Corpus Linguistics (L615)
Corpus Annotation Tools

Markus Dickinson
Department of Linguistics, Indiana University
Spring 2013
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

- 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/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

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 ...

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 [O[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.
Cathy, *hen*, and *zwaaien* are all dependents of *zag*. 
Negra format

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?
EXMARaLDA tools

Good for multi-layer annotation of corpora
(http://www.exmaralda.org/en_downloads.html)

Three main components:

1. Partitur-Editor: for annotating data
2. Corpus manager (Coma): for combining annotated files into a corpus & adding meta-data
3. EXAKT: for searching through the data

All files use an XML format

- The documentation is mostly in German, but the example corpus helps figure some things out:
Partitur

EXMARaLDA Partitur-Editor 1.5.1

<table>
<thead>
<tr>
<th>v</th>
<th>The</th>
<th>quick</th>
<th>fox</th>
<th>jumped</th>
<th>over</th>
<th>the</th>
<th>lazy</th>
<th>dog</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>DT</td>
<td>JJ</td>
<td>NN</td>
<td>VBD</td>
<td>IN</td>
<td>DT</td>
<td>JJ</td>
<td>NN</td>
</tr>
</tbody>
</table>

Done.

Segmentation: GENERIC
Player: ELAN-Quicktime-Player
**.exb format (1)**

```xml
<basic-body>
<common-timeline>
<tlid="T0"/>
<tlid="T1"/>
<tlid="T2"/>
<tlid="T3"/>
<tlid="T4"/>
<tlid="T5"/>
<tlid="T6"/>
<tlid="T7"/>
<tlid="T8"/>
<tlid="T9"/>
</common-timeline>
<tierid="TIE0" category="v" type="t" display-name="[v]">
<event start="T0" end="T1">The</event>
<event start="T1" end="T2">quick</event>
<event start="T2" end="T3">fox</event>
</tierid="TIE0" category="v" type="t" display-name="[v]">
```

- **Corpus Linguistics**
- **Corpus Annotation Tools**
- **Annotation formats**
- **EXMARaLDA tools**
- **UAM Corpus Tool**
- **brat**
- **Other tools**
  - MMAX2
  - GATE
  - WordFreak
  - CLaRK
  - NITE
.exb format (2)

```xml
<event start="T3" end="T4">jumped</event>
<event start="T4" end="T5">over</event>
<event start="T5" end="T6">the</event>
<event start="T6" end="T7">lazy</event>
<event start="T7" end="T8">dog</event>
<event start="T8" end="T9">.</event>
</tier>
<tier id="TIE1" category="POS" type="a"

display-name="[POS]">
<event start="T0" end="T1">DT</event>
<event start="T1" end="T2">JJ</event>
<event start="T2" end="T3">NN</event>
<event start="T3" end="T4">VBD</event>
<event start="T4" end="T5">IN</event>
<event start="T5" end="T6">DT</event>
<event start="T6" end="T7">JJ</event>
<event start="T7" end="T8">NN</event>
<event start="T8" end="T9">.</event>
</tier></basic-body>
```
Partitur

Splitting cells
Splitting cells: XML

```xml
<common-timeline>
  <tli id="T0"/>
  <tli id="T1"/>
  <tli id="T2"/>
  <tli id="T3"/>
  <tli id="T10"/>
  <tli id="T4"/>
  <tli id="T5"/>
  <tli id="T6"/>
  <tli id="T7"/>
  <tli id="T8"/>
  <tli id="T9"/>
</common-timeline>
...

<tier id="TIE2" category="morph" type="a" display-name="[morph]">"PST</event>
</tier>
```
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

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**
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
PTB Scheme (extract 1)

```xml
<SYSTEM>
  <NAME>NOUN-TYPE</NAME>
  <EC>noun</EC>
  <FEATURES>
    <FEATURE>
      <NAME>common</NAME>
      <STATE>active</STATE>
    </FEATURE>
    <FEATURE>
      <NAME>proper</NAME>
      <STATE>active</STATE>
    </FEATURE>
  </FEATURES>
</SYSTEM>
```
PTB Scheme (extract 2)

```xml
<SYSTEM>
  <NAME>COMMON-TYPE</NAME>
  <EC>common</EC>
  <FEATURES>
    <FEATURE>
      <NAME>nn</NAME>
      <STATE>active</STATE>
      <GLOSS>0/7/03</GLOSS>
    </FEATURE>
    <FEATURE>
      <NAME>nns</NAME>
      <STATE>active</STATE>
    </FEATURE>
  </FEATURES>
  ...
</SYSTEM>
```
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

Produced by G. Fuhrman

LEAVES OF GRASS
By Walt Whitman

Come, said my soul,
Such verses for my Body let us write, (for we are one,)
That should I after return,
Or, long, long hence, in other spheres,
There to some group of mates the chants resuming,
(Tallying Earth's soil, trees, winds, tumultuous waves,) Ever with pleas'd smile I may keep on,
Ever and erewit the verses o'ning, as first I here and now:

[POS tagging interface shown with words like 'temp', 'token', 'word', 'noun', 'common', and 'nn', 'nnp', 'nns', 'nnps' in the COMMON-TYPE column]
UAM Corpus Tool
Searching the Corpus

- Basic searching
- Searching across layers
- Fancier string searching (for English)
  - Wilcard 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, ...*
Searching

Enter Search Query Below:

noun +

Show

Only Partially Coded * Segment with Comment

Leaves of Grass/leaves-of-gra:
soul
Corpus Linguistics
Corpus Annotation Tools
Annotation formats
EXMARaLDA tools
UAM Corpus Tool

UAM Corpus Tool
Corpus statistics

<table>
<thead>
<tr>
<th>Type of Study:</th>
<th>Describe a dataset</th>
<th>Aspect of Interest:</th>
<th>General Text Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Units:</td>
<td>register</td>
<td>Show</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Number of segments:</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Words in segments:</td>
<td>126665</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text Complexity:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Av. Word Length (chars):</td>
<td>4.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Av. Word Length (syllables):</td>
<td>1.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Av. Segment Length (tokens):</td>
<td>126665.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Min. Segment Length (tokens):</td>
<td>99999</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Max. Segment Length (tokens):</td>
<td>150180</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexical Density:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Lexemes per segment:</td>
<td>63467.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Lexemes % of text:</td>
<td>50.11%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subjectivity:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Subjective Positivity:</td>
<td>0.120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Subjective Strength:</td>
<td>0.658</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference Density:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- 1p Reference:</td>
<td>4.333%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- 2p Reference:</td>
<td>1.551%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- 3p Reference:</td>
<td>2.721%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| 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
Corpus Linguistics
Corpus Annotation Tools
Annotation formats
EXMARaLDA tools
UAM Corpus Tool

Other tools
MMAX2
GATE
WordFreak
CLaRK
NITE

brat screenshot
brat standoff format

T241    VROOT 625 629    ROOT
T243    NN2 630 637    Details
T245    II 638 640    of
T247    DD1 641 645    each
T249    NN1 646 650    plan
T251    VBR 651 654    are
T253    II 655 657    as
T255    VVZ 658 665    follows
T257    Y 666 667    .
R119    SUBJ Arg1:T251 Arg2:T243
R120    NJCT Arg1:T243 Arg2:T245
R121    QUANT Arg1:T249 Arg2:T247
R122    POBJ Arg1:T245 Arg2:T249
R123    ROOT Arg1:T241 Arg2:T251
R124    CPZR Arg1:T255 Arg2:T253
R125    CPRED Arg1:T251 Arg2:T255
R126    PUNCT Arg1:T251 Arg2:T257
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 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)
Preprocessing

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

- Markable files look like the following:

```xml
<?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
<?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 [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
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
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)
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

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

Here's the “corpus” file (fox.xml):

```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:

```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 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