

The Computer and Natural Language (Ling 445/515)

Introduction to Corpus Linguistics

Markus Dickinson
Dept. of Linguistics, Indiana
Autumn 2008

Computers and
Language
Introduction to
Corpus Linguistics

Semantic Web
Semantic annotation

Corpora
Linguistics
Lexicography
Language Learning
Types

Corpus Annotation
Encoding
Linguistic Annotation

Lemmatization

POS Tagging
Tokenization
Tagset
Automatic Tagging

Parsing

1 / 56

Improving Searching

What do we need?

- ▶ Better understanding of the user's query (i.e., its meaning/semantics)
- ▶ Better understanding of a webpage's content
- ▶ More uniform way of representing data
 - ▶ Two webpages could have tables of the same information (e.g., weather data) stored differently

We'll look in a very cursory way at how the Semantic Web handles some of these problems

- ▶ This will lead into corpus annotation & corpus linguistics

Computers and
Language
Introduction to
Corpus Linguistics

Semantic Web
Semantic annotation

Corpora
Linguistics
Lexicography
Language Learning
Types

Corpus Annotation
Encoding
Linguistic Annotation

Lemmatization

POS Tagging
Tokenization
Tagset
Automatic Tagging

Parsing

2 / 56

The Semantic Web

The semantic web tries to link data on a global scale, such that the data is usable for everyone.

- ▶ The content of a webpage needs to be able to be processed by computers in a meaningful way.
- ▶ Information needs to be given explicit meaning ... and natural language is not explicit enough.
- ▶ A **Resource Description Framework (RDF)** provides a way to organize data
 - ▶ Real-world data—elements with unique identifiers (e.g., webpages)—have relations among one another
 - ▶ From these relations, can make inferences

Computers and
Language
Introduction to
Corpus Linguistics

Semantic Web
Semantic annotation

Corpora
Linguistics
Lexicography
Language Learning
Types

Corpus Annotation
Encoding
Linguistic Annotation

Lemmatization

POS Tagging
Tokenization
Tagset
Automatic Tagging

Parsing

3 / 56

Relations among data

RDFs are graph-based, meaning that we connect different **nodes** (concepts) with **links** (relations)

- ▶ A webpage is annotated with a “class” described as a *clown*
- ▶ A different webpage specifies that a *comic* is a “subclass” of *clown*
- ▶ And a final webpage states that *JohnSmith* is a “type” of *comic*

We can thus infer that *JohnSmith* is a type of *clown*

<http://www.w3.org/TR/2002/WD-rdf-concepts-20021108/>

Computers and
Language
Introduction to
Corpus Linguistics

Semantic Web
Semantic annotation

Corpora
Linguistics
Lexicography
Language Learning
Types

Corpus Annotation
Encoding
Linguistic Annotation

Lemmatization

POS Tagging
Tokenization
Tagset
Automatic Tagging

Parsing

4 / 56

Better searching: ontologies!

What we're describing above is a hierarchy of data; the OWL Web Ontology Language makes specific what kinds of relations are allowed

An **ontology** specifies the concepts within a world and their relationships. For example:

- ▶ Class
- ▶ Individual
- ▶ equivalentClass
- ▶ inverseOf
- ▶ Restriction

Computers and
Language
Introduction to
Corpus Linguistics

Semantic Web
Semantic annotation

Corpora
Linguistics
Lexicography
Language Learning
Types

Corpus Annotation
Encoding
Linguistic Annotation

Lemmatization

POS Tagging
Tokenization
Tagset
Automatic Tagging

Parsing

5 / 56

Towards semantic annotation

Now that we have webpages with more refined annotation (at least in theory), we can state the relations among them

- ▶ If a webpage lists a Ford show, a search for car shows should link the two up.

(We'll skip all the details here ...)

But who comes up with the semantic annotation in the first place?

- ▶ We need some way to say what the annotation should be
- ▶ We need somebody to mark up a text with the appropriate annotation

Computers and
Language
Introduction to
Corpus Linguistics

Semantic Web
Semantic annotation

Corpora
Linguistics
Lexicography
Language Learning
Types

Corpus Annotation
Encoding
Linguistic Annotation

Lemmatization

POS Tagging
Tokenization
Tagset
Automatic Tagging

Parsing

6 / 56

Ontologies and Linguistic Data Sources

These aren't new questions: linguists have been asking for quite some time how to

- ▶ And here's where we're going to leave the Semantic Web and investigate linguistic annotation

For example, WordNet (<http://wordnet.princeton.edu/>) describes semantic relations between words in a cognitively plausible manner

- ▶ We'll examine its hierarchical structure when we talk about semantics for machine translation

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Semantic annotation
- Corpora
- Linguistics
- Lexicography
- Language Learning
- Types
- Corpus Annotation
- Encoding
- Linguistic Annotation
- Lemmatization
- POS Tagging
- Tokenization
- Tagset
- Automatic Tagging
- Parsing

Databases and Corpora

Resources like WordNet are nice because they provide a database capturing properties of language

- ▶ But databases only say so much
- ▶ We often need to see how the data works in the real world.
 - ▶ e.g., knowing that *jalopy* is a type of *car* tells me very little about how people use that word.

So, we turn to **corpora** ...

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Semantic annotation
- Corpora
- Linguistics
- Lexicography
- Language Learning
- Types
- Corpus Annotation
- Encoding
- Linguistic Annotation
- Lemmatization
- POS Tagging
- Tokenization
- Tagset
- Automatic Tagging
- Parsing

What is a Corpus?

CORPUS:

- (1) A collection of texts, especially if complete and self-contained: the corpus of Anglo-Saxon verse.
- (2) In linguistics and lexicography, a body of texts, utterances, or other specimens considered more or less representative of a language, and usually stored as an electronic database.

Currently, computer corpora may store many millions of running words

(from *The Oxford Companion to the English Language*, ed. McArthur & McArthur, 1992)

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Semantic annotation
- Corpora
- Linguistics
- Lexicography
- Language Learning
- Types
- Corpus Annotation
- Encoding
- Linguistic Annotation
- Lemmatization
- POS Tagging
- Tokenization
- Tagset
- Automatic Tagging
- Parsing

Why Are Electronic Corpora Useful?

- ▶ as a collection of examples for linguists
- ▶ as a data resource for lexicographers
- ▶ as instruction material for language teachers and learners
- ▶ as training material for natural language processing applications
 - ▶ training of speech recognizers
 - ▶ training of statistical part-of-speech taggers and parsers
 - ▶ training of example-based and statistical machine translation systems

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Semantic annotation
- Corpora
- Linguistics
- Lexicography
- Language Learning
- Types
- Corpus Annotation
- Encoding
- Linguistic Annotation
- Lemmatization
- POS Tagging
- Tokenization
- Tagset
- Automatic Tagging
- Parsing

Bad Start for Corpus Linguistics

Noam Chomsky (1957) *Syntactic Structures*:

- ▶ p. 15: "... it is obvious that the set of grammatical sentences cannot be identified with any particular corpus of utterances ...
- ... a grammar mirrors the behavior of the speaker, who, on the basis of a finite and accidental experience with language, can produce or understand an indefinite number of new sentences."

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Semantic annotation
- Corpora
- Linguistics
- Lexicography
- Language Learning
- Types
- Corpus Annotation
- Encoding
- Linguistic Annotation
- Lemmatization
- POS Tagging
- Tokenization
- Tagset
- Automatic Tagging
- Parsing

Bad Start for Corpus Linguistics (2)

Noam Chomsky (1957) *Syntactic Structures*:

- ▶ p. 16/17: "... one's ability to produce and recognize grammatical utterances is not based on notions of statistical approximations or the like.
- ... If we rank the sequences of a given length in order of statistical approximation to English, we will find both grammatical and ungrammatical sequences scattered throughout the list; there appears to be no particular relation between the order of approximations and grammaticality."

What that means for us is that we must approach corpora with care, but corpora have proven themselves highly useful ...

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Semantic annotation
- Corpora
- Linguistics
- Lexicography
- Language Learning
- Types
- Corpus Annotation
- Encoding
- Linguistic Annotation
- Lemmatization
- POS Tagging
- Tokenization
- Tagset
- Automatic Tagging
- Parsing

Examples for Linguists

Examples for English noun phrases from the Penn treebank:

- ▶ USX's transition from Big Steel to Big Oil
- ▶ Pittsburgh instead of New York or Findlay, Ohio, Marathon 's home
- ▶ his concern about boosting shareholder value
- ▶ the modest goal of becoming tax manager by the age of 46
- ▶ a move that, in effect, raised the cost of a \$7.19 billion Icahn bid by about \$3 billion

Computers and Language
Introduction to Corpus Linguistics

Semantic Web
Semantic annotation

Corpora

Linguistics

Lexicography
Language Learning
Types

Corpus Annotation

Encoding
Linguistic Annotation

Lemmatization

POS Tagging

Tokenization
Tagset

Automatic Tagging

Parsing

13/56

Examples for Linguists (2)

- ▶ an undistinguished college student who dabbled in zoology until he concluded that he couldn't stand cutting up frogs
- ▶ the sale of the reserves of Texas Oil & Gas, which was acquired three years ago and hasn't posted any significant operating profits since
- ▶ not just its reserves of about 1.2 trillion cubic feet of natural gas and 28 million barrels of oil but also its pipeline, gas-gathering and contract-drilling operations

Computers and Language
Introduction to Corpus Linguistics

Semantic Web
Semantic annotation

Corpora

Linguistics

Lexicography
Language Learning
Types

Corpus Annotation

Encoding
Linguistic Annotation

Lemmatization

POS Tagging

Tokenization
Tagset

Automatic Tagging

Parsing

14/56

Data for Lexicographers

How many senses does the word *line* have?

14 (according to Webster's New Encyclopedic Dictionary, 1994):

1. a comparatively strong slender cord
2. a cord, wire, or tape used in measuring and leveling
3. piping for conveying a fluid
4. a row of words, letters, numbers or symbols that are written, printed, or displayed
5. something that is distinct, elongated, and narrow
6. a state of agreement (bring ideas into line)
7. a course of conduct, action, or thought (a political line)
8. limit, restraint (overstep the line of good taste) . . .

Computers and Language
Introduction to Corpus Linguistics

Semantic Web
Semantic annotation

Corpora

Linguistics

Lexicography

Language Learning
Types

Corpus Annotation

Encoding
Linguistic Annotation

Lemmatization

POS Tagging

Tokenization
Tagset

Automatic Tagging

Parsing

15/56

Concordancer

- ▶ search tool for text corpora
- ▶ allows to search for occurrences of words / sequences of words
- ▶ generally displays results in KWIC format = key word in context
- ▶ i.e. target expression is centered and set off from context
- ▶ context is displayed in a fixed length

Computers and Language
Introduction to Corpus Linguistics

Semantic Web
Semantic annotation

Corpora

Linguistics

Lexicography

Language Learning
Types

Corpus Annotation

Encoding
Linguistic Annotation

Lemmatization

POS Tagging

Tokenization
Tagset

Automatic Tagging

Parsing

16/56

Concordancer

Let's play being a lexicographer: Find out how many different meanings the word *line* has.

Use: <http://ysomeya.hp.infoseek.co.jp/>

Computers and Language
Introduction to Corpus Linguistics

Semantic Web
Semantic annotation

Corpora

Linguistics

Lexicography

Language Learning
Types

Corpus Annotation

Encoding
Linguistic Annotation

Lemmatization

POS Tagging

Tokenization
Tagset

Automatic Tagging

Parsing

17/56

Instruction for Language Learning

How do you say in English: think about or think on?

If in doubt, ask google:

163,000,000 hits for think about (9/7/07)

2,080.000 hits for think on (9/7/07)

Computers and Language
Introduction to Corpus Linguistics

Semantic Web
Semantic annotation

Corpora

Linguistics

Lexicography

Language Learning
Types

Corpus Annotation

Encoding
Linguistic Annotation

Lemmatization

POS Tagging

Tokenization
Tagset

Automatic Tagging

Parsing

18/56

Types of Corpora

- ▶ mono-lingual versus multi-lingual corpora
- ▶ special-purpose, domain-specific corpora versus general-purpose, large-scale corpora
- ▶ spoken language corpora versus collections of written text
- ▶ ad-hoc corpus collections versus balanced, representative corpora
- ▶ raw text versus marked-up documents
- ▶ unannotated versus annotated corpora
- ▶ WWW as a corpus

Computers and Language
Introduction to Corpus Linguistics

Semantic Web
Semantic annotation
Corpora
Linguistics
Lexicography
Language Learning

Types
Corpus Annotation
Encoding
Linguistic Annotation
Lemmatization
POS Tagging
Tokenization
Tagset
Automatic Tagging
Parsing

19 / 56

English Corpora

- ▶ **Brown Corpus:** 1 million words of written American English texts from various genres, dating from 1961
- ▶ **Lancaster-Oslo-Bergen (LOB) Corpus:** 1 million words of written British English texts, dating from 1961. Genres are parallel to the Brown Corpus.
- ▶ **British National Corpus:** 100 mio. words of written and spoken language, balanced corpus of current British English
- ▶ **Internation Corpus of English (ICE):** national or regional varieties of English; one million word collections of contemporary spoken and written English (Great Britain, USA, Australia, South Africa, Canada, Hong Kong, India, etc.)

Computers and Language
Introduction to Corpus Linguistics

Semantic Web
Semantic annotation
Corpora
Linguistics
Lexicography
Language Learning

Types
Corpus Annotation
Encoding
Linguistic Annotation
Lemmatization
POS Tagging
Tokenization
Tagset
Automatic Tagging
Parsing

20 / 56

Non-English Corpora

- ▶ **IPI PAN Polish Corpus:** 300 mio. words
- ▶ **Czech National Corpus:** 100 mio. words
- ▶ **Hungarian National Corpus:** 80 mio. words
- ▶ **Croatian National Corpus:** 30 mio. words
- ▶ **Hellenic National Corpus:** 20 mio. words
- ▶ **METU Turkish Corpus:** 10 mio. words
- ▶ **Sinica Corpus:** 5 mio. words
- ▶ ...

Computers and Language
Introduction to Corpus Linguistics

Semantic Web
Semantic annotation
Corpora
Linguistics
Lexicography
Language Learning

Types
Corpus Annotation
Encoding
Linguistic Annotation
Lemmatization
POS Tagging
Tokenization
Tagset
Automatic Tagging
Parsing

21 / 56

Parallel Corpora

- ▶ **MULTEXT-East:** for Bulgarian, Croatian, Czech, English, Estonian, Hungarian, Lithuanian, Resian, Romanian, Russian, Slovene, and Serbian. For most languages: Orwell's 1984.
- ▶ **Hansard Corpus:** from the official records (Hansards) of the 36th Canadian Parliament [1997-2000], 3 mio. words
- ▶ **Europarl:** extracted from the proceedings of the European Parliament; includes versions in 11 European languages: Romanic (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish. Ca. 20 mio. words.

Computers and Language
Introduction to Corpus Linguistics

Semantic Web
Semantic annotation
Corpora
Linguistics
Lexicography
Language Learning

Types
Corpus Annotation
Encoding
Linguistic Annotation
Lemmatization
POS Tagging
Tokenization
Tagset
Automatic Tagging
Parsing

22 / 56

The Need for Corpus Mark-Up

Annotation guidelines are needed in order to facilitate the accessibility and reusability of corpus resources.

Minimal information:

- ▶ authorship of the source document
- ▶ authorship of the annotated document
- ▶ language of the document
- ▶ character set and character encoding used in the corpus

Computers and Language
Introduction to Corpus Linguistics

Semantic Web
Semantic annotation
Corpora
Linguistics
Lexicography
Language Learning
Types

Corpus Annotation
Encoding
Linguistic Annotation
Lemmatization
POS Tagging
Tokenization
Tagset
Automatic Tagging
Parsing

23 / 56

Text Encoding Initiative (TEI)

- ▶ project sponsored by the Association for Computational Linguistics, the Association for Literary and Linguistic Computing, and the Association for Computers in the Humanities
- ▶ encoding guidelines
- ▶ link: <http://www.tei-c.org>
- ▶ define how documents should be marked-up with the mark-up language SGML (or more recently XML)

Computers and Language
Introduction to Corpus Linguistics

Semantic Web
Semantic annotation
Corpora
Linguistics
Lexicography
Language Learning
Types

Corpus Annotation
Encoding
Linguistic Annotation
Lemmatization
POS Tagging
Tokenization
Tagset
Automatic Tagging
Parsing

24 / 56

XML

- ▶ XML: Extensible Markup Language
 - ▶ Used to structure data
 - ▶ similar to HTML
- ▶ has no fixed “semantics”: user defines what tags mean
 - ▶ formally verifiable via document type definitions (DTD) of XML Schema Definitions (XSD)
- ▶ recognized as international ISO standard
- ▶ tools available for editing, displaying, querying

25 / 56

XML – Example

```
<CATALOG>
  <CD>
    <TITLE>Empire Burlesque</TITLE>
    <ARTIST>Bob Dylan</ARTIST>
    <COUNTRY>USA</COUNTRY>
    <COMPANY>Columbia</COMPANY>
    <PRICE>10.90</PRICE>
    <YEAR>1985</YEAR>
  </CD>
  <CD>
    <TITLE>Greatest Hits</TITLE>
    <ARTIST>Dolly Parton</ARTIST>
    <COUNTRY>USA</COUNTRY>
    <COMPANY>RCA</COMPANY>
    <PRICE>9.90</PRICE>
    <YEAR>1982</YEAR>
  </CD>
</CATALOG>
```

27 / 56

TEI Guidelines

Each text that is conform with the TEI guidelines consists of two parts– a header and the text itself.

The header contains information such as:

- ▶ author, title, and date
- ▶ the edition or publisher used in creating the machine-readable text
- ▶ information about the encoding practices adopted

28 / 56

XML Annotated Text

```
<text>
<body>
<div type="BODY">
<div type="Q">
<head>Subject: The staffing in the Commission of the European Communities
</head>
<p>Can the Commission say:</p>
<p>1. how many temporary officials are working at the Commission?</p>
<p>2. who they are and what criteria were used in selecting them?</p>
</div>
<div type="R">
<head>Answer given by <name type="PERSON"><abbr rend="TAIL-SUPER">Mr</ABBR> Cardoso e Cunha</name> on behalf of the Commission
<date>(22 September 1992)</date></head>
<p>1 and 2. The Commission will send tables showing the number of temporary staff working for the Commission directly to the Honourable Member and to Parliament's Secretariat.</p>
</div></div></body></text>
```

30 / 56

Levels of Linguistic Annotation

- ▶ morphological annotation (e.g. inflection, derivation, compounding)
- ▶ morpho-syntactic annotation: part-of-speech (POS) tagging
- ▶ syntactic annotation (e.g. named entities, phrasal chunking, full syntactic analysis)
- ▶ semantic annotation (e.g. word-sense disambiguation, anaphora and coreference resolution, information structure)
- ▶ discourse annotation (e.g. dialog turns, speech acts)

31 / 56

Why Do We Need Annotation?

- ▶ for training NLP tools
- ▶ for finding examples
 - ▶ what is the plural form of *fish*?
 - ▶ which nouns can occur as bare nouns, without a determiner?
 - ▶ are there subjectless sentences in German? – Yes, e.g. *Mir ist kalt.* (To me is cold.)
 - ▶ is it possible in English to have something between a noun and its modifying relative clause?

32 / 56

Step by Step Annotation

- ▶ tokenization
- ▶ lemmatization / morphological analysis
- ▶ part-of-speech tagging
- ▶ named-entity recognition
- ▶ partial parsing
- ▶ full syntactic parsing
- ▶ semantic and discourse processing

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Corpora
- Corpus Annotation
- Linguistic Annotation
- Lemmatization**
- POS Tagging
- Parsing

Lemmatization

- ▶ refers to the process of relating individual word forms to their citation form (lemma) by means of morphological analysis
 - ▶ e.g. *stopped* ⇒ *stop*
- ▶ provides a means to distinguish between the total number of word tokens and distinct lemmata that occur in a corpus
 - ▶ e.g. helps to find all occurrences of *buy*
- ▶ is indispensable for highly inflectional languages which have a large number of distinct word forms for a given lemma

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Corpora
- Corpus Annotation
- Linguistic Annotation
- Lemmatization**
- POS Tagging
- Parsing

Lemmatization – German Example

Lemmas in second column

wie	wie	+Adv+Wh+#lex+COWIE
wie	wie	+Conj+Coord+#lex+COWIE
wie	wie	+Conj+Subord+#lex+COWIE
sie	sie	+Pron+Pers+3P+Pl+Fem+Nom+#lex+PERSPRO
sie	sie	+Pron+Pers+3P+Sg+Fem+Nom+#lex+PERSPRO
offenbar	offenbaren	+Verb+Imp+2P+Sg+#lex+VVFIN
offenbar	offenbar	+Adj+Pos+Pred+#lex+ADJD
gedacht	gedenken	+Verb+PPast+#lex+VVPP
gedacht	dachen	+Verb+PPast+#lex+VVPP
gedacht	denken	+Verb+PPast+#lex+VVPP
hat	haben	+Verb+Indc+3P+Sg+Pres+#lex+VAFIN

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Corpora
- Corpus Annotation
- Linguistic Annotation
- Lemmatization**
- POS Tagging
- Parsing

Tools for Lemmatization

- ▶ **XEROX Morphological Analyzer:** comprehensive morphological analyzers for many languages including English, French, Dutch, German, Hungarian, Italian, Portuguese, Czech, Danish, Finnish, Norwegian, Polish, Russian, Turkish.
 - ▶ link: <http://www.xrce.xerox.com/competencies/content-analysis/toolhome.en.html>
- ▶ **Lingsoft:** morphological analyzers for English, Danish, German, Swedish, and Finnish
 - ▶ link: <http://www.lingsoft.fi/demos.html>

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Corpora
- Corpus Annotation
- Linguistic Annotation
- Lemmatization**
- POS Tagging
- Parsing

Morphological Analysis

Xerox:

half half+Adj
 half half+Adv
 half half+Noun+Sg

Lingsoft:

"<half>"
 "half" <Quant> DET PRE SG/PL @QN>
 "half" <NonMod> <Quant> PRON SG/PL
 "half" N NOM SG
 "half" ADV

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Corpora
- Corpus Annotation
- Linguistic Annotation
- Lemmatization**
- POS Tagging
- Parsing

Part of Speech Tagging

POS Tagging = Assigning word class information to words

ex: *the* *man* *bought* *a* *book*
 determiner noun verb determiner noun

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Corpora
- Corpus Annotation
- Linguistic Annotation
- Lemmatization**
- POS Tagging**
- Parsing

Linguistic Questions

- ▶ How do we divide the text into individual **word tokens**?
- ▶ How do we choose a **tagset** to represent all words?
- ▶ How do we select appropriate **tags** for individual **words**?

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Corpora
- Corpus Annotation
- POS Tagging
- Tagset
- Automatic Tagging
- Parsing

Tokenization

- Tokenization has to deal with multiwords, merged words, and compounds
- ▶ Multiwords
 - ▶ e.g., *in spite of* the firm promise
 - ▶ Merged words
 - ▶ he *couldn't* come, *dunno* == do not know
 - ▶ Compounds
 - ▶ a *hundreds-of-billions-of-yen* market
 - ▶ *dieses Gestern-fand-ich-sie-noch-dooof-heute-gehen-sie-schon-in-Ordnung-Feeling*

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Corpora
- Corpus Annotation
- POS Tagging
- Tokenization
- Tagset
- Automatic Tagging
- Parsing

Possible Solution: Layered Analysis

```
<w pos=in> in spite of</w>

<w pos=md+rb> shouldn't</w>

<w pos=jj>
  <w pos=nns> hundreds</w>
  <w pos=in> of</w>
  <w pos=nns> billions</w>
  <w pos=in> of</w>
  <w pos=nns> yen</w></w>
<w pos=nn> market</w>
```

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Corpora
- Corpus Annotation
- POS Tagging
- Tagset
- Automatic Tagging
- Parsing

Selecting a Tagset

simple: nouns, verbs, adjectives, adverbs

all	conference	rooms	are	pretty	much	booked
DT	NN	NN	VBP	RB	RB	VBN

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Corpora
- Corpus Annotation
- POS Tagging
- Tokenization
- Tagset
- Automatic Tagging
- Parsing

Issues in Selecting a Tagset

- ▶ **conciseness**: short labels better than long ones
prep ⇒ preposition
- ▶ **perspicuity**: labels that are easily interpreted are better
prep ⇒ in
- ▶ **analysability**: should be possible to decompose in different parts
vmfin: verb, modal, finite
pds: pronoun, demonstrative, substituting

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Corpora
- Corpus Annotation
- POS Tagging
- Tagset
- Automatic Tagging
- Parsing

POS Representations

Horizontal Format

I/PP will/MD then/RB maybe/RB travel/VB directly/RB on/IN to/IN Berlin/NP

Vertical Format

I	PP
will	MD
then	RB
maybe	RB
travel	VB
directly	RB
on	IN
to	IN
Berlin	NP

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Corpora
- Corpus Annotation
- POS Tagging
- Tokenization
- Tagset
- Automatic Tagging
- Parsing

Tagset Size

- ▶ English:
 - TOSCA 32
 - Penn treebank 36
 - BNC C5 61
 - Brown 77
 - LOB 132
 - London-Lund Corpus 197
 - TOSCA-ICE 270
- ▶ Romanian: 614
- ▶ Hungarian: ca. 2 100

Computers and Language

Introduction to Corpus Linguistics

Semantic Web
Semantic annotation

Corpora
Linguistics
Lexicography
Language Learning
Types

Corpus Annotation
Encoding
Linguistic Annotation

Lemmatization

POS Tagging
Tokenization

Tagset

Automatic Tagging

Parsing

48 / 56

Penn Treebank Tagset

CC	Coord. conjunction	RB	Adverb
CD	Cardinal number	RBR	Adverb, comparative
DT	Determiner	RBS	Adverb, superlative
EX	Existential there	RP	Particle
FW	Foreign word	SYM	Symbol
IN	Prep. / subord. conj.	TO	to
JJ	Adjective	UH	Interjection
JJR	Adjective, comparative	VB	Verb, base form
JJS	Adjective, superlative	VBD	Verb, past tense
LS	List item marker	VBG	Verb, gerund / present part
MD	Modal	VBN	Verb, past part.
NN	Noun, singular or mass	VBP	Verb, non-3rd p., sing. pres.
NNS	Noun, plural	VBZ	Verb, 3rd p. sing. pres.
NP	Proper noun, singular	WDT	Wh-determiner
NPS	Proper noun, plural	WP	Wh-pronoun
PDT	Predeterminer	WP	Possessive wh-pronoun
POS	Possessive ending	WRB	Wh-adverb
PRP	Personal pronoun	,	Comma
PRP\$	Possessive pronoun	.	Sentence-final punctuation

Computers and Language

Introduction to Corpus Linguistics

Semantic Web
Semantic annotation

Corpora
Linguistics
Lexicography
Language Learning
Types

Corpus Annotation
Encoding
Linguistic Annotation

Lemmatization

POS Tagging
Tokenization

Tagset

Automatic Tagging

Parsing

49 / 56

Annotating POS Tags

two fundamentally different approaches:

- ▶ start from scratch, find characteristics in words or context (= rules) which give indication of word class
i.e. if word ends in ``ion``, tag it as noun
- ▶ accumulate lexicon, disambiguate words with more than one tag
i.e. possible categories for ``about``: preposition, adverb, particle

Computers and Language

Introduction to Corpus Linguistics

Semantic Web
Semantic annotation

Corpora
Linguistics
Lexicography
Language Learning
Types

Corpus Annotation
Encoding
Linguistic Annotation

Lemmatization

POS Tagging
Tokenization

Tagset

Automatic Tagging

Parsing

50 / 56

Automatic POS Tagging

Assumption: local context is sufficient

examples:

- ▶ for the man: noun or verb?
- ▶ we will man: noun or verb?
- ▶ I can put: verb base form or past?
- ▶ re-cap real quick: adjective or adverb?

Computers and Language

Introduction to Corpus Linguistics

Semantic Web
Semantic annotation

Corpora
Linguistics
Lexicography
Language Learning
Types

Corpus Annotation
Encoding
Linguistic Annotation

Lemmatization

POS Tagging
Tokenization

Tagset

Automatic Tagging

Parsing

51 / 56

Bigram Tagging

- ▶ basic assumption: POS tag only depends on word itself and on the POS tag of the previous word
- ▶ use lexicon to retrieve **ambiguity class** for words
e.g. word: *beginning*, ambiguity class: [JJ, NN, VBG]
- ▶ for unknown words: use heuristics, e.g. all open class POS tags
- ▶ disambiguation: look for most likely path through possibilities

Computers and Language

Introduction to Corpus Linguistics

Semantic Web
Semantic annotation

Corpora
Linguistics
Lexicography
Language Learning
Types

Corpus Annotation
Encoding
Linguistic Annotation

Lemmatization

POS Tagging
Tokenization

Tagset

Automatic Tagging

Parsing

52 / 56

Bigram Tagging – Basic Facts

- ▶ ambiguity: 40% of word types and 70% of word tokens are ambiguous (in Brown corpus)
- ▶ accuracy of taking the most likely tag: ca. 90%!!!

Computers and Language

Introduction to Corpus Linguistics

Semantic Web
Semantic annotation

Corpora
Linguistics
Lexicography
Language Learning
Types

Corpus Annotation
Encoding
Linguistic Annotation

Lemmatization

POS Tagging
Tokenization

Tagset

Automatic Tagging

Parsing

53 / 56

Bigram Tagging – Counter-Examples

- ▶ start before
- ▶ start before the course **or** start before he is done
- ▶ real quick
- ▶ re-cap real quick **or** a real quick lunch
- ▶ barely changed
- ▶ he was barely changed **or** he barely changed his contents
- ▶ that beginning
- ▶ that beginning part **or** that beginning frightened the students **or** with that beginning early, he was forced ...

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Corpora
- Corpus Annotation
- Lemmatization
- POS Tagging
- Automatic Tagging
- Parsing

54 / 56

Available POS Taggers

- ▶ Amalgam tagger - Email: mail-in tagger platform, supports different corpora and tagsets for English
link: <http://www.comp.leeds.ac.uk/amalgam/amalgam/amalgtag3.html>
- ▶ Brill tagger: transformation-based error driven learning
link: <http://research.microsoft.com/users/brill/>
- ▶ TnT (Tags and Trigrams): Hidden Markov Model, best tagger available
link: <http://www.coli.uni-saarland.de/~thorsten/tnt/>
- ▶ TreeTagger: decision tree tagger
link: <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Corpora
- Corpus Annotation
- Lemmatization
- POS Tagging
- Automatic Tagging
- Parsing

55 / 56

Parsing

- (1) Analyzing a sentence into its constituents, identifying in greater or less detail the syntactic relations and parts of speech.
- (2) Describing a word in a sentence, identifying its part of speech, inflectional form, and syntactic function.

McArthur (1992) The Oxford Companion to the English Language.
We will discuss parsing more when we get to grammar checking.

- Computers and Language
- Introduction to Corpus Linguistics
- Semantic Web
- Corpora
- Corpus Annotation
- Lemmatization
- POS Tagging
- Parsing

56 / 56