Words and Tokens

Lecture 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

May 5, 2026

Reference

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/`

Vocabulary Growth Problem (VGP) in NLP

Definition:
The Vocabulary Growth Problem occurs when the number of unique words in a corpus keeps growing as more text is added. This makes language modeling and NLP tasks challenging.

Why It Happens:

- Word-based tokenization treats every unique word as a separate token.
- Morphological variations create many word forms.
- Named entities and proper nouns constantly introduce new tokens.
- Social media and typos introduce rare and new words.

Example: Morphological Variations (English)

Sentence:

I connect networks. She is connecting nodes.

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

- "connect" and "connecting" are treated as separate tokens.
- Each morphological variant increases vocabulary unnecessarily.

The Exploding Vocabulary

Corpus	**Vocabulary Size ($\|V\|$)**	**Total Tokens (N)**
Shakespeare	31,000	884,000
Brown Corpus	38,000	1 million
Switchboard	20,000	2.4 million
COCA	2 million	440 million
Google N-grams	13+ million	1 trillion

Key Observation

Larger corpora → More unique words

Vocabulary size grows without bound!

Why Vocabulary Growth is a Problem

- **Data Sparsity:** Many words occur very rarely → hard to learn good statistics.
- **Memory and Computation:** Large vocabulary → large embeddings → slower training.
- **OOV (Out-Of-Vocabulary) Words:** New words in test data may not exist in training → cannot handle unseen words.

Heaps' Law: Vocabulary Growth Pattern

Heaps' Law (Herdan's Law)

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

Where:

- $|V|$ = Vocabulary size
- N = Corpus size in tokens
- $\beta \approx 0.5$ (English)
- k = Language constant

Interpretation:

- Vocabulary grows with $\sqrt{N}$
- Double corpus $\rightarrow$ $\sqrt{2} \approx 1.4\times$ more words
- Empirical observation

Heaps' Law Plot

$|V|$ vs N on log-log scale

Slope = $\beta \approx 0.5$

Interpretation of Heaps' Law

Key Idea:

- Vocabulary does **not grow linearly** with the number of words.
- Growth is **sub-linear**, roughly proportional to the square root of the corpus size:

$$|V| \sim \sqrt{N}$$

Example:

- A corpus of 100,000 words has approximately 20,000 unique words: $|V| \approx 20,000$
- Doubling the corpus to 200,000 words:

$$|V| \approx 20,000 \times \sqrt{2} \approx 28,300$$

- Only 8,300 new words are added, not another 20,000

Observation: This pattern is observed empirically in many languages and domains.

Example of Vocabulary Growth

English Corpus Example:

Corpus Size (N)	Vocabulary Size (|V|)
10,000	5,000
100,000	20,000
1,000,000	63,000

Observation:

- Vocabulary grows sub-linearly.
- Doubling corpus does **not** double vocabulary.
- Explains why word-based models face vocabulary explosion.

Vocabulary Explosion in Word-Based Models

1. **Word-Based Tokenization**
   - In traditional word-based NLP models, **each unique word is treated as a separate token**.
   - **Example:** connect, connected, connecting, connection $\rightarrow$ 4 separate tokens

2. **Morphological Variations**
   - Languages with rich morphology (German, Finnish, Arabic, Chinese compounding) produce many forms from a single root.
   - Each form is treated as a new word.

3. **Named Entities / Proper Nouns**
   - New people, places, products constantly appear in text.
   - **Example:** Yao Ming, Lionel Messi, Cristiano Ronaldo $\rightarrow$ each is a new token

4. **Social Media / Typos / Slang**
   - Misspellings and informal writing increase the number of rare tokens.
   - **Example:** soooo happy, happpy, h4ppy $\rightarrow$ all different tokens

5. **Heaps' Law Effect**
   - Vocabulary size grows sub-linearly with corpus size, but keeps increasing as more text is added.
   - The larger the corpus, the more unique words appear $\rightarrow$ leads to **exploding vocabulary**.

Consequences of Vocabulary Explosion

- **Large Embedding Matrices:** More memory required
- **Data Sparsity:** Many rare words $\rightarrow$ harder to learn meaningful embeddings
- **OOV Words:** Words in test data may not exist in training $\rightarrow$ model fails on unseen words

Solutions to Vocabulary Explosion

- **Subword Tokenization:** Byte Pair Encoding (BPE), WordPiece
  Example: 世界总决赛 → 世界 + 总决赛
- **Character-Level Models:** Fixed small vocabulary, no OOV
- **Lemmatization / Stemming:** Reduce variants to base form
  Example: connect, connected, connecting → connect

Zipf's Law: Word Frequency Distribution

Zipf's Law

$$f(r) \propto \frac{1}{r}, \quad r = \text{rank of word by frequency}$$

Observations:

- Few words occur **very frequently** (e.g., "the", "and")
- Most words occur **very rarely**
- Long tail of **hapax legomena** (words occurring once)

Consequences for NLP:

- Vocabulary is never complete
- Always unknown words (**OOV**)
- Need techniques like subword embeddings, BPE, or fastText

Example from Brown Corpus

The Unknown Word Problem

No Corpus is Complete!

Reasons for OOV:

1. New words constantly emerge
   - "COVID-19" (2020)
   - "tweet" (new verb)
   - "selfie" (2013)
2. Domain-specific vocabulary
3. Proper names
4. Typos and variations

Consequences:

1. Models can't process unseen words
2. Degraded performance
3. Need fallback strategies
4. Fundamental limitation of word-based approaches

This motivates subword approaches!

Can't list all words → Learn to build words from parts

What is a Morpheme?

Definition:
A morpheme is the **smallest meaningful unit of language**.

Key Idea:

- Cannot be divided further without losing meaning
- Words may contain one or more morphemes

Example:

- cats = cat + s

Types of Morphemes

1. **Free Morphemes**
   - Can stand alone as a word
   - Examples: `book`, `run`, `happy`
2. **Bound Morphemes**
   - Cannot stand alone
   - Must attach to another morpheme
   - Examples: `-s`, `-ed`, `un-`, `-ness`

Morphemes: The Building Blocks

Morpheme = Smallest meaning-bearing unit

Examples:

- **fox**: 1 morpheme
- **cats**: 2 morphemes
  - *cat* (animal)
  - *-s* (plural)
- **unhappiness**: 3 morphemes
  - *un-* (not)
  - *happy* (emotion)
  - *-ness* (state of)

Types:

- **Root**: Core meaning
  - *work*, *glass*, *believe*
- **Affix**: Attached to roots
  - Prefix: *un-* (undo)
  - Suffix: *-ed* (worked)
  - Infix: *-bl-* (fan-bloody-tastic)

What is a Syllable?

Definition:
A syllable is a **unit of sound** in spoken language.

Key Idea:

- Built around a vowel sound
- Represents one rhythmic beat in speech

Example:

- cat $\rightarrow$ 1 syllable
- table $\rightarrow$ ta + ble (2 syllables)

Structure of a Syllable

A typical syllable has three parts:

- **Onset** – Beginning consonant(s)
- **Nucleus** – Vowel sound (core part)
- **Coda** – Ending consonant(s)

Example: cat

- Onset: c
- Nucleus: a
- Coda: t

Examples of Syllable Counts

One Syllable

- book
- run
- cat

Two Syllables

- happy $\rightarrow$ hap + py
- table $\rightarrow$ ta + ble
- student $\rightarrow$ stu + dent

Three Syllables

- computer $\rightarrow$ com + pu + ter
- beautiful $\rightarrow$ beau + ti + ful

Syllable vs Morpheme

Important Distinction

Example 1: table

- Syllables: ta + ble (2)
- Morphemes: table (1)

Example 2: cats

- Syllables: cats (1)
- Morphemes: cat + s (2)

Conclusion:

- Syllables $\rightarrow$ sound units
- Morphemes $\rightarrow$ meaning units

Importance of Syllables in NLP

- **Speech Recognition**
  - Helps model pronunciation patterns
- **Text-to-Speech (TTS)**
  - Improves natural rhythm and stress patterns
- **Pronunciation Modeling**
  - Used with CMU Pronouncing Dictionary
- **Readability Metrics**
  - Flesch Reading Ease uses syllable count
- **Poetry and Meter Analysis**
  - Rhythm and stress detection

Why Morphemes Matter in NLP

- **Stemming**
  - connected, connecting, connection $\rightarrow$ connect
- **Lemmatization**
  - running $\rightarrow$ run
- **Vocabulary Reduction**
  - Helps reduce vocabulary explosion
- **Subword Models**
  - Used in modern NLP (e.g., BPE, WordPiece)

Key Takeaways

- A syllable is a **unit of pronunciation**.
- It usually contains one vowel sound.
- It is different from a morpheme.
- Syllables are crucial in **speech-based NLP systems**.

Syllables model rhythm. Morphemes model meaning.

Morphological Typology

How Languages Build Words

Definition:
Morphological typology classifies languages based on:

- How many morphemes appear in a word
- How grammatical information is expressed
- How clearly morpheme boundaries are marked

Key Question:

How much meaning is packed into a single word?

Isolating Languages

Core Idea:

- Usually one morpheme per word
- Very little inflection
- Grammar expressed using word order

Example: Mandarin Chinese

他说这是房子
tā shuō zhè shì fángzi
"He said this is a house"

Each word one morpheme.

NLP Challenge:

- No spaces → word segmentation problem

Isolating (Analytic) Languages

Core Idea:

- Words usually consist of a single morpheme
- Very little or no inflection
- Grammar expressed through word order and function words

Examples of Isolating Languages:

Language	Region	NLP Challenge
Mandarin Chinese	China	No spaces → word segmentation
Cantonese	China/Hong Kong	No spaces, multiple scripts
Vietnamese	Vietnam	Tone marks, compound words
Thai	Thailand	No spaces, word segmentation
Lao	Laos	Word segmentation
Burmese	Myanmar	Script without spaces
Khmer	Cambodia	Complex syllable structures
Yoruba	Nigeria	Tone marks affect meaning
Igbo	Nigeria	Tone marks, limited NLP resources
Ewe	Ghana/Togo	Tone-dependent meaning
English	Worldwide	Mostly analytic, but some inflection remains
Hmong-Mien	SE Asia	Tone and orthography variation

Note: Isolating languages present unique NLP challenges, mainly **word segmentation** and **tone handling**.

Morphological Typology of Languages

How Languages Build Words

Isolating

- Few morphemes per word
- Example: Mandarin Chinese
- 他说这是房子
  tā shuō zhè shì fángzi
  "He said this is a house"

Polysynthetic

- Many morphemes per word
- Example: Inuktitut
- *qangatasuuqtuq*
  "he goes to the upper world"
  (Single word = whole clause)

Agglutinative

- Clean boundaries
- Example: Turkish
- *ev-ler-im-de*
  house-PL-my-in
  "in my houses"

Fusional

- Combined meanings
- Example: Russian
- (*stol-om*)
  table-SG-INSTR-DECL1
  (Instrumental case)