Lecture 11: Linguistic Annotation

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

- AMR review
- Linguistic annotation
  - Types of linguistic annotation
  - Annotation formats
  - Annotation tools

AMR example

- Guidelines:
  - https://github.com/amrisi/amr-guidelines/blob/master/amr.md

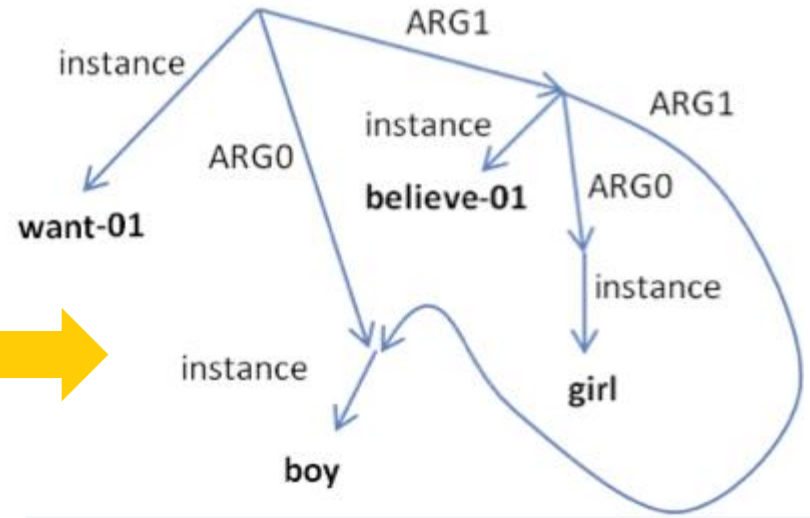
The boy desires the girl to believe him.

The boy desires to be believed by the girl.

The boy has a desire to be believed by the girl.

The boy's desire is for the girl to believe him.

The boy is desirous of the girl believing him.



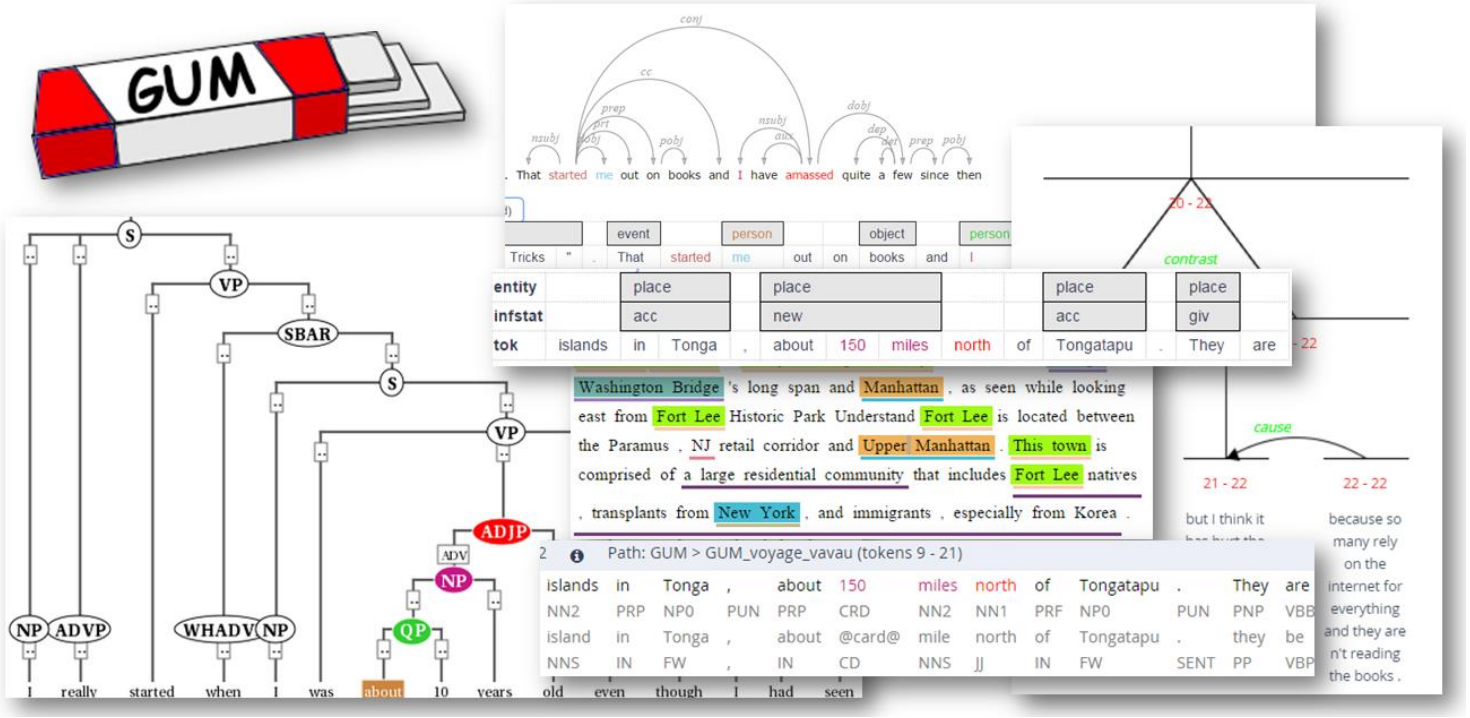

```
(w / want-01
   :ARG0 (b / boy)
   :ARG1 (b2 / believe-01
             :ARG0 (g / girl)
             :ARG1 b))
```

AMR annotated corpora

- https://amr.isi.edu/download.html
- *The Little Prince* by Antoine de Saint-Exupéry is annotated in AMR in full.
  - English
  - Chinese
- Why build such corpora?

Linguistic annotation: what types?

- What types of linguistic annotation have we seen so far?
- GUM: **The Georgetown University Multilayer Corpus**
  - https://gucorpling.org/gum/index.html
  - A corpus with *all* levels of linguistic knowledge annotated!!



Why annotate?

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

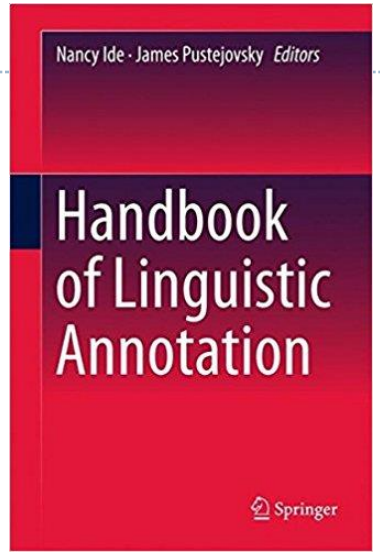
  ← Linguistic annotation represents linguistic knowledge of humans that AI agents learn through machine learning, which they then mimic

What are linguists' roles in all this?

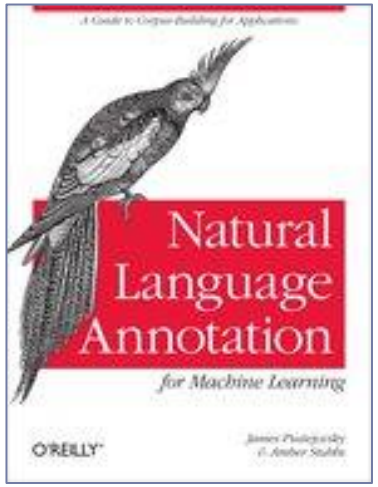
- Doing the annotation
  - Linguistics undergrads and grads make excellent annotators.
- Leading annotation projects
  - Design annotation schemes
  - Develop annotation guidelines
  - Train and supervise annotators
  - An example https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/penn-etb-2-style-guidelines.pdf
- 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
- Increasingly: Be a community-minded steward of language data. Address concerns of ethics and representation.

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



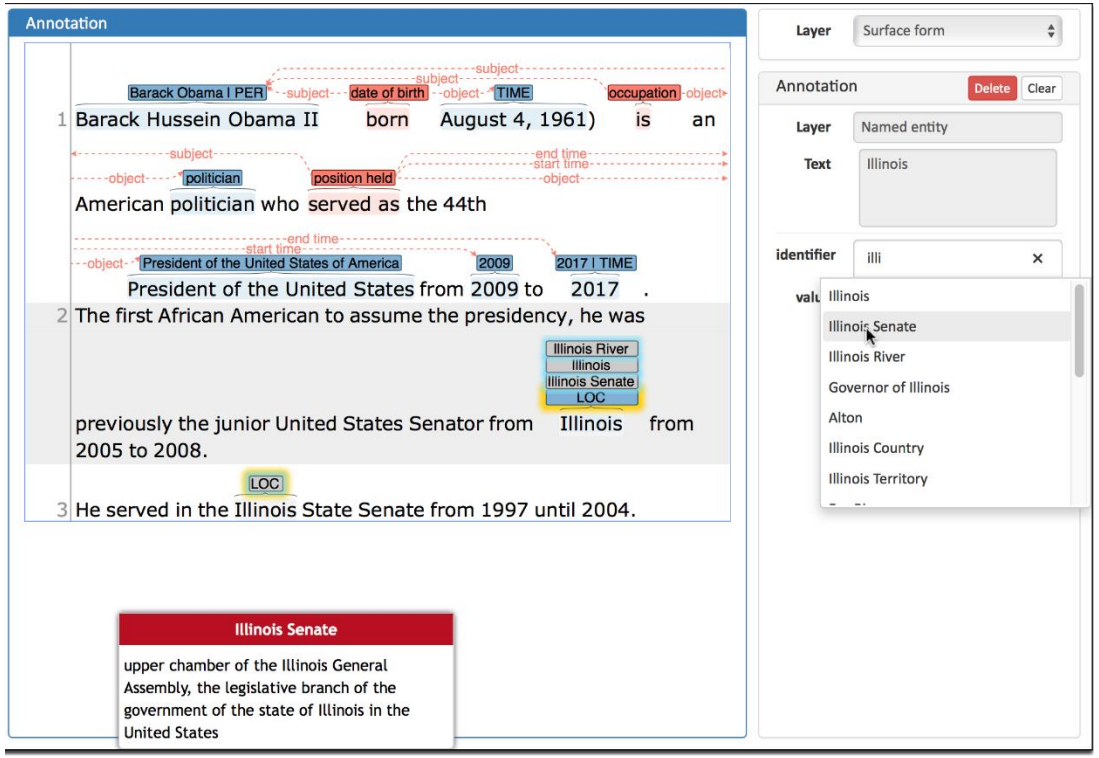
- *Natural Language Annotation for Machine Learning* (2012)
  - James Pustejovsky, Amber Stubbs
  - https://www.oreilly.com/library/view/natural-language-annotation/9781449332693/ch01.html



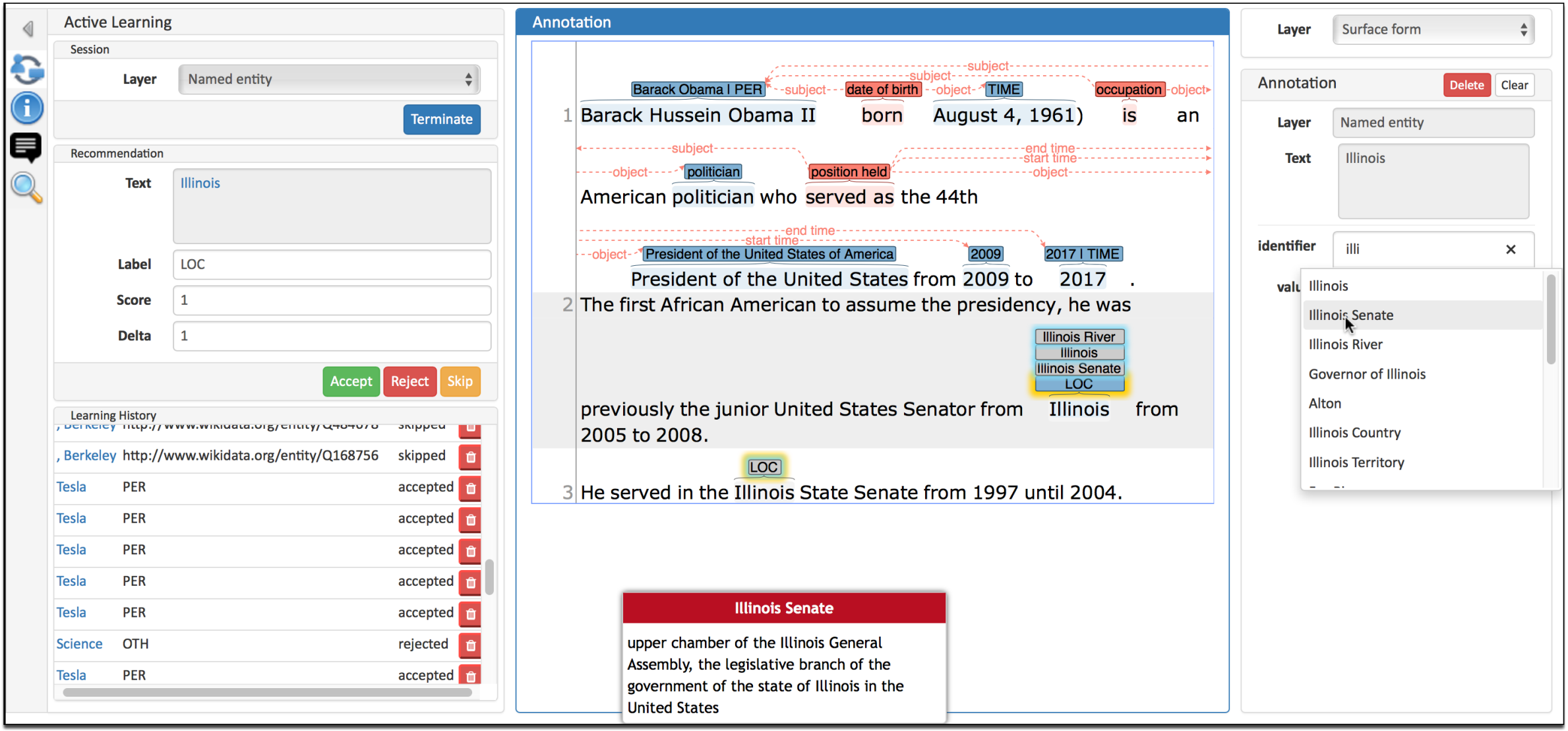
Annotation interface: browser-based

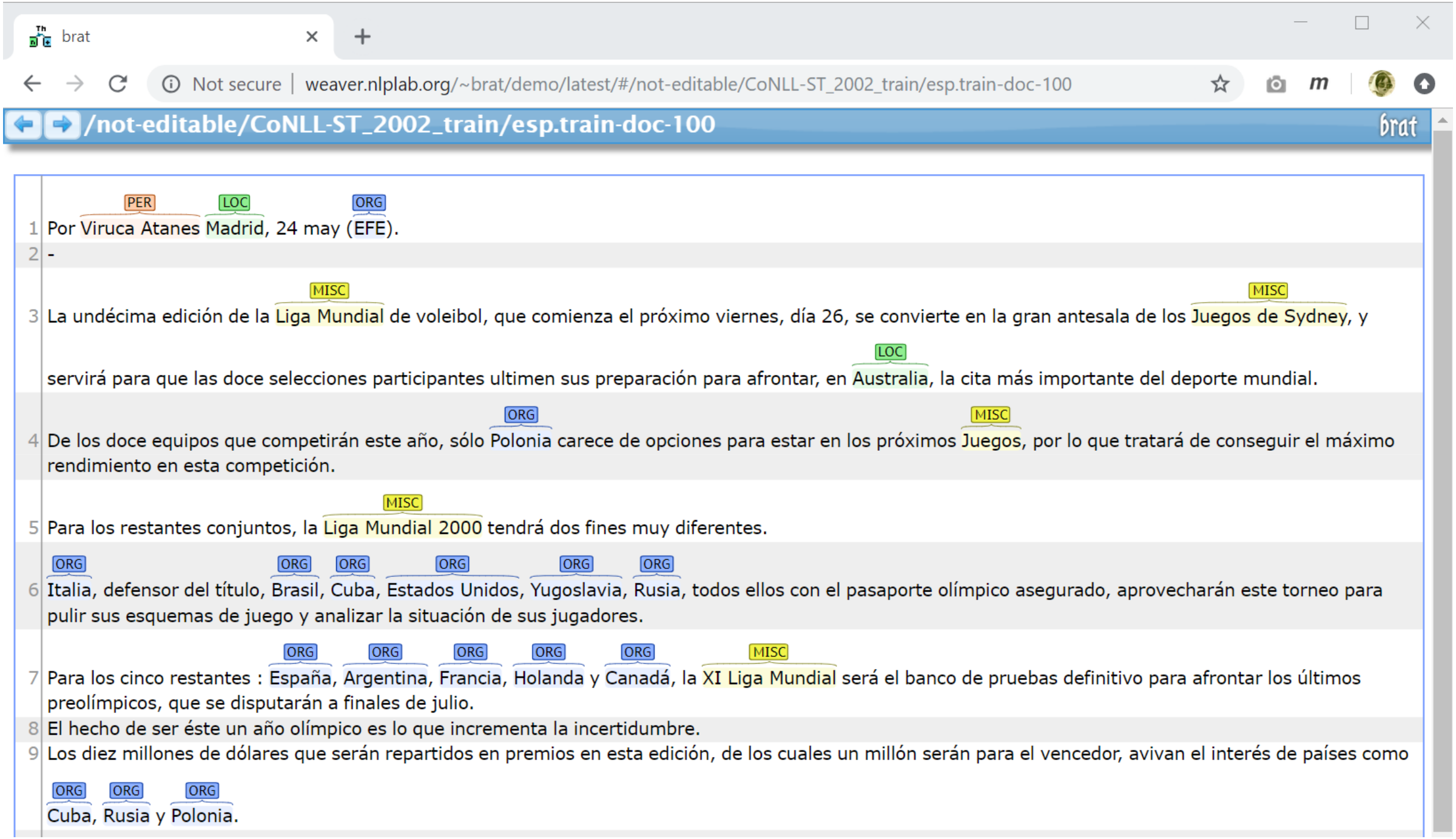
- Text editor programs (Notepad++, Atom) do not cut it as an annotation platform. Why?

- Often, large-scale annotation projects involve a centrally managed annotation interface, accessible via a browser
  - WebAnno
  - INCEpTION
    - Georgetown University's GUM Corpus used it for annotation: https://inception-project.github.io/use-cases/gum/



INCEpTION annotation interface




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
  - Annotation wrap
  - Machine learning: regression
- Your project
  - Progress Report #1 due this Friday!