# Words and Tokens

## Lecture 2

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/`

# Ways to Instruct a Machine for NLP Automation

**Machines can be instructed through multiple input modalities:**

- **Text Input:**
  - Direct typing (chatbots, search engines)
  - Structured text/scripts (JSON, XML, CSV)
- **Voice / Speech:**
  - Spoken commands via microphone
  - Requires Automatic Speech Recognition (ASR)
- **Gestures / Touch:**
  - Touchscreens, swipes, pointing gestures
  - Example: Selecting text to summarize
- **Visual Input (Images / Video):**
  - OCR: text from printed/handwritten sources
  - Sign language recognition using NLP + computer vision
- **Brain-Computer Interfaces (BCI):**
  - Neural signals translated into text/commands
  - Useful for dictation or accessibility
- **Sensor / IoT Input:**
  - Smart devices sending commands to NLP systems
  - Example: "Add milk to shopping list"
- **Haptic/Mechanical Input:**
  - Button presses or mechanical triggers
  - Example: Press button to activate voice assistant
- **Multimodal Input:**
  - Combination of text, speech, images, gestures
  - Example: "Show me the report on this chart" + pointing gesture

**Summary:** NLP systems interpret these inputs to perform tasks like translation, summarization, sentiment analysis, or question answering.

# Examples of AI Agents by Input Modality

## Text Input

- Perplexity – AI answer engine: https://www.perplexity.ai
- ChatGPT – conversational assistant: https://chat.openai.com

## Voice / Speech

- Amazon Alexa – smart speaker assistant: https://www.amazon.com/alexa
- Google Assistant – mobile/smart speaker: https://assistant.google.com

## Gestures / Touch

- Google Assistant on Android (gesture / long-press trigger)
- Apple Siri with touch UI: https://www.apple.com/siri

## Visual Input (Images / Video)

- ChatGPT with vision (GPT-4 class): https://chat.openai.com
- Google Lens – visual search: https://lens.google

## Brain–Computer Interfaces (BCI)

- Research BCI typing systems (AI cursor/typing agent): https://pmc.ncbi.nlm.nih.gov/articles/PMC12597088/

## Sensor / IoT Input

- Home Assistant voice shopping lists: https://www.home-assistant.io
- Alexa Smart Home controls: https://www.amazon.com/alexa-smart-home

## Haptic / Mechanical Input

- Echo device buttons to trigger Alexa: https://www.amazon.com/echo
- Side/power button for Siri or Google Assistant

## Multimodal Input

- Perplexity (text + web + multimodal search): https://www.perplexity.ai
- ChatGPT (GPT-4o, text+image+audio): https://chat.openai.com

# Linguistic Basics in NLP

1. **Morphology – Word Structure**
   - Study of prefixes, suffixes, roots.
   - Example: **unhappiness** = un- (negation) + happy (root) + -ness (noun)
2. **Syntax – Sentence Structure**
   - Rules for arranging words in sentences.
   - Example:
     Correct: "The cat chased the mouse."
     Incorrect: "Chased cat the mouse."
3. **Semantics – Meaning**
   - Meaning of words and sentences.
   - Example:
     "He kicked the bucket."
     Literal: physically kicked a bucket
     Idiomatic meaning: he died
4. **Pragmatics – Contextual Meaning**
   - How context affects interpretation.
   - Example: "Can you pass the salt?"
     Literal: ability question
     Pragmatic: polite request
5. **Discourse – Connected Text**
   - Structure of conversations or paragraphs.
   - Example:
     A: "Did you watch the match yesterday?"
     B: "Yes! It was amazing, especially the last goal."

# Text Preprocessing in NLP

1. **Tokenization – Splitting Text**
   - Breaking text into words, sentences, or subwords.
   - Example: "I love NLP." → ["I", "love", "NLP"]
2. **Normalization – Standardizing Text**
   - Lowercasing, stemming, lemmatization, removing punctuation.
   - Example: "Running, RUNS, run!" → "run run run"
3. **Stopwords – Common Words Removed**
   - Words that carry little meaning for analysis.
   - Example: "I love the NLP course" → "love NLP course"
4. **Stemming – Root Form Reduction**
   - Reduce words to their stem (may be crude).
   - Example: "running" → "run", "connections" → "connect"
5. **Lemmatization – Dictionary Form**
   - Reduce words to dictionary form, considering context.
   - Example: "better" → "good", "running" → "run"
6. **POS Tagging – Part-of-Speech**
   - Label words with their grammatical roles.
   - Example: "I/PRON love/VERB NLP/NOUN"
7. **Named Entity Recognition (NER)**
   - Identify proper nouns, locations, organizations, dates.
   - Example: "Barack Obama was born in Hawaii." → **Barack Obama/Person**, **Hawaii/Location**

# Text Representations in NLP

1. **Bag-of-Words (BoW)**
   - Represent text as a vector of word frequencies.
   - Example: "I love NLP. NLP is fun."
     Vocabulary: [I, love, NLP, is, fun]
     BoW vector: [1, 1, 2, 1, 1]
2. **TF-IDF (Term Frequency-Inverse Document Frequency)**
   - Weight words by importance in a corpus.
   - Common words get lower weight, rare informative words get higher weight.
3. **Word Embeddings**
   - Dense vector representation capturing semantic similarity.
   - Examples: **word2vec, GloVe, fastText**
     "king" - "man" + "woman" $\approx$ "queen"
4. **Contextual Embeddings**
   - Word representations change depending on context.
   - Examples: **ELMo, BERT**
     "He went to the **bank**" (river vs. finance) $\rightarrow$ different embeddings
5. **One-Hot Encoding**
   - Binary vector indicating presence of a word.
   - Example: Vocabulary: [I, love, NLP, is, fun]
     "I love NLP" $\rightarrow$ [1, 1, 1, 0, 0]

# Language Modeling in NLP

1. **n-grams – Word Sequences**
   - Sequence of **n consecutive words** used to predict the next word.
   - Example (bigram, n=2):
     Text: "I love NLP" → Bigrams: ["I love", "love NLP"]
     Text: "I love NLP. NLP is fun." → Trigrams: ["I love NLP", "love NLP NLP", "NLP NLP is", ...]
2. **Smoothing – Handling Zero Probabilities**
   - Prevents assigning zero probability to unseen n-grams.
   - Example: If "deep learning rocks" was unseen, smoothing assigns it a small non-zero probability instead of 0.
3. **Perplexity – Model Evaluation Metric**
   - Measures how well a language model predicts a sample.
   - Lower perplexity → better prediction.
   - Example: Model A perplexity = 50, Model B perplexity = 200 → Model A is better.

# Common NLP Tasks

1. **Text Classification**
   - Assign predefined categories to text.
   - Examples: Spam detection, sentiment analysis (positive/negative reviews)
2. **Machine Translation**
   - Translating text from one language to another.
   - Example: English $\rightarrow$ French: "I love NLP" $\rightarrow$ "J'aime le NLP"
3. **Summarization**
   - Extract key points or generate a shorter version of text.
   - Example: Long article $\rightarrow$ 3–4 sentence summary
4. **Question Answering (QA)**
   - Find answers to questions based on a given text or context.
   - Example: Text: "The Eiffel Tower is in Paris."
     Question: "Where is the Eiffel Tower?" $\rightarrow$ Answer: "Paris"
5. **Text Generation**
   - Generate coherent and contextually relevant text.
   - Example: GPT models generating essays, stories, or code

# Challenges in NLP

1. **Ambiguity – Multiple Meanings**
   - Words can have different meanings depending on context.
   - Example: "bank" $\rightarrow$ river bank vs. financial bank
2. **Out-of-Vocabulary (OOV) Words**
   - Words not seen during model training can cause issues.
   - Example: Model trained on common words may not recognize "cryptocurrency"
3. **Data Sparsity**
   - Rare words or phrases occur infrequently, reducing model accuracy.
   - Example: Rare medical terms in text may be poorly represented
4. **Context Dependence**
   - Meaning of words can change based on surrounding text.
   - Example: "He saw the man with a telescope." Who has the telescope? Ambiguity resolved only with context

# Evaluation Metrics in NLP

1. **Accuracy, Precision, Recall, F1-Score (Classification)**
   - **Accuracy:** Fraction of correct predictions. Example: 90 correct out of 100 $\rightarrow$ Accuracy = 0.9
   - **Precision:** Correct positive predictions / All predicted positives. Example: Spam detection $\rightarrow$ Precision = # correctly identified spam / # predicted spam
   - **Recall:** Correct positive predictions / All actual positives. Example: Spam detection $\rightarrow$ Recall = # correctly identified spam / # actual spam
   - **F1-Score:** Harmonic mean of Precision and Recall. $\text{F1} = 2 \times \frac{\textbf{Precision} \times \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}}$
2. **BLEU / ROUGE (Translation Summarization)**
   - **BLEU:** Measures similarity of generated text to reference translations. Example: Machine Translation output vs. human reference.
   - **ROUGE:** Measures overlap of n-grams, recall, and F1 for summaries. Example: Generated summary vs. reference summary

# Importance of Analyzing Diverse Languages

| Dimension / Language | English | Chinese | Arabic | Spanish | Hindi | German | Finnish |
|---|---|---|---|---|---|---|---|
| Word segmentation challenges | Low | High | Medium | Low | Medium | Low | Medium |
| Script complexity | Latin | Logographic | Abjad | Latin | Abugida | Latin | Latin |
| Morphological richness | Low–medium | Low | High | Medium | High | Medium–high | Very high |
| Resource availability | Very high | High | Medium | High | Medium–high | High | Medium |
| Typological distance from English | – | High | High | Low | Medium | Low | High |
| NLP priority for inclusion/fairness | Well served | High | High | Medium | High | Medium | High |

- **Word segmentation challenges**: How hard it is to split raw text into meaningful tokens (words or subwords). Languages without spaces or with clitics/compounds make this step non-trivial.
- **Script complexity**: How complex the writing system is for recognition and processing (e.g., Latin alphabet vs. logographic characters vs. abjads/abugidas). This affects OCR, tokenization, and character-level models.
- **Morphological richness**: How much grammatical information is encoded in word forms (inflections, agglutination, etc.). Rich morphology increases vocabulary size and sparsity, impacting tagging, parsing, and language modeling.
- **Resource availability (data, tools)**: How many corpora, lexicons, pretrained models, and annotated datasets exist for the language. High-resource languages are easier to support; low-resource ones need special methods.
- **Typological distance from English**: How structurally different a language is from English in word order, morphology, syntax, etc. Larger distance stresses models and reveals whether they truly generalize across language types.
- **NLP priority for inclusion/fairness**: How important it is to invest in NLP for that language to avoid digital inequality. Under-served languages often need prioritization so their speakers are not excluded from AI-enabled services.

# Why Are Words Important in NLP?

- Words are the **basic units of meaning** in natural language.
- They carry **semantic information** (dog vs. cat).
- Word order affects **syntactic meaning**.
- Most NLP models operate on **word-level representations**.

**Example:**

- "Dog bites man" $\neq$ "Man bites dog"

# What is a Word?

**Simple Question, Complex Answer**

"Words" seem intuitive to humans but pose challenges for computers

**Human Perspective:**

- Intuitive units of meaning
- Separated by spaces (in many languages)
- Psychological reality

**Computational Perspective:**

- Ambiguous boundaries
- Language-dependent rules
- Practical processing challenges

# Human Perspective of a Word in NLP

## 1. Words as Intuitive Units of Meaning

- Humans naturally perceive words as meaning-bearing units.
- Even children interpret meaning word-by-word.

**Example 1:**

*The cat is sleeping.*

Humans instantly understand:

- **cat** $\rightarrow$ animal
- **sleeping** $\rightarrow$ action

**Example 2:**

*Fire!*

One word can signal danger, shooting, or excitement.

# Human Perspective of a Word in NLP

2. **Words Separated by Spaces (Many Languages)**

- In English and many languages, words are visually separated by spaces.
- Humans segment sentences effortlessly.

**Example (English):**

*NLP is interesting*

Humans see:

- NLP
- is
- interesting

**Contrast Example (Chinese – No Spaces):**

我喜欢人工智能

*Wǒ xǐhuān réngōng zhìnéng*

"I like Artificial Intelligence"

Humans interpret:

- 我 (Wǒ) – I
- 喜欢 (xǐhuān) – like
- 人工智能 (réngōng zhìnéng) – Artificial Intelligence

**Observation:** Word segmentation is psychologically natural but computationally challenging.

# Human Perspective of a Word in NLP

## 3. Psychological Reality of Words

- Words exist in the human **mental lexicon**.
- They are organized in interconnected networks.

**Example 1: Semantic Association**

- doctor → nurse (fast reaction)
- doctor → banana (slow reaction)

**Example 2: Category Activation**

*Apple, Mango, Banana*

Automatically activates the concept: **"Fruits"**

**Example 3: Slip of Tongue**

*"He baked the car" instead of "braked"*

Shows words are stored as psychological units.

# Computational Perspective of a Word in NLP

1. **Ambiguous Word Boundaries**
   - Computers do not inherently know where words begin or end.
   - Word boundaries are not always computationally obvious.

**Example 1:**

*New York-based startup*

Is it:

- New + York + based + startup?
- New York + based + startup?

**Example 2:**

*I can't do this.*

Should "can't" be treated as:

- can + not?
- a single token?

**Observation:** Word boundaries are often ambiguous from a computational perspective.

# Computational Perspective of a Word in NLP

## 2. Language-Dependent Rules

- The definition of a "word" varies across languages.
- Tokenization rules are language-specific.

**Example 1: German Compounds**

*Donaudampfschifffahrtsgesellschaft*

A single orthographic word containing multiple semantic units.

**Example 2: Chinese (No Spaces)**

我喜欢机器学习

Segmentation must be determined:

- 机器学习 (Machine Learning)
- 机器 + 学习 ?

**Example 3: Hindi Postpositions**

*राम ने खाना खाया ।*

Handling particles such as "ने" requires linguistic knowledge.

# Computational Perspective of a Word in NLP

## 3. Practical Processing Challenges

- Real-world text is noisy and inconsistent.
- Words appear in many surface forms.
- Preprocessing decisions affect downstream tasks.

**Example 1: Social Media Noise**

*soooo happppyyyy!!!*

Normalize to:

- so happy ?

**Example 2: Contractions**

*don't, I'm, they've*

Tokenize as:

- do + n't ?
- I + 'm ?

# Computational Perspective of a Word in NLP

## 3. Practical Processing Challenges

- Real-world text is noisy and inconsistent.
- Words appear in many surface forms.
- Preprocessing decisions affect downstream tasks.

**Example 1: Social Media Noise**

*soooo happppyyyy!!!*

Normalize to:

- so happy ?

**Example 2: Contractions**

*don't, I'm, they've*

Tokenize as:

- do + n't ?
- I + 'm ?

**Example 3: Morphological Variations**

*connect, connected, connecting, connection*

Reduce to common lemma:

- connect ?

**Example 4: Hashtags and Mixed Tokens**

*#MachineLearningRocks*

Segment as:

- Machine + Learning + Rocks ?

# How Many Words Here?

**Example Sentence**

**"They picnicked by the pool, then lay back on the grass and looked at the stars."**

**16 Words**

If we **don't count** punctuation

**18 Units**

If we **count** punctuation marks

- Commas, periods — are they "words"?
- Different tasks may need different answers

# Task-Specific Counting

## Why do the answers change?

The definition of a "word" depends on what the computer needs to do with the text.

**Task A: Sentiment Analysis**
Counts **18 units** (includes punctuation). Symbols like "!" or "..." carry emotional meaning.

**Task B: Machine Translation**
Counts **16 words**, but groups forms like *picnicked* and *picnic* as one "root" word.

**Task C: Search Engines (Google)**
Counts only ∼**9 keywords**. It ignores "stop words" (the, by, on, at) to save database space.

## Conclusion

One size does **not** fit all in Natural Language Processing!

# Spoken Language Challenges

**Real Speech Example**

"I do uh main- mainly business data processing"

## Disfluencies:

- Filled pauses: "uh", "um"
- Fragments: "main-"
- Repetitions: "I-I think"

## Computational Questions:

- Should we count these as words?
- How to normalize them?
- Impact on language models?

# Types vs. Tokens

## Same Sentence Analysis

"They picnicked by the pool, then lay back on the grass and looked at the stars."

**Types:**

- Unique vocabulary items
- Vocabulary size = $|V|$
- **14 types** in example

**Tokens:**

- Running text occurrences
- Total count = $N$
- **16 tokens** in example

## Important Distinction

"the" appears 3 times = **1 type**, **3 tokens**

# The Contraction Problem

I'm

**Orthographic View**

1 word

- Single written unit
- Space boundaries
- Dictionary entry

**Grammatical View**

2 words

- Subject pronoun: I
- Verb contraction: 'm (am)
- Separate syntactic roles

Similar issues: don't, can't, we'll, they've

# Languages Without Word Spaces

## Not all languages use spaces!

**Chinese**
我喜欢自然语言处理
*wǒ xǐhuān zìrán yǔyán chǔlǐ*
"I like NLP"

**Japanese**
自然言語処理が好きです
*shizen gengo shori ga suki desu*
"I like NLP"

**Thai**
**Chan chop gan bpra-muuan pon paa-saa**
*(Thai script omitted for compatibility)*
"I like NLP"

**Fundamental Challenge**

The concept of "word" as space-separated unit doesn't apply universally

# Chinese Word Segmentation Challenge

姚明进入总决赛

*Yáo Míng jìnrù zǒngjuésài*

"Yao Ming reaches the finals"

| | | | | | |
|---|---|---|---|---|---|
| **3 words:** | 姚明 | 进入 | 总决赛 | | |
| | YaoMing | reaches | finals | | |
| **5 words:** | 姚 | 明 | 进入 | 总 | 决赛 |
| | Yao | Ming | reaches | overall | finals |
| **7 words:** | 姚 | 明 | 进 | 入 | 总 | 决 | 赛 |
| | Yao | Ming | enter | enter | overall | decision | game |

**Common Solutions**

- Character-based (simplest)
- Statistical segmentation
- Dictionary-based approaches

# Why Chinese Word Segmentation Matters in NLP

1. **Vocabulary Size Depends on Segmentation**

**Word-Based Segmentation**

姚明 进入 总决赛

- Large vocabulary
- Many compound words:
  - 总决赛
  - 半决赛
  - 世界总决赛
- Sparsity problem

**Character-Based Segmentation**

姚 明 进 入 总 决 赛

- Small vocabulary
- Less sparsity
- Longer sequences

# Sparsity Problem in Word-Based Segmentation

**What is Sparsity?**

- Many words appear very rarely in the corpus.
- Vocabulary size grows rapidly.
- Most words have very low frequency.

**Core Idea**

When vocabulary grows faster than available data, the model cannot learn reliable statistics.

# Example: Chinese Compound Words

**Word-Based Segmentation**

- 总决赛 (Grand Final)
- 半决赛 (Semi-final)
- 世界总决赛 (World Grand Final)
- 全国总决赛 (National Grand Final)

**Frequency Illustration**

| Word | Frequency |
|---|---|
| 总决赛 | 50 |
| 半决赛 | 30 |
| 世界总决赛 | 3 |
| 全国总决赛 | 2 |

Rare words → Poor statistical estimation

# Mathematical Perspective

**Language Model Probability**

$$P(w) = \frac{\text{Count}(w)}{N}$$

- If $\text{Count}(w)$ is very small:
  - Probability estimate is unreliable
  - High variance
- If $\text{Count}(w) = 0$:
  - Out-of-Vocabulary (OOV) problem

**Sparsity Effect**

Many rare words $\rightarrow$ Weak probability estimates

# How Subword Models Reduce Sparsity

**Instead of:**

世界总决赛

Treating as one rare word,

**We Split Into:**

- 世界 + 总决赛
- or characters: 世 + 界 + 总 + 决 + 赛

- Smaller vocabulary
- Units appear more frequently
- Better generalization

# Why Chinese Word Segmentation Matters in NLP

**2. Impact on Downstream Tasks**

姚明进入总决赛 "Yao Ming reaches the finals"

**Key Idea**

Segmentation is not just preprocessing.
It directly affects every NLP task that follows.

# Why Chinese Word Segmentation Matters in NLP

**2. Impact on Downstream Tasks**

**Impact on POS (Part of Speech) Tagging**

**Correct Segmentation:**

姚明 / 进入 / 总决赛

- 姚明 → Proper Noun
- 进入 → Verb
- 总决赛 → Noun

**Incorrect Segmentation:**

姚 / 明 / 进入 / 总 / 决赛

- Broken word boundaries
- Incorrect or inconsistent POS tags

# Common POS Categories

| POS | Meaning | Example |
|---|---|---|
| Noun (NN) | Person/Place/Thing | dog, school |
| Verb (VB) | Action/State | run, eat |
| Adjective (JJ) | Describes noun | big, happy |
| Adverb (RB) | Modifies verb/adj | quickly |
| Pronoun (PRP) | Replaces noun | he, they |
| Preposition (IN) | Relation | in, on |
| Conjunction (CC) | Connects words | and, but |

# What is POS Tagging?

**POS Tagging**

Assigning each word in a sentence its grammatical label.

**Example (English)**

*They play football.*

- They → Pronoun (PRP)
- play → Verb (VB)
- football → Noun (NN)

# POS Example in Chinese

姚明进入总决赛

**Correct Segmentation**

- 姚明 → Proper Noun
- 进入 → Verb
- 总决赛 → Noun

**Important**

If segmentation is wrong, POS tagging will also be incorrect.

# Impact on Parsing

**Correct Segmentation**

- Subject → Verb → Object

  *[姚明] [进入] [总决赛]*

**Incorrect Segmentation**

姚 + 明 + 进入 + 总 + 决赛

- Subject not clearly identified
- Broken dependency structure
- Noisy parse trees

**Important**

If segmentation is wrong, POS tagging will also be incorrect.

# Impact on Machine Translation (MT)

**Original Sentence:**

姚明进入总决赛
*Yao Ming reaches the finals*

**Incorrect Segmentation:**

姚 / 明 / 进入 / 总 / 决赛

**MT Output:**

*"Yao bright enters overall final"*

- Named entity split incorrectly ( 姚明 $\rightarrow$ Yao + bright)
- Compound words misinterpreted ( 总决赛 $\rightarrow$ overall + final)
- Meaning of sentence distorted

# Impact on Named Entity Recognition (NER)

**Named Entity:**

姚明
*Yao Ming*

- Correct segmentation: 姚明 → one entity
- Incorrect segmentation: 姚 | 明 → entity detection fails
- Entity linking and knowledge base lookup break

**Key Insight**

Segmentation errors propagate to all downstream NLP tasks:

- Machine Translation
- Named Entity Recognition
- Parsing, POS tagging, etc.

# Common Solutions to Chinese Segmentation

1. **Character-Based Models**
   - Treat every character as a token
   - Small vocabulary
   - Avoid segmentation errors
   - Used in early neural models

2. **Dictionary-Based Segmentation**
   - Use lexicons to match words
   - Works for known vocabulary
   - Fails for new words or names

3. **Statistical / Neural Segmentation**
   - Learn boundaries from annotated corpora
   - BiLSTM-CRF / Transformer-based models
   - Most modern approach
   - Still not perfect due to ambiguity