Revista Española de Lingüística

EVALUACIÓN MORFOLÓGICA DE LOS VOCABULARIOS DE SUBPALABRAS UTILIZADOS POR LOS GRANDES MODELOS DE LENGUAJE

Óscar García-Sierra[1]
Universidad Complutense de Madrid; dezzai

Ana Fernández-Pampillón Cesteros[2]
Universidad Complutense de Madrid

Miguel Ortega-Martín[3]
Universidad Complutense de Madrid; dezzai

Resumen

Con el auge de los grandes modelos del lenguaje neuronales, especialmente aquellos basados en *Transformers*, la tradicional segmentación en palabras y morfemas que empleaba reglas lingüísticas ha sido reemplazada por algoritmos de segmentación estadísticos. Estos algoritmos son mucho más eficientes y, sin necesidad de intervención humana, son capaces de, a partir de corpus de millones de palabras, construir el vocabulario de palabras y subpalabras que necesitan los grandes modelos del lenguaje monolingües o multilingües. Ocurre, sin embargo, que estas subpalabras no se corresponden siempre con morfemas y esto repercute negativamente en el funcionamiento de los modelos del lenguaje que utilizan estos segmentadores. Cuánto se alejan los vocabularios estadísticos de un vocabulario real de palabras y morfemas de una lengua –lo que denominamos calidad morfológica del vocabulario–, y cuánto repercute esta falta de calidad en la eficacia de los grandes modelos del lenguaje son cuestiones todavía sin resolver. Este artículo aborda la primera cuestión, la calidad morfológica de los vocabularios, aportando un método de evaluación basado en tres medidas de calidad –relevancia, coherencia y corrección morfológica–, y un procedimiento para evaluarlas. El método se aplica para medir la calidad de los vocabularios generados por tres algoritmos de segmentación en subpalabras, *BPE*, *WordPiece* y *Unigram*, utilizados mayoritariamente para la construcción de los grandes modelos del lenguaje. Los resultados

1. oscargo2@ucm.es; https://orcid.org/0000-0002-8828-7338
2. apampi@filol.ucm.es; https://orcid.org/0000-0002-6606-0159
3. m.ortega@ucm.es; https://orcid.org/0000-0002-1880-5048

RSEL 54/1 · 2024 · pp. 103-130 · DOI: https://doi.org/10.31810/rsel.54.1.4
 · ISSN 0210-1874 · e-ISSN 2254-8769

que hemos obtenido indican que la calidad morfológica de los mismos es muy baja, por lo que merece la pena buscar nuevas soluciones para mejorar la calidad de los vocabularios de los grandes modelos del lenguaje.

Palabras clave: segmentación; morfemas; subpalabras; grandes modelos del lenguaje; lengua española.

MORPHOLOGICAL EVALUATION OF SUBWORD VOCABULARIES USED BY LARGE LANGUAGE MODELS

Abstract

Traditional tokenization methods using linguistic rules have been replaced by statistical segmentation algorithms. Although these algorithms show a higher efficiency and are capable of building subword vocabularies from large corpora without human supervision, these subwords do not consistently correspond to morphemes. This paper addresses this issue by proposing an evaluation methodology and applying it to the morphological quality of Spanish vocabularies produced by three prominent subword tokenization algorithms –*BPE*, *WordPiece*, and *Unigram*– commonly used in Large Language Models (LLMs). Three gold standards were created to measure relevance, coherence, and morphological accuracy of vocabularies of six tokenizers trained on Spanish corpus, exploring different vocabulary sizes. Evaluation results indicate that none of the three algorithms is suitable for accurately representing Spanish morphology.

Keywords: tokenizing; morphemes; subwords; large language models; Spanish language.

Recibido: 08/03/2024 Aprobado: 30/06/2024

1. INTRODUCCIÓN

La segmentación (*tokenization* en inglés) es una de las fases más relevantes en el procesamiento del lenguaje natural. Implica la división del texto en unidades más pequeñas llamadas *tokens* (Friedman, 2023). Estos *tokens* pueden ser morfemas, palabras, sintagmas, frases o incluso caracteres, dependiendo de la aplicación específica. La segmentación es también un componente clave en la preparación de los datos de entrada para entrenar las redes neuronales artificiales en la medida en que representan las unidades discretas de información a partir de las cuales se construyen los modelos del lenguaje (Friedman, 2023). Por ejemplo, el modelo *GPT-2* depende de un vocabulario de cincuenta mil *tokens,* aproximadamente.

Así, en el entrenamiento de las actuales redes neuronales de tipo *Transformer* (Vaswani *et al.*, 2017) para generar (también se utiliza el término *aprender*), los grandes modelos del lenguaje (en adelante, utilizaremos *modelos del lenguaje*) se utilizan vocabularios en los que los *tokens* son palabras y, mayoritariamente, subpalabras. Una subpalabra es una parte de la palabra que tiene una frecuencia relevante de aparición en el texto segmentado y que no se corresponde necesariamente con un morfema. También, para poder utilizarse (como generador de texto, clasificador, traductor o sistema de diálogo, entre otros), los modelos del lenguaje necesitan utilizar el vocabulario de palabras y subpalabras con el que han sido creados.

Los segmentadores que se emplean actualmente para generar los vocabularios de los modelos del lenguaje utilizan estrategias estadísticas que aplican a enormes corpus de texto, en vez del conocimiento lingüístico formalizado en las gramáticas tradicionales basadas en reglas. Principalmente, se fundamentan en la frecuencia con la que ciertas cadenas de caracteres (palabras o subpalabras) aparecen en el corpus de entrenamiento, lo que produce como resultado que las palabras con una frecuencia relevante en el corpus se segmenten como un solo *token*, y, sin embargo, las palabras menos frecuentes se dividan en varios *tokens* de tipo subpalabra con una mayor frecuencia de aparición. Por ejemplo, el segmentador *WordPiece* del modelo de lenguaje *BETO* trata la palabra *población* como un solo *token*, pero descompone *desprestigiar* en varios *tokens*: *despre*, *##st*, *##igi*, *##ar*.

Entre los segmentadores utilizados para generar los vocabularios de los modelos del lenguaje destacan tres por ser los que se utilizan mayoritariamente: *Byte-Pair Encoding (BPE)* (Sennrich, 2015), *WordPiece* (Schuster, 2012; Wu, 2016) y *Unigram* (Kudo, 2018). Así, por ejemplo, el modelo *BERT* (Devlin *et al.*, 2019) utiliza *WordPiece*, *GPT* (Radford *et al.*, 2018) o *RoBERTa* (Liu *et al.*, 2019) utilizan *BPE*, y *Albert* (Lan *et al.*, 2019) emplea *Unigram*. Son estos tres segmentadores los que hemos seleccionado para llevar a cabo la evaluación empírica de calidad que presentamos en este artículo.

La ventaja indudable de los actuales segmentadores estadísticos frente a los simbólicos basados en reglas lingüísticas es su eficacia: son capaces de construir vocabularios de decenas de miles de *tokens* a partir de corpus de texto de gran tamaño sin intervención humana alguna. Además, la segmentación estadística permite obtener vocabularios independientes de las lenguas, lo que los hace especialmente útiles cuando se trabaja con modelos multilingües.

El inconveniente es, como demuestran diversos estudios, que la segmentación en subpalabras no se corresponde, en un alto porcentaje, con morfemas, perdiéndose el conocimiento lingüístico básico sobre el que se construyen los modelos del lenguaje y, en consecuencia, perdiéndose la eficacia y fiabilidad de dichos modelos (Chuch, 2020; Bostrom y Durret, 2020; Hoffman *et al.*, 2021; Park, 2020). En este sentido, Church (2020) analiza cualitativamente algunas palabras complejas en inglés y llega a la conclusión de que estas se descomponen en demasiadas subpalabras que no se corresponden ni con palabras ni con morfemas reales, a pesar de lo que señalan algunos autores (Song *et al.*, 2020). Por su parte, Bostrom y Durret (2020) comparan los segmentadores *BPE* y *Unigram* en inglés y japonés, y concluyen que el segundo segmentador genera mejores vocabularios de morfemas en ambos idiomas que el primero. Hoffman *et al.* (2021) analizan el tratamiento que hace *WordPiece* de las palabras complejas en inglés, concluyen también que generan subpalabras que no se corresponden con morfemas reales, y comprueban que, en ese mismo idioma, un modelo con un vocabulario de morfemas mejora los resultados respecto de un modelo que utiliza subpalabras puramente estadísticas. Por último, Park (2020) realiza un estudio similar al de Hoffman *et al.* (2021), pero para el coreano.

Denominamos *calidad morfológica del vocabulario de un modelo del lenguaje en una lengua* a su grado de semejanza respecto de un vocabulario real de morfemas y palabras de esa misma lengua. El problema que surge, antes de abordar la mejora de la calidad morfológica de los vocabularios de los modelos del lenguaje, es que –hasta donde conocen los autores de este trabajo– no existe un método para evaluar y contrastar de forma objetiva la calidad lingüística de estos vocabularios.

En este artículo proponemos una solución a la cuestión de cómo evaluar la calidad morfológica, basándonos en los resultados de los trabajos previos, y la aplicamos a la evaluación de los tres segmentadores (*BPE*, *WordPiece* y *Unigram*) trabajando con textos del español. El artículo se ha organizado de la forma siguiente: en la sección segunda presentamos el método de evaluación de la calidad morfológica; en la sección tercera se describe su aplicación a los tres segmentadores; en las secciones cuarta y quinta se discuten los resultados, primero los valores cuantitativos obtenidos para cada criterio de evaluación y, después, cualitativamente mediante un análisis de errores. En la sexta y última sección se presentan las conclusiones y el trabajo futuro.

2. PROPUESTA DE MÉTODO DE EVALUACIÓN DE LA CALIDAD MORFOLÓGICA DE LOS VOCABULARIOS DE LOS MODELOS DEL LENGUAJE

El método de evaluación de la calidad morfológica se fundamenta en dos elementos: i) tres criterios de calidad: la relevancia, coherencia y corrección morfológicas; y, ii), tres conjuntos datos de validación para medir el grado de cumplimiento de cada uno de los tres criterios: primero, una lista de pares palabra-segmentación que sea una muestra equilibrada de palabras de las diferentes categorías morfosintácticas y que contenga todos los morfemas de la lengua junto con sus segmentaciones morfológicas para evaluar la corrección morfológica; segundo, una lista de los morfemas de la lengua, para evaluar la coherencia morfológica; y, finalmente, a partir de las listas anteriores, una lista de pares palabra-morfema que servirá para evaluar la relevancia morfológica. La definición de cada uno de los criterios de evaluación se describe en la subsección 2.1, mientras que el procedimiento para medir su cumplimiento se describe en la subsección 2.2.

2.1. *Los criterios de evaluación*

Recogiendo y ampliando los trabajos anteriormente mencionados de Bostrom y Durret (2020), Church (2020), Hoffman *et al.* (2021) y Park (2020) proponemos utilizar tres criterios o parámetros para la evaluación de la calidad de los vocabularios de los modelos del lenguaje. Estos criterios son: (1) la relevancia morfológica, (2) la coherencia morfológica y, (3) la corrección morfológica.

(1) *La relevancia morfológica*

La relevancia morfológica mide cuánto de morfológico es el vocabulario. Se calcula como la intersección entre el vocabulario que genera el segmentador –y que posteriormente utilizará el modelo de lenguaje– y los morfemas reales de la lengua (Figura 1). Así, por ejemplo, en el vocabulario del segmentador del modelo del lenguaje *BETO* está incluida la subpalabra *códi* que no es un morfema real del español, no está incluido el sufijo del español *-érrimo* y sí está incluido el morfema del español *-idad*. Este último morfema sería el que contaría para calcular la relevancia morfológica.

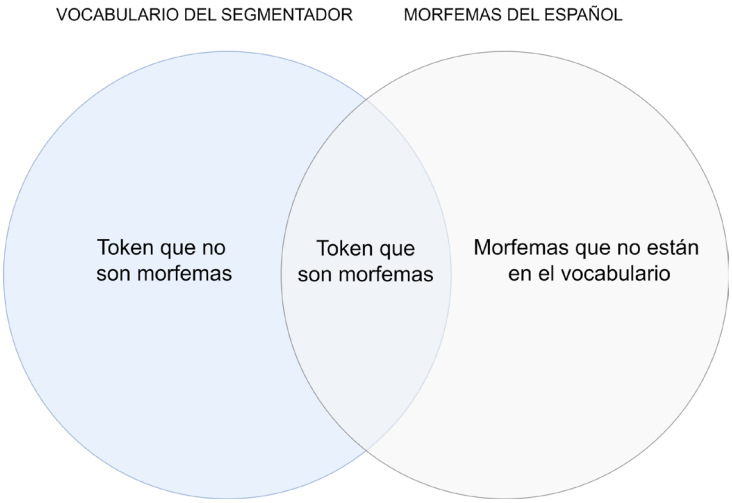

Figura 1. Esquema de concepto de relevancia morfológica

Para medir la relevancia proponemos el uso de las métricas tradicionales de precisión, cobertura y valor F1[4]. Los valores se encuentran en un rango entre 0 (cuando no hay intersección entre vocabulario del modelo y conjunto de morfemas) y 1 (cuando la intersección es total). También pueden utilizarse valores porcentuales.

Las métricas de precisión, cobertura o valor F1 han sido ya utilizados para analizar la calidad de los vocabularios en otras lenguas. Así, para la lengua coreana, Park *et al.* (2020) utilizan la precisión para medir qué porcentaje de los *tokens* del vocabulario del segmentador son morfemas correctos del coreano. Bostrom y Durret (2020) utilizan las tres métricas, precisión, cobertura y F1, para medir la intersección entre los *tokens* generados por un segmentador *BPE* y una porción de morfemas reales del inglés. Por último, Hoffman *et al.*, (2022) emplean un corpus de segmentación morfológica que incluye la derivación y composición, llamado *CELEX* (Van de Wouden, 1990) para extraer una lista de morfemas del inglés, y posteriormente emplean la medida de cobertura para comprobar qué porcentaje de esos elementos están en el vocabulario de su segmentador.

4. La precisión, en este caso, es el número de *tokens* que son morfemas del vocabulario del segmentador (la intersección en la Figura 1) dividido entre el total de *tokens* del vocabulario del segmentador. La cobertura (*recall* en inglés) es el número de *tokens* que son morfemas del vocabulario del segmentador (la intersección) dividido por el total de morfemas de la lengua. El valor F1 es la media armónica de la precisión y la cobertura.

En nuestra propuesta, a diferencia de los trabajos previos, la relevancia morfológica se medirá con las tres métricas para poder comparar de forma más objetiva este criterio en diferentes vocabularios.

(2) *La coherencia morfológica*

El segundo criterio de evaluación que proponemos es la coherencia morfológica, que mide con qué frecuencia las palabras con una misma estructura morfológica (palabras que comparten un determinado morfema) se segmentan, de acuerdo con alguna de las opciones siguientes (Figura 2):

- como un solo *token* (no hay segmentación)
- como varios *tokens* de modo que el morfema correcto se corresponde con uno de los *tokens* obtenidos en la segmentación
- como varios *tokens* de modo que el morfema correcto no se corresponde con ningún *token* de la segmentación.

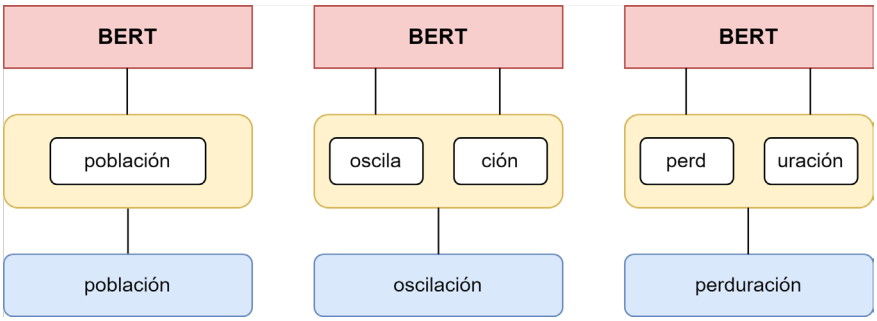

Figura 2. Falta de coherencia morfológica en vocabulario del segmentador utilizado por el modelo del lenguaje *BETO*

La Figura 2 muestra un ejemplo real de falta de coherencia morfológica de los *tokens* obtenidos por el segmentador *WordPiece* que utiliza el modelo del lenguaje *BETO* (Cañete *et al.*, 2023). En concreto, la figura muestra cómo tres palabras que comparten el sufijo *-ción* son segmentadas incoherentemente de tres formas diferentes:

- aparece sin segmentar en la palabra población, debido a que es frecuente en el corpus de entrenamiento;
- aparece correctamente segmentado como dos subpalabras en el caso de oscilación (oscila y *-ción*); y,

– en la palabra perduración, la segmentación es errónea en perd y uración, con lo que el sufijo *-ción* aparece perdido en la subpalabra uración.

En este sentido, el modelo de lenguaje *BETO* aprende que esas tres palabras no comparen el morfema, por lo que se pierde la posibilidad de modelar la similitud morfológica que realmente existe entre las tres palabras *población*, *oscilación* y *perduración*.

En los trabajos previos de Church (2020) y Hoffman *et al.* (2021) se apunta la existencia del problema de falta de coherencia morfológica al segmentar textos en inglés, pero ninguno de los trabajos lo evalúa cuantitativamente. Hoffman *et al.* (2021) analizan cualitativamente el sufijo *-ize*, y comprueban que muchas palabras que lo contienen no tienen en la práctica ningún token en común, puesto que se dan los tres casos ya explicados para el sufijo en español. Church (2020) analiza los casos en los que al añadir un prefijo a una palabra se pierde la coherencia entre las segmentaciones de la forma base y la forma derivada.

Nosotros proponemos medir, para cada tipo de morfema, qué porcentaje del total de palabras representan cada uno de los siguientes casos:

– palabras que contienen el morfema y que se segmentan como un solo token.
– palabras que contienen el morfema y se segmentan como varios tokens, pero el morfema no se corresponde con un token.
– palabras que contienen el morfema y se segmentan como varios tokens en las que morfema se corresponde con un token.

(3) *Corrección morfológica*

El criterio de corrección en la segmentación morfológica tiene como objetivo evaluar cómo funcionan los segmentadores a la hora de segmentar completamente una palabra en morfemas. A diferencia del criterio anterior, aquí se mide si el segmentador divide *correctamente* los morfemas de la palabra en cuestión. Así, puede ocurrir que un segmentador separe de forma totalmente coherente una palabra, pero los *tokens* que obtiene son erróneos porque no se corresponden con las palabras y morfemas correctos de la lengua. Por ejemplo, el segmentador del modelo del lenguaje *BETO* segmenta bien el prefijo de la palabra *contrapuntos*, pero no el resto de la palabra: [*contra*, *pu*, *n*, *tos*].

Para medir la corrección morfológica proponemos utilizar la métrica de corrección. El resultado de una segmentación es correcto si todos los *tokens* segmentados coinciden con los morfemas etiquetados. El rango de la corrección varía entre 0 y

100%, indicando con 0 que ninguna palabra se ha segmentado correctamente y con valor 100 que todas las palabras han sido segmentadas correctamente.

En los trabajos previos de Bostrom y Durret (2020) y Hoffman *et al.* (2021) se evalúa cualitativamente la corrección de los segmentadores de subpalabras a la hora de segmentar morfológicamente palabras en inglés, pero ninguno lo hace cuantitativamente.

2.2. *Procedimiento de evaluación*

Proponemos que, para la evaluación de la calidad de los segmentadores y sus vocabularios, los criterios anteriormente definidos de relevancia morfológica de los vocabularios, coherencia y corrección morfológica se calculen utilizando el procedimiento siguiente en cinco pasos:

Paso 1. Creación de tres conjuntos de datos de evaluación, uno para cada criterio:

i) Para el criterio de relevancia morfológica de los vocabularios se debe crear un conjunto de datos que consiste en la lista de morfemas de la lengua, a ser posible divididos por su tipología: prefijo, sufijo y raíz.
ii) Para el criterio de la coherencia morfológica se debe construir una lista de pares palabra y morfema. Esta lista se crea a partir del conjunto de datos anterior de morfemas y debe contener todos los morfemas de la lengua.
iii) Para evaluar la corrección morfológica se crea una lista de palabras aleatoria etiquetadas con su(s) correspondiente(s) segmentación(es) morfológica(s). Para que sea una muestra representativa, esta lista debe contener todos los morfemas de la lengua al menos una vez y mantener la proporción de categorías gramaticales del *Diccionario de la lengua española* (en adelante, *DLE*; RAE-ASALE, en línea).

En la sección siguiente, se detallará un posible método de construcción, así como el contenido de los tres conjuntos de datos.

Paso 2. Selección de los algoritmos de segmentación de tokens en subpalabras que se van a evaluar.

Paso 3. Selección del corpus de entrenamiento de los segmentadores que se van a evaluar (seleccionados en el paso 2). Básicamente, se debe tener en cuenta que el corpus sea suficientemente representativo de la lengua y que se ajuste a los

recursos informáticos disponibles. Este corpus solo se utilizará para entrenar los segmentadores y generar los vocabularios.

Paso 4. Generación de los vocabularios con cada segmentador a evaluar mediante el entrenamiento de dichos segmentadores (seleccionados en el paso 2) con el corpus de entrenamiento creado en el paso 3.

Paso 5. Cálculo del nivel de cumplimiento de cada criterio utilizando los conjuntos de datos de evaluación creados en el paso 1. Además, en el caso del conjunto de datos de para evaluar la corrección morfológica, también es interesante almacenar el número de tokens utilizados por cada segmentador a la hora de segmentar, para posteriormente calcular la media de tokens utilizados por cada segmentador.

3. APLICACIÓN A LA EVALUACIÓN DE LOS VOCABULARIOS *BPE*, *WORDPIECE* Y *UNIGRAM* EN ESPAÑOL

Con el fin de probar la viabilidad del método de evaluación de la calidad morfológica de los segmentadores hemos llevado a cabo una prueba de concepto utilizando los tres segmentadores, *BEP*, *WordPiece* y *Unigram*, y generando vocabularios de dos tamaños diferentes por cada segmentador. En esta sección detallamos el proceso de aplicación del método y en la sección siguiente se muestran los resultados y su discusión. Así, el proceso de evaluación se llevó a cabo en los cinco pasos propuestos en la sección 2 de la forma siguiente:

Paso 1. En este paso se construyen los tres conjuntos de datos que servirán para realizar la evaluación: conjunto de datos para la relevancia morfológica, para la coherencia morfológica y para la corrección morfológica. Se comienza construyendo el conjunto de datos para evaluar la corrección.

En primer lugar, se ha construido manualmente el conjunto de datos para evaluar la corrección morfológica, puesto que este conjunto de datos se ha usado también para evaluar un segmentador propio basado en reglas morfológicas del español, que ha permitido construir automáticamente los conjuntos de datos de evaluación restantes con la corrección suficiente.

Para ello, se han anotado manualmente 1.231 palabras del español con todas sus segmentaciones morfológicas posibles, siguiendo las reglas de la *Nueva gramática de lengua española* (en adelante, *NGLE*). La selección de palabras se ha realizado aleatoriamente teniendo en cuenta que todos los morfemas del español estén representados al menos una vez y se ha respetado la proporción de categorías

gramaticales del *DLE*. La Tabla 1 contiene los totales y un ejemplo de cada categoría gramatical.

Categoría gramatical	Total	Ejemplos
Nombres	481	['camiones': {'NOUN': ['camion', 'es']}]
Verbos	445	['dejan': {'VERB': ['dej', 'an']}]
Adjetivos	299	['adecuadas': {'ADJ': ['adecu', 'ad', 'a', 's']}]
Pronombres	61	['estos': {'PRON': ['est', 'o', 's']}]
Adverbios	59	['aquí': {'ADV': ['aquí']}]
Determinantes	30	['mis': {'DET': ['mi', 's']}]
Preposiciones	22	['desde': {'ADP': ['desde']}]
Conjunciones	14	['pero': {'CONJ': ['pero']}]

Tabla 1. Conjunto de datos para evaluar la corrección morfológica

En segundo lugar, se ha construido el conjunto de datos para medir la relevancia morfológica de los vocabularios. Los morfemas han sido extraídos automáticamente de la *NGLE* y del *Diccionario de uso del español* (Moliner, 1967/2009; en adelante, *DUE*), mientras que las raíces han sido generadas automáticamente por un segmentador de autoría propia y que ha sido evaluado en el conjunto de datos de 1.231 palabras con una corrección superior al 99%. En concreto se han seleccionado las 5.000 raíces más frecuentes del *DLE*. La Tabla 2 contiene los morfemas totales y un ejemplo de cada tipo de morfema.

Tipo de morfema	Total	Ejemplos
Prefijos	61	['des', 're', ...]
Sufijos	175	['mos', 'ción', ...]
Raíces	5,000	['dej', 'alt', ...]

Tabla 2. Conjunto de datos para evaluar la relevancia morfológica

Por último, para evaluar la coherencia morfológica se ha construido un tercer conjunto de datos de evaluación. Partiendo del conjunto de prefijos, raíces y sufijos empleados para evaluar la relevancia morfológica (Tabla 2) y las palabras del *DLE* se ha creado la lista de pares (palabra, morfema). La Tabla 3 contiene los totales de pares (palabra, morfema) para cada tipo de morfema, así como una muestra de ejemplos. Hay que señalar que una misma palabra puede encontrarse varias veces en este conjunto de datos, puesto que aparecerá tantas veces como morfemas tenga. Así,

por ejemplo, en el caso de la palabra *remodelación* el conjunto de datos contendrá el par (*remodelación*, *re-*), (*remodelación*, *modela*) y (*remodelación*, *-ción*). Además, el conjunto de datos contiene 3.920 verbos con pronombres enclíticos que se han descargado del corpus *Spanish Web 2018* mediante la herramienta de análisis textual *Sketch Engine*. Se ha elegido este recurso por su total accesibilidad y por ser suficientemente representativo del vocabulario actual del español.

Tipo de morfema	Total de palabras	Ejemplos
Prefijos	205.792	(*remodelación*, *re-*)
Raíces	234.564	(*ubicásemos*, *ubic*)
Sufijos	1.575.207	(*fijación*, *-ción*)
Clíticos	3.920	(*dilo*, *lo*)

Tabla 3. Conjunto de datos de evaluación de la coherencia morfológica

Como resultado de este paso se dispone de los tres conjuntos de evaluación para evaluar en el paso 5.

Paso 2. Se seleccionan los algoritmos que se van a evaluar que, como hemos indicado anteriormente, son *WordPiece*, *BPE* y *Unigram* por ser los principales algoritmos de segmentación en subpalabras utilizados actualmente por los grandes modelos del lenguaje (Church, 2020) y por estar disponibles en la plataforma *Hugging Face*.

Paso 3. Para la selección del corpus de entrenamiento hemos revisado la sección de conjuntos de datos disponible en la plataforma *Hugging Face* que es la principal plataforma de modelos y conjuntos de datos de la inteligencia artificial, donde existen más de 400 corpus para modelar lenguaje en español. El corpus, en nuestro caso, debe ajustarse a las limitaciones de almacenamiento de *Google Collab*, la plataforma que tenemos disponible para el entrenamiento de los segmentadores (12gb de RAM y 100gb de disco duro).

Se ha elegido el corpus *Oscar* (Ortiz Suárez *et al.*, 2019; Ortiz Suárez *et al.*, 2020), que es el segundo corpus más utilizado para modelar lenguaje en español de esta plataforma solo ligeramente por detrás del de *Wikipedia*[5]. Es un corpus con el contenido de diferentes webs preparado para utilizarse (limpio y postprocesado). Se ha preferido *Oscar* frente a *Wikipedia* porque este último contiene

5. https://huggingface.co/datasets?task_ids=task_ids:language-modeling&language=language:es&sort=trending

presumiblemente más nombres propios que no interesan desde un punto de vista morfológico. En concreto, se ha utilizado la versión pequeña de *Oscar*[6], que consta de unas 600.000 frases, tiene un tamaño de 10Gb y se adapta mejor a nuestras limitaciones computacionales.

El resultado de este paso son 600.000 frases en español, limpias y disponibles para entrenar con ellas los tres segmentadores. Las frases se agrupan en lotes de 1.000 frases para facilitar y optimizar su posterior procesamiento durante el entrenamiento, siguiendo el proceso propuesto por *Hugging Face*[7].

Paso 4. Creación de los vocabularios. Los vocabularios se generan mediante el entrenamiento de los segmentadores conforme la guía propuesta en *Hugging Face*. Los únicos hiperparámetros ajustables son el tamaño de los lotes (que, como ya se ha mencionado se ha fijado en 1.000 frases), y el de los tamaños de los vocabularios (que, como también se ha explicado, se han fijado en 31.000 y 52.000 *tokens*).

Para cada segmentador se han creado, por lo tanto, dos versiones de vocabulario, versión 31 y versión 52, correspondientes a los tamaños 31.000 y 52.000 *tokens*. De esta forma, se puede estudiar si el tamaño del vocabulario mejora o empeora la calidad morfológica del mismo. Los tamaños se han elegido porque se corresponden con tamaños de los modelos del lenguaje reales. Por ejemplo, *BETO* tiene 31.000 *tokens* y *RoBERTa* cuenta con 50.000. Además, el vocabulario de todos los morfemas del español tiene un tamaño aproximado de entre 38.000 y 52.000 *tokens*, según los cálculos que hemos realizado al construir los conjuntos de datos en el paso 2. La Tabla 4 muestra los segmentadores entrenados.

Vocabulario	Corpus	Tamaño corpus (frases)	Tamaño vocabulario (tokens)
wordpiece_oscar_31	oscar small	600.000	31.000
wordpiece_oscar_52	oscar small	600.000	52.000
BPE_oscar_31	oscar small	600.000	31.000
BPE_oscar_52	oscar small	600.000	52.000
unigram_oscar_31	oscar small	600.000	31.000
unigram_oscar_52	oscar small	600.000	52.000

Tabla 4. Listado de los vocabularios evaluados

6. https://huggingface.co/datasets/nthngdy/oscar-small
7. https://colab.research.google.com/github/huggingface/notebooks/blob/master/examples/tokenizer_training.ipynb

Paso 5. En este paso se han evaluado los seis segmentadores con los tres conjuntos de datos de evaluación creados en el paso 2 y se ha procedido a analizar los resultados. Esta tarea se ha centrado en la comparación entre los tres segmentadores y en los diferentes tamaños de vocabulario. Los resultados se muestran y discuten en la sección 4.

Como se ha indicado en la sección anterior, para medir la relevancia morfológica se han utilizado las métricas de precisión, cobertura y valor F1 basados en la intersección entre el conjunto de datos de evaluación de la relevancia morfológica (Tabla 2) y los seis vocabularios generados (Tabla 4).

Respecto a la coherencia morfológica, además de calcular la coherencia morfológica, para estudiar los errores, para cada tipo de morfema, prefijos, sufijos, raíces y clíticos, se han creado tres listas llamadas *un token*, *varios - token correcto* y *varios - token incorrecto* (Tabla 5). Las listas se han creado de la forma siguiente:

i) se procesan los pares de palabra, morfema, segmentando la palabra;
ii) si la palabra se segmenta como un *token*, se añade la palabra a la lista *un token*; si se segmenta en varios y el morfema es uno de ellos, se añade a *varios - token correcto*: y, si se segmenta en varios y el morfema no está entre ellos, se añade a la tercera lista de *varios-token incorrecto*;
iii) una vez procesadas todas las palabras, se calcula el tamaño de cada una de las tres listas y se normaliza en forma de porcentaje respecto al total de *tokens*.

La coherencia morfológica se corresponde con la lista *varios - token_correcto*.

	Un *token*	Varios-*token* correcto	Varios-*token* incorrecto
Prefijos	*retirar: [retirar]*	*circunvalar: [circun, vala, r]*	*desprogramar: [despro, grama, r]*
Raíces	*generar: [generar]*	*cebar: [ceb, ar]*	*suturar: [su, tura, r]*
Sufijos	*calmar: [calmar]*	*destapar: [des, tap, ar]*	*aburrir: [abur, rir]*
Clíticos	*ayudarte: [ayudarte]*	*llamarme: [llamar, me]*	*sáquelo: [s, áquel, o]*

Tabla 5. Ejemplos de las listas generadas para el análisis de errores del segmentador *Wordpiece_31*. En negrita el morfema evaluado para cada palabra

Por último, para evaluar la corrección morfológica se utiliza, como se ha indicado anteriormente, la métrica de corrección morfológica. Para calcular

esta medida se utiliza el conjunto de datos de corrección morfológica (Tabla 1). Se compara cada par (palabra-segmentación) con el conjunto de datos y se cuenta una segmentación como correcta cuando se comprueba que todos los morfemas de la palabra se han segmentado correctamente.

Además, en este paso se calcula la media de *tokens* utilizados para segmentar las palabras de este conjunto de datos, con el fin de averiguar qué segmentadores utilizan un mayor número de *tokens* por palabra.

En la siguiente sección se muestran y discuten los resultados de este proceso de evaluación respecto a cada criterio.

4. RESULTADOS Y DISCUSIÓN

4.1. *Relevancia morfológica*

La Tabla 6 muestra los resultados para precisión, cobertura y valor F1 de relevancia morfológica de los seis vocabularios según los prefijos, sufijos y raíces contenidos en los vocabularios, y según el total de morfemas para poder estudiar, si existen, las diferencias respecto de cada tipo de morfema.

Como se observa, los valores de precisión son, en todos los casos, muy bajos. Estos valores bajos confirman que, efectivamente, el funcionamiento estadístico de los segmentadores no es capaz de extraer las unidades lingüísticas básicas del lenguaje natural, en este caso del español. La cobertura, sin embargo, presenta mejores valores porque mide qué proporción de los morfemas reales están en los vocabularios, aun cuando presenta una gran cantidad de subpalabras que no son morfemas. El valor F1 es muy bajo porque es la media armónica de la precisión y cobertura y la precisión es muy baja. Básicamente, lo que ocurre es que estos tres segmentadores mejoran la cobertura mediante la inclusión de más subpalabras frecuentes lo que aumenta la probabilidad de contener los morfemas de la lengua, pero a costa de disminuir su precisión.

Tipo	Totales	Vocabulario	Precisión (%)	Cobertura (%)	F1 (%)
Prefijos	61	wordpiece_31	0.17	88.52	0.35
		wordpiece_52	0.11	96.72	0.23
		BPE_31	0.18	90.16	0.37
		BPE_52	0.11	96.72	0.23
		unigram_31	0.15	77.05	0.30
		unigram_52	0.10	83.61	0.20
Sufijos	175	wordpiece_31	0.41	73.56	0.82
		wordpiece_52	0.26	77.01	0.51
		BPE_31	0.41	72.41	0.81
		BPE_52	0.25	73.56	0.49
		unigram_31	0.33	59.20	0.66
		unigram_52	0.21	63.22	0.42
Raíces	5.000	wordpiece_31	8.59	9.97	9.23
		wordpiece_52	7.00	16.02	9.74
		BPE_31	7.75	8.99	8.32
		BPE_52	5.57	12.74	7.75
		unigram_31	5.75	6.68	6.18
		unigram_52	4.15	9.50	5.78
Totales	5.236	wordpiece_31	4,15	24,55	7,09
		wordpiece_52	3,13	31,08	5.69
		BPE_31	3,25	19,24	5.56
		BPE_52	2.36	23,40	4,28
		unigram_31	2,56	15,15	4,38
		unigram_52	1,99	19,79	3,62

Tabla 6. Resultados de la evaluación de la relevancia morfológica de los vocabularios. Se han marcado en gris los valores más altos

De forma más pormenorizada respecto al tipo de morfema se comprueba lo siguiente:

– En el caso de los prefijos, los vocabularios del segmentador *BPE* son los que presentan mejor relevancia morfológica con un valor F1 de 0.37 y hasta un 90% de cobertura para los prefijos del español. Los vocabularios de *WordPiece* ocupan el segundo lugar con un valor F1 de 0.35% y los de *Unigram* resultan ser los de peor rendimiento.

– En el caso de los sufijos, los vocabularios de *WordPiece* superan ligeramente a los de *BPE* y, de nuevo, claramente a los de *Unigram*.
– En el caso de las raíces, los vocabularios de *WordPiece* siguen siendo los mejores, de tal manera que cuando se utiliza un tamaño de vocabulario de 52.000 *tokens* se almacena hasta un 16% de las raíces evaluadas.

Respecto al tamaño del vocabulario, observamos que, como era de esperar, al aumentar el tamaño aumenta también el porcentaje de morfemas reales presentes en ellos y, por tanto, la cobertura. Sin embargo, como recoge la Tabla 6, los valores de F1 empeoran para todos los casos al ampliar los tamaños, puesto que la precisión no crece en la misma proporción que la cobertura. Así, al aumentar el tamaño del vocabulario, se observa que el ritmo de aumento del número de *tokens* que no son morfemas reales del español es superior al aumento de morfemas reales de la lengua.

A pesar de ello, cabe destacar que, al aumentar el tamaño, los vocabularios de *WordPiece* logran los mejores resultados, igualando a los de *BPE* en los prefijos (0.23% de F1) y superándolos en los sufijos y en las raíces. Los vocabularios de *Unigram* son los peores cuando tienen 52.000 *tokens*.

En resumen, de los resultados de la evaluación de la relevancia morfológica se puede concluir que ninguno de los vocabularios destaca significativamente respecto a este criterio de relevancia morfológica, aunque parece claro que los vocabularios de *Unigram* son los que ofrecen peores resultados independientemente del tamaño de vocabulario utilizado. El aumento de tamaño parece favorecer a los vocabularios generados por *WordPiece*, mientras que tamaños más reducidos benefician a los vocabularios de *BPE*. En ninguno de los casos el aumentar el tamaño del vocabulario hace que mejore el valor F1, lo que implica que al aumentar el tamaño del vocabulario los *tokens* que no son morfemas crecen más que los que sí lo son. Los vocabularios generados por los tres segmentadores tienen una cobertura aceptable, siendo los mejores los generados por *WordPiece* y *BPE* con un vocabulario de 52.000 *tokens* en la inclusión de los prefijos (incluyen más del 96%), pero su cobertura disminuye en la inclusión de los sufijos (73-77%) y empeora en la inclusión de las raíces. Los valores de precisión son extremadamente bajos porque, como hemos indicado, al ser una construcción puramente estadística, los morfemas reales representan una parte pequeña del total de los *tokens*.

4.2. *Coherencia morfológica*

La coherencia morfológica se refiere a la cantidad de veces que las palabras con una estructura morfológica similar (es decir, las que comparten un morfema) tienen también un *token* en común, siendo el *token* dicho morfema. Para analizar los resultados de la evaluación de la coherencia morfológica se han recogido en la Tabla 7 los resultados por tipo de morfema y se han añadido los clíticos. Como se observa, respecto a los tipos de morfema, todos los segmentadores tienden a usar varios *tokens* para segmentar las palabras del corpus de evaluación, lo que ofrece valores de coherencia realmente bajos, que llegan, en el mejor de los casos, a casi un 66% de coherencia en *BPE_31*. Valores semejantes de coherencia morfológica se obtienen en el caso de los clíticos, siendo *Unigram_31* el que mejores resultados presenta (casi un 67%). Es también relevante observar que el número de palabras únicas que incluyen estos vocabularios es muy bajo.

Categoría	Vocabulario	Total palabras	La palabra es un solo *token* (%)	Varios *tokens* (subpalabras)	
				El morfema es un *token* (%)	El morfema no es un *token* (%)
Prefijos	wordpiece_31	205.792	0,71	13,45	85,84
	wordpiece_52		1,41	9,03	89,56
	BPE_31		0,04	65,86	34,10
	BPE_52		0,10	60,45	39,45
	unigram_31		1,11	42,73	56,16
	unigram_52		1,78	38,61	59,62
Raíces	wordpiece_31	234.584	1,64	16,03	82,33
	wordpiece_52		2,69	18,75	78,56
	BPE_31		0,40	4,66	94,94
	BPE_52		0,64	5,50	93,86
	unigram_31		1,79	9,08	89,13
	unigram_52		2,62	10,93	86,45

Sufijos	wordpiece_31	1.575.207	0,74	15,20	84,05
	wordpiece_52		1,34	14,07	84,59
	BPE_31		0,11	10,10	89,79
	BPE_52		0,20	7,80	92,00
	unigram_31		0,92	20,74	78,34
	unigram_52		1,43	18,49	80,08
Clíticos	wordpiece_31	3.920	5,61	61,40	32,99
	wordpiece_52		16,61	51,66	31,73
	BPE_31		0,69	48,57	50,74
	BPE_52		1,12	44,21	54,67
	unigram_31		4,64	66,87	28,49
	unigram_52		11,94	63,44	24,62
Totales	wordpiece_31	2.019.504	0,86	15,21	83,93
	wordpiece_52		1,54	14,17	84,29
	BPE_31		0,14	15,23	84,64
	BPE_52		0,24	12,97	86,79
	unigram_31		1,05	21,71	77,24
	unigram_52		1,62	19,75	78,63

Tabla 7. Resultados de la evaluación de la coherencia morfológica. Se indican en gris los valores más altos

Respecto al tamaño del vocabulario, se observa que al ampliar los tamaños de 31.000 a 52.000 *tokens* empeora la coherencia morfológica (es decir, para todos los segmentadores, disminuye el porcentaje de casos en los que el morfema evaluado se corresponden con un *token* correcto). Cabe recordar que, al ampliar los tamaños de los vocabularios, los valores de F1 de la tarea de relevancia morfológica también decrecían en todos los casos debido a la disminución de la precisión.

También es relevante señalar que, aunque en todos los segmentadores el porcentaje de palabras segmentadas como un solo *token* es muy bajo, en el caso del segmentador *BPE* estos valores son especialmente bajos.

Finalmente, parece que podría existir cierta correlación entre los valores de relevancia y coherencia morfológica de los vocabularios para prefijos y raíces. Así, *BPE* presenta los mejores resultados de relevancia morfológica en los prefijos, y eso se traduce en mejores resultados de coherencia morfológica respecto a los prefijos. En las raíces ocurre lo mismo con *WordPiece*. En los sufijos, en cambio,

se rompe esa tendencia, ya que *WordPiece* es el que mayor relevancia morfológica presenta y, sin embargo, no es el que tiene mayor coherencia morfológica (son los segmentadores de *Unigram*).

En resumen, los tres segmentadores segmentan de forma poco coherente –el valor máximo de coherencia es de 66% de BPE_31– y los vocabularios más grandes tienen peor coherencia morfológica. En los prefijos y las raíces parece que existe una correlación entre la relevancia morfológica respecto a prefijos y raíces de los vocabularios y la coherencia morfológica. En los prefijos *BPE* es el que mejor relevancia morfológica del vocabulario presenta, y también es el que mejores resultados de coherencia presenta para este tipo de morfema. En las raíces, *WordPiece* es el mejor en relevancia morfológica y también lo es en coherencia. Sin embargo, en los sufijos, *Wordpiece* es el mejor en relevancia del vocabulario, mientras que *Unigram* lo supera en coherencia. Esto se debe a que, como se verá en el análisis de errores, aunque un morfema forme parte del vocabulario de un segmentador, no significa necesariamente que el segmentador vaya a usarlo para segmentar palabras que lo contienen.

4.3. *Corrección morfológica*

En el caso de la corrección morfológica, los resultados de todos los segmentadores son, de nuevo, realmente bajos. La Tabla 8 muestra los resultados. Como se observa, ninguno de los vocabularios supera el 15% de corrección.

Vocabulario	Totales	Corrección (%)	*Tokens* utilizados de media por palabra
wordpiece_31	1231	14,54	1,75
wordpiece_52		14,94	1,54
BPE_31		8,69	2,39
BPE_52		10,23	2,19
unigram_31		14,70	1,83
unigram_52		15,10	1,59

Tabla 8. Resultados de corrección morfológica y media de *tokens* utilizados por palabra

De forma más detallada, se observa que el segmentador *Unigram* obtiene los mejores resultados y que *BPE* es el peor de los tres. Respecto al tamaño de vocabulario, aumentarlo hace que mejoren ligeramente los resultados.

El análisis de errores nos ha permitido encontrar una explicación a estos valores tan bajos. Se ha comprobado que todos los aciertos de segmentación se corresponden con palabras segmentadas como un solo *token*. Es decir, con las palabras que no requieren ser segmentadas porque son muy frecuentes en el corpus. En cambio, ninguno de los seis segmentadores segmenta correctamente en morfemas la mayor parte de las palabras que requieren ser segmentadas. Por esta razón, el segmentador *Unigram,* que tiene tendencia a no dividir las palabras, obtiene una mejor corrección. Esto se aprecia también en la Tabla 8, en la columna de *tokens* usados de media en cada segmentación. Los segmentadores *Unigram*, que son los que obtienen mejores resultados, utilizan pocos *tokens* de media para segmentar cada palabra.

Respecto al tamaño del vocabulario, al aumentar el tamaño se observa que contienen más palabras completas, lo que lleva a que aumente la corrección del vocabulario.

Con relación al tipo de segmentador, la evaluación de la corrección morfológica muestra resultados diferentes de los obtenidos en los criterios de relevancia y coherencia morfológica. En primer lugar, el algoritmo *BPE* es el que ofrece peores resultados a la hora de segmentar una palabra en todos sus morfemas. *Unigram* es, en este caso, el mejor, lo que confirma que no puede señalarse que un algoritmo sea superior a los otros en el global de la evaluación.

En segundo lugar, las palabras que no necesitan segmentación y que con éxito se segmentan como un solo *token* son el motivo de la superioridad de *Unigram* en este criterio. Todos los aciertos de todos los segmentadores se corresponden con este tipo de palabras monomorfemáticas, y todos fallan completamente a la hora de segmentar morfológicamente las palabras que requieren ser descompuestas en morfemas. En la sección siguiente se muestra el análisis de errores que justifica estos resultados.

5. RESULTADOS DEL ANÁLISIS DE ERRORES

Para completar la evaluación se ha realizado un análisis de errores utilizando los resultados de la evaluación de la corrección morfológica (Tabla 7). Estos resultados se obtuvieron segmentado las palabras del conjunto de datos de corrección morfológica (1.231 palabras) con cada segmentador (en el paso 5 de la sección 3). El análisis de estos errores ha permitido identificar cuatro tipos de errores (Tabla 9):

Segmentador\Error	Tipo 1	Tipo 2	Tipo 3	Tipo 4	Total
Wordpiece_31	436	16	352	248	1.052
Wordpiece_52	568	11	251	217	1.047
BPE_31	98	85	591	350	1.124
BPE_52	149	65	513	378	1.105
Unigram_31	459	15	424	152	1.050
Unigram_52	575	10	335	125	1.045

Tabla 9. Resultados cuantitativos del análisis de errores de los segmentadores. Se han marcado en azul y gris los dos tipos de errores más comunes en cada tipo de segmentador

Tipo de error 1. Infrasegmentación. Las palabras frecuentes que necesitan segmentación no se segmentan porque se identifican en el corpus como muy frecuentes y se incluyen en el vocabulario como un solo *token*. Esto hace que no se incluya la segmentación en morfemas en muchas palabras que sí lo requerirían. Esto ocurre por ejemplo con la palabra *población*, palabras que los seis segmentadores tratan como un solo *token*. En la Tabla 9 se puede observar que la mayor parte de los errores de los algoritmos de segmentación de *WordPiece* y *Unigram* son de este tipo.

Tipo de error 2. Sobresegmentación. Algunas palabras poco frecuentes en el corpus de evaluación que no necesitarían realmente ser segmentadas se descomponen erróneamente en varios *tokens*. Esto ocurre, por ejemplo, con la palabra *tortilla* que el segmentador *BPE* descompone en [*tor*, *tilla*]. Este tipo de error es el menos frecuente en todos los algoritmos de segmentación.

Tipo de error 3. Ausencia del morfema. El morfema correcto no está en vocabulario, por lo que hay que recurrir a otros *tokens o* subpalabras para realizar las segmentaciones. Esto ocurre, por ejemplo, en todos los segmentadores con el sufijo *-érrimo*: ninguno lo tiene en su vocabulario. Los dos *WordPiece* producen la segmentación [*pau, pé, rr, imo*], los dos *Unigram* [*pau, pé, rri, mo*], *BPE_31* [*pa, up, ér, ri, mo*] y *BPE_52* [*pa*, *up*, *ér*, *rimo*]. Este tipo de error parece ser el más crítico, puesto que es el que se comete con más frecuencia en los segmentadores de tipo *BPE*, pero también es el segundo tipo de error más común de *Unigram* y *WordPiece*.

Tipo de error 4. No uso del morfema. No se utiliza el morfema correcto en una segmentación, aunque dicho morfema está incluido en el vocabulario aprendido por el segmentador. Según el tipo de segmentador pueden darse varias situaciones:

1. Los segmentadores *BPE* (Seinrich, 2015) y *Unigram* (Kudo, 2018) contienen en sus archivos las probabilidades aprendidas durante el entrenamiento para las diferentes subpalabras. Así, en el momento de segmentar una palabra, se elige la segmentación cuyas subpalabras son las más probables, lo que a veces puede implicar que el morfema real no sea el elegido. Por ejemplo, al procesar la palabra *escogiendo*, el segmentador *Unigram_52* tendría varias opciones: [*es, cogiendo*] o [*es, cog, iendo*]. Todas estas subpalabras están en su vocabulario aprendido en el entrenamiento. La segunda segmentación sería la correcta desde un punto de vista morfológico. Sin embargo, según las probabilidades que guarda el segmentador en su configuración, la primera es más probable en el corpus de entrenamiento, por lo que es la segmentación elegida finalmente.
2. A diferencia de ellos, WordPiece no almacena en ningún archivo las probabilidades de aparición de las subpalabras en el corpus de entrenamiento. A la hora de segmentar una palabra, WordPiece va buscando la subpalabra más larga de su vocabulario que coincida con el comienzo de la palabra (Song et al., 2021). Por ejemplo, hemos encontrado que el segmentador WordPiece de 31.000 tokens (palabras y subpalabras) de vocabulario segmenta la palabra desamor como [desa, mor] a pesar de que des y amor son tokens de su vocabulario. La razón es que el segmentador busca en el vocabulario, en primer lugar, los tokens desamor, desamo y desam. Como ninguno de ellos está en el vocabulario, se busca, con éxito, desa. A continuación, se repite el procedimiento con el resto de la palabra. Puesto que mor también es un token del vocabulario, se elige [desa, mor] como segmentación final. Este tipo de errores son los que explican que en los prefijos WordPiece presente una coherencia morfológica tan baja, a pesar de la alta cobertura morfológica que presenta su vocabulario: muchos prefijos se pierden al buscar el token inicial más largo.

6. CONCLUSIONES Y TRABAJO FUTURO

Este trabajo aporta una solución original, hasta donde sabemos, para resolver el problema de evaluar cuán eficaces son los vocabularios que utilizan los actuales grandes modelos del lenguaje. Estos vocabularios son generados automáticamente por segmentadores basados en estrategias estadísticas a partir de grandes corpus de texto. El vocabulario es un componente clave para la eficacia y fiabilidad de un modelo del lenguaje porque contiene las unidades lingüísticas que el modelo es capaz

de reconocer y combinar. Uno de los problemas actuales de los grandes modelos del lenguaje es que sus vocabularios están formados por secuencias de caracteres (las denominadas *subpalabras*) que tienen una alta frecuencia de aparición en el corpus de texto y que no siempre se corresponden con unidades lingüísticas de tipo morfema o palabra.

Para poder medir cuán bueno es un vocabulario y, así poder mejorarlo, hemos desarrollado un método de evaluación de su calidad morfológica. Entendemos que la calidad morfológica es el grado de similitud de un vocabulario de morfemas y palabras comparado con el vocabulario real de la lengua. Para valorar/evaluar la calidad lingüística, hemos propuesto medir tres criterios: la relevancia, la coherencia y la corrección morfológica del vocabulario.

Además, hemos desarrollado un procedimiento para realizar estas mediciones. Hemos aplicado el método a la medición de la calidad de seis vocabularios de dos tamaños diferentes generados con los tres segmentadores principales utilizados en la creación de los actuales modelos del lenguaje: *WordPiece* (usado, entre otros, por el modelo *BERT*), *BPE* (utilizado, entre otros, por los modelos *GPT*) y *Unigram* (empleado, por ejemplo, por el modelo del lenguaje *Albert*). Además, se ha explorado la posible influencia del tamaño de los vocabularios en la calidad morfológica.

Los resultados obtenidos indican que:

1. La calidad morfológica de los vocabularios es muy baja. Estos segmentadores, por lo tanto, no parecen apropiados para conservar conocimiento morfológico del español. De forma más concreta, los resultados son los siguientes:
    a. Respecto a la relevancia morfológica de los vocabularios generados, solo para la medida de cobertura se obtienen valores relevantes (el mejor vocabulario de todos, *Wordpiece_32*, tiene una cobertura media total de 31,08%) y, respecto a ellos, *BPE* y *Wordpiece* son superiores al segmentador *Unigram*.
    b. Respecto a la coherencia morfológica, los resultados son malos y aún menos concluyentes respecto a cuál podría ser el mejor. Aunque no merece la pena elaborar un ránquin, se puede indicar que *BPE* es el mejor en la obtención de prefijos, *WordPiece* en raíces y *Unigram* en sufijos y clíticos.
    c. Respecto a la corrección morfológica, los valores también son realmente bajos, con un valor máximo de un 15% de corrección. De nuevo, aunque con estos valores no merece la pena realizar una clasificación, *Unigram* y *WordPiece* parecen superiores a *BPE*. En todos los vocabularios, los

aciertos coinciden con las palabras formadas por un solo morfema y que, por tanto, no tienen segmentación en morfemas.

2. En cuanto a los tamaños de los vocabularios generados, no se puede concluir que aumentar el tamaño mejore los resultados en ninguno de los tres criterios.
3. Respecto al tipo de errores observados se puede concluir que son de cuatro tipos: i) no se segmentan las palabras muy frecuentes en el corpus de entrenamiento del segmentador; ii) sobresegmentación de las palabras poco frecuentes; iii) el morfema correcto no está en el vocabulario; y (iv) el funcionamiento de los segmentadores lleva a elegir otra combinación de tokens que no incluye al morfema correcto, a pesar de que este es un token del vocabulario.

El tipo de error más frecuente encontrado en las segmentaciones generadas por todos los algoritmos evaluados es el tipo 3 (no han aprendido alguno de los morfemas de la lengua), seguido del tipo 1 (infrasegmentación), que es frecuente en *WordPiece* y *Unigram* y el tipo 4 (no utilizan los morfemas, aunque los han aprendido), que es el más frecuente en *BPE*.

En conclusión, se ha encontrado que los algoritmos estadísticos *WordPiece*, *BPE* y *Unigram* utilizados de forma mayoritaria para la creación y funcionamiento de los grandes modelos del lenguaje actuales no son óptimos para segmentar correctamente los morfemas del español y generan vocabularios con subpalabras que no se corresponden con unidades lingüísticas en esta lengua. Para mejorar los resultados de estos algoritmos podría ser útil diseñar soluciones dirigidas a ayudarles a aprender los morfemas de la lengua de forma que disminuyan los errores más frecuentes (tipo 3) y a priorizar su utilización en las segmentaciones para disminuir el segundo tipo de error (tipo 4) que más frecuentemente cometen.

Para terminar, destacamos como contribuciones principales de este artículo:

1. La elaboración de un nuevo método de evaluación de la calidad morfológica de los vocabularios de los grandes modelos del lenguaje generados por segmentadores de naturaleza estadística.
2. La evaluación de la calidad morfológica de los vocabularios en español generados por los tres principales segmentadores actuales.
3. La constatación empírica de que la mejora de los tres principales segmentadores actuales puede dirigirse hacia el aumento y uso de su conocimiento morfológico de la lengua.

En qué medida la calidad morfológica de los vocabularios de los grandes modelos del lenguaje afecta a la eficacia y fiabilidad de dichos modelos es una cuestión todavía no resuelta. Nuestro trabajo actual y futuro se dirige a responder esta cuestión.

Declaración de contribución de autoría

Óscar García-Sierra: investigación, redacción–borrador original.

Ana Fernández-Pampillón Cesteros: supervisión, metodología, redacción–revisión y edición.

Miguel Ortega-Martín: supervisión, redacción–revisión y edición.

Agradecimientos

A nuestros compañeros de Dezzai. A los revisores anónimos, por sus comentarios y sugerencias.

El trabajo ha contado con el apoyo del proyecto del Ministerio de Ciencia e Innovación PID2022-140897OB-I00 ROBOT-TALK.

REFERENCIAS

Bostrom, K., y Durrett, G. (2020). Byte pair encoding is suboptimal for language model pretraining. *arXiv Preprint arXiv:2004. 03720*.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., y Pérez, J. (2023). Spanish pre-trained bert model and evaluation data. *arXiv Preprint arXiv:2308. 02976*.

Church, K. W. (2020). Emerging trends: Subwords, seriously? *Natural Language Engineering*, *26*(3), 375–382.

Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv [Cs.CL]*. Retrieved from http://arxiv.org/abs/1810.04805

Fang, H., Ostendorf, M., Baumann, P., y Pierrehumbert, J. (2015). Exponential language modeling using morphological features and multi-task learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(12), 2410–2421.

Friedman, R. (2023). Tokenization in the Theory of Knowledge. *Encyclopedia, 3*(1), 380-386.

Hofmann, V., Pierrehumbert, J., y Schütze, H. (2021). Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3594–3608.

Kudo, T., y Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv Preprint arXiv:1808. 06226*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., y Soricut, R. (2019). Albert: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv [Cs.CL]*. Retrieved from http://arxiv.org/abs/1907.11692

Moliner, M. (1967/2012). *Diccionario de uso del español*. Madrid: Gredos.

Park, K., Lee, J., Jang, S., y Jung, D. (2020). An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks. *arXiv Preprint arXiv:2010. 02534*.

RAE-ASALE (=Real Academia Española-Asociación de Academias de la Lengua Española) (2009). *Nueva gramática de la lengua española*. Madrid: Espasa.

RAE-ASALE (=Real Academia Española-Asociación de Academias de la Lengua Española) (en línea). *Diccionario de la lengua española*. https://dle.rae.es/

Radford, A., Narasimhan, K., Salimans, T., y Sutskever, I. (2018). *Improving language understanding by generative pre-training*. https://paperswithcode.com/paper/improving-language-understanding-by

Sennrich, R., Haddow, B., y Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv Preprint arXiv:1508. 07909*.

Schuster, M., y Nakajima, K. (2012). Japanese and korean voice search. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5149–5152. IEEE.

Song, X., Salcianu, A., Song, Y., Dopson, D., y Zhou, D. (2020). Fast wordpiece tokenization. *arXiv preprint arXiv:2012.15524*.

Suárez, P. J. O., Sagot, B., y Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Suárez, P. J. O., Romary, L., y Sagot, B. (2020). A monolingual approach to contextualized word embeddings for mid-resource languages. *arXiv Preprint arXiv:2006. 06202*.

Van der Wouden, T. (1990). Celex: Building a multifunctional polytheoretical lexical data base. *Proceedings of BudaLex*, *88*, 363–373.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W. y Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv Preprint arXiv:1609. 08144*.