Lecture 10: Data Formats

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

- Linguistic annotation: inter-annotator agreement
  - ← In Lecture 9 slides, we'll wrap up
- Corpus data: standard and popular formats
  - Review of common data formats

Data standards & exchange formats

	What	Notes, reference
CSV	Comma-separated values	For tabular data. Compatible with Excel
TSV	Tab-separated values	
HTML	Web pages	Not meant as data format
XML	For markup and text encoding	A Gentle Introduction to XML by TEI
JSON	JavaScript Object Notation (Twitter, Jupyter Notebook)	Introducing JSON JSON example (vs. XML)

Let's review examples you found in To-do #10!

What makes these formats true STANDARDS?

Not true standards, but widely used in NLP and linguistics

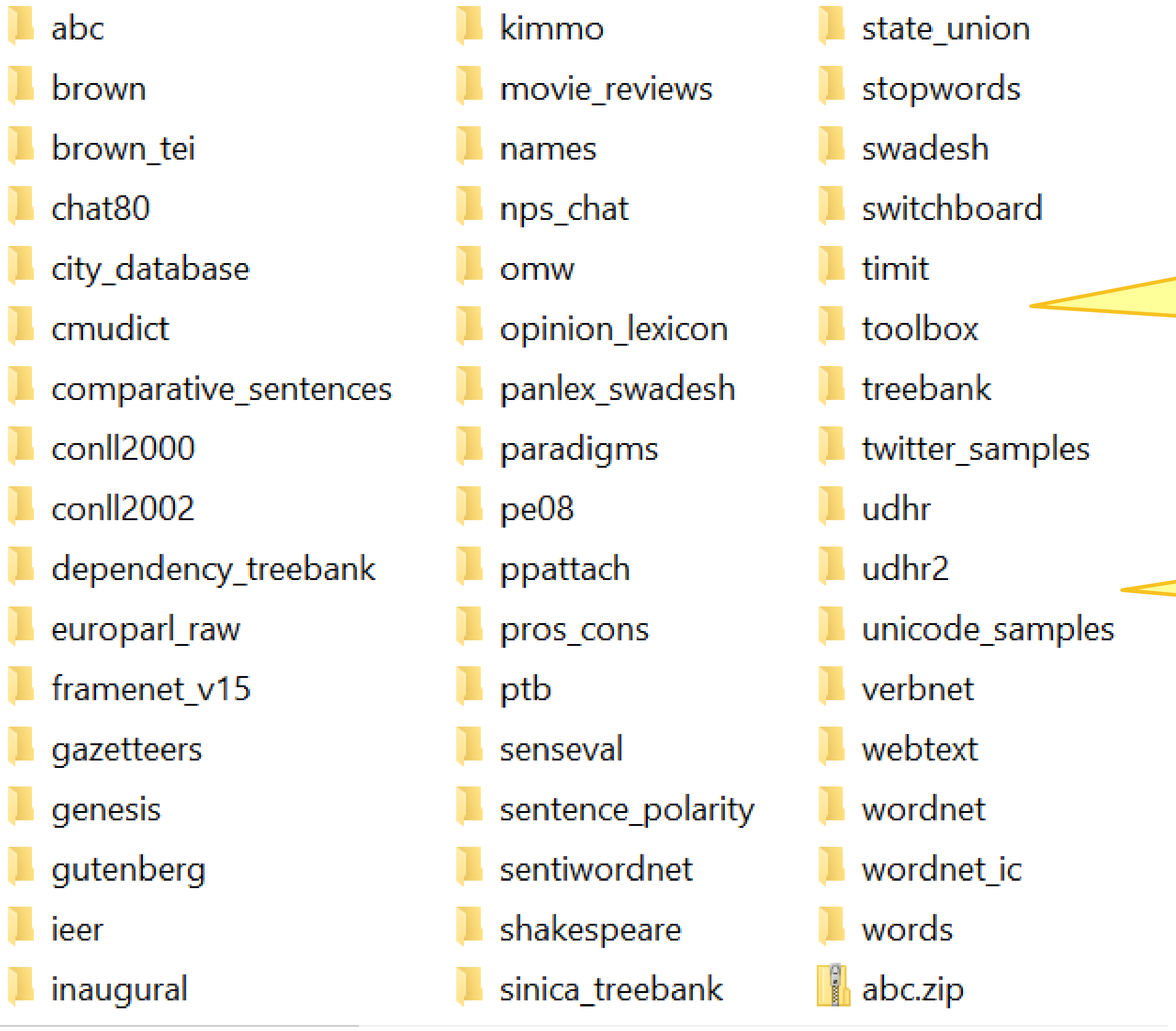
	What	Notes, reference
Penn Treebank	Constituent syntax trees with ( ) bracketing	We worked with these trees in LING 1330
CoNLL-U	One token per line, with multiple tab-separated columns for dependency syntax, lemma, POS, etc.	Common in NLP, used by Universal Dependencies project and more
TextGrid	PRAAT's native data format	For speech data! Time stamps, etc.
CHAT	CMU's CHILDES and TALKBANK projects, and more	Popular in SLA (second language acquisition)

... and many, many more!

They are all TEXT files.

- Underneath it all, these files are all TEXT files with **special formatting syntax** and **special characters** designated for formatting purposes.
  - In command line, you can `cat` and `less` through the files. Also: `head`, `tail`
  - You can open them up in a **text editor** (Atom, Notepad++) and edit.
  - Some editors/applications are aware of format-specific syntax and will display **rendered view** (cf. "source view"). Some will apply **syntax highlighting**.
    - Unlike, say, PDF files, style attributes are NOT part of the source data files themselves. (e.g., markdown file)
- As TEXT files, they have two important aspects:
  - **Encoding**: Latin-1 (=ISO-8859-1), ASCII, UTF-8, UTF-16, CP-1252 (Windows-1252), ANSI...
  - **Line endings**: **LF** (`'\n'`: OS X & Linux), **CRLF** (`'\r\n'`: Windows)

A brief tour of NLTK's many "corpus" data



- abc
- brown
- brown_tei
- chat80
- city_database
- cmudict
- comparative_sentences
- conll2000
- conll2002
- dependency_treebank
- europarl_raw
- framenet_v15
- gazetteers
- genesis
- gutenberg
- ieer
- inaugural
- kimmo
- movie_reviews
- names
- nps_chat
- omw
- opinion_lexicon
- panlex_swadesh
- paradigms
- pe08
- ppattach
- pros_cons
- ptb
- senseval
- sentence_polarity
- sentiwordnet
- shakespeare
- sinica_treebank
- state_union
- stopwords
- swadesh
- switchboard
- timit
- toolbox
- treebank
- twitter_samples
- udhr
- udhr2
- unicode_samples
- verbnet
- webtext
- wordnet
- wordnet_ic
- words
- abc.zip

Many of them are language data, not corpora per se

Diverse genres and data formats represented!

Wrapping up

- No To-do out
  - Work on your project!
- Your project: 1st Progress Report
  - Due date moved to next Wednesday
  - Data work typically takes MUCH LONGER! Start NOW.