Lecture 9: Annotation, data-mining web & social media

LING 1340/2340: Data Science for Linguists

Jevon Heath

Batch processing through shell scripting

- Your command line is actually running a programming environment: bash shell.
- You can *program* in command line, even for loops!

```
narae@T450s MINGW64 ~/Desktop/inaugural
$ for file in *.txt
> do
> iconv -f US-ASCII -t UTF-16 $file > try/$file
> echo $file complete
> done
1789-Washington.txt complete
1793-Washington.txt complete
1797-Adams.txt complete
1801-Jefferson.txt complete
1805-Jefferson.txt complete
1809-Madison.txt complete
1813-Madison.txt complete
1817-Monroe.txt complete
1821-Monroe.txt complete
1825-Adams.txt complete
```

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)
  - To-do #8
  - Getting acquainted with JSON format
  - Data Analysis using Twitter API and Python, The Code Way tutorial
  - And a couple more 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_Int
        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',
```

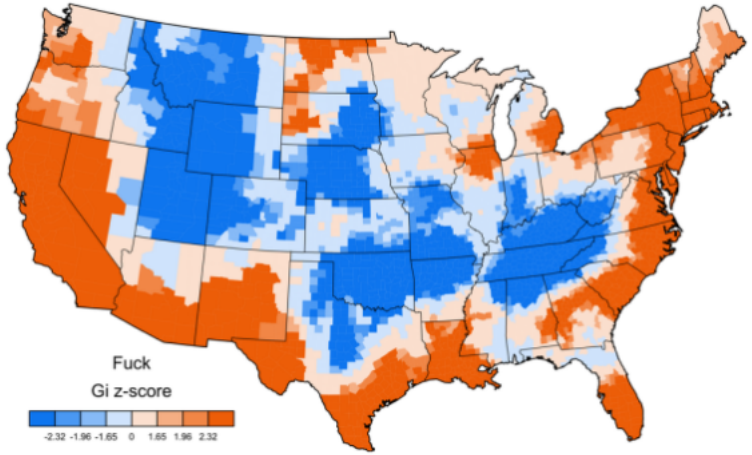
Web mining

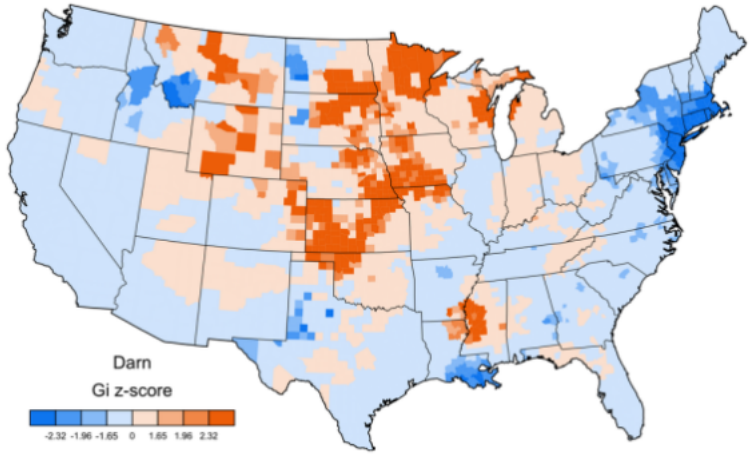
- Involves "web crawling" "web spyder", ...
- `scrapy` is the most popular library.
  - https://scrapy.org/

  ← You will have to install it first.
- Scrapy tutorial:
  - Official Scrapy:
    - https://doc.scrapy.org/en/latest/intro/tutorial.html
  - Digital Ocean:
    - https://www.digitalocean.com/community/tutorials/how-to-crawl-a-web-page-with-scrapy-and-python-3
- 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.
  - Beautiful Soup (bs4) is capable of parsing XML and HTML files.

Mining social media for swear words

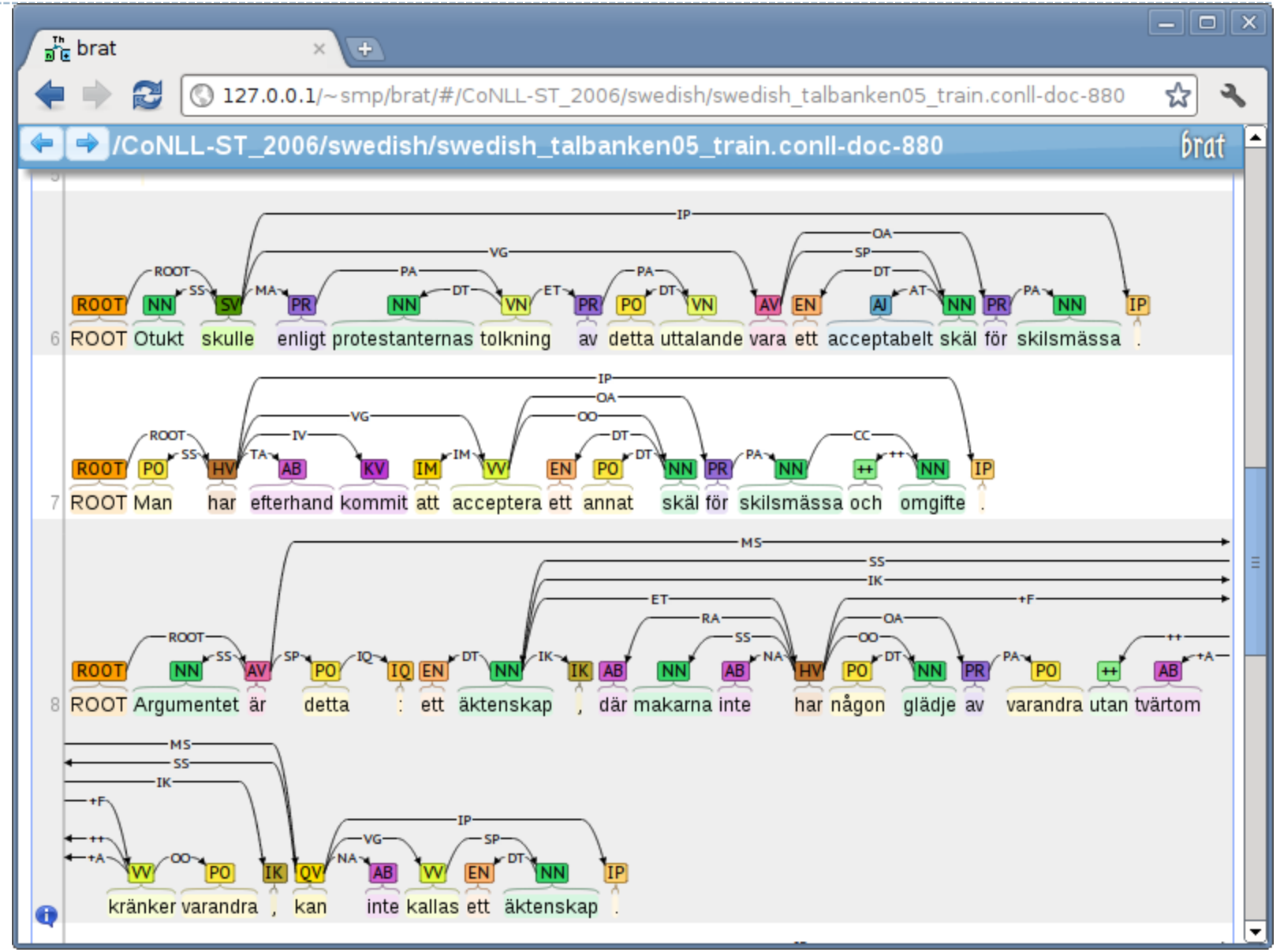
- https://stronglang.wordpress.com/2015/07/28/mapping-the-united-swears-of-america/
  - Jack Grieve mined Twitter and mapped prominent swear words by geographic regions within the US

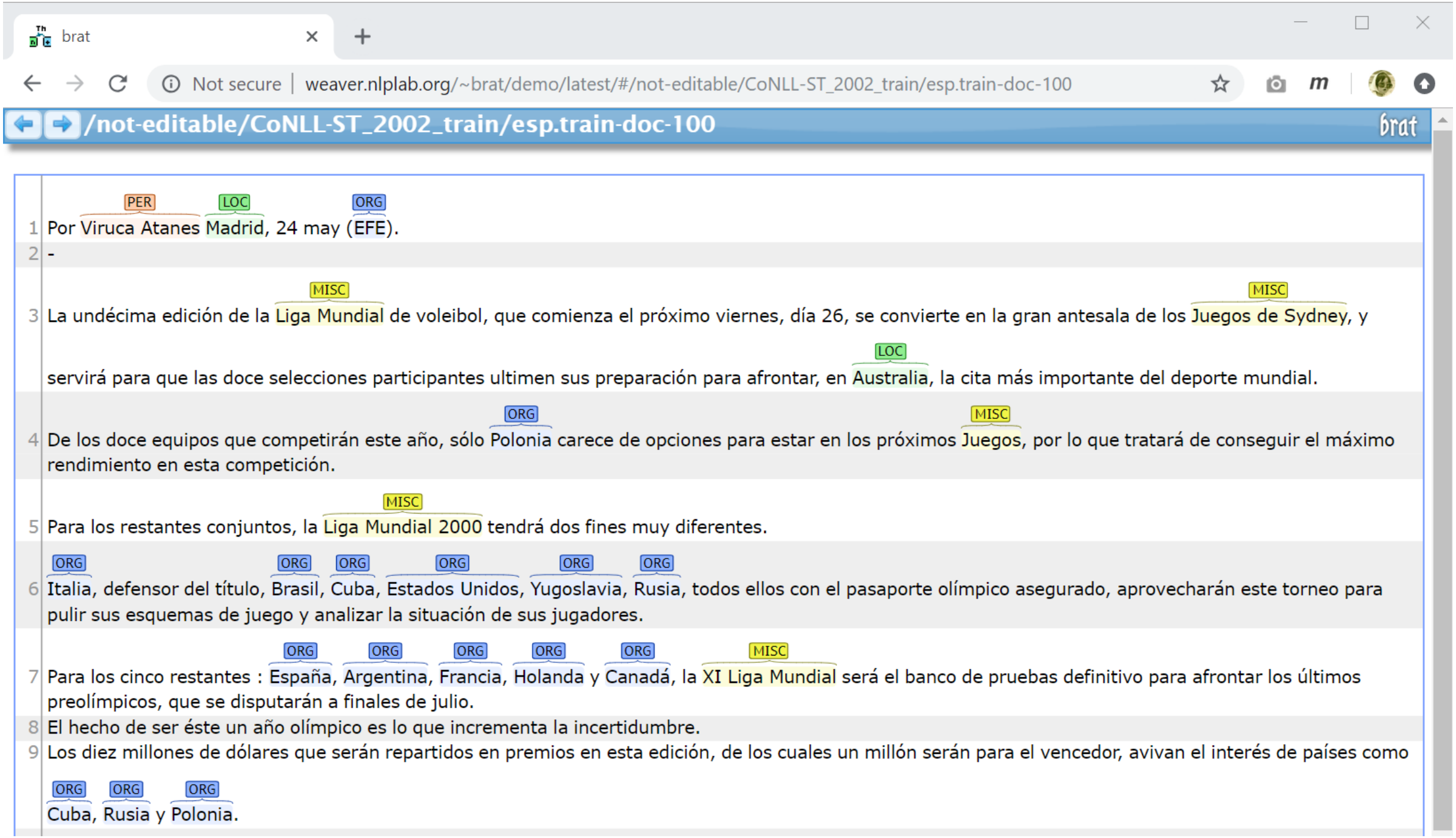




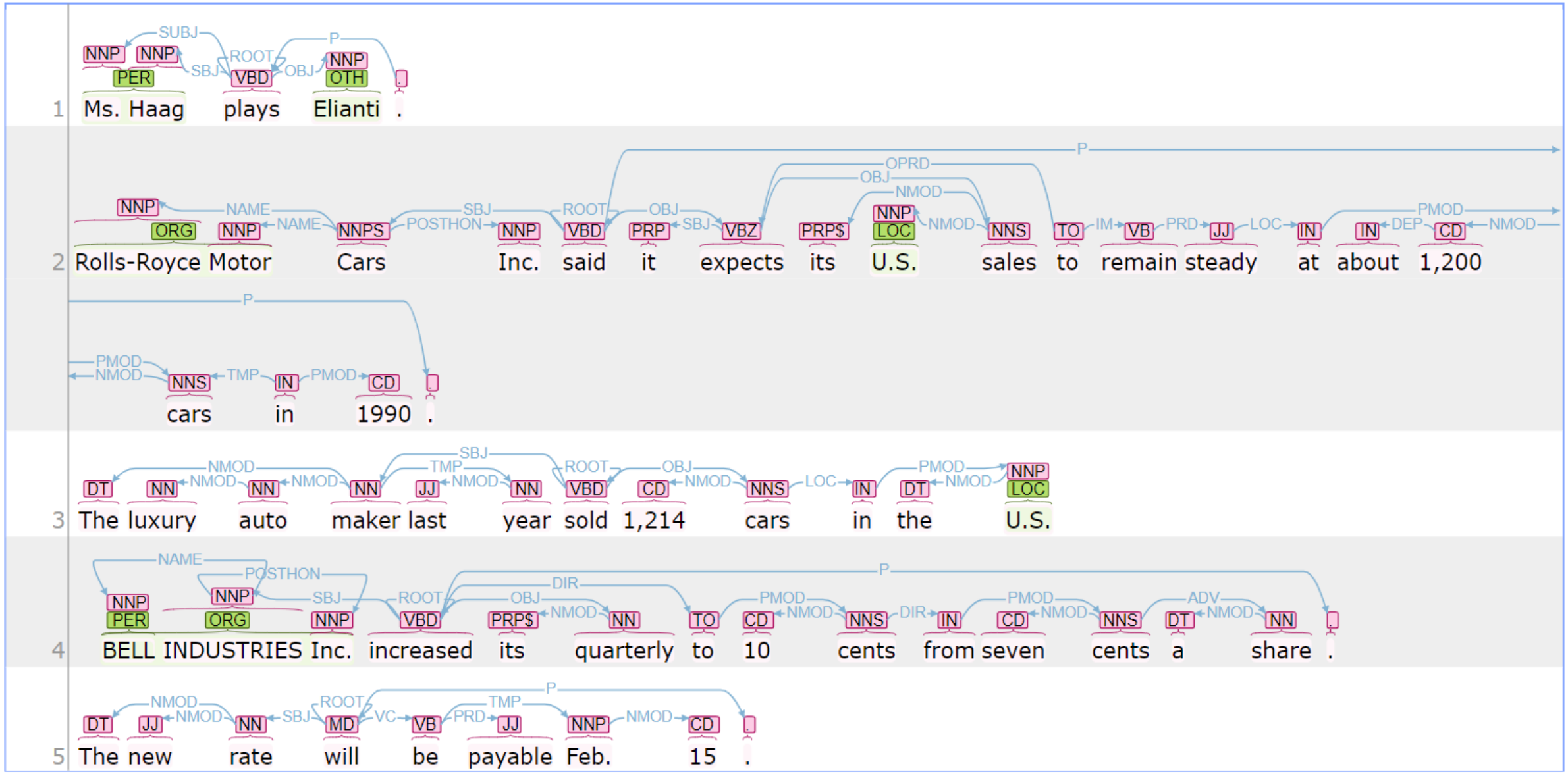
Back to annotation

- "brat" annotation interface




brat
Not secure | weaver.nlplab.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).
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.
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.
5 Para los restantes conjuntos, la Liga Mundial 2000 tendrá dos fines muy diferentes.
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.
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.
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.

Annotation


1 Ms. Haag plays Elianti .
2 Rolls-Royce Motor Cars Inc. said it expects its U.S. sales to remain steady at about 1,200 cars in 1990 .
3 The luxury auto maker last year sold 1,214 cars in the U.S.
4 BELL INDUSTRIES Inc. increased its quarterly to 10 cents from seven cents a share .
5 The new rate will be payable Feb. 15 .

Dependency annotation: format

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

```
# sent_id = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713-0002
# 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	_
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	_
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	_
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	SpaceAfter=No
19	.	.	PUNCT	.	_	5	punct	5:punct	_
```

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

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?

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 дру
    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">восточныя</err>
    услугам, у нас все равно есть просвещение в отлицие от предедущих времён.
  </tunit>
 ▼<tunit>
    Язык у нас так-же полнастью не похож на те-же Азиатские эроглифы.
  </tunit>
 ▼<tunit>
    К мнению что основная часть России в Азии все равно не повод не считать Россиян Европейцам
  </tunit>
</essay>
```

Annotation format

- Inline or stand-off?
  - **Inline annotation** has annotations occurring alongside the text.
    - 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.
  - **Stand-off annotation** has an annotation existing in a separate layer, typically as a separate file. Annotation points to an *offset* or a *span*.

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

Annotation process

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

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?

Evaluation and quality control

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

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**: Student essays are rated from 0 to 5. 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) = **0.4**
Case B:
K = (0.7 - 0.17) / (1 - 0.17) = **0.64**

Usage

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

1. By what means and under what conditions will the data be available to users?
2. What are the expected usages of the annotated data?
3. Will the data be used for machine learning, and if so what types of task?

Wrapping up

- New topic: machine learning
  - Start learning!

- 1st progress report due on Tuessday