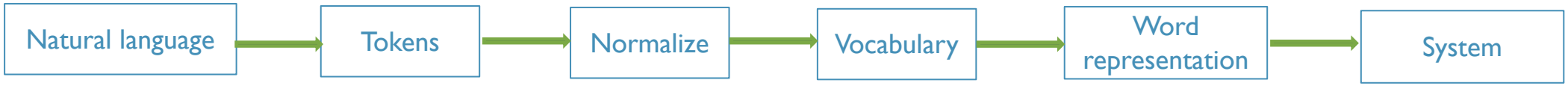
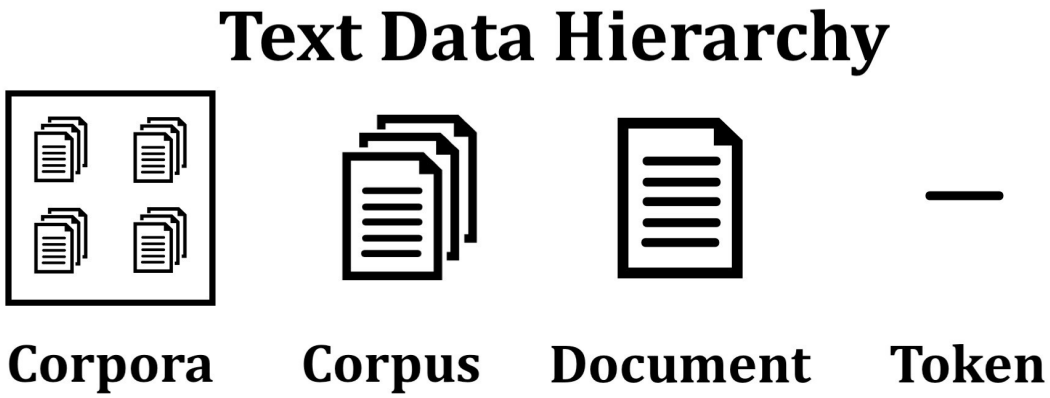
# Lecture 2
# Tokenization and Vocabulary

# OVERVIEW

1. Corpus
2. Word definition
3. Tokenization
4. Vocabulary

# GENERAL WORKFLOW OF AN NLP SYSTEM



# CORPUS IN NLP

An unbiased copy of text collected in a natural communicative setting for a specific purpose.

- monolingual corpus
- multilingual corpus

Quality and quantify of the corpus is greatly affects the quality of the NLP system.

Automatically constructed from existing resources

- Wikipedia, patent and legal text, …..

Manually constructed for domains which don't have enough data.

# CORPUS DESIGN

Should be a representative sample of the language under investigation.

Representativeness: findings from corpus can be generalized.

Balance: cover a wide range of text categories.

Sampling: samples should cover variability in the language or text.

Corpus size: no scientific guidelines should be task dependent.

# EXAMPLES OF NLP CORPORA

- The Brown corpus (US English)
  - a million tokens
- The Lancaster-Oslo-Bergen corpus (British English)
  - a million tokens
- The British National Corpus (BNC)
  - ≈100 million tokens
- The Bank of English corpus
  - 650 million tokens

# HOW TO IDENTIFY TOKENS IN TEXT

***Word (English):*** any sequence of characters between ”*whitespaces*”.

- Text boundaries can be relative to..
  - words
  - phrase
  - Sentences
  - Paragraph

***Whitespaces are not sufficient.*** → ***[Whitespaces, are, not, sufficient, .]***

# WHITESPACES AS WORD BOUNDARIES

## This is good for first approximation but insufficient

- What about compound words:

  *New York-based* ⇔ *[New York, based] or [New, York-based]*

  *ice scream or ice-scream* ⇔ *[ice, scream]*

- Are these one or two words?

## What about **Punctuations?**

- What about contractions like *couldn't, 've, he's*?

  *he's* ⇔ *[he, s] [he, is], [he, has]*

- What about abbreviations like

  *Mr. Smith went to D.C. Ms. Johnson went to Chicago instead*

***Challenge:*** *punctuation marks in abbreviations (Mr, D.C, Ms,…)*

# HOW DO WE IDENTIFY WORDS IN A TEXT?

- Languages without whitespaces (e.g., Chinese and Japanese)

**Chinese:** 我开始写小说 = **我** **开始** **写** 小说

*I* *start(ed)* *writing* *novel(s)*

- Further Japanese have multiple alphabets intermingled
- In Turkish there are single word that represent an entire sentence

*nasilsin* ⇔ *how are you?*

# TOKENIZATION – IDENTIFYING BOUNDARIES

- Tokenization - splitting a phrase, sentence, paragraph, or an entire text document into atomic units of meaning called *tokens*.
  - tokens can be words, phrases, sentence, paragraphs etc…

- Example: using whitespaces, we can form the following tokens from the sentence.

*What time is it? <=> [What, time, is, it?]*

- Main question is how to handle *it*, should it be stored as *it*, *it.* or *it?*

# WHAT IS A WORD?

## Two words form exist

- **surface forms** that occur in text; *books, wants, beginners*.
- **lemmas** that are the uninflected or stem forms of words; *book, beginner, take*.

## Inflection morphology creates different forms of the same word

## Verbs

- Infinitive/present tense: walk, go
- 3rd person singular present tense (s-form): walks, goes
- Simple past: walked, went
- Present participle (ing-form): walking, going

## Nouns

- Inflect for number: *book (singular) vs. books(plural)*
- Inflect for person, number gender; *I saw him; he saw me; you saw her; we saw them; they saw us.*

# DERIVATIONAL MORPHOLOGY

Derivation creates different words from the same lemma:

- Nominalization:
  - V + -ation: *computer* ⇔ *computerization*
  - V+ -er: *sing* ⇔ *singer*
- Negation:
  - un-: *kind* ⇔ *unkind, do* ⇔*undo*
  - mis-: *mistake, misplaced*
- Adjectivization:
  - V+ -able: *doable*
  - N + -al: *national*

# WORD FORMS

Words as atomic symbols

- each word form its own symbol
- add generalization by mapping different forms of a word to the same atomic symbol

Different words forms consist of a ***stem*** + ***affixes*** (***prefixes*** or ***suffixes***)

*dis-grace-ful-ly*

prefix-stem-suffix-suffix

# HOW DO WE REPRESENT THE STRUCTURE OF WORDS?

**1. Normalization:** map all variants of the same word (form) to the same canonical variant

- lowercase everything, normalize spellings, perhaps spell-check

  *US-based, US based, U.S.-based, U.S. based* ⇔ *us-based*

  *labor , labour* ⇔ *labour*

2. Lemmatization:

- reduce inflections or variant forms to base form
- A lemma maybe a word, (lemmatized text is no longer grammatical).

  *am, are, is* ⇔ *be*

  *Car, cars, car's, cars'* ⇔ *car*
- Lemmatization finds correct dictionary headwords
- resulting sentence may not be be grammatically correct.

  *The boy's cars are different colours* ⇔ *the boy car be different colour*

# HOW DO WE REPRESENT THE STRUCTURE OF WORDS?

3. Stemming:

- remove endings that differ among word forms
- no guarantee that the resulting symbol is an actual word)

Reduces words/terms to their stem (crude chopping of affixes)

*Automates, automatic, automation* ⇔ *automat*

- Examples:

Original: *for example, compressed and compression are both accepted as equivalent to compress.*

Stemming: *for exampl compress and compress are both accept as equival to compress*

# HOW DO WE REPRESENT THE STRUCTURE OF A WORDS?

4. Represent structure of each word

- takes into account things like part of speech
- requires a morphological analyser (more on this later)

*"books" => "book N pl" or "book V 3rd sg"*

- The output of such representation is often
  - a lemma ("book") plus
  - morphological information ("N pl" i.e. plural noun)

# SENTENCE SEGMENTATION

*The staff was great. The receptionists were very helpful and answered all our questions. The room was clean and bright, and the room service was always on time. Will be coming back! Will I recommend N.Y.? Hotel? Definitely!*

*Finding sentence boundaries that include fractions like .02 or 4.3 are difficult. It becomes more complicated when the last word is an abbreviation like Dr. or D.C.*

**!, ? are relatively unambiguous**

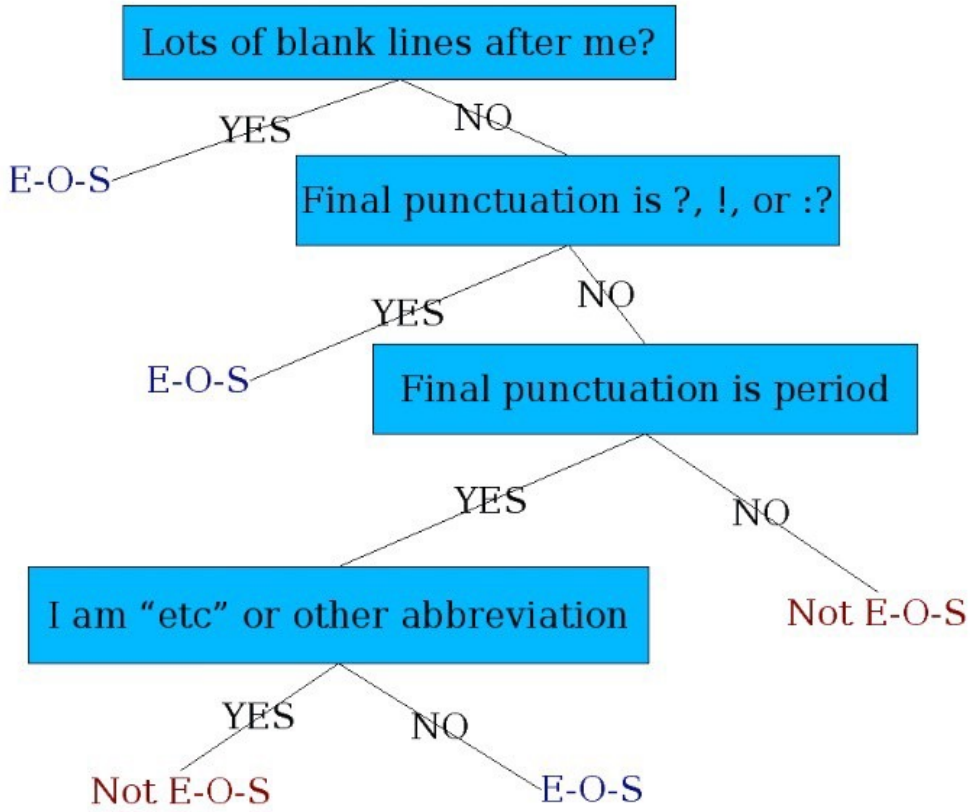
**Period "." is quite ambiguous**

Sentence boundary

Abbreviations like N.Y.

Numbers like .02% or 4.3

# DETERMINING END-OF-SENTENCE – DECISION TREE



# VOCABULARY

- The vocabulary is a holding area for processed text before it is transformed into some representation for the impending NLP task or systems.

- The vocabulary of a language is the set of unique word types in the text corpus.
  - Unique word types after normalization

$$V = \{a, able \ \ldots., zebra\}$$

- The tokens in a document include all occurrences of the word types
- The frequency of a word (type) in a document is the number of occurrences (tokens) of that type.

# VOCABULARY – COUNTING WORDS

- How large is the vocabulary of the English language:

  *Vocabulary size = number of distinct word types (forms)*

- For most corpus of text in the English language
  - close class words are very frequent (the, be, to, of, and, a, in, that,...)
    - Referred to as stop words and often discarded
  - all open class words are very rare

**Word frequency**: the number of occurrences of a word type in a text (or in a collection of texts)

- Biased towards open class words