

# 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>

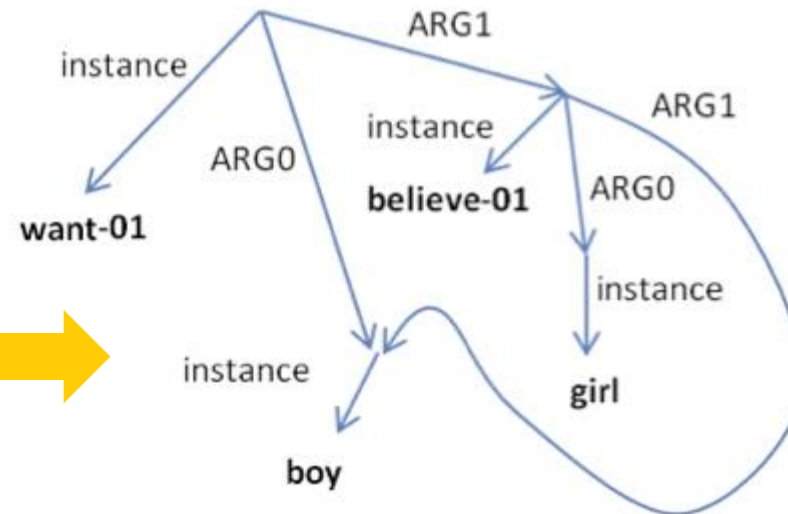
The boy desires the girl to believe him.

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!!

**GUM**

That started me out on books and I have amassed quite a few since then

Washington Bridge's long span and Manhattan, as seen while looking east from Fort Lee Historic Park Understand Fort Lee is located between the Paramus, NJ retail corridor and Upper Manhattan. This town is comprised of a large residential community that includes Fort Lee natives, transplants from New York, and immigrants, especially from Korea.

entity	place	place	place	place	place
instat	acc	new		acc	giv
tok	islands	in	Tonga	, about 150 miles north	of Tongatapu . They are

Path: GUM > GUM\_voyage\_vavau (tokens 9 - 21)

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
islands	in	Tonga	,	about	150	miles	north	of	Tongatapu	.	They	are								
NN2	PRP	NPO	PUN	PRP	CRD	NN2	NN1	PRF	NPO	PUN	PNP	VBB								
island	in	Tonga	,	about	@card@	mile	north	of	Tongatapu	.	they	be								
NNS	IN	FW	IN	CD	NNS	JJ	IN	FW	SENT	PP	VBP									

# 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?

---

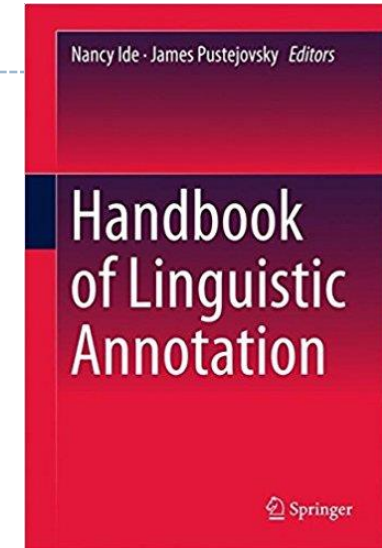
- ▶ Doing the annotation
  - ◆ Linguistics undergrads and grads make excellent annotators.
- ▶ Leading annotation projects
  - ◆ Design annotation schemes
  - ◆ Develop annotation guidelines
  - ◆ Train and supervise annotators
  - ◆ An example <https://www ldc.upenn.edu/sites/www ldc.upenn.edu/files/penn-etb-2-style-guidelines.pdf>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

---

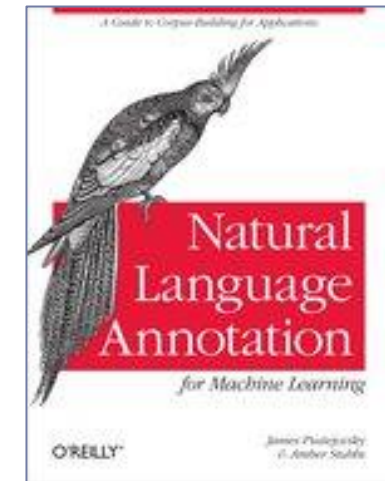
## ▶ *Handbook of Linguistic Annotation* (2017)

- ◆ Nancy Ide, James Pustejovsky (eds)
- ◆ [https://link.springer.com/chapter/10.1007/978-94-024-0881-2\\_1](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/>

The screenshot displays the INCEpTION annotation interface. The main window shows three sentences with various annotations:

- 1 Barack Hussein Obama II born August 4, 1961) is an American politician who served as the 44th President of the United States from 2009 to 2017 .
- 2 The first African American to assume the presidency, he was previously the junior United States Senator from Illinois from 2005 to 2008.
- 3 He served in the Illinois State Senate from 1997 until 2004.

Annotations include named entities (e.g., Barack Obama I PER, August 4, 1961), relations (e.g., date of birth, born, is an, position held, President of the United States of America), and locations (e.g., Illinois River, Illinois Senate, LOC). A sidebar on the right shows the current layer is 'Surface form' and the current annotation is 'Named entity' with the text 'Illinois'. A dropdown menu is open, showing a list of identifiers starting with 'illi', with 'Illinois Senate' selected. A tooltip for 'Illinois Senate' is visible at the bottom, providing a definition: 'upper chamber of the Illinois General Assembly, the legislative branch of the government of the state of Illinois in the United States'.

# INCEpTION annotation interface

### Active Learning

Session

Layer: Named entity

Terminate

Recommendation

Text: Illinois

Label: LOC

Score: 1

Delta: 1

Accept Reject Skip

Learning History

Berkeley	http://www.wikidata.org/entity/Q484078	skipped	
Berkeley	http://www.wikidata.org/entity/Q168756	skipped	
Tesla	PER	accepted	
Tesla	PER	accepted	
Tesla	PER	accepted	
Tesla	PER	accepted	
Tesla	PER	accepted	
Science	OTH	rejected	
Tesla	PER	accepted	

### Annotation

1 Barack Hussein Obama II born August 4, 1961) is an American politician who served as the 44th President of the United States from 2009 to 2017 .

2 The first African American to assume the presidency, he was previously the junior United States Senator from Illinois from 2005 to 2008.

3 He served in the Illinois State Senate from 1997 until 2004.

Annotations: Barack Obama I PER, date of birth TIME, occupation, politician, position held, 2009 TIME, 2017 TIME, LOC

Layer: Surface form

Annotation: Delete Clear

Layer: Named entity

Text: Illinois

identifier: illi

value: Illinois, Illinois Senate, Illinois River, Governor of Illinois, Alton, Illinois Country, Illinois Territory

**Illinois Senate**

upper chamber of the Illinois General Assembly, the legislative branch of the government of the state of Illinois in the United States

1 Por Viruca Atanes <sup>PER</sup> <sup>LOC</sup> Madrid, <sup>ORG</sup> 24 may (EFE).

2 -

3 La undécima edición de la <sup>MISC</sup> Liga Mundial de voleibol, que comienza el próximo viernes, día 26, se convierte en la gran antesala de los <sup>MISC</sup> Juegos de Sydney, y servirá para que las doce selecciones participantes ultimen sus preparación para afrontar, en <sup>LOC</sup> Australia, la cita más importante del deporte mundial.

4 De los doce equipos que competirán este año, sólo <sup>ORG</sup> Polonia carece de opciones para estar en los próximos <sup>MISC</sup> Juegos, por lo que tratará de conseguir el máximo rendimiento en esta competición.

5 Para los restantes conjuntos, la <sup>MISC</sup> Liga Mundial 2000 tendrá dos fines muy diferentes.

6 <sup>ORG</sup> Italia, defensor del título, <sup>ORG</sup> Brasil, <sup>ORG</sup> Cuba, <sup>ORG</sup> Estados Unidos, <sup>ORG</sup> Yugoslavia, <sup>ORG</sup> 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 : <sup>ORG</sup> España, <sup>ORG</sup> Argentina, <sup>ORG</sup> Francia, <sup>ORG</sup> Holanda y <sup>ORG</sup> Canadá, la <sup>MISC</sup> 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

<sup>ORG</sup> Cuba, <sup>ORG</sup> Rusia y <sup>ORG</sup> Polonia.

# Wrapping up

---

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