Corpus Linguistics (L615)
Corpus Annotation Tools

Markus Dickinson
Department of Linguistics, Indiana University
Spring 2013

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

- See also http://tiny.cc/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/LinguisticAnnotation

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

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

Knowing a scripting language like Perl can help you convert between these formats.

TnT format

A way to simply encode POS tags:

```
%% Comments appear after beginning-of-line '%%'
%% s1 ... The DT man NN ran VB ... 
%% s2 ... Fulton County
```

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!
```

SUSANNE format

```
A01:0010.03 - YB <minbrk> - [Oh.Oh]
A01:0010.06 - AT The the [05[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.*
Cathy, hen, and zwaaien are all dependents of zag.

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

Some considerations:

- Is the interface intuitive?
- How does meta-data fit in?
- Does this tool allow me to consistently annotate data?
- Are multiple layers of different types easily allowed?
- Is it easy to plug external technology in to this tool?
- Is the documentation thorough and will there be support two years from now?
UAM Corpus Tool

Main Steps

1. Start a new project
2. Add (an) annotation layer(s)
   - You can use some pre-built annotation schemes or design your own
3. Add files
   - Incorporated files are ones you have already started annotating
4. Annotate

UAM Corpus Tool

Defining the Annotation Scheme

If you are creating a new scheme, you are given the opportunity to specify this when creating an annotation layer
- system = name of the type of annotation (cf. attribute)
- feature = alternatives for each system (cf. value)

Or you can re-use a scheme

UAM Corpus Tool

Partitur

Splitting cells

Splitting cells: XML

UAM Corpus Tool

UAM Corpus Tool (http://www.wagsoft.com/CorpusTool/) is designed to be usable by non-computational linguists
- See write-up: http://www.aclweb.org/anthology-new/P/P08/P08-4004.pdf

The manual that comes with the download is very thorough & easy to use
- We will walk through using some of it
UAM Corpus Tool
Annotating the Corpus

- Annotate document
- Annotate segments (e.g., words)
  - Segment definitions are flexible
  - Note: I had freezing issues when I tried to have automatic segmentation for the annotation layer
- Automatic analysis with Stanford parser (English only)
  - There are various options to automatically annotate the data, which we won’t go over

POS Annotation

- Basic searching
- Searching across layers
- Fancier string searching (for English)
  -Wildcard token (*): ca* matches cat, caffeine, etc.
  - Classes: ca*noun matches nouns starting with ca
    - see the appendix for features which can be searched
    - note: this can take some time
  - % matches inflections: e.g., be% matches be, is, ...
Corpus Linguistics
Corpus Annotation
Tools
Annotation formats
EXMARaLDA tools
UAM Corpus Tool
brat
Other tools
MMAX2
GATE
WordFreak
CLaRK
NITE

brat rapid annotation tool is a web-based interface to allow for collaborative annotation

▶ Encoded in a standoff format
▶ Works well for structured annotation
  ▶ Defined in terms of text spans
  ▶ Allows for relations between entities
▶ Has been used for:
  ▶ entity mention detection, event extraction, coreference resolution, normalization, chunking, dependency syntax, meta-knowledge, ...

See: http://brat.nlplab.org

brat screenshot

brat standoff format

Other tools

We probably won’t have time to go through these other tools, which I covered the last time I taught L615

▶ I include notes here for completeness
▶ Note that some information may be out-of-date

MMAX2

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.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.

Markables

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)

Adding annotation

To add annotation, you need to change the scheme files
- Here is what my POS_scheme.xml file now looks like:

  ```xml
  <?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:

Changing displays

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.

GATE

“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:
- by specifying an architecture, or organisational structure, for language processing software;
- by providing a framework, or class library, that implements the architecture and can be used to embed language processing capabilities in diverse applications;
- 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](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

Adding annotation

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

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

WordFreak

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)

CLaRK

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 have to slightly tweak either the ClarkSystem.uni or ClarkSystem.bat file

Getting a corpus in the right format

Take a look at: ClarkSystem/resources/Demo/fox/
Here’s the “corpus” file (fox.xml):

```xml
<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
DTD

And here's the grammar.dtd file:

```xml
<!DOCTYPE GrammarExample [
  <!ELEMENT GrammarExample (S)+ >
  <!ELEMENT S ANY >
  <!ELEMENT NP ANY >
  <!ELEMENT VP ANY >
  <!ELEMENT PP ANY >
  <!ELEMENT P #PCDATA >
  <!ELEMENT N #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.

NITE XML Toolkit

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