# Corpus Construction and Annotation

Why are annotated corpora important for computational linguists?

- training and evaluation of NLP tools
  - classification (POS, word sense)
  - parsing (syntactic structure)
  - extraction (named entity, semantic role, coreference, events)
- make it possible to search for particular linguistic phenomena

# Annotation Process

- target phenomena
- corpus selection
- annotation efficiency and consistency
- (annotation infrastructure)
- annotation evaluation

# Annotation Process: Target Phenomena

What linguistic phenomena do you want to annotate?

- Do we really need manual annotation or can this be done automatically?
- What resources and prior annotation are needed?
  - syntactic annotation often depends on POS annotation
  - semantic annotation often depends on syntactic annotation

# Annotation Process: Corpus Selection

What data do you want to annotate?

- Written or spoken data? (Transcribed spoken data?)
- Single genre or mixed genres? Representative sampling of genres?
- How much data do you need to find enough examples of your phenomena?
  - a 1-million-word corpus doesn't always contain enough occurrences of particular words for semantic role labelling or sense tagging

# Annotation Process: Efficiency and Consistency

How difficult is the annotation task?

- What kind of annotation guidelines need to be written?
  - OntoNotes verb sense annotation: 11 pages
  - Penn Treebank syntactic annotation guidelines: 300 pages
- How much training do the annotators need?
  - several weeks?
  - degree in linguistics?
- How consistent are the annotators?
  - are errors due to carelessness/fatigue or lack of clear guidelines?

# Annotation Process: Evaluation

Two perspectives:

- external: does this annotation improve performance for a certain task?
- internal: do human annotators agree with each other?
  → inter-annotator agreement (ITA)
  - simplest method: percentage of cases where annotators agree
  - more useful/meaningful approaches take into account how often annotators would have agreed by chance

# A Brief Look at Existing Corpora

There are many different types of corpora, including corpora with:

- large amounts of representative data
- syntactic annotation
- semantic annotation
  - word senses
  - semantic roles
- temporal annotation

# Focus on Large Amounts of Representative Data

- large amounts of data = no annotation or only automatic POS tagging

A few English and German corpora:

- 1960s: Brown Corpus (1 million words, American English, written)
- 1980s: Lancaster/Oslo/Bergen Corpus (1 million words, British English, written)
- British National Corpus (100 million words, 10% spoken data)
  - BNC Sampler (2 million words, more detailed hand-checked tags)
- American National Corpus (22 million words, written and spoken)
- DeReKo German Reference Corpus (4 billion* words)

# British National Corpus

Uses inline annotation: the annotation is inserted into the text.

```
<head>
<s n="1"><w NN2>Surnames <w CJC>and <w DPS>their
<w NN2>meanings
</head>
<p>
<s n="2"><w AT0>The <w NN1>study <w PRF>of <w AT0>the
<w NN2>surnames <w PRF>of <w NN0>people <w VVG>living
<w PRP-AVP>in <w AT0>a <w NN1>place <w VM0>can
<w VBI>be <w CRD>one <w PRF>of <w AT0>the <w AV0>most
<w AJ0>time-consuming<c PUN>, <w AJ0>frustrating<c PUN>,
<w VVG-AJ0>baffling<c PUN>, <w CJC>but
<w AJ0-VVG>rewarding <w NN2>activities <w AT0>a
<w NN1>researcher <w VM0>can <w VVI>undertake<c PUN>.
```

# American National Corpus

Uses standoff annotation: the text and annotation are in separate files.

- one file contains the text (with a little document structure)

```
<turn who="A" start="54.18" end="55.32" id="t1">
   <u id="t1u1">So, how are you?</u>
</turn>
```

- another file contains the linguistics annotation

```
<chunk type="utterance" xml:base="#t1u1">
 <tok xlink:href="xpointer(string-range('', 0, 2))">
    <msd>ql++++</msd>
    <base>so</base>
 </tok>
 <tok xlink:href="xpointer(string-range('', 2, 3))">
    <base>,</base>
    <msd>,+clp+++</msd>
 </tok>
```

# Treebanks: Syntactic Annotation

A few corpora with syntactic annotation:

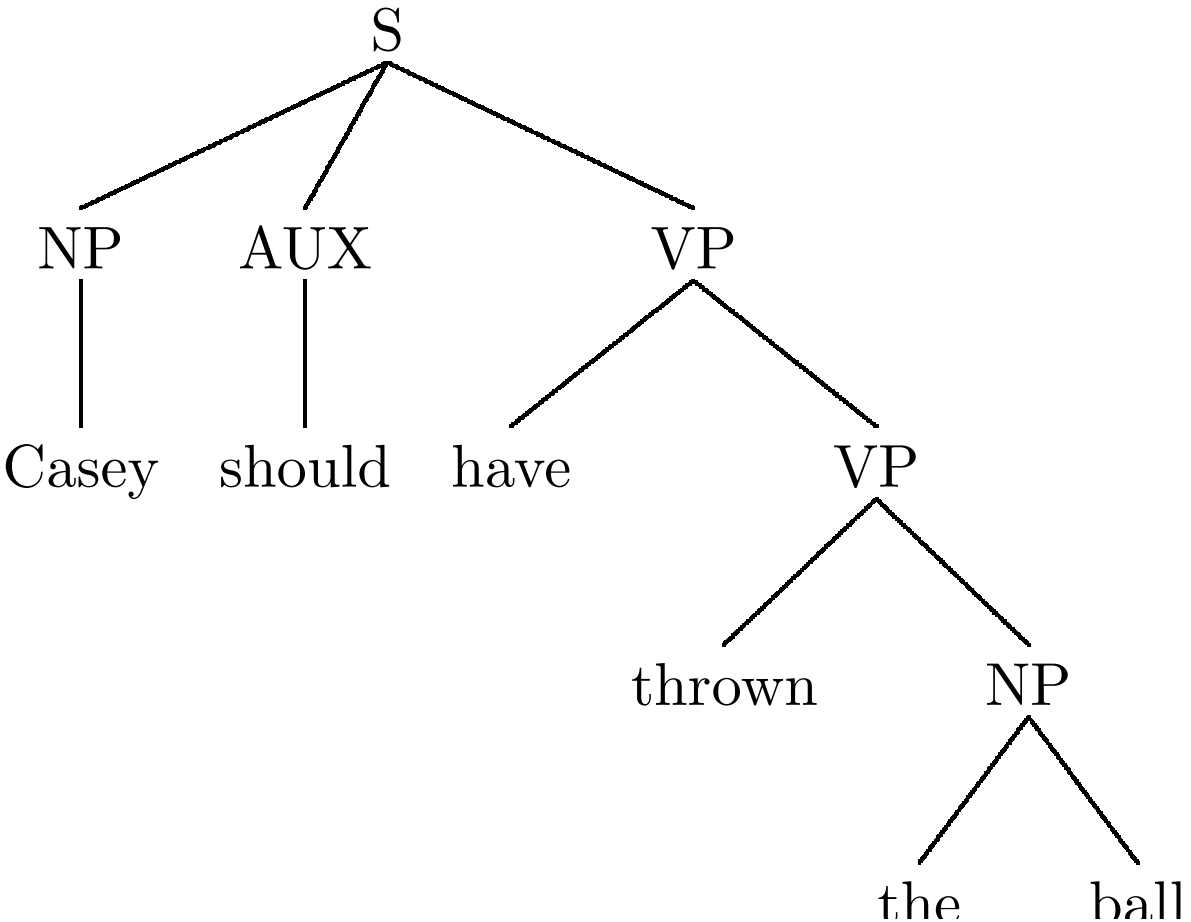
- Penn Treebank (1 million words, WSJ)
- NEGRA/TIGER (20,000/50,000 sentences)
- TueBa-D/Z (50,000 sentences)
- Prague Dependency Treebank (2 million words)

# Penn Treebank Format

Syntactic phrase structure annotated with parentheses:

```
(S (NP Casey)
   (AUX should)
   (VP have
       (VP thrown
           (NP the ball))))
```

# Penn Treebank Format: Converted to Tree



# Penn Treebank Format

Separate files for:

- raw text
- POS tags
- parsed

# Penn Treebank Format

```
The forest-products concern currently has about 38
million shares outstanding.
```

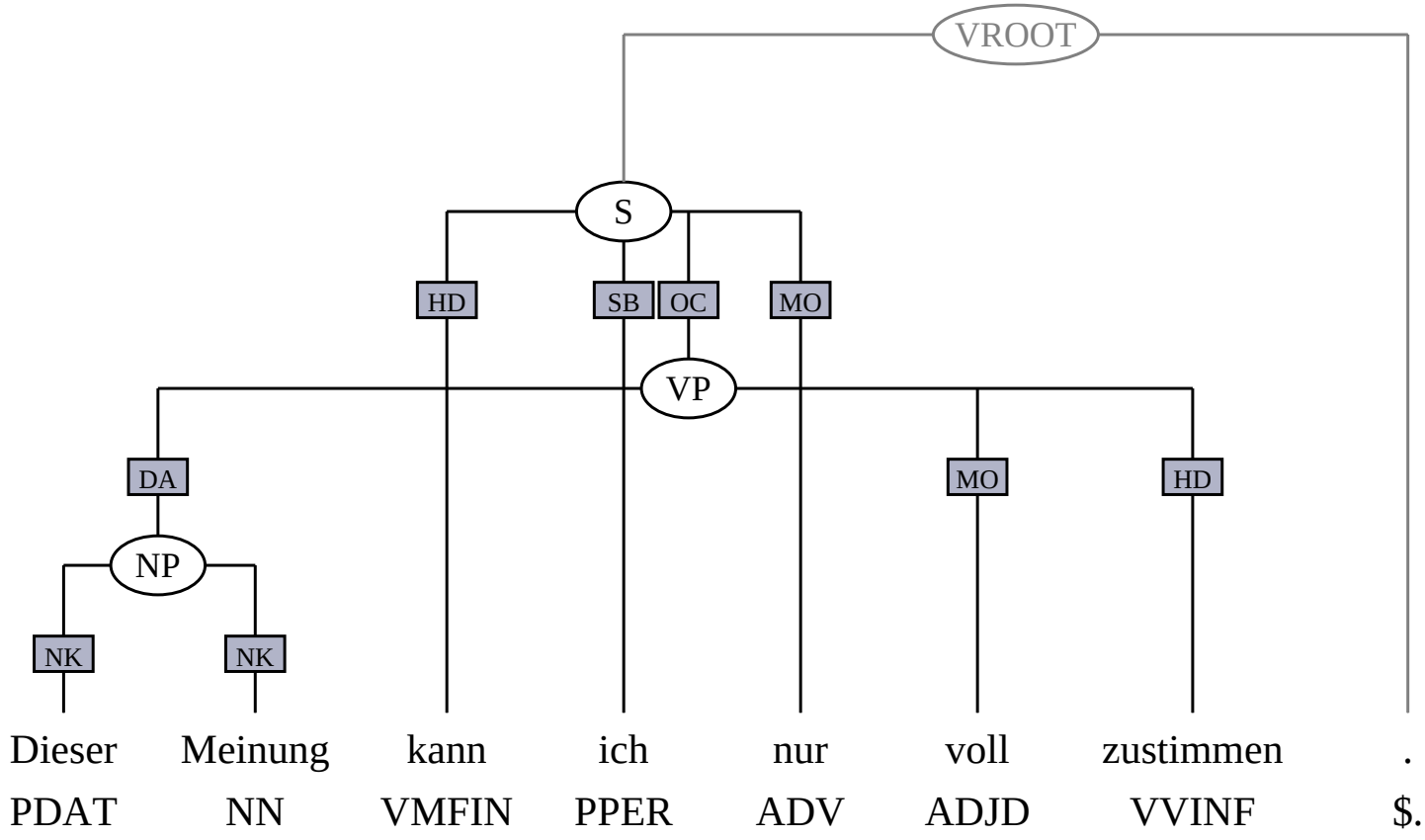
```
[ The/DT forest-products/NNS concern/NN ]
currently/RB has/VBZ about/RB
[ 38/CD million/CD shares/NNS ]
outstanding/JJ ./.
```

```
( (S (NP-SBJ The forest-products concern)
     (ADVP-TMP currently)
     (VP has
         (NP (NP (QP about 38 million) shares)
             (ADJP outstanding)))
     .))
```

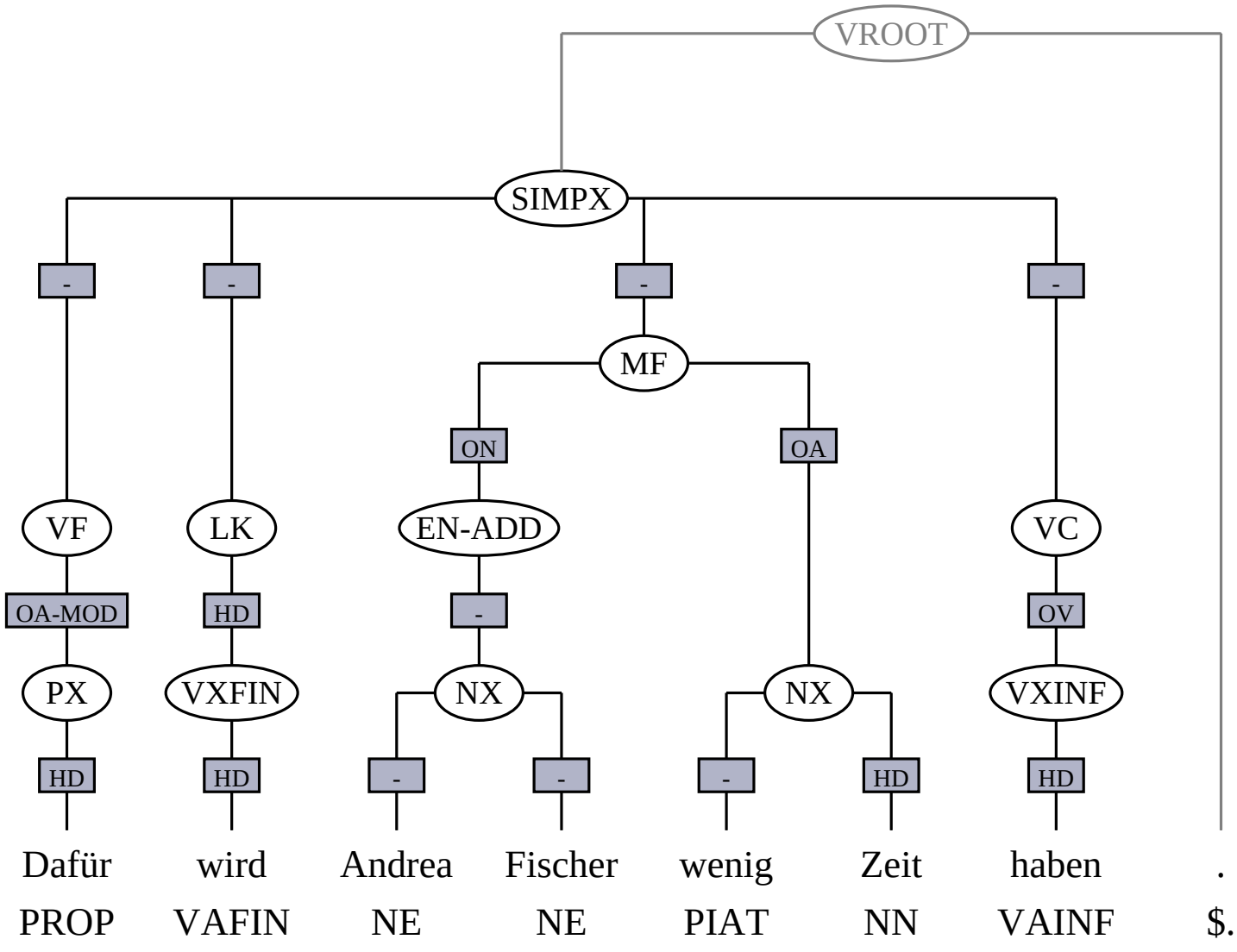
# Penn Treebank Format: Merged

```
( (S
    (NP-SBJ (DT The) (NNS forest-products) (NN concern) )
    (ADVP-TMP (RB currently) )
    (VP (VBZ has)
      (NP
        (NP
          (QP (RB about) (CD 38) (CD million) )
          (NNS shares) )
        (ADJP (JJ outstanding) )))
    (. .) ))
```

# NEGRA/TIGER Treebanks



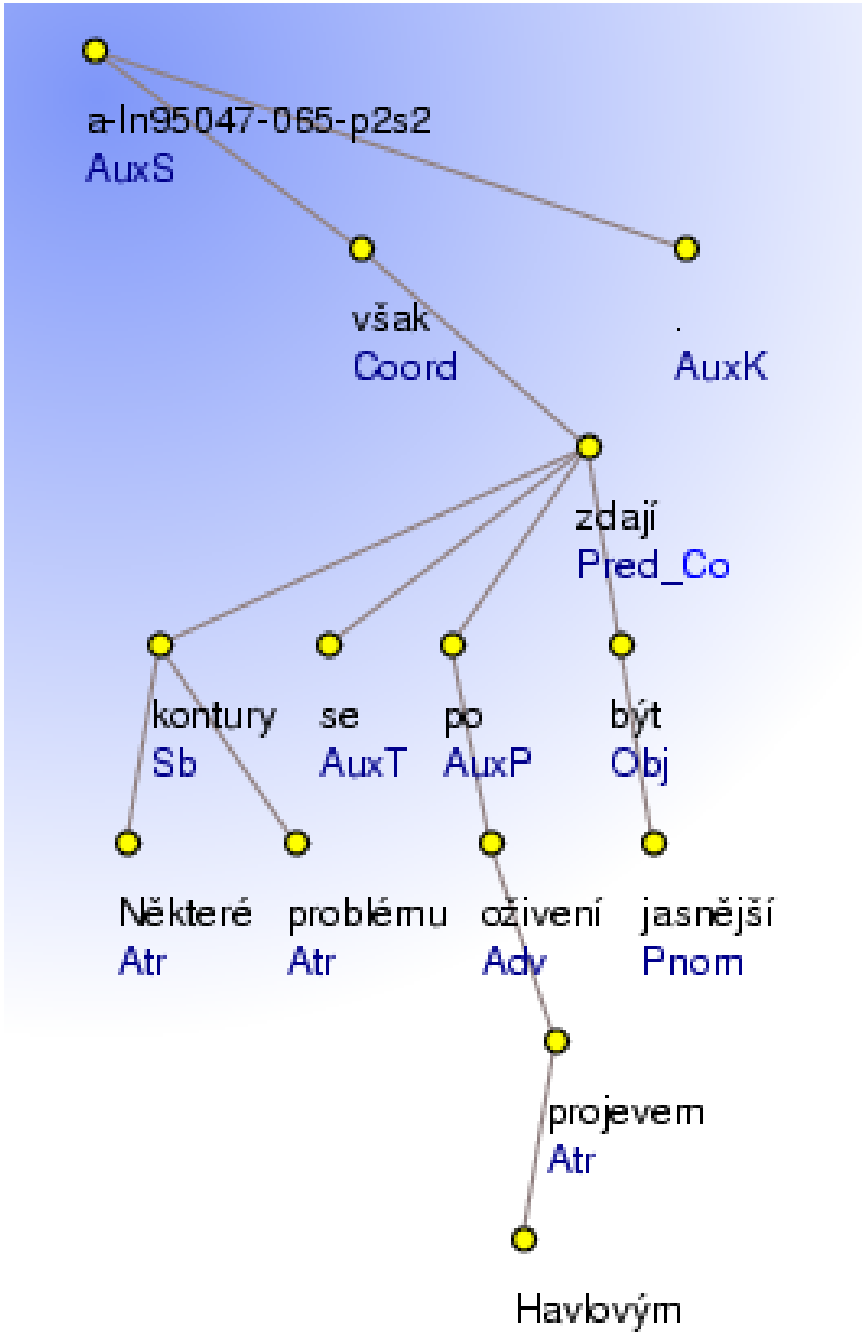
# TüBa-D/Z Treebank



# Treebank Search Tools

- token annotation: Corpus Query Processor (CQP)
- syntactic annotation: TIGERSearch

# Prague Dependency Treebank



# Semantic Annotation

A few English corpora with semantic annotation:

- word senses: SemCor (360,000 words, Brown)
- semantic relations: PropBank (113,000 verbs, Penn Treebank WSJ)

# Word Sense Annotation: Semantic Concordance

Texts from the Brown Corpus annotated with WordNet senses:

```
<s snum=1>
<wf pos=JJ wnsn=1 lexsn=3:00:02::>Most</wf>
<wf pos=NN wnsn=1 lexsn=1:04:00::>recreation</wf>
<wf pos=NN wnsn=1 lexsn=1:04:00::>work</wf>
<wf pos=VB wnsn=2 lexsn=2:42:00::>calls_for</wf>
<wf pos=DT>a</wf>
<wf pos=NN wnsn=1 lexsn=1:23:00::>good_deal</wf>
<wf pos=IN>of</wf>
<wf pos=JJ wnsn=0 lexsn=5:00:00:preceding(a):00>pre</wf>
<wf pos=NN wnsn=1 lexsn=1:04:00::>planning</wf>
<punc>.</punc>
</s>
```

# Word Sense Annotation: Semantic Concordance

<wf pos=NN wnsn=1 lexsn=1:04:00::>recreation</wf>
WordNet entry:

- S: (n) diversion#1, recreation#1 (an activity that diverts or amuses or stimulates)
  “scuba diving is provided as a diversion for tourists”
- S: (n) refreshment#2, recreation#2 (activity that refreshes and recreates; activity that renews your health and spirits by enjoyment and relaxation)
  “time for rest and refreshment by the pool”

# Semantic Roles: PropBank

Frame File for the verb *expect*:

*Roles:*
*Arg0: expecter*
*Arg1: thing expected*

Example: Transitive, active:
*Portfolio managers expect further declines in interest rates.*

```
Arg0:        Portfolio managers
REL:         expect
Arg1:        further declines in interest rates
```

Example: Transitive, passive:
*Regulatory approval is expected soon by everyone.*

```
Arg0:         everyone
REL:          is expected
Arg1:         Regulatory approval
```

# Where to Find Corpora and Linguistic Resources?

- Linguistic Data Consortium (LDC):
  `http://www.ldc.upenn.edu`
- European Language Resources Association (ELRA):
  `http://www.elra.info`
- Within SfS: `/afs/sfs/resource`