COR: Corpus Linquistics

Markup and Annotation

Francis Bond
Department of Asian Studies
Palacký University
https://fcbond.github.io/
bond@ieee.org

https://github.com/bond-lab/Corpus-Linguistics

Overview

- Revision of Introduction
  - What is Corpus Linguistics
- Mark-up
- Annotation
- Regular Expressions
- **Lab One**

Revision

What is a Corpus?

corpus (pl: ***corpora***):

- In linguistics and lexicography, a body of texts, utterances, or other specimens considered more or less representative of a language, and usually stored as an electronic database.
  - machine-readable (i.e., computer-based)
  - authentic (i.e., naturally occurring)
  - sampled (bits of text taken from multiple sources)
  - representative of a particular language or language variety.

- Sinclair's (1996) definition:

  A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.

Why Are Electronic Corpora Useful?

- as a collection of examples for linguists
  - intuition is unreliable
- as a data resource for lexicographers
  - use natural data to exemplify usage
- as instruction material for language teachers and learners
- as training material for natural language processing applications
  - training of speech recognizers, parsers, MT

The British National Corpus (BNC)

- 100 million words of written and spoken British English
- Designed to represent a wide cross-section of British English from late 20th century: balanced and representative
- POS tagging (2 million word sampler hand checked)

Written	Domain	Date	Medium
(90%)	Imaginative (22%)	1960-74 (2%)	Book (59%)
	Arts (8%)	1975-93 (89%)	Periodical (31%)
	Social science (15%)	Unclassified (8%)	Misc. published (4%)
	Natural science (4%) ...		Misc. un-pub (4%)
Spoken	Region	Interaction type	Context-governed
(10%)	South (46%)	Monologue (19%)	Informative (21%)
	Midlands (23%)	Dialogue (75%)	Business (21%)
	North (25%) ...	Unclassified (6%)	Institutional (22%) ...

General vs. specialized corpora

- General corpora (such as "national" corpora) are a huge undertaking. These are built on an institutional scale over the course of many years.

- Specialized corpora (ex: corpus of English essays written by Japanese university students, medical dialogue corpus) can be built relatively quickly for the purpose at hand, and therefore are more common

- Characteristics of corpora:

  1. Machine-readable, authentic
  2. Sampled to be balanced and representative

- Trend: for specialized corpora, criteria in (2) are often weakened in favor of quick assembly and large size
  Rare phenomena only show up in large collections

Mark Up

Mark-Up vs. Corpus Annotation

- **Mark up** provides objectively verifiable information
  - Authorship
  - Publication dates
  - Paragraph boundaries
  - Source text (URL, Book, ...)

- **Annotation** provides interpretive linguistic information
  - Sentence/Utterance boundaries
  - Tokenization
  - Part-of-speech tags, Lemmas
  - Sentence structure
  - Domain, Genre

Many people use the terms interchangeably.

The Need for Corpus Mark-Up

Mark up and Annotation guidelines are needed in order to facilitate the accessibility and reusability of corpus resources.

- Minimal information:
  - authorship of the source document
  - authorship of the annotated document
  - language of the document
  - character set and character encoding used in the corpus
  - conditions of licensing

Dublin Core Ontology

- Goals
  - Provides a semantic vocabulary for describing the "core" information properties of resources (electronic and "real" physical objects)
  - Provide enough information to enable intelligent resource discovery systems
- History
  - A collaborative effort started in 1995
  - Initiated by people from computer science, librarianship, on-line information services, abstracting and indexing, imaging and geospatial data, museum and archive control.

Dublin Core - 15 Elements

- Content (7)
  - Title, Subject, Description, Type, Source, Relation and Coverage
- Intellectual property (4)
  - Creator, Publisher, Contributor, Rights
- Instantiation (4)
  - Date, Language, Format, Identifier

Dublin Core – discussion

- Widely used to catalog web data
  - OLAC: Open Language Archives Community
  - LDC: Linguistic Data Consortium
  - ELRA: European Language Resources Archive
  - ...

International Standard Language Resource Number (ISLRN)

- An attempt to give Language Resources (such as corpora) a unique Identifier
- Like ISBNs for books

  "The main purpose of the metadata schema used in ISLRN, is the identification of LRs. Inspired by the broadly known OLAC schema, a minimal set of metadata was chosen to ensure that any resource can be correctly distinguished and identified. We emphasize on the simplicity of the fields, that are easy and quick to fill in beyond misunderstanding." http://www.islrn.org/basic_metadata/

- Only a distributor, creator or rights holder for a resource is entitled to submit it for ISLRN assignment.

Task: Pick a corpus, write its MetaData

ISLRN Metadata schema

Metadata	Description	Example
Title	The name given to the resource.	1993-2007 United Nations Parallel Text
Full Official Name	The name by which the resource is referenced in bibliography.	1993-2007 United Nations Parallel Text
Resource Type	The nature or genre of the content of the resource from a linguistic standpoint.	Primary Text
Source/URL	The URL where the full metadata record of the resource is available.	https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/FinClarinSiteUEF
Format/MIME Type	The file format (MIME type) of the resource. Examples: text/xml, video/mpeg, etc.	text/xml
Size/Duration	The size or duration of the resource.	21416 KB
Access Medium	The material or physical carrier of the resource.	Distribution: 3 DVDs
Description	A summary of the content of the resource.	
Version	The current version of the resource.	1.0
Media Type	A list of types used to categorize the nature or genre of the resource content.	Text
Language(s)	All the languages the resource is written or spoken in.	eng (English)
Resource Creator	The person or organization primarily responsible for making the resource.	Ah Lian; NTU; lian@ntu; Singapore
Distributor	The person or organization responsible for making the resource available.	Ah Beng; NUS; Beng@nus; Singapore
Rights Holder	The person or organization owning or managing rights over the resource.	Lee, KY; Gov; LKY@gov; Singapore
Relation	A related resource.	

Annotation

Geoffrey Leech's Seven Maxims of Annotation

1. It should be possible to remove the annotation from an annotated corpus in order to revert to the raw corpus.

2. It should be possible to extract the annotations by themselves from the text. This is the flip side of maxim 1. Taking points 1. and 2. together, the annotated corpus should allow the maximum flexibility for manipulation by the user.

3. The annotation scheme should be based on guidelines which are available to the end user.

4. It should be made clear how and by whom the annotation was carried out.

5. The end user should be made aware that the corpus annotation is not infallible, but simply a potentially useful tool.

6. Annotation schemes should be based as far as possible on widely agreed and theory-neutral principles.

7. No annotation scheme has the a priori right to be considered as a standard. Standards emerge through practical consensus.

Types of Corpus Annotation

- Tokenization, Lemmatization
- Parts-of-speech
- Syntactic analysis
- Semantic analysis
- Discourse and pragmatic analysis
- Phonetic, phonemic, prosodic annotation
- Error tagging

How is Corpus Annotation Done?

- Three ways:

  1. Manual (done entirely by human annotators)
  2. Semi-automatic (done first by computer programs; post-edited)
  3. Automatic (done entirely by computer programs)

- Labor intensive: $1 > 2 > 3$

- Some types of annotation can be reasonably reliably produced by computer programs alone

  - Part-of-speech tagging: accuracy of 97%
  - Lemmatization

  Computer programs for other annotation types are not yet good enough for fully automatic annotation

Lemmatization

- *unit* and *units* are two word-forms belonging to the same lemma ***unit***.

- Same lemmas are shared for:
  - Plural morphology: *unit/units*: ***unit***, *child/children*: ***child***
  - Verbal morphology: *eat/eats/ate/eaten/eating*: ***eat***
  - Comparative/superlative morphology of adjectives
  - *many/more/most*: ***many***, *slow/slower/slowest*: ***slow***
    but also *much/more/most*: ***much***

- Lemmatization does not affect:
  - derived words that belong to different part-of-speech groups
  - *quick/quicker/quickest*: ***quick***, *quickly*: ***quickly***
  - *Korea*: ***Korea***, *Korean*: ***Korean***

- Lemmatization for English can be performed reasonably reliably and accurately using automated programs: as long as the POS is correct.

- Why do we need lemmatization?

Part-of-Speech (POS) Tagging

- (POS) Tagging: adding part-of-speech information (tag) to words *Colorless/JJ green/JJ ideas/NNS sleep/VBP furiously/RB ./.*
- Useful in searches that need to distinguish between different POSs of same word (ex: *work* can be both a noun and a verb)
- In the US, the Penn Treebank POS set is de-facto standard:
  - `http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html`
  - 45 tags (including punctuation)
- In Europe, CLAWS tagset is popular (also used by **`english-corpora.org`**):
  - `http://ucrel.lancs.ac.uk/claws7tags.html`
  - 137 tags (without punctuation)

- For English, many POS taggers with good performance are available for automated corpus annotation:
  - CLAWS (Lancaster University, 96-97% accuracy)
  - TnT tagger/Hunpos (Saarland University, 94-97% accuracy)
  - Averaged Perceptron (NLTK)/RNNs do a little better

Penn Treebank Examples

Tag	Description	Tag	Description
NN	Noun, singular or mass	VB	Verb, base form
NNS	Noun, plural	VBD	Verb, past tense
NNP	Proper noun, singular	VBG	Verb, gerund or present participle
NNPS	Proper noun, plural	VBN	Verb, past participle
PRP	Personal pronoun	VBP	Verb, non-3rd person singular present
IN	Preposition	VBZ	Verb, 3rd person singular present
TO	*to*	.	Sentence Final punct (*.,?,!*)

- The tags include inflectional information
  - If you know the tag, you can generally find the lemma
- Some tags are very specialized: I/PRP wanted/VBD to/TO go/VB ./.

Universal Tagset

Tag	Explanation	Example
VERB	verb (all tenses and modes)	
AUX	auxiliary	have, be
NOUN	nouns (common and proper)	
PRON	pronouns	
PROPN	proper noun	Francis
ADJ	adjectives	
ADV	adverbs	
ADP	adpositions (prepositions and postpositions)	
CCONJ	coordinating conjunction	
SCONJ	subordinating conjunction	
NUM	cardinal numbers	twenty-four, fourth, 1991, 14:24
PART	particles or other function words	at, on, out, over per, that, up, with
INTJ	Interjection	Ahoj!
SYM	symbol	
X	other: foreign words, typos, abbreviations	ersatz, esprit, dunno, gr8, univeristy
PUNCT	. , ; ! punctuation	

Task UPOS

Find a tagset in a language you speak — map it to the universal tagset. (format tag—utag e.g. NNS—Noun). Try to do the most exotic language you know.

- Make a group of two or three
- Pick a language (good if we can have two or more groups)
- Find a tagset online (with some documentation)
  - try not to look at existing mappings
- Map it to UPOS
- See if you can find a mapping online to compare to.

30 minutes?

Parsing (Syntactic Annotation)

- Parsing: adding phrase-structure information (parse) to sentences

```
(S (NP (N Claudia))
   (VP (V sat)
       (PP (P on)
           (NP (AT a)
               (N stool)))))
```

- Useful for corpus investigation of grammatical structures
- Parsed corpora are sometimes known as treebanks
  - The Penn Treebank, Hinoki Treebank, ...
- Parsed corpora are used for "training" automated parsers
- Stanford Parser and CMU parser were trained on the Penn Treebank corpora

Corpus vs. Annotation Software

- The two help each other. How?

  1. An annotated corpus is built, entirely by humans
  2. Then a computer program is **trained** on this corpus
  3. Now new corpora can be automatically annotated using this program

- In practice, normally start off with a simple program and then correct its output

Training a Parser/Learning a Model

- A corpus can be used to **train** a computer program. A program **learns** from corpus data. What does this mean?
  - *work* is a noun (NN) in some contexts, and a verb (VB) in some others.
  - When *work* follows an adjective (ADJ), it is likely to be a noun.
  - When *work* follows a plural noun (NNS), it is likely to be a verb.
    * *nice/ADJ work/NN*, *beautiful/ADJ work/NN*
    * *they/NNS work/VB at a hospital*
    * *my parents/NNS work/VB too much*

    These patterns can be extracted from a corpus, and the "trained" computer program makes a statistical model with them to predict the POS of *work* in a new text

Semantic Annotation

- Word sense disambiguation between homonyms
- ex. *lie* in:
  - *The boy lies$_1$ to his parents.*
  - *Mary lies$_2$ down after lunch for a nap.*
- ex. *share* in:
  - *They all did their share$_1$ of the work.*
  - *The Twitter share$_2$ holders were disappointed.*

Semantic Role Labeling

- A semantic role is the relationship that a syntactic constituent has with a predicate: e.g., Agent, Patient, Instrument, Locative, Temporal, Manner, Cause, ...

- An example from the PropBank corpus:

  [A0 He] [AM-MOD would] [AM-NEG n't] [V accept] [A1 anything of value] from [A2 those he was writing about] .

  - V: verb
  - A0: acceptor
  - A1: thing accepted
  - A2: accepted-from
  - A3: attribute
  - AM-MOD: modal
  - AM-NEG: negation

Time Annotation

- Temporal expressions tell us:
  - When something happened
  - How long something lasted
  - How often something occurs
- Examples
  - He wrapped up a three-hour meeting with the Iraqi president in Baghdad today.
  - The king lived 4,000 years ago.
  - I'm a creature of the 1960s, the days of free love.

Discourse and Pragmatic Annotation

- Annotating for discourse information (usually spoken dialogue corpora):
  - speech acts (ex. accept, acknowledge, answer, confirm, correct, direct, echo, exclaim, greet)
  - speech act forms (ex. declarative, yes-no question, imperative, etc.)

- Coreference annotation:
  - keep track of an entity that is mentioned throughout a text
    - ex: Kim_4 said ... Sandy_6 told him_4 that she_6 would ...
    - ex: <COREF ID='100'>The Kenya Wildlife Service</COREF> estimates that <COREF ID="101" TYPE=IDENT REF='100'>it</COREF> loses $1.2 million a year in park entry fee...

Error Tagging

- Error tagging is often done on learner corpora
- Cambridge Learner Corpora (CLC) and the Longman Learner's Corpus are tagged for errors, as well as numerous other learner corpora
- Error types used for CLC include:
  - wrong word form used
  - something missing
  - word/phrase that needs replacing
  - unnecessary word/phrase
  - wrongly derived words

- Example:

  ```
  For example, my friend told me if I knew about Shakespare.
  But, <TIP id=17-56 etype=24 tutor="I knew">I know</TIP>
  about him <TIP id=17-57 etype=10 tutor="a little bit">little
  bit</TIP>, so I couldn 't <TIP id=17-56679-5 etype=15
  tutor="explain it">explain</TIP> fairly to her.
  ```

Inter Annotator Agreement

➢ For most annotation we don't know the answer

Q How can we test whether the annotation is correct (and reproducible)?

A Tag with multiple annotators, measure the agreement

Approaches to Annotation

- Multiple annotators, discard outliers
  - Check for weird distribution
  - Often crowd-sourced quality control is an issue
- Multiple annotators, majority tag
- Two annotators, adjudication for disputes
- Single annotator, adjudication vs model
- Single annotator

Cohen's Kappa Coefficient

- Cohen's kappa coefficient, also known as the Kappa statistic, is a better way of measuring agreement that takes into account the probability of agreeing by change. κ is defined as:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \tag{1}$$

- $\Pr(a)$ is the relative observed agreement among raters

- $\Pr(e)$ is the probability of chance agreement, calculated using the annotated data to estimate the probabilities of each observer randomly saying each category

- If the raters are in complete agreement then $\kappa = 1$

Text Encoding Initiative (TEI)

A project sponsored by the Association for Computational Linguistics, the Association for Literary and Linguistic Computing, and the Association for Computers in the Humanities encoding guidelines: link:http://www.tei-c.org

It defines how documents should be marked-up with the mark-up language SGML (or more recently XML)

XML

- XML: Extensible Markup Language similar to HTML has no fixed semantics: user defines what tags mean
- recognized as international ISO standard
- formally verifiable via document type definitions (DTD)
- tools available for editing, displaying,

XML – Example

```
<CATALOG>
<CD>
<TITLE>Empire Burlesque</TITLE>
<ARTIST>Bob Dylan</ARTIST>
<COUNTRY>USA</COUNTRY>
<COMPANY>Columbia</COMPANY>
<PRICE>10.90</PRICE>
<YEAR>1985</YEAR>
</CD>
<CD>
<TITLE>Greatest Hits</TITLE>
<ARTIST>Dolly Parton</ARTIST>
<COUNTRY>USA</COUNTRY>
<COMPANY>RCA</COMPANY>
<PRICE>9.90</PRICE>
<YEAR>1982</YEAR>
</CD></CATALOG>
```

TEI Guildelines

Each text that is conformant with the TEI guidelines consists of two parts

- Header
  - author
  - title
  - date the edition or publisher used in creating the machine-readable text
  - information about the encoding practices adopted

  ...
- Body
  - The actual annotated text

XML Annotated Text

```
<text>
<body>
<div type="BODY">
<div type="Q">
<head>Subject: The staffing in the Commission of the
  European Communities
</head>
<p>Can the Commission say:</p>
<p>1. how many temporary officials are working at the Commission?</p>
<p>2. who they are and what criteria were used in selecting them?</p>
</div>
<div type="R">
<head>Answer given by <name type="PERSON">
<abbr rend="TAIL-SUPER">Mr</ABBR>
Cardoso e Cunha</name> on behalf of the Commission
```

```
<date>(22 September 1992)</date></head>
<p>1 and 2. The Commission will send tables showing the number
 of temporary staff working for the Commission directly to the
  Honourable Member and to Parliament's Secretariat.</p>
</div></div></body></text>
```

Stand off Annotation

- separate the tags and text
- link back to character (or byte positions)
  - This is a pen
  - <pos='noun' cfrom='0' cto='4'>
  - <pos='verb' cfrom='5' cto='7'>

Lab 1

Non-exact matching

- Often you will want to match not just a word or sequence of words but some kind of pattern
- Various corpus interface tools make it easy to do this
- A standard way is to match **regular expressions**
- A good text editor should allow regular expression matching
  e.g., EMACS, notepad++

Regular Expressions

Regular Expressions

➢ Regular expressions: a formal language for matching things.

Symbol	Matches
.	any single character
[ ]	a single character that is contained within the brackets.
	[a-z] specifies a range which matches any letter from "a" to "z".
[^]	a single character not in the brackets.
^	the starting position within the string/line.
$	the ending position of the string/line.
*	the preceding element zero or more times.
?	the preceding element zero or one time.
+	the preceding element one or more times.
|	either the expression before or after the operator.
\	escapes the following character.

Regular Expression Examples

- `.at` matches any three-character string ending with "at", including "hat", "cat", and "bat".
- `[hc]at` matches "hat" and "cat".
- `[^b]at` matches all strings matched by `.at` except "bat".
- `^[hc]at` matches "hat" and "cat", but only at the beginning of the string or line.
- `[hc]at$` matches "hat" and "cat", but only at the end of the string or line.
- `\[.\]` matches any single character surrounded by "[" and "]" since the brackets are escaped, for example: "[a]" and "[b]".

Wild Cards

- a wildcard character substitutes for any other character or characters in a string.
  - Files and directories (Unix, CP/M, DOS, Windows)
    - `*` matches zero or more characters
    - **?** matches one character
    - **[ ]** matches a list or range of characters
    - E.g.: Match any file that ends with the string ".txt" or ".tex".

      `ls *.txt *.tex`
  - Structured Query Language (SQL)
    - **%** matches zero or more characters
    - _ matches a single character

english-corpora.org Interface Specialties

Pattern	Explanation	Example	Matches
word	One exact word	mysterious	mysterious
pos	Part of speech	[vvg]	going, using
pos*	Part of speech	[v*]	find, does, keeping, started
WORD	Lemma	SING	sing, singing, sang
=word	Synonyms	=strong	formidible, muscular, fervent
=WORD	Synonym+Lemma	=STRONG	formidible, muscular, fervent, strongest
word|wurd	Any of the words	stunning|gorgeous	stunning, gorgeous
x?xx*	wildcards	on*ly	only, ontologically, on-the-fly,
x?xx*	wildcards	s?ng	sing, sang, song
-word	negation	-[nn*]	the, in, is
word_pos	Word AND pos	can_v*	can, canning, canned (verbs)
		can_n*	can, cans (nouns)

Acknowledgments

- Thanks to Na-Rae Han for inspiration for some of the slides (from *LING 2050 Special Topics in Linguistics: Corpus linguistics*, U Penn).

- Thanks to Sandra Kübler for some of the slides from her *RoCoLi*[1] *Course: Computational Tools for Corpus Linguistics*

- Definitions from WordNet 3.0

[1] *Romania Computational Linguistics Summer School*