Corpus-based Methods

Corpus

- a huge collection of texts

Text Corpus Structure

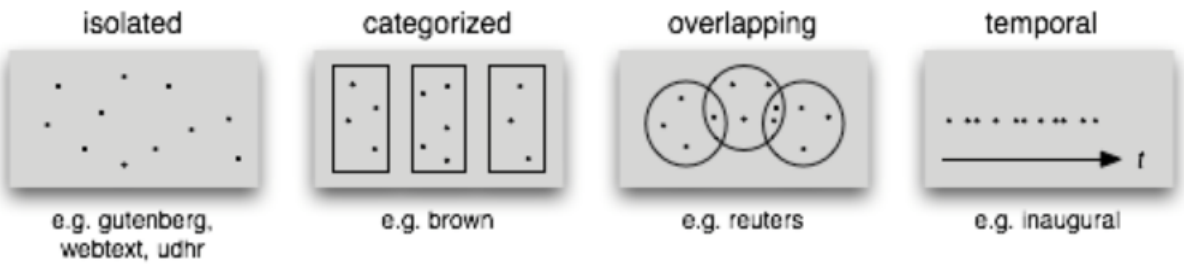


Figure 1: corporastructures

Open Source Corpora

- Gutenberg Corpus - 25,000 free ebooks
- Brown Corpus
- Web and Chat Text
- Reuters Corpus

Kinds of Corpora

1. Monolingual
2. Parallel
3. Multilingual

Preprocessing

junk items

- includes header, titles, tables, figures, equations, diagrams, pictures

misspelled/misrecognized words

- must be corrected, not removed

text cases

- abbreviations
- depends on the language

periods (.)

- used in abbreviation, decimal points, email addresses, domain, etc.

hyphens (-)

- removal of this may change meaning of word

forward slashes (/)

- removal of this can change meaning or make an unrecognizable word

single apostrophes (')

- contracted words
- possessive words

whitespaces

- phrasal verbs: *"work things out"*
- phrases: *"in fact"*, *"on the other hand"*

word segmentation

日本の岸田文雄新首相は経済の成長だけでなく、成果の分配にも目配りする「新し
資本主義」を掲げている。実は、中国も成長重視から分配重視の「共同富裕」を掲
げ、企業が多額の資金を拠出する新たな分配政策を導入した。国家資本主義と評さ
るほど成長重視で走ってきた中国だが、来年秋の共産党大会でトップへの再任を目
す習近平（シー・ジンピン）氏の指導部は格差是正にカジを切った。もっとも、野
図に企業に資金拠出を求...

- language-dependent

Morphology

- study of word forms
- helps in selecting the word form and storing words to a dictionary or lexicon (Stemming and Lemmatization)

Declension

- employ different endings for singular, plural, and for different cases and tenses

	singular	plural
Nom.	ὁ ἀδελφ-ός	οἱ ἀδελφ-οί
Voc.	ὦ ἄδελφ-ε	ὦ ἀδελφ-οί
Acc.	τὸν ἀδελφ-όν	τοὺς ἀδελφ-ούς
Gen.	τοῦ ἀδελφ-οῦ	τῶν ἀδελφ-ῶν
Dat.	τῷ ἀδελφ-ῷ	τοῖς ἀδελφ-οῖς

Sample declension for the Greek phrase "the brother"

Nominative, Genitive, Dative, Accusative

Conjugation

- inflection of verbs

	Singular	Plural
1st Person	was	were
2nd Person	were	were
3rd Person	was	were

sample conjugation of verb *'be'* in the past tense

Affixation

- prefix, infix, suffix

Sentence

- series of words forming a complete thought

sentence boundaries

- punctuations such as period(.), exclamation(!), or question mark(?)

Tokenization

- dividing the collected body of text into units (tokens)

POS Tagging

- parts-of-speech tagging
- marked-up data

Sentence	CLAWS c5	Brown	Penn Treebank	ICE
she	PNP	PPS	PRP	PRON(pers,sing)
was	VBD	BEDZ	VBD	AUX(pass,past)
told	W N	VBN	VBN	V(ditr,edp)
that	CJT	c s	IN	CONJUNC(subord)
the	AT0	AT	DT	ART(def)
journey	NN1	NN	NN	N(com,sing)
might	VM0	MD	MD	AUX(modal,past)
kill	W I	VB	VB	V(montr,infin)
her	PNP	PPO	PRP	PRON(poss,sing)
	PUN			PUNC(per)

- different corpora have different tag sets

Lexical Resources

- a collection of words and/or phrases along with associated information
- lexical entry
  - headword (lemma) + addl info (e.g. POS, sense definition)

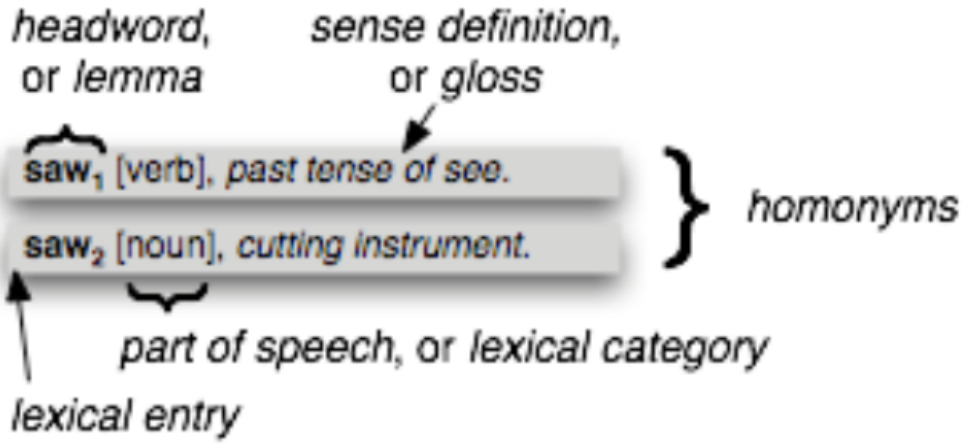


Figure 2: lexical-entry

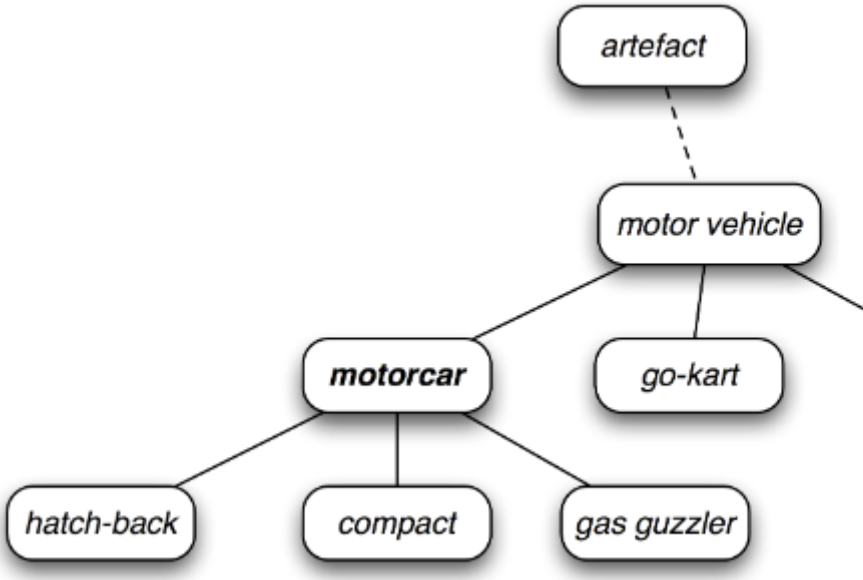
Some lexical resources

- WordList corpora - used to find unusual or mis-spelt words

- Pronouncing dictionary - a table containing the word with its properties such as phonetic codes
- WordNet - semantically-oriented dictionary of English

WordNet

- a network of words that correspond to abstract concepts



- synset: collection of synonymous words

Lexical Relations

- synonyms - words with closely related meanings
- antonyms - words with opposite meanings
- homonymns - one form has two or more unrelated meanings
- polysemy - two or more words with the same form and related meanings
- hyponyms - one form is included in the meaning of another
- holonyms - a word is the whole of which another word is part of
- hypernyms - describes the general term of another
- meronyms - a part of something but is used to refer to the whole
- metonymy - words with close connection from a daily experience
- homophones - words with different spelling and meaning but the same pronunciation
- collocations - words that tend to occur with other words

Pros & Cons

Benefits

- data-driven
- representative
- multi-disciplinary

Challenges

- data collection
- data preprocessing
- representativeness

Group Assignment

- Think about the subject of your NLP project.
- Start looking for datasets.