Lecture 8: Annotation

LING 1340/2340: Data Science for Linguists

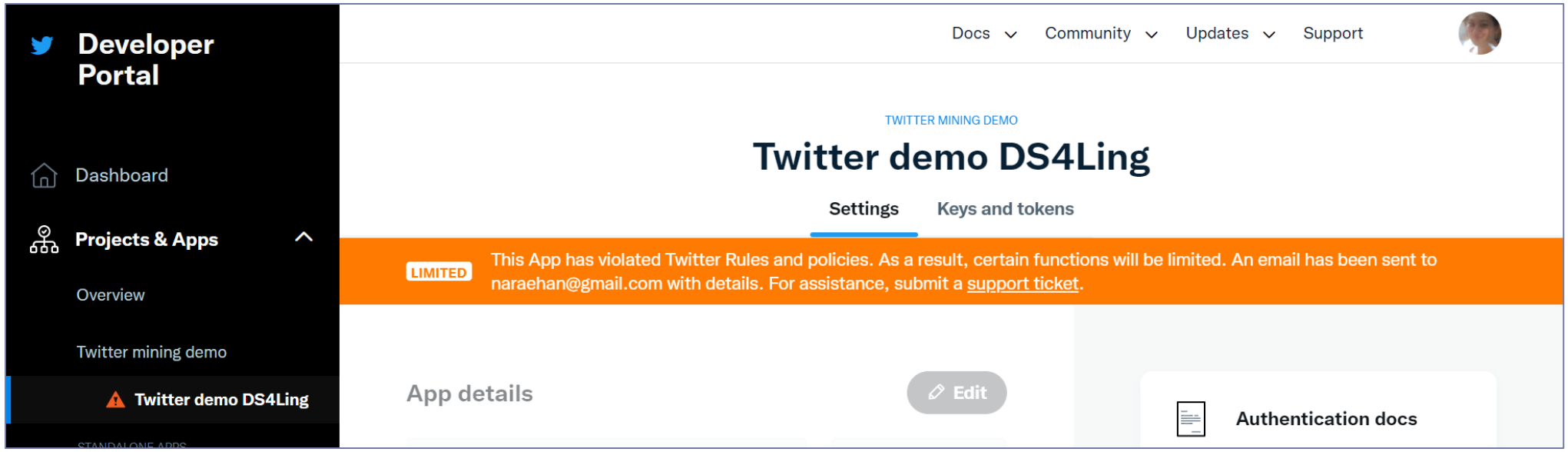
Na-Rae Han

Objectives

- To-do #6 Twitter mining
- Linguistic research using Twitter data?
- Linguistic annotation
  - Types of linguistic annotation
  - Annotation formats
  - Annotation tools
    - Hands-on with Webanno

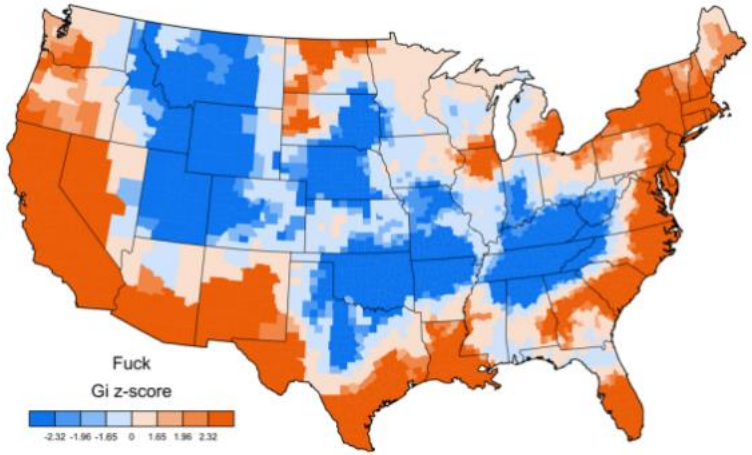
To-do #6 Twitter Mining

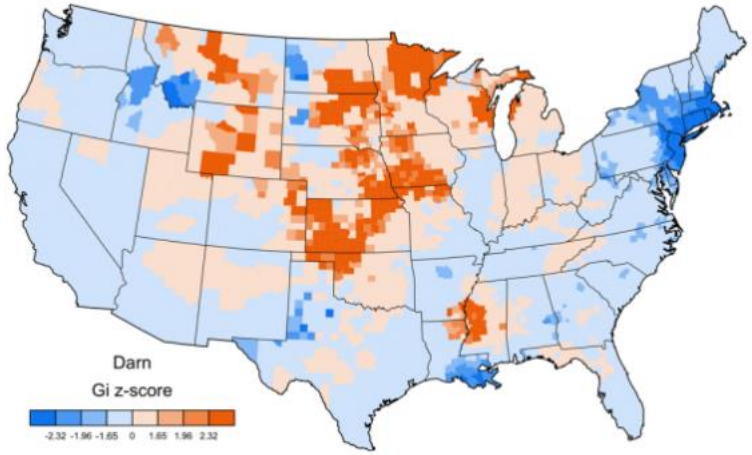
- How did it go?
- Twitter API v1.1 vs. 2.0
- Anyone got… suspended?



Mining social media for swear words

- https://stronglang.wordpress.com/2015/07/28/mapping-the-united-swears-of-america/
  - Jack Grieve mined Twitter and mapped prominent swear words by geographic regions within the US





Linguistic annotation

What types of linguistic annotation have we seen so far?

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

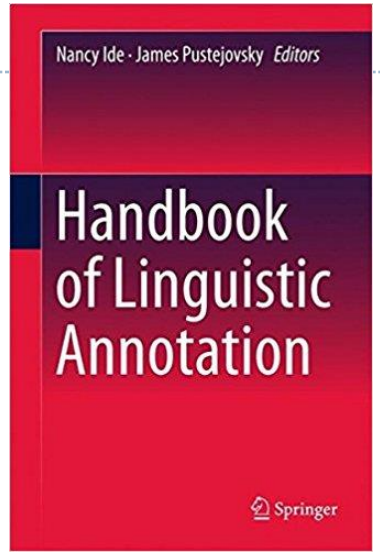
  ← Linguistic annotation represents linguistic knowledge of humans that AI agents learn through machine learning, which they then mimic

What are linguists' roles in all this?

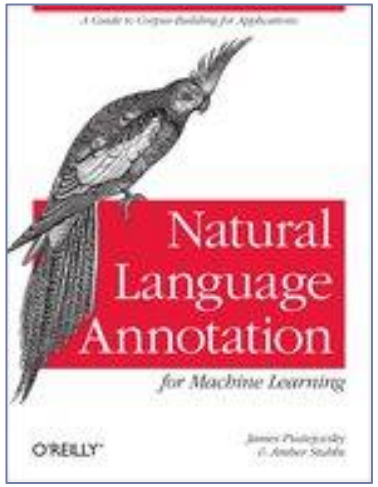
- Doing the annotation
  - Linguistics undergrads and grads make excellent annotators.
- Leading annotation projects
  - Design annotation schemes
  - Develop annotation guidelines
  - Train and supervise annotators
  - An example https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/penn-etb-2-style-guidelines.pdf
- 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) )))))
    (. .) ))
```

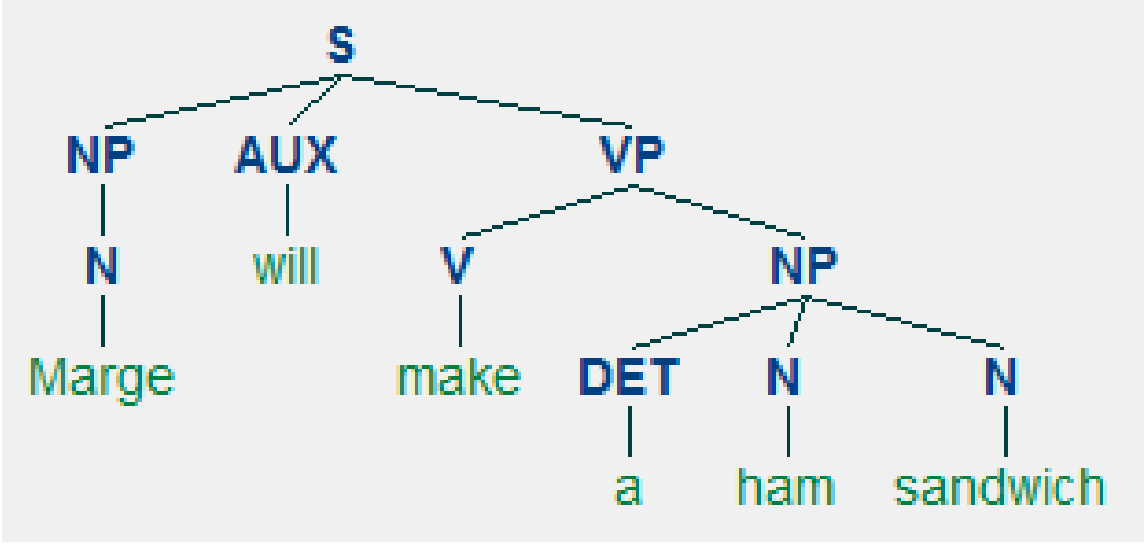
http://languagelog.ldc.upenn.edu/nll/?p=3594

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

Context-free grammar

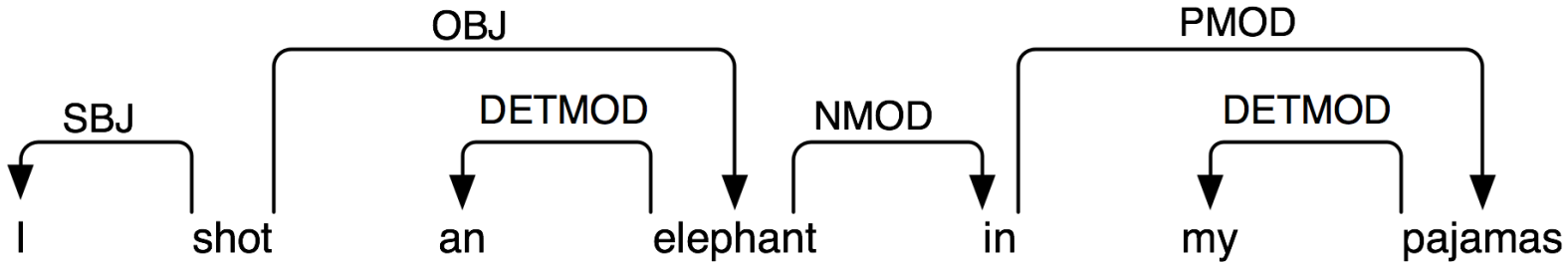
- Phrase-structure grammar is based upon constituency.
- Each local constituent can be expressed through **context-free grammar**.

```
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

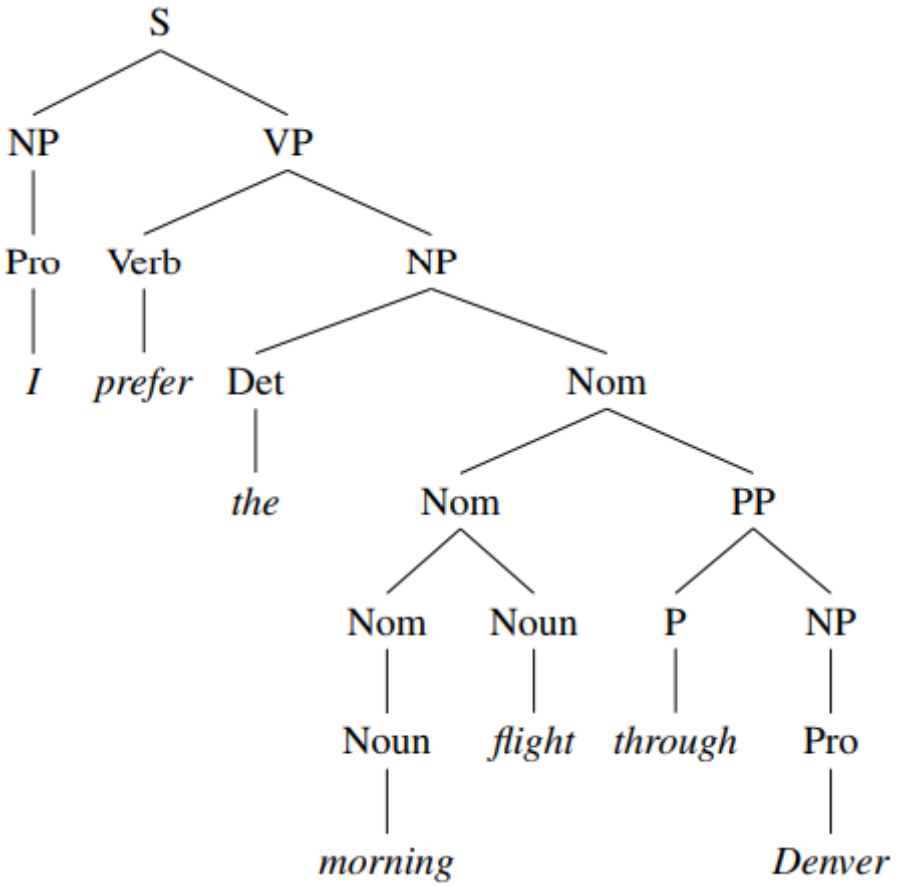
- **Phrase structure grammar** is all about **constituents**: phrasal units that words combine into.
- **Dependency grammar**, on the other hand, focuses on how words *relate* to other words: dependency relation between the headword and its dependents.

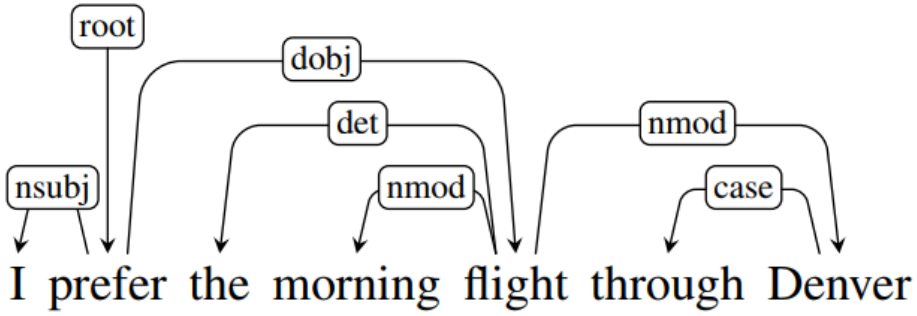


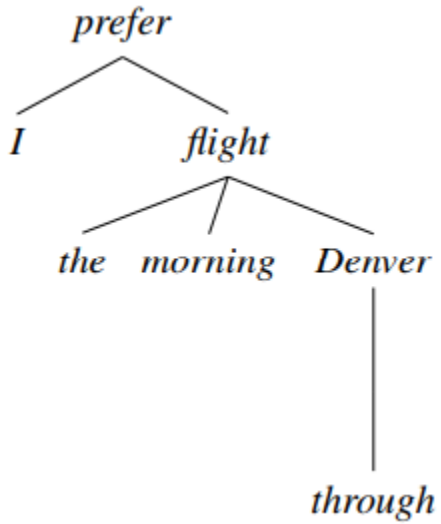
- NLTK book chapter: Dependency and Dependency Grammar
  - http://www.nltk.org/book/ch08.html#dependencies-and-dependency-grammar

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!

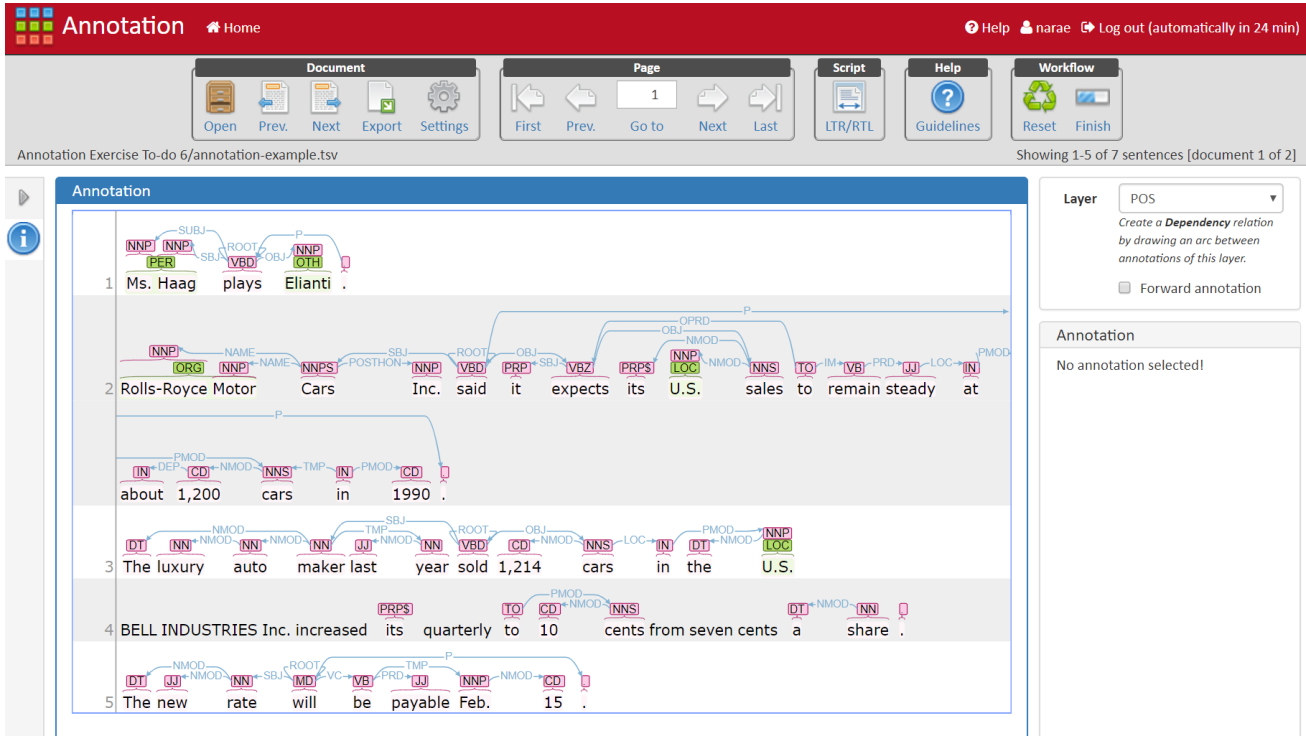
Dependency annotation: format

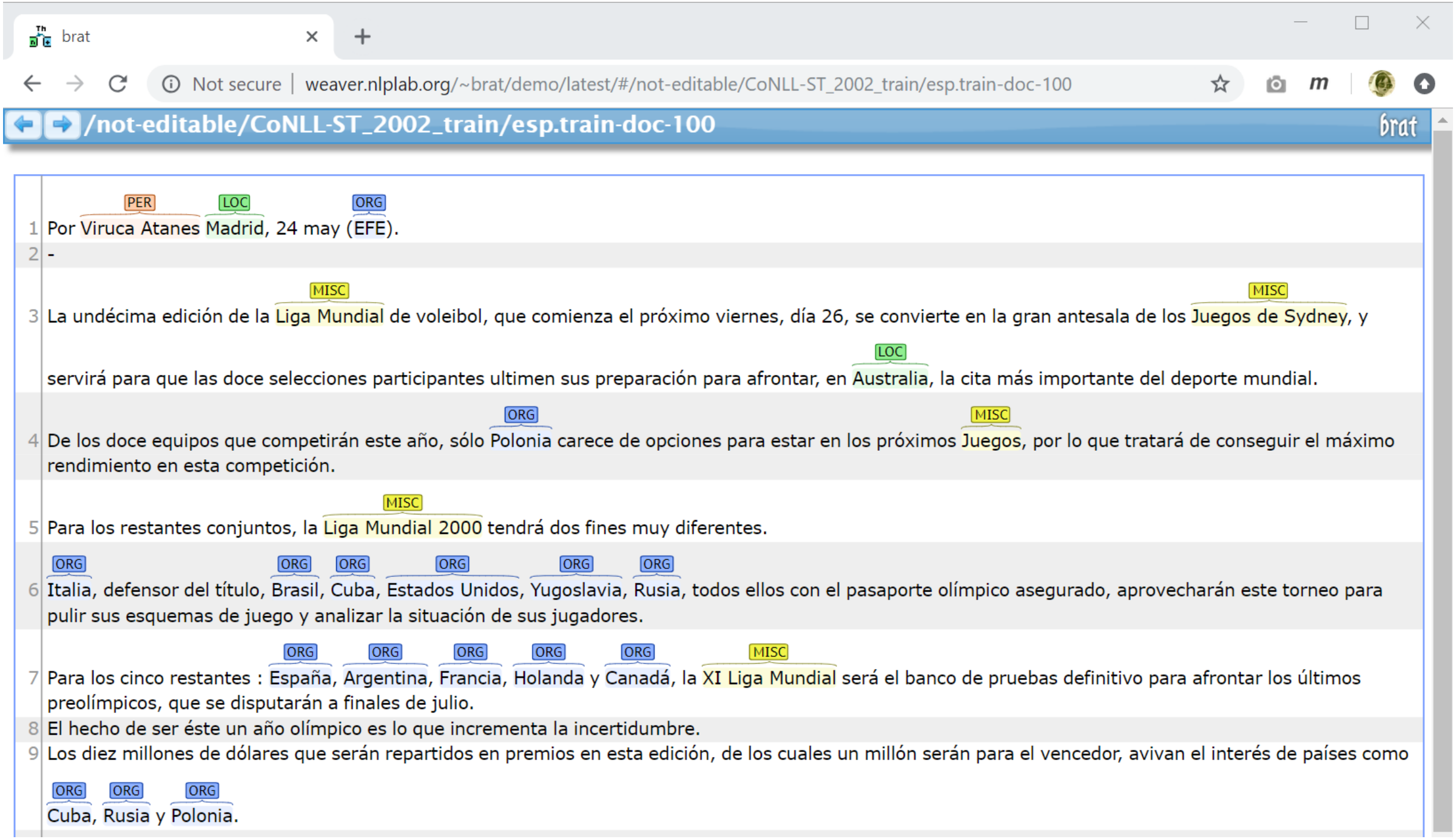
- https://raw.githubusercontent.com/UniversalDependencies/UD_English-EWT/dev/en_ewt-ud-dev.conllu

```
# sent_id = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713-0002
# newpar id = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713-p0002
# 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:on	_
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:to	_
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:on	_
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:in	SpaceAfter=No
19	.	.	PUNCT	.	_	5	punct	5:punct	_
```

Annotation interface: browser-based

- Text editor programs (Notepad++, Atom) do not cut it as an annotation platform
  - Why?
- Often, large-scale annotation projects involve a centrally managed annotation interface, accessible via a browser
  - Brat Rapid Annotation Tool
  - WebAnno




brat
Not secure | weaver.nlplab.org/~brat/demo/latest/#/not-editable/CoNLL-ST_2002_train/esp.train-doc-100
/not-editable/CoNLL-ST_2002_train/esp.train-doc-100
brat
1 Por Viruca Atanes Madrid, 24 may (EFE).
2 -
3 La undécima edición de la Liga Mundial de voleibol, que comienza el próximo viernes, día 26, se convierte en la gran antesala de los Juegos de Sydney, y servirá para que las doce selecciones participantes ultimen sus preparación para afrontar, en Australia, la cita más importante del deporte mundial.
4 De los doce equipos que competirán este año, sólo Polonia carece de opciones para estar en los próximos Juegos, por lo que tratará de conseguir el máximo rendimiento en esta competición.
5 Para los restantes conjuntos, la Liga Mundial 2000 tendrá dos fines muy diferentes.
6 Italia, defensor del título, Brasil, Cuba, Estados Unidos, Yugoslavia, Rusia, todos ellos con el pasaporte olímpico asegurado, aprovecharán este torneo para pulir sus esquemas de juego y analizar la situación de sus jugadores.
7 Para los cinco restantes : España, Argentina, Francia, Holanda y Canadá, la XI Liga Mundial será el banco de pruebas definitivo para afrontar los últimos preolímpicos, que se disputarán a finales de julio.
8 El hecho de ser éste un año olímpico es lo que incrementa la incertidumbre.
9 Los diez millones de dólares que serán repartidos en premios en esta edición, de los cuales un millón serán para el vencedor, avivan el interés de países como Cuba, Rusia y Polonia.

Wrapping up

- Next class
  - Corpus linguistics, annotation
- To-do #7
  - Annotation try-out
- Your project
  - Work on it! Focus on DATA.