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
  - How to plan and run an annotation project
    - An anatomy of annotation project

AMR example

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

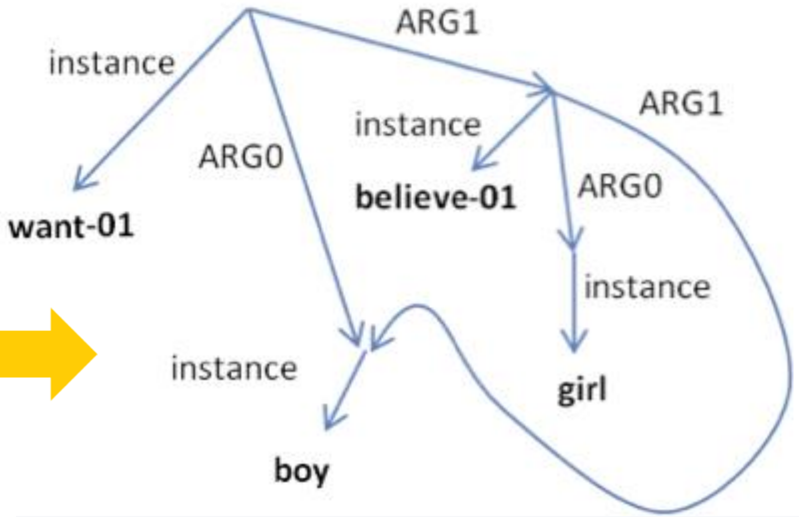
The boy desires the girl to believe him.

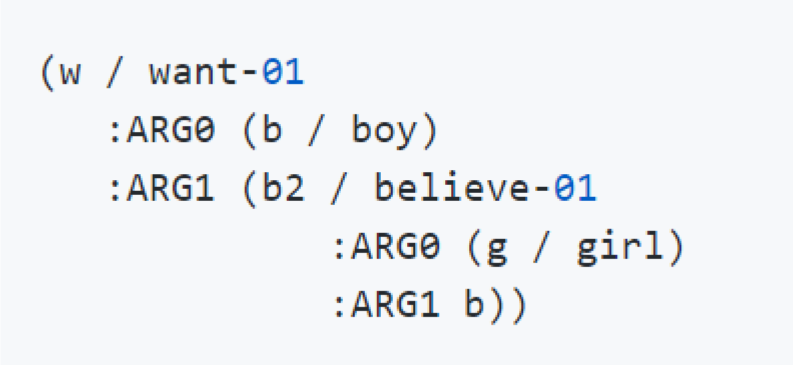
The boy desires to be believed by the girl.

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.





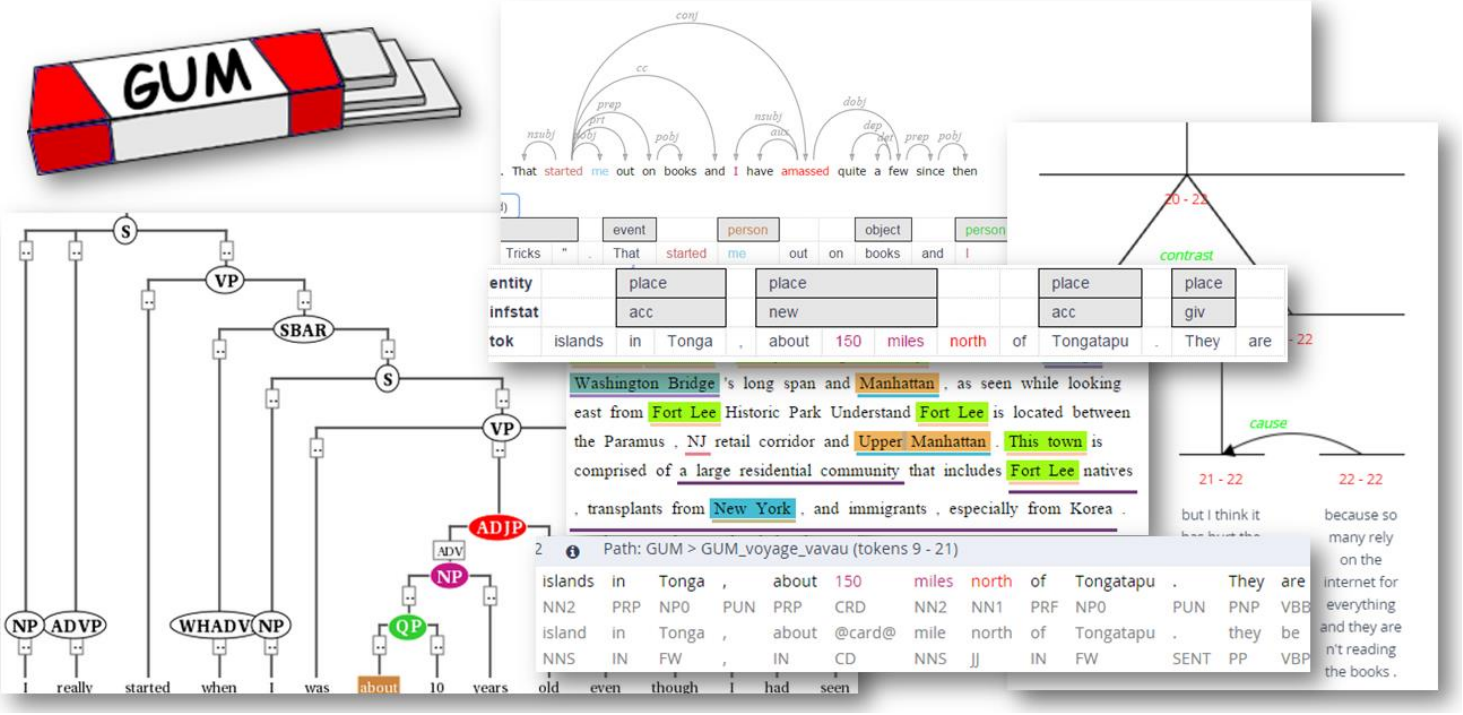
```
(w / want-01
   :ARG0 (b / boy)
   :ARG1 (b2 / believe-01
            :ARG0 (g / girl)
            :ARG1 b))
```

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!!



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

		event		person			object		person
Tricks	"	. That	started	me	out	on	books	and	I

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

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 .

but I think it

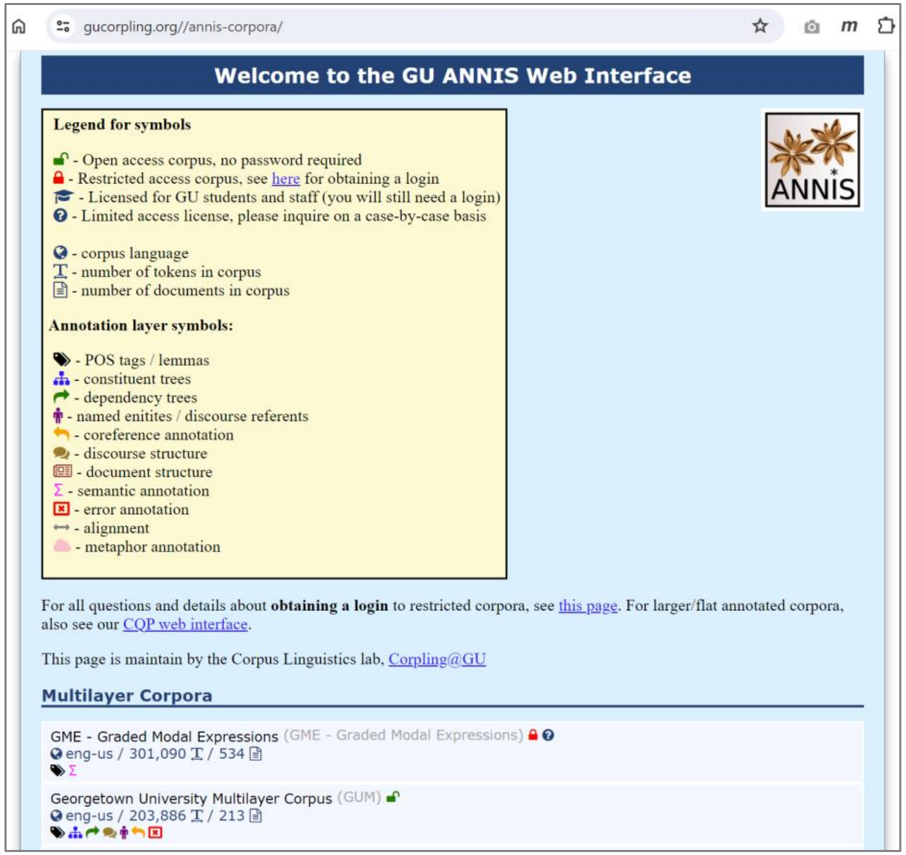
because so many rely on the internet for everything and they are n't reading the books .

2 Path: GUM > GUM_voyage_vavau (tokens 9 - 21)

islands	in	Tonga	,	about	150	miles	north	of	Tongatapu	.	They	are
NN2	PRP	NP0	PUN	PRP	CRD	NN2	NN1	PRF	NP0	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

Using GUM

- How to explore and use the GUM corpus?
  - Download the source on GitHub: https://github.com/amir-zeldes/gum then process it yourself
  - Or: use the **ANNIS** interface
    - A web browser-based search and visualization architecture for complex multilayer linguistic corpora with diverse types of annotation
    - https://corpus-tools.org/annis/
    - GU ANNIS Web Interface: https://gucorpling.org/annis-corpora/



gucorpling.org//annis-corpora/

Welcome to the GU ANNIS Web Interface

Legend for symbols

🔓 - Open access corpus, no password required
🔒 - Restricted access corpus, see here for obtaining a login
🎓 - Licensed for GU students and staff (you will still need a login)
❓ - Limited access license, please inquire on a case-by-case basis

🌐 - corpus language
T - number of tokens in corpus
📄 - number of documents in corpus

Annotation layer symbols:

🏷 - POS tags / lemmas
🖧 - constituent trees
↪ - dependency trees
🚹 - named enitites / discourse referents
↩ - coreference annotation
💬 - discourse structure
📰 - document structure
Σ - semantic annotation
☒ - error annotation
↔ - alignment
☁ - metaphor annotation

For all questions and details about **obtaining a login** to restricted corpora, see this page. For larger/flat annotated corpora, also see our CQP web interface.

This page is maintain by the Corpus Linguistics lab, Corpling@GU

Multilayer Corpora

GME - Graded Modal Expressions (GME - Graded Modal Expressions) 🔒 ❓
🌐 eng-us / 301,090 T / 534 📄
🏷 Σ

Georgetown University Multilayer Corpus (GUM) 🔓
🌐 eng-us / 203,886 T / 213 📄
🏷 🖧 ↪ 💬 🚹 ↩ ☒

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.

Wrapping up

- Next class
  - An anatomy of annotation project
  - Annotation wrap
  - Machine learning: regression
- Your project
  - Progress Report #1 due this Friday!