería de análisis para encontrar su forma singular, aplicamos primero una serie de filtros que reducen considerablemente la tasa de error de entrada del algoritmo. Estos filtros se aplican primero por eficiencia computacional.

En primer lugar, aplicamos una expresión regular para detectar y descartar formas verbales. En castellano encontramos terminaciones en *-s* en formas verbales de distintos tiempos y modos, como en segunda persona del singular (*vienes, cantes, salgas*, etc.) y en primera persona del plural (*venimos, cantamos, salgamos*, etc.) así como en los pronombres enclíticos (*cantarles, llamábales, arreglándoselas*, etc.).

Optamos por detectar estas terminaciones por medio de expresiones regulares. Si de esta forma se consigue una coincidencia, entonces el script cambia la etiqueta NC (o ADJ) del etiquetador reemplazándola por la categoría verbo. En esta ocasión no hemos resuelto recuperar el infinitivo del verbo en cuestión, ya que consideramos que esa es otra tarea que dejamos para un trabajo futuro.

El siguiente filtro es el de los elementos que típicamente no son plurales en castellano, y que aparecen en el caso de los corpus especializados con mayor o menor frecuencia dependiendo del dominio:

```
(us|[oipeas][lrxtfsd]is|[íó][dtf][aei][dnl]esus|[oipeas]
[lrxtfsd]is|[íó][dtf][aei][dnl]es)$
```

Al igual que en el caso de los verbos, las unidades que coincidan con este modelo morfológico serán clasificadas como elementos no plurales, y se asignará como lema la misma forma encontrada.

Un tercer filtro es el de las formas en inglés. Ocurre con frecuencia, particularmente cuando se trabaja con corpus especializados, que fragmentos del corpus están escritos en otra lengua, usualmente en inglés, debido a los resúmenes, las palabras clave, las citas y los títulos bibliográficos que aparecen en los artículos especializados. Idealmente, deberíamos implementar un detector de fragmentos en inglés, pero en lugar de eso nos pareció más sencillo aplicar una nueva regla de filtrado para detectar la morfología característica del plural en inglés, y que m

```
(ys[ei]s|nces|tr?ics|ions|[nl]ess|ties|tions|en[td]s|[ae]
cts|sters|oids|ishes|ous|ers|[csr]ies|ants|[aoe]ss)$
```

Si una unidad léxica terminada en *-s* $W_p$ coincide con estas terminaciones se le asigna la etiqueta "Palabra en inglés", y queda así también fuera del análisis. Si, en cambio, $W_p$ supera estos filtros, se registra su frecuencia en el corpus-objetivo y es sometida al resto del proceso.

### *3.4.3 Consulta a la tabla de parejas singular-plural y procedimientos auxiliares*

La tabla de parejas plural-singular, cuyo fragmento se mostró en el Cuadro 4, es leída y cargada en memoria como *hash table* al principio del proceso y ofrece dos informaciones: por un lado, la forma singular del elemento en cuestión y, por el otro, la frecuencia de aparición de cada elemento en el corpus de referencia.

Definimos en (4) una función $m(W_f)$ exclusivamente para los casos en los que encontramos que la forma singular termina con la letra *a*, lo que podría ser indicativo de que se trata de una forma en femenino tal como sucede normalmente en el caso de los adjetivos. La manera de determinar esto es comprobar si existe en el corpus de referencia la forma masculina correspondiente.

$$m(W_f) = \begin{cases} 1 & \text{máx}\,(c(W_o), c(W_s)) > u \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

La función $m(W_f)$ toma en consideración entonces la frecuencia de aparición en el corpus de referencia de las formas masculina y femenina del término en cuestión. Así, si $W_f$ es una palabra terminada en *a*, definimos un candidato a forma masculina que puede ser con o sin *o* ($W_o$ y $W_s$, respectivamente) en función de la frecuencia de aparición en el corpus de referencia ($c(W_o)$ y $c(W_s)$). Para ello se exploran dos posibilidades:

1. Reemplazar la última letra *a* por *o*.
2. Eliminar directamente la letra *a*.

Si alguna de las dos formas resultantes tiene una frecuencia igual o superior al umbral de frecuencia $u$ ya definido anteriormente, entonces $m(W_f)$ decide "masculinizar" la forma terminada en *-a*. Si ambos cumplen esta condición, se elige el más frecuente. En cambio, si $m(W_f) = 0$, se asume que no se trata de un adjetivo y por tanto debe lematizarse con la forma terminada en *-a*. Naturalmente, este será el caso de sustantivos femeninos, como *rosa*, que debe ser lematizado de esa manera, pero también casos como el de *pediatra*, que termina en *-a* pero no es un sustantivo femenino y su lema también permanece inalterado.

### *3.4.4 Estrategias de repliegue*

A pesar del gran tamaño del corpus de referencia utilizado, es normal observar que el corpus-objetivo contiene formas léxicas que no aparecen en el de referencia. La primera estrategia de repliegue en este caso es reintentar el proceso, pero esta vez reemplazando el corpus de referencia por el mismo corpus-objetivo. Particularmente, esto permite resolver casos de terminología propia del dominio.

En los casos en que una determinada palabra no se ha observado en ninguno de los dos corpus y no ha podido ser clasificada con ninguna de las estrategias anteriores, aplicamos entonces la segunda estrategia de repliegue, que consiste en el análisis de la posible prefijación de la palabra para los casos de un elemento léxico que sí es conocido, pero que ha sufrido un proceso de derivación mediante la adición de prefijos. Aplicamos en estos casos una regla de determinación de prefijos, para lo cual utilizamos la siguiente lista de prefijos en castellano (Real Academia Española, 2006):

```
^(ad|ana|anti|auto|cata|co|cuasi|de|di|em|en
|entre|ex|extra|hetero|hiper|hipo|in|infra
|inter|macro|meta|micro|mono|neuro|para|per
|poli|pos|post|pre|pro|p?seudo|psico|radio
|re|sin|son|sub|super|tele|tran?s|ultra)
```

No es una lista exhaustiva, pero reúne los prefijos de alta productividad en dominios de especialidad. Si con la ayuda de esta lista podemos separar la palabra en dos partes y descubrir la unidad léxica conocida, sometemos esa unidad al mismo proceso descrito en las secciones anteriores, con la única diferencia de que al lema resultante volvemos a añadir el prefijo encontrado. Así, por ejemplo, encontramos que la forma *antidopaminérgicos* no aparece con la frecuencia mínima en los corpus, pero luego de la eliminación del prefijo *anti-* se descubre una forma que sí es conocida (*dopaminérgicos*) y que sí admite un singular con nuestra metodología (*dopaminérgico*). De esta forma, se restituye el prefijo elidido anteriormente y se recupera de esta forma el singular *antidopaminérgico*.

Si esta segunda estrategia de repliegue también falla, es decir, si no estamos ante un caso de derivación por prefijos, entonces esta circunstancia nos obliga a dar cuenta de estos elementos e intentar también ofrecer un lema aunque se trate de una entidad puramente teórica. Cabe esperar que entren en esta categoría los errores de tipeo encontrados en el corpus. No hemos pretendido dar solución aquí el problema del error de tipeo porque consideramos que es una investigación diferente (*cf.* comentarios al respecto en la Sección 5, cuando mencionamos posibilidades de trabajo futuro). En lugar de esto, optamos por proporcionar un lema teórico debido que entre los errores de tipeo cabe esperar también la aparición de unidades terminológicas genuinas que aun no han sido registradas (neologismos).

Claramente se trata de una estrategia arriesgada en este último caso por lo que, cuando esto ocurre, añadimos a la lematización la etiqueta UNKNOWN para que el usuario sepa que el sistema aquí está "inventando" un lema cuya existencia no ha sido documentada. Esto se consigue aplicando la misma batería de reglas descrita en la Sección 3.2, pero ahora ya sin el apoyo empírico que fundamenta la decisión de clasificarlo como un lema genuino.

## 4 Resultados

El primer paso para la evaluación de los resultados fue constituir un corpus especializado para ser utilizado como corpus-objetivo. Con la autorización de la *Revista Chilena de Neuropsiquiatría*, conformamos un corpus-objetivo a partir una muestra de artículos aleatoria (800.000 palabras) desde su sitio web[2]. Transformamos el material a texto plano y, por medio de expresiones regulares, eliminamos resúmenes y palabras clave en inglés, y así quedó así conformado el corpus-objetivo[3]. Para nuestros experimentos, etiquetamos este corpus con TreeTagger, que arrojó un tokenizado de 805.624 líneas, y luego con UDPipe, que verticalizó el corpus en 797.812 tokens. Las diferencias entre ambos resultados se deben a las distintas estrategias de tokenización de cada programa. En particular, difieren en la forma de encapsular, por ejemplo, locuciones o marcadores discursivos como una sola unidad léxica (*en general, por ejemplo, sin embargo*, etc.).

Utilizamos estos dos etiquetadores por las razones expuestas en la Sección 2.2: principalmente, lo extendido que se encuentra su uso en la actualidad y el hecho de que representen estrategias de lematización distintas (probabilístico vs. neuronal). Sin embargo, como ya advertimos, no es el propósito de esta investigación hacer una evaluación exhaustiva de todos los etiquetadores actualmente existentes, ya que el hecho de que un etiquetador tenga mejor desempeño que otro no tiene relación directa con el método que propone-

[2] La revista se puede consultar en línea y los artículos son *open access*: http://www.sonepsyn.cl

[3] Las pruebas de desarrollo del algoritmo se hicieron en un proyecto de extracción de terminología en un corpus de biología en castellano. Las dificultades allí encontradas con el plural motivaron esta investigación.

mos. Dicho de forma más general, dado un corpus etiquetado con una calidad $X$, nuestro método producirá una cantidad $Y$ de reducción del error ($1 - X$ en un intervalo comprendido entre 0 y 1).

El procedimiento para evaluación consistió entonces en aplicar nuestro script al resultado obtenido con cada etiquetador. Cada vez que nuestro script encuentra una forma terminada en *-s* se considera que esto es indicio de que podría tratarse de una forma plural. A partir de aquí, los siguientes resultados son posibles:

1. El algoritmo decide que la forma corresponde efectivamente a un plural. En este caso se propone el lema correspondiente en singular.
2. Decide que la forma no es un plural, por lo tanto deja como lema la misma forma.
3. Decide que la palabra está en inglés, por tanto es un elemento no analizable (el lema queda igual a la forma y se identifica con la etiqueta "Palabra en inglés").
4. Decide que la palabra es una forma verbal (no corresponde por tanto aplicar las reglas para la detección de plural y se identifica con la etiqueta "verbo").

Después de la aplicación de los etiquetadores y de nuestro script encontramos, en el caso de TreeTagger, un total de 9.763 formas distintas (types) terminadas en *-s*. En más de la mitad de estos casos (5.081) nuestro script modificó el lema originalmente asignado por el etiquetador. En el caso de UDPipe, en tanto, encontramos 9.576 types, de los cuales 2.364 tuvieron una modificación del lema por parte de nuestro script. Este primer dato ya arroja una estimación de la diferencia en el desempeño de los etiquetadores, ya que en el segundo caso fueron requeridas menos modificaciones. Es necesario tener en cuenta, sin embargo, que la discrepancia no significa necesariamente un error en la lematización. Es posible asignar lemas distintos y que ambos sean aceptables. El caso más frecuente sería el de los participios, que pueden lematizarse con un infinitivo (en la lectura de forma verbal) o bien como un adjetivo (dejando el participio en la forma masculino singular).

Para llevar a cabo la evaluación, tomamos muestras aleatorias de 600 casos de cada uno de los dos resultados, únicamente cuando el lema propuesto por nuestro script es distinto al lema producido por los etiquetadores, ya que suponemos que si el lema es el mismo entonces el resultado debe ser correcto. Hicimos un muestreo dividido por categorías: 300 casos de formas reconocidas como plural (que es el fenómeno que más nos interesa estudiar en esta investigación); 100 casos de formas no plurales; 100 casos de formas clasificadas por nuestro script como palabras en inglés, y finalmente 100 casos de formas verbales.

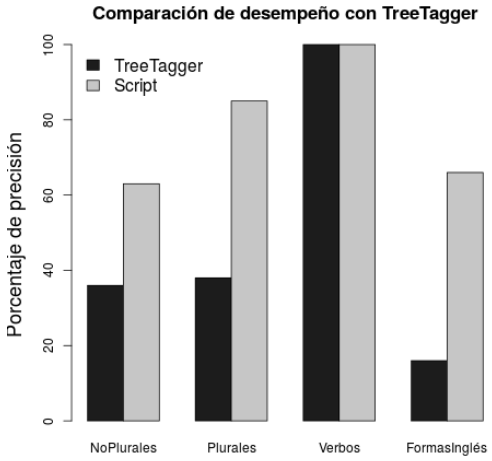


Figura 2: Diagrama de barras con la comparación del desempeño entre TreeTagger y el script

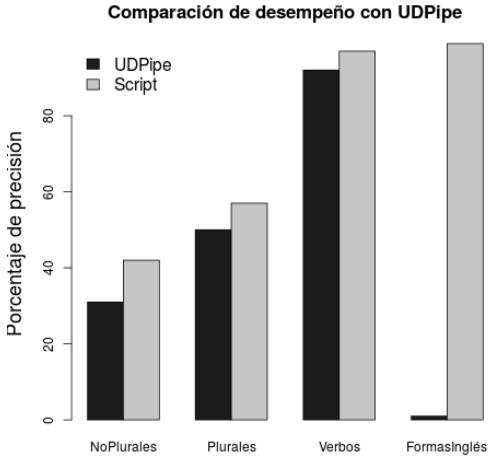


Figura 3: Diagrama de barras con la comparación del desempeño entre UDPipe y el script

Las figuras 2 y 3 muestran diagramas de barras que comparan el desempeño de nuestro prototipo con el de TreeTagger y UDPipe, respectivamente. Es necesario recalcar aquí que los porcentajes de precisión corresponden a las muestras de palabras terminadas en *-s* en las que existe una discrepancia entre el lema asignado por el etiquetador y el asignado por nuestro script. En ambas figuras se puede observar que cuando hay disparidad en la lematización, en general nuestro script tiende a ser más preciso, tendencia que se mantiene en las cuatro categorías de error analizadas.

En el caso de las formas que fueron clasificadas como plurales y fueron lematizadas, encontramos que, en la comparación con TreeTagger, un 85 % de las veces en que hubo disparidad con este etiquetador, el lema propuesto por el script fue correcto, frente a un 38 % de este etiquetador (los porcentajes no suman 100 porque, como ya se indicó, hay casos en que lematizaciones distintas son aceptables). Esta diferencia se reduce considerablemente en el caso de UDPipe. En parte, y como ya vimos, el desacuerdo con nuestro script es mucho menor. Pero cuando existe desacuerdo, el desempeño de nuestro script es mejor, aunque el margen es más reducido: 57 % de nuestro script frente a 50 % de UDPipe.

La mayoría de los errores cometidos por el script en esta categoría corresponden a palabras en inglés, con un 35 % (*caregivers, remarks, substances*, etc.). A veces los lemas que propone el script son igualmente correctos aunque estén en inglés, pero esto es irrelevante porque es un resultado casual, no parte del diseño, y debemos considerarlos por tanto como errores, ya que el sistema debería haberlos clasificado como formas en inglés. El porcentaje restante de errores lo conforman apellidos con el 11 % (como *Hodges, Hopkins, Duprés*, etc.), siglas y abreviaturas con el 5 % (*irss, mgrs, tlrs*, etc.), entre otros fenómenos. En cuanto a los resultados de la detección de no-plurales, el porcentaje de precisión en los casos en los que hay discrepancia con el etiquetador, la precisión fue de 63 % contra 36 % en el caso de TreeTagger, y 42 % frente a 31 % en el caso de UDPipe.

A modo de ilustración, el Cuadro 5 ofrece un fragmento de los resultados en la categoría "plurales". Allí, el único error registrado es de tipeo *cuaiïdades* [*sic*], que recibió el "lema" *cuaiïdad*, y que tenemos que marcar como error del procedimiento, ya que es consecuencia de apostar por una forma cuya existencia no está documentada en corpus. Se trata de un problema que podría tener solución con un algoritmo de corrección ortográfica, pero esto es, como ya hemos señalado, un problema distinto. En ninguno de los casos etiquetados como elementos desconocidos se encontraron errores de otro tipo. El Cuadro 6, por su parte, muestra resultados aleatorios en la categoría "NoPlurales". Allí, la columna "Error" muestra los errores que marcamos durante la revisión manual de los resultados. Tal como se puede apreciar, los errores cometidos son principalmente formas en inglés o nombres propios.

Los datos también permiten hacer observaciones sobre la comparación del desempeño entre los dos etiquetadores. En general se observa que UD-

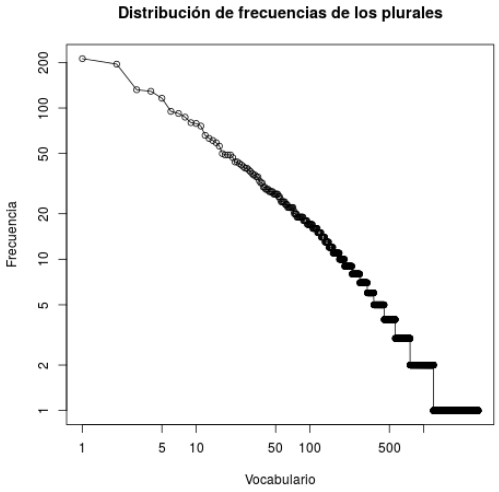


Figura 4: Distribución de frecuencias de los plurales encontrados (escala logarítmica en ambos ejes)

Pipe tiene mejor desempeño porque el porcentaje de mejora o de corrección de error que introduce nuestro script es menor. Sin embargo, hay ciertas categorías en las que tiende a cometer más errores que TreeTagger, que no intenta lematizar formas que no conoce (asigna el lema "unknown"). UDPipe sí lo hace y comete más errores, lo que se ve en particular en el caso de las formas en inglés, en donde el desempeño de UDPipe se deteriora notablemente.

Finalmente, la Figura 4 muestra la distribución de frecuencias de los elementos corregidos en el caso de TreeTagger. Como cabía esperar, la figura muestra que se trata de una distribución típicamente zipfeana.

## 5 Conclusiones

En esta propuesta metodológica hemos presentado una formalización de las reglas para la pluralización de sustantivos y adjetivos en el español y un algoritmo que ofrece una solución de alta efectividad para el problema de las UND. Consideramos que el sistema será de utilidad para el trabajo de los terminólogos en corpus especializado en castellano.

En este trabajo hemos optado por acotar el problema al caso del plural debido a su alta incidencia y su efecto perjudicial en tareas más avanzadas como la extracción de terminología o la revisión ortográfica en procesadores de texto. En ese sentido, nuestra investigación es de interés desde un punto de vista eminentemente práctico. Sin embargo, más allá de este aspecto práctico, hay que destacar la simplicidad del algoritmo,

| Forma | Lema | Desconocido | Error |
|---|---|---|---|
| ... | ... | | |
| hipersensibles | [hipersensible] | * | |
| escretoras | [escretor] | * | |
| anorectics | [PALABRA EN INGLÉS] | | |
| fuentes | [fuente] | * | |
| cuaiidades | [cuaiidad] | * | ! |
| perdonamos | [FORMA VERBAL] | | |
| anatómicos | [anatómico] | | |
| inexplicadas | [inexplicado] | | |
| intangibles | [intangible] | * | |
| contraindicadas | [contraindicado] | | |
| teriovenosas | [teriovenoso] | * | |
| existenciarios | [existenciario] | | |
| fetales | [fetal] | | |
| ... | ... | | |

Cuadro 5: Fragmento de salida del algoritmo para la categoría "plurales", con indicación de elementos desconocidos y eliminación de formas verbales y palabras en inglés

| Forma | Error |
|---|---|
| ... | ... |
| sarcoidosis | |
| angus | |
| dermis | |
| epistaxis | |
| linfocitosis | |
| crisis | |
| periartritis | |
| rawlins | ! |
| polyhidramnios | ! |
| neurogénesis | |
| meningitis | |
| losephs | ! |
| enuresis | |
| alcalosis | |
| ... | ... |

Cuadro 6: Fragmento de salida del script para la categoría de "NoPlurales"

sobre todo en comparación con la complejidad de los modelos basados en redes neuronales. Esto tiene importancia práctica, la que se traduce en facilidad de uso y rapidez de ejecución, pero creemos que también tiene un atractivo metodológico y conceptual: el seguir el principio de parsimonia (*keep it small and simple*). Se requieren pocos conocimientos de programación para adaptar esta implementación a otra lengua, suponiendo que es posible sistematizar también reglas de formación de plural.

El presente trabajo sugiere, además, varias líneas de trabajo futuro, puesto que representa una invitación a enfrentar con un razonamiento similar otros tipos de error cometidos por los etiquetadores. Por ejemplo, una línea de investigación que consideramos complementaria y de gran interés sería la determinación del género de los sustantivos, que en este trabajo apenas hemos gestionado. Otra posibilidad en la misma línea sería corregir las etiquetas gramaticales que asignan los etiquetadores, tema que tampoco dejamos resuelto en este artículo, aunque ya damos algunas orientaciones sobre cómo podría hacerse al buscar las desinencias verbales, por ejemplo. Otro caso de esto último sería comprobar si una etiqueta de NC asignada por el etiquetador no debería ser en realidad ADJ cuando se observa que la palabra en cuestión tiene flexión de género además de la de número. Esto, naturalmente, sin que se observe la presencia de determinantes en la posición inmediatamente anterior a la palabra analizada con una frecuencia significativa en el corpus de referencia, lo que acusaría su uso como sustantivo.

En este artículo hemos mencionado también, pero no resuelto, el tema de la corrección ortográfica, ya que los errores de tipeo o de ortografía son frecuentes incluso en las publicaciones especializadas. Pensamos que se podría explorar, por ejemplo, una medida de similitud ortográfica entre palabras para descubrir lo que el autor quiso decir (ej. *cualidades* por *cuaiidades*). Sin embargo, en un caso así, una medida de similitud ortográfica no sería suficiente, ya que se debería complementar con una medida de similitud distribucional para controlar que no se ofrezcan palabras ortográficamente similares pero semánticamente distintas.

Finalmente, consideramos también la siguiente línea de trabajo futuro, inspirada en parte en el artículo de Brill (1992): investigar hasta qué punto se puede utilizar el etiquetado que ya trae el corpus de referencia para crear el modelo de la pluralización a partir de las etiquetas dejadas por este. El corpus etiquetado funcionaría así como un material de entrenamiento para un algoritmo de clasificación. De esta manera, es posible aprender a reconocer la morfología del plural (o la del género, etc.) a través de las palabras que sí son reconocidas por el etiquetador, para extrapolar esa morfología aprendida hacia las palabras que no son reconocidas. Esta línea de investigación es interesante sobre todo por las posibilidades de generalización hacia otras lenguas.

## Agradecimientos

Este trabajo ha sido posible gracias a la financiación del Proyecto Fondecyt Regular 1191481: "Inducción automática de taxonomías de marcadores discursivos a partir de corpus multilingües", dirigido por el primer autor (Fondo Nacional de Desarrollo Científico y Tecnológico, Gobierno de Chile). Queremos agradecer también a los revisores Gerardo Sierra y Marcos García, ya que con sus comentarios mejoraron sustancialmente el artículo.

## Referencias

Adams, Greg, Beth Millar, Eric Neufeld & Tim Philip. 1994. Ending-based strategies for part-of-speech tagging. En *Uncertainty in artificial Intelligence*, 1–7. Elsevier. 10.1016/B978-1-55860-332-5.50005-5.

Agerri, Rodrigo, Josu Bermudez & German Rigau. 2014. IXA pipeline: Efficient and ready to use multilingual NLP tools. En *Language Resources and Evaluation Conference (LREC)*, 3823–3828.

Alcina, Juan & José Manuel Blecua. 1975. *Gramática española*, vol. 1991. Barcelona: Ariel.

Alemany, José. 1920. *Tratado de la formación de palabras en la lengua castellana*. Madrid: Librería general de Victoriano Suárez.

Allauzen, Alexandre & Hélene Bonneau-Maynard. 2008. Training and evaluation of POS taggers on the French MULTITAG corpus. En *Language Resources and Evaluation Conference (LREC)*, s/p.

Ambadiang, Théophile. 1999. La flexión nominal: género y número. En *Gramática descriptiva de la lengua española*, 4843–4914. Espasa Calpe.

Asociación de Academias de la Lengua Española. 2005. *Diccionario panhispánico de dudas*. Real Academia Española.

Bazzi, Issam. 2002. *Modelling out-of-vocabulary words for robust speech recognition*: Massachusetts Institute of Technology. Tesis Doctoral.

Bazzi, Issam & James Glass. 2002. A multiclass approach for modelling out-of-vocabulary words. En *7th International Conference on Spoken Language Processing*, 1613–1616.

Bello, Andrés. 1847. *Gramática de la lengua castellana destinada al uso de los americanos*. Chile: Imprenta del Progreso.

Biemann, Chris. 2006. Unsupervised part-of-speech tagging employing efficient graph clustering. En *21st Conference on Computational Linguistics and 44th Meeting of the Association for Computational Linguistics: student research workshop*, 7–12.

Bohnet, Bernd & Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. En *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1455–1465.

Brill, Eric. 1992. A simple rule-based part of speech tagger. En *3rd Conference on Applied Natural Language Processing*, 152–155. 10.3115/974499.974526.

Brill, Eric. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics* 21(4). 543–565.

Carreras, Xavier, Isaac Chao, Lluis Padró & Muntsa Padró. 2004. FreeLing: An open-source suite of language analyzers. En *Language Resources and Evaluation Conference*, 239–242.

Cedeño, Rafael A. Núñez, Sonia Colina & Travis G. Bradley. 2014. *Fonología generativa contemporánea de la lengua española*. Washington, DC: Georgetown University Press.

Charniak, Eugene, Curtis Hendrickson, Neil Jacobson & Mike Perkowitz. 1993. Equations for part-of-speech tagging. En *11th Conference on Artificial Intelligence (AIII)*, vol. 93, 784–789.

Church, Kenneth Ward. 1989. A stochastic parts program and noun phrase parser for unrestricted text. En *International Conference on Acoustics, Speech, and Signal Processing*, 695–698. 10.3115/974235.974260.

Creutz, Mathias, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar & Andreas Stolcke. 2007. Analysis of morph-based speech recognition and the modeling of out-of-vocabulary words across languages. En *Proceedings of Human Language Technologies Conference*, 380–387.

De Nebrija, Antonio. 1492. *Gramática castellana*. Salamanca.

Dixon, Robert. 2009. *Basic linguistic theory volume 1: Methodology*. Oxford: Oxford University Press.

Foley, James. 1967. Spanish plural formation. *Language* 43(2). 486–493. 10.2307/411548.

Greene, Barbara B. & Gerald M. Rubin. 1971. *Automatic grammatical tagging of English*. Providence, Rhode Island: Department of Linguistics, Brown University.

Hernández Alonso, César. 1984. *Gramática funcional del español*. Madrid: Gredos.

Honnibal, Mathew. 2016. Spacy. https://spacy.io/. Accessed: 2018-10-30.

Honnibal, Matthew & Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. En *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1373–1378. 10.18653/v1/D15-1162.

Kilgarriff, Adam & Irene Renau. 2013. esTenTen, a vast web corpus of Peninsular and American Spanish. *Procedia - Social and Behavioral Sciences* 95. 12 – 19. 10.1016/j.sbspro.2013.10.617.

de León, Ramón Zacarías Ponce. 2015. Flexión de número en la composición nominal del español: estructura morfológica y rutinización. *Anuario de Letras. Lingüística y Filología* 2(2). 101–131.

Ling, Wang, Chris Dyer, Alan W Black, Isabel Trancoso, Ramón Fermandez, Silvio Amir, Luís Marujo & Tiago Luís. 2015. Finding function in form: Compositional character models for open vocabulary word representation. En *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1520–1530. 10.18653/v1/D15-1176.

Manning, Christopher D. 2011. Part-of-speech tagging from 97 % to 100 %: is it time for some linguistics? En *International Conference on Intelligent Text Processing and Computational Linguistics*, 171–189. 10.1007/978-3-642-19400-9_14.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA, USA: MIT Press.

Miłkowski, Marcin. 2010. Developing an open-source, rule-based proofreading tool. *Software: Practice and Experience* 40(7). 543–566.

Moreno, Antonio & José M Goni. 1995. GRAMPAL: a morphological processor for spanish implemented in prolog. *arXiv preprint cmp-lg/9507004* .

Moyna, María Irene. 2011. *Compound words in spanish: theory and history*, vol. 316. John Benjamins Publishing.

Müller, Thomas, Ryan Cotterell, Alexander Fraser & Hinrich Schütze. 2015. Joint lemmatization and morphological tagging with lemming. En *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2268–2274. 10.18653/v1/D15-1272.

Parada, Carolina, Mark Dredze & Frederick Jelinek. 2011. OOV sensitive named-entity recognition in speech. En *12th Conference of the International Speech Communication Association (INTERSPEECH)*, s/p.

Parra Escartín, Carla & Héctor Martínez Alonso. 2015. Choosing a spanish part-of-speech tagger for a lexically sensitive task. *Procesamiento del Lenguaje Natural* 54. 29–36.

Plank, Barbara, Anders Søgaard & Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. En *54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 412–418. 10.18653/v1/P16-2067.

Qi, Peng, Timothy Dozat, Yuhao Zhang & Christopher D. Manning. 2018. Universal dependency parsing from scratch. En *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 160–170.

Qin, Long. 2013. *Learning out-of-vocabulary words in automatic speech recognition*: School of Computer Science, Carnegie Mellon University. Tesis Doctoral.

Real Academia Española. 1741. *Orthographia española, compuesta, y ordenada por la real academia española*. Madrid: Imprenta de la Real Academia Española.

Real Academia Española. 1920. *Gramática de la lengua castellana*. Madrid: Perlado Páez y compañía, Impresores y Libreros de la Real Academia Española.

Real Academia Española. 2006. *Diccionario esencial de la lengua española*. Espasa Calpe.

Real Academia Española. 2009. *Nueva gramática de la lengua española*. Espasa Libros.

Sánchez Corrales, Víctor. 1994. La categoría morfosintáctica número en el sustantivo español. *Revista de filología y lingüística de la Universidad de Costa Rica* 20(1). 155–168.

Santos, Cicero D & Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. En *31st International Conference on Machine Learning (ICML)*, 1818–1826.

Saporta, Sol. 1965. Ordered rules, dialect differences, and historical processes. *Language* 41(2). 218–224. 10.2307/411875.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. En *International Conference on New Methods in LanguageProcessing*, 25–36.

Stockwell, Robert P, J Donald Bowen & John W Martin. 1965. *The grammatical structures of english and spanish*. University of Chicago Press.

Straka, Milan. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. En *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 197–207. 10.18653/v1/K18-2020.

Straka, Milan, Jan Hajič & Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-u files performing tokenization, morphological analysis, POS tagging and parsing. En *10th Language Resources and Evaluation Conference (LREC)*, 4290–4297.

Straka, Milan & Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. En *CoNLL Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99. 10.18653/v1/K17-3009.