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
  - *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

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

```
(ROOT
  (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 occassion 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

1	in plaats van
2	onder leiding van
3	op basis van
4	ten opzichte van
5	op het gebied van
6	aan het eind van
7	in tegenstelling tot
8	op weg naar
9	op grond van
10	naar aanleiding van

11	met behulp van
12	na afloop van
13	aan de hand van
14	in verband met
15	in opdracht van
16	in het kader van
17	in ruil voor
18	op verzoek van
19	in de loop van
20	ten koste van

- Bouma and Villada, *Corpus-based acquisition of collocational prepositional phrases*, CLIN 2001.

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

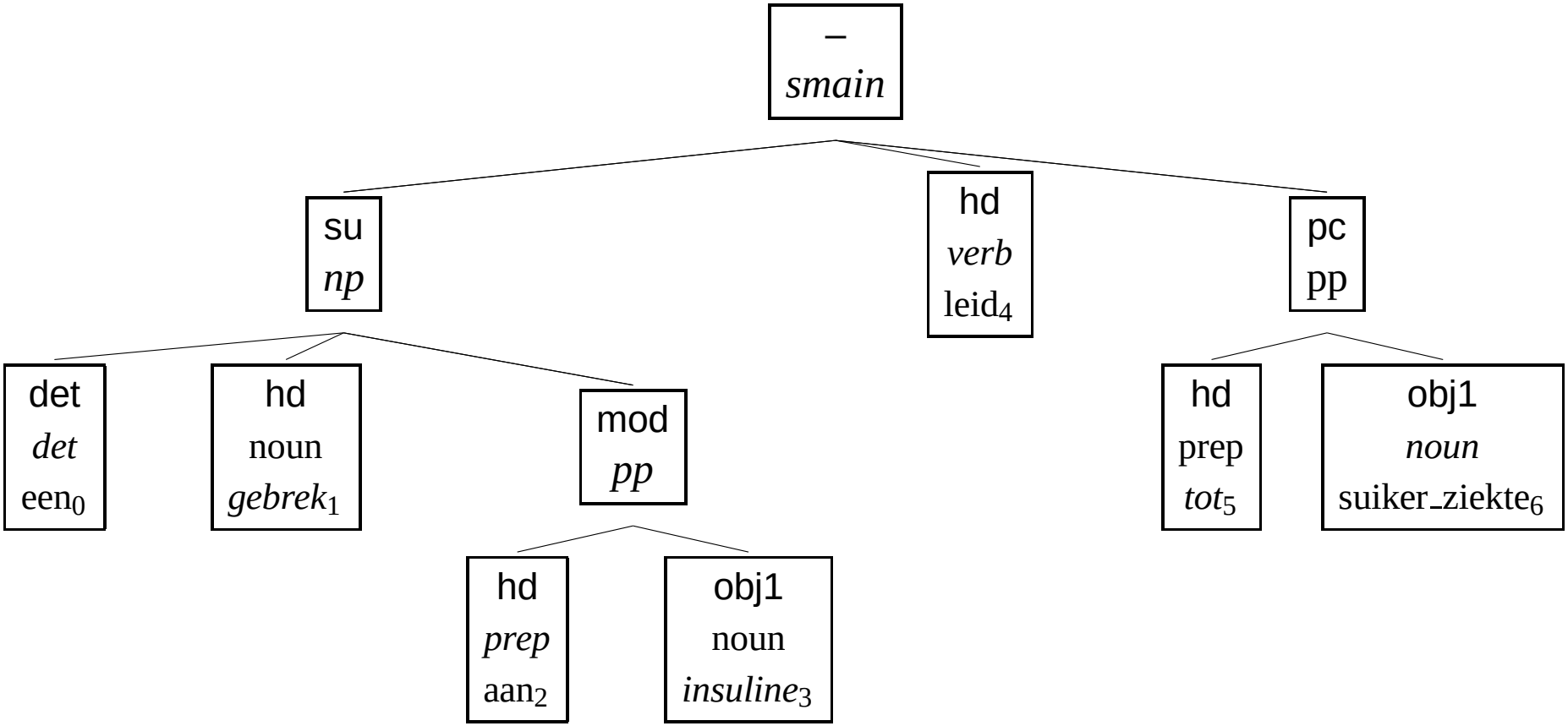
Applications using Syntactic Information

- 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

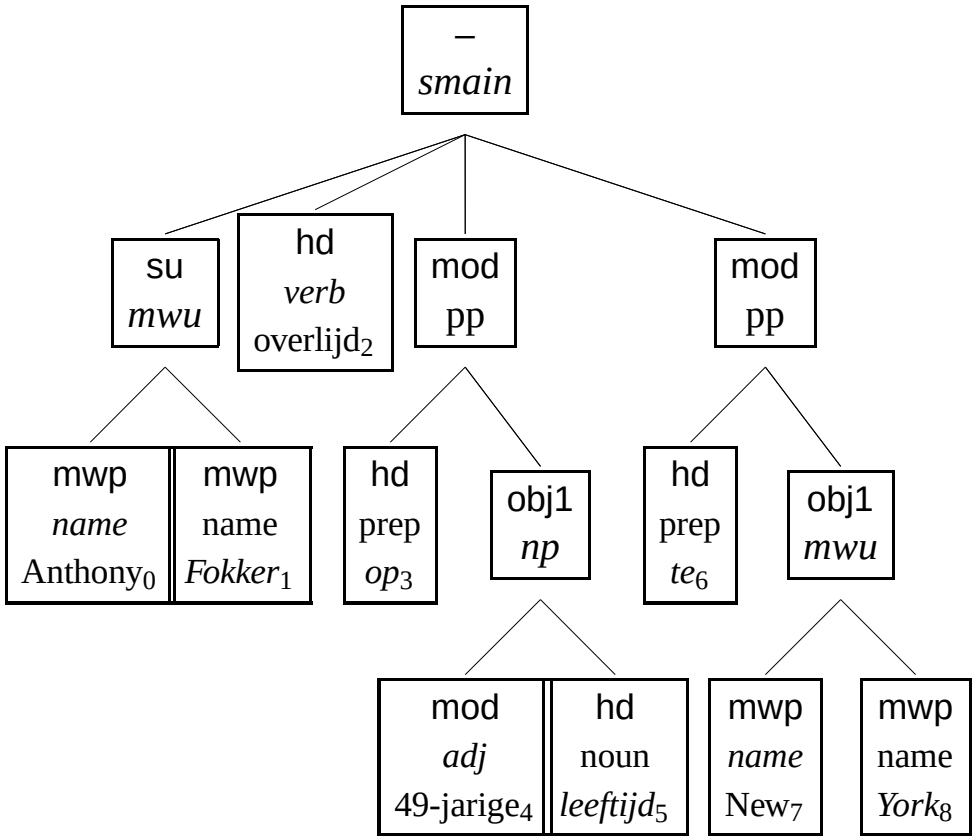
- Each sentence consists of constituents
- Each constituent may consist of smaller constituents
- The smallest constituent is a single word
- Each constituent has a dependency label
  - subject, direct object, indirect object, modifier, verbal complement, determiner, prepositional complement, locative complement.

Dependency Trees



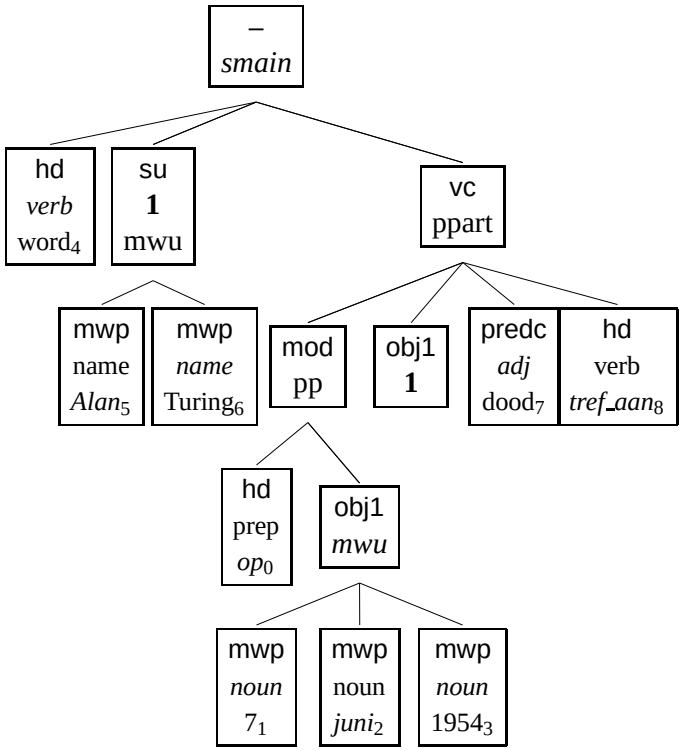
Een gebrek aan insuline leidt tot suikerziekte (*A shortage of insuline causes diabetes*)

Dependency Trees



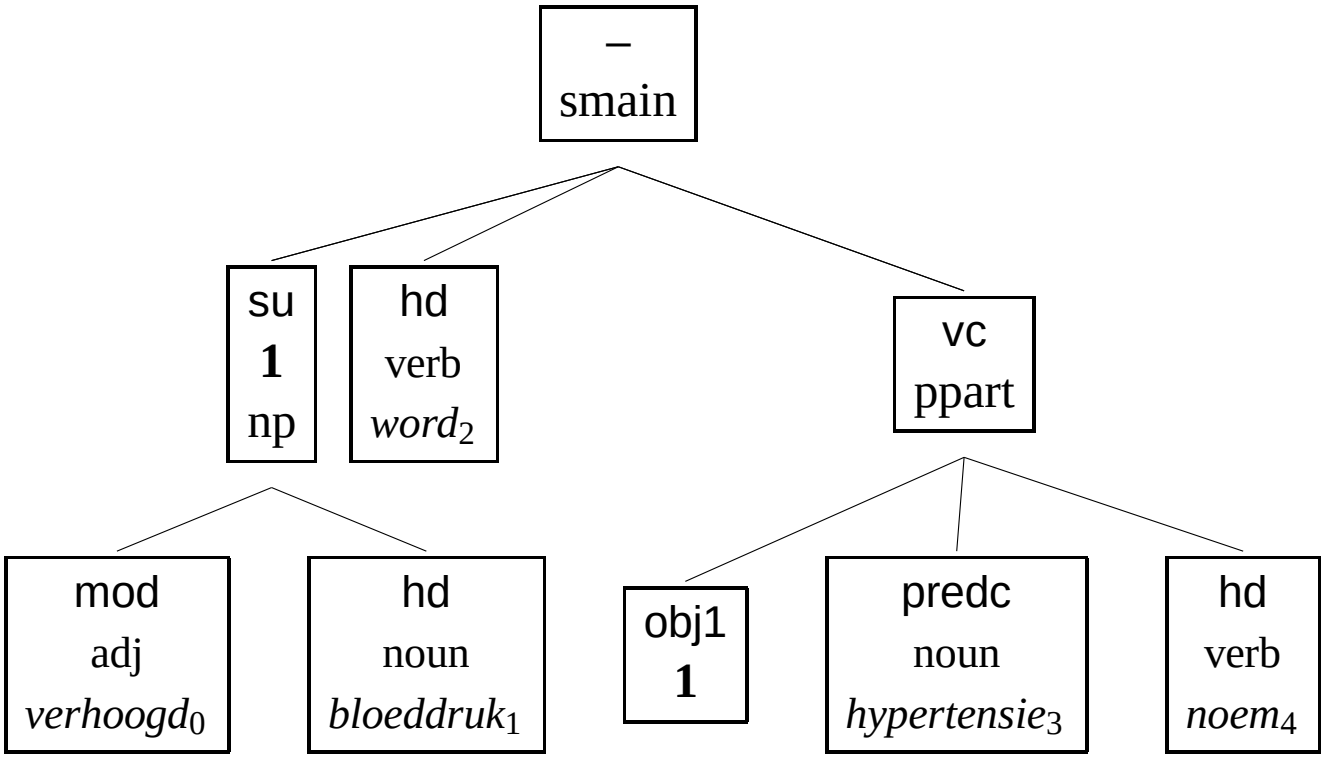
Anthony Fokker overlijdt op 49-jarige leeftijd te New York (*Anthony Fokker dies at age 49 in New York*)

Dependency Trees



Alan Turing wordt op 7 juni 1954 dood aangetroffen (*Alan Turing is found dead on June, 7th, 1954*)

Dependency Trees



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 pavenders))
      (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)))))))))
   (. .)))
```

Stanford Parser Dependencies

Dependency Relations

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)
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)          39 dobj(won, medal)
 69 dobj(made, debut)          39 dobj(changed, name)
 57 dobj(won, pole)            35 dobj(holds, people)
 47 dobj(take, place)          32 dobj(started, career)
 47 dobj(began, career)        32 dobj(expanding, it)
 45 dobj(has, population)      31 dobj(help, Wikipedia)
 44 dobj(had, population)      30 dobj(fill, vacancy)
 44 dobj(customised, stamp)    29 dobj(made, appearances)
 40 dobj(takes, place)
```

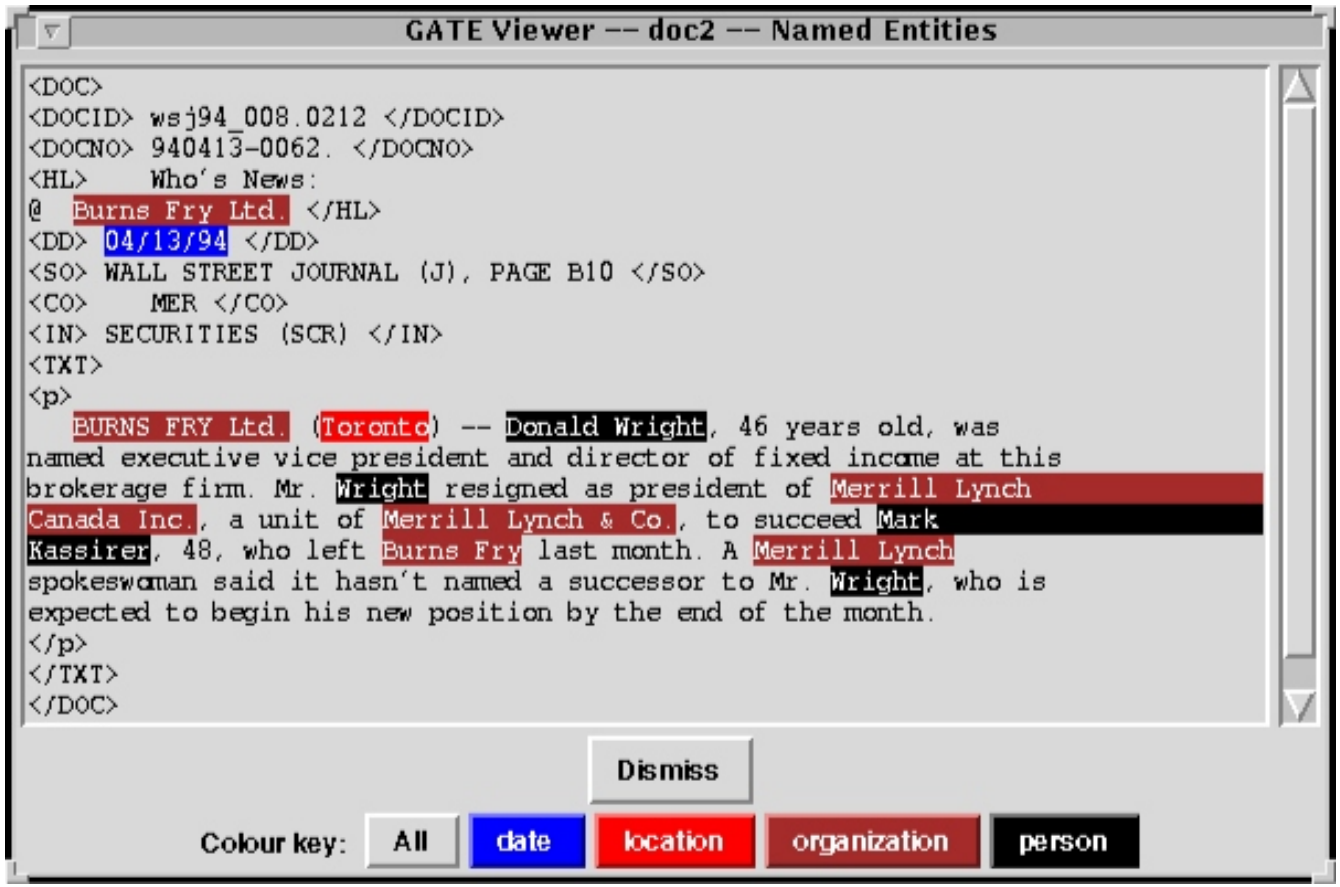
Objects of take

47	dobj(take, place)	7	dobj(take, position)
40	dobj(takes, place)	7	dobj(taken, control)
19	dobj(take, part)	7	dobj(take, it)
13	dobj(take, advantage)	6	dobj(take, care)
11	dobj(take, control)	5	dobj(take, world)
10	dobj(taken, place)	5	dobj(take, them)
10	dobj(take, him)	5	dobj(takes, time)
9	dobj(take, action)	5	dobj(takes, them)
7	dobj(takes, name)	5	dobj(takes, origin)
7	dobj(takes, control)	5	dobj(takes, it)

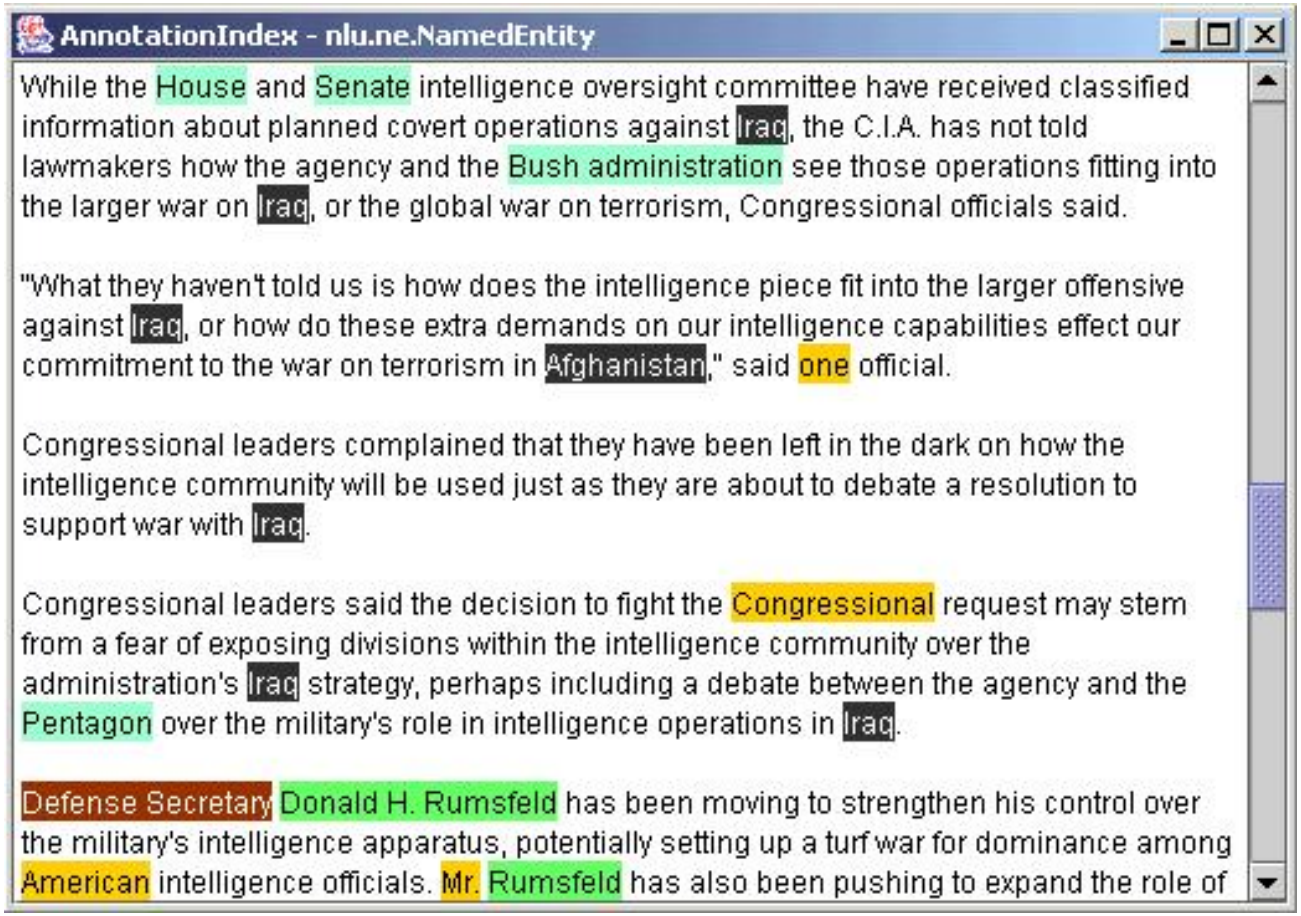
Semantic Annotation

- 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

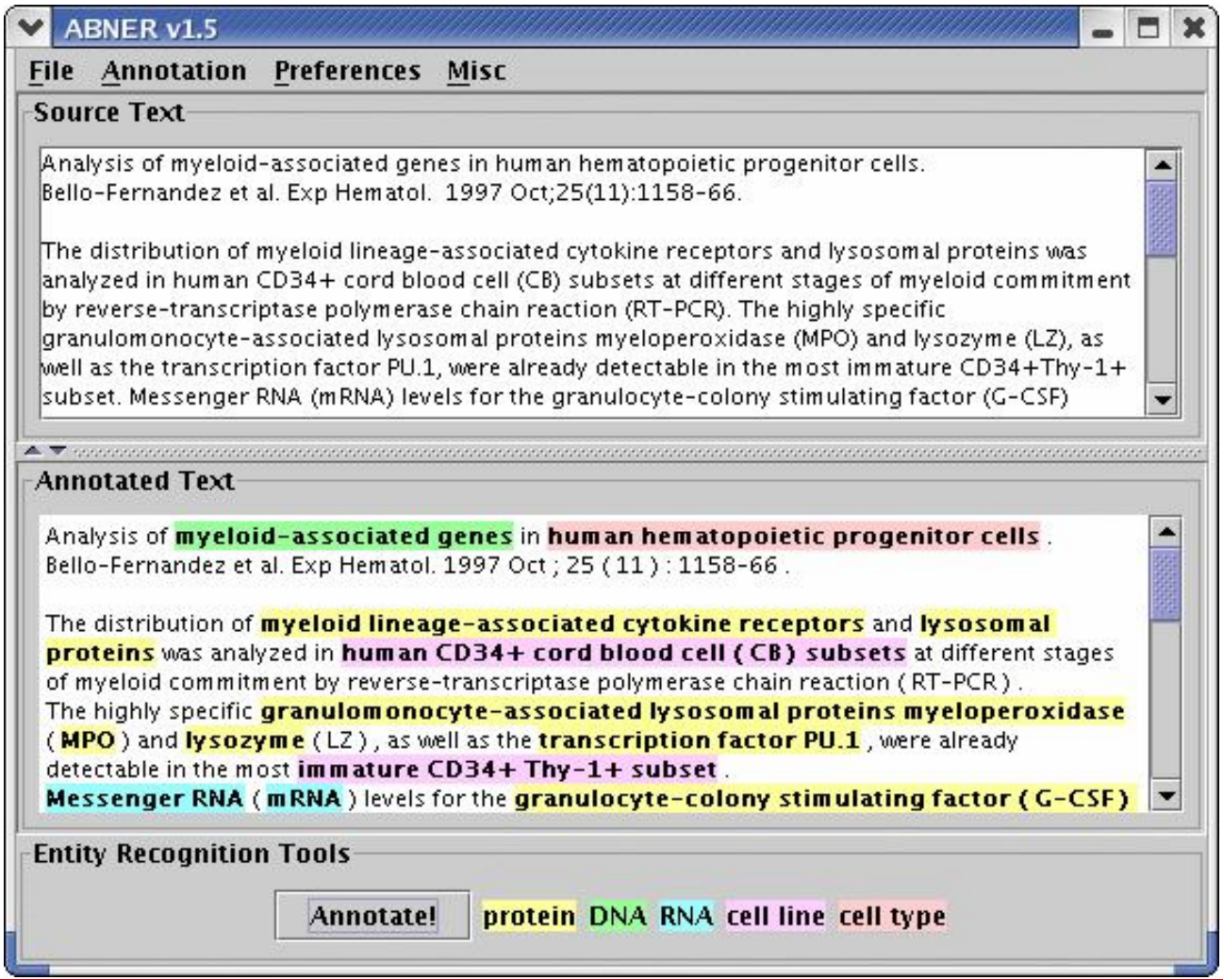
Named Entity Classes



Named Entity Classes



Named Entity Classes



Word Sense Disambiguation

- 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

Google Translate BETA

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

Google Translate BETA

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

Google Translate BETA

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**
  David Crystal, *A Dictionary of Linguistics and Phonetics*, Blackwell, 3rd Edition, 1991.
- **A collection of naturally occurring language text, chosen to characterize a state or variety of a language**
  John Sinclair, *Corpus, Concordance, Collocation*, OUP, 1991

  (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**
  David Crystal, *A Dictionary of Linguistics and Phonetics*, Blackwell, 3rd Edition, 1991.
- **A collection of naturally occurring language text, chosen to characterize a state or variety of a language**
  John Sinclair, *Corpus, Concordance, Collocation*, OUP, 1991

  (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

- Brown Corpus: ± 1 M words, consisting of samples of ± 2000 words (H. Kučera & W. Francis: *Computational Analysis of Present-Day American English* Brown University Press, 1967).
- LOB (London Oslo Bergen) Corpus: ca. 1 M words of British English, consisting of samples of ca. 2000 words (K. Hofland & S. Johansson: *Word Frequencies in British and American English* Norwegian Computing Centre for the Humanities 1982).

Dutch

- Eindhoven Corpus, almost 1 M words (P.C. Uit den Boogaart (ed.): *Woordfrequenties in geschreven en gesproken Nederlands* Oosthoek, Scheltema & Holkema 1975).

Examples (old but still used widely)

English

- Brown Corpus: $\pm$ 1 M words, consisting of samples of $\pm$ 2000 words (H. Kučera & W. Francis: *Computational Analysis of Present-Day American English* Brown University Press, 1967).
- LOB (London Oslo Bergen) Corpus: ca. 1 M words of British English, consisting of samples of ca. 2000 words (K. Hofland & S. Johansson: *Word Frequencies in British and American English* Norwegian Computing Centre for the Humanities 1982).

Dutch

- Eindhoven Corpus, almost 1 M words (P.C. Uit den Boogaart (ed.): *Woordfrequenties in geschreven en gesproken Nederlands* Oosthoek, Scheltema & Holkema 1975).

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 labou
\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 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

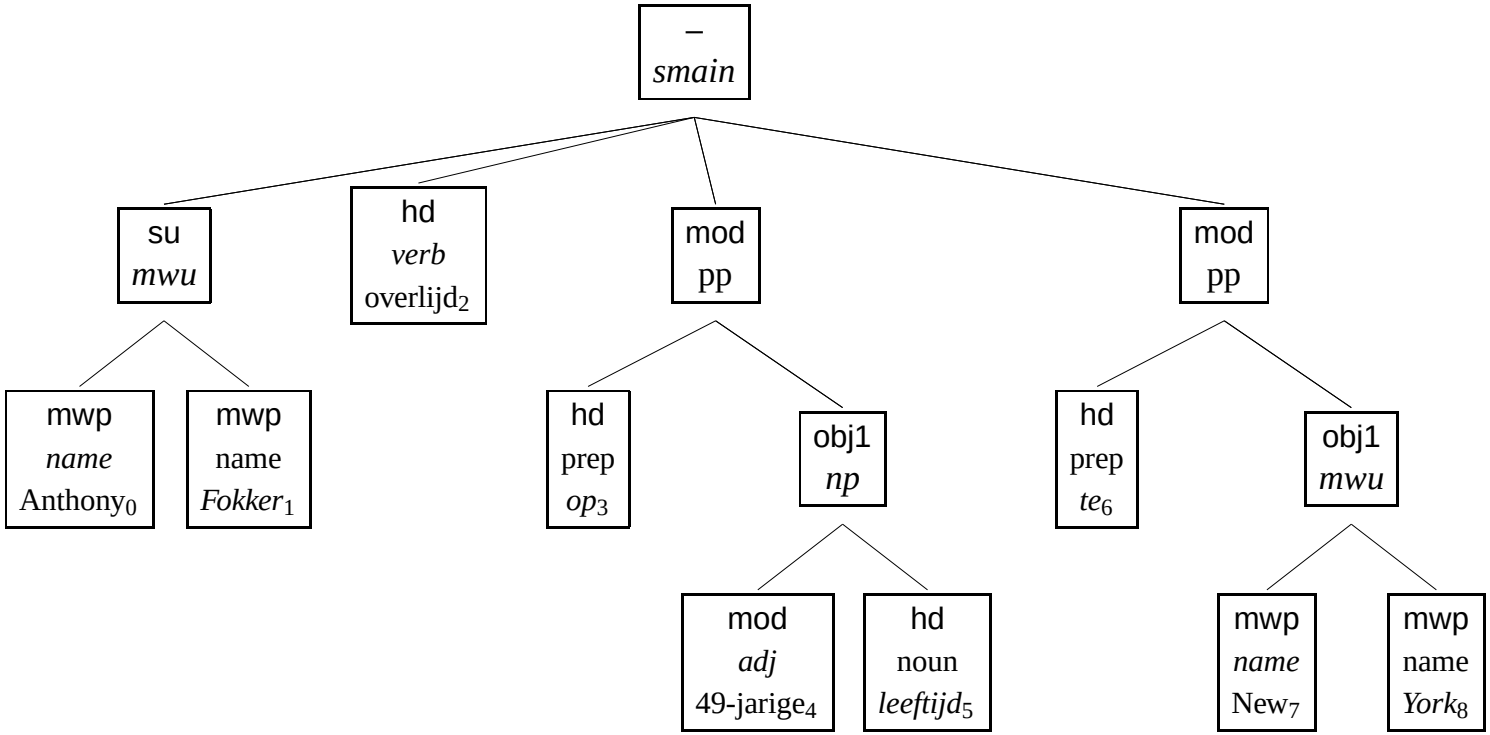
- Institut für Deutsche Sprache
- eines Korpus der Gegenwartssprache von ca. 1,6 Milliarden Textwörtern

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

How many words are there in this text?

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.
Sinds 1998 meet Intraval in opdracht van de gemeente jaarlijks de temperatuur van de binnenstad. Winkeliers, bezoekers overdag, bewoners, horecaondernemers en bezoekers 's avonds krijgen vragenlijsten voorgelegd. In de categorie bezoekers overdag, winkeliers en bewoners zegt 87 procent zich nooit onveilig te voelen in het centrum, 90 procent vindt de binnenstad gezellig. In 1998 lagen deze percentages op 94 en 87.
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

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

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

Tokens and Types

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

11	de	3	het
8	en	3	1998
7	in	2	winkeliers
6	binnenstad	2	voelen
5	procent	2	vindt
4	van	2	veilig
4	bezoekers	2	...
3	zich	1	zijn
3	intraval	1	zegt
		1	zeer

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

# Tokens (x 1000)	Wikipedia Types	TT ratio
100	17.360	0.173
200	27.775	0.138
300	37.656	0.125
400	47.721	0.119
500	55.227	0.110
600	61.644	0.103
700	70.535	0.101
800	76.014	0.095
900	82.488	0.092
1000	87.954	0.087

Type/Token Ratio decreases as text size increases

Type/Token Ratio

TT Ratio for increasing text sizes

# Tokens (x 1000)	Wikipedia	
	Types	TT ratio
100	17.360	0.173
200	27.775	0.138
300	37.656	0.125
400	47.721	0.119
500	55.227	0.110
600	61.644	0.103
700	70.535	0.101
800	76.014	0.095
900	82.488	0.092
1000	87.954	0.087

Type/Token Ratio decreases as text size increases

Type/Token Ratio

Wikipedia vs Newspaper (AD 1999)

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

There is more repetition (less variation) in AD then in Wikipedia

Type/Token Ratio

Wikipedia vs Newspaper (AD 1999)

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

There is more repetition (less variation) in AD then in Wikipedia

Most Frequent Word in Dickens, A Tale of Two Cities

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

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

Word Frequencies

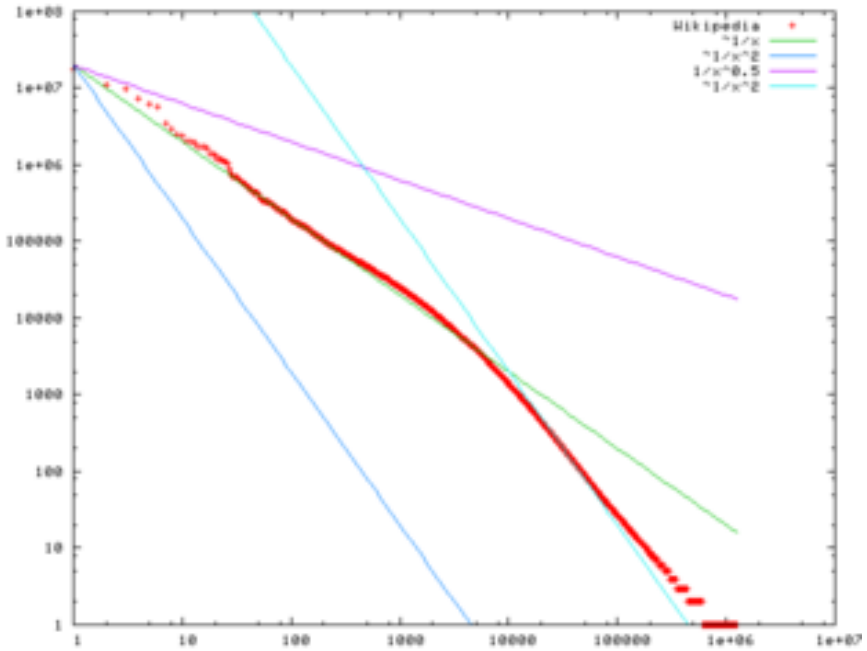
- Few words very frequent (the, a, and, in, on, that, ...)
- Many low-frequency words

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.

$$freq(W_R) = \alpha \frac{freq(W_1)}{R}$$

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 occurences. 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

	O&vdW	G&N	Google
Dutch	150M	622M	
English		47.2B	1.024B

Current Size of the Dutch Web

Frequency of *eens (once)*

Corpus	Size	Eens	Freq
INL	720k	966	1/730
Wikipedia	58M	8064	1/6250
AD 1999	14.1M	9249	1/1566
Average ?			1/1000

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

Engine	Count	Websize
Google (domain NL)	2.0M	2B
Google (lg Dutch)	2.3M	2.3B
Ilse	5.5M	5.5B
Google	48.0M	48.0B
Yahoo (lg Dutch)	106.0M	106.0B
Yahoo	110.0M	110.0B

Current Size of the Dutch Web

Frequency of *eens (once)*

Corpus	Size	Eens	Freq
INL	720k	966	1/730
Wikipedia	58M	8064	1/6250
AD 1999	14.1M	9249	1/1566
Average ?			1/1000

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

Engine	Count	Websize
Google (domain NL)	2.0M	2B
Google (lg Dutch)	2.3M	2.3B
Ilse	5.5M	5.5B
Google	48.0M	48.0B
Yahoo (lg Dutch)	106.0M	106.0B
Yahoo	110.0M	110.0B

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

+s	count	-s	count	English
weer**s**voorspelling	295K	weervoorspelling	125K	*wheather forecast*
spelling**s**regel	1020	spellingregel	909	*spelling rule*
besturing**s**systeem	1.25M	besturingsysteem	108K	*operating system*
dood**s**kist	57.7K	doodkist	6.9K	*coffin*
drug**s**beleid	87K	drugbeleid	10K	*drugs policy*
moeder**s**dag	8K	moederdag	700K	*mothersday*

Using Web Counts

Dialectal Variation

	NL	BE
eens (once)	2M	753K
alweer (again)	603	282K
weeral (again)	71K	267K
vast en zeker (for sure)	263K	68K
zeker en vast (for sure)	65K	174K
nootmuskaat (nutmeg)	83K	20K
muskaatnoot (nutmeg)	606	17K

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