

Corpus Linguistics (L615)

Corpus Annotation Tools

Markus Dickinson

Department of Linguistics, Indiana University
Spring 2009

Big picture

I want you to see a range of possible tools for working with corpus data

- ▶ Some are XML-based; some are not

We won't walk through too many specifics

- ▶ on your next assignment, you will have to thoroughly use one of them (of your choice)
- ▶ See also <http://devoted.to/corpora> → Software, Tools, ...
 - ▶ Scroll to “Tools & Resources for Transcribing, Annotating or Analysing texts”
- ▶ Or check out the Linguistic Annotation Wiki:
http://annotation.exmaralda.org/index.php/Linguistic_Annotation

We'll also look at various annotation formats, to help us understand what it is exactly that a tool is trying to encode.

Annotating basic text files

Annotation formats

GATE

EXMARaLDA tools

MMA2

WordFreak

CLaRK

NITE

We can annotate basic text files by adding column data, and this can even include structural data

- ▶ We will first look at some examples of this
- ▶ Then we will look at various tools designed to help with corpus annotation

As you see the different formats, hopefully, you'll start to realize that knowing Perl can help you convert between them.

A way to simply encode POS tags:

%% Comments appear after beginning-of-line '%%'

%% s1 ...

The DT

man NN

ran VB

. .

%% s2 ...

CHAT (CHILDES) format

*MAR: I wanted a toy.

%mor: PRO|I&2S V|want-PAST DT|a&INDEF N|toy.

*MOT: well go get it!

%spa: \$IMP \$REF \$INS

%mor: ADV|well V|go&PRES V|get&PRES PRO|it!

A01:0010.03	-	YB	<minbrk>	-	[0h.0h]
A01:0010.06	-	AT	The	the	[0[S[Nns:S.
A01:0010.09	-	NP1s	Fulton	Fulton	[Nns.
A01:0010.12	-	NNL1cb	County	county	.Nns]
A01:0010.15	-	JJ	Grand	grand	.
A01:0010.18	-	NN1c	Jury	jury	.Nns:s]
A01:0010.21	-	VVDv	said	say	[Vd.Vd]
A01:0010.24	-	NPD1	Friday	Friday	[Nns:t.Nns:t]
...					

Fulton County is a phrase of category Nns.

1	Cathy	Cathy	N	N	eigen ev neut	2	subject	WordFreak	_	_
2	zag	zie	V	V	trans ovt 1of2of3 ev	0	ROOT	MARK	_	_
3	hen	hen	Pron	Pron	per 3 mv datofacc	2	object	TE	_	_
4	wild	wild	Adj	Adj	attr stell onverv	5	modifier		_	_
5	zwaaien	zwaai	N	N	soort mv neut	2	verb		_	_
6	.	.	Punc	Punc	punt	5	punct		_	_

Cathy, *hen*, and *zwaaien* are all dependents of *zag*.

```
#BOS 614 0 1091718495 1 %% @SB2AV@
Nach nach APPR -- AC 500
der der ART Dat.Sg.Fem NK 500
Wende Wende NN Dat.Sg.Fem NK 500
wollte wollen VMFIN 3.Sg.Past.Ind HD 505
Bonn Bonn NE Nom.Sg.Neut SB 505
...
#500 -- PP -- MO 505
...
#505 -- S -- -- 0
#EOS 614
```

Nach der Wende forms a PP, which is itself a daughter of S.

Specifics of annotation tools

The tools we're going to look at offer different kinds of features, and none is perhaps perfect for your needs

Important considerations:

- ▶ Does this tool allow me to consistently annotate data?
- ▶ Is it easy to plug external technology in to this tool?
- ▶ Is it easy to include previously-annotated layers of annotation?

“GATE [<http://gate.ac.uk/>] is an infrastructure for developing and deploying software components that process human language. GATE helps scientists and developers in three ways:

1. by specifying an architecture, or organisational structure, for language processing software;
2. by providing a framework, or class library, that implements the architecture and can be used to embed language processing capabilities in diverse applications;
3. by providing a development environment built on top of the framework made up of convenient graphical tools for developing components.”

From the user's guide, <http://www.gate.ac.uk/sale/tao/index.html>

Getting and using GATE

It's pretty easy to get GATE going ...

1. Download the software (probably the binaries, unless you have other preferences)
2. Run the application

Loading a(n unannotated) corpus

1. Right-click on 'Language Resources' and choose 'New', then 'GATE Document'.
2. In the dialog box choose the file you want to open in GATE or type a URL.
3. Change 'markupAware' to false, if you do not want GATE to analyse the document format.
4. Provide a document name or leave blank to use an automatically generated name.
5. Click OK.
6. The document will appear under the list of Language Resources loaded in the system.
7. To view its content, double click on its name.

<http://gate.ac.uk/demos/movies.html>

GATE is ideally designed for:

- ▶ Running a pipeline of NLP tools on a corpus
 - ▶ Load processing resources
 - ▶ Run a corpus pipeline over the document
 - ▶ This puts annotation into annotation sets
- ▶ Working with Java plug-ins

You can walk through a document by following the instructions at: <http://gate.ac.uk/demos/movies.html>

There are a variety of tools available for doing multi-layer annotation of corpora at:

http://www.exmaralda.org/en_downloads.html

- ▶ Probably the main one to look at is: Corpus manager (Coma)
 - ▶ nb: need Java 1.6 to run it
- ▶ The documentation is mostly in German, but if you download the example corpus, you can figure some things out:
 - ▶ http://www.exmaralda.org/corpora/en_demokorpus.html

MMAX2 is an XML-based tool that is particularly useful for anaphora annotation

- ▶ MMAX2 is fairly easy to obtain and install; simply download and unpack the appropriate files at: <http://mmax2.sourceforge.net/>
 - ▶ For documentation, see the doc/ folder, as well as the paper available at this site

Loading a text file

1. Run `./startmmax.sh` (unix) or `startmmax.bat` (windows)
2. Tools → Project Wizard
 - 2.1 Text Input File: Pick file and click on *Analyse File*
 - 2.2 Tokenization: select “one token per line” and click on *Tokenize*
 - 2.3 Markable level: Click on *Add level* for each level to be added
 - ▶ Make word level
 - ▶ Can make, e.g., POS level (or POS can be an attribute of the word level)
 - 2.4 .MMAX Project: Pick a project path; you’ll likely want `basedata`, `scheme`, etc. as daughter directories of this path.

See also p. 22 of the `mmax2quickstart.pdf` file, which walks you through using the wizard.

What is a markable?

- ▶ A markable is an item from the corpus which can be marked.
 - ▶ For POS annotation, this corresponds to words
 - ▶ For other annotations, this might be more than one word
- ▶ Annotation is either an *attribute* or a *relation* of the markable
 - ▶ An attribute is a property (e.g., POS tag) with a particular value for that markable.
 - ▶ A relation relates one markable to another
 - ▶ Can have MARKABLE_SETs (unordered relations) or MARKABLE_POINTERs (ordered relations)

So, when we create a word level of annotation, we have word markables that can be annotated

- ▶ Markable files look like the following:

```
<?xml version="1.0" encoding="US-ASCII"?>
<!DOCTYPE markables SYSTEM "markables.dtd">
<markables xmlns="www.eml.org/NameSpaces/word">
<markable mmax_level="word" id="markable_1" span="word_1"/>
<markable mmax_level="word" id="markable_2" span="word_2"/>
...
</markables>
```

- ▶ MMAX2 creates this automatically, but it really isn't that hard to convert data into this format

Adding annotation

To add annotation, you need to change the scheme files

- ▶ Here is what my POS_scheme.xml file now looks like:

```
<?xml version="1.0" encoding="UTF-8"?>
<annotationscheme>
<attribute id="tag_level" name="tag" type="freetext">
  <value name="tag"/>
</attribute>
</annotationscheme>
```

Note the use of `freetext` as the type: this allows me to create new POS tags on the fly (but could lead to more errors)

- ▶ Useful slides:
<http://homepages.inf.ed.ac.uk/olemon/mullerslides.pdf>

Changing displays

Corpus Linguistics

Corpus Annotation
Tools

Annotation formats

GATE

EXMARaLDA tools

MMAX2

WordFreak

CLaRK

NITE

When using annotation, it is often useful to change displays

- ▶ You can do this through style sheets and, for things like color, through the customization file.
- ▶ See the `mmax2stylesheets.pdf` documentation.

1. <http://wordfreak.sourceforge.net/>
2. `java -jar wordfreak-2.2.jar`
 - ▶ Look at the help contents for some help, especially the quick-start guide
 - ▶ Getting a file up and running

WordFreak is a bit more limited in its capacity (e.g., it's harder to change tagsets)

The CLaRK system is a fairly robust system for encoding syntactic annotation:

<http://www.bultreebank.org/clark/index.html>

After downloading it, you'll want to read the `readme.txt` file for installation instructions (in `ClarkSystem/`)

- ▶ Namely, you'll have to slightly tweak either the `ClarkSystem.uni` or `ClarkSystem.bat` file

Getting a corpus in the right format

Corpus Linguistics

Corpus Annotation
Tools

Annotation formats

GATE

EXMARaLDA tools

MMA2

WordFreak

CLaRK

NITE

Take a look at: `ClarkSystem/resources/Demo/fox/`

Here's the "corpus" file (`fox.xml`):

```
<GrammarExample>  
<S>the quick brown fox jumps over the lazy dog</S>  
<S>the man saw the boy with the telescope in the garden</S>  
</GrammarExample>
```

In other words: you'll need your corpus in the right format

And here's the grammar.dtd file:

```
<!DOCTYPE GrammarExample [  
<!ELEMENT GrammarExample (S)+ >  
<!ELEMENT S ANY >  
<!ELEMENT NP ANY >  
<!ELEMENT VP ANY >  
<!ELEMENT PP ANY >  
<!ELEMENT P #PCDATA >  
<!ELEMENT N #PCDATA >  
<!ELEMENT D #PCDATA >  
<!ELEMENT V #PCDATA >  
<!ELEMENT Aux #PCDATA >  
<!ELEMENT Pron #PCDATA >  
<!ELEMENT Adj #PCDATA >  
>]
```

Opening the demo

1. First, compile the DTD: DTD → Compile DTD
2. Then, load the file: File → Import XML

You can also import pure text with the “Import text” function

- ▶ You'll still need an appropriate DTD if you're going to include particular features.
- ▶ Of course, you can always change the DTD later, as you edit the file.

The NITE XML toolkit is particularly useful for multi-modal data: <http://www.ltg.ed.ac.uk/NITE/>

- ▶ Read the documentation for more information

A related set of tools, LT TTT2, is available here:
<http://www.ltg.ed.ac.uk/software>

- ▶ Regardless of what annotation software you want to use, these can tokenize & tag English data in a variety of ways