Goals of this Course

Automatic Linguistic Analysis of Large Corpora

- Part-of-Speech Tagging and Morphological Analysis
  - word classes, root forms, compounds, suffixes
- Syntactic Analysis
  - constituents, dependency relations
- Semantic Analysis
  - word senses, thematic roles, coreference, discourse relations

Using Automatically Annotated Corpora in Linguistics

- Corpora provide usage and frequency information
- Linguistic research (theoretical linguistics, psycholinguistics, corpus linguistics) requires annotation of words, constituents, semantics
- Many questions involve all of these at the same time
  - semantic characteristics of objects of the verb to cure...
- Many questions require large corpora (100 M words or more)
  - Beyond the scope of manual annotation projects

Using Automatically Annotated Corpora for Applications

- Lexical Acquisition
  - synonyms, hypernyms, class labels
- Information Extraction
  - Who bought what, What causes what, who founded what, ...?
- Question Answering
  - Which African capitals have more than 1 million inhabitants?
Course Overview

Monday  
Introduction to Computational Linguistics and Corpus Linguistics

Tuesday  
(Gertjan van Noord): Syntactic Analysis, Dependency Trees, Disambiguation, Statistical Models

Wednesday  
Corpus-based Linguistic Research

Thursday  
Using Syntactically Annotated Corpora for Lexical Acquisition, Information Extraction, and Question Answering

Friday  
Beyond Syntax. Thematic Roles, Word Senses, Semantic Classes, Coreference Resolution, Discourse Relations

Today

Introduction

Computational Linguistics  
Part of Speech tagging  
Syntax and Dependencies  
Semantic Annotation

Corpus Linguistics  
What is a corpus  
Frequencies  
Web as a Corpus

Searching Large Corpora

- Obtaining large amounts of text is relatively easy (internet)
  - English Wikipedia (nov 2006): approx 500 M words
  - Dutch Wikipedia (jul 2008): approx 120 M words
- But working with text almost always requires a certain amount of normalization and annotation

Tomica Wright is the widow of late rapper Eazy-E. She now owns her husband’s record label, Ruthless Records. She took ownership of Ruthless after her husband’s death in 1995. Although her husband died from AIDS, Tomica Wright is HIV negative as well as her kids fathered by Eric "Eazy-E" Wright.

Searching Large Corpora

(Linguistic Preprocessing and Annotation)

- running text → sentences, tokens, root forms
- Linguistic Information → POS-tags, constituent boundaries, dependency relations, ...
Searching for Linguistic Patterns

Find all sentences with the verb walk
- I walk to the store
- Kim walks/walked was walking to the store
- Kim went for a walk

Find all sentences with verb promise followed by that or to
- He promised that the cases would be withdrawn
- Their album was promising to be the most demanded CD

Linguistic Search
- How to find all forms of a verb? → Add root form
- How to distinguish verbs from nouns? → Add Part of Speech information

Part of Speech Tagging
- Assign a Part of Speech tag to each word in a sentence
- Example below from English Wikipedia, parsed using the Stanford Parser (Manning and Klein)
  - POS-tagging is a prerequisite for (or side-effect of) syntactic parsing

Eindhoven Corpus
- Eindhoven corpus is a 1M word Dutch corpus constructed manually in the seventies

How many Dutch sentences start with a subject/direct object/indirect object/...?
- Kim gaf het boek aan Sandy (Kim gave the book to Sandy)
- Het boek gaf Kim aan Sandy (The book, Kim gave to Sandy)
- Aan Sandy gaf Kim het boek (To Sandy, Kim gave the book)

How often does an indirect object occur with aan?
- Kim geeft het boek aan Sandy (Kim gives the book to Sandy)
- Kim geeft Sandy het boek (Kim gives Sandy the book)

Linguistic Search
- How to locate the subject? → Add dependency relations

Part of Speech Tagging

(NNS Manassas)
(VBD were)
(DT a)
(JJ seventies)
(NN rock)
(NN band)
(VBN formed)
(IN by)
(NNP Stephen)
(NNP Stills)
(IN in)
(CD 1971)
( . . )
Part of Speech Tagging

Use a dictionary?
- But many words belong to more than one PoS category
- Counts from BNC (British National Corpus) fragment
  - attack: Noun (109), Verb (59)
  - attempt: Noun (135), Verb (82)
  - before: Adv (143), Conj (305), Prep (434)
- Many words not present in a dictionary

Part of Speech Tagging

Three Methods
- Human, manual, annotation
  - Expensive
  - But very accurate (99% agreement)
- Automatically
  - Cheap
  - Relatively accurate (97% accuracy)
- Semi-automatic
  - Humans correct errors in automatically annotated material
  - Annotation tools suggest alternatives

Phrasal Prepositions in Dutch

Combination of
- preposition + (determiner) + noun + preposition

More or less fixed combinations
- Archaic (old) prepositions: ten opzichte van (in comparison with), ten gevolge van (as consequence of)
- Strange nouns: aan de vooravond van (on the eve of), bij monde van (according to), ..

Can we find more examples in large corpus?
- Requires searching for frequent preposition + (determiner) + noun + preposition combinations

Phrasal Prepositions in Dutch

ten opzichte van ‘with respect to’
in tegenstelling tot ‘as opposed to’
in verband met ‘in connection with’
in plaats van ‘instead of’
op basis van ‘on the basis of’
naar aanleiding van ‘in response to’
ter gelegenheid van ‘on the occasion of’
te midden van ‘amidst’
in het kader van ‘on the basis of’
Searching for Phrasal Prepositions in Dutch

Find all preposition + (determiner) + noun + preposition patterns

- Method 1: write a (Perl,...) script to collect all sequences of 3 or 4 lines with relevant PoS tags
- Method 2: Use specialized software
  - IMS Open Corpus Workbench (cwb.sourceforge.net)
  - GSearch (www.hcrc.ed.ac.uk/gsearch)
- Do some statistical analysis on the results
  - Frequency
  - Other tests (Mutual Information, $X^2$, log-likelihood)
  - Using Ngram-package (ngram.sourceforge.net),...

Searching for Phrasal Prepositions in Dutch

- Highest ranked phrasal prepositions according to log-likelihood
- Using 16 M word newspaper corpus, and a frequency cut-off of 10

<table>
<thead>
<tr>
<th>Rank</th>
<th>Preposition Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>in plaats van</td>
</tr>
<tr>
<td>2</td>
<td>onder leiding van</td>
</tr>
<tr>
<td>3</td>
<td>op basis van</td>
</tr>
<tr>
<td>4</td>
<td>ten opzichte van</td>
</tr>
<tr>
<td>5</td>
<td>op het gebied van</td>
</tr>
<tr>
<td>6</td>
<td>aan het eind van</td>
</tr>
<tr>
<td>7</td>
<td>in tegenstelling tot</td>
</tr>
<tr>
<td>8</td>
<td>op weg naar</td>
</tr>
<tr>
<td>9</td>
<td>op grond van</td>
</tr>
<tr>
<td>10</td>
<td>naar aanleiding van</td>
</tr>
<tr>
<td>11</td>
<td>met behulp van</td>
</tr>
<tr>
<td>12</td>
<td>na afloop van</td>
</tr>
<tr>
<td>13</td>
<td>aan de hand van</td>
</tr>
<tr>
<td>14</td>
<td>in verband met</td>
</tr>
<tr>
<td>15</td>
<td>in opdracht van</td>
</tr>
<tr>
<td>16</td>
<td>in het kader van</td>
</tr>
<tr>
<td>17</td>
<td>in ruil voor</td>
</tr>
<tr>
<td>18</td>
<td>op verzoek van</td>
</tr>
<tr>
<td>19</td>
<td>in de loop van</td>
</tr>
<tr>
<td>20</td>
<td>ten koste van</td>
</tr>
</tbody>
</table>


Syntactic Analysis

Syntactic Analysis (*Parsing*) assigns grammatical structure to sentences. Instead of working with strings of words, you have constituents (*Noun Phrases, Prepositional Phrases, Clauses, Adverbial Phrases, ...*), and grammatical functions (*Subject, Object, Modifier, ...*).

- Grammar Rules
  - Specify Syntactic Structures of the Language
- Lexicon
  - List Words and their properties (Part of Speech, ...)
- Parser
  - Given an input string, compute the (most likely) syntactic structure

Applications using Syntactic Information

- Information Extraction:
  - *Which topics do which Dutch politicians talk about?*
- Question Answering
  - *What is the capital of Togo?*
  - *How much did Man United pay for Berbatov?*
- Summarization
  - *Give an overview of the recent Duyvendak-affair*
- All these tasks can benefit from syntactic analysis
Each sentence consists of constituents
Each constituent may consist of smaller constituents
The smallest constituent is a single word
Each constituent has a dependency label

- subject, direct object, indirect object, modifier, verbal complement, determiner, prepositional complement, locative complement.

Een gebrek aan insuline leidt tot suikerziekte (A shortage of insuline causes diabetes)

Anthony Fokker overlijdt op 49-jarige leeftijd te New York (Anthony Fokker dies at age 49 in New York)

Alan Turing wordt op 7 juni 1954 dood aangetroffen (Alan Turing is found dead on June, 7th, 1954)
Verhoogde bloeddruk wordt hypertensie genoemd (High blood pressure is called hypertension)
Manassas were a seventies rock band formed by Stephen Still in 1971.

```
(ROOT
  (S
    (NP (NNS Manassas))
    (VP (VBD were)
      (NP
        (NP (DT a) (JJ seventies) (NN rock) (NN band))
        (VP (VBN formed)
          (PP (IN by)
            (NP
              (NP (NNP Stephen) (NNP Stills))
              (PP (IN in)
                (NP (CD 1971)))))
        (.))))
```

**Using Dependency Relations**

- Find all verb - object pairs, return head noun of the object
  
- grep dobj (and remove string positions and sort and count)

**Verb-Object pairs**

- searched 140 K Wikipedia sentences
  
- 102 dobj(took, place)
- 69 dobj(made, debut)
- 57 dobj(won, medal)
- 47 dobj(entered, career)
- 47 dobj(has, population)
- 44 dobj(customised, stamp)
- 44 dobj(takes, place)
## Objects of take

<table>
<thead>
<tr>
<th>Count</th>
<th>dobj</th>
<th>Count</th>
<th>dobj</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>(take, place)</td>
<td>7</td>
<td>(take, position)</td>
</tr>
<tr>
<td>40</td>
<td>(takes, place)</td>
<td>7</td>
<td>(taken, control)</td>
</tr>
<tr>
<td>19</td>
<td>(take, part)</td>
<td>7</td>
<td>(take, it)</td>
</tr>
<tr>
<td>13</td>
<td>(take, advantage)</td>
<td>6</td>
<td>(take, care)</td>
</tr>
<tr>
<td>11</td>
<td>(take, control)</td>
<td>5</td>
<td>(take, world)</td>
</tr>
<tr>
<td>10</td>
<td>(taken, place)</td>
<td>5</td>
<td>(take, them)</td>
</tr>
<tr>
<td>10</td>
<td>(take, him)</td>
<td>5</td>
<td>(takes, time)</td>
</tr>
<tr>
<td>9</td>
<td>(take, action)</td>
<td>5</td>
<td>(takes, them)</td>
</tr>
<tr>
<td>7</td>
<td>(takes, name)</td>
<td>5</td>
<td>(takes, origin)</td>
</tr>
<tr>
<td>7</td>
<td>(takes, control)</td>
<td>5</td>
<td>(takes, it)</td>
</tr>
</tbody>
</table>

## Semantic Annotation

- Some applications benefit from semantic information
  - Relation Extraction: Find relations between e.g. genes and diseases
  - Machine translation: translate Dutch *gerecht* as *dish* or *courthouse*
- Many forms of semantic information can be added
  - Named entity classes
  - Word senses (meanings)
  - Coreference relations
  - Discourse relations
  - Thematic roles
Some (most) words have more than one meaning or sense

- house, bug, danish, ....
- Word Sense Disambiguation is the task of selecting the correct meaning of a word
  - There was a bug in the room
  - There was a bug in the code

Ik hou niet van golf
Ik ben goed in golf
Ik speel golf

Een hoge golf sloeg op het strand
De golf maakte hem nat

I do not like golf
I am good at golf
I play golf

A high wave hit on the beach
The wave made him wet
Word Sense Disambiguation

Ik hou niet van golf
I do not like golf

Ik ben goed in golf
I am good at golf

Ik speel golf
I play golf

Een hoge golf sloeg op het strand
A high wave hit on the beach

De golf maakte hem nat
The wave made him wet

Golf na golf rolde naar de kust
Golf after wave rolled to the coast

What is a corpus?

▶ A collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language

▶ A collection of naturally occurring language text, chosen to characterize a state or variety of a language
(quotations courtesy British National Corpus web site)

What is a corpus? (Cont’d)

▶ There’s nothing particularly new in large collections of texts for academic research: for centuries people have been collecting manuscripts, books and newspapers for analysis of a very laborious nature. Thankfully, as technological advances make the computerized storage and access of large quantities of information easier, so the construction and use of text corpora continue to increase, and the potential for research has widened considerably.
(quotations courtesy British National Corpus web site)

What is a corpus?

A Corpus is

1. A collection of language data
   ▶ spoken or written

2. collected for linguistic purposes
   ▶ representative for your research question
   ▶ therefore, with a certain size

3. Electronically accessible
What is a corpus?

A Corpus is
1. A collection of language data
   ▶ spoken or written
2. collected for linguistic purposes
   ▶ representative for your research question
   ▶ therefore, with a certain size
3. Electronically accessible

What is not a corpus?
1. a single newspaper article
2. Last night's tv news
3. a single novel
4. a dictionary

Examples (old but still used widely)

English

Dutch
stop electing life peers by Trevor Williams. A move to stop Mr. Gaitskell from nominating any more labour peers is to be made at a meeting of labour MPs tomorrow.

Mr. Michael Foot has put down a resolution on the subject and he is to be backed by Mr. Will Griffiths, for Manchester Exchange though they may gather some left-wing support, a large majority of MPs are likely to turn down the Foot-Griffiths resolution to abolish Lords.

Examples (recent)

English
British National Corpus
- ca. 100 M words, both written and spoken language – but no sound files

Dutch
Corpus Gesproken Nederlands (CGN), Corpus of Spoken Dutch
- 10M words, only spoken language,
- Sound, phonemic transcriptions, Part-of-Speech, Constituents

Multilingual
CHILDES
- Children (and parents) in many languages, transcribed speech, 300 M characters
JEA: xxx vandaag?
ABE: he.
JEA: geen snor drinken.
JEA: xxx.
GER: moet ik helpen, Abel?
%com: ABE puts the sugar in the teacups.
ABE: ja.
ABE: en ik heb &6 een van mama.
GER: oh, oh.
JEA: maar ik hoef geen suiker, hoor.
ABE: xx hoef geen suiker.
GER: oh, ze hoeft geen suiker.
GER: ja.
GER: ja, ok.
ABE: da(t) (i)s lekker.
GER: ja.
GER: ik hou wel van een beetje suiker in me thee.
GER: ja, goed.
ABE: heb ik klaar [= nou ben ik klaar].
GER: nee, ik wil nog een beetje.

Examples (Under Construction)

German
IDS Corpus
▷ Institut für Deutsche Sprache
▷ eines Korpus der Gegenwartssprache von ca. 1,6 Milliarden Textwörtern

Dutch
LASSY
▷ Informatiekunde Groningen, Universiteit Leuven
▷ 500M words
▷ Syntactic Annotation (Part-of-Speech, Constituents)

LASSY syntactic annotation

Antony Fokker overleed op 49-jarige leeftijd te New York
Antony Fokker died at age 49 in New York
De Groninger binnenstad scoort onveranderd hoge waarderingscijfers bij haar gebruikers. Dat meldt het Groningse onderzoeksbureau Intraval in zijn jaarlijkse thermometer van de binnenstad. Volgens Intraval voelen ondernemers en bezoekers zich zeer veilig en is er grote tevredenheid over de sfeer van de binnenstad.

### Type/Token Ratio

- The number of types divided by the number of tokens
- TTR(binnenstad) = 76/133 = 0.571
- How does type/token ratio correlate with text length?
- What does type/token ratio tell us about a text?

#### TT Ratio for increasing text sizes

<table>
<thead>
<tr>
<th># Tokens (x 1000)</th>
<th>Wikipedia Types</th>
<th>TT Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>17.360</td>
<td>0.173</td>
</tr>
<tr>
<td>200</td>
<td>27.775</td>
<td>0.138</td>
</tr>
<tr>
<td>300</td>
<td>37.656</td>
<td>0.125</td>
</tr>
<tr>
<td>400</td>
<td>47.721</td>
<td>0.119</td>
</tr>
<tr>
<td>500</td>
<td>55.227</td>
<td>0.110</td>
</tr>
<tr>
<td>600</td>
<td>61.644</td>
<td>0.103</td>
</tr>
<tr>
<td>700</td>
<td>70.535</td>
<td>0.101</td>
</tr>
<tr>
<td>800</td>
<td>76.014</td>
<td>0.095</td>
</tr>
<tr>
<td>900</td>
<td>82.488</td>
<td>0.092</td>
</tr>
<tr>
<td>1000</td>
<td>87.954</td>
<td>0.087</td>
</tr>
</tbody>
</table>

Type/Token Ratio decreases as text size increases

#### Wikipedia vs Newspaper (AD 1999)

<table>
<thead>
<tr>
<th># Tokens (x 1000)</th>
<th>Wikipedia Types</th>
<th>TT Ratio</th>
<th>AD 1999 Types</th>
<th>TT Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>17.360</td>
<td>0.173</td>
<td>17.038</td>
<td>0.170</td>
</tr>
<tr>
<td>200</td>
<td>27.775</td>
<td>0.138</td>
<td>26.706</td>
<td>0.134</td>
</tr>
<tr>
<td>300</td>
<td>37.656</td>
<td>0.125</td>
<td>34.172</td>
<td>0.113</td>
</tr>
<tr>
<td>400</td>
<td>47.721</td>
<td>0.119</td>
<td>40.293</td>
<td>0.101</td>
</tr>
<tr>
<td>500</td>
<td>55.227</td>
<td>0.110</td>
<td>46.181</td>
<td>0.092</td>
</tr>
<tr>
<td>600</td>
<td>61.644</td>
<td>0.103</td>
<td>51.607</td>
<td>0.086</td>
</tr>
<tr>
<td>700</td>
<td>70.535</td>
<td>0.101</td>
<td>56.175</td>
<td>0.080</td>
</tr>
<tr>
<td>800</td>
<td>76.014</td>
<td>0.095</td>
<td>60.968</td>
<td>0.076</td>
</tr>
<tr>
<td>900</td>
<td>82.488</td>
<td>0.092</td>
<td>65.751</td>
<td>0.073</td>
</tr>
<tr>
<td>1000</td>
<td>87.954</td>
<td>0.087</td>
<td>70.005</td>
<td>0.070</td>
</tr>
</tbody>
</table>

There is more repetition (less variation) in AD then in Wikipedia.
**Type/Token Ratio**

Wikipedia vs Newspaper (AD 1999)

<table>
<thead>
<tr>
<th># Tokens (x 1000)</th>
<th>Wikipedia</th>
<th>AD 1999</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Types</td>
<td>TT ratio</td>
</tr>
<tr>
<td>100</td>
<td>17.360</td>
<td>0.173</td>
</tr>
<tr>
<td>200</td>
<td>27.775</td>
<td>0.138</td>
</tr>
<tr>
<td>300</td>
<td>37.656</td>
<td>0.125</td>
</tr>
<tr>
<td>400</td>
<td>47.721</td>
<td>0.119</td>
</tr>
<tr>
<td>500</td>
<td>55.227</td>
<td>0.110</td>
</tr>
<tr>
<td>600</td>
<td>61.644</td>
<td>0.103</td>
</tr>
<tr>
<td>700</td>
<td>70.535</td>
<td>0.101</td>
</tr>
<tr>
<td>800</td>
<td>76.014</td>
<td>0.095</td>
</tr>
<tr>
<td>900</td>
<td>82.488</td>
<td>0.092</td>
</tr>
<tr>
<td>1000</td>
<td>87.954</td>
<td>0.087</td>
</tr>
</tbody>
</table>

There is more repetition (less variation) in AD than in Wikipedia.

**Word Frequencies**

- Few words very frequent (the, a, and, in, on, that, ...)
- Many low-frequency words

**Zipf’s Law (Wikipedia)**

Zipf’s law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Thus the most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, etc. For example, in the Brown Corpus "the" is the most frequently occurring word, and all by itself accounts for nearly 7% of all word occurrences (69971 out of slightly over 1 million). True to Zipf’s Law, the second-place word "of" accounts for slightly over 3.5% of words (36411 occurrences), followed by "and" (28852). Only 135 vocabulary items are needed to account for half the Brown Corpus.

\[
\text{freq}(W_R) = \alpha \frac{\text{freq}(W_1)}{R}
\]
Zipf’s Law (Wikipedia)

A plot of word frequency in Wikipedia (November 27, 2006). The plot is in log-log coordinates. x is rank of a word in the frequency table; y is the total number of the word’s occurrences. Most popular words are the, of and and, as expected. Zipf’s law corresponds to the upper linear portion of the curve, roughly following the green (1/x) line.

Web as a Corpus

When do you have enough data?
1. Corpora are limited in size.
2. Some questions require large amounts of data.
3. Web is much larger than largest corpus.
4. Can we use the web as a corpus?

Estimate Size of the Web (in Words)
1. Identify language-specific, general (domain-independent), words
2. Estimate frequency in a corpus of known size
3. Collect web search engine counts
4. Estimate size of the web for the given language

Size of the Web

- Oostendorp & van der Wouden, Corpus Internet, 1998
  - Counts for the word eens on the Web, in corpora
- Grefenstette & Nioche, Estimation of English and non-English Language Use on the WWW, 2000
- Google N-grams database
  - (http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html)

Estimation of Web Size

<table>
<thead>
<tr>
<th>Language</th>
<th>O&amp;vdW</th>
<th>G&amp;N</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch</td>
<td>150M</td>
<td>622M</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>47.2B</td>
<td>1.024B</td>
<td></td>
</tr>
</tbody>
</table>

Current Size of the Dutch Web

Frequency of eens (once)

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size</th>
<th>Eens</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>INL</td>
<td>720k</td>
<td>966</td>
<td>1/730</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>58M</td>
<td>8064</td>
<td>1/6250</td>
</tr>
<tr>
<td>AD 1999</td>
<td>14.1M</td>
<td>9249</td>
<td>1/1566</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td>1/1000</td>
</tr>
</tbody>
</table>

Estimation of Web Size

- Google N-grams database
  - (http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html)
### Current Size of the Dutch Web

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size</th>
<th>Eens</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>INL</td>
<td>720k</td>
<td>966</td>
<td>1/730</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>58M</td>
<td>8064</td>
<td>1/6250</td>
</tr>
<tr>
<td>AD 1999</td>
<td>14.1M</td>
<td>9249</td>
<td>1/1566</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td>1/1000</td>
</tr>
</tbody>
</table>

### Frequency of *eens (once)*

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size</th>
<th>Eens</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>INL</td>
<td>720k</td>
<td>966</td>
<td>1/730</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>58M</td>
<td>8064</td>
<td>1/6250</td>
</tr>
<tr>
<td>AD 1999</td>
<td>14.1M</td>
<td>9249</td>
<td>1/1566</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td>1/1000</td>
</tr>
</tbody>
</table>

### Web as a Corpus

**Is the Web useful for linguistic research?**

- Using the web as a corpus has **many disadvantages**
  - No control of content, selection
  - Not linguistically annotated
- But it is **much larger in size** than any controlled and/or annotated corpus
  - *There is no data like more data!*

### Reliability of Web Data

- **Web data is noisy**
  - Newspaper vs Google
  - Anyone can place stuff in the internet..
    - Blogs are notorious...
  - How to interpret Google/Yahoo?MSN/.. data?

### Reliability of Web Counts

**Google** Web Counts are unreliable

- Veronis, Beaver, Liberman (Language Log)
- Illogical behaviour of OR
  - Chirac: 3.2 mln, Chirac or Sarkozy 1.7 mln, Chirac and Sarkozy 1.6 mln, Chirac and Chirac: 1.7 mln, Chirac Chirac: 1.7 mln
  - Bouma: 457.000, Bouma OR Bouma: 503.000
- Number of Hits fluctuates strongly
  - the (Feb): 8 bln, the (Mar): 3.2 bln
Using Web Counts

Spelling Variants: Compounds with -s or not?
- In Dutch many compounds optionally take an -s

<table>
<thead>
<tr>
<th>+s count</th>
<th>-s count</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>weervoorspelling 295K</td>
<td>weervoorspelling 125K</td>
<td>wheather forecast</td>
</tr>
<tr>
<td>spellingsregel 1020</td>
<td>spellingregel 909</td>
<td>spelling rule</td>
</tr>
<tr>
<td>besturingsysteem 1.25M</td>
<td>besturingsysteem 108K</td>
<td>operating system</td>
</tr>
<tr>
<td>doodskist 57.7K</td>
<td>doodskist 6.9K</td>
<td>coffin</td>
</tr>
<tr>
<td>drugsbeleid 87K</td>
<td>drugsbeleid 10K</td>
<td>drugs policy</td>
</tr>
<tr>
<td>moedersdag 8K</td>
<td>moedersdag 700K</td>
<td>mothersday</td>
</tr>
</tbody>
</table>

Dialectal Variation

<table>
<thead>
<tr>
<th>NL</th>
<th>BE</th>
</tr>
</thead>
<tbody>
<tr>
<td>eens (once) 2M</td>
<td>753K</td>
</tr>
<tr>
<td>alweer (again) 603</td>
<td>282K</td>
</tr>
<tr>
<td>weeral (again) 71K</td>
<td>267K</td>
</tr>
<tr>
<td>vast en zeker (for sure) 263K</td>
<td>68K</td>
</tr>
<tr>
<td>zeker en vast (for sure) 65K</td>
<td>174K</td>
</tr>
<tr>
<td>nootmuskaat (nutmeg) 83K</td>
<td>20K</td>
</tr>
<tr>
<td>muskaatnoot (nutmeg) 606</td>
<td>17K</td>
</tr>
</tbody>
</table>

Summary

Computational Linguistics
- Offers the tools to annotate large text collections automatically
- Useful for applications
- Useful for linguistic research

Corpus Linguistics
- Study of linguistics using real language data
- Corpora can be manually or automatically annotated
- Corpora vary widely in size