

Palabras y tokens

Clase 3

Adarsh Kumar

Department of Industrial Mining Engineering and ICT (EMIT),
Manresa School of Engineering (EPSEM),
Polytechnic University of Catalonia, Manresa, Barcelona, Spain

adarsh.kumar@upc.edu

5 de mayo de 2026

Co-funded by
the European Union

[Diapositiva 2: Referencia]

VGP (problema del crecimiento del vocabulario) | Morfema

Referencia

Jurafsky, D., & Martin, J. H. (2026). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models (3rd ed.). Retrieved from https://web.stanford.edu/~jurafsky/slp3/

[Diapositiva 3: El problema del crecimiento del vocabulario (VGP) en NLP]

El problema del crecimiento del vocabulario (VGP) en NLP

Definición:

El problema del crecimiento del vocabulario (Vocabulary Growth Problem, VGP) se produce cuando el número de palabras únicas de un corpus sigue creciendo a medida que se añade más texto. Esto dificulta el modelado del lenguaje y las tareas de NLP.

Por qué ocurre:

- La tokenización basada en palabras trata cada palabra única como un token distinto.
- Las variaciones morfológicas generan muchas formas de una misma palabra.
- Las entidades nombradas y los nombres propios introducen constantemente nuevos tokens.
- Las redes sociales y las erratas introducen palabras raras y nuevas.

[Diapositiva 4: Ejemplo: variaciones morfológicas (inglés)]

Ejemplo: variaciones morfológicas (inglés)

Oración:

I connect networks. She is connecting nodes.

Tokens: I, connect, networks, She, is, connecting, nodes

- “connect” y “connecting” se tratan como tokens distintos.
- Cada variante morfológica aumenta el vocabulario de forma innecesaria.

[Diapositiva 5: El vocabulario que se dispara]

El vocabulario que se dispara

Co-funded by the European Union

Corpus	Tamaño del vocabulario (|V|)	Total de tokens (N)
Shakespeare	31.000	884.000
Brown Corpus	38.000	1 millón
Switchboard	20.000	2,4 millones
COCA	2 millones	440 millones
Google N-grams	más de 13 millones	1 billón (10^{12})

Observación clave

Corpus más grandes → más palabras únicas

¡El tamaño del vocabulario crece sin límite!

[Diapositiva 6: Por qué el crecimiento del vocabulario es un problema]

Por qué el crecimiento del vocabulario es un problema

- **Dispersión de datos (data sparsity):** muchas palabras aparecen muy raramente → es difícil aprender buenas estadísticas.
- **Memoria y cómputo:** vocabulario grande → embeddings grandes → entrenamiento más lento.
- **Palabras OOV (fuera de vocabulario):** las palabras nuevas de los datos de prueba pueden no existir en el entrenamiento → no se pueden manejar las palabras no vistas.

[Diapositiva 7: Ley de Heaps: patrón de crecimiento del vocabulario]

Ley de Heaps: patrón de crecimiento del vocabulario

Ley de Heaps (ley de Herdan)

$|V| = k \times N^{\beta}$

Donde:

- $|V|$ = tamaño del vocabulario
- N = tamaño del corpus en tokens
- $\beta \approx 0,5$ (inglés)
- k = constante del idioma

Interpretación:

- El vocabulario crece como √N

Co-funded by the European Union

- Al duplicar el corpus → √2 ≈ 1,4 veces más palabras
- Observación empírica

Gráfico de la ley de Heaps

|V| frente a N en escala log-log

Pendiente = β ≈ 0,5

[Diapositiva 8: Interpretación de la ley de Heaps]

Interpretación de la ley de Heaps

Idea clave:

- El vocabulario no crece linealmente con el número de palabras.
- El crecimiento es sublineal, aproximadamente proporcional a la raíz cuadrada del tamaño del corpus: |V| ~ √N

Ejemplo:

- Un corpus de 100.000 palabras tiene aproximadamente 20.000 palabras únicas: |V| ≈ 20.000
- Al duplicar el corpus a 200.000 palabras: |V| ≈ 20.000 × √2 ≈ 28.300
- Solo se añaden 8.300 palabras nuevas, no otras 20.000

Observación: este patrón se observa empíricamente en muchos idiomas y dominios.

[Diapositiva 9: Ejemplo de crecimiento del vocabulario]

Ejemplo de crecimiento del vocabulario

Ejemplo con un corpus en inglés:

Tamaño del corpus (N)	**Tamaño del vocabulario (|V|)**
10.000	5.000
100.000	20.000
1.000.000	63.000

Observación:

- El vocabulario crece de forma sublineal.
- Duplicar el corpus no duplica el vocabulario.

Co-funded by the European Union

- Explica por qué los modelos basados en palabras sufren una explosión del vocabulario.

[Diapositiva 10: Explosión del vocabulario en los modelos basados en palabras]

Explosión del vocabulario en los modelos basados en palabras

1. Tokenización basada en palabras

- En los modelos de NLP tradicionales basados en palabras, cada palabra única se trata como un token distinto.
- Ejemplo: connect, connected, connecting, connection → 4 tokens distintos

2. Variaciones morfológicas

- Los idiomas con morfología rica (alemán, finés, árabe, composición en chino) producen muchas formas a partir de una sola raíz.
- Cada forma se trata como una palabra nueva.

3. Entidades nombradas / nombres propios

- Constantemente aparecen en los textos nuevas personas, lugares y productos.
- Ejemplo: Yao Ming, Lionel Messi, Cristiano Ronaldo → cada uno es un token nuevo

4. Redes sociales / erratas / jerga

- Las faltas de ortografía y la escritura informal aumentan el número de tokens raros.
- Ejemplo: soooo happy, happpy, h4ppy → todos son tokens distintos

5. Efecto de la ley de Heaps

- El tamaño del vocabulario crece de forma sublineal con el tamaño del corpus, pero sigue aumentando a medida que se añade más texto.
- Cuanto mayor es el corpus, más palabras únicas aparecen → lleva a una explosión del vocabulario.

[Diapositiva 11: Consecuencias de la explosión del vocabulario]

Consecuencias de la explosión del vocabulario

- **Matrices de embeddings grandes:** se necesita más memoria
- **Dispersión de datos:** muchas palabras raras → es más difícil aprender embeddings significativos
- **Palabras OOV:** las palabras de los datos de prueba pueden no existir en el entrenamiento → el modelo falla con las palabras no vistas

Co-funded by the European Union

[Diapositiva 12: Soluciones a la explosión del vocabulario]

Soluciones a la explosión del vocabulario

- **Tokenización por subpalabras:** Byte Pair Encoding (BPE), WordPiece
    - Ejemplo: 世界总决赛 → 世界 + 总决赛
- **Modelos a nivel de carácter:** vocabulario pequeño y fijo, sin OOV
- **Lematización / stemming:** reducen las variantes a la forma base
    - Ejemplo: connect, connected, connecting → connect

[Diapositiva 13: Ley de Zipf: distribución de la frecuencia de las palabras]

Ley de Zipf: distribución de la frecuencia de las palabras

Ley de Zipf

$f(r) \propto 1/r$, r = rango de la palabra según su frecuencia

Observaciones:

- Pocas palabras aparecen con mucha frecuencia (p. ej., "the", "and")
- La mayoría de las palabras aparecen muy raramente
- Larga cola de hapax legomena (palabras que aparecen una sola vez)

Consecuencias para el NLP:

- El vocabulario nunca está completo
- Siempre hay palabras desconocidas (OOV)
- Se necesitan técnicas como los embeddings de subpalabras, BPE o fastText

Ejemplo del corpus Brown

[Diapositiva 14: El problema de las palabras desconocidas]

El problema de las palabras desconocidas

¡Ningún corpus está completo!

Motivos de las palabras OOV:

1. Surgen constantemente palabras nuevas

Co-funded by the European Union

- "COVID-19" (2020)
- "tweet" (nuevo verbo)
- "selfie" (2013)

2. Vocabulario específico de un dominio

3. Nombres propios

4. Erratas y variaciones

Consecuencias:

1. Los modelos no pueden procesar palabras no vistas

2. Rendimiento degradado

3. Se necesitan estrategias de reserva (fallback)

4. Limitación fundamental de los enfoques basados en palabras

¡Esto motiva los enfoques de subpalabras!

No se pueden enumerar todas las palabras → Aprender a construir palabras a partir de sus partes

---

[Diapositiva 15: ¿Qué es un morfema?]

¿Qué es un morfema?

Definición:

Un morfema es la unidad más pequeña con significado de una lengua.

Idea clave:

- No se puede dividir más sin perder el significado
- Las palabras pueden contener uno o más morfemas

Ejemplo:

cats = cat + s

[Diapositiva 16: Tipos de morfemas]

Tipos de morfemas

1. Morfemas libres

- Pueden aparecer solos como una palabra
- Ejemplos: book, run, happy

2. Morfemas ligados

- No pueden aparecer solos
- Deben unirse a otro morfema
- Ejemplos: -s, -ed, un-, -ness

[Diapositiva 17: Los morfemas: los bloques de construcción]

Los morfemas: los bloques de construcción

Morfema = unidad mínima que aporta significado

Ejemplos:

- fox: 1 morfema
- cats: 2 morfemas
    - cat (animal)
    - -s (plural)
- unhappiness: 3 morfemas
    - un- (negación, «no»)
    - happy (emoción)
    - -ness (estado de)

Tipos:

- **Raíz:** significado central
    - work, glass, believe
- **Afijo:** se une a las raíces
    - Prefijo: un- (undo, «deshacer»)
    - Sufijo: -ed (worked)
    - Infijo: -bl- (fan-bloody-tastic)

[Diapositiva 18: ¿Qué es una sílaba?]

¿Qué es una sílaba?

Definición:

Una sílaba es una unidad de sonido de la lengua hablada.

Co-funded by the European Union

Idea clave:

- Se construye en torno a un sonido vocálico
- Representa un pulso rítmico del habla

Ejemplo:

- cat → 1 sílaba
- table → ta + ble (2 sílabas)

[Diapositiva 19: Estructura de una sílaba]

Estructura de una sílaba

Una sílaba típica tiene tres partes:

- **Ataque (onset)** – consonante(s) iniciales
- **Núcleo (nucleus)** – sonido vocálico (parte central)
- **Coda** – consonante(s) finales

Ejemplo: cat

- Ataque: c
- Núcleo: a
- Coda: t

[Diapositiva 20: Ejemplos de recuento de sílabas]

Ejemplos de recuento de sílabas

Una sílaba

- book
- run
- cat

Dos sílabas

- happy → hap + py
- table → ta + ble
- student → stu + dent

Tres sílabas

- computer → com + pu + ter
- beautiful → beau + ti + ful

[Diapositiva 21: Sílaba frente a morfema]

Sílaba frente a morfema

Distinción importante

Ejemplo 1: table

- Sílabas: ta + ble (2)
- Morfemas: table (1)

Ejemplo 2: cats

- Sílabas: cats (1)
- Morfemas: cat + s (2)

Conclusión:

- Sílabas → unidades de sonido
- Morfemas → unidades de significado

[Diapositiva 22: Importancia de las sílabas en NLP]

Importancia de las sílabas en NLP

Reconocimiento del habla

- Ayuda a modelar los patrones de pronunciación

Síntesis de voz (Text-to-Speech, TTS)

- Mejora el ritmo natural y los patrones de acentuación

Modelado de la pronunciación

- Se usa con el CMU Pronouncing Dictionary

Métricas de legibilidad

- El índice de facilidad de lectura de Flesch (Flesch Reading Ease) utiliza el recuento de sílabas

Análisis de poesía y métrica

- Detección del ritmo y de la acentuación

Co-funded by the European Union

[Diapositiva 23: Por qué importan los morfemas en NLP]

Por qué importan los morfemas en NLP

Stemming

- connected, connecting, connection → connect

Lematización

- running → run

Reducción del vocabulario

- Ayuda a reducir la explosión del vocabulario

Modelos de subpalabras

- Se usan en el NLP moderno (p. ej., BPE, WordPiece)

[Diapositiva 24: Conclusiones clave]

Conclusiones clave

- Una sílaba es una unidad de pronunciación.
- Normalmente contiene un sonido vocálico.
- Es distinta de un morfema.
- Las sílabas son fundamentales en los sistemas de NLP basados en el habla.

Las sílabas modelan el ritmo. Los morfemas modelan el significado.

[Diapositiva 25: Tipología morfológica]

Tipología morfológica

Cómo construyen las palabras los idiomas

Definición:

La tipología morfológica clasifica los idiomas según:

- Cuántos morfemas aparecen en una palabra
- Cómo se expresa la información gramatical
- Con qué claridad se marcan los límites entre morfemas

Pregunta clave:

¿Cuánto significado se concentra en una sola palabra?

Co-funded by the European Union

[Diapositiva 26: Idiomas aislantes]

Idiomas aislantes

Idea central:

- Normalmente un morfema por palabra
- Muy poca flexión
- La gramática se expresa mediante el orden de las palabras

Ejemplo: chino mandarín

他说这是房子

tā shuō zhè shì fángzi

“Dijo que esto es una casa”

Cada palabra, un morfema.

Desafío para el NLP:

- Sin espacios → problema de segmentación de palabras

[Diapositiva 27: Idiomas aislantes (analíticos)]

Idiomas aislantes (analíticos)

Idea central:

- Las palabras suelen constar de un único morfema
- Muy poca flexión o ninguna
- La gramática se expresa mediante el orden de las palabras y las palabras funcionales

Ejemplos de idiomas aislantes:

Idioma	**Región**	**Desafío para el NLP**
Chino mandarín	China	Sin espacios → segmentación de palabras
Cantonés	China/Hong Kong	Sin espacios, varias escrituras
Vietnamita	Vietnam	Marcas de tono, palabras compuestas

Tailandés	Tailandia	Sin espacios, segmentación de palabras
Lao	Laos	Segmentación de palabras
Birmano	Myanmar	Escritura sin espacios
Jemer	Camboya	Estructuras silábicas complejas
Yoruba	Nigeria	Las marcas de tono afectan al significado
Igbo	Nigeria	Marcas de tono, recursos de NLP limitados
Ewe	Ghana/Togo	Significado dependiente del tono
Inglés	En todo el mundo	Mayormente analítico, aunque conserva algo de flexión
Hmong-Mien	Sudeste asiático	Variación de tono y de ortografía

Nota: los idiomas aislantes plantean desafíos de NLP propios, principalmente la segmentación de palabras y el tratamiento de los tonos.

[Diapositiva 28: Tipología morfológica de los idiomas]

Tipología morfológica de los idiomas

Cómo construyen las palabras los idiomas

Aislante

- Pocos morfemas por palabra
- Ejemplo: chino mandarín
- 他说这是房子
- tā shuō zhè shì fángzi
- “Dijo que esto es una casa”

Polisintético

- Muchos morfemas por palabra
- Ejemplo: inuktitut
- qangatasuuqtuq

- “va al mundo de arriba”
- (Una sola palabra = una cláusula completa)

Aglutinante

- Límites claros
- Ejemplo: turco
- ev-ler-im-de
- casa-PL-mi-en
- “en mis casas”

Fusional

- Significados combinados
- Ejemplo: ruso
- (stol-om)
- mesa-SG-INSTR-DECL1
- (caso instrumental)

Co-funded by
the European Union