Computational Linguistics
Introduction

Gosse Bouma

Information Science
University of Groningen

LOT Winterschool 2009
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
Goals of this Course

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
Goals of this Course

Using Automatically Annotated Corpora for Applications

- Lexical Acquisition
  - synonyms, hypernyms, class labels
- Information Extraction
- 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

1. Introduction

2. Computational Linguistics
   - Part of Speech tagging
   - Syntax and Dependencies
   - Semantic Annotation

3. 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

(LIST
  (S
    (NP (PRP She))
    (ADVP (RB now))
    (VP (VBZ owns))
    (NP
      (NP (NP (PRP$ her) (NN husband) (POS ’s))
        (NN record) (NN label))
      (, ,)
      (NP (NNP Ruthless) (NNPS Records)))
    (, .)))

Linguistic Preprocessing and Annotation

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

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

- 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

(NNS Manassas)
(VBD were)
(DT a)
(JJ seventies)
(NN rock)
(NN band)
(VBN formed)
(IN by)
(NNP Stephen)
(NNP Stills)
(IN in)
(CD 1971)
Eindhoven Corpus

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

```
Dit Pron
in Prep
verband N
met Prep
de Art
gemiddeld Adj
langere Adj
levensduur N
van Prep
de Art
vrouw N
```
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>Phrasal Preposition</th>
<th>Rank</th>
<th>Phrasal Preposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>in plaats van</td>
<td>11</td>
<td>met behulp van</td>
</tr>
<tr>
<td>2</td>
<td>onder leiding van</td>
<td>12</td>
<td>na afloop van</td>
</tr>
<tr>
<td>3</td>
<td>op basis van</td>
<td>13</td>
<td>aan de hand van</td>
</tr>
<tr>
<td>4</td>
<td>ten opzichte van</td>
<td>14</td>
<td>in verband met</td>
</tr>
<tr>
<td>5</td>
<td>op het gebied van</td>
<td>15</td>
<td>in opdracht van</td>
</tr>
<tr>
<td>6</td>
<td>aan het eind van</td>
<td>16</td>
<td>in het kader van</td>
</tr>
<tr>
<td>7</td>
<td>in tegenstelling tot</td>
<td>17</td>
<td>in ruil voor</td>
</tr>
<tr>
<td>8</td>
<td>op weg naar</td>
<td>18</td>
<td>op verzoek van</td>
</tr>
<tr>
<td>9</td>
<td>op grond van</td>
<td>19</td>
<td>in de loop van</td>
</tr>
<tr>
<td>10</td>
<td>naar aanleiding van</td>
<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
Dependency Trees

- 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)
Stanford Parser Dependencies

Phrase Structure

The Pevensie children eat pavenders when stranded on the island of Cair Paravel in the novel Prince Caspian.
Stanford Parser Dependencies

(ROOT
  (S
    (NP (DT The) (NNP Pevensie) (NNS children))
    (VP (VBP eat)
      (NP (NNS pavers))
      (SBAR
        (WHADVP (WRB when))
        (S
          (VP (VBN stranded)
            (PP (IN on)
              (NP
                (NP (DT the) (NN island))
                (PP (IN of)
                  (NP
                    (NP (NNP Cair) (NNP Paravel))
                    (PP (IN in)
                      (NP (DT the) (JJ novel) (NNP Prince) (NNP Caspian))))))))))
    (.) (. .)))
Stanford Parser Dependencies

Dependency Relations

The Pevensie children eat pavenders when stranded on the island of Cair Paravel in the novel Prince Caspian

det(children-3, The-1)
nn(children-3, Pevensie-2)
nsubj(eat-4, children-3)
dobj(eat-4, pavenders-5)
advmod(stranded-7, when-6)
dep(eat-4, stranded-7)
prep(stranded-7, on-8)
det(island-10, the-9)
pobj(on-8, island-10)

prep(island-10, of-11)
nn(Paravel-13, Cair-12)
pobj(of-11, Paravel-13)
prep(Paravel-13, in-14)
det(Caspian-18, the-15)
amod(Caspian-18, novel-16)
nn(Caspian-18, Prince-17)
pobj(in-14, Caspian-18)
Stanford Parser Dependencies

Phrase Structure

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))))))))
  (. .))))
Manassas were a seventies rock band formed by Stephen Still in 1971.

nsubj(band-6, Manassas-1)  
cop(band-6, were-2)  
det(band-6, a-3)  
amod(band-6, seventies-4)  
nn(band-6, rock-5)  
partmod(band-6, formed-7)  
prep(formed-7, by-8)  
nn(Stills-10, Stephen-9)  
pobj(by-8, Stills-10)  
pobj(by-8, Stills-10)  
prep(Stills-10, in-11)  
pobj(in-11, 1971-12)
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, pole)
  - 47 dobj(take, place)
  - 47 dobj(began, career)
  - 45 dobj(has, population)
  - 44 dobj(had, population)
  - 44 dobj(customised, stamp)
  - 40 dobj(takes, place)
  - 39 dobj(won, medal)
  - 39 dobj(changed, name)
  - 35 dobj(holds, people)
  - 32 dobj(started, career)
  - 32 dobj(expanding, it)
  - 31 dobj(help, Wikipedia)
  - 30 dobj(fill, vacancy)
  - 29 dobj(made, appearances)
Objects of take

47 dobj(take, place)  
40 dobj(takes, place)  
19 dobj(take, part)  
13 dobj(take, advantage)  
11 dobj(take, control)  
10 dobj(taken, place)  
10 dobj(take, him)  
9 dobj(take, action)  
7 dobj(takes, name)  
7 dobj(takes, control)  

7 dobj(take, position)  
7 dobj(taken, control)  
7 dobj(take, it)  
6 dobj(take, care)  
5 dobj(take, world)  
5 dobj(take, them)  
5 dobj(takes, time)  
5 dobj(takes, them)  
5 dobj(takes, origin)  
5 dobj(takes, it)
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
Named Entity Classes

GATE Viewer -- doc2 -- Named Entities

<BDOC>
<DODID> wsj94 008.0212 </DODID>
<DODNO> 940413-0062. </DODNO>
<HL> Who's News:

@ Burns Fry Ltd </HL>
<DD> 04/13/94 </DD>
<SO> WALL STREET JOURNAL (J), PAGE B10 </SO>
<CO> MER </CO>
<IN> SECURITIES (SCR) </IN>
<TXT>
<p> BURNS FRY Ltd. (Toronto) -- Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokesman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month. </p>
</TXT>
</DOC>
Named Entity Classes

While the House and Senate intelligence oversight committee have received classified information about planned covert operations against Iraq, the C.I.A. has not told lawmakers how the agency and the Bush administration see those operations fitting into the larger war on Iraq, or the global war on terrorism, Congressional officials said.

"What they haven't told us is how does the intelligence piece fit into the larger offensive against Iraq, or how do these extra demands on our intelligence capabilities effect our commitment to the war on terrorism in Afghanistan," said one official.

Congressional leaders complained that they have been left in the dark on how the intelligence community will be used just as they are about to debate a resolution to support war with Iraq.

Congressional leaders said the decision to fight the Congressional request may stem from a fear of exposing divisions within the intelligence community over the administration's Iraq strategy, perhaps including a debate between the agency and the Pentagon over the military's role in intelligence operations in Iraq.

Defense Secretary Donald H. Rumsfeld has been moving to strengthen his control over the military's intelligence apparatus, potentially setting up a turf war for dominance among American intelligence officials. Mr. Rumsfeld has also been pushing to expand the role of
Named Entity Classes

Analysis of myeloid-associated genes in human hematopoietic progenitor cells.

The distribution of myeloid lineage-associated cytokine receptors and lysosomal proteins was analyzed in human CD34+ cord blood cell (CB) subsets at different stages of myeloid commitment by reverse-transcriptase polymerase chain reaction (RT-PCR). The highly specific granulomonocyte-associated lysosomal proteins myeloperoxidase (MPO) and lysozyme (LZ), as well as the transcription factor PU.1, were already detectable in the most immature CD34+Thy-1+ subset. Messenger RNA (mRNA) levels for the granulocyte-colony stimulating factor (G-CSF)
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
Word Sense Disambiguation

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

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

Een hoge golf sloeg op het strand
De golf maakte hem nat
Golf na golf rolde naar de kust

A high wave hit on the beach
The wave made him wet
Golf after wave rolled to the coast
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
Word Sense Disambiguation

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
Golf na golf rolde naar de kust

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
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?

- 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 not a corpus?

1. a single newspaper article
2. Last night's tv news
3. a single novel
4. a dictionary
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

Examples (old but still used widely)

**English**


**Dutch**

Eindhoven Corpus

<samp_tel_2-10-1-cdb> Misc (markup)
<zin> Misc (markup)
Dit Pron (aanw, neut, zelfst)
in Prep (voor)
verband N (soort, ev, neut)
met Prep (voor)
de Art (bep, zijd_of_mv, neut)
gemiddeld Adj (adv, stell, onverv)
langere Adj (attr, vergr, verv_neut)
levensduur N (soort, ev, neut)
van Prep (voor)
de Art (bep, zijd_of_mv, neut)
vrouw N (soort, ev, neut)
.
Punc (punt)
LOB corpus

stop_VV0 electing_VBG life_NN peers_NNS . by_IO Trevor_NP Williams_NP .
a_AT1 move_NN to_TO stop_VV0 \0Mr_NNSB1 Gaitskell_NP from_IO nominating_VBG any_DD more_DA labour_NN life_NN peers_NNS is_VBZ to_TO be_VB0 made_VBN at_IO a_AT1 meeting_NN of_IO labour_NN \0MPs_NNSB2 tomorrow_NN1 . \0Mr_NNSB1 Michael_NP Foot_NP has_VHZ put_VBN down_RP a_AT1 resolution_NN on_IO the_AT1 subject_NN and_CC he_PPHO1 is_VBZ to_TO be_VB0 backed_VBN by_IO \0Mr_NNSB1 Will_NP Griffiths_NP ,_, \0MP_NNSB1 for_IO Manchester_NP Exchange_NP though_CS they_PPHS2 may_VM gather_VV0 some_DD left-wing_JB support_NN ,_, a_AT1 large_JJ majority_NN of_IO labour \0MPs_NNSB2 are_VBR likely_JJ to_TO turn_VV0 down_RP the_AT1 Foot-Griffiths_NP resolution_NN abolish_VV0 Lords_NNSB2 .
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
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
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
CHILDES DUTCH

*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 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.
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)
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.


Bij de horecaondernemers en avondbezoekers voelt 73 procent zich veilig in de binnenstad en vindt 90 procent het gezellig. In 1998 was dat nog respectievelijk 60 en 84 procent.
Types and Tokens

How many words does a text contain?

- **Tokens**
  - The number of word tokens in a text is the total number of words that occur in a text
  - if *the* occurs 10 times, it is counted 10 times

- **Types**
  - The number of word types in a text is the total number of different words that occur in a text
  - if *the* occurs 10 times, it is counted only once
1 Word per Line

```bash
$$ tr '' \n < binnenstad.txt | tr 'A-Z' 'a-z' \
   | sed 's/[,.]//' > binnenstad.lst
$$ less binnenstad.lst

de
groninger
binnenstad
scoort
onveranderd
hoge
waarderingscijfers
bij
haar
gebruikers
dat
meldt
```

Gosse Bouma
## Tokens and Types

```bash
$ wc -l binnenstad.lst
133
%
$ sort -u binnenstad.lst | wc -l
76
$ sort binnenstad.lst | uniq -c | sort -nr
```

<table>
<thead>
<tr>
<th>Count</th>
<th>Token</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>de</td>
</tr>
<tr>
<td>8</td>
<td>en</td>
</tr>
<tr>
<td>7</td>
<td>in</td>
</tr>
<tr>
<td>6</td>
<td>binnenstad</td>
</tr>
<tr>
<td>5</td>
<td>procent</td>
</tr>
<tr>
<td>4</td>
<td>van</td>
</tr>
<tr>
<td>4</td>
<td>bezoekers</td>
</tr>
<tr>
<td>3</td>
<td>zich</td>
</tr>
<tr>
<td>3</td>
<td>intraval</td>
</tr>
<tr>
<td>3</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>voelen</td>
</tr>
<tr>
<td>2</td>
<td>vindt</td>
</tr>
<tr>
<td>2</td>
<td>veilig</td>
</tr>
<tr>
<td>1</td>
<td>zijn</td>
</tr>
<tr>
<td>1</td>
<td>zegt</td>
</tr>
<tr>
<td>1</td>
<td>zeer</td>
</tr>
</tbody>
</table>

Gosse Bouma
Types and Tokens

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?
# Type/Token Ratio

## 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.
Type/Token Ratio

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
### 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 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 then in Wikipedia.
Most Frequent Word in Dickens, A Tale of Two Cities

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Count</th>
<th>% of text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the</td>
<td>8017</td>
<td>5.89</td>
</tr>
<tr>
<td>2</td>
<td>and</td>
<td>4928</td>
<td>3.62</td>
</tr>
<tr>
<td>3</td>
<td>of</td>
<td>4015</td>
<td>2.95</td>
</tr>
<tr>
<td>4</td>
<td>to</td>
<td>3462</td>
<td>2.54</td>
</tr>
<tr>
<td>5</td>
<td>a</td>
<td>2921</td>
<td>2.14</td>
</tr>
<tr>
<td>6</td>
<td>in</td>
<td>2581</td>
<td>1.89</td>
</tr>
<tr>
<td>7</td>
<td>it</td>
<td>2003</td>
<td>1.47</td>
</tr>
<tr>
<td>8</td>
<td>his</td>
<td>2002</td>
<td>1.47</td>
</tr>
<tr>
<td>9</td>
<td>i</td>
<td>1901</td>
<td>1.39</td>
</tr>
<tr>
<td>10</td>
<td>that</td>
<td>1884</td>
<td>1.38</td>
</tr>
<tr>
<td>11</td>
<td>he</td>
<td>1830</td>
<td>1.34</td>
</tr>
<tr>
<td>12</td>
<td>was</td>
<td>1761</td>
<td>1.29</td>
</tr>
<tr>
<td>13</td>
<td>you</td>
<td>1372</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Count</th>
<th>% of text</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>with</td>
<td>1307</td>
<td>0.96</td>
</tr>
<tr>
<td>15</td>
<td>had</td>
<td>1298</td>
<td>0.95</td>
</tr>
<tr>
<td>16</td>
<td>as</td>
<td>1139</td>
<td>0.83</td>
</tr>
<tr>
<td>17</td>
<td>her</td>
<td>1036</td>
<td>0.76</td>
</tr>
<tr>
<td>18</td>
<td>at</td>
<td>1030</td>
<td>0.75</td>
</tr>
<tr>
<td>19</td>
<td>him</td>
<td>964</td>
<td>0.70</td>
</tr>
<tr>
<td>20</td>
<td>for</td>
<td>949</td>
<td>0.69</td>
</tr>
<tr>
<td>21</td>
<td>on</td>
<td>920</td>
<td>0.67</td>
</tr>
<tr>
<td>22</td>
<td>not</td>
<td>838</td>
<td>0.61</td>
</tr>
<tr>
<td>23</td>
<td>is</td>
<td>809</td>
<td>0.59</td>
</tr>
<tr>
<td>24</td>
<td>be</td>
<td>762</td>
<td>0.55</td>
</tr>
<tr>
<td>25</td>
<td>have</td>
<td>737</td>
<td>0.54</td>
</tr>
</tbody>
</table>
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.

\[
freq(W_R) = \alpha \frac{freq(W_1)}{R}
\]
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>1/1000</td>
<td></td>
</tr>
</tbody>
</table>

Web-counts (april 2008) for *eens (once)*

<table>
<thead>
<tr>
<th>Engine</th>
<th>Count</th>
<th>Websize</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google (domain NL)</td>
<td>2.0M</td>
<td>2B</td>
</tr>
<tr>
<td>Google (lg Dutch)</td>
<td>2.3M</td>
<td>2.3B</td>
</tr>
<tr>
<td>Ilse</td>
<td>5.5M</td>
<td>5.5B</td>
</tr>
<tr>
<td>Google</td>
<td>48.0M</td>
<td>48.0B</td>
</tr>
<tr>
<td>Yahoo (lg Dutch)</td>
<td>106.0M</td>
<td>106.0B</td>
</tr>
<tr>
<td>Yahoo</td>
<td>110.0M</td>
<td>110.0B</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>

Web-counts (april 2008) for *eens* (*once*)

<table>
<thead>
<tr>
<th>Engine</th>
<th>Count</th>
<th>Websize</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google (domain NL)</td>
<td>2.0M</td>
<td>2B</td>
</tr>
<tr>
<td>Google (lg Dutch)</td>
<td>2.3M</td>
<td>2.3B</td>
</tr>
<tr>
<td>Ilse</td>
<td>5.5M</td>
<td>5.5B</td>
</tr>
<tr>
<td>Google</td>
<td>48.0M</td>
<td>48.0B</td>
</tr>
<tr>
<td>Yahoo (lg Dutch)</td>
<td>106.0M</td>
<td>106.0B</td>
</tr>
<tr>
<td>Yahoo</td>
<td>110.0M</td>
<td>110.0B</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
Spelling Variants: Compounds with -s or not?

- In Dutch many compounds optionally take an -s

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

In some cases, the counts and spellings differ, but the meaning is preserved.
# Using Web Counts

## Dialectal Variation

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