Lecture 8: Linguistic Data and Annotation

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

- Your term project
  - Plan submitted, repo created!
- Linguistic annotation
  - Types of linguistic annotation
  - Annotation formats

Your term project

- Everyone's project repo is at our GitHub org.
- First progress report is due next!
  - 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?
    - Depends on your license!

Linguistic data, corpus, annotation, coding

- How language exists in the world:

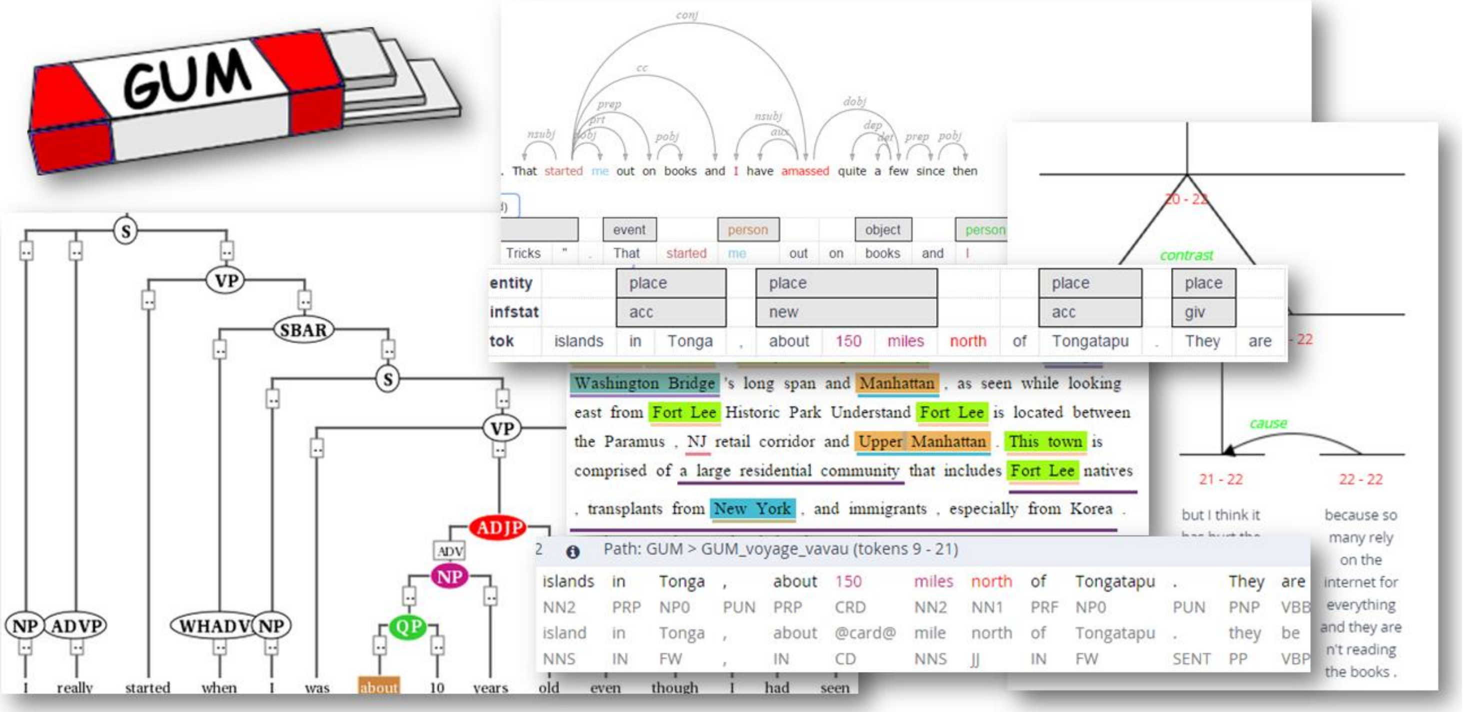
- How it is captured, becomes language DATA:
  - RAW data: spoken, written, visual
  - Experimental data
- Added (layers of) information:
  - Metadata
  - Transcription, time alignment
  - Acoustic measurements
  - Annotation
  - Coding

What are the differences?

Linguistic annotation: what types?

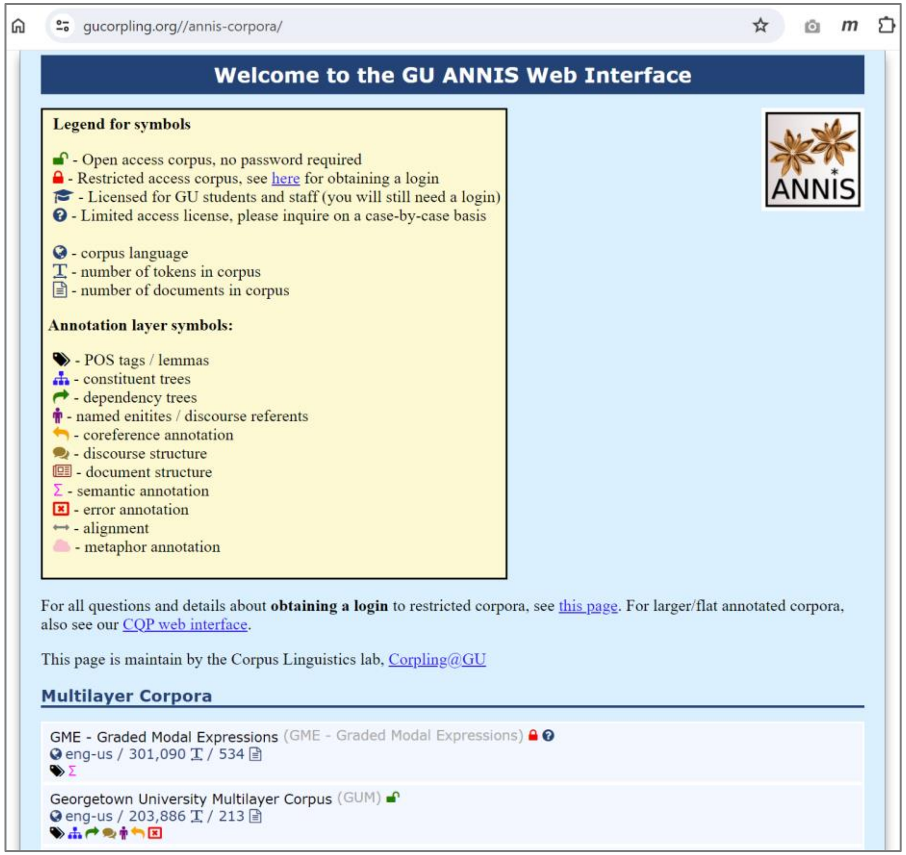
- What types of linguistic annotation have we seen so far?

- GUM: **The Georgetown University Multilayer Corpus**
  - https://gucorpling.org/gum/index.html
  - A corpus with *all* levels of linguistic knowledge annotated!!


GUM
conj
cc
prep
nsubj
dobj
pobj
That started me out on books and I have amassed quite a few since then
event
person
object
entity
infstat
tok
place
acc
new
giv
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 .
Path: GUM > GUM_voyage_vavau (tokens 9 - 21)
20 - 22
contrast
cause
21 - 22
22 - 22
but I think it
because so many rely on the internet for everything and they are n't reading the books .
S
VP
SBAR
NP
ADVP
WHADVP
ADJP
ADV
QP
I really started when I was about 10 years old even though I had seen

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's ANNIS Web Interface: https://gucorpling.org/annis-corpora/
    - Example queries on this page (scroll down): https://gucorpling.org/gum/search.html



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, analysis

🡐 Assists empirical linguistic inquiries

- Develop and evaluate (statistically based) NLP technologies

🡐 Becomes the basis of "language models" in NLP applications

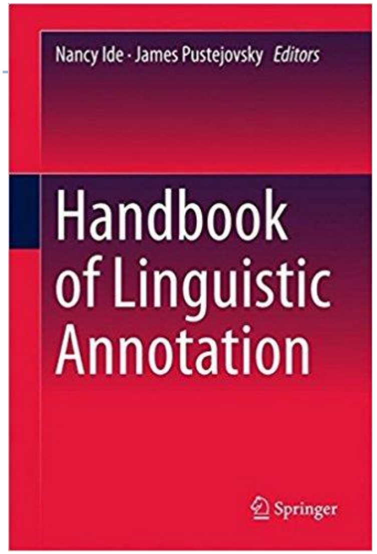
🡐 Linguistic annotation represents linguistic knowledge of humans that AI agents learn through machine learning, which they then mimic

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

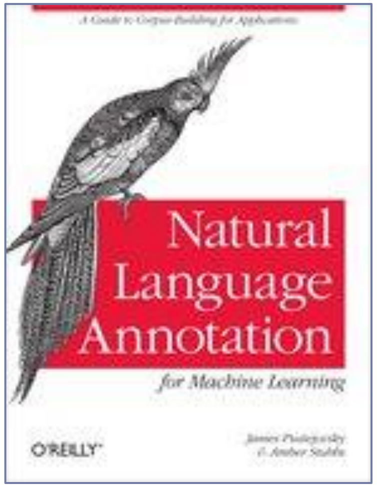


Nancy Ide · James Pustejovsky *Editors*

Handbook of Linguistic Annotation

Springer

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



A Guide to Corpus-Building for Applications

Natural Language Annotation

for Machine Learning

O'REILLY

James Pustejovsky & Amber Stubbs

Wrapping up

- To-do #9
  - Explore APLS (Archive of Pittsburgh Language and Speech)
- Friday
  - Guest presentation by Maya Asher, introduction to APLS project
- Your project
  - Work on it! Focus on DATA.