Lecture 14: Linguistic Annotation

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

- Linguistic annotation
  - Types of linguistic annotation
    - Part-of-speech
    - Syntax
  - Annotation formats
- Annotation tools

Linguistic annotation

- Why annotate text with linguistic information?
- Development and testing of linguistic theories

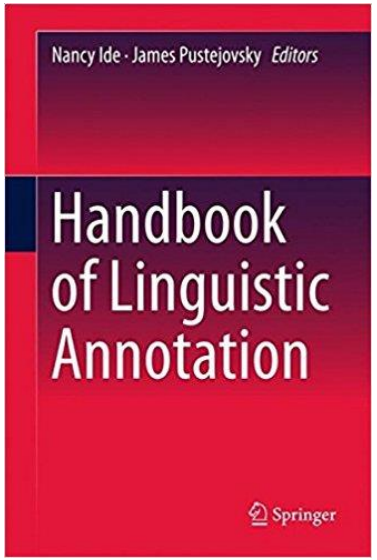
  ← Assists empirical linguistic inquiries
- Develop and evaluate (statistically based) NLP technologies

  ← Becomes the basis of "language models" in NLP applications

  ← Linguistic annotation represents linguistic knowledge of humans that AI agents learn through machine learning, which they then mimic

All about Linguistic Annotation

- *Handbook of Linguistic Annotation* (2017)
  - Nancy Ide, James Pustejovsky (eds)
  - https://link.springer.com/chapter/10.1007/978-94-024-0881-2_1
  - Offers in-depth coverage on the topic of linguistic annotation



The Brown Corpus

```
The/at Fulton/np-tl County/nn-tl Grand/jj-tl Jury/nn-tl said/vbd
Friday/nr an/at investigation/nn of/in Atlanta's/np$ recent/jj
primary/nn election/nn produced/vbd ``/`` no/at evidence/nn
''/'' that/cs any/dti irregularities/nns took/vbd place/nn ./.

The/at jury/nn further/rbr said/vbd in/in term-end/nn
presentments/nns that/cs the/at City/nn-tl Executive/jj-tl
Committee/nn-tl ,/, which/wdt had/hvd over-all/jj charge/nn
of/in the/at election/nn ,/, ``/`` deserves/vbz the/at praise/nn
and/cc thanks/nns of/in the/at City/nn-tl of/in-tl Atlanta/np-tl
''/'' for/in the/at manner/nn in/in which/wdt the/at election/nn
was/bedz conducted/vbn ./.
```

- Linguistic information: POS tag
- Tag set: The Brown Corpus Tagset
- Format: ad-hoc, embedded tags with designated delimiter

POS tagsets

- There are multiple POS tagsets in use.
  - Some are larger, some are smaller.
- **The Brown Corpus tagset** (87 tags)
  - http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM
- In NLP, **the Penn Treebank tagset** (45 tags) has become de facto standard.
  - https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
- Lately, "**Universal**" **POS tagset** is gaining grounds
  - http://universaldependencies.org/u/pos/

Universal POS tags

- "**Universal" POS tagset** is gaining grounds
  - http://universaldependencies.org/u/pos/

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

- Tags mark the core POS categories; additional grammatical properties are relegated to features
- What do you think? Truly universal?

The Penn Treebank

```
( (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) )))
    (. .) ))
( (S
    (NP-SBJ (NNP Mr.) (NNP Vinken) )
    (VP (VBZ is)
      (NP-PRD
        (NP (NN chairman) )
        (PP (IN of)
          (NP
            (NP (NNP Elsevier) (NNP N.V.) )
            (, ,)
            (NP (DT the) (NNP Dutch) (VBG publishing) (NN group) )))))
    (. .) ))
```

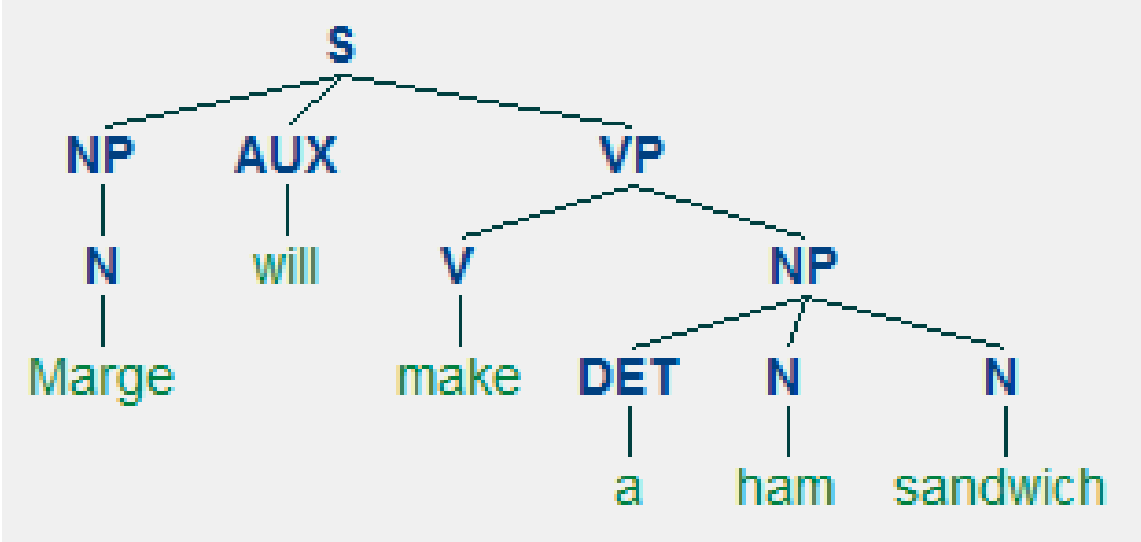
http://languagelog.ldc.upenn.edu/nll/?p=3594

Penn Treebank is based upon **phrase structure grammar** framework

Context-free grammar

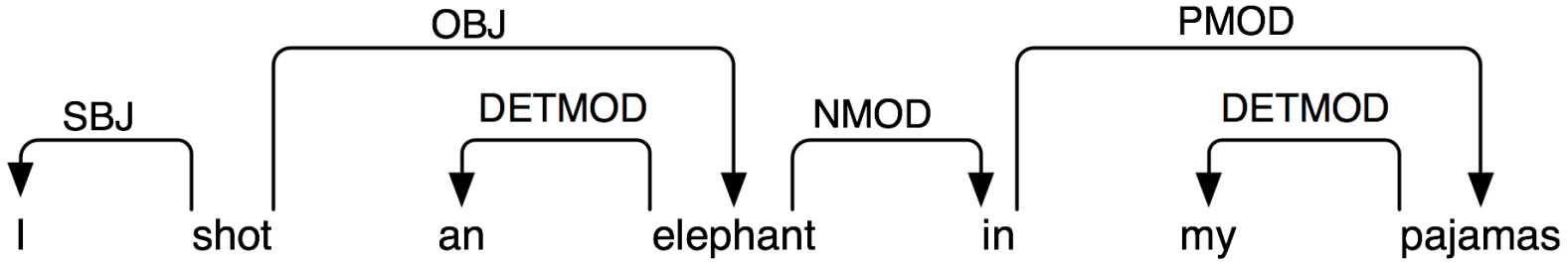
- Phrase-structure grammar is based upon constituency.
- Each local constituent can be expressed through **context-free grammar**.

```
S -> NP AUX VP
NP -> N
VP -> V NP
NP -> DET N N
N -> 'Marge'
Aux -> 'will'
V -> 'make'
DET -> 'a'
N -> 'ham' | 'sandwich'
```



A paradigm shift: dependency grammar

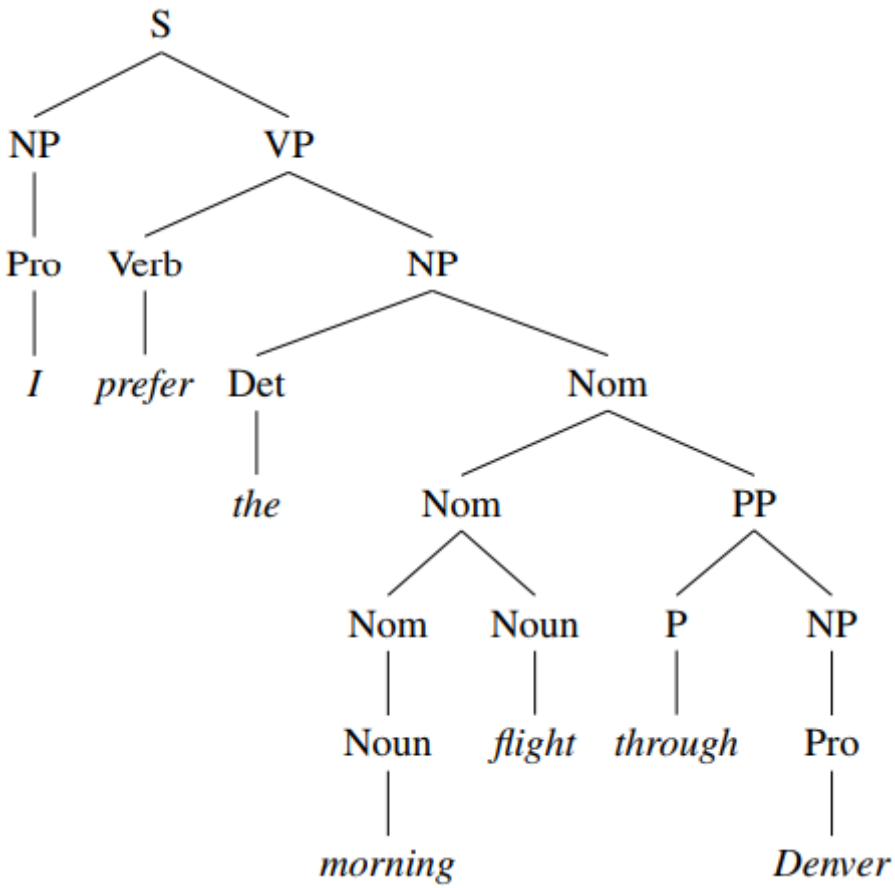
- **Phrase structure grammar** is all about **constituents**: phrasal units that words combine into.
- **Dependency grammar**, on the other hand, focuses on how words *relate* to other words: dependency relation between the headword and its dependents.

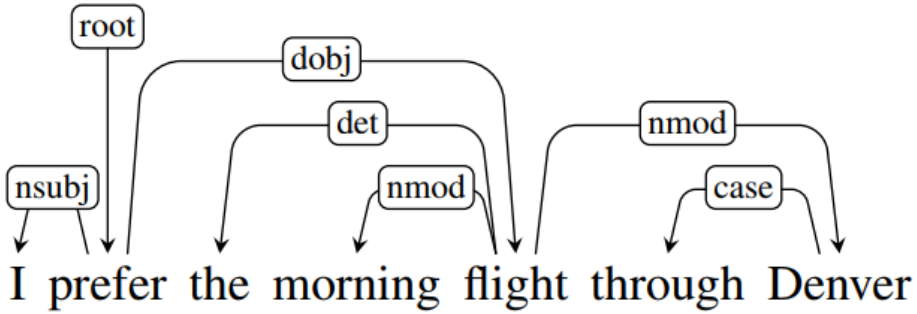


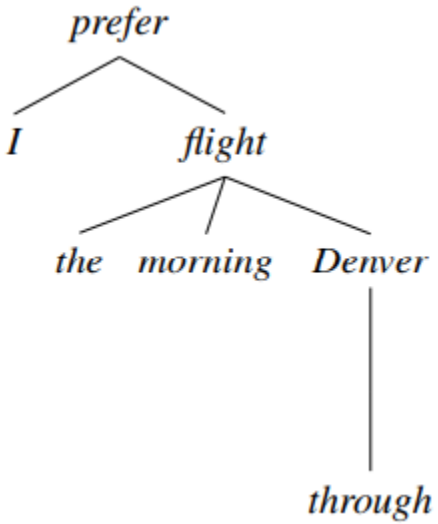
- NLTK book chapter: Dependency and Dependency Grammar
  - http://www.nltk.org/book/ch08.html#dependencies-and-dependency-grammar

A comparison

Constituency grammar vs. Dependency grammar







Universal dependencies

- Dependency grammar and parsing have become increasingly popular.
- Dependency grammar is thought to be more suited to languages with flexible word order.

← Could it be a better candidate for a truly universal grammar formalism?

← Linguistic theory aside, does it offer an engineering-side advantage?

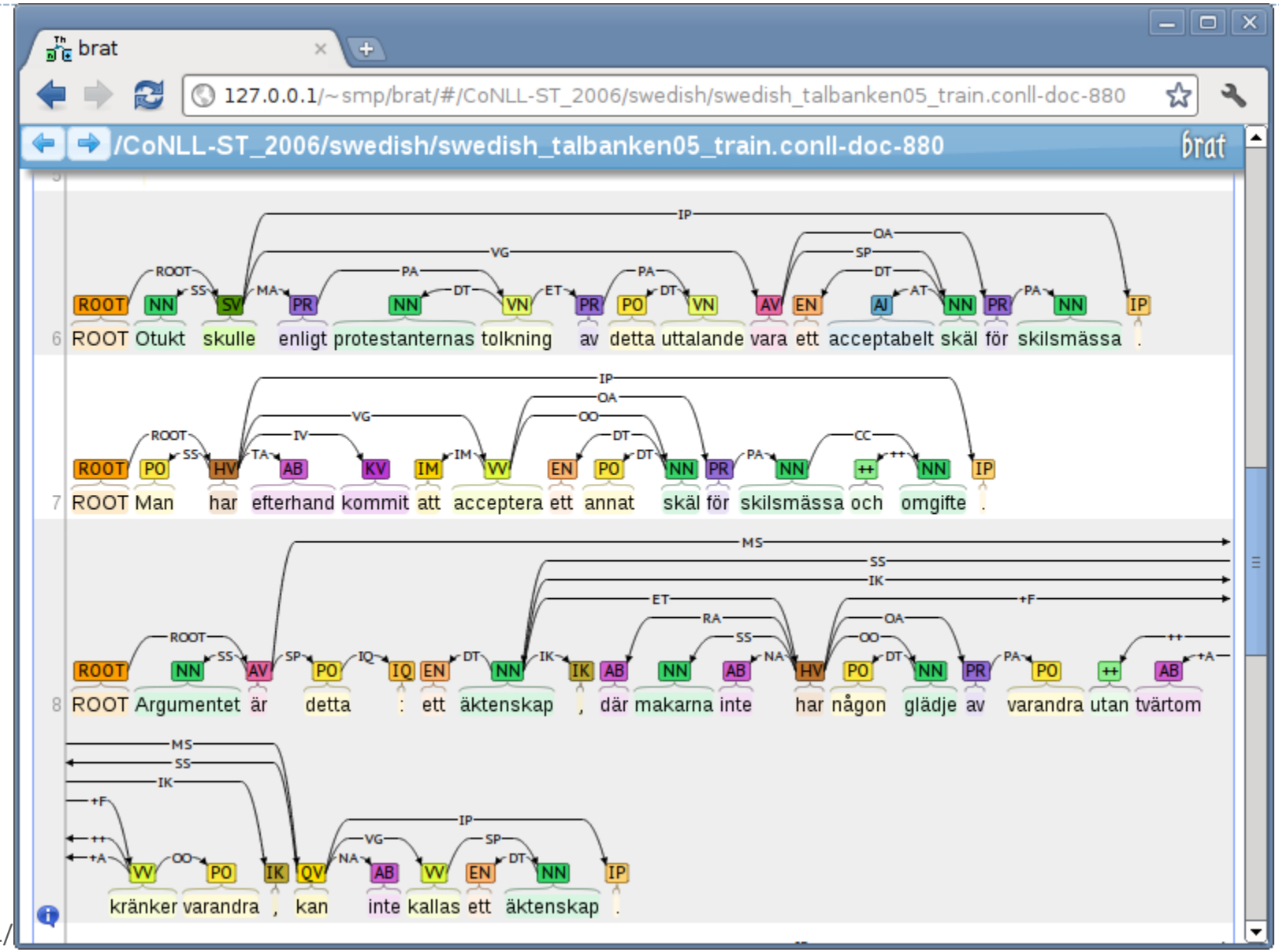
- **Universal Dependencies** working group
  - http://universaldependencies.org/introduction.html
  - A wide variety of languages represented!

Dependency annotation: example

- https://raw.githubusercontent.com/UniversalDependencies/UD_English/master/en-ud-dev.conllu

```
# sent_id = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713-0002
# text = President Bush on Tuesday nominated two individuals to replace retiring jurists on federal courts in the Washington
area.
1       President       President       PROPN   NNP     Number=Sing     5       nsubj   5:nsubj _
2       Bush    Bush    PROPN   NNP     Number=Sing     1       flat    1:flat  _
3       on      on      ADP     IN      _       4       case    4:case  _
4       Tuesday Tuesday PROPN   NNP     Number=Sing     5       obl     5:obl   _
5       nominated       nominate        VERB    VBD     Mood=Ind|Tense=Past|VerbForm=Fin        0       root    0:root  _
6       two     two     NUM     CD      NumType=Card    7       nummod  7:nummod        _
7       individuals     individual      NOUN    NNS     Number=Plur     5       obj     5:obj   _
8       to      to      PART    TO      _       9       mark    9:mark  _
9       replace replace VERB    VB      VerbForm=Inf    5       advcl   5:advcl _
10      retiring        retire  VERB    VBG     VerbForm=Ger    11      amod    11:amod _
11      jurists jurist  NOUN    NNS     Number=Plur     9       obj     9:obj   _
12      on      on      ADP     IN      _       14      case    14:case _
13      federal federal ADJ     JJ      Degree=Pos      14      amod    14:amod _
14      courts  court   NOUN    NNS     Number=Plur     11      nmod    11:nmod _
15      in      in      ADP     IN      _       18      case    18:case _
16      the     the     DET     DT      Definite=Def|PronType=Art       18      det     18:det  _
17      Washington      Washington      PROPN   NNP     Number=Sing     18      compound        18:compound     _
18      area    area    NOUN    NN      Number=Sing     14      nmod    14:nmod SpaceAfter=No
19      .       .       PUNCT   .       _       5       punct   5:punct _
```

Annotation interface



What are linguists' roles in all this?

- Doing the annotation
  - Linguistics undergrads and grads make excellent annotators.
- Leading annotation projects
  - Design annotation schemes
  - Develop annotation guidelines
  - Train and supervise annotators
  - An example: ftp://ftp.cis.upenn.edu/pub/ircs/tr/01-10/01-10.pdf
- As part of NLP community, help keep linguistic knowledge representation in balance with engineering-side considerations
- Be a USER of linguistically annotated data by conducting empirical research
  - An example: https://web.stanford.edu/~bresnan/qs-submit.pdf

Wrapping up

- To-Do #13
  - Two more visits!
- Work on your term project!
  - Come see me.
- 3rd progress report due after Thanksgiving
  - Guidelines will be updated shortly.
- Presentation schedule
  - 11/28 (Tue) Margaret
  - 11/30 (Thu) Ben, Paige, Andrew
  - 12/5 (Tue) Alicia, Chris, Katherine
  - 12/7 (Thu) Dan, Robert Kyle

  ← Guidelines will also be posted.