Basic Text Analysis Tools

Markus Dickinson

Dept. of Linguistics, Indiana University
Catapult Workshop Series; February 1, 2013
Basic text analysis

Before any sophisticated analysis, we want ways to get a “sense” of text data

▷ online searching
▷ concordancing
▷ word frequency counting
▷ annotating

I recommend working with a scripting language like Perl or Python to get even more capabilities for these things

A good resource to find various tools: http://tiny.cc/corpora (see “Software, Tools, Freq Lists, etc.”)
Corpus linguistics

The material I’ll be presenting is adapted from my course, L615 (Corpus Linguistics)

- Corpus linguistics is the study of how to gather & analyze linguistic data
- Feel free to grab whatever materials you find useful at: http://cl.indiana.edu/~md7/13/615/
Online search tools

You’ll often find web interfaces to corpora, which allow for searching

1. Pros: nice interfaces, useful for generating basic statistics, etc. access to corpora you might not otherwise have access to

2. Cons: You’re limited by what the designer wants you to search for; you can’t run a tagger or parser on it; etc.
Some search tools

- Mark Davies’ corpus search interface: http://corpus.byu.edu/
- SketchEngine: http://www.sketchengine.co.uk
- Intellitext: http://corpus.leeds.ac.uk
Concordancing

AntConc

Concordancing is simply viewing words by their contexts, a.k.a. Keyword in Context (KWIC)

We’ll look specifically at AntConc, by Laurence Anthony (http://www.antlab.sci.waseda.ac.jp/software.html)

- Installation is relatively straightforward
- The README files provide a lot of information

For this example, we’ll use *Les Misérables* by Victor Hugo

- Available from Project Gutenberg: http://www.gutenberg.org/ebooks/135
AntConc

File loading & concordancing

Load in a file (File → Open File(s))

- Make sure this is *plain text*, i.e., not a Word document or anything with extra mark-up

1. Enter search term
2. Determine window size
3. Start the search
4. Then, select sorting options & sort

Clicking on a word allows you to see the original context

- *Advanced* options: can use word fragments, multiple search terms, regular expressions, context words

Note that you can export your search results to text files
Rivoli serves as pattern even in the sewer. However, if the geometrical line is in place anywhere, it is certainly in the drainage trench of a great city. There, everything should be subordinated to the shortest road. The sewer has, nowadays, assumed a certain official aspect. The very police reports, of which it sometimes forms the subject, no longer want in respect towards it. The words which characterize it in administrative language are sonorous and dignified. What used to be called a gut is now called a gallery; what used to be called a hole is now called a surveying orifice. Villon would no longer meet with his ancient temporary provisional lodging. This net-work of cellars has its immemorial population of prowlers, rodents, swarming in greater numbers than ever; from time to time, an aged and veteran rat risks his head at the window of the sewer and surveys the Parisians; but even these vermin grow tame, so satisfied are they with their subterranean palace. The cesspool no longer retains anything of its primitive ferocity. The rain which in former days soiled the sewer, now washes it. Nevertheless, do not trust yourself too much to it. Miasmas still inhabit it. It is more hypocritical than irreproachable. The prefecture of police and the commission of health have done their best. But, in spite of all this, processes of disinfection, it exhales, a vague, suspicious odor like Tartuffe after confession.

Let us confess, that, taking it all in all, this sweeping is a homage which the sewer pays to civilization, and as, from this point of view, Tartuffe's conscience is a progress over the Augustan stables, it is certain that the sewers of Paris have been improved.
AntConc

Concordance plots

Concordance plots allow one to see distribution at a glance
AntConc

Clusters & N-grams

You can search for:

- Clusters involving a particular word
- All “n-grams” of a particular size
- Collocations involving a particular word
  - Note: there are a lot of methods & tools for finding collocations (e.g., http://www.collocations.de, http://www.d.umn.edu/~tpederse/nsp.html)
Basic Text Analysis Tools

Online Search
Concordancing
AntConc
Word Frequency
Annotating
UAM Corpus Tool
Automatically annotating
### Basic Text Analysis Tools

- **Online Search**
- **Concordancing**
- **AntConc**
- **Word Frequency**
- **Annotating**

#### Screenshot

![Screenshot of AntConc 3.2.4m](image)

**Total No. of N-Grams Types:** 213301  **Total No. of N-Grams Tokens:** 523771

<table>
<thead>
<tr>
<th>Rank</th>
<th>Freq</th>
<th>N-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5176</td>
<td>of the</td>
</tr>
<tr>
<td>2</td>
<td>3221</td>
<td>in the</td>
</tr>
<tr>
<td>3</td>
<td>1937</td>
<td>on the</td>
</tr>
<tr>
<td>4</td>
<td>1728</td>
<td>to the</td>
</tr>
<tr>
<td>5</td>
<td>1140</td>
<td>of a</td>
</tr>
<tr>
<td>6</td>
<td>1069</td>
<td>at the</td>
</tr>
<tr>
<td>7</td>
<td>1075</td>
<td>he had</td>
</tr>
<tr>
<td>8</td>
<td>1057</td>
<td>and the</td>
</tr>
<tr>
<td>9</td>
<td>1022</td>
<td>Jean Valjean</td>
</tr>
<tr>
<td>10</td>
<td>802</td>
<td>in a</td>
</tr>
<tr>
<td>11</td>
<td>775</td>
<td>from the</td>
</tr>
<tr>
<td>12</td>
<td>756</td>
<td>to be</td>
</tr>
<tr>
<td>13</td>
<td>746</td>
<td>with the</td>
</tr>
<tr>
<td>14</td>
<td>732</td>
<td>he was</td>
</tr>
<tr>
<td>15</td>
<td>701</td>
<td>had been</td>
</tr>
<tr>
<td>16</td>
<td>674</td>
<td>it is</td>
</tr>
<tr>
<td>17</td>
<td>658</td>
<td>with a</td>
</tr>
<tr>
<td>18</td>
<td>654</td>
<td>that he</td>
</tr>
<tr>
<td>19</td>
<td>640</td>
<td>was a</td>
</tr>
<tr>
<td>20</td>
<td>633</td>
<td>did not</td>
</tr>
<tr>
<td>21</td>
<td>627</td>
<td>in his</td>
</tr>
<tr>
<td>22</td>
<td>622</td>
<td>it was</td>
</tr>
<tr>
<td>23</td>
<td>604</td>
<td>It was</td>
</tr>
</tbody>
</table>

**Search Term:** sewer

**N-Gram Size**

- **Min. Size:** 2
- **Max. Size:** 2

**Min. N-Gram Frequency:** 1
<table>
<thead>
<tr>
<th>Rank</th>
<th>Freq</th>
<th>Freq(L)</th>
<th>Freq(R)</th>
<th>Stat</th>
<th>Collocate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>148</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>sewer</td>
</tr>
<tr>
<td>2</td>
<td>75</td>
<td>72</td>
<td>3</td>
<td>2.99067</td>
<td>the</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>16</td>
<td>6</td>
<td>4.41132</td>
<td>The</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>2</td>
<td>20</td>
<td>2.11920</td>
<td>of</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>sewer</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>0</td>
<td>13</td>
<td>3.06929</td>
<td>is</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
<td>13</td>
<td>0</td>
<td>10.87364</td>
<td>Grand</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>11.54858</td>
<td>belt</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>0</td>
<td>10</td>
<td>1.46549</td>
<td>and</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>2.61492</td>
<td>which</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1.59661</td>
<td>in</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>7</td>
<td>1</td>
<td>1.19836</td>
<td>a</td>
</tr>
<tr>
<td>13</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>1.18882</td>
<td>was</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>1.41042</td>
<td>that</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>11.25002</td>
<td>Montmartre</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>1.39703</td>
<td>with</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>3.75650</td>
<td>its</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>3.45371</td>
<td>has</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>9.32809</td>
<td>collecting</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>3.52863</td>
<td>A</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>-0.80089</td>
<td>to</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2.94938</td>
<td>This</td>
</tr>
<tr>
<td>23</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4.72842</td>
<td>THE</td>
</tr>
</tbody>
</table>

Search Term: sewer
Window Span: 1L 1R
Min. Collocate Frequency: 1
Saved Window
Exit
AntConc

Word lists

Word lists are easy to generate

- This is an easy way to get word frequency counts

Note: you can set the preferences so as to use lemmas, if you have a file listing lemmas
Basic Text Analysis Tools

Online Search
Concordancing
AntConc
Word Frequency
Annotating
UAM Corpus Tool
Automatically annotating
You can also generate Keyword Lists, which show unusually (in)frequent words as compared to some reference corpus.

- To do this, you have to load your reference corpus (or word list) under (Tool) Preferences → Keyword (List)

The next two slides show:

1. The keywords of *Leaves of Grass*, as compared to a book called *Camping*
   - unusually frequent, given the reference corpus

2. The negative keywords of *Leaves of Grass*, as compared to a book called *Camping*
   - unusually infrequent, given the reference corpus
Basic Text Analysis

Tools

Online Search
Concordancing
AntConc
Word Frequency
Annotating
UAM Corpus Tool
Automatically annotating

Screenshot
Word Lists & Frequency

There are other tools on the David Lee site to calculate word frequency

- There are also word lists separate from corpus data, should you need those
- e.g., most frequent academic words

Calculating word frequencies on your own is easy with something like Perl
Annotating data

Corpus **annotation** is information about the corpus included in the corpus itself

- Meta-data helps recover the original context & allows for a wider range of questions to be explored
- Linguistic annotation:
  - helps train NLP tools
  - helps find & categorize examples
    - what is the plural form of fish?
    - are there subjectless sentences in German? (Yes, e.g., *Mir ist kalt.* (‘To me is cold.’))

Annotation makes it possible to find phenomena that would otherwise disappear in masses of data
Benefits of annotation

Corpus annotation adds much value to a corpus

- Extracting information is easier:
  - Can easily isolate *left* in its adjective uses
  - Can find information on a language you don’t speak
- Corpus is more reusable
  - Insights are more accessible to others
- Corpus is more multifunctional
- Annotation provides a clear record of analysis, open to future scrutiny
  - Decisions are more reproducible
Annotation tools

A lot of annotation tools are available, though many are geared towards people with computational skill

▶ See http://tiny.cc/corpora → Software, Tools, …
  ▶ Scroll to “Tools & Resources for Transcribing, Annotating or Analysing texts”
  ▶ Many tools for working with corpus are XML-based; some are not

▶ Or check out the Linguistic Annotation Wiki:
  http://annotation.exmaralda.org/index.php/Linguistic_Annotation

We’ll focus on the UAM Corpus Tool, which is a little more user-friendly than many
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

Note on annotation tools:

▶ They tend to work best by splitting your corpus into many smaller sub-corpora/small files

▶ Especially if the tool also automatically processes part of the corpus
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
Penn Treebank (PTB) Part-of-Speech (POS) 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 C. 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 ever yet the verses crying, oo, first, I hear and now.

Assigned

<table>
<thead>
<tr>
<th>temp</th>
<th>token</th>
<th>word</th>
<th>noun</th>
<th>common</th>
</tr>
</thead>
</table>

COMMON-TYPE

<table>
<thead>
<tr>
<th>nn</th>
<th>nnp</th>
<th>nns</th>
<th>nnps</th>
</tr>
</thead>
</table>

Gloss

Comment:
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*, ...

Note: to search for classes, do the search on a document-level annotation layer

- Segment-level annotation only searches across segments you have defined

Some of the work you can do in AntConc you can do here
Searching

Enter Search Query Below:

noun + Show

Only Partially Coded * Segment with Comment

Leaves of Grass/leaves-of-gra:

soul
### UAM Corpus Tool

#### Corpus statistics

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

You might want to explore using some automatic tools to be able to refer to linguistic classes (e.g., nouns)

- A good collection of NLP tools are the Stanford ones: http://nlp.stanford.edu/software/index.shtml

- There will be a tutorial on the Natural Language Toolkit (NLTK) on March 8: http://www.nltk.org