Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 29-40 recibido 22-03-2025 revisado 05-05-2025 aceptado 16-05-2025

Un Segmentador Morfológico para la Generación de los Vocabularios de Grandes Modelos de Lenguaje en Español

A Morphological Tokenizer for Generating Vocabularies for Large Language Models in Spanish

Óscar García-Sierra,[1,2] Ana Fernández-Pampillón Cesteros,[1] Miguel Ortega-Martín[1,2]
[1]Universidad Complutense de Madrid
[2]dezzai
{oscarg02, apampi, m.ortega}@ucm.es

Resumen: En este artículo presentamos un segmentador para la generación de vocabularios de grandes modelos del lenguaje que, a diferencia de las aproximaciones estadísticas actuales, está basado en conocimiento morfológico y léxico del español. El objetivo es tratar de mejorar la eficacia de estos modelos, especialmente en tareas de carácter lingüístico, resolviendo los problemas de falta de relevancia, coherencia y corrección morfológica que presentan sus vocabularios estadísticos. El segmentador morfológico que presentamos divide los textos de entrada en morfemas reales del español en vez de en subpalabras frecuentes que no siempre coinciden con morfemas. Para ello utiliza un lexicón generado semiautomáticamente de 755.467 palabras y morfemas del español y una gramática, generada manualmente, de 234 reglas morfológicas. Hemos evaluado el segmentador y su vocabulario aplicando una metodología desarrollada en nuestro trabajo previo, y hemos podido comprobar que el segmentador morfológico genera un vocabulario con una corrección morfológica del 98% en un tiempo razonable de forma que pueda utilizarse con grandes modelos de lenguaje.
Palabras clave: segmentación, morfología, subpalabras, LLM.

Abstract: Here we present a tokenizer for the generation of vocabularies for large language models which, unlike current statistical approaches, is based on morphological and lexical knowledge of Spanish. The goal is to improve the effectiveness of these models, especially in linguistically-oriented tasks, by addressing issues related to the lack of relevance, coherence, and morphological accuracy found in statistical vocabularies. The morphological tokenizer we present splits input texts into actual Spanish morphemes rather than frequent subword units. To achieve this, it uses a semi-automatically generated lexicon of 755,467 Spanish words and morphemes, along with a manually crafted grammar containing 234 morphological rules. We evaluated the tokenizer and its vocabulary using a methodology developed in our previous work, and found that the morphological tokenizer produces a vocabulary with 98% morphological accuracy in a reasonable amount of time, making it suitable for its use with large language models.
Keywords: tokenization, morphology, subwords, LLM.

1 Introducción

La segmentación de texto, conocida en inglés como *tokenization*, es el proceso de descomponer el texto en unidades menores llamadas token (Friedman, 2023). Actualmente, la segmentación juega un papel crucial en la preparación de los datos que se utilizan para entrenar los modelos del lenguaje neuronales, ya que define las unidades discretas de información sobre las que se construyen estos modelos (Friedman, 2023).

En el entrenamiento de los actuales grandes modelos del lenguaje (en adelante LLM, del inglés *Large Language Models*) (Jurafsky y James, 2025) basados en arquitecturas

ISSN 1135-5948 DOI 10.26342/2025-75-2

Transformer (Vaswani et al., 2017), los vocabularios se construyen principalmente a partir de la segmentación estadística en palabras y subpalabras de los textos de entrenamiento.

Los algoritmos de segmentación empleados por los LLM se basan en enfoques estadísticos sin conocimiento lingüístico. Estos algoritmos analizan grandes corpus de texto y determinan las unidades de segmentación en función de la frecuencia con la que aparecen las cadenas de caracteres. Como resultado, las palabras frecuentes suelen mantenerse como un único token, mientras que las menos comunes se dividen en varios token de subpalabras. Por ejemplo, el segmentador WordPiece del modelo BETO (Cañete et al., 2023) trata la palabra "población" como un único token, pero divide "perduración" en las subpalabras 'perd', y '##uración', ninguna de ellas morfemas del español.

Entre los algoritmos de segmentación más utilizados actualmente por los LLM destacan Byte-Pair Encoding (BPE) (Sennrich, 2015), WordPiece (Schuster, 2012; Wu, 2016) y Unigram (Kudo, 2018). Estos algoritmos utilizados por LLM de referencia como BERT (Devlin et al., 2019), que emplea WordPiece, GPT (Radford et al., 2018) o RoBERTa (Liu et al., 2019), que utilizan BPE, y Albert (Lan et al., 2019), que utiliza Unigram.

La principal ventaja de los segmentadores estadísticos frente a los basados en modelos lingüísticos radica en su eficiencia: permiten construir rápidamente vocabularios de gran tamaño a partir de extensos corpus sin necesidad de intervención humana. Sin embargo, diversos estudios han demostrado que la segmentación en subpalabras no siempre se corresponde con los morfemas reales. Así, Church (2020) observa que algunas palabras complejas en inglés se dividen en demasiadas subunidades sin correspondencia clara con morfemas o palabras. Bostrom y Durret (2020) comparan BPE y Unigram en inglés y japonés, concluyendo que Unigram logra una segmentación más alineada con los morfemas de ambos idiomas. En García Sierra et al. (2025a y b), se evalúa, para el español, la calidad de los vocabularios generados por los tres algoritmos BPE, WordPiece y Unigram obteniéndose que la relevancia morfológica, la coherencia y la corrección morfológicas son muy bajas. Hoffman et al. (2021) analizan el comportamiento de WordPiece con palabras complejas en inglés y confirman que la segmentación generada no siempre coincide con morfemas reales. Nzeyimana y Rubungo (2022) llegan a conclusiones similares para el caso del Kinyarwanda. Mager et al. (2022) utilizaron el segmentador Morfessor (Virpioja et al., 2013) para comparar los resultados de modelos que emplean subpalabras estadísticamente relevantes frente a modelos que emplean segmentadores morfológicos en la traducción a varias lenguas polisintéticas. Como resultado encontraron que el LLM con un vocabulario morfológico mejora los resultados del LLM con un vocabulario estadístico -generado por el segmentador BPE- en cinco de las ocho lenguas. Asimismo, Park (2020) llega a conclusiones similares en su estudio sobre el coreano.

La pregunta que surge, entonces, es ¿hasta qué punto mejorar el vocabulario de un LLM puede mejorar dicho LLM?

Responder a esta pregunta no es sencillo. En primer lugar, es necesario disponer de un segmentador que genere vocabularios con las palabras y morfemas de la lengua o, al menos, los más frecuentes si fuese necesario reducir el tamaño del vocabulario. En segundo lugar, será necesario evaluar el grado de influencia de la calidad del vocabulario en la calidad del LLM.

En este artículo abordamos el primer problema, la búsqueda de un segmentador para LLM que genere, a partir de los textos del corpus de entrenamiento del LLM, vocabularios de palabras y morfemas. Nos centramos en la lengua española porque presenta una complejidad morfológica superior a la del inglés, y, además, por ser una de las lenguas más habladas en el mundo.

Antes de presentar nuestra propuesta, revisamos en la sección segunda, los segmentadores disponibles en español. En la tercera sección presentamos el segmentador morfológico para el español que hemos creado. En la sección cuarta mostramos los resultados de la evaluación del segmentador y su discusión. Finalmente, en la sección quinta se presentan las conclusiones y el trabajo en curso.

2 Segmentadores morfológicos del español. Antecedentes

Existen varias aplicaciones que procesan el lenguaje natural a nivel morfológico como son los lematizadores (Balakrishnan y Lloyd-Yemoh, 2014), los extractores de raíces (Porter, 2001) o los analizadores morfológicos (Carreras

et al., 2004; Honnibal y Montaine, 2017). Sin embargo, el número de soluciones disponibles para segmentar un texto en palabras o morfemas son más escasas, especialmente en español.

Uno de los primeros trabajos sobre segmentación morfológica en español es SMORPH (Aït-Mokhtar y Rodrigo Mateos, 1995). Este trabajo aborda la segmentación de forma parcial centrándose únicamente en la flexión. Utiliza un lexicón de lemas, una lista de prefijos y de sufijos flexivos y reglas de flexión, pero no indica el tamaño de estos recursos ni ofrece datos sobre la evaluación del segmentador.

Morfessor (Virpioja et al., 2013), por su parte, aborda la segmentación morfológica de forma completa mediante una estrategia estadística. Utiliza un algoritmo de aprendizaje supervisado basado en el modelo de campos aleatorios condicionales. Versiones posteriores de Morfessor, como Flatcat (Grönroos et al., 2014) emplean Modelos Ocultos de Markov, con lo que, en inglés y finés, mejora los resultados de Morfessor original. Este sistema, además del análisis morfológico, devuelve la segmentación en morfemas.

Aunque inicialmente Morfessor fue creado para el inglés, el finés, el árabe, el turco y el alemán, es fácilmente entrenable para otros idiomas. En español el trabajo de Méndez-Cruz et al. (2016) evaluó varios segmentadores basados en Morfessor, un segmentador estadístico no supervisado de desarrollo propio y otro basado en el algoritmo Paramor (Monson et al., 2007). La evaluación se realizó utilizando el enfoque BPR (*Boundary Precision Recall*), que se basa en comparar los índices de las fronteras entre los morfemas usando las métricas tradicionales de precisión, cobertura y F1. Para ello se utilizó un conjunto de 1600 palabras con los morfemas anotados manualmente. De todos los segmentadores que evaluaron, el mejor fue el algoritmo creado por ellos, logrando un 0,736 de precisión, 0,688 de cobertura y un 0,711 de valor-F. En Gutierrez-Vasques et al. (2019) se presenta una versión de Morfessor 2.0 que obtiene, utilizando también BPR, un 0,84 precisión, 0,806 de cobertura y un 0,823 de valor-F. El corpus de evaluación, sin embargo, es diferente de Méndez-Cruz et al. (2016) y ni los segmentadores de ambos trabajos ni los corpus de evaluación son accesibles.

En este contexto, inicialmente, construimos una nueva versión de Morfessor en español. Para ello, llevamos a cabo un entrenamiento utilizando como corpus todos los lemas del español extraídos de Diccionario de la Lengua Española (DLE) (Real Academia Española, s.f) y un conjunto de 664.394 verbos conjugados del español (Molino de ideas, 2012). Sin embargo, solo logramos que el segmentador llegase hasta un 68% de corrección (*accuracy*) utilizando un corpus de evaluación que denominamos “de corrección morfológica” y que se describe en García-Sierra et al., (2024a). Este resultado tan bajo nos llevó a descartar esta solución.

Dado que, hasta donde llega nuestro conocimiento, no parece que existan otras alternativas, decidimos abordar la construcción de un nuevo segmentador morfológico del español que tuviera una precisión por encima del 84% conseguido por Gutiérrez-Vasques et al. (2019), y, con una velocidad adecuada para su integración con los LLM actuales.

3 Segmentador morfológico para generar vocabularios de LLM en español

La idea básica que hemos aplicado para crear el segmentador morfológico es cambiar la estrategia estadística por una estrategia basada en utilizar conocimiento léxico y morfológico para guiar la segmentación. El segmentador creado se encuentra disponible en el repositorio github.[1]

3.1 El algoritmo de segmentación

El segmentador se apoya en un índice, previamente construido, de 755.467 palabras y morfemas del español (§ 3.3.5) y en 49 reglas de una gramática morfológica más amplia con 234 reglas (§ 3.3.3) que describe los sufijos flexivos, prefijos, sufijos apreciativos y pronombres clíticos del español. El lexicón se configura y utiliza como un índice de palabra-morfemas. Estos recursos se describen, posteriormente, en la sección 3.4.

El funcionamiento del segmentador se resume en la figura 1. Dado un texto de entrada, en un primer paso, se presegmenta en token utilizando espacios en blanco y puntuación. El resultado de esa presegmentación será, en su

[1] https://github.com/ogarciasierra/spanish-morph-tokenizer

mayor parte una secuencia de palabras, aunque también estarán incluidos números, fechas, siglas, entidades nombradas y otros elementos no lingüísticos. Como el segmentador se centra sólo en las palabras de la lengua, para los token que no son palabras, devuelve una segmentación en caracteres y así evita que se incremente sustancialmente el vocabulario con token no léxicos.

En un segundo paso se busca en el índice de palabra-morfemas cada uno de los token de la presegmentación. Si está en él, se obtiene la segmentación correspondiente. Si no está, se aplican las 49 reglas de segmentación de la gramática morfológica para la identificación de los morfemas de tipo:

1) morfemas de número,
2) morfemas de género,
3) prefijos,
4) apreciativos,
5) clíticos

Las reglas se aplican en orden. Cuando se aplica una regla y se identifica un morfema, en un tercer paso, se busca, en el índice, la forma base que es la forma resultante de eliminar el morfema identificado. En determinados nombres y adjetivos obtener la forma base puede implicar añadir las vocales finales de género masculino "-o"/"-e". Por ejemplo, si tuviésemos la palabra "guapa" y hubiésemos identificado la "a" final, buscaríamos la palabra "guapo" en el índice. En el caso de "presidenta" encontraríamos en el índice la palabra "presidente".

Si la forma base está en el índice se obtiene su segmentación. Sino, la palabra se segmenta en caracteres puesto que no se ha reconocido como palabra de la lengua. Esta solución, por lo tanto, es la que se aplica cuando un token no es reconocido como una palabra de la lengua como es el caso de los números, fechas, nombres propios, etc.

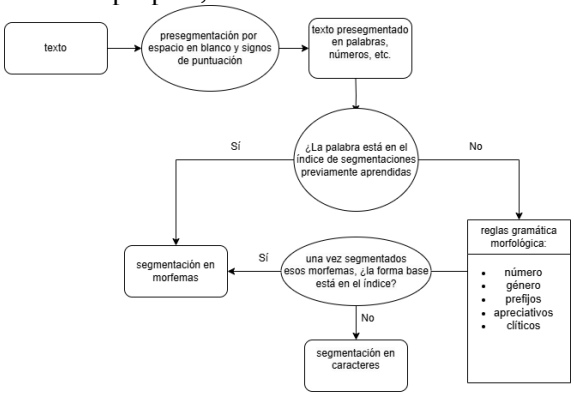

Figura 1. Algoritmo del segmentador.

Para ilustrar el funcionamiento del algoritmo de segmentación tomamos como ejemplo la frase:

"me compré una casita en 1999".

En el primer paso, la frase se presegmenta usando los espacios en blanco, lo que da lugar a seis token:

["me", "compré", "una", casita", "en" y "1999"]

En el segundo paso, token por token, se busca el token en el índice; en caso de no estar, se aplica la gramática y, finalmente, si tampoco se puede aplicar la gramática, en un tercer paso, se segmenta en caracteres. Así,

i) la primera palabra "me" es un pronombre personal que aparece en el índice, por lo que se accede a su segmentación, que es un solo morfema: ["me"].

ii) La segunda palabra, "compré" se busca en el índice y, como también está, se obtiene la segmentación en dos morfemas ["compr" y "##é"]. Usamos, como en Wordpiece, la marca "##" para indicar que un token no es comienzo de palabra.

iii) Con la tercera palabra, "una", una vez comprobado que no está en el índice, se aplica la gramática de reglas: (i) identificar el morfema de número, sin éxito, y, (ii), identificar, con éxito, el morfema de género en la vocal "-a". Una vez identificado el morfema se busca la cadena restante, "un" en el índice. Al encontrarse "un" en el índice se devuelve el resultado, que en este caso es el propio lema, ["un"], y "a" como morfema flexivo de género, dando como resultado ["un", "##a"]. Si no se encontrase "un", se probaría a buscar "uno", añadiendo la "-o" del masculino.

iv) La cuarta palabra, "casita", tampoco está en el índice, así que debe aplicarse la gramática. El proceso de aplicación de las reglas morfológicas es: (1) la búsqueda de morfemas de número no arroja resultados, por lo que se busca, (2) el de género. Como antes, se encuentra la "a" final, pero en este caso la forma resultante "casit" no está en el índice, por lo que se busca, (3) prefijos, sin tener éxito, así que se aplica, (4) sufijos apreciativos y se encuentra el sufijo "it" que es una marca de

diminutivo. Se extrae ese morfema y se busca la forma base resultante en el índice. Esta búsqueda de la forma base implica la reconstrucción de la forma base añadiendo la vocal final "a" a la raíz, puesto que previamente la hemos identificado como la marca de género. Si esta no estuviese, se probaría a buscar con la vocal del masculino (como en "guapa" y "guapo") y sin ella (como en "perdedora" y "perdedor").

Una vez obtenida la forma base, la palabra "casa", en un tercer paso, se busca en el índice y se encuentra su segmentación ["cas", "##a"]. Con esta segmentación se construye la segmentación final: la raíz de la forma base "cas", el morfema derivativo "it" y el de género "a", dando lugar a ["cas", "##it", "##a"].

v) La quinta palabra, "en", está en el índice y su segmentación es ["en"].

vi) Por último, "1999" es un ejemplo de un token que no tiene segmentación posible en morfemas lo que se obtiene tras haber pasado, sin éxito, por los dos pasos de búsqueda en el índice y de aplicación de la gramática. En este caso se segmenta en caracteres, dando lugar a ["1", "##9", "##9", "##9"].

3.2. Creación del vocabulario del LLM

Del tamaño de los vocabularios de los LLM depende el número de parámetros posibles del modelo. Así, los modelos basados en Transformers suelen emplear vocabularios que van desde 30.000 token, como BETO, hasta, por ejemplo, los 100.000 token, como GPT-4.

Los segmentadores basados en algoritmos estadísticos limitan el tamaño del vocabulario, simplemente, seleccionando los token en orden descendente a su frecuencia de aparición en el corpus de entrenamiento.

En nuestro caso, el vocabulario con todos los morfemas del español tiene un tamaño de 52.000 token (morfemas). Lo ideal sería que los LLM pudieran utilizar estos 52.000 token. Sin embargo, necesitamos limitar el tamaño a 31.000 token para poder realizar las evaluaciones de calidad (García-Sierra et al., 2024b)

Para limitar el tamaño del vocabulario además de tener en cuenta la mayor frecuencia del token (morfema) en el corpus, tenemos en cuenta la frecuencia de los token (morfemas) en el diccionario general DLE. Hemos establecido empíricamente la proporción de morfemas del corpus y morfemas del diccionario DLE en, aproximadamente, 90%-10%. La idea es tener en cuenta los morfemas más representativos de la lengua española y corregir, en cierta medida, el sesgo hacia el corpus de entrenamiento.

Así, el vocabulario de tamaño 31.000 token se construye con los 28.000 morfemas más frecuentes del corpus (aquí hay que incluir, además, signos de puntuación, números que también deben ser token del vocabulario para no provocar casos de token fuera de vocabulario) del corpus de entrenamiento y los 3000 morfemas más frecuentes del diccionario

DLE. La calidad de este vocabulario se evalúa en la sección 4ª.

En las siguientes subsecciones se detalla la creación del índice y los recursos léxicos del segmentador.

3.3. Creación del índice del segmentador

El índice que utiliza el segmentador se construye con el objetivo de agilizar la búsqueda de las segmentaciones en morfemas de las palabras. Se trata de recoger las segmentaciones previamente aprendidas para todos los lemas del DLE y la lista de 664.394 verbos conjugados del español (Molino de ideas, 2012), de forma que, si una palabra se encuentra en el índice, el obtener la segmentación es más rápido que aplicar las reglas de la gramática morfológica del español. Estas reglas, como hemos visto, solo se aplican si la palabra no está en el índice, lo que representa un 20,7% de los casos en los experimentos que hemos realizado.

Para generar el índice se crea, en primer lugar, automáticamente, un lexicón del español con 755.467 formas; en segundo lugar, se crea, también un repertorio de morfemas del español y, en tercer lugar, se crea, manualmente, una gramática con 234 reglas morfológicas de las cuales 49 son utilizadas, también, por el segmentador para identificar los morfemas de número, género, prefijos, apreciativos y clíticos.

3.3.1. El lexicón

El lexicón está formado por todos los lemas del DLE, extraídos de su versión en línea, y por la

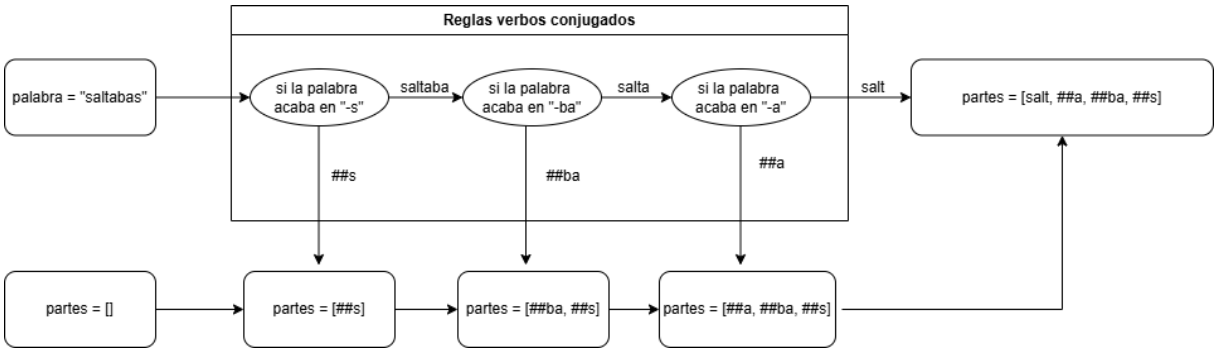

Figura 2. Ejemplo de las reglas de segmentación de los verbos conjugados.

lista de 664.394 verbos conjugados.[2] Se organiza por categorías gramaticales y, en el caso de los nombres y el de los adjetivos, también se separan aquellas palabras que tienen marca de género de las que no. La tabla 1 muestra el total de palabras por categoría gramatical y marcas.

Categoría gramatical	Palabras
Nombres con marca de género	47.676
Nombres sin marca de género	2.159
Adjs. con marca de género	16.318
Adjs. sin marca de género	6.477
Verbos en infinitivo	16.289
Verbos conjugados	664.392
Adverbios	2.001
Pronombres	76
Determinantes	44
Preposiciones	20
Conjunciones	15
Total	755.467

Tabla 1. Composición del lexicón.

3.3.2. Repertorio de morfemas

Los morfemas que serán utilizados por las reglas de segmentación han sido obtenidos manualmente de la Nueva Gramática de Lengua Española[3] (Real Academia Española, 2009). Están etiquetados con su tipo, prefijo, sufijo derivativo, sufijo flexivo, sufijo apreciativo, clítico, y con la categoría gramatical a la que pueden añadirse. La tabla 2 muestra la composición del repertorio de morfemas

Tipo de morfema	Total
Prefijos	61
Sufijos derivativos	156
Sufijo flexivo	26
Apreciativos	17
Clíticos	11

Tabla 2. Composición del repertorio de morfemas.

3.3.3. La gramática morfológica

Esta gramática se utiliza para crear el índice, y una parte de ella, 49 reglas referidas a flexión de número, género, prefijos, apreciativos y clíticos, la utiliza el segmentador.

En la gramática, cada una de las categorías gramaticales de tipo variable, nombres, adjetivos, verbos y adverbios, tiene su correspondiente conjunto de reglas de segmentación morfológica. Hay un total de 234 reglas en la gramática. Todas las reglas han sido extraídas manualmente de la Nueva Gramática de la Lengua Española. Además, se establece un orden de aplicación de las reglas:

1. Sufijos derivativos
2. Sufijos flexivos
3. Prefijos
4. Sufijos apreciativos (aumentativos, diminutivos y peyorativos)
5. Clíticos

Tanto para la construcción del índice como en el segmentador se aplican buscando los morfemas que afectan a dicha categoría gramatical bien en la terminación de la palabra en el caso de los sufijos, o bien, en el comienzo de la palabra en el caso de los prefijos. La

[2] Proporcionados por la empresa "Molino de ideas": https://www.molinodeideas.com/

[3] https://www.rae.es/gramática/

figura 2 muestra un ejemplo de cómo se aplican las reglas a un verbo conjugado.

Además de las 234 reglas, una serie de reglas adicionales controlan las excepciones que se pueden producir al añadir determinados morfemas, como pueden ser los cambios de la sílaba tónica que repercuten en cambios en las tildes o los cambios de algunas consonantes finales, como puedes ser la "-z" final al pasar a plural.

En el caso de las categorías invariables, al ser monomorfemáticas, la segmentación resultante se corresponde con la propia palabra, puesto que no tienen segmentación morfológica.

3.3.4. Algoritmo de generación del índice

Para crear el índice, como muestra la figura 3, dada una palabra de entrada proveniente del lexicón, se siguen dos pasos: (1) se busca en el lexicón para extraer su categoría gramatical, y, (2) se aplican las reglas correspondientes a la categoría gramatical para buscar los morfemas de dicha palabra. Si una palabra tiene más de una categoría, la búsqueda del paso (1) se realiza hasta no encontrar más categorías gramaticales para dicha palabra, por lo que se localizan todas las categorías posibles y, en consecuencia, en el paso (2) se generan, para dicha palabra, varias segmentaciones, una para cada categoría.

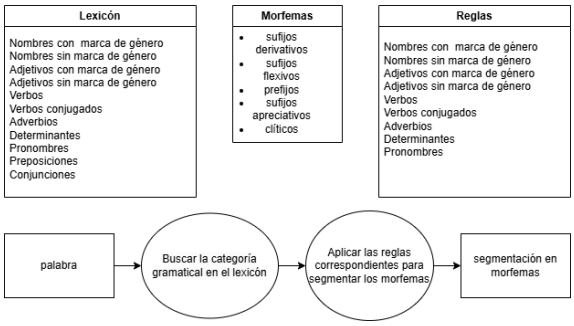

Figura 3. Arquitectura y funcionamiento del generador del índice.

Un aspecto a tener en cuenta del funcionamiento del generador del índice es que permite sobresegmentar la vocal final de aquellos nombres y adjetivos que pueden llegar a perderla final debido a algún proceso morfológico. Por ejemplo, la palabra "casa" debería segmentarse como un solo token porque la "a" final no es un morfema de género en su caso. Sin embargo, si activamos el modo sobresegmentación del generador, a la hora de construir el índice se buscarían en el lexicón palabras que tuviesen la misma raíz, como podría ser "caserío", y "casa" pasaría a segmentarse como ["cas", "a"]. Si el modo sobresegmentación no se activa, "casa" se segmentaría como ["casa"]. El objetivo de esto es: (i) que ambas palabras compartan la raíz y, (ii), poder generar un vocabulario de menor tamaño reduciendo el número de palabras monomorfemáticas.

3.3.5. El índice

El resultado de usar el generador es un índice que contiene una segmentación para cada una de las 755.467 palabras del lexicón.

A pesar de que el generador puede crear más de una segmentación para cada palabra, por razones de eficiencia, en la versión del segmentador que hemos construido, solo hemos utilizado una segmentación por palabra. Si se necesita incluir todas las segmentaciones posibles de una palabra con más de una categoría gramatical sería necesario incluir un análisis morfológico que identifique estas categorías, lo que aumenta la complejidad y el tiempo de procesamiento de la generación del índice. Asimismo, en el caso del segmentador, sería necesario también incluir un paso de análisis morfológico que desambigüe la categoría gramatical de la palabra.

Para abordar los casos en los que una palabra puede tener varias categorías gramaticales, actualmente, priorizamos las categorías invariables y después, por orden de palabras con dicha categoría en el lexicón: verbos, nombres y adjetivos,

Hemos comprobado que, para la segmentación de mil palabras, el uso del índice ha reducido el tiempo de segmentación diez veces respecto a la opción de aplicar la gramática completa con las 234 reglas morfológicas.

La tabla 3 contiene una muestra del índice.

Palabra	Segmentación
habitación	habit, ##ación
guapo	guap, ##o
desprestigiar	des, ##prestig, ##iar
desde	desde

Tabla 3. Muestra del índice.

4 Evaluación de la calidad morfológica del vocabulario

Definimos la *calidad morfológica* como el grado de semejanza de un vocabulario con el vocabulario real de morfemas y léxico de una lengua. Para evaluar la calidad morfológica vamos a utilizar el método de evaluación propuesto en nuestro trabajo de investigación previo (García-Sierra et al., 2024a). Adicionalmente utilizamos las métricas de precisión, cobertura y F1 del enfoque BPR para comparar nuestro segmentador con los antecedentes de la sección 2.

4.1. Método de evaluación

El método de evaluación de la calidad morfológica está basado en la medición de cuatro criterios de calidad:

1) La *relevancia morfológica del vocabulario*: mide cuántos de los token del vocabulario corresponden a morfemas reales de la lengua.
2) La *coherencia morfológica*: evalúa con qué frecuencia palabras que comparten un mismo morfema son segmentadas con ese morfema.
3) La *corrección morfológica de la segmentación*: evalúa si el segmentador divide correctamente las palabras en sus morfemas reales.
4) La *precisión, cobertura y F-score* siguiendo el enfoque *BPR* para poder establecer comparaciones con otros segmentadores morfológicos en español.

El procedimiento de evaluación consiste en:

1) Generación del vocabulario a partir de un corpus del corpus de entrenamiento del LLM mediante el segmentador.
2) Utilización de los tres conjuntos de evaluación creados en García-Sierra et al., (2024a), uno para cada criterio descrito anteriormente. Están disponibles online en español.[4]
3) Medida de la relevancia morfológica del vocabulario, utilizando los morfemas reales de la lengua disponibles en el repositorio de morfemas (sección 3.3.2).
4) Medida de la coherencia morfológica, como se detalla en García-Sierra et al. (2024a). Para cada tipo de morfema, prefijos, sufijos, raíces y clíticos se crean tres listas llamadas un token, varios-token correcto y varios-token incorrecto (Tabla 5). Las listas se generan 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) se calcula la longitud de cada lista en forma de porcentaje respecto al total de palabras. La coherencia morfológica se corresponde con la lista “varios -token_correcto”.
5) La *corrección morfológica (accuracy)*. Para ello se compara cada par (palabra-segmentación) con el conjunto de datos de evaluación y se cuenta una segmentación como correcta cuando se comprueba que todos los morfemas de la palabra se han segmentado correctamente.
6) La *precisión, cobertura y valor-F1 del BPR*. Para ello se utiliza el conjunto de datos de evaluación y se mide el número de “fronteras” de los segmentos obtenidos respecto a las “fronteras” correctas.

4.2. Evaluación del vocabulario morfológico

Aplicamos el método de evaluación de la calidad morfológica al vocabulario generado por el segmentador morfológico seleccionando un tamaño de 31.000 token y el corpus de entrenamiento *oscar-small*[5], compuesto por 600.000 frases en español. Las razones de estas elecciones fueron la disponibilidad del modelo, del corpus y la capacidad de los equipos informáticos que utilizamos.

Para comparar los resultados de la evaluación seleccionamos el segmentador estadístico *Wordpiece* que es, junto con BPE y Unigram, utilizado mayoritariamente por los LLM actuales (entre otros, por el LLM BETO) y, en media fue el que mejor resultados obtuvo en las evaluaciones realizadas en nuestro

[4] https://github.com/ogarciasierra/spanish-subwords-evaluation

[5] https://huggingface.co/datasets/nthngdy/oscar-small

trabajo anterior (García-Sierra et al, 2024a). Asimismo, seleccionamos, de los antecedentes revisados, los dos segmentadores morfológicos del español de los que se dispone de métricas de evaluación Méndez-Cruz et al. (2016), Gutierrez-Vasques et al. (2019).

Las tablas 4 a 8 presentan los resultados en comparación con los del segmentador Wordpiece entrenado en nuestro trabajo anterior para generar un vocabulario de tamaño 31.000 token, Wordpiece_31, que ya fue evaluado en García-Sierra et al. (2024a).

Los resultados que hemos obtenido son:

1) respecto a la *relevancia morfológica* del vocabulario (tabla 4), se obtiene un valor de cobertura de 100%, lo que indica que todos los morfemas del conjunto de datos de evaluación están presentes en el vocabulario del modelo.

Tipo		Total	Cobertura (%)
prefijos	M	61	**100**
	W		88,5
sufijos	M	175	**100**
	W		73,56
raíces	M	5.000	**100**
	W		9,97
total	M	5.236	**100**
	W		22,55

Tabla 4. Resultados de la evaluación de la relevancia morfológica del segmentador morfológico (M) en comparación con los del segmentado Wordpiece_31 (W).

2) Respecto a la *coherencia morfológica* (tabla 5), se observa que en la totalidad de los casos el segmentador segmenta de forma que palabras que comparten un morfema siempre segmentarán este morfema. El segmentador morfológico mejora los resultados del segmentador Wordpiece.

Tipo		Un token (%)	Varios token	
			El morfema es un token (%)	El Morfema no es un token (%)
prefijos	M	**0**	**100**	**0**
	W	0,71	13,4	85,84
raíces	M	**0**	**100**	**0**
	W	1,64	16,0	83,33
sufijos	M	**0**	**100**	**0**
	W	0,74	15,2	84,05
clíticos	M	**0**	**100**	**0**
	W	5,61	61,4	32,99
totales	M	**0**	**100**	**0**
	W	0,86	15,2	83,93

Tabla 5. Resultados evaluación coherencia morfológica del segmentador morfológico (M) en comparación con los del segmentador Wordpiece_31 (W).

3) Respecto al criterio de *corrección morfológica* (tabla 6), el vocabulario tiene una corrección del 98,5%, muy por encima del 15% en torno al cual se situaban los resultados de los segmentadores estadísticos BPE, WordPiece y Unigram evaluados en nuestros trabajos García-Sierra et al. (2024a) y García-Sierra et al. (2024b).

Segmentador	Total palabras	Corrección morfológica (*accuracy*) (%)
Morfológico	1.231	**98,5**
Wordpiece_31		14,54

Tabla 6. Resultados evaluación corrección morfológica en comparación con los del segmentador Wordpiece_31.

4) Respecto a los resultados de precisión, cobertura y F1 del BPR (tabla 7) parece que mejoran sensiblemente el estado de la cuestión. No es posible, sin embargo, una comparativa real porque las medidas se han tomado con corpus de evaluación diferentes al no estar disponibles los corpus de evaluación de Méndez-Cruz et al. (2016) ni Gutiérrez-Vasques et al. (2019).

Segment-.	Precisión	Cobertura	F1
Morf.	0.9855	0.9935	0.9895
Méndez-Cruz et al. (2016)	0,736	0,688	0,711
Gutierrez-Vasques et al. (2019)	0,84	0,806	0,823

Tabla 7. Resultados BPR de precisión, cobertura y F1 del segmentador y resultados del estado de la cuestión.

Finalmente, la comparativa de los tiempos de segmentación del corpus de entrenamiento (*oscar-small*) muestra que el segmentador está dentro de los rangos de velocidad adecuados para un LLM, como muestra la tabla 8.

Segmentador	Tiempo (s) para la

	segmentación de 1231 palabras
Wordpiece_31	2.58
Morfológico	2.59

Tabla 8. Velocidad de segmentación del segmentador morfológico en comparación con Wordpiece_31.

Con estos resultados podemos considerar que el segmentador morfológico constituye una solución posible para poder estudiar el grado de influencia de la calidad del vocabulario en la calidad de los LLM.

4.3. Estudio de errores

En nuestro trabajo previo (García-Sierra et al., 2024) detectamos cuatro tipos de errores que cometían los segmentadores estadísticos:

1) Infrasegmentación: no se segmenta una palabra con morfemas porque es muy frecuente.
2) Sobresegmentación: una palabra monomorfémica se divide en varios token.
3) El morfema real no está en el vocabulario.
4) No se usa el morfema, aunque está en el vocabulario por la estrategia de segmentación.

La tabla 9 contiene el total de errores de cada tipo del segmentador morfológico en comparación con el segmentador Wordpiece de 31.000 token de vocabulario. En todos los tipos de errores el segmentador morfológico mejora considerablemente los del segmentador Wordpiece_31.

Segmentador	Tipo 1	Tipo 2	Tipo 3	Tipo 4	Total
Morfológico	0	10	2	6	18
Wordpiece_31	436	16	352	248	1.052

Tabla 9. Estudio de errores del segmentador morfológico en comparación con Wordpiece_31.

Los errores de tipo 1 de infrasegmentación se reducen por completo.

Los errores de tipo 2, sobresegmentación, afectan a prefijos y sufijos derivacionales en los casos excepcionales en que no deben usarse. Por ejemplo, el sufijo “era” en los nombres sin marca de género, como “borrachera” o “leonera” no es morfema de “madera”. Este error se puede corregir usando una lista de excepciones.

Los errores de tipo 3, el morfema no está en el vocabulario, ocurre en raíces poco frecuentes como, por ejemplo, en “melifluo”.

Por último, los errores de tipo 4, algún segmento no es un morfema, se dan, bien cuando segmenta un morfema que no es el correcto, o bien, porque son excepciones como por ejemplo, cuando es necesario añadir una consonante intermedia, como en “ensanchar”.

5. Conclusiones y trabajo en curso

El objetivo de este trabajo era encontrar de un segmentador que genere vocabularios en español de alta calidad para ser utilizados por los LLM. El objetivo se puede considerar logrado en la medida en que se ha desarrollado un segmentador morfológico del español con una corrección de 98,5%, una coherencia del 100% y una cobertura total. Estos valores son significativamente superiores a los obtenidos por el segmentador estadístico *WordPiece*, uno de los tres más utilizados por los LLM actuales, y por los anteriores segmentadores del español con métricas publicadas. Se ha comprobado además que, en el vocabulario generado, se han reducido substancialmente todos los tipos de errores estudiados respecto a *WordPiece*. La velocidad de funcionamiento del segmentador morfológico es, también, adecuada para su integración con los LLM.

Este trabajo aporta, además, tres recursos lingüísticos del español necesarios para crear el segmentador: un lexicón con 755.467 palabras, una gramática morfológica 234 reglas y un generador de índices palabra-morfemas.

Respecto a las limitaciones, es relevante señalar que el segmentador se ha creado con el léxico del diccionario general del español DLE lo que significa que no está preparado para trabajar con textos de especialidad. En este caso requería incluir el léxico especializado.

Actualmente, estamos utilizando el segmentador morfológico para comprobar si, efectivamente, una mayor calidad morfológica del vocabulario y del segmentador repercute en mejorar los resultados del LLM que los utilice.

Agradecimientos

A nuestros compañeros de la empresa *Dezzai*. A Eduardo Basterrechea, fundador y director ejecutivo de la empresa *Molino de Ideas* por cedernos la lista de verbos conjugados en español.

Estos resultados son parte del proyecto de I+D+i ROBOT-TALK PID2022-140897OB-I00 financiado por

MCIN/AEI/10.13039/501100011033/ y FEDER/UE.

Bibliografía

Aït Mokhtar, S., y J. L. Rodrigo Mateos. (1995). Segmentación y análisis morfológico de textos en español utilizando el sistema SMORPH. Revista de Procesamiento del lenguaje natural. Nº 17 (sept. 1995), pp. 29-41.

Balakrishnan, V., y E. Lloyd-Yemoh. (2014). Stemming and lemmatization: A comparison of retrieval performances. Lecture notes on software engineering, 2(3), 262-267.

Bostrom, K., y G. Durrett. (2020). Byte pair encoding is suboptimal for language model pretraining. arXiv Preprint arXiv:2004.03720.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, y J. Pérez. (2023). Spanish pre-trained BERT model and evaluation data. arXiv Preprint arXiv:2308.02976.

Carreras, X., I. Chao, L. Padró, y M. Padró. (2004, May). FreeLing: An Open-Source Suite of Language Analyzers. In LREC (pp. 239-242).

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

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

Fang, H., M. Ostendorf, P. Baumann, y J. Pierrehumbert. (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.

Jurafsky, D., y J. H. Martin. (2025). Large language models (Chap. 10). En Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models (3.ª ed.). Manuscrito en línea publicado el 12 de enero de 2025. https://web.stanford.edu/~jurafsky/slp3

García-Sierra, O., A. Fernández-Pampillón, y M. Ortega-Martín. (2024a). Evaluación morfológica de los vocabularios de subpalabras utilizados por los grandes modelos de lenguaje. Revista Española de Lingüística, 54(1), 103-129.

García-Sierra, O., A. Fernández-Pampillón, y M. Ortega-Martín. (2024b). Morphological evaluation of subwords vocabulary used by BETO language model. arXiv preprint arXiv:2410.02283.

Grönroos, S. A., S. Virpioja, P. Smit, y M. Kurimo. (2014, August). Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers (pp. 1177-1185).

Gutierrez-Vasques, X., A. Medina-Urrea, y G. Sierra. (2019). Morphological segmentation for extracting Spanish-Nahuatl bilingual lexicon. Procesamiento del Lenguaje Natural, 63, 41-48.

Hofmann, V., J. Pierrehumbert, y H. Schütze. (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.

Honnibal, M., y I. Montani. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Kudo, T., y J. Richardson. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv Preprint arXiv:1808.06226.

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

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

Mager, M., A. Oncevay, E. Mager, K. Kann, y N. T. Vu. (2022). BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. arXiv preprint arXiv:2203.08954.

Méndez-Cruz, C. F., A. Medina-Urrea, y G. Sierra. (2016). Unsupervised morphological segmentation based on affixality measurements. Pattern Recognition Letters, 84, 127-133.

Monson, C., J. Carbonell, A. Lavie, y L. Levin. (2007). ParaMor: Minimally Supervised Induction of Paradigm Structure and Morphological Analysis. In Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, pages 117–125, Prague, Czech Republic. Association for Computational Linguistics.

Molino de Ideas. (2012). Los verbos en español. Biblioteca Molino de Ideas. ISBN 978-84-937706-1-7. https://www.onoma.es/

Park, K., J. Lee, S. Jang, y D. Jung. (2020). An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks. arXiv Preprint arXiv:2010.02534.

Porter, M. F. (2001). Snowball: A language for stemming algorithms.

Radford, A., K. Narasimhan, T. Salimans, I. Sutskever, y Otros. (2018). Improving language understanding by generative pre-training.

Real Academia Española. (s.f.). Cultura. En Diccionario de la lengua española. Recuperado en 10 de febrero de 2019, de https://dle.rae.es/cultura?m=form

Real Academia Española. (2009). Nueva gramática de la lengua española (Vol. 2). Madrid: Espasa Libros.

Sennrich, R., B. Haddow, y A. Birch. (2015). Neural machine translation of rare words with subword units. arXiv Preprint arXiv:1508.07909.

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

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

Suárez, P. J. O., B. Sagot, y L. Romary. (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., L. Romary, y B. Sagot. (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., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, … I. Polosukhin. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

Virpioja, S., P. Smit, S. A. Grönroos, y M. Kurimo. (2013). Morfessor 2.0: Python implementation and extensions for Morfessor Baseline.

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, … Otros. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv Preprint arXiv:1609.08144.