

Lecture 7: Corpus Linguistics, Annotation

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

▶ Corpus linguistics

- ◆ Review of corpora and corpus tools
- ◆ Your own data plans for your project

▶ Linguistic annotation

- ◆ Types of linguistic annotation
- ◆ Annotation formats
- ◆ Annotation tools
 - ◆ Hands-on with Webanno
- ◆ Inter-annotator agreement

Corpus linguistics

- ▶ To-do #7 corpora and tools:
 - ◆ https://github.com/Data-Science-for-Linguists-2021/Class-Lounge/blob/main/corpora_tools_list.md
 - ◆ What exciting corpora and tools did you discover?

Your term project

- ▶ Your project is now on GitHub
 - ◆ <https://github.com/Data-Science-for-Linguists-2021>
- ▶ First progress report is due Thursday
 - ◆ 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! For now, include an appropriate amount of samples.

Data-mining web & social media

- ▶ Twitter sample corpus
 - ◆ Static corpus: download from the [NLTK data page](#)
- ▶ How does one data-mine Twitter?
 - ◆ Answer: through **API** (**Application Program Interface**)
 - ◆ Getting acquainted with JSON format
 - ◆ Tutorials on on the Learning Resource page
- ▶ Libraries used: [tweepy](#), [json](#)
- ▶ How did you like Twitter Mining?

Processing a static Twitter corpus

- ▶ "Twitter Samples" corpus can be downloaded from http://www.nltk.org/nltk_data/

```
In [3]: # One json object per line
jfile = 'D:/Corpora/twitter_samples/positive_tweets.json'
jlines = open(jfile).readlines()
jlines[0]
```

```
Out[3]: '{"contributors": null, "coordinates": null, "text": "#FollowFriday @France_Inte
e @PKuchly57 @Milipol_Paris for being top engaged members in my community this
week :)", "user": {"time_zone": "Paris", "profile_background_image_url": "htt
```

```
In [5]: # using json library to read line.
import json
json.loads(jlines[0])
```

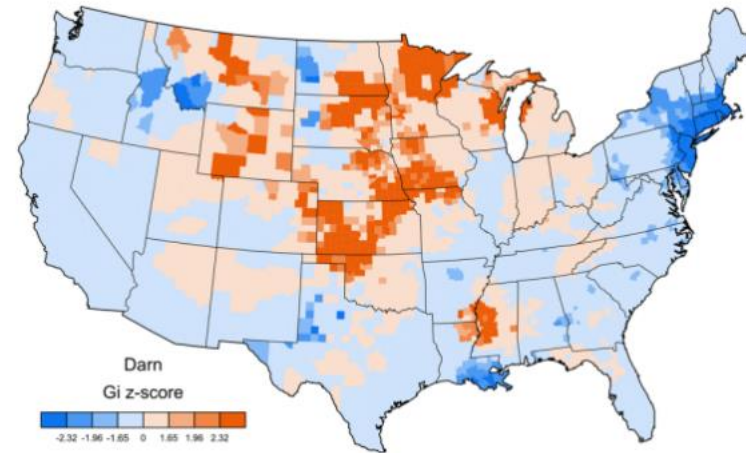
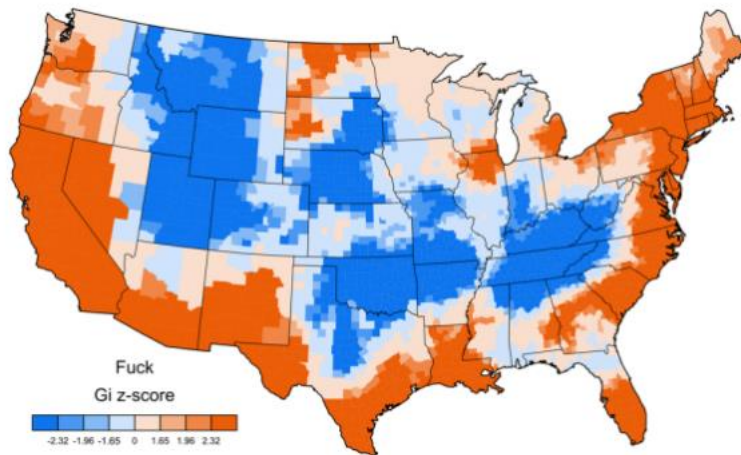
```
Out[5]: {'contributors': None,
'coordinates': None,
'created_at': 'Fri Jul 24 08:23:36 +0000 2015',
'entities': {'hashtags': [{'indices': [0, 13], 'text': 'FollowFriday'}]},
'symbols': [],
'urls': [],
'user_mentions': [{'id': 3222273608,
'id_str': '3222273608',
'indices': [14, 26],
'name': 'France International',
```

Web mining

- ▶ Involves "web crawling" "web spyder", ...
- ▶ **scrapy** is the most popular library.
 - ◆ <https://scrapy.org/>
 - ← You will have to install it first.
- ▶ You have collected a set of web pages. Now what?
 - ◆ A web page typically has tons of non-text, extraneous data such as headers, scripts, etc.
 - ◆ You will need to parse each page to extract textual data.
 - ◆ BeautifulSoup (bs4) is capable of parsing XML and HTML files.
- ▶ OK, so you've processed the web pages as data. Now what?
 - ◆ Linguistic analysis?

Mining social media for swear words

- ▶ <https://stronglang.wordpress.com/2015/07/28/mapping-the-united-swears-of-america/>
 - ◆ Jack Grieve mined Twitter and mapped prominent swear words by geographic regions within the US



Linguistic annotation

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: <ftp://ftp.cis.upenn.edu/pub/ircs/tr/01-10/01-10.pdf>

▶ As part of the NLP community, help keep linguistic knowledge representation in balance with engineering-side considerations

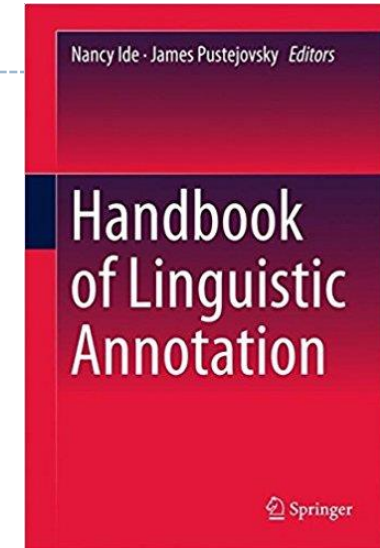
▶ Be a USER of linguistically annotated data by conducting empirical research

- ◆ An example: <https://web.stanford.edu/~bresnan/qs-submit.pdf>

All about Linguistic Annotation

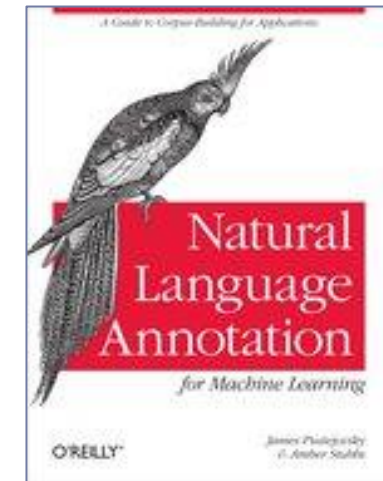
▶ *Handbook of Linguistic Annotation* (2017)

- ◆ 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>



POS tagsets

- ▶ There are multiple POS tagsets in use.
 - ◆ Some are larger, some are smaller.
- ▶ **The Brown Corpus tagset** (87 tags)
 - ◆ <http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM>
- ▶ In NLP, **the Penn Treebank tagset** (45 tags) has become de facto standard.
 - ◆ https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
- ▶ Lately, **"Universal" POS tagset** is gaining grounds
 - ◆ Next slide

Universal POS tags

- ▶ **"Universal" POS tagset** is gaining grounds
 - ◆ <http://universaldependencies.org/u/pos/>

| Open class words | Closed class words | Other |
|------------------|--------------------|--------------|
| <u>ADJ</u> | <u>ADP</u> | <u>PUNCT</u> |
| <u>ADV</u> | <u>AUX</u> | <u>SYM</u> |
| <u>INTJ</u> | <u>CCONJ</u> | <u>X</u> |
| <u>NOUN</u> | <u>DET</u> | |
| <u>PROPN</u> | <u>NUM</u> | |
| <u>VERB</u> | <u>PART</u> | |
| | <u>PRON</u> | |
| | <u>SCONJ</u> | |

- ▶ Tags mark the core POS categories; additional grammatical properties are relegated to features
- ▶ What do you think? Truly universal?

Syntactic annotation: the Penn Treebank

<http://languagelog ldc.upenn.edu/nll/?p=3594>

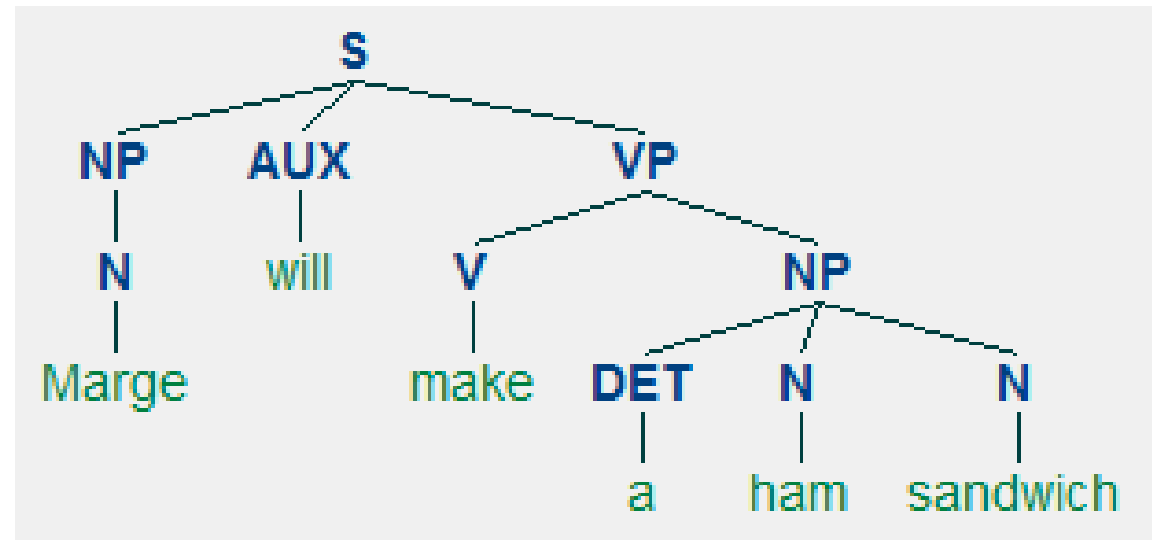
Penn Treebank is based upon **phrase structure grammar** framework

```
( (S
  (NP-SBJ
    (NP (NNP Pierre) (NNP Vinken) )
    ( , , )
    (ADJP
      (NP (CD 61) (NNS years) )
      (JJ old) )
    ( , , ) )
  (VP (MD will)
    (VP (VB join)
      (NP (DT the) (NN board) )
      (PP-CLR (IN as)
        (NP (DT a) (JJ nonexecutive) (NN director) ))
      (NP-TMP (NNP Nov.) (CD 29) )))
  ( . . ) ) )
( (S
  (NP-SBJ (NNP Mr.) (NNP Vinken) )
  (VP (VBZ is)
    (NP-PRD
      (NP (NN chairman) )
      (PP (IN of)
        (NP
          (NP (NNP Elsevier) (NNP N.V.) )
          ( , , )
          (NP (DT the) (NNP Dutch) (VBG publishing) (NN group) )))))
    ( . . ) ) ) )
```

Context-free grammar

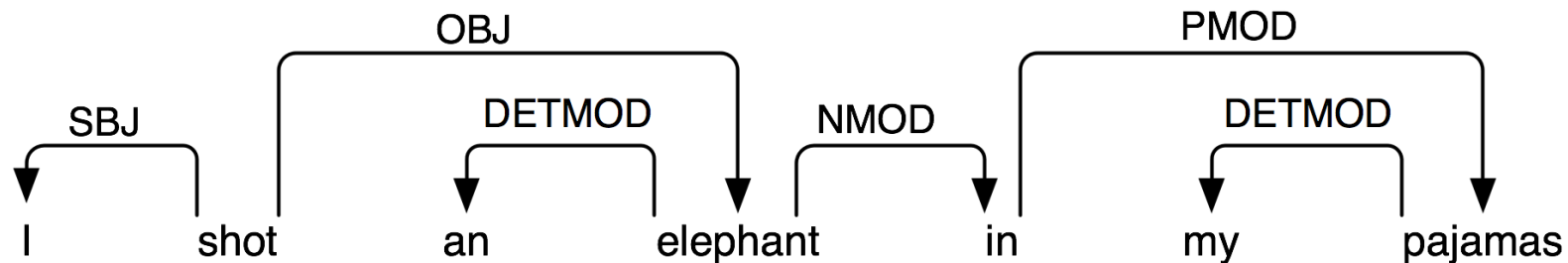
- ▶ Phrase-structure grammar is based upon constituency.
- ▶ Each local constituent can be expressed through **context-free grammar**.

```
S -> NP AUX VP
NP -> N
VP -> V NP
NP -> DET N N
N -> 'Marge'
AUX -> 'will'
V -> 'make'
DET -> 'a'
N -> 'ham' | 'sandwich'
```



A paradigm shift: dependency grammar

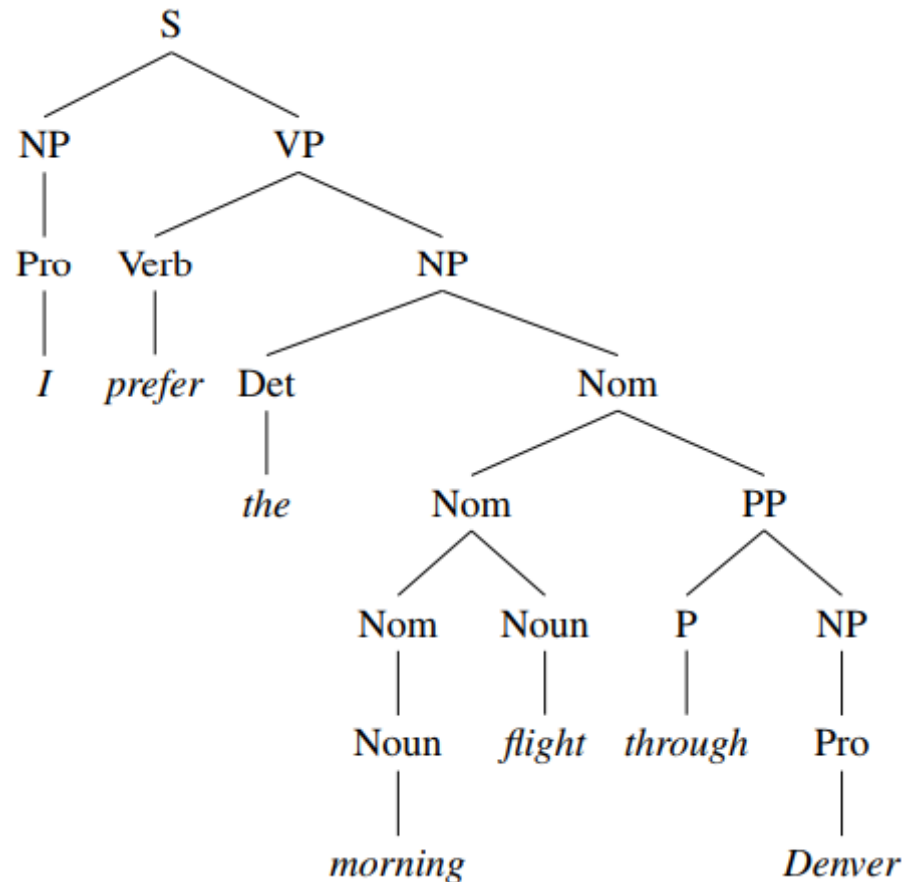
- ▶ **Phrase structure grammar** is all about **constituents**: phrasal units that words combine into.
- ▶ **Dependency grammar**, on the other hand, focuses on how words *relate* to other words: **dependency relation** between the **headword** and its **dependents**.



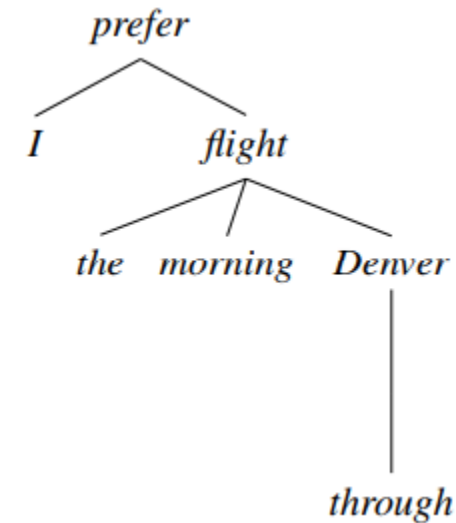
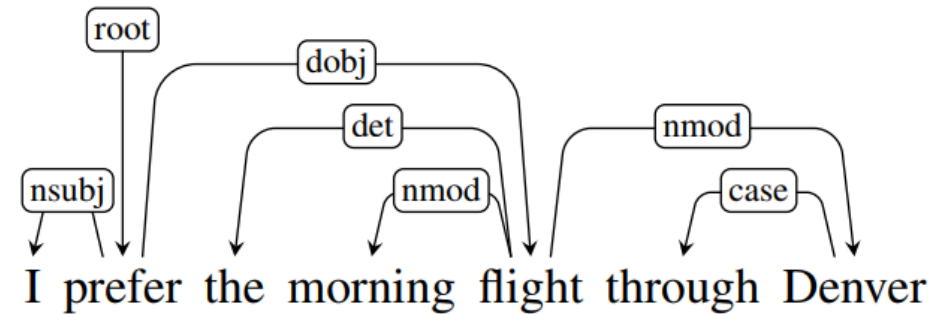
- ▶ NLTK book chapter: Dependency and Dependency Grammar
 - ◆ <http://www.nltk.org/book/ch08.html#dependencies-and-dependency-grammar>

A comparison

Constituency grammar



vs. Dependency grammar



Universal dependencies

- ▶ Dependency grammar and parsing have become increasingly popular.
- ▶ Dependency grammar is thought to be more suited to languages with flexible word order.
- ← Could it be a better candidate for **a truly universal grammar formalism**?
- ← Linguistic theory aside, does it offer an engineering-side advantage?

- ▶ **Universal Dependencies** working group
 - ◆ <http://universaldependencies.org/introduction.html>
 - ◆ A wide variety of languages represented!

Dependency annotation: format

- ▶ https://raw.githubusercontent.com/UniversalDependencies/UD_English-EWT/dev/en_ewt-ud-dev.conllu

```
# sent_id = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713-0002
# newpar id = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713-p0002
# text = President Bush on Tuesday nominated two individuals to replace retiring jurists on federal courts in the Washington area.
1   President      President      PROPN  NNP      Number=Sing  5      nsubj  5:nsubj  _
2   Bush    Bush    PROPN  NNP      Number=Sing  1      flat   1:flat  _
3   on      on      ADP    IN        _      4      case   4:case  _
4   Tuesday Tuesday PROPN  NNP      Number=Sing  5      obl    5:obl:on  _
5   nominated  nominate     VERB   VBD      Mood=Ind|Tense=Past|VerbForm=Fin  0      root   0:root  _
6   two      two      NUM    CD        NumType=Card  7      nummod 7:nummod  _
7   individuals individual  NOUN   NNS      Number=Plur  5      obj    5:obj  _
8   to      to      PART   TO        _      9      mark   9:mark  _
9   replace replace  VERB   VB        VerbForm=Inf  5      advcl  5:advcl:to  _
10  retiring  retire    VERB   VBG      VerbForm=Ger  11     amod   11:amod  _
11  jurists  jurist   NOUN   NNS      Number=Plur  9      obj    9:obj  _
12  on      on      ADP    IN        _      14     case   14:case  _
13  federal federal  ADJ    JJ        Degree=Pos   14     amod   14:amod  _
14  courts  court   NOUN   NNS      Number=Plur  11     nmod   11:nmod:on  _
15  in      in      ADP    IN        _      18     case   18:case  _
16  the     the     DET    DT        Definite=Def|PronType=Art  18     det    18:det  _
17  Washington Washington PROPN  NNP      Number=Sing  18     compound 18:compound  _
18  area    area    NOUN   NN        Number=Sing  14     nmod   14:nmod:in  SpaceAfter=No
19  .      .      PUNCT  .        _      5      punct  5:punct  _
```

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
 - ◆ [Brat Rapid Annotation Tool](#)
 - ◆ [WebAnno](#)

The screenshot displays the Brat Rapid Annotation Tool interface. At the top, there is a red header bar with the title "Annotation" and a "Home" button. To the right of the header, there are links for "Help", "narae", and "Log out (automatically in 24 min)". Below the header, there are several toolbars: "Document" (Open, Prev, Next, Export, Settings), "Page" (First, Prev, Go to, Next, Last), "Script" (LTR/RTL), "Help" (Guidelines), and "Workflow" (Reset, Finish). The main content area shows a document titled "Annotation Exercise To-do 6/annotation-example.tsv" with five sentences. Each sentence is accompanied by a dependency parse tree. The sentences are: 1. Ms. Haag plays Elianti .; 2. Rolls-Royce Motor Cars Inc. said it expects its U.S. sales to remain steady at; 3. about 1,200 cars in 1990 .; 4. The luxury auto maker last year sold 1,214 cars in the U.S.; 5. BELL INDUSTRIES Inc. increased its quarterly to 10 cents from seven cents a share .; 6. The new rate will be payable Feb. 15 . The interface also includes a "Layer" dropdown menu set to "POS" and a "No annotation selected!" message.

1 Por Viruca Atanes ^{PER} ^{LOC} Madrid, ^{ORG} 24 may (EFE).

2 -

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

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

5 Para los restantes conjuntos, la ^{MISC} Liga Mundial 2000 tendrá dos fines muy diferentes.

6 ^{ORG} Italia, defensor del título, ^{ORG} Brasil, ^{ORG} Cuba, ^{ORG} Estados Unidos, ^{ORG} Yugoslavia, ^{ORG} 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 : ^{ORG} España, ^{ORG} Argentina, ^{ORG} Francia, ^{ORG} Holanda y ^{ORG} Canadá, la ^{MISC} 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

^{ORG} Cuba, ^{ORG} Rusia y ^{ORG} Polonia.

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
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?
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
 - ◆ <http://proycon.github.io/folia/>
 - ◆ XML-based architecture
 - ◆ Software support, Python libraries etc.!

Wrapping up

- ▶ New topic: machine learning
 - ◆ Start learning!
- ▶ 1st progress report due on Thursday