1 Introduction and Acknowledgment

This handout accompanies the presentation in DIGIlab, 4pm January 18th 2019.

It exemplifies some of the capabilities of the new corpus server, with special emphasis on applications in Linguistics and the Digital Humanities.

For an account on the server, please contact John Hale via <linglab@uga.edu>

Table 1: Thank you, sponsors:

<table>
<thead>
<tr>
<th>Center for Teaching and Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Department of Linguistics</td>
</tr>
<tr>
<td>DIGIlab</td>
</tr>
</tbody>
</table>

Special thanks also to Dr. Margaret Renwick, Dr. Chad Howe, Kyle Vanderniet and Donald Dunagan for their help realizing this project.

2 the New UGA corpus server

There are many corpora installed on the server. A subset of them can be accessed using the IMS Open Corpus Workbench. The workbench provides an interactive, command line search tool called CQP or "corpus query processor." This handout is not itself a CQP tutorial, however a tutorial is available, referred to below simply as T. T is essential reading for those seeking to use the system for real research.

First, connect to the corpus server following the instructions in Utilizing the Kucera Server by Donald Dunagan. Then, at the linux prompt, start CQP like this:

kucera $ cqp -eC

The -e flag gives you some limited line editing capabilities, such as accessing the previous command with the up-arrow key. -C makes CQP more colorful. All CQP commands should be ended with a semicolon. Here’s how to ask it what corpora are available:

[no corpus]> show corpora ;
[no corpus]> info WACKYPEDIA
988 million words from Wikipedia in 2009.

You can leave the CQP program with the exit command. Use info to learn about a particular corpus. One of most famous is the British National Corpus or BNC.

```plaintext
info BNC;
```

Use the spacebar to page through long output page by page, the return key to go line-by-line, and "q" to get out — especially at the end when it says END! You are really using the unix pager program "less." (Exit cqp and type man less at the shell prompt to get more information about navigation shortcuts in less)

To work with a particular corpus, type its name. For instance:

```plaintext
[no corpus]> BNC;
```

This selects the BNC as the corpus against which queries will be evaluated.

### 3 a lovely way to start off

The simplest queries are about individual words. Entering a word in quotes searches for that word in the currently-selected corpus, printing a concordance as output.

```plaintext
BNC> "lovely";
```

You can export results for inclusion e.g. in an academic paper. Redirect the keyword-in-context display for the Last query to a file using the greater-than sign.

```plaintext
BNC> set PrintMode ascii;
BNC> cat Last > "lovely.txt";
```

The BNC includes metadata, for instance about the demographics of speaker or writer. text_author_sex_ is an annotation that indicates the gender of the writer. Use the group command to get a frequency distribution over values of this annotation.
Another piece of metadata has to do with the modality of the language included in the BNC, i.e. written or spoken.

Now, considering just the spoken portion do women really say "lovely" more?

Women in the BNC say "lovely" at a rate of \( \frac{(1265 \times 1M)}{11983120} = 105.5 \) per million whereas men say it \( \frac{(552 \times 1M)}{11983120} = 46 \) per million.

Table 2: contingency table for "lovely" as uttered by men and women in the British National Corpus

<table>
<thead>
<tr>
<th></th>
<th>female</th>
<th>male</th>
</tr>
</thead>
<tbody>
<tr>
<td>attestations of &quot;lovely&quot;</td>
<td>1265</td>
<td>751</td>
</tr>
<tr>
<td>corpus size</td>
<td>11983120</td>
<td>11983120</td>
</tr>
</tbody>
</table>

We can verify that this difference is statistically significant by computing a Log-Likelihood ratio (G2). Andrew Hardie provides a web calculator with this functionality, and so do Stefan Evert and Paul Rayson. Some basic documentation on how to use Rayson’s calculator can be found under "Doing a significance test" in *Corpus Linguistics: method, theory and practice.*

Is it the young or the old women who are saying "lovely" so much?
BNC> WOMEN = "lovely" :: match.u_sex="female";
BNC> group WOMEN match u_age_group ;
#---------------------------------------------------------------
(none) 60+ 375
 45-59 256
 25-34 228
 35-44 178
 15-24 98
 unknown 95
 0-14 35

4 linguistic annotations: part of speech tags

A typical kind of annotation in corpus linguistics is the part of speech tag. This just means that each word is marked with its grammatical category. We don’t need to approximate past tense verbs as words whose spelling ends in -ed

BNC> ED = [ word = ".+ed" ] ;
BNC> reduce ED to 500 ;
BNC> group ED match word ;

We can instead search directly for words tagged as verbs. The CLAWS C5 tagset specifies that VVD is the tag for past tense lexical verbs.

BNC> [ pos = "VVD" ] ;

Notice that the lefthand side of the constraint in the query mentions pos rather than word. The latter is what you get when you simply enter a word in quotes as the query.

One can actually get quite a lot of (morpho-)syntax done just using part of speech tags! [Biber (1988)] defines 67 different linguistic features in terms of POS patterns, let’s examine just one – his rule 18 for agentless passives.

The basic idea is that there has to be some form of "be" (POS tags VBB, VBD, VBG . . . ) followed by a past participle (VVN).
But actually Biber specifies that there can be up to two adverbials in between the BE and the past participle. This catches examples like "very well educated" and "more urgently needed." The Up To Two constraint is specified by the postfix expression \{0,2\}.

We can insist that the passive be truly agentless by disallowing by-phrases after the past participle. This negative specification is done using the nonmatch operator !=.

And actually, we don’t mind if there are some prepositions or adverbs after the participle, as long as they don’t involve by. This allows us to capture phrasal verbs.

Martin Weisser notices that the agentless passive construction comes in a variant where phrases like "in turn" or "to some extent" are inserted between the helping verb and the past participle. We can find those inside a window of up to 4 nonverbs \{0,4\} with a query like this:

Adding the qualifier within s prevents matches across sentence boundaries.

We could refine this forever, but for now let’s just confirm with Biber that agentless passives are way more frequent in writing compared to spoken language.
Moving on, what are the top ten instances of single-word coordination? Here is Weisser’s query, slightly adapted. It shows off CQP’s ability to enforce equality constraints, i.e. that the part of speech of the two conjuncts must be the same.

\[
\text{BNC}\left\{ \text{a: [ pos = "N.*|J.*|V.*|AV.*" & pos != "NP0" ] [ word="and|an"%c & pos="CJC"] [pos=a.pos] ; \right\}
\]

\[
\text{BNC} \text{ count by word cut 10}
\]

| 1873 | men and women [#202946-#204818] |
| 1151 | up and down [#333492-#334642] |
| 1066 | more and more [#209355-#210420] |

5 "strong" and "powerful" in newspaper text

At the dawn of statistical NLP, Church and Mercer 1993 made an observation about the word "strong" and "powerful" which, on the face it, seem to have pretty much the same meaning. Both are adjectives. But consider the nouns that they modify in The New York Times. These queries take a bit longer because the corpus comprises 1.3 billion tokens.

\[
\text{NYT}\left\{ \text{info NYT}
\right\}
\]\n
\[
\text{NYT}\left\{ \text{define $punc = ". . . , ? ’ ’ “ ; - -- --- !";}\right\}
\]\n
\[
\text{NYT}\left\{ \text{S = "strong" @[ word != $punc & pos = "NN.*" ]}
\right\}
\]\n
\[
\text{NYT}\left\{ \text{group S target word cut 50;}
\right\}
\]\n
\[
\text{NYT}\left\{ \text{P = "powerful" @[ word != $punc & pos = "NN.*" ]}
\right\}
\]\n
\[
\text{NYT}\left\{ \text{group P target lemma cut 50;}
\right\}
\]

Hanks, a lexicographer ... hypothesized that strong is an intrinsic quality whereas powerful is an extrinsic one. Thus, for example, any worthwhile politician or cause can expect strong supporters, who are enthusiastic, convinced, vociferous, etc., but far more valuable are powerful supporters, who will bring others with them. They are also, according to the AP news, much rarer—or at any rate, much less often mentioned. This is a fascinating hypothesis that deserves further investigation.

We have 265 strong supporters but only 46 powerful supporters.
6 body language as a contributor to characterization

Let's turn now to a more literary example. Michaela Mahlberg (2013) studies a kind of coordination used in the works of Charles Dickens, the as if comparison.

DICKENS> ASIF = "as" "if" ;
DICKENS> group ASIF match novel_title ;

Mahlberg explains that as if typically introduces a comment by the narrator (rather than by a character). She identifies several meaning groups

<table>
<thead>
<tr>
<th>Meaning group</th>
<th>Collocates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action verbs</td>
<td>made, turned, went, stood, shook, stopped, came</td>
</tr>
<tr>
<td>Body part nouns</td>
<td>head, hand, hands, eyes, face, back, mouth, arms, arm, lips</td>
</tr>
<tr>
<td>Settings</td>
<td>door, fire, room, side, chair, wall</td>
</tr>
<tr>
<td>Manner</td>
<td>manner, way, air, seemed</td>
</tr>
<tr>
<td>LOOK</td>
<td>looked, looking, look, looks</td>
</tr>
<tr>
<td>SPEAK</td>
<td>said, speaking, spoke, voice</td>
</tr>
<tr>
<td>FEEL</td>
<td>felt, feel</td>
</tr>
<tr>
<td>Other</td>
<td>time, moment, man, round, down, again, little, almost, very, great, now, well, quite, still</td>
</tr>
</tbody>
</table>

Figure 1: meaning groups identified by Mahlberg (2013)

which we can verify. Let's formulate a set of "content words" CW that are either lexical verbs or common nouns.

DICKENS> LexicalVerbsBeforeAsIf = @[ pos = "VB.*" & lemma != "be|do|have"] []{0,5} "as" "if" ;
DICKENS> CommonNBeforeAsIf = @[ pos = "NN(S)?" ] []{0,5} "as" "if" ;
DICKENS> CW = union CommonNBeforeAsIf LexicalVerbsBeforeAsIf;
DICKENS> group CW target lemma cut 10 ;
This immediately pops out Mahlberg’s LOOK, FEEL and MANNER meaning groups. We can drill down on the body part meaning group with the help of a word list, $body$.

\[
\text{DICKENS}> \text{define }\ $body = "\text{head hand eye face back mouth arm lip}"; \\
\text{DICKENS}> \text{BodyBeforeAsIf} = @( \text{lemma} = $body & \text{pos} = "\text{NN(S)}?" ) [0,5] \"as\ " \"if\" \text{within s}; \\
\text{DICKENS}> \text{size BodyBeforeAsIf}; \\
335
\]

And in a similar way count up each meaning group.

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>all lexical verbs and common nouns</td>
<td>3580</td>
</tr>
<tr>
<td>action verbs</td>
<td>205</td>
</tr>
<tr>
<td>body part nouns</td>
<td>335</td>
</tr>
<tr>
<td>Settings</td>
<td>107</td>
</tr>
<tr>
<td>Manner</td>
<td>125</td>
</tr>
<tr>
<td>LOOK</td>
<td>292</td>
</tr>
<tr>
<td>SPEAK</td>
<td>120</td>
</tr>
<tr>
<td>FEEL</td>
<td>102</td>
</tr>
<tr>
<td>other</td>
<td>298</td>
</tr>
</tbody>
</table>

Whereas there are 335 attestations of body part nouns, if we choose eight random words from CW, we should expect a number of attestations equal to eight times the average attestation rate over all words. That’s about 18.

The log-likelihood ratio between the 335 actual attestations of body part nouns and the 18 attestations that we would expect from eight randomly-selected content words is 361.92. This is statistically significant. It shows that body language is indeed a major Dickensian technique for externalizing characters through narrator comment (see Korte 1997 cited in Mahlberg 2013).

But how exhaustively has Mahlberg carved up the space of left-as-if collocates? Here are the "residual" content words that weren’t in any of her lists. Despite not being listed, the vast majority of these collocates plainly do fall into one of her existing meaning groups.

Indeed, her analysis generalizes in large part to words that weren’t specifically studied.
Jonathan Culpeper (2014) suggests that pronominal patterns facilitate the Bard's characterization. We can check this suggestion for ourselves. What is the pronoun distribution like among Juliette's lines in Romeo and Juliette?

FOLGER> JPRO = [ pos = "p(n|o).*"] :: match.speaker_who = "Juliet_Rom";
FOLGER> count JPRO by word %c cut 5 > "juliette-pronoun-dist.txt";

Compare that to pronouns in the Corpus of Contemporary American English

COCA> PRO = [ pos = "appgel(appge_)?(pp(h|i|x|y).*?|(pp(h|i|x|y).*)_appge"]
COCA> count PRO by word %c cut 10000 > "coca-pronoun-dist.txt";

Use some unix commands to put the pronouns in a standard (alphabetical) order

awk '{ print NR,$1,$2}' coca-pronoun-dist.txt | sort -k3 > coca-pronouns-sorted.txt
awk '{ print $2,$1}' juliette-pronoun-dist.txt | sort -k1 > juliette-pronouns-sorted.txt

The join them, put back COCA-attestation-rate order, and normalize for corpus size

join -1 3 coca-pronouns-sorted.txt juliette-pronouns-sorted.txt | sort -k2 -n | awk '{print $2,$1,$3/27318237,$4/638}' > coca-vs-juliette.txt

and use gnuplot to visualize

set offsets 1,1,0,0;
set ylabel "relative frequency"
plot 'coca-vs-juliette.txt' using 3:xticlabels(2) title "COCA" lt rgb "blue", 'coca-vs-juliette.txt' using 4:xticlabels(2) title "Juliette" lt rgb "orange"

Culpeper:

Although this is not conclusive evidence, it is consistent with the idea that Juliet spends much time in the play bearing her soul.
8 French infinitival verb clusters and what alignment means

The CQP installation at UGA is configured to search the Europarl 3 collection of EU Parliamentary debates between 1996 and 2003. This collection includes about 40 million words in each of six languages: English, German, French, Spanish, Italian and Dutch. Parts of speech were assigned automatically using TreeTagger; the tagsets for each language are documented on the TreeTagger website. Consider the French tags which distinguish infinitive verbs from other forms.

EUROPARL-FR> show +pos ;
EUROPARL-FR> InfVCluster = @[ pos = "VER:infi"] [ pos = "VER.infi"]+ ;
EUROPARL-FR> group InfVCluster target word cut 100;

Did you find "faire"? What do these infinitive verb clusters mean? We can grope for a meaning by examining the aligned corpora.

The "context descriptor" indicates which corpora are aligned with the one you are currently using.

EUROPARL-FR> show cd;
== Context Descriptor=========================================

Did you find "faire"? What do these infinitive verb clusters mean? We can grope for a meaning by examining the aligned corpora.

The "context descriptor" indicates which corpora are aligned with the one you are currently using.

EUROPARL-FR> show cd;
== Context Descriptor=========================================

The "show" command switches on particular alignments. This means that each match will be accompanied by the corresponding sentences in the same alignment bead.
EUROPARL-FR> show +europarl-en ;
EUROPARL-FR> cat faireVinf cut 100;
Table 3: Residual left-collocates of "as if" that were not explicitly listed by Mahlberg

<table>
<thead>
<tr>
<th>word</th>
<th>attestations</th>
<th>Mahlberg group</th>
</tr>
</thead>
<tbody>
<tr>
<td>glance</td>
<td>25</td>
<td>LOOK</td>
</tr>
<tr>
<td>hold</td>
<td>22</td>
<td>action verbs</td>
</tr>
<tr>
<td>take</td>
<td>18</td>
<td>action verbs</td>
</tr>
<tr>
<td>think</td>
<td>17</td>
<td>action verb?</td>
</tr>
<tr>
<td>boy</td>
<td>16</td>
<td>analogous to &quot;man&quot; -&gt; Other</td>
</tr>
<tr>
<td>put</td>
<td>16</td>
<td>action verbs</td>
</tr>
<tr>
<td>sit</td>
<td>15</td>
<td>action verbs</td>
</tr>
<tr>
<td>smile</td>
<td>15</td>
<td>action? body?</td>
</tr>
<tr>
<td>see</td>
<td>13</td>
<td>LOOK</td>
</tr>
<tr>
<td>talk</td>
<td>13</td>
<td>SPEAK</td>
</tr>
<tr>
<td>breath</td>
<td>12</td>
<td>action? body?</td>
</tr>
<tr>
<td>wave</td>
<td>12</td>
<td>action? body?</td>
</tr>
<tr>
<td>Mrs</td>
<td>11</td>
<td>X honorific title</td>
</tr>
<tr>
<td>begin</td>
<td>11</td>
<td>action verb</td>
</tr>
<tr>
<td>expression</td>
<td>11</td>
<td>manner</td>
</tr>
<tr>
<td>hair</td>
<td>11</td>
<td>body part nouns</td>
</tr>
<tr>
<td>place</td>
<td>11</td>
<td>action verbs</td>
</tr>
<tr>
<td>stare</td>
<td>11</td>
<td>manner</td>
</tr>
<tr>
<td>try</td>
<td>11</td>
<td>all but 3 are &quot;try to look&quot; -&gt; LOOK</td>
</tr>
<tr>
<td>draw</td>
<td>10</td>
<td>half are &quot;draw...breath&quot; -&gt; breath</td>
</tr>
<tr>
<td>lady</td>
<td>10</td>
<td>analogous to &quot;man&quot; -&gt; Other</td>
</tr>
</tbody>
</table>
Figure 2: the COCA pronoun distribution vs Juliette’s pronoun distribution
Figure 3: Aligned corpora in Europarl 3 is at the sentence level (from Koehn 2005)