



DATA SCIENCE FOR DIGITAL HUMANITIES

TEXT REPRESENTATIONS

PROF. DR. GORAN GLAVAŠ



Challenges of Natural Language Processing

Natural Language Processing Challenges

- Why is working with **unstructured, text data** (notoriously) **difficult**?
- Text is written in **natural language**
- **Natural language is a tough nut to crack**
 - It is **complex**
 - It is **ambiguous**
 - It is **vague**
 - It relies on **common-sense knowledge**
- Full understanding of natural language is an **AI-complete** problem
- On top of this: dealing with large amounts of textual data poses serious **technical challenges**

Language ambiguity

- Ambiguity increases the number of possible interpretations (combinatorial explosion of search space)
- **Categorial ambiguity:**
 - *Flying planes can be dangerous*
- **Word sense ambiguity:**
 - *I saw her run to the bank*
 - *The thief was charged by the police and had to pay a fine.*
- **Structural ambiguity:**
 - *I saw a boy on the hill with a telescope*
- **Referential ambiguity:**
 - *Ann and Lisa gave John and Mark some apples because they liked them*
 - *Lisa gave Ann a present and she said thanks*

Ambiguous Named Entities

John Williams

Richard Kaufman goes a long way back with **John Williams**. Trained as a classical violinist, Californian Kaufman started doing session work in the Hollywood studios in the 1970s. One of his movies was Jaws, with **Williams** conducting his score in recording sessions in 1975...

Michael Phelps

Debbie Phelps, the mother of swimming star **Michael Phelps**, who won a record eight gold medals in Beijing, is the author of a new memoir, ...

Michael Phelps is the scientist most often identified as the inventor of PET, a technique that permits the imaging of biological processes in the organ systems of living individuals. **Phelps** has ...



John Williams	author	1922-1994
J. Lloyd Williams	botanist	1854-1945
John Williams	politician	1955-
John J. Williams	US Senator	1904-1988
John Williams	Archbishop	1582-1650
John Williams	composer	1932-
Jonathan Williams	poet	1929-

Michael Phelps	swimmer	1985-
Michael Phelps	biophysicist	1939-

Natural Language Processing and statistics

- Embrace ambiguity and use statistical methods to appropriately model ambiguity *in context*
- That is, we rely on *probabilistic models built from language data*
- Example: **machine translation**
 - P(“maison” → “house”) **high**
 - P(“L’avocat général” → “the general avocado”) **low**

Context is Everything!

avocat

Auxiliaire de justice dont la mission consiste à assister et à représenter en justice une personne qui se présente à lui... ⇒ lawyer



Fruit de l'avocatier, comestible, vert ou violet, en forme de poire, à peau grumeleuse, à chair fondante et savoureuse... ⇒ avocado

L'**avocat**, riche en lipides, apporte très majoritairement des acides gras insaturés bénéfiques pour la santé cardio-vasculaire.



Text Representations

Processing text with computers

- **A text is „one big string”**
 - Needs to be transformed into a more suitable format for text analysis / processing
- **Preprocessing**
 - Set of methods that transform text into a „format” appropriate for text analysis algorithms

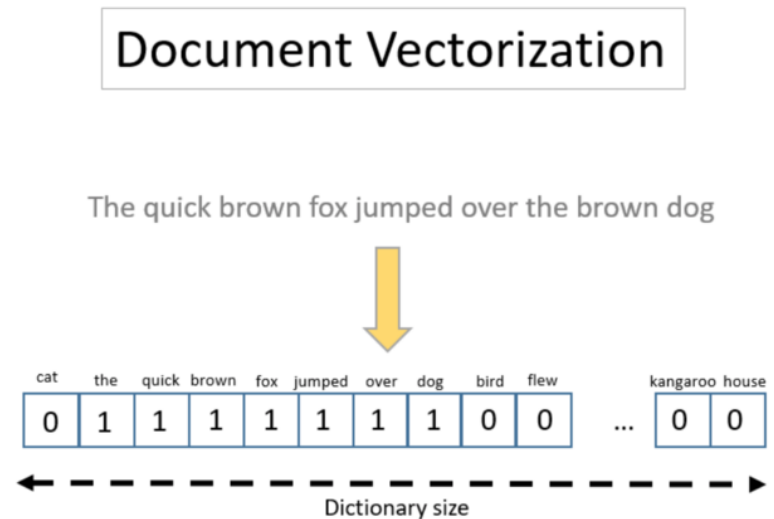


Image from: <https://www.kdnuggets.com/2019/11/text-encoding-review.html>

Preprocessing: introducing „some” structure

- **A text is „one big string”**: convert into something more „manageable”
- **Preprocessing**: may introduce some level of structure
 - **Text representations**:
 - **Unstructured**: Bag-of-words
 - **Weakly-structured**: e.g., bag-of-nouns, bag-of-named-entities
 - **Structured**: a tree or a graph

Text representations: unstructured

1. Unstructured representation

- Text represented as an **unordered set of terms** (the so-called **bag of words**)
- Considerable **oversimplification**
 - We are ignoring the syntax, semantics, and pragmatics of text
 - Is this problematic?

Q: „Revenue of Apple”

D: „*Apple* Pencil 2 'to launch in March 2017'...

Microsoft faces drop in *revenue* in the 3rd quarter...”

- Despite oversimplifying, BoW representations yield good IR performance
- **BoW** was *de facto* the standard representation in traditional TA/TM
 - Until 2013 / 2014 when neural text representations took over

Text representations: unstructured

Document snippet

„Fiat Chrysler and PSA on Monday unveiled the logo of Stellantis - the company resulting from their planned merger. And Fiat Chrysler and PSA described this as a further step towards the finalisation of the deal.”

Corresponding bag (multiset) of words (BoW)

{ (Fiat, 2), (Chrysler, 2), (and, 3), (PSA, 2), (on, 1), (Monday, 1), (unveil, 1), (the, 4), (logo, 1), (of, 2), (Stellantis, 1), (company, 1), (result, 1), (from, 1), (their, 1), (plan, 1), (merger, 1), (describe, 1), (this, 1), (as, 1), (a, 1), (further, 1), (step, 1), (towards, 1), (finalisation, 1), (deal, 1) }

Text representations: weakly structured

2. Weakly-structured representations

- Certain groups of terms given more importance – e.g., nouns or named entities
- Other terms' contribution is either downscaled or completely ignored

Some natural language processing (NLP) tools required

- Part-of-speech (POS) tagger to identify nouns or
- Named entity recognizer (NER) to identify named entities

- Additional preprocessing can be costly

Text representations: weakly structured

Document snippet

„Fiat Chrysler and PSA on Monday unveiled the logo of Stellantis - the company resulting from their planned merger. And Fiat Chrysler and PSA described this as a further step towards the finalisation of the deal.”

Bag of nouns:

{ (Fiat, 2), (Chrysler, 2), (PSA, 2), (Monday, 1), (logo, 1), (Stellantis, 3), (company, 1), (merger, 1), (step, 1), (finalisation, 1), (deal, 1) }

Bag of named entities:

{ (Fiat Chrysler, 2), (PSA, 2), (Stellantis, 1) }

Text representations: structured

3. Structured representations

- Graphs or trees
- Nodes represent some words or concepts
- Edges capture relations between words and concepts (e.g., semantic, temporal, or causal)

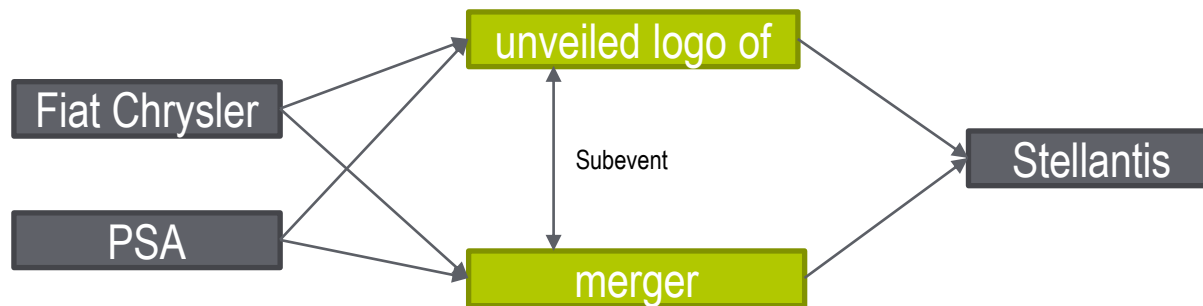
- Sophisticated information extraction (IE) and NLP tools needed to induce structure
- Such IE models may not be accurate enough for the envisioned application
- Very few tasks in which fully structured representations are used

Text representations: weakly structured

Document snippet

„Fiat Chrysler and PSA on Monday unveiled the logo of Stellantis - the company resulting from their planned merger. And Fiat Chrysler and PSA described this as a further step towards the finalisation of the deal.”

Structured representation:





Linguistic Structure

Linguistic structure

In all natural languages:

- words have **parts of speeches**
- **Sentences** have **grammatical structure: parse**

- **Shallow structure: Part-of-speech tagging**
 - Annotate each word in a sentence with a part-of-speech marker.
 - Lowest level of syntactic analysis

- **Deep structure: Syntactic parsing**
 - Induce a **syntactic tree** of a sentence
 - Phrase structure tree or dependency tree (different formalisms)

Shallow linguistic structure: POS-tagging

Fiat Chrysler and PSA on Monday unveiled the logo of Stellantis - the company resulting from their planned merger. And Fiat Chrysler and PSA described this as a further step towards the finalisation of the deal

Course-grained POS tagging

POS tagging


Enter a complete sentence (no single words!) and click at "POS-tag!". The tagging works better when grammar and orthography are correct.

Text:

" Fiat Chrysler and PSA on Monday unveiled the logo of Stellantis - the company resulting from their planned merger . And Fiat Chrysler and PSA described this as a further step towards the finalisation of the deal . "

 Edit text



English 

Adjective

Adverb

Conjunction

Determiner

Noun

Number

Preposition

Pronoun

Verb

Shallow linguistic structure: POS-tagging

Fiat Chrysler and PSA on Monday unveiled the logo of Stellantis - the company resulting from their planned merger. And Fiat Chrysler and PSA described this as a further step towards the finalisation of the deal

Fine-grained POS tagging

Fiat/**NNP** Chrysler/**NNP** and/**CC** PSA/**NNP** on/**IN** Monday/**NNP** unveiled/**VBD**
the/**DT** logo/**NN** of/**IN** Stellantis/**NNP** -/: the/**DT** company/**NN** resulting/**VBG**
from/**IN** their/**PRP\$** planned/**VBN** merger/**NN** ./ . And/**CC** Fiat/**NNP** Chrysler/**NNP**
and/**CC** PSA/**NNP** described/**VBD** this/**DT** as/**IN** a/**DT** further/**JJR** step/**NN**
towards/**IN** the/**DT** finalisation/**NN** of/**IN** the/**DT** deal/**NN**

Shallow linguistic structure: POS-tagging

Coarse-grained or fine-grained POS tagging?

Depends on your target application (type of analysis)

- E.g., do you need to differentiate common nouns from proper nouns?
- E.g., do you need to differentiate tenses of verbs (e.g., past from present)?

" Fiat Chrysler and PSA on Monday unveiled the logo of Stellantis - the company resulting from their planned merger . And Fiat Chrysler and PSA described this as a further step towards the finalisation of the deal . "

Vs.

Fiat/*NNP* Chrysler/*NNP* and/*CC* PSA/*NNP* on/*IN* Monday/*NNP* unveiled/*VBD* the/*DT* logo/*NN* of/*IN* Stellantis/*NNP* -/: the/*DT* company/*NN* resulting/*VBG* from/*IN* their/*PRP*\$ planned/*VBN* merger/*NN* ./ . And/*CC* Fiat/*NNP* Chrysler/*NNP* and/*CC* PSA/*NNP* described/*VBD* this/*DT* as/*IN* a/*DT* further/*JJR* step/*NN* towards/*IN* the/*DT* finalisation/*NN* of/*IN* the/*DT* deal/*NN*

Parts of Speech (EN)

- Noun (person, place or thing)
 - Singular (NN): cat, table
 - Plural (NNS): cats, tables
 - Proper (NNP, NNPS): Natasha, New York
 - Personal pronoun (PRP): I, you, he, she, it
 - Wh-pronoun (WP): who, what
- Verb (actions and processes)
 - Base, infinitive (VB): jump
 - Past tense (VBD): jumped
 - Gerund (VBG): jumping
 - Past participle (VBN): jumped
 - Non 3rd person singular present tense (VBP): jump
 - 3rd person singular present tense: (VBZ): jumps
 - Modal (MD): should, can
 - Verbal To (TO): to (to eat)

Parts of Speech (EN)

- Adjective (modify nouns)
 - Basic (JJ): red, tall
 - Comparative (JJR): redder, taller
 - Superlative (JJS): reddest, tallest
- Adverb (modify verbs)
 - Basic (RB): quickly
 - Comparative (RBR): quicker
 - Superlative (RBS): quickest
- Preposition (IN): on, in, by, to, with
- Determiner:
 - Basic (DT) a, an, the
 - WH-determiner (WDT): which, that
- Coordinating Conjunction (CC): and, but, or,
- Particle (RP): off (took off), up (put up)

Closed vs. Open Class

- **Closed class** categories are composed of a small, fixed set of grammatical function words for a given language.
 - *Pronouns, Prepositions, Modals, Determiners, Particles, Conjunctions*
- **Open class categories** have large number of words and new ones are easily invented.
 - *Nouns, Verbs, Adjectives, Adverbs*

Shallow linguistic structure: POS-tagging

Approaches:

- **Rules**: Human crafted rules based on lexical and other linguistic knowledge.
- **Machine Learning**: Trained on human annotated corpora (e.g., the Penn Treebank).
 - **Traditional statistical models**: Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF)
 - **Neural networks**: Recurrent networks like Long Short Term Memory (LSTMs)
- Machine learning-based POS-tagging is **more effective** (more accurate)

Deep linguistic structure: Parsing

- **Parsing**: inducing a full (hierarchical) grammatical structure of a sentence – this structure is always a **tree**

Two commonly used parsing paradigms:

1. **Constituency parsing**

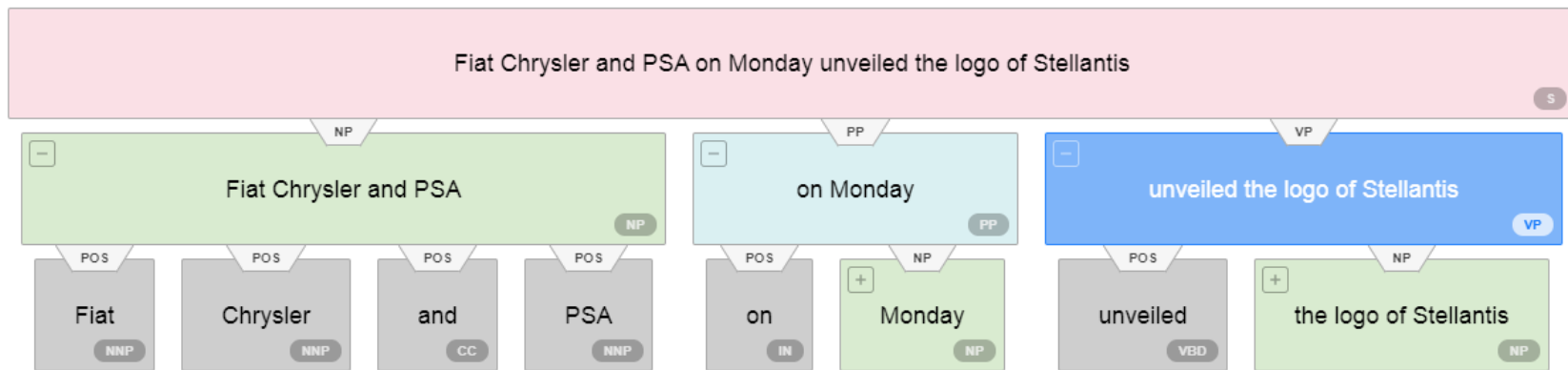
- **Root node** is a sentence node
- **Intermediate nodes** represent syntactic substructures of a sentence (phrases, clauses)
- **Leaves** are words from the sentence

2. **Dependency parsing**

- **All nodes are words from the sentence**
- **Direct dependencies between two words**
 - › **Governing word and dependent word**

Constituency parsing

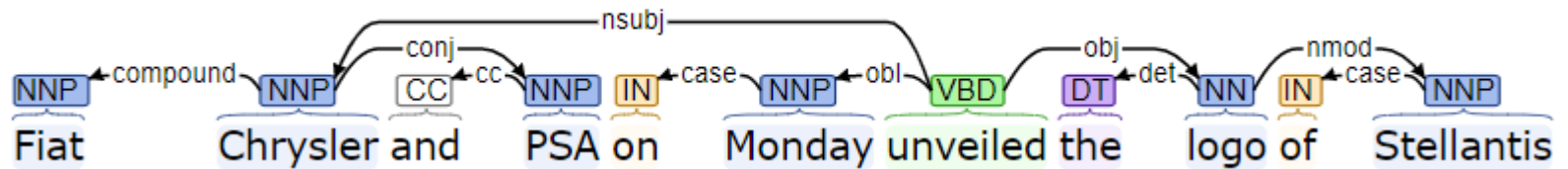
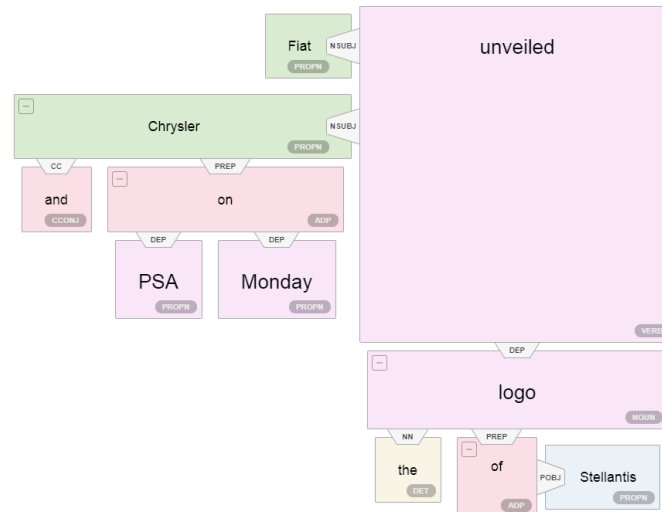
„Fiat Chrysler and PSA on Monday unveiled the logo of Stellantis”



(S (NP (NP (NP Fiat) Chrysler)
and (NP PSA))
on Monday
(VP unveiled
(NP (NP the logo)
(PP of
(NP Stellantis))))))

Dependency parsing

„Fiat Chrysler and PSA on Monday unveiled the logo of Stellantis”



Deep linguistic structure: Parsing

- **Dependency parsing** is more popular than constituency parsing in recent years
 - It does not mean it is better, depends on the use
- **How do we obtain a parser?**
- **Machine Learning**: Trained on human annotated trees for sentences – the so called **treebanks**
- **State of the art parsing performance with neural networks:**
 - **Neural biaffine (dependency) parser**

Why induce linguistic structure?

What do we need POS-tagging and dependency parsing for?

- These additional linguistic annotations can support a wider range of analysis of textual data (i.e., searchers over corpora)
- E.g., **rule-based information extraction**
 - [NNP] merged with [NNP]
 - [NNP] acquired [NNP]

What about „*GM speedily acquired Ford*”?



Why induce linguistic structure?

What do we need POS-tagging and dependency parsing for?

- These additional linguistic annotations can support a wider range of analysis of textual data (i.e., searchers over corpora)
- Additional features for machine learning-based:
 - Information extraction
 - Question answering
- Useful for **candidate ranking** and **error correction** in **speech recognition** and **statistical machine translation**
 - Parser **fails to parse** a sentence → the sentence (transcription in ASR or translation in SMT) is most likely **grammatically incorrect**



Text Preprocessing

Text Preprocessing

In (traditional) TA, we most often use unstructured text representations

Text is represented as **unordered set of terms** (i.e., **bag of words, BoW**)

However, many details about the **exact representation of text** are still undefined

- › How do we „**split**” text into words?
 - › Can this be done in more than one way?
- › Do we consider **all terms**, or do we want to **eliminate some**?
 - › E.g., functional words that have little meaning like *articles* and *prepositions*?
- › How do we treat different forms of the same word?
 - › E.g., should „**house**” be treated the same as „**houses**”? What about „**housing**”?

Text Preprocessing

The preprocessing (i.e., preparing text for the retrieval process) usually involves the following steps:

1. Extracting pure textual content (e.g., from HTML, PDF, Word)
2. Language detection
Optional – if you're dealing with multilingual text collections
3. **Tokenization** (separating text into character sequences)
4. **Morphological normalization** (lemmatization or stemming)
5. **Stopword removal**

Tokens and terms

- **Word** is a delimited string of characters as it appears in the text
- **Term** is a normalized form of the word (accounting for morphology, spelling, etc.)
 - Word and term are in the same equivalence class – in informal speech they are often used interchangeably
- **Token** is an instance of a word or term occurring in a document
 - Tokens are „words” in the general sense
 - But numbers, punctuation, and special characters are also tokens
- **Tokenization** is a process, typically automated, of breaking down the text (one long string) into a sequence of tokens (shorter strings)

Tokenization

Two types of methods for tokenization

- Rule-based (i.e., heuristic)
- Based on **supervised machine learning** models
 - › Learn from manually tokenized texts
- Tokenization might seem simple, but it's **not always unambiguous**
 - E.g., a simple rule: split string on all whitespaces
 - › „Hewlett-Packard declared losses”: „Hewlett-Packard” or „Hewlett” and „Packard”? What about „lower-case”?
 - › What about „Denmark’s mountains”: „Denmark” and „’s”, or „Denmarks”, or „Denmark”?
- What about tokenizing numbers and punctuation?
 - „19/1/2017”, „55 B.C.”, „+49 176 832 40 332”, „IP: 192.168.0.1”
 - Sometimes spaces are not an indication of an end of a token

Issues in tokenization

What about different languages?

- German has numerous compounds
 - › „Lebensversicherungsgesellschaftsangestellter” (life insurance company employee)
 - › Is this a single token or 4 tokens?
- TA systems for German greatly benefit from a compound splitting module
- How about languages that don't segment text using whitespaces at all?
 - › E.g., Chinese
 - › „莎拉波娃现在居住在美国东南部的佛罗里达”

Normalization

- **Normalization** or **standardization** can involve various changes to the token
 - **Error/spelling correction** – repairing the incorrect word
 - **Case-folding** – converting all letters to lower case
 - › „Morgen will ich in MIT” – is this German preposition „mit”?
 - › Sometimes best to lower case all text
 - **How does Google do it?**
 - › „C.A.T.” (information need: Caterpillar Inc.) returns **cat (animal)** as the first result
- **Morphological normalization**
 - Reducing different forms of the „same” word to a common representative form
 - E.g., {*ate, eaten, eats, eat*} -> *eat*



Morphological normalization

- **Inflectional normalization** (or **lemmatization**) reduces all lexico-syntactic forms of the same word to one standard form, **lemma** (headword form)
 - Nouns: singular form in „nominative” case
 - Verbs: infinitive form
 - E.g., „houses” -> „house”, „tried” -> „try”
- **Derivational normalization** reduces all words syntactically derived from some word to the original word (even if the derived word has different meaning)
 - Derivational operators often change the part-of-speech of the word
 - E.g., „destruction” -> „destroy”
- Most TA systems perform inflectional but not derivational normalization

Stemming

- Lemmatization reduces words to dictionary headword entries
 - I.e., the resulting lemma is a string that is a **valid word** in the language
- **Stemming** is the procedure of reducing the word to its grammatical root
 - The result of stemming is not necessarily a valid word of the language
 - › E.g., „recognized” -> „recogniz”, „incredibly” -> „incredibl”
 - Stemming removes suffixes with heuristics
 - › E.g., „automates”, „automatic”, „automation” will all be reduced to „automat”
 - Stemming is „**more aggressive**” than lemmatization and „**less aggressive**” than derivational normalization
- Stemming is more frequently used in **information retrieval** (IR) than in TA systems

Porter's algorithm

- **Most common algorithm for English stemming**
 - **Rule-based algorithm**
 - Grammatical conventions and 5 phases of reduction
 - Phases are executed sequentially, one at a time
 - Each phase consists of a set of concurrent suffix-trimming rules
 - › If multiple rules apply, use the one that removes the longest suffix
- More on Porter's stemmer:
- <http://snowball.tartarus.org/algorithms/porter/stemmer.html>
- Similar algorithms have been developed for other languages as well

Porter's algorithm

- Examples of rules
 - „-ing” -> „”
 - „ly” -> „”
 - „sses” -> „ss”
 - „ational” -> „ate”
 - „tional” -> „tion”
- Rules are sensitive to the measure of „how much of a word” a string is
- Rules consider sequences of consonants and vowels, e.g., $[C][VC]^m[V]$
- Rules also often take into account the length of the remaining „root”
- E.g., „ement” -> „” is valid only if the remaining word has more than one syllable
 - „replacement” -> „replac” but
 - „cement” -> „cement”

Stopword removal

- **Stopwords** are semantically poor terms such as articles, prepositions, conjunctions, pronouns, etc.
 - Removal of stopwords is one of the most common preprocessing steps in text analysis and information retrieval
- **Q:** Why would we want to remove the stopwords?
 - **A:** Because stopwords have very little meaning, they do not reflect the content / meaning / topic of the text
 - **A:** Removing stopwords reduces the size of vocabulary and makes text representation space smaller (good for ML methods)
 - **A:** Including stopwords may lead to **false positives** because of stopword matches between query and documents
- **Stopword lists** for a number of languages:
 - <http://www.ranks.nl/stopwords>