CMPE 409 Machine Translation

Words, Sentence, Corpora

Murat ORHUN

Istanbul Bilgi University

March 16, 2022

Word

- In linguistics, a word of a spoken language can be defined as the smallest sequence of phonemes that can be uttered in isolation with objective or practical meaning.
- In many languages, words also correspond to sequences of graphemes ("letters") in their standard writing systems that are delimited by spaces wider than the normal inter-letter space, or by other graphical conventions

Words

- Simple Words
- Compound words
- Binomials
- Idoms
- Phrases

Simple Words

- English words
  - book
  - house
  - cat
  - dog
- Turkish words
  - kitap
  - ev
  - kedi
  - köpek

Simple Words

- Chinese words
  - 书
  - 房间
  - 猫
  - 狗
- Japanese words
  - 本
  - 部屋
  - 猫
  - 犬

Compound Words

- bullfrog
- snowball
- mailbox
- grandmother
- railroad
- sometimes
- inside
- upstream

Turkish Compound Words

- birbiri
- affetmek
- karadul (örümcek)
- kadınbudu (köfte)
- balköpüğü
- Ertuğrulgazi
- aşçıbaşı
- Genelkurmay
- karadut

English Binomials

- by and large
- give or take
- pros and cons
- Wine and dine
- leggy-peggy
- rusty-dusty
- tick–tack
- drippity–droppity
- Willy-nilly

Turish Binomials

- koşa koşa
- gürül gürül
- gizli saklı
- gelenek görenek
- irili ufaklı
- ezik büzük
- ıvır zıvır
- Tıkır tıkır
- para mara

Idiom

- A group of words established by usage as having a meaning not deducible from those of the individual words
- For example, if you say someone has “cold feet,” it doesn’t mean their toes are actually cold. Rather, it means they’re nervous about something
- Idioms can’t be deduced merely by studying the words in the phrase

English idioms

- dog-tired

  Exhausted.

  I'm always dog-tired after a day at the amusement park.

- dog eat dog

  Characterized by ruthless behavior and competition...

  1. It's dog eat dog right now at school because all the top students are competing to be valedictorian.
  2. Don't expect this kind of consideration in the real world|it's dog eat dog out there.
  3. Just be careful|it's a dog-eat-dog industry,so everyone will only be looking out for themselves.

Turkish idioms

- Dost acı söyler

  The proverb "Dost acı söyler," which translates into English as "A friend says what hurts," means that a real friend tells the bitter truth and it is used when someone needs to soften the blow of having to deliver or receive unfortunate news from a close buddy.

Phrases

- In linguistic analysis, a phrase is a group of words (or possibly a single word) that functions as a constituent in the syntax of a sentence, a single unit within a grammatical hierarchy
- A phrase is a group of words that form a single unit in a sentence
- It does not have both a subject and a verb.

English Phrases

- Who ate **the last sandwich**
- **All passengers with tickets** can board now.
- The students were **really bored with the film**.
- The window was **behind a large brown sofa**

Turkish Phrases

- İyi akşamlar
- Güle güle
- Çok yaşayın
- Sıra sizde
- Merāk etmeyin

Tokenization

- Tokenization is the process of breaking text down into simpler units.
- For most text, we are concerned with isolating words
- Tokens are split based on a set of delimiters
- These delimiters are frequently whitespace characters (Not always)

Tokenization- English

- We don't have lecture tomorrow.

  `We | do | n't| have | lecture | tomorrow`

- Istanbul is the biggest city of Turkey.

  `Istanbul | is | the | biggest | city | of | Turkey`

Chinese- Chinese

新华社北京2月27日电（记者高敬、胡璐）野生动物伤了人怎么办？如何

民群众生命财产安全？相关部门统计评估野猪等野生动物致害情况，研究部

去年5月至7月，全国人大常委会组织开展了全面禁止野生动物非法交易

执法检查。常委会第二十一次会议听取和审议了该项执法检查报告。本次常

家林业和草原局局长关志鸥作了国务院关于研究处理该项执法检查报告及审

Japanese-Japanese

東京都で337人感染 重症者数68人

菅首相の主張届かず…元官僚嘆き

給与デジタル払いで地銀は三重苦

海外で評価される日本の保健体育

新築マンションに価格載らぬワケ

コロナでキャッシュレス需要がUP PR

GoTo割引きを告知 楽天の言い分

夕食は缶詰2缶…食料配布の現場

正常な生活の再開 米は来春以降?

春頃にコロナ禍の流れ変わる予感

保健所は業務逼迫 現場の悲鳴

globeが復活する可能性はあるか

Word Segmentation (Tokenization)

- Word segmentation is the problem of dividing a string of written language into its component words.
- In English and many other languages using some form of the Latin alphabet, the space is a good approximation of a word divider
- Japanese, Chinese, Thai, Vietnam

https://en.wikipedia.org/wiki/Text_segmentation#Word_segmentation

English contractions

can't	cannot
'cause	because
could've	could have
couldn't	could not
couldn't've	could not have
daren't	dare not / dared not
daresn't	dare not
dasn't	dare not
didn't	did not

Word Segmentation is Complicated

- **Language**: Different languages present unique challenges.
- **Text format**: Text is often stored or presented using different formats.
- **Stopwords**: Commonly used words might not be important for some NLP tasks such as general searches
- **Text expansion:** For acronyms and abbreviations, it is sometimes desirable to expand them so that postprocesses can produce better quality results

Word Segmentation is Complicated (Cont.)

- **Case**: The case of a word (upper or lower) may be significant in some situations. For example, the case of a word can help identify proper nouns.
- **Stemming and lemmatization**: These processes will alter the words to get to their "roots".

Word Segmentation is Complicated (Cont.)

- Idioms
- Phrases
- Binomials
- hyphenated words

Vocabulary

- It is the set of words which constitute a language
- **Types**:Types are the number of distinct words in a corpus; if the set of words in the vocabulary is V , the number of types is the vocabulary size |V|
- The following Brown sentence has 16 tokens and 14 types:

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

English Corpus

Corpus	Tokens = N	Types = $\|V\|$
Shakespeare	884 thousand	31 thousand
Brown corpus	1 million	38 thousand
Switchboard telephone conversations	2.4 million	20 thousand
COCA	440 million	2 million
Google N-grams	1 trillion	13 million

Functional and nonfunctional words

- Functional words
  - In linguistics, function words are words that have little lexical meaning or have ambiguous meaning and express grammatical relationships among other words within a sentence, or specify the attitude or mood of the speaker.
  - eterminers, conjunctions, prepositions, pronouns, auxiliary verbs, modals, qualifiers, and question words
  - Stopwords, close-class
- Content Words
  - Content words are words with specific meanings, such as nouns, adjectives, adverbs, and main verbs (those without helping verbs.)
  - Content words, open-class

Some Functional words in Turkish

- **Yalın bağlaçlar:** ve, ama, ile, eğer, de, hem, yani
- **Bileşik bağlaçlar:** öyleyse, yoksa, nitekim, sanki, oysa, kim bilir
- **Öbekleşmiş bağlaçlar:** ya da, hem de, nerede kaldı ki, değil mi ki
- **Türemiş bağlaçlar:** anlaşılan, gerçekten, kısacası, mesela, örneğin, üstelik

https://tr.wikipedia.org/wiki/Ba

Stem or root

A part of the word to which suffixes and prefixes can be attached

- book
- books
- Mike's
- kitap
- kitaplar
- kitaplarım
- uygarlaştıramadıklarımızdanmışsınızcasına

English Vocabulary

FEBRUARY 1, 2020 BY ADMIN

Number of Words in English

1,062,759.4

Number of Words in the English Language, January 1, 2021, estimate

'Millionth English word' declared

https://languagemonitor.com/number-ofwords/number-of-words-in-english/

POS

- **Parsing (Tagging)** means taking an input (word, sentence) and produce some sort of linguistic structure for it.
- In traditional grammar, a part of speech or part-of-speech (abbreviated as POS or PoS) is a category of words (or, more generally, of lexical items) that have similar grammatical properties.

English POS

- noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection, numeral, article, or determiner.
- The cow jumped over the moon.
- The/DT cow/NN jumped/VBD over/IN the/DT moon./NN

Tags in English and Spanish

English		Spanish		
Input	**Morphological Parse**	**Input**	**Morphological Parse**	**Gloss**
cats	cat +N +PL	pavos	pavo +N +Masc +Pl	'ducks'
cat	cat +N +SG	pavo	pavo +N +Masc +Sg	'duck'
cities	city +N +Pl	bebo	beber +V +PInd +1P +Sg	'I drink'
geese	goose +N +Pl	canto	cantar +V +PInd +1P +Sg	'I sing'
goose	goose +N +Sg	canto	canto +N +Masc +Sg	'song'
goose	goose +V	puse	poner +V +Perf +1P +Sg	'I was ab
gooses	goose +V +1P +Sg	vino	venir +V +Perf +3P +Sg	'he/she c
merging	merge +V +PresPart	vino	vino +N +Masc +Sg	'wine'
caught	catch +V +PastPart	lugar	lugar +N +Masc +Sg	'place'
caught	catch +V +Past			

Tags

- uygarlaştıramadıklarımızdanmışsınızcasına
- uygar +laş +tır +ama +dık +lar +ımız +dan +mış +sınız +casına
- civilized + BEC + CAUS + NABL + PART + PL + P1PL + ABL + PAST +2PL +AsIf
- (behaving) as if you are among those whom we could not civilize

Modified English TreeTagger part-of-speech tagset

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
CDZ	possesive pronoun	one's
DT	determiner	the
EX	existential there	there is
FW	foreign word	d'hoevre
IN	preposition, subordinating conjunction	in, of, like
IN/that	that as subordinator	that
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NNSZ	possessive noun plural	people's, women's
NNZ	possessive noun, singular or mass	year's, world's
NP	proper noun, singular	John
NPS	proper noun, plural	Vikings

Morphological tags in the METU-Sabanc Turkish treebank data.

A1pl	NotState	A1sg	Noun
A2pl	Num	A2sg	Opt
A3pl	Ord	A3sg	P1pl
Abl	P1sg	Able	P2pl
Acc	P2sg	Acquire	P3pl
Adj	P3sg	Adv	Pass
Agt	Past	AfterDoingSo	PastPart
Aor	PCAbl	As	PCAcc
AsIf	PCDat	Become	PCGen
ByDoingSo	PCIns	Card	PCNom
Caus	PersP	Cond	Pnon
Conj	Pos	Cop	Postp
Dat	Pres	Demons	PresPart
DemonsP	Prog1	Desr	Prog2
Det	Pron	Distrib	Prop
Dup	Punc	Equ	Ques
FitFor	QuesP	Fut	Range
FutPart	Real	Gen	Recip
Hastily	Reflex	Imp	ReflexP
InBetween	Rel	Inf	Related

METU-Sabanc Turkish treebank data.

- 5600 sentences
- 60000 words

https://www.aclweb.org/anthology/W12-3620.pdf

Morphology

- In linguistics, morphology is the study of words, how they are formed, and their relationship to other words in the same language.[2][3] It analyzes the structure of words and parts of words, such as stems, root words, prefixes, and suffixes
- A morpheme is the smallest linguistic part of a word that can have a meaning. In other words, it is the smallest meaningful part of a word.

https://en.wikipedia.org/wiki/Morphology_(linguistics)

Morpheme

- Free morpheme:a morpheme which can be joined with other morphemes (such as unbreakable) or on its own (such as break)
- Bound morpheme: a morpheme which can only be used when joined to other morphemes (such as unbreakable)
- Derivational morpheme: a morpheme which can be derived (added) to another morpheme to create a new word (such as adding -ness to happy to form the new word happiness)

https://simple.wikipedia.org/wiki/Morpheme

Morpheme (Cont.)

- Inflectional morpheme: a morpheme which can change a word's tense, number, etc. (such as adding -s to dog to form the plural dogs)
- Allomorphs: different types of the same morpheme (for example, the morpheme ed can have the sound 'id' in the word hunted, the sound 't' in the word fished or the sound 'd' in the word buzzed)

https://simple.wikipedia.org/wiki/Morpheme

online morphological analyzer

- Turkish: http://tools.nlp.itu.edu.tr/MorphAnalyzer
- English: https://cloud.gate.ac.uk/shopfront/displayItem/pos-tagging-and-morphological-analysis
- Multi-Language: https://langrid.org/playground/morphological-analyzer.html

Turkish morphological analyzer

- Ben gözlük kullanıyorum
- ```
  Ben ben+Pron+Pers+A1sg+Pnon+Nom
  Ben ben+Noun+A3sg+Pnon+Nom
  gözlük göz+Noun+A3sg+Pnon+Nom^DB+Adj+Fitfor
  gözlük+Noun+A3sg+Pnon+Nom
  kullanıyorum kullan+Verb+Pos+Prog1+A1sg
  ```

http://tools.nlp.itu.edu.tr/MorphAnalyzer
  ```

Discussion about words

- open class
- closed class
- words in a language (Turkish, Arabic, Chinese etc)
- Functional words
- Compare English with Turkish
- Compare Turkish with other Turkic languages

Recourse

- Jurafsky, D. and J. H. Martin. Speech and language processing*:* An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition, Upper Saddle River, NJ: Prentice-Hall, 2008.
- AshishSingh Bhatia and Richard M. Reese, **Natural Language Processing with Java**: Techniques for Building Machine Learning and Neural Network Models for NLP, 2nd Edition, ISBN 9781788993494, packt, July 2018.