Lecture 8:
Corpus Linguistics, Annotation

LING 1340/2340: Data Science for Linguists

Jevon Heath

Objectives

- Corpus linguistics
  - Review of corpora and corpus tools
  - Your own data plans for your project

- Linguistic annotation
  - Types of linguistic annotation
  - Annotation formats
  - Annotation tools
  - Inter-annotator agreement

Corpus linguistics

- To-do #6 corpora and tools:
  - https://github.com/Data-Science-for-Linguists-2020/Class-Plaza/blob/master/corpora_tools_list.md
- What exciting corpora and tools did you discover?

Your term project

- Your project is now on GitHub
  - https://github.com/Data-Science-for-Linguists-2020
- First progress report is due in a couple of weeks
  - Focus on data: sourcing, curation and cleaning
- Managing your data
  - You will be manipulating and processing your data.
  - Should you include your data set in your GitHub repo?

    GOOD QUESTION. Next slide ➔

Licensing, public vs. private

- Your data:
  - Your original data source: what kind of license does it come with?
  - Can you re-distribute the data?
  - "Derivative" data: are you allowed to distribute?
  - How about samples?
  - How to best *present* the outcome and ensure *reproducibility* if you cannot share your data in full?
- Your code:
  - Will you allow other people to use your code? Re-distribute?
  - Will you allow other people to turn your code into a commercial product? Patent it?

Licensing, public vs. private

- As a principle, your term project -- including code and data -- should be **as public and open as possible**.
  - Your repo should be **public**.
  - For now, store your data files in a directory that's ignored through `.gitignore` . Suggestion: private/ or data/.

Licensing, public vs. private

- Do your research on copyright and licensing.
  - http://www.library.pitt.edu/copyright
  - https://choosealicense.com/
- Document, document, document!
  - You should **document and justify** your sharing and licensing decisions. It is an important part of your project.

Data standards & exchange formats

	What	Notes, reference
CSV	Comma-separated values	Compatible with Excel
TSV	Tab-separated values	
HTML	Web pages	
XML	For markup and text encoding	A Gentle Introduction to XML by TEI
JSON	JavaScript Object Notation (Twitter, Jupyter Notebook)	Introducing JSON JSON example (vs. XML)

They are all TEXT files.

- Encoding: Latin-1, ASCII, UTF-8, UTF-16, CP1252, ...
- Line endings:
  - LF (`'\n'`: OS X & Linux) , CRLF (`'\r\n'`: Windows)
- But underneath it all, these files are all TEXT files with **special formatting syntax** and **special characters** designated for formatting purposes.
  - In command line, you can `cat` and `less` through the files.
  - You can open them up in a text editor (Atom, Notepad++) and edit.
  - Some editors/applications are aware of the format-specific syntax and will highlight/render accordingly.
    - Unlike, say, PDF files, style attributes are NOT part of the files themselves. (e.g., markdown file)

Format conversion

- When dealing with corpora, you may need to convert 100+ files at once.
  - On-line services are too cumbersome.
  - Try batch-processing through command line.
- Automatic tools available on command line.
  - Encoding conversion: `iconv` (Linux, OS X, on Git Bash)
  - Line ending conversion: `unix2dos`, `dos2unix`
  - Pandoc http://www.pandoc.org/
    - Universal document coverter
    - HTML, XML, PDF, LaTeX, Markdown, Epub, MS Doc, ...
    - After installation, you can use it via command line

Resource-specific (ad-hoc) formats

- 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 ./.
```

- Korean Treebank corpus:

```
;;05:127: 저는 그 일을 할 수 있는 한 빨리 하겠습니다 .

(S (NP-SBJ 저/NPN+는/PAU)
   (VP (NP-OBJ-LV 그/DAN
                  일/NNC+을/PCA)
       (VP (NP-ADV (S (NP-SBJ (S (NP-SBJ *pro*)
                                 (VP 하/VV+ㄹ/EAN))
                              (NP 수/NNX))
                         (ADJP 있/VJ+는/EAN))
                   (NP 한/NNX))
           (ADVP 빨리/ADV)
           (VP (LV 하/VV+겠/EPF+습니다/EFN))))
   ./SFN)
```

It is up to end users to write code to parse data files.

Refer to documentation!

Do not re-invent the wheel.

- Don't try and parse them manually.
- There are Python libraries. Import and use them.
  - CSV & TSV: `pandas`
  - HTML & XML: Beautiful Soup (`bs4`)
  - JSON:
    - `json` library
    - `pandas.read_json`
- NLP-specific formats (Treebank, Universal Dependency, CoNLL):
  - Look at NLTK, see if it has reader
  - If not, chances are there is parser library written by someone somewhere (likely on GitHub)

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

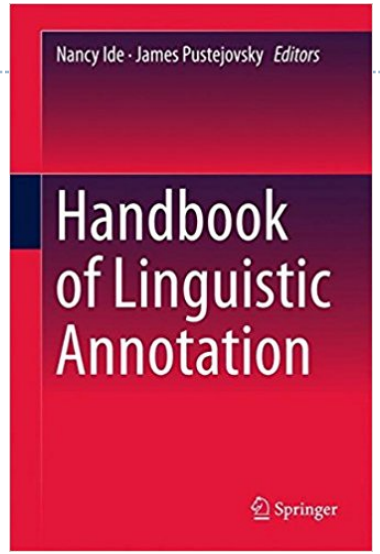
  ← Linguistic annotation represents linguistic knowledge of humans that AI agents learn through machine learning, which they then mimic

What are linguists' roles in all this?

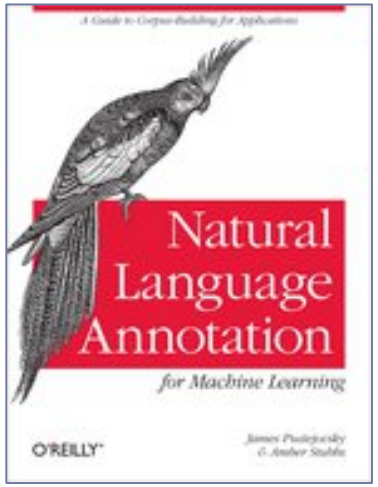
- Doing the annotation
  - Linguistics undergrads and grads make excellent annotators.
- Leading annotation projects
  - Design annotation schemes
  - Develop annotation guidelines
  - Train and supervise annotators
  - An example: ftp://ftp.cis.upenn.edu/pub/ircs/tr/01-10/01-10.pdf
- As part of the 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

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



- *Natural Language Annotation for Machine Learning* (2012)
  - James Pustejovsky, Amber Stubbs
  - https://www.oreilly.com/library/view/natural-language-annotation/9781449332693/ch01.html



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
  - Next slide

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?

Syntactic annotation: 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) )))))
    (. .) ))
```

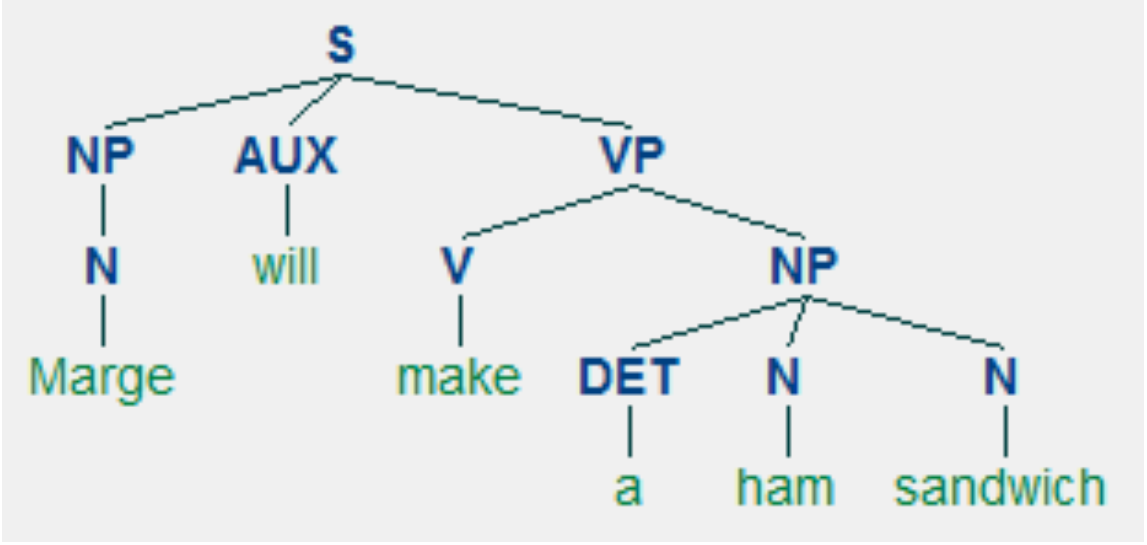
http://languagelog.ldc.upenn.edu/nll/?p=3594

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

Context-free grammar

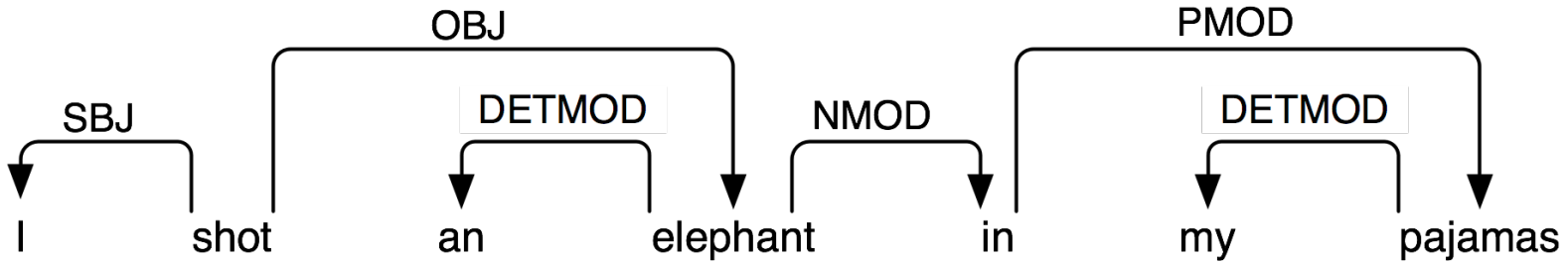
- Phrase-structure grammar is based upon constituency.
- Each local constituent can be expressed through **context-free grammar**.

```
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

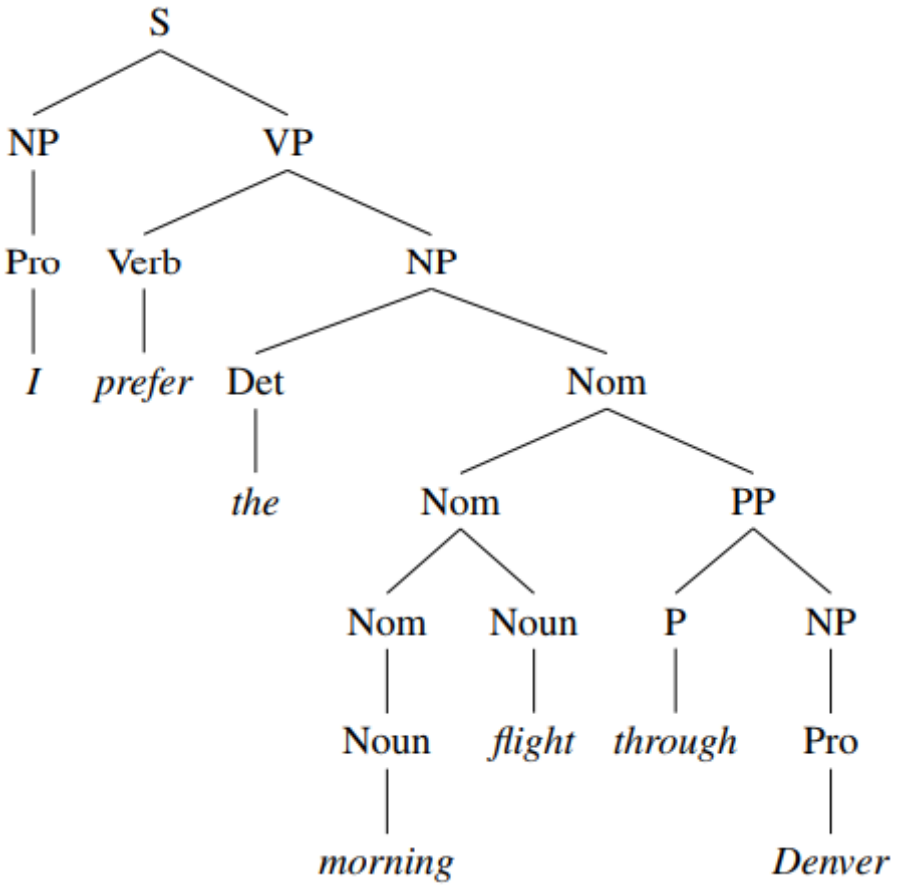
- **Phrase structure grammar** is all about **constituents**: phrasal units that words combine into.
- **Dependency grammar**, on the other hand, focuses on how words *relate* to other words: dependency relation between the headword and its dependents.

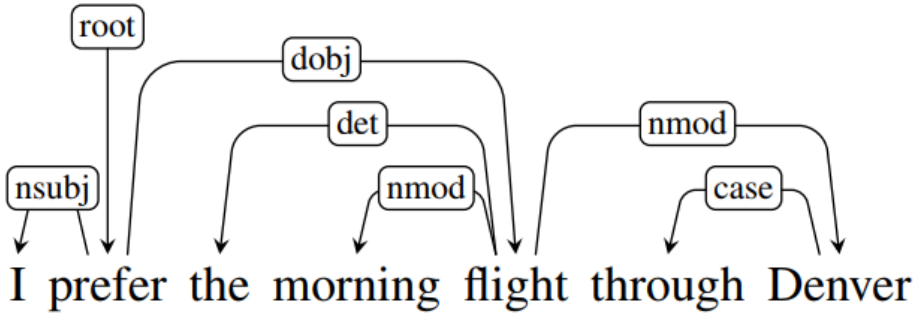


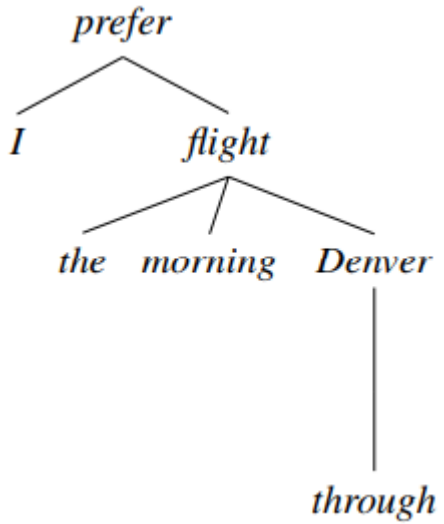
- NLTK book chapter: Dependency and Dependency Grammar
  - http://www.nltk.org/book/ch08.html#dependencies-and-dependency-grammar

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-EWT/master/en_ewt-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	_
```

Another licensed data set

- TIMIT Acoustic-Phonetic Continuous Speech Corpus
  - https://catalog.ldc.upenn.edu/ldc93s1
  - In "Licensed-Data-Sets" repo
  - Is this a "corpus"...?

Wrapping up

- Next class: guest lecture by Lauren Collister and Dominic Bordelon
  - Submit your question via To-do 6!
  - Think about licensing issues for your project
- CSLX talk today at 4:45 by Dan Villareal
  - CL G-17 (LMC)
- Reminder:
  - You should WORK ON YOUR PROJECT!