

Lecture 12: Annotation Projects

LING 1340/2340: Data Science for Linguists

Na-Rae Han

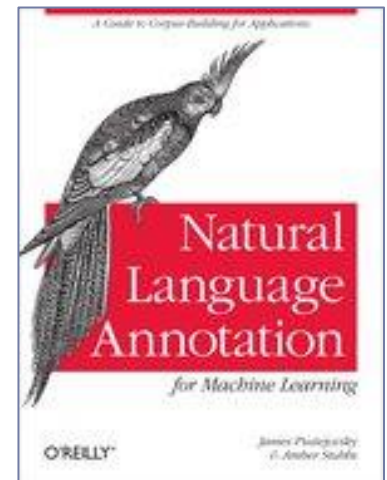
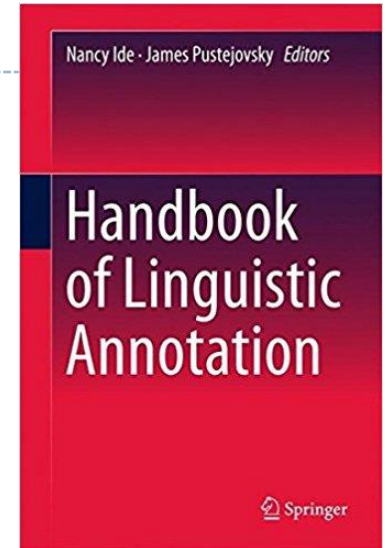
Objectives

- ▶ Linguistic annotation projects
 - ◆ Annotation tools
 - ◆ How to plan and run an annotation project
 - ◆ An anatomy of annotation project

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 platform, 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 semantic annotations. Sentence 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". Sentence 2: "The first African American to assume the presidency, he was previously the junior United States Senator from Illinois from 2005 to 2008." Sentence 3: "He served in the Illinois State Senate from 1997 until 2004." Annotations include named entities (PER, TIME, LOC), relations (subject, object, date of birth, position held, start time, end time), and a dropdown menu for "Illinois" with options like "Illinois Senate", "Illinois River", "Governor of Illinois", "Alton", "Illinois Country", and "Illinois Territory". A tooltip for "Illinois Senate" is visible at the bottom, defining it as the upper chamber of the Illinois General Assembly.

INCEpTION annotation interface

The screenshot displays the INCEpTION annotation interface. On the left, the 'Active Learning' panel shows a 'Session' with 'Layer' set to 'Named entity' and a 'Recommendation' for 'Illinois' with 'Label' 'LOC', 'Score' '1', and 'Delta' '1'. Below this is a 'Learning History' table with columns for text, label, and status.

Text	Label	Status
berkeley		skipped
Berkeley		skipped
Tesla	PER	accepted
Tesla	PER	accepted
Tesla	PER	accepted
Tesla	PER	accepted
Tesla	PER	accepted
Science	OTH	rejected
Tesla	PER	accepted

The main 'Annotation' panel shows three sentences with various entities and relations highlighted. Sentence 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 .'. Sentence 2: 'The first African American to assume the presidency, he was previously the junior United States Senator from Illinois from 2005 to 2008.'. Sentence 3: 'He served in the Illinois State Senate from 1997 until 2004.'. A dropdown menu for 'Illinois' is open, showing options like 'Illinois Senate', 'Illinois River', 'Governor of Illinois', etc.

Learning History Table:

Text	Label	Status
berkeley		skipped
Berkeley		skipped
Tesla	PER	accepted
Tesla	PER	accepted
Tesla	PER	accepted
Tesla	PER	accepted
Tesla	PER	accepted
Science	OTH	rejected
Tesla	PER	accepted

What sorts of functionality are needed in this tool?

brat

Not secure | weaver.nlp.lab.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).
PER LOC ORG
- 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.
MISC MISC
LOC
- 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.
ORG MISC
- 5 Para los restantes conjuntos, la Liga Mundial 2000 tendrá dos fines muy diferentes.
MISC
- 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.
ORG ORG ORG ORG ORG ORG
- 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.
ORG ORG ORG ORG ORG MISC
- 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.
ORG ORG ORG

An anatomy of annotation project

▶ Suppose you are tasked to start up an annotation project:

- Error annotation of a set of essays written by ESL learners
- Audio files of sociolinguistic interviews
- A set of videos featuring ASL content

▶ What should you be figuring out?

1. Annotation scheme
2. Physical representation + software tool
3. Annotation process
4. Evaluation and quality control
5. Usage

I agree greatly this topic mainly because I think that English becomes an official language in the not too distant. Now, many people can speak English or study it all over the world, and so more people will be able to speak English. Before the Japanese fall behind other people, we should be able to speak English, therefore, we must study English not only junior high school students or over but also pupils. Japanese education system is changing such a program. ...

Adapted from p.9 of Ide & Pustejovsky eds. (2017), *Handbook of Linguistic Annotation*

Annotation scheme

- Error annotation of a set of essays written by ESL learners
- Audio files of sociolinguistic interviews
- A set of videos featuring ASL content

1. Is there an underlying theory? What is it?
2. What features should be targeted and how should they be organized?
3. What is the process of annotation scheme development?
4. Should the potential use of the annotations inform development of the annotation scheme?
5. Will development of the scheme inform the development of linguistic theories or knowledge?

Physical representation

- Error annotation of a set of essays written by ESL learners
- Audio files of sociolinguistic interviews
- A set of videos featuring ASL content

1. How is the annotation represented? What **format**? Standards?
2. What are the reasons for the particular representation chosen?
 - ◆ What are the advantages/disadvantages of the chosen representation that may have come to light through its use?
 - ◆ Is the chosen format easily convertible into some other format down the line?
3. What **annotation software tools** are capable of handling them?

Linguistic annotation format: standardize?

- ▶ Ad-hoc formats mean different linguistic annotations are often incompatible
- ▶ Converting back and forth between them wastes resource
- ▶ Solution: Standardized format for linguistic annotation
- ▶ FoLiA: Format for Linguistic Annotation
 - ◆ <https://proycon.github.io/fofia/>
 - ◆ XML-based architecture
 - ◆ Software support, Python libraries etc.!

Example: semantic role

- ▶ [https://folia.readthedocs.io/en/latest/semrole notation.html](https://folia.readthedocs.io/en/latest/semrole%20notation.html)

```
24     <provenance>
25         <processor xml:id="p1" name="proycon" type="manual" />
26     </provenance>
27 </metadata>
28 <text xml:id="example.text">
29     <p xml:id="example.p.1">
30         <s xml:id="example.p.1.s.1">
31             <t>The Dalai Lama greeted him.</t>
32             <w xml:id="example.p.1.s.1.w.1"><t>The</t></w>
33             <w xml:id="example.p.1.s.1.w.2"><t>Dalai</t></w>
34             <w xml:id="example.p.1.s.1.w.3"><t>Lama</t></w>
35             <w xml:id="example.p.1.s.1.w.4"><t>greeted</t></w>
36             <w xml:id="example.p.1.s.1.w.5" space="no"><t>him</t></w>
37             <w xml:id="example.p.1.s.1.w.6"><t>.</t></w>
38         <semroles>
39             <predicate class="greet">
40                 <semrole class="agent">
41                     <wref id="example.p.1.s.1.w.2" />
42                     <wref id="example.p.1.s.1.w.3" />
43                 </semrole>
44                 <semrole class="patient">
45                     <wref id="example.p.1.s.1.w.5" />
46                 </semrole>
47             </predicate>
48         </semroles>
49     </s>
50 </p>
51 </text>
52 </FoLiA>
```

Annotation format

► To XML or not to XML?

◆ Gina Peirce's [Russian learner corpus](#)

```
▼ <essay>
  ▼ <tunit>
    Россия является частью Европы потому-что Россияни одеваются обычно по моде, так-же как другие
    страны Европы, и так-же многие считают что они более подобны белой Европе чем Азии.
  </tunit>
  ▼ <tunit>
    Политика в России отличается от Китая и например Индии.
  </tunit>
  ▼ <tunit>
    У нас нет систем
    <err cf="каст" pos="nn" gnd="fm" cs="g" num="pl" t="cs">касты</err>
    .
  </tunit>
  ▼ <tunit>
    Даже если Россия чуть опаздывает от Европы по моде или например
    <err cf="восточным" pos="adj" gnd="ms" num="pl" cs="d" t="cs num">восточная</err>
    услугам, у нас все равно есть просвещение в отличие от предыдущих времён.
  </tunit>
  ▼ <tunit>
    Язык у нас так-же полностью не похож на те-же Азиатские эроглифы.
  </tunit>
  ▼ <tunit>
    К мнению что основная часть России в Азии все равно не повод не считать Россиян Европейцами.
  </tunit>
</essay>
```

Annotation format

▶ Inline or stand-off?

- ◆ **Inline annotation** has annotations occurring alongside the text. Often used for describing a single structural element (ex. per-token)
 - ◆ Example: The Brown corpus, Gina Peirce's corpus
 - ◆ Pros: simple, self-contained. An XML parser is all you need.
 - ◆ Cons: May not be suitable for multi-layer annotations.
 - ◆ Folia page on In-line annotation:
https://folia.readthedocs.io/en/latest/inline_annotation_category.html
- ◆ **Stand-off annotation** has an annotation existing in a separate layer, typically as a separate file. Annotation points to an *offset* or a *span*.
 - ◆ Folia page on Span annotation:
https://folia.readthedocs.io/en/latest/span_annotation_category.html

Stand-off annotation: an example

- ▶ Original text: "Mia visited Seoul to look me up yesterday."

```
<maf xmlns:"http://www.iso.org/maf">
<seg type="token" xml:id="token1">Mia</seg>
<seg type="token" xml:id="token2">visited</seg>
<seg type="token" xml:id="token3">Seoul</seg>
<seg type="token" xml:id="token4">to</seg>
<seg type="token" xml:id="token5">look</seg>
<seg type="token" xml:id="token6">me</seg>
<seg type="token" xml:id="token7">up</seg>
<seg type="token" xml:id="token8">yesterday
</seg>
<pc>.</pc>
</maf>
```

Word tokens:
inline segmentation

```
<isoTimeML xmlns:"http://www.iso.org/isoTimeML">
<TIMEX3 xml:id="t0" type="DATE" value="2009-10-20"
functionInDocument="CREATION_TIME"/>
<EVENT xml:id="e1" target="#token2" class="OCCURRENCE" tense="PAST"/>
<EVENT xml:id="e2" target="#token5 #token7" class="OCCURRENCE" tense="NONE"
vForm="INFINITIVE"/>
<TIMEX3 xml:id="t1" type="DATE" value="2009-10-19"/>
<TLINK eventID="#e1" relatedToTime="#t0" relType="BEFORE"/>
<TLINK eventID="#e1" relatedToTime="#t1" relType="ON_OR_BEFORE"/>
<TLINK eventID="#e2" relatedToTime="#t1" relType="IS_INCLUDED"/>
</isoTimeML>
<tei-isoFSR xmlns:"http://www.iso.org/tei-isoFSR">
<fs xml:id="t0"><f name="Type" value="2009-10-20"/></fs>
</tei-isoFSR>
```

Time Event Annotation:
stand-off annotation

TimeML
annotation
standard

Annotation process

1. Will the annotation be done *manually, automatically*, or via some combination of the two?
2. Manual annotation:
 - ◆ How many annotators? Their background?
 - ◆ What annotation environment/platform will be used?
 - ◆ What are the exact steps? Multiple passes involving multiple annotators? Pipeline?
 - ◆ How will inter-annotator agreement be computed?
3. Automatic annotation:
 - ◆ What software will be used to generate the annotations?
 - ◆ How well does this software generally perform? Will it be a good fit with your data?
 - ◆ Any additional pre- or post-processing steps to enhance accuracy?

Evaluation and quality control

1. Systematic scaffolding to minimize human error?
2. By what method(s) will the quality of the annotations evaluated?
 - ◆ Inter-annotator agreement (IAA)
3. What is the threshold for the quality of annotations?

Inter-annotator agreement

- ▶ An important part of quality control
- ▶ Necessary to demonstrate the **reliability** of annotation.
- ▶ Common practices:
 - ◆ Create "**gold**" **annotation** (deemed "correct") to evaluate individual annotators' output against
 - ◆ Designate a portion of data to be annotated by **multiple annotators**, then measure **inter-annotator agreement**
 - ◆ **Pre-** and **post-adjudication** agreement: do disagreements persist after an adjudication process?

Inter-annotator agreement: factors

- ▶ Agreement rate depends on two main factors:
 - ◆ Quality of annotators: how well-trained the annotators are
 - ◆ Complexity of task: how difficult or abstract the annotation task at hand is, how easy it is to clearly delineate the category
- ← IMPORTANT because human agreement (esp. post-adjudication) is considered a **CEILING** for performance of machine-learning!

How much will humans agree?

- ▶ POS tagging
 - ◆ Via [Universal Dependency POS tagset](#)?
 - ◆ Using the [Penn Treebank tagset](#)?
- ▶ Syntactic tree bracketing for Penn Treebank
 - ◆ Reported to be about 88% (F-score)
- ▶ Scoring TOEFL essays, 0 to 5
 - ◆ Reported to be about 80% (Cohen's kappa)
 - ◀ Is there hope for automated essay grading?

Cohen's kappa

- ▶ Good or bad level of agreement?
 - ◆ **Case A:** Movie reviews are annotated as "rotten" or "fresh". Two annotators agree 70% of the time.
 - ◆ **Case B:** Tokens are labeled N, V, ADJ, ADV, P, DET. Two annotators agree 70% of the time.
- ▶ **Cohen's kappa (K) coefficient** is one of the most widely used measures of inter-annotator agreement.
 - ◆ Accounts for "chance" agreement.

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e}$$

P_o : observed agreement
 P_e : probability of chance agreement

P_e is **0.5** in Case A, **0.17** in Case B.

Case A:

$$K = (0.7 - 0.5) / (1 - 0.5) = \mathbf{0.4}$$

Case B:

$$K = (0.7 - 0.17) / (1 - 0.17) = \mathbf{0.64}$$

Weighted Cohen's kappa

- ▶ Good or bad level of agreement?
 - ◆ **Case B:** Tokens are labeled N, V, ADJ, ADV, P, DET. Two annotators agree 70% of the time.
 - ◆ **Case C:** Student essays are rated from 0 to 5. Two annotators agree 70% of the time.
 - ◆ Case B is **nominal**: no order among the labels, and C is **ordinal**: $0 < 1 < 2 < 3 < 4 < 5$
 - ◆ Case C: disagreement of 2 vs. 5 is worse than 2 vs. 3...
- ▶ Use **Weighted Cohen's kappa** for ordinal categories:

$$\kappa_w = 1 - \frac{\sum w_{ij} \cdot f_{oij}}{\sum w_{ij} \cdot f_{eij}}$$

Weighting factors

Observe frequencies

Expected frequencies

More here:
<https://datatab.net/tutorial/weighted-cohens-kappaer-annotator-agreement/>

Wrapping up

- ▶ Your project

- ◆ Progress Report #1 due this Friday!