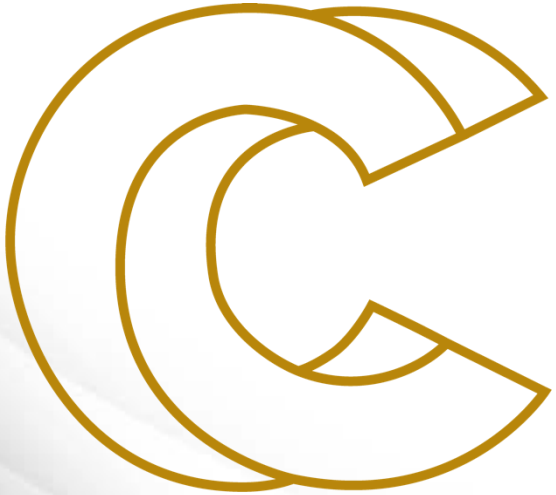
TÜBİTAK

ITU NLP | NATURAL LANGUAGE PROCESSING RESEARCH GROUP



EURO^2

Lect. Tuğba Pamay Arslan

ITUNLP Research Group

AI & Data Engineering, İstanbul Technical University

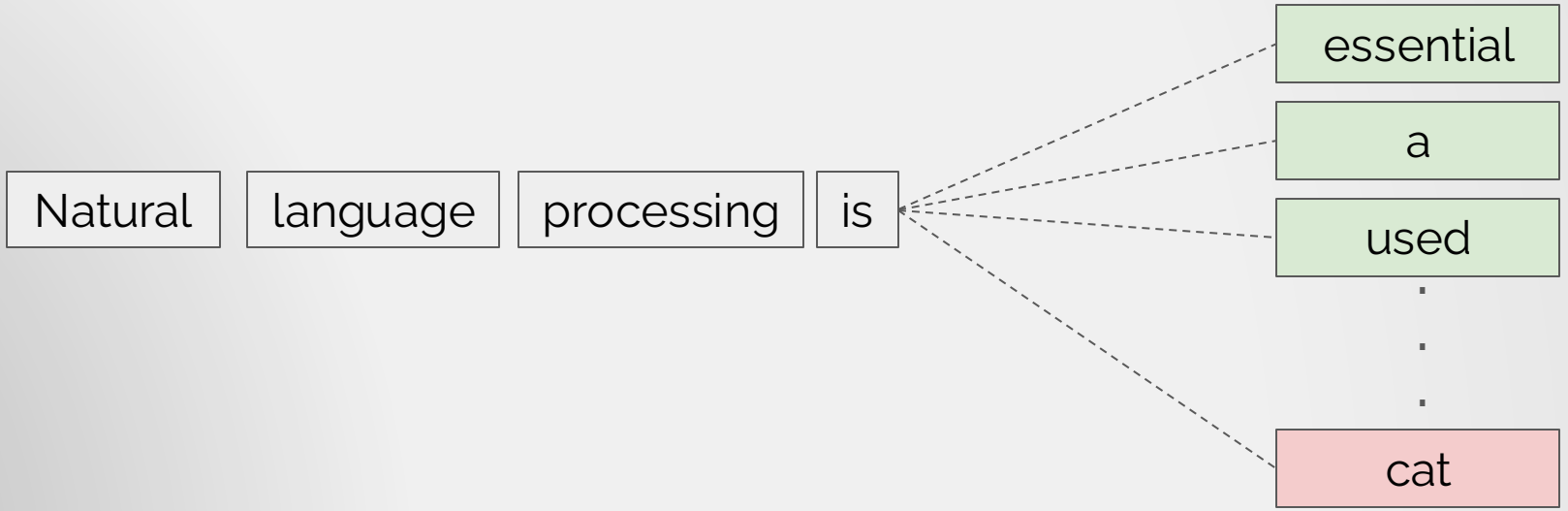
ncc@ulakbim.gov.tr

Introduction to Large Language Models

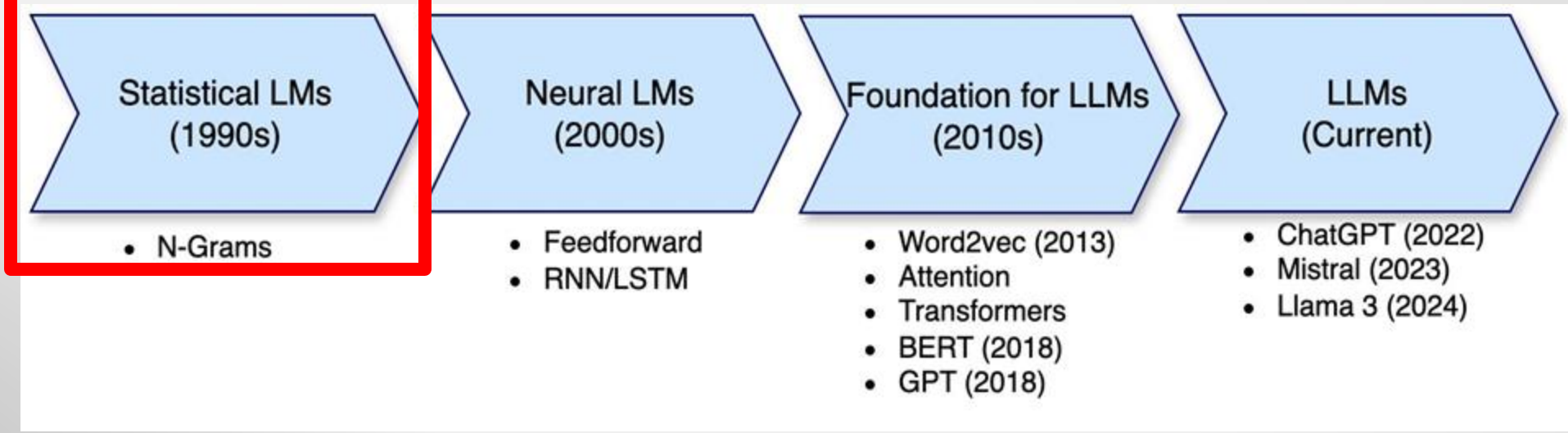
What is a Language Model ?

Models that assign a probability to each possible next word.

Definition [1]



Evolution of Language Models



Statistical LMs

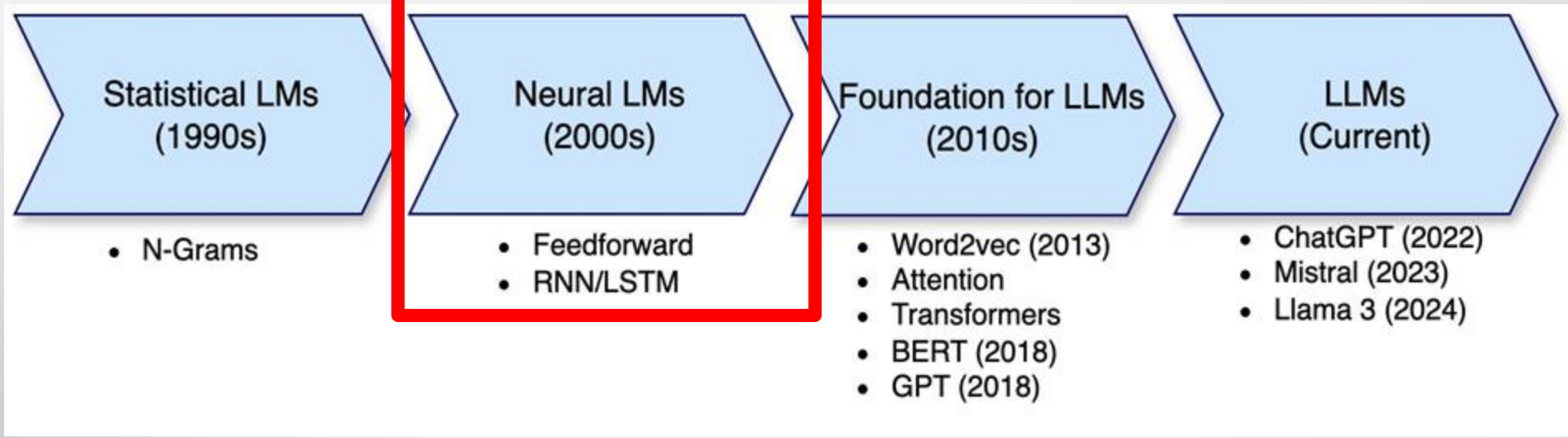
- Modeling languages using simpler statistical methods.
- Often consider a few preceding words to predict the likelihood of the next word.

- Methods:
  - N-Gram
    - To understand the short-term context of a language.
  - Markov Chains
    - Calculate the probability of transitioning from one state (or word) to another.

- Advantages
  - Simple and fast.
  - Work with small datasets to derive probabilities.

- Disadvantages
  - Fail to understand long-term context.
  - Sparsity and overfitting while working with large datasets

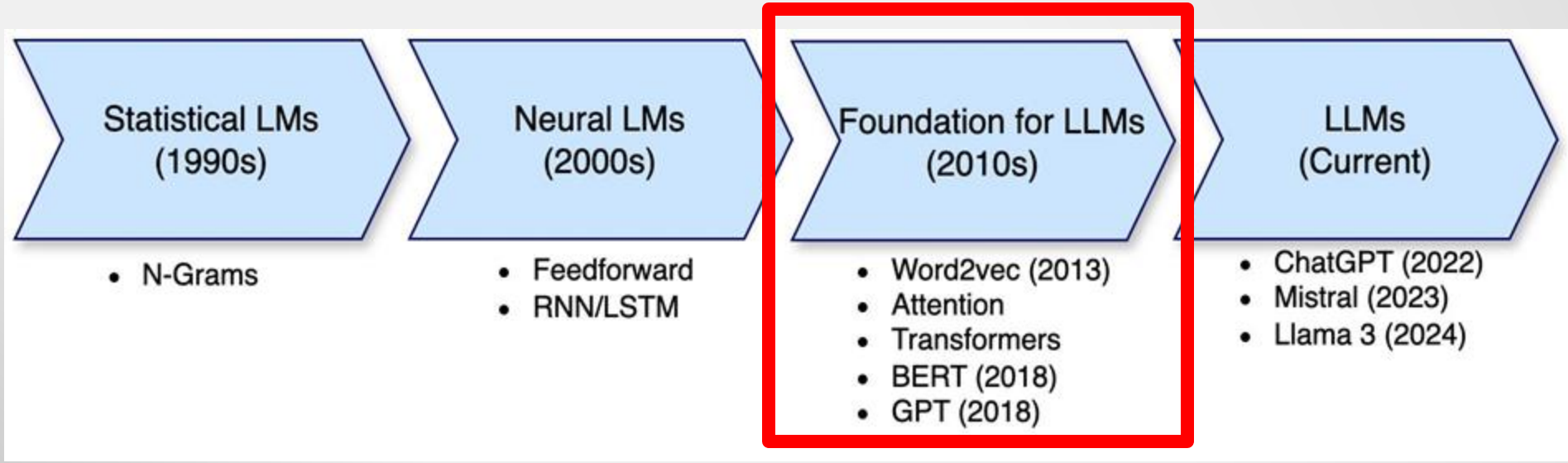
Evolution of Language Models



Neural LMs

- Use deep learning techniques to better understand the context of language.
  - Consider longer contexts within a text.
  - Generate language in a more natural and fluent manner.
- Recurrent Neural Networks:
  - Designed to model sequential relationships in a language.
- Long Short-Term Memory
  - Advanced versions of RNNs (Remove the long-term dependency issue in RNNs)
- Disadvantages
  - Computational Cost
  - Data Dependency

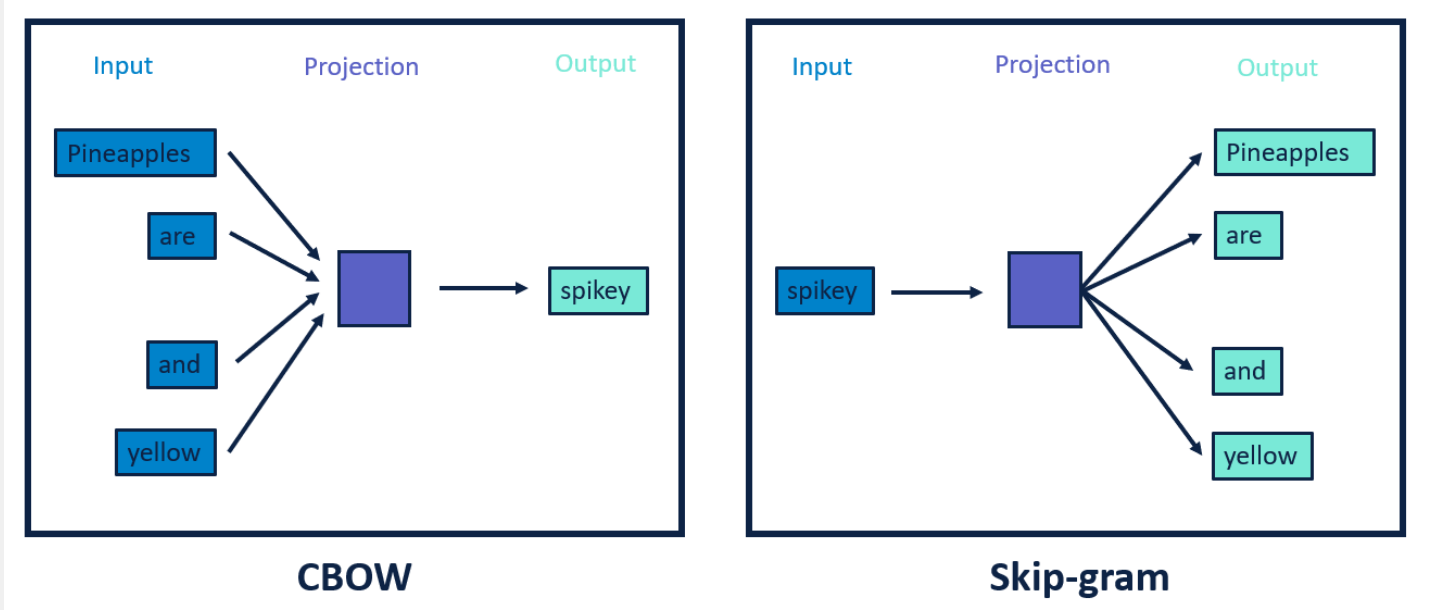
Evolution of Language Models



Foundation for LLMs

- Word Embeddings
  - Words are transformed into mathematical representations through vectors.
  - The first approach for creating semantic word representations.
  - 2 approaches:
    - CBOW

    (Continuous Bag of Words)
    - Skip-Gram

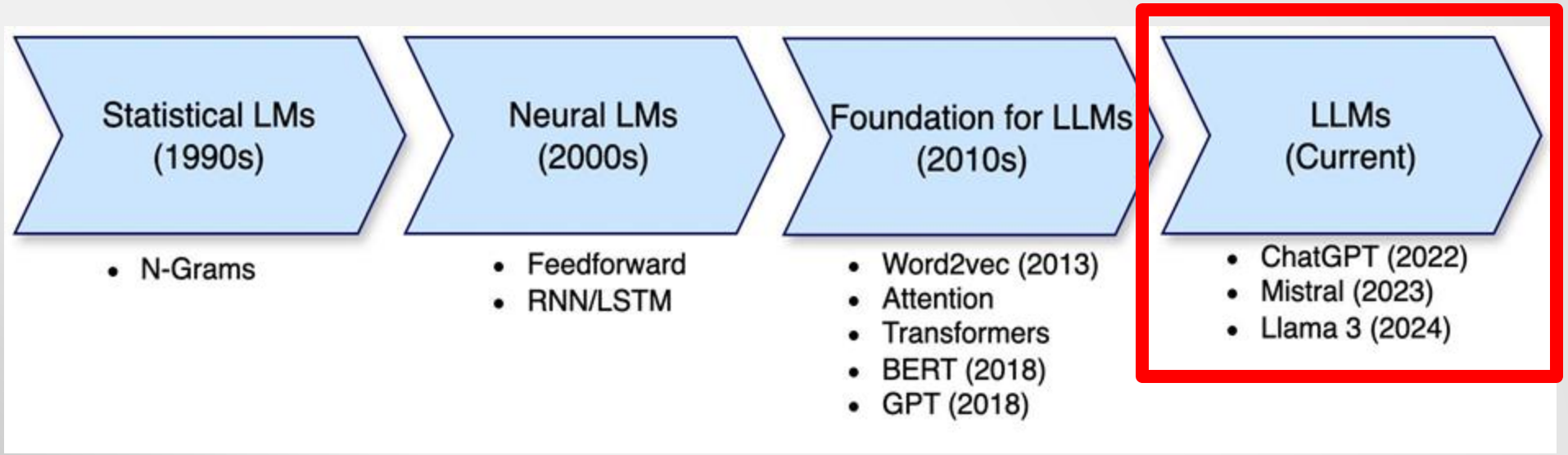


https://community.alteryx.com/t5/Data-Science/Word2vec-for-the-Alteryx-Community/ba-p/305285

Foundation for LLMs

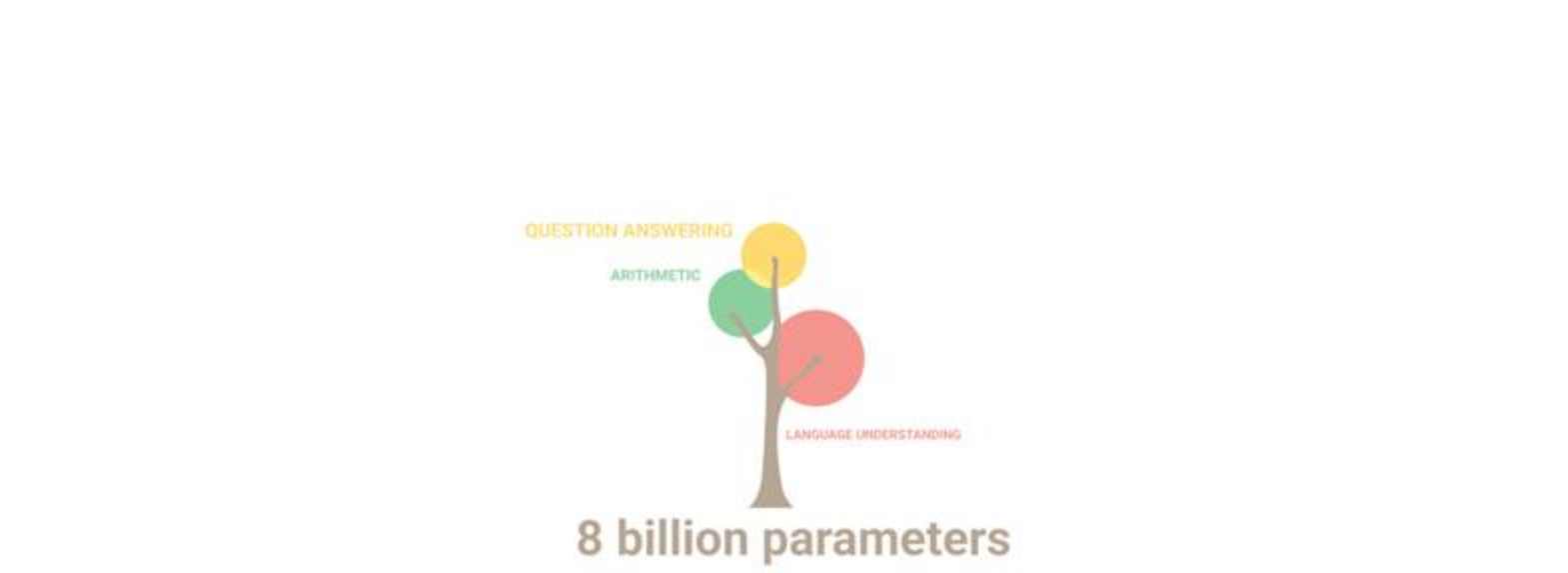
- Attention Mechanism
  - Model can decide which words in a text are more important.
- Transformers
  - Built on attention mechanisms
  - Significant milestone in modern LLMs.
  - Advantages:
    - Parallel Processing
    - Self-Attention

Evolution of Language Models



<Large> Language Models

- ↑ Parameters and Data
- Architecture Advancements
  - with Transformers and Self-Attention
- ↑ Capabilities
  - More powerful & complex applications



As the scale of the model increases, the performance improves across tasks while also unlocking new capabilities.
https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

Linguistic Units in NLP

Word

- Basic element of language that gives a meaning. ,
- It may consist of a single morpheme or a combination of morphemes.

Morpheme

- Smaller meaningful parts of a word.

Example:

Word: arabasında

Morphemes: araba - (s)ı - (n)da

Linguistic Units in NLP

Corpus

- A computer-readable collection of text (or speech).
- It can be a single document or a collection. Its plural form is corpora.

Vocabulary

- A collection of unique words defined for a natural language.
- If the set of words in a vocabulary is V, the number of types (i.e., unique words) is equal to the vocabulary size, |V|.

Token

- Each unit produced after the tokenization process.
- Tokenization is the task of segmenting a piece of text (e.g., sentence, paragraph.) into smaller-pieces (e.g., words, characters, sub-words, morphemes.).

Linguistic Units in NLP

Word Embedding

- Transformed version of words into a fixed-sized numerical vector.