

# Corpus Linguistics (L615)

## POS and Syntactic Annotation

Markus Dickinson

Department of Linguistics, Indiana University  
Spring 2009

# Def. Part of Speech Tagging

## Definition

Tokenization

Tagset Design

Automatic POS  
Tagging

Bigram tagging

Syntactic  
annotation

POS Tagging = Assigning word class  
information to words

ex:	<i>the</i>	<i>man</i>	<i>bought</i>	<i>a</i>	<i>book</i>
	determiner	noun	verb	determiner	noun

## Definition

Tokenization

Tagset Design

Automatic POS  
Tagging

Bigram tagging

Syntactic  
annotation

- ▶ 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**?

# Tokenization: Multiwords

*in spite of* the firm promise  
*because of* technical problems  
*bis zum* dritten Oktober

*so daß* es nicht zustande kam  
lower rents may seem surprising *given that* some  
communities . . .

agents will *look for* under-reported income  
er *kam* aus verschiedenen Gründen nicht *vorbei*

# Tokenization: Merged Words

*I don't know*  
*he couldn't come*

*am* vierten Oktober

*dunno* == do not know

*didn't* he know?  $\Rightarrow$  *did* he *not* know?

# Tokenization: Compounds

*Great Northern Nekoosa Corp.*

a daily *Chicago-Paris* flight  
an *Atlanta-based forest-products company*

a *hundreds-of-billions-of-yen* market  
diese *Rühr-mich-nicht-an* Haltung

# 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>

# Issues in Tagset Design

Corpus Linguistics

POS and Syntactic  
Annotation

Definition

Tokenization

Tagset Design

Automatic POS  
Tagging

Bigram tagging

Syntactic  
annotation

- ▶ define which words are considered multiwords: no multiwords, ditto tags, layered annotation
- ▶ describe how mergers are treated: combined tags, splitting up
- ▶ describe how compounds are treated: surface oriented, layered annotation
- ▶ can the solutions be implemented automatically

# Selecting a Tagset

simple: nouns, verbs, adjectives, adverbs

Corpus Linguistics

POS and Syntactic  
Annotation

Definition

Tokenization

Tagset Design

Automatic POS  
Tagging

Bigram tagging

Syntactic  
annotation

# Selecting a Tagset

simple: nouns, verbs, adjectives, adverbs

all conference rooms are pretty much booked

# Selecting a Tagset

simple: nouns, verbs, adjectives, adverbs

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

# Issues in Selecting a Tagset

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

## 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

# 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

# Penn Treebank Tagset

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

# Penn Treebank decisions

We'll look at some Penn Treebank (PTB) principles for the POS tagset, to help understand what POS annotation can be

- ▶ Recoverability: do not encode properties which can be automatically recovered (eliminate redundancy)
  - ▶ Lexical recoverability: merged VB (*be*), VD (*do*), VH (*have*), and VV (other verbs) into VB
  - ▶ Syntactic recoverability: IN is preposition or subordinating conjunction: tree structure tells them apart (NP or S following)
  - ▶ Note that syntactic recoverability depends upon the quality of the syntactic annotation
- ▶ Syntactic function: e.g., tag *one* as NN when used nominally; Brown corpus always tags *one* as CD
- ▶ Indeterminacy: allow for indecision of annotator: e.g., JJ|VBG
  - ▶ Certain pairs of tags are commonly indeterminate (and commonly confused)

# Penn Treebank tagging process

Corpus Linguistics

POS and Syntactic  
Annotation

Definition

Tokenization

Tagset Design

Automatic POS  
Tagging

Bigram tagging

Syntactic  
annotation

1. Tag the corpus automatically with the PARTS tagger
  - ▶ 3-4% error rate
  - ▶ Modified Brown tagset, which has to be mapped to PTB tags (introducing a little more error)
2. Manually correct, tracking common changes

Experiments show that manual correction is faster, more accurate, and more consistent than manual annotation

# Annotating POS Tags

The first step for many tagging projects is thus automatic annotation

- ▶ To help understand what the tags mean (and the limitations of searching through POS-tagged text), let's look at how automatic tagging works

Two basic 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

# Automatic POS Tagging

General assumption: local context is sufficient

Some examples where this seems to hold true:

- ▶ 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?

# 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

In some sense, this is fairly easy:

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

# 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 ...

# 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>

# Syntactic annotation

Penn Treebank

Corpus Linguistics

POS and Syntactic  
Annotation

Definition

Tokenization

Tagset Design

Automatic POS  
Tagging

Bigram tagging

Syntactic  
annotation

We'll focus more on syntactic annotation later in the semester, but look briefly at the Penn Treebank (PTB) today

- ▶ See how it relates to POS (morphosyntactic) tags
- ▶ See some design decisions

The PTB started by first using the Fidditch parser, which:

- ▶ provides exactly one analysis
- ▶ never attaches uncertain clauses
- ▶ has good coverage & is fairly accurate

The remaining task is generally to “glue” structures together

# Example tree

```
( (S
  (NP-SBJ
    (NP (NNP Pierre) (NNP Vinken) )
    (, ,)
    (ADJP
      (NP (CD 61) (NNS years) )
      (JJ old) )
    (, ,) )
  (VP (MD will)
    (VP (VB join)
      (NP (DT the) (NN board) )
      (PP-CLR (IN as)
        (NP (DT a) (JJ nonexecutive) (NN director) ))
      (NP-TMP (NNP Nov.) (CD 29) )))
  (. .) ))
```

# Skeletal syntax

To achieve fast annotation, no distinctions were added to the parser output (at least for PTB-1)

- ▶ In fact, the goal was to create a less specified structure
  - ▶ e.g., internal NP structure is not specified
- ▶ Combinations of structures provide more information: e.g., SBAR with *to* before the VP is infinitival.

Some details:

- ▶ Null elements are used to indicate (underlying) predicate-argument structure
- ▶ “While these null elements correspond more directly to entities in some grammatical theories more than others, it is not our intention to lean toward one or another theoretical view in producing our corpus.”

Q: To what extent can (syntactic) annotation be theory-neutral?

# Post-editing

Learning how to bracket the data takes significantly longer than POS tagging; particularly difficult:

- ▶ argument/adjunct distinctions
- ▶ attachment points
- ▶ coordinate structures

Process:

1. Run Fidditch parser
2. Automatically clean up Fidditch trees by removing POS tags, nonbranching lexical nodes, and some phrasal nodes
3. Modify trees
  - ▶ Attach unattached (?) constituents
  - ▶ Promote a constituent up
  - ▶ Delete or add brackets
  - ▶ Change the label

# Basic statistics

Corpus Linguistics

POS and Syntactic  
Annotation

Definition

Tokenization

Tagset Design

Automatic POS  
Tagging

Bigram tagging

Syntactic  
annotation

For the PTB-3, there are now over a million words of annotated text in the Wall Street Journal (WSJ) subcorpus

Other subcorpora (see [/Volumes/Data/Corpora/en/penn-treebankv3/parsed/mrg/](#)):

- ▶ Switchboard
- ▶ Brown
- ▶ ATIS