Overview

- Words
- Sentences
- Corpora
Overview

• Words

• Sentences

• Corpora
Words

- It is the basic unit of meaning

  house
  cat
  happy
  run
Syllables

• They are parts of a word
• They produce no mental image

house
hous-es
hap-py
run
**Compound word**

- It is the junction of two or more words
- One or two words?
  - house's = house + 's
  - rooftop = roof + top
  - doesn't = does not

the house's rooftop

the house doesn't have a rooftop
**Idiomatic expressions**

- It's a expression which has a meaning different from the words it contains

- „house of cards“
  - an organization or a plan that is very weak and can easily be destroyed

(http://idioms.thefreedictionary.com/a+house+of+cards
http://www.imdb.com/title/tt1856010/
Tokenization

- It is the task of breaking a text (sentence) into words (tokens)
- Pre-processing steps for many NLP tasks

The house doesn't have a rooftop

The | house | does | n't | has | a | rooftop
Tokenization = Word segmentation

- For writing systems without space separators between words

(http://www.basistech.com/better-multilingual-search/)
Word segmentation

- Contractions also need to be segmented

<table>
<thead>
<tr>
<th>Contraction</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ain’t</td>
<td>am not</td>
</tr>
<tr>
<td>aren’t</td>
<td>are not</td>
</tr>
<tr>
<td>can’t</td>
<td>can not</td>
</tr>
<tr>
<td>could’ve</td>
<td>could have</td>
</tr>
<tr>
<td>couldn’t</td>
<td>could not</td>
</tr>
<tr>
<td>couldn’t’ve</td>
<td>could not have</td>
</tr>
<tr>
<td>didn’t</td>
<td>did not</td>
</tr>
<tr>
<td>doesn’t</td>
<td>does not</td>
</tr>
<tr>
<td>don’t</td>
<td>do not</td>
</tr>
<tr>
<td>hadn’t</td>
<td>had not</td>
</tr>
</tbody>
</table>

Word segmentation

• Should hyphenated words be separated?
  – They lose their original meaning when separated

<table>
<thead>
<tr>
<th>accident-prone</th>
<th>to ice-skate</th>
</tr>
</thead>
<tbody>
<tr>
<td>good-looking</td>
<td>to booby-trap</td>
</tr>
<tr>
<td>sugar-free</td>
<td>to spot-check</td>
</tr>
<tr>
<td>power-driven</td>
<td>to court-martial</td>
</tr>
<tr>
<td>quick-thinking</td>
<td></td>
</tr>
</tbody>
</table>
Word segmentation

• Should compound words be separated?
  – They lose their original meaning when separated

Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz

(“has now been confined to the linguistic history books by authorities in Mecklenburg-Vorpommern.”)

(http://www.theguardian.com/world/2013/jun/03/indfleischetikettierungsberwachungsaufgabenbertragungsgesetz-word-germany)
More very long words...

**Pneumonoultramicroscopicsilicovolcanoconiosis** A lung disease caused by inhalation of silica dust. The longest word listed in an English language

**Floccinaucinihilipilification** Act or habit of estimating as worthless. Longest in Hansard, by Jacob Rees-Mogg

**Taumatawhakatangihangakoauauotamateaturipukakapikimaunga-horonukupokaiwhenuakitanatahu** Longest-named place in the world, given to a New Zealand hill by the Maoris

**Hippopotomonstrosesquipedaliophobia** The fear of long words

(https://www.theguardian.com/world/2013/jun/03/indfleischetikettierungsberwachungsaufgabenbertragungsgesetz-word-germany)
Lowercasing and truecasing

- Lowercasing: converting the word to lowercase
  - House = house = HOUSE

- Truecasing: keeping uppercase letters in names
  - Mr. Fischer ≠ fisher
Translation processing pipeline

- Inverse tasks
  - detokenization
  - decasing

The boy's house is small

the boy's house is small  (tokenization, lowercasing)
das haus des jungens ist klein  (translation)
Das Haus des Jungens ist klein  (decasing)
das haus von dem jungen ist klein
Das Haus vom Jungen ist klein  (detokenization, decasing)
Vocabulary (V)

- It is the set of words which constitute a language
- A text is a sequence of the words from the vocabulary
Size of the vocabulary (language)

- Fluid concept
  - Words are coined every day!

The number of words in the English language is: 1,025,109.8.

This is the estimate by the Global Language Monitor on January 1, 2014.

The English Language passed the Million Word threshold on June 10, 2009 at 10:22 a.m. (GMT).

The Millionth Word was the controversial ‘Web 2.0’.

Currently there is a new word created every 98 minutes or about 14.7 words per day.

Size of the vocabulary (corpus)

- It is the number of words in a corpus

---

Start with which corpus?

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size (words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>American</td>
<td>155 billion</td>
</tr>
<tr>
<td>British</td>
<td>34 billion</td>
</tr>
<tr>
<td>1,000,000 books</td>
<td>89 billion</td>
</tr>
<tr>
<td>Spanish</td>
<td>45 billion</td>
</tr>
</tbody>
</table>
Word tokens vs. word types

- Word tokens: each word in the corpus
- Word types: each unique word in the corpus, no repetition

- Google N-Gram corpus
  - 1,024,908,267,229 word tokens
  - 13,588,391 word types

- Why so many word types?
  - Large English dictionaries have around 500k word types
Functional vs. non-functional words

- **Functional**
  - Words that have none or little semantic meaning
  - Stopwords, close-class
  - Articles, pronouns, conjunctions

- **Non-functional words**
  - Words that have semantic meaning
  - Content words, open-class
  - Verbs, nouns, adjectives

The boy's house is small
Distribution of functional and non-functional words

- Functional vs. non-functional words
Lemma vs. stem

• Lemma
  - Canonical or dictionary form of a word
  - are, is, am, was, were, being → be

• Stem or root (Porter stemmer)
  - A part of the word to which suffixes and prefixes can be attached
  - wants, wanted, wanting, unwanted → want
  - am → am
  - computer → comput

(http://text-processing.com/demo/stem/)
Zipf's law

- The frequency of any word is inversely proportional to its rank in the frequency table

(https://finnaarupnielsen.wordpress.com/2013/10/22/zipf-plot-for-word-counts-in-brown-corpus/)
Zipf's law

• The most frequent word will occur approximately
  - twice as often as the second most frequent word,
  - three times as often as the third most frequent word,
  - ...

• Rank of a word times its frequency is approximately a constant
  - Rank · Freq ≈ c
  - c ≈ 0.1 for English
### Zipf's law

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Count</th>
<th>Freq(%)</th>
<th>Freq x Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The</td>
<td>69970</td>
<td>6.8872</td>
<td>0.06887</td>
</tr>
<tr>
<td>2</td>
<td>of</td>
<td>36410</td>
<td>3.5839</td>
<td>0.07167</td>
</tr>
<tr>
<td>3</td>
<td>and</td>
<td>28854</td>
<td>2.8401</td>
<td>0.08520</td>
</tr>
<tr>
<td>4</td>
<td>to</td>
<td>26154</td>
<td>2.5744</td>
<td>0.10297</td>
</tr>
<tr>
<td>5</td>
<td>a</td>
<td>23363</td>
<td>2.2996</td>
<td>0.11498</td>
</tr>
<tr>
<td>6</td>
<td>in</td>
<td>21345</td>
<td>2.1010</td>
<td>0.12606</td>
</tr>
<tr>
<td>7</td>
<td>that</td>
<td>10594</td>
<td>1.0428</td>
<td>0.07299</td>
</tr>
<tr>
<td>8</td>
<td>is</td>
<td>10102</td>
<td>0.9943</td>
<td>0.07954</td>
</tr>
<tr>
<td>9</td>
<td>was</td>
<td>9815</td>
<td>0.9661</td>
<td>0.08694</td>
</tr>
<tr>
<td>10</td>
<td>He</td>
<td>9542</td>
<td>0.9392</td>
<td>0.09392</td>
</tr>
<tr>
<td>11</td>
<td>for</td>
<td>9489</td>
<td>0.9340</td>
<td>0.10274</td>
</tr>
<tr>
<td>12</td>
<td>it</td>
<td>8760</td>
<td>0.8623</td>
<td>0.10347</td>
</tr>
<tr>
<td>13</td>
<td>with</td>
<td>7290</td>
<td>0.7176</td>
<td>0.09328</td>
</tr>
<tr>
<td>14</td>
<td>as</td>
<td>7251</td>
<td>0.7137</td>
<td>0.09991</td>
</tr>
<tr>
<td>15</td>
<td>his</td>
<td>6996</td>
<td>0.6886</td>
<td>0.10329</td>
</tr>
<tr>
<td>16</td>
<td>on</td>
<td>6742</td>
<td>0.6636</td>
<td>0.10617</td>
</tr>
<tr>
<td>17</td>
<td>be</td>
<td>6376</td>
<td>0.6276</td>
<td>0.10669</td>
</tr>
<tr>
<td>18</td>
<td>at</td>
<td>5377</td>
<td>0.5293</td>
<td>0.09527</td>
</tr>
<tr>
<td>19</td>
<td>by</td>
<td>5307</td>
<td>0.5224</td>
<td>0.09925</td>
</tr>
<tr>
<td>20</td>
<td>l</td>
<td>5180</td>
<td>0.5099</td>
<td>0.10198</td>
</tr>
</tbody>
</table>

\( \text{Freq} \cdot \text{Rank} \approx c \)
Zipf's law

- Not very accurate for very frequent and very infrequent words

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Count</th>
<th>Freq(%)</th>
<th>Freq x Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The</td>
<td>69970</td>
<td>6.8872</td>
<td>0.06887</td>
</tr>
<tr>
<td>2</td>
<td>of</td>
<td>36410</td>
<td>3.5839</td>
<td>0.07167</td>
</tr>
<tr>
<td>3</td>
<td>and</td>
<td>28854</td>
<td>2.8401</td>
<td>0.08520</td>
</tr>
<tr>
<td>4</td>
<td>to</td>
<td>26154</td>
<td>2.5744</td>
<td>0.10297</td>
</tr>
<tr>
<td>5</td>
<td>a</td>
<td>23363</td>
<td>2.2996</td>
<td>0.11498</td>
</tr>
</tbody>
</table>
Zipf's law

- ... and very infrequent words

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Count</th>
<th>Freq(%)</th>
<th>Freq x Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>current</td>
<td>104</td>
<td>0.0102</td>
<td>0.10200</td>
</tr>
<tr>
<td>1001</td>
<td>spent</td>
<td>104</td>
<td>0.0102</td>
<td>0.10210</td>
</tr>
<tr>
<td>1002</td>
<td>eight</td>
<td>104</td>
<td>0.0102</td>
<td>0.10220</td>
</tr>
<tr>
<td>1003</td>
<td>covered</td>
<td>104</td>
<td>0.0102</td>
<td>0.10230</td>
</tr>
<tr>
<td>1004</td>
<td>Negro</td>
<td>104</td>
<td>0.0102</td>
<td>0.10240</td>
</tr>
<tr>
<td>1005</td>
<td>role</td>
<td>104</td>
<td>0.0102</td>
<td>0.10251</td>
</tr>
<tr>
<td>1006</td>
<td>played</td>
<td>104</td>
<td>0.0102</td>
<td>0.10261</td>
</tr>
<tr>
<td>1007</td>
<td>I'd</td>
<td>104</td>
<td>0.0102</td>
<td>0.10271</td>
</tr>
<tr>
<td>1008</td>
<td>date</td>
<td>103</td>
<td>0.0101</td>
<td>0.10180</td>
</tr>
<tr>
<td>1009</td>
<td>council</td>
<td>103</td>
<td>0.0101</td>
<td>0.10190</td>
</tr>
<tr>
<td>1010</td>
<td>race</td>
<td>103</td>
<td>0.0101</td>
<td>0.10201</td>
</tr>
</tbody>
</table>
Challenges for MT

• Functional (closed-class)
  - limited number of words, do not grow usually

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auxiliary</td>
<td>can, should, have</td>
</tr>
<tr>
<td>Article/Determiner</td>
<td>the, some</td>
</tr>
<tr>
<td>Conjunction</td>
<td>and, or</td>
</tr>
<tr>
<td>Pronoun</td>
<td>he, my</td>
</tr>
<tr>
<td>Preposition</td>
<td>to, in</td>
</tr>
<tr>
<td>Particle</td>
<td>off, up</td>
</tr>
<tr>
<td>Interjection</td>
<td>Ow, Eh</td>
</tr>
</tbody>
</table>
Challenges for MT

- Functional (closed-class)
  - Languages differ greatly on them
    - „the“ vs. „der“, „die“, „das“, „dem“, „den“, „des“

The house of the boy is small

Das Haus des Jungen ist klein

Das Haus vom (von dem) Jungen ist klein
Challenges for MT

• Functional (closed-class)
  – Languages differ greatly on them
  • No equivalent to „the“ in Chinese and Japanese
    – Choose among „the“, „a“ or nothing

The house of the boy is small
The house of a boy is small
A house of the boy is small

Boy of the house is small
少年の家は小さいです
少年の家は小さいです

(http://translate.google.de)
Challenges for MT

- Functional (closed-class)
  - Languages differ greatly on them
    - „wa“ particle in Japanese
      - Next to which noun phrase to place it?

The boy's house is small

少年の家は小さいです

Boy of the house is small

少年の家は小さいです

(http://translate.google.de)
Challenges for MT

- Content (open-class) words
  - Unlimited number of words
  - New ones are coined every day

<table>
<thead>
<tr>
<th>Noun</th>
<th>book/books, nature, Germany, Sony</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>eat, wrote</td>
</tr>
<tr>
<td>Adjective</td>
<td>new, newer, newest</td>
</tr>
<tr>
<td>Adverb</td>
<td>well, urgently</td>
</tr>
</tbody>
</table>
Part-of-speech tags

- They are classes of words which have similar grammatical properties
  - Verbs, adjectives, pronouns, etc.
## Penn TreeBank Tagset

<table>
<thead>
<tr>
<th>POS Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>coordinating conjunction</td>
<td>and</td>
</tr>
<tr>
<td>CD</td>
<td>cardinal number</td>
<td>1, third</td>
</tr>
<tr>
<td>DT</td>
<td>determiner</td>
<td>the</td>
</tr>
<tr>
<td>EX</td>
<td>existential there</td>
<td>there is</td>
</tr>
<tr>
<td>FW</td>
<td>foreign word</td>
<td>d’hoere</td>
</tr>
<tr>
<td>IN</td>
<td>preposition/subordinating conjunction</td>
<td>in, of, like</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
<td>green</td>
</tr>
<tr>
<td>JJR</td>
<td>adjective, comparative</td>
<td>greener</td>
</tr>
<tr>
<td>JJ$</td>
<td>adjective, superlative</td>
<td>greenest</td>
</tr>
<tr>
<td>LS</td>
<td>list marker</td>
<td>i)</td>
</tr>
<tr>
<td>MD</td>
<td>model</td>
<td>could, will</td>
</tr>
<tr>
<td>NN</td>
<td>noun, singular or mass</td>
<td>table</td>
</tr>
<tr>
<td>NNS</td>
<td>noun plural</td>
<td>tables</td>
</tr>
<tr>
<td>NNP</td>
<td>proper noun, singular</td>
<td>John</td>
</tr>
<tr>
<td>NNP$</td>
<td>proper noun, plural</td>
<td>Vikings</td>
</tr>
<tr>
<td>PDT</td>
<td>predeterminer</td>
<td>both the boys</td>
</tr>
<tr>
<td>POS</td>
<td>possessive ending</td>
<td>friend’s</td>
</tr>
<tr>
<td>FRP</td>
<td>personal pronoun</td>
<td>I, he, it</td>
</tr>
<tr>
<td>FRP$</td>
<td>possessive pronoun</td>
<td>my, his</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
<td>however, usually, naturally, here, good</td>
</tr>
<tr>
<td>RBR</td>
<td>adverb, comparative</td>
<td>better</td>
</tr>
<tr>
<td>RBS</td>
<td>adverb, superlative</td>
<td>best</td>
</tr>
<tr>
<td>RP</td>
<td>particle</td>
<td>give up</td>
</tr>
<tr>
<td>TO</td>
<td>to</td>
<td>to go, to him</td>
</tr>
<tr>
<td>UH</td>
<td>interjection</td>
<td>uhuhuhuh</td>
</tr>
<tr>
<td>VBD</td>
<td>verb, base form</td>
<td>take</td>
</tr>
<tr>
<td>VBG</td>
<td>verb, gerund/present participle</td>
<td>taking</td>
</tr>
<tr>
<td>VBN</td>
<td>verb, past participle</td>
<td>taken</td>
</tr>
<tr>
<td>VBP</td>
<td>verb, sing. present, non-3d</td>
<td>take</td>
</tr>
<tr>
<td>VBZ</td>
<td>verb, 3rd person sing. present</td>
<td>takes</td>
</tr>
<tr>
<td>WDT</td>
<td>wh-determiner</td>
<td>which</td>
</tr>
<tr>
<td>WP</td>
<td>wh-pronoun</td>
<td>who, what</td>
</tr>
<tr>
<td>WP$</td>
<td>possessive wh-pronoun</td>
<td>whose</td>
</tr>
<tr>
<td>WRB</td>
<td>wh-adverb</td>
<td>where, when</td>
</tr>
</tbody>
</table>

[http://www.americannationalcorpus.org/OANC/penn.html](http://www.americannationalcorpus.org/OANC/penn.html)
Part-of-speech in MT
Part-of-speech in MT

„I like ice-cream“

„Ich mag Eis“

„Ich wie Eis“
Morphology

- It is the analysis and identification of the morphemes

- Morpheme is the smallest grammatical unit
  - Free morphemes: „dog“ in „doghouse“
  - Bound morphemes (affixes):
    - Suffixes: „exactly“
    - Prefixes: „unhappiness“
Morphology

- **Derivational morpheme**
  - Changes the meaning or the functional role (POS tag)
    - „kind“ vs. „unkind“ (opposite meanings)
    - „happy“ vs. „happiness“ (adjective to noun)

- **Inflectional morphemes**
  - Do not affect meaning nor the part-of-speech tag
    - „house“ vs. „houses“
    - „want“ vs. „wanted“
Morphology – aglutinative

Avrupa- -laş- -tir- -ama- -dik- -lar- -imiz- -dan
Europe -an become -ize NEG whom those we one. of

"Are you one of those whom we could not Europeanize?"

(http://allthingslinguistic.com/post/50939757945/morphological-typology-illustrations-from)
Morphology – polysynthetic

(If only you had not been able to make him take it all at from under me again for them.)

(http://allthingslinguistic.com/post/50939757945/morphological-typology-illustrations-from)
Ubykh – extinct language

“The language's last native speaker, Tevfik Esenç, died in 1992.”

Challenges for MT

• Cases
  – Give more flexibility for the sentence structure

  Der Löwe frißt das Zebra
  Den Löwe frißt das Zebra
  Das Zebra frißt den Löwe
  Das Zebra frißt der Löwe

Finnish has 15 cases!
Challenges for MT

• Prepositions
  – Change the meaning of the sentence

  I go to the house
  I go from the house
  I go in the house
  I go on the house
  I go by the house
  I go through the house
Challenges in MT

- Gender
  - „the“ vs. „der“, „die“, „das“, „dem“, „den“, „des“
Lexical semantics

- The meaning of a word
- band (music group)
- band (material)
- band (wavelength)
- fall (season)
- fall (verb)
WordNet

Noun

- S: (n) fall, autumn (the season when the leaves fall from the trees) "in the fall of 1973"
- S: (n) spill, tumble, fall (a sudden drop from an upright position) "he had a nasty spill on the ice"
- S: (n) Fall (the lapse of mankind into sinfulness because of the sin of Adam and Eve) "women have been blamed ever since the Fall"
- S: (n) descent, declivity, fall, decline, declination, declension, downslope (a downward slope or bend)
- S: (n) fall (a lapse into sin; a loss of innocence or of chastity) "a fall from virtue"
- S: (n) fall, downfall (a sudden decline in strength or number or importance) "the fall of the House of Hapsburg"
- S: (n) fall (a movement downward) "the rise and fall of the tides"
- S: (n) capitulation, fall, surrender (the act of surrendering (usually under agreed conditions)) "they were protected until the capitulation of the fort"
- S: (n) twilight, dusk, gloaming, gloam, nightfall, evenfall, fall, crepuscle, crepuscle (the time of day immediately following sunset) "he loved the twilight"; "they finished before the fall of night"

(https://wordnet.princeton.edu/)
WordNet

Verb

- **S**: (v) **fall** (descend in free fall under the influence of gravity) "The branch fell from the tree"; "The unfortunate hiker fell into a crevasse"
- **S**: (v) **descend**, **fall**, **go down**, **come down** (move downward and lower, but not necessarily all the way) "The temperature is going down"; "The barometer is falling"; "The curtain fell on the diva"; "Her hand went up and then fell again"
- **S**: (v) **fall** (pass suddenly and passively into a state of body or mind) "fall into a trap"; "She fell ill"; "They fell out of favor"; "Fall in love"; "fall asleep"; "fall prey to an imposter"; "fall into a strange way of thinking"; "she fell to pieces after she lost her work"
- **S**: (v) **fall**, **come** (come under, be classified or included) "fall into a category"; "This comes under a new heading"
- **S**: (v) **precipitate**, **come down**, **fall** (fall from clouds) "rain, snow and sleet were falling"; "Vesuvius precipitated its fiery, destructive rage on Herculaneum"
- **S**: (v) **fall** (suffer defeat, failure, or ruin) "We must stand or fall"; "fall by the wayside"
- **S**: (v) **fall** (die, as in battle or in a hunt) "Many soldiers fell at Verdun"; "Several deer have fallen to the same gun"; "The shooting victim fell dead"

(https://wordnet.princeton.edu/)
### Semantics in MT

<table>
<thead>
<tr>
<th>Verben</th>
</tr>
</thead>
<tbody>
<tr>
<td>to fall</td>
</tr>
<tr>
<td>to fall</td>
</tr>
<tr>
<td>to fall</td>
</tr>
<tr>
<td>to fall</td>
</tr>
<tr>
<td>to fall</td>
</tr>
<tr>
<td>to fall</td>
</tr>
<tr>
<td>to fall</td>
</tr>
<tr>
<td>to fall</td>
</tr>
<tr>
<td>to fall</td>
</tr>
</tbody>
</table>

- **fallen | fiel, gefallen |**
- **stürzen | stürzte, gestürzt |**
- **sinken | sank, gesunken | - z. B. Druck, Spannung**
- **zurückgehen | ging zurück, zurückgegangen |**
- **abnehmen | nahm ab, abgenommen |**
- **absinken | sank ab, abgesunken |**
- **hinfallen | fiel hin, hingefallen |**
- **abstürzen | stürzte ab, abgestürzt |**
- **nachgeben | gab nach, nachgegeben |**
- **zerfallen | zerfiel, zerfallen |**

(http://www.leo.org/)
Word sense disambiguation

- It is the task of determining the right sense of a word in a text

- Usually make use of dictionaries (definitions) and/or the context
Overview

- Words
- Sentences
- Corpora
Sentence structure

- Verb is the central element of a sentence

  Jane swims.
  Jane bought the house.
  Jane gave Joe the book.
Sentence structure

• Verbs may have many valencies, arguments, adjuncts, ...

Jane swims.
Jane bought the house.
Jane bought the house from Jim.
Jane gave Joe the book.
Jane bought the beautiful house.
Jane bought the beautiful house in the city center.
Sentence structure

- Sentences can include Recursion
  - Nested construction of constituents

Jane, who recently won the lottery, bought the house that was just put in the market.

- Jane bought the house.
- Jane recently won the lottery.
- The house was just put in the market.
Sentence structure

- Clause
  - Part of sentence which is composed of a verbs and its arguments or adjuncts

Jane bought the house that was just put in the market.
Sentence structure

- Structural ambiguity

  Joe eats steak with a knife.

  Jim eats steak with ketchup.

  Jane watches the man with the telescope.

  Jim washes the dishes and watches television with Jane.
Grammar

- Chunking or shallow parsing
  - Noun phrases, verbal phrases, etc.

(NP My dog) (ADVP also) (VP likes) (VP eating) (NP sausage) (. .)
Grammar

- Parse tree (full parsing)
  - It represents the syntactic structure of a sentence

```
NP   ADVP   VP
|     |      |
PRP$  NN   RB    VBZ
   |  also  likes |
My   dog

S     .
|      |
S     NP

VP
|      |
VBG   NN
|      |
eating sausage
```
Grammar

• Context-free grammars
  – Non-terminal symbols (POS tags and phrases)
  – Terminal symbols (words)

S → NP VP
S → VP
NP → NN
NP → PRP
NP → DT NN
NP → NP NP
NP → NP PP

PRP$ → My
NN → dog
RB → also
VBZ → likes
VBG → eating
NN → sausage
Grammar

- Probabilistic context-free grammars
  - Probabilities are associated to rules and POS tags

\[
\begin{align*}
0.9 & \quad S \rightarrow NP \ VP \\
0.1 & \quad S \rightarrow VP \\
0.3 & \quad NP \rightarrow NN \\
0.4 & \quad NP \rightarrow PRP \\
0.1 & \quad NP \rightarrow DT \ NN \\
0.2 & \quad NP \rightarrow NP \ NP \\
0.1 & \quad NP \rightarrow NP \ PP \\
& \ldots \\
1.0 & \quad PRP$ \rightarrow \ My \\
0.9 & \quad NN \rightarrow \ dog \\
1.0 & \quad RB \rightarrow \ also \\
0.6 & \quad VBZ \rightarrow \ likes \\
0.5 & \quad VBG \rightarrow \ eating \\
0.8 & \quad NN \rightarrow \ sausage \\
\end{align*}
\]
Grammar

• Dependency tree
  - Alternative syntactic structure
  - It shows relationships between words

(ROOT-0)
  \[\begin{array}{c}
  \text{ROOT-0} \\
  \text{likes-4} \\
  \text{eating-5} \\
  \text{also-3} \\
  \text{dog-2} \\
  \text{sausage-6} \\
  \text{My-1}
  \end{array}\]

(http://nlp.stanford.edu:8080/parser/index.jsp)
Grammar

- Head words
  - They are words on which others depend
Translation of sentence structure

- Reordering of words
- Insertion and deletion of function words
Pierre Vinken will join the board.
Translation of sentence structure

Vinken will join the board

Vinken wird dem Vorstand beitreten
Discourse

- Co-reference
  - Anaphora: denotes the act of referring *backwards* in a dialog or text
  - Cataphora: sees the act of referring *forward* in a dialog or text

Susan dropped the *plate*. It shattered loudly.
The *music* stopped, and *that* upset everyone.
Fred was *angry*, and *so* was I.
If Sam *buys* a new bike, I will do *it* as well.

Because *he* was very cold, *David* put on his coat.
Although Sam might do *so*, I will not *buy a new bike*.  

(https://en.wikipedia.org/wiki/Anaphora_(linguistics))
Overview

- Words
- Sentences
- Corpora
Domains

- Politics (European Parliament)
- News
- Conversations
- Biomedicine (neuroscience)
- ...

67
Modality

• Spoken language is different of written language.
Acquiring parallel corpora

- Parallel corpora

SAN FRANCISCO – It has never been easy to have a rational conversation about the value of gold. Lately, with gold prices up more than 300% over the last decade, it is harder than ever.

Acquiring parallel corpora

- Parallel corpora
  - Web pages
  - Books in many translations (Bible)
  - Manuals of products
  - ...

Crawling for parallel corpora

- Web technical issues, HTML parsing

This paper presents the state of the art of self-etch adhesive systems. Four topics are shown in this review and included: the historic of this category of bonding agents, bonding mechanism, characteristics/properties and the formation of acid-base resistant zone at enamel/dentin-adhesive interfaces. Also, advantages regarding etch-and-rinse systems and classifications of self-etch adhesive systems according to the number of steps and acidity are addressed. Finally, issues like the potential durability and clinical importance are discussed. Self-etch adhesive systems are promising materials because they are easy to use, bond chemically to tooth structure and maintain the dentin hydroxyapatite, which is important for the durability of the bonding.

Keywords:
Self-etch adhesive systems; adhesion; dentin.
Document alignment

- Some resources provide documents in more than one language

- **Topical Intrapocket Anesthesia During Scaling and Root Planing: a Randomized Clinical Trial**
  Antoniezz, Raquel Pippi; Cargnelutti, Bruna; Freitas, Daniela Napoleão; Guimarães, Magáii Beck; Zanatta, Fabricio Batistin; Feldens, Carlos Alberto

  - abstract in English | Portuguese
  - text in English
  - English (pdf epdf)

- **Is There a Relationship Between Periodontal Disease and Causes of Death? A Cross Sectional Study**
  Natto, Zuhair S.; Aladawy, Majdi; Alasqah, Mohammed; Papas, Athena

  - abstract in English | Portuguese
  - text in English
  - English (pdf epdf)

- **Phenytoin-Induced Gingival Overgrowth Management with Periodontal Treatment**
  Gurgel, Bruno César de Vasconcelos; Morais, Carlos Roberto Batista de; Rocha-Neto, Pedro Carlos da; Dantas, Euler Maciel; Pinto, Lêao Pereira; Costa, Antonio de Lisboa Lopes

  - abstract in English | Portuguese
  - text in English
  - English (pdf epdf)

(http://www.scielo.br/scielo.php?script=sci_issuetoc&pid=0103-644020150001&lng=en&nrm=iso)
Document alignment

- Are documents really aligned???
  - Mix of languages in the publication

English

Portuguese

Document alignment

- Web pages in more than one language

- But sometimes they are not perfect translation of each other
Kunduz bombing: MSF demands Afghan war crimes probe

Aid agency Medecins Sans Frontieres is seeking to invoke a never-used body to investigate the US bombing of its hospital in the Afghan city of Kunduz.

MSF said it did not trust internal military inquiries into the bombing that killed at least 22 people.

The International Humanitarian Fact-Finding Commission (IHFFC) was set up in 1991 under the Geneva Conventions.

MSF appelle à une enquête indépendante

Médecins Sans Frontières (MSF) appelle à une enquête internationale sur le bombardement américain la semaine dernière contre un hôpital appartenant à l'ONG dans la ville afghane de Kunduz.

MSF exhorte les Etats membres à saisir la Commission internationale humanitaire d'établissement des faits (CIHEF) afin que des investigations indépendantes soient menées.

Selon l'ONG, le bombardement, qui a tué 22 personnes, contrevenait à la Convention de Genève.

...
Comparable corpora

- Documents on the same topics
  - Same events, same date, etc.

- But no direct translations!
Comparable corpora

Chemie-Nobelpreis für Forschung zu DNA-Reparatur

Tomas Lindahl, Paul Modrich und Aziz Sancar haben herausgefunden, wie der Körper beschädigtes Erbgut wieder herstellt - ein wichtiger Grundstein für die Entwicklung von Krebsmedikamenten.

Der Nobelpreis für Chemie geht in diesem Jahr an die DNA-Forscher Paul Modrich aus den USA, Tomas Lindahl aus Schweden und an den türkisch-amerikanischen Wissenschaftler Aziz Sancar.

.....

(https://www.sueddeutsche.de/wissen/-chemie-nobelpreis-fuer-forschung-zu-dna-reparatur-1.2680858
Sentence alignment

- It is an essential step in corpus preparation

1 <=> 1  one sentence aligned to another
2,3 <=> 2
4 <=> 3
5 <=> 4
6,7 <=> 5
8,9 <=> 6
10 <=> 7
11 <=> omitted  one sentence not aligned to any other
12 <=> 8
13,14 <=> 9  two sentences aligned to one sentence
15 <=> 10
16 <=> 11

(output of the GMA tool: http://nlp.cs.nyu.edu/GMA/)
Sentence alignment

- All sentences need to be accounted
- Each sentence may occur in only one alignment

1 <=> 1
2,3 <=> 2
4 <=> 3
5 <=> 4
6,7 <=> 5
8,9 <=> 6
10 <=> 7
11 <=> omitted
12 <=> 8
13,14 <=> 9
15 <=> 10
16 <=> 11

(output of the GMA tool: http://nlp.cs.nyu.edu/GMA/)
Summary – Natural Language Processing steps

- Morphology
- POS Tagging
- Chunking
- Parsing
- Semantics

The dog jumped from a chair.
Summary – Natural Language Processing steps

- Semantics
- Parsing
- Chunking
- POS Tagging
- Morphology

The dog jumped from a chair.

The dog jump-[Past-Sing-3rd] from a chair.
Summary – Natural Language Processing steps

- Semantics
- Parsing
- Chunking
- POS Tagging
- Morphology

The/DT dog/NN jumped/VBD from/IN a/DT chair/NN ./.

The dog jumped from a chair.
Summary – Natural Language Processing steps

- Morphology
- POS Tagging
- Chunking
- Parsing
- Semantics

The dog jumped from a chair.

(http://nlp.stanford.edu:8080/parser/index.jsp)
Summary – Natural Language Processing steps

- Morphology
- POS Tagging
- Chunking
- Parsing
- Semantics

The dog jumped from a chair.
Summary – Natural Language Processing steps

Increased complexity of processing...

- Morphology
- POS Tagging
- Chunking
- Parsing
- Semantics

### Noun
- **S**: (n) chair (a seat for one person, with a support for the back) "he put his coat over the back of the chair and sat down"
- **S**: (n) professorship, chair (the position of professor) "he was awarded an endowed chair in economics"
- **S**: (n) president, chairman, chairwoman, chair, chairperson (the officer who presides at the meetings of an organization) "address your remarks to the chairperson"
- **S**: (n) electric chair, chair, death chair, hot seat (an instrument of execution by electrocution; resembles an ordinary seat for one person) "the murderer was sentenced to die in the chair"
- **S**: (n) chair (a particular seat in an orchestra) "he is second chair violin"

### Verb
- **S**: (v) chair, chairman (act or preside as chair, as of an academic department in a university) "She chaired the department for many years"
- **S**: (v) moderate, chair, lead (preside over) "John moderated the discussion"

The dog jumped from a chair.
Suggested reading

- Chapter 2