Corpus Linguistics (L615)

Concordancing & Basic Tools

Markus Dickinson

Department of Linguistics, Indiana University

Fall 2015

Basic text analysis

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

- concordancing
- word frequency counting
- categorizing
- online searching [already discussed]

Our explorations with Perl will give you even more capabilities for these types of things

A good resource to find such tools: http://tiny.cc/corpora (see "Software, Tools, Freq Lists, etc.")

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.laurenceanthony.net/software/antconc/)

- 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
- Download the *Plain Text UTF-8* version

AntConc

File loading & concordancing

Load in a file (`File` → `Open File(s)`)

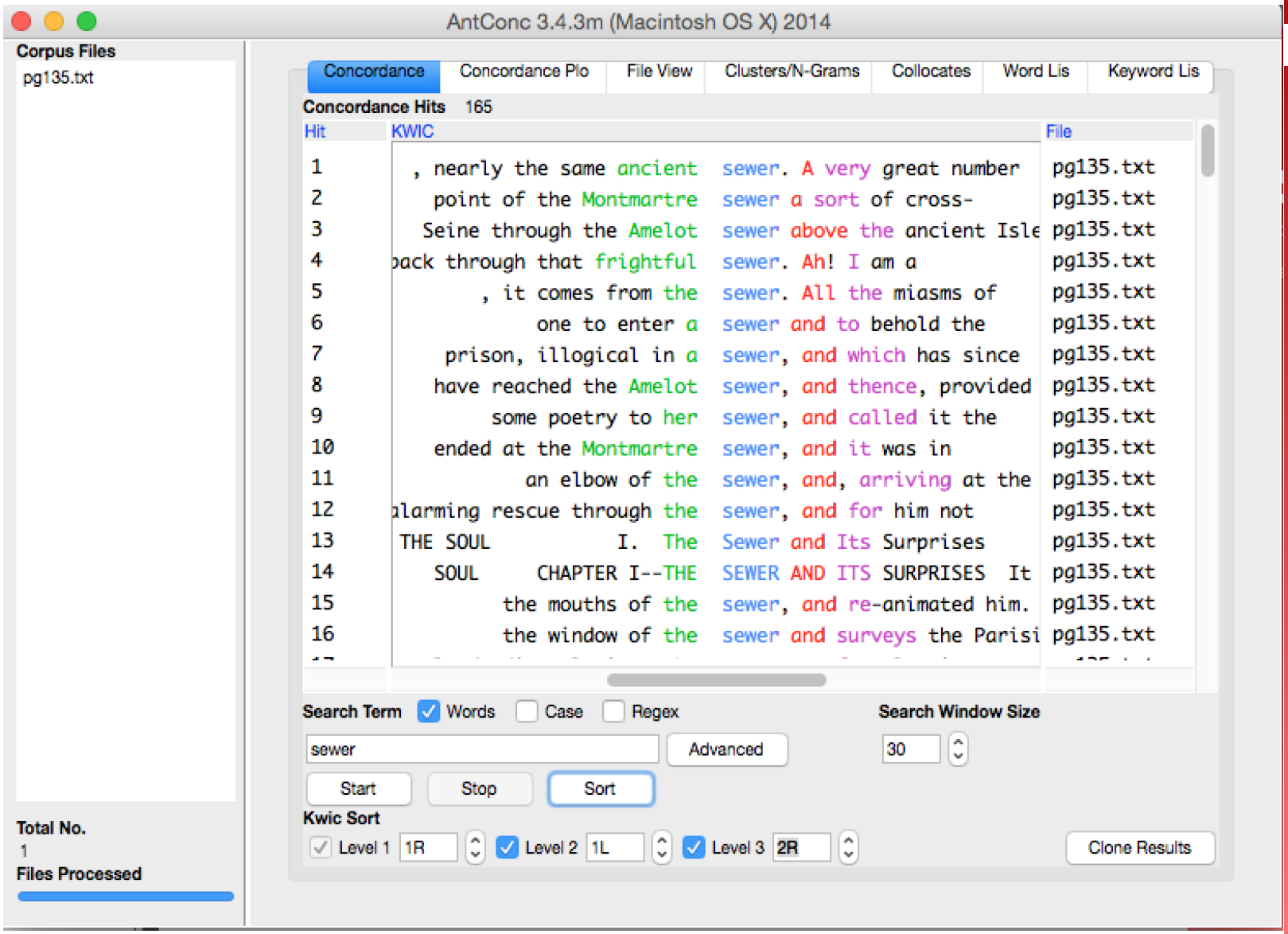
1. Enter search term
2. Determine window size
3. Start the search
4. Then, select sorting options & sort

Clicking on a word allows you to see the original context

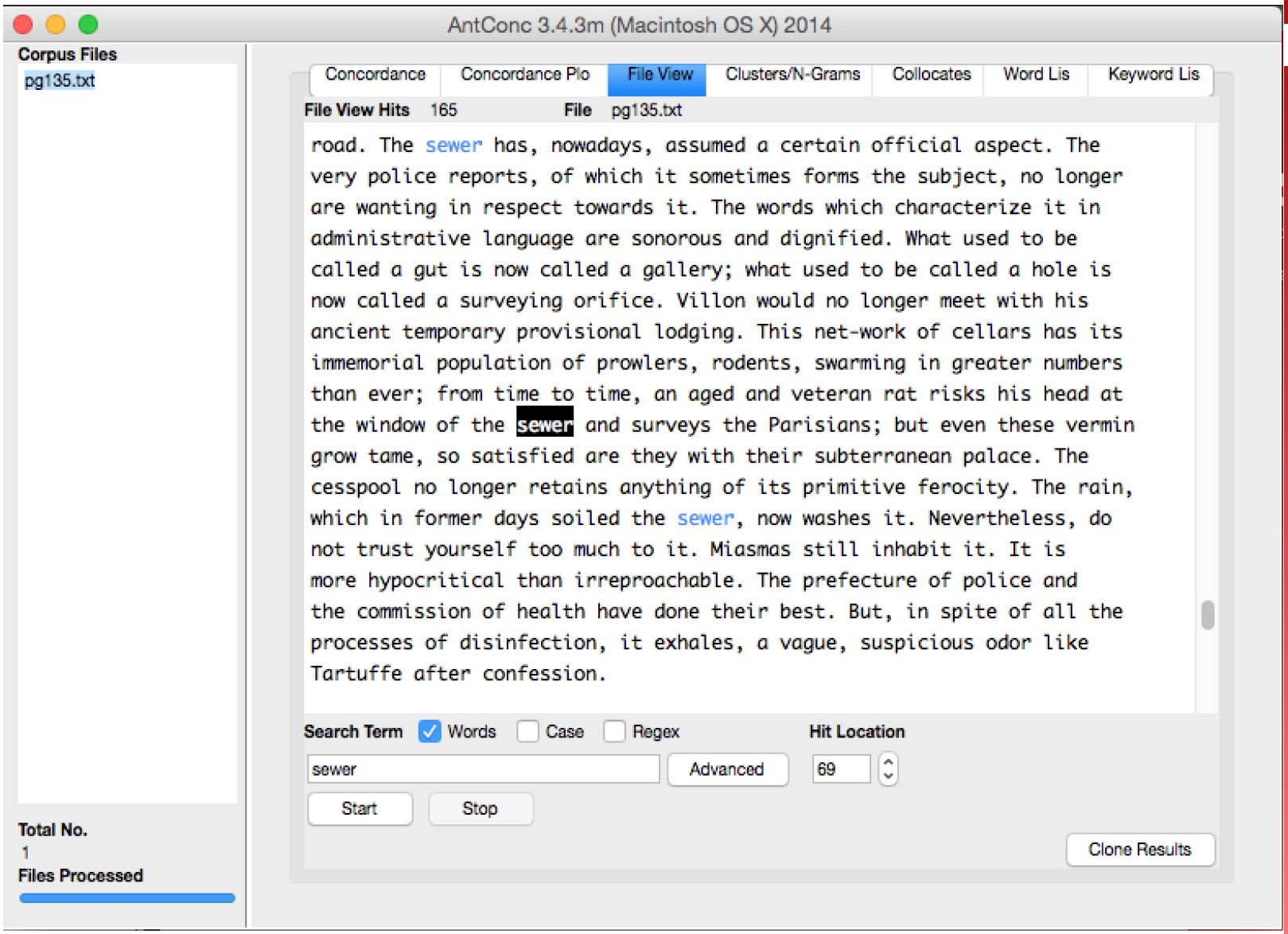
- *Advanced* options: can use word fragments, multiple search terms, regular expressions, context words

Note that you can export your search results to text files

Screenshot


AntConc 3.4.3m (Macintosh OS X) 2014
Corpus Files
pg135.txt
Total No.
1
Files Processed
Concordance
Concordance Plo
File View
Clusters/N-Grams
Collocates
Word Lis
Keyword Lis
Concordance Hits 165
Hit
KWIC
File
1 , nearly the same ancient sewer. A very great number pg135.txt
2 point of the Montmartre sewer a sort of cross- pg135.txt
3 Seine through the Amelot sewer above the ancient Isle pg135.txt
4 back through that frightful sewer. Ah! I am a pg135.txt
5 , it comes from the sewer. All the miasms of pg135.txt
6 one to enter a sewer and to behold the pg135.txt
7 prison, illogical in a sewer, and which has since pg135.txt
8 have reached the Amelot sewer, and thence, provided pg135.txt
9 some poetry to her sewer, and called it the pg135.txt
10 ended at the Montmartre sewer, and it was in pg135.txt
11 an elbow of the sewer, and, arriving at the pg135.txt
12 alarming rescue through the sewer, and for him not pg135.txt
13 THE SOUL I. The Sewer and Its Surprises pg135.txt
14 SOUL CHAPTER I--THE SEWER AND ITS SURPRISES It pg135.txt
15 the mouths of the sewer, and re-animated him. pg135.txt
16 the window of the sewer and surveys the Parisi pg135.txt
Search Term
Words
Case
Regex
Search Window Size
sewer
Advanced
30
Start
Stop
Sort
Kwic Sort
Level 1 1R
Level 2 1L
Level 3 2R
Clone Results

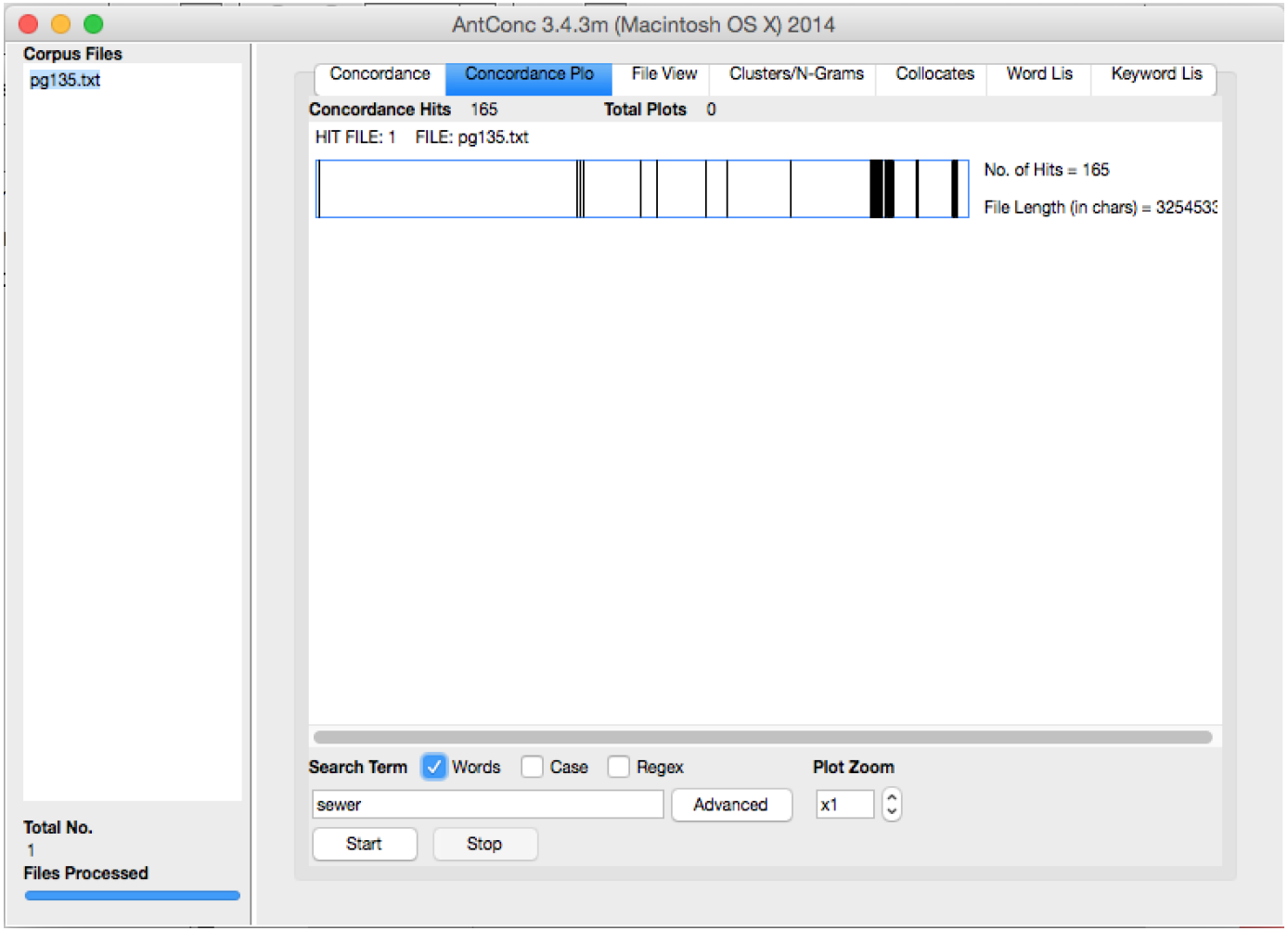
Screenshot


AntConc 3.4.3m (Macintosh OS X) 2014
Corpus Files
pg135.txt
Total No.
1
Files Processed
Concordance
Concordance Plo
File View
Clusters/N-Grams
Collocates
Word Lis
Keyword Lis
File View Hits 165
File pg135.txt
road. The sewer has, nowadays, assumed a certain official aspect. The very police reports, of which it sometimes forms the subject, no longer are wanting 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 the processes of disinfection, it exhales, a vague, suspicious odor like Tartuffe after confession.
Search Term
Words
Case
Regex
Hit Location
sewer
Advanced
69
Start
Stop
Clone Results

AntConc

Concordance plots

Concordance plots allow one to see distribution at a glance



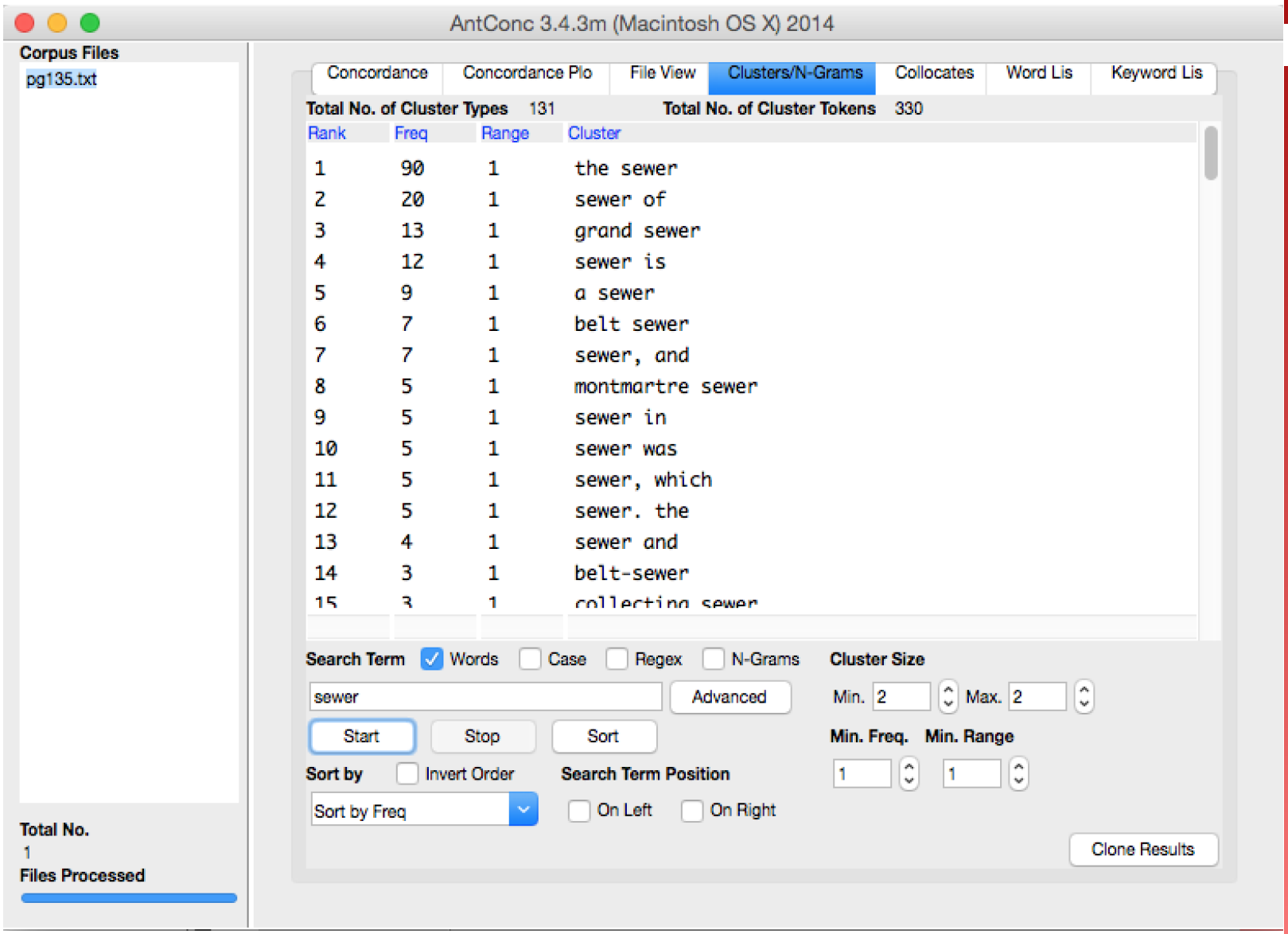
AntConc

Clusters & N-grams

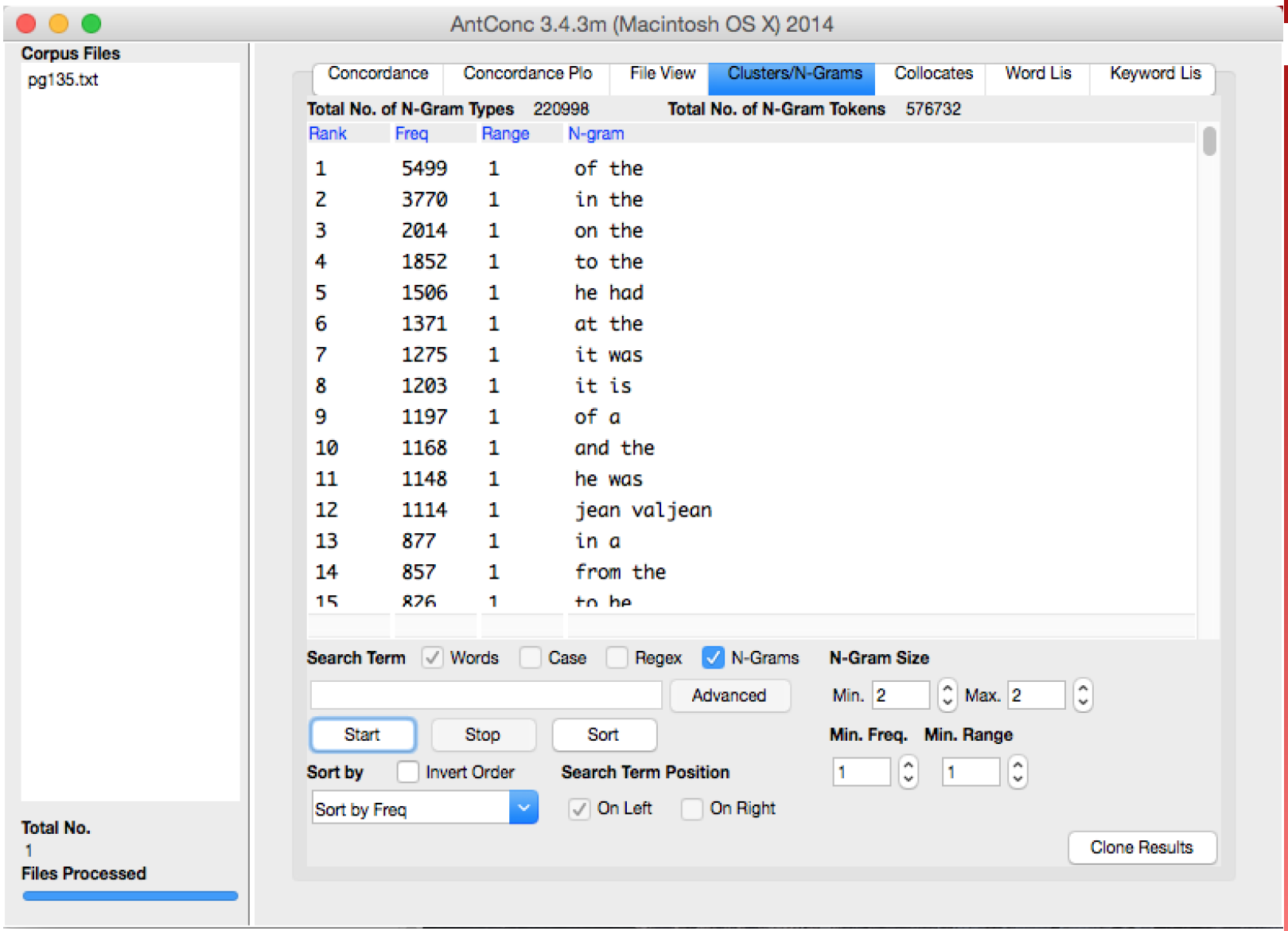
You can search for:

- Clusters involving a particular word
- All “n-grams” of a particular size
- Collocations involving a particular word

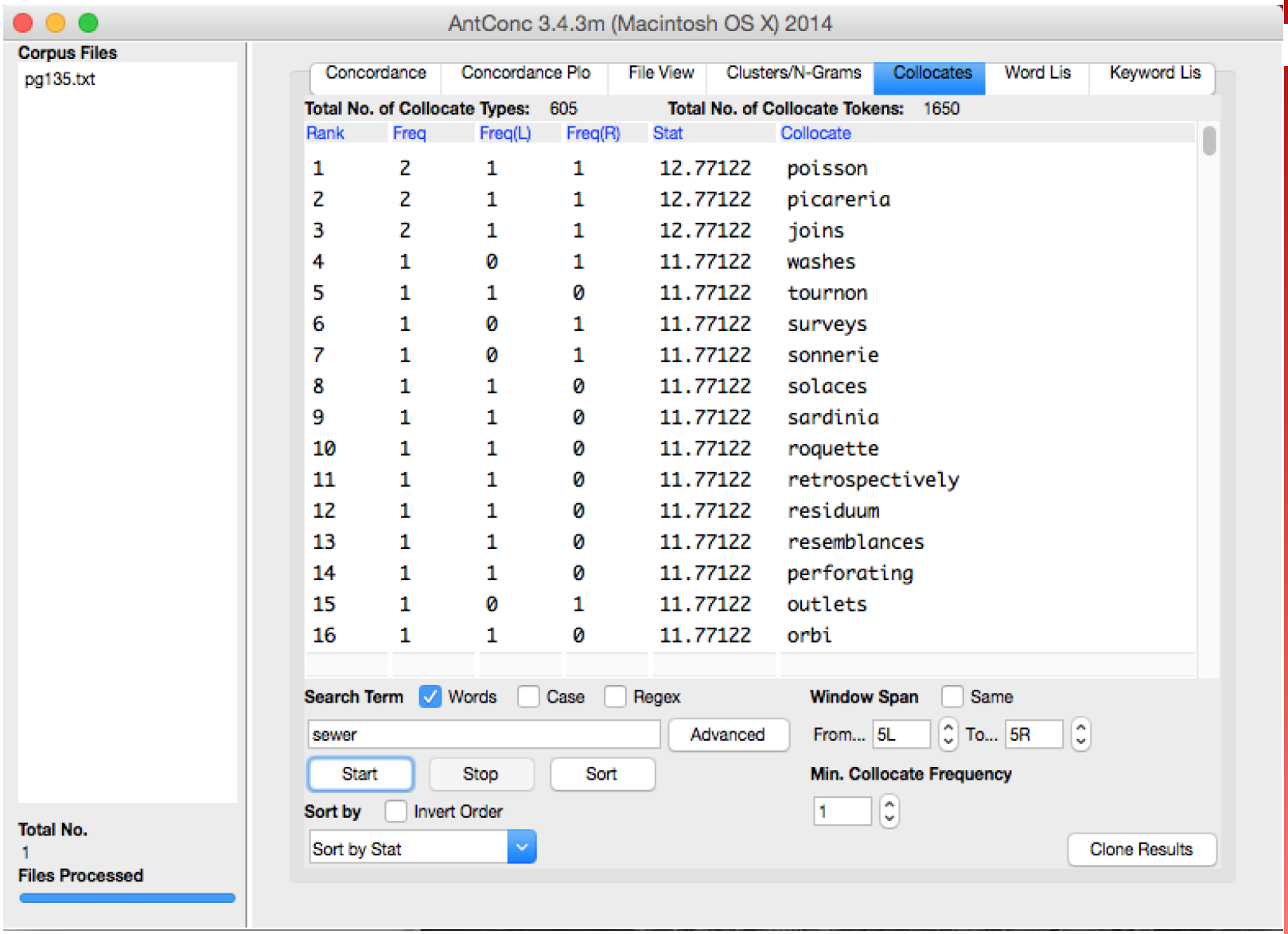
Screenshot


AntConc 3.4.3m (Macintosh OS X) 2014
Corpus Files
pg135.txt
Total No.
1
Files Processed
Concordance
Concordance Plo
File View
Clusters/N-Grams
Collocates
Word Lis
Keyword Lis
Total No. of Cluster Types 131
Total No. of Cluster Tokens 330
Rank Freq Range Cluster
1 90 1 the sewer
2 20 1 sewer of
3 13 1 grand sewer
4 12 1 sewer is
5 9 1 a sewer
6 7 1 belt sewer
7 7 1 sewer, and
8 5 1 montmartre sewer
9 5 1 sewer in
10 5 1 sewer was
11 5 1 sewer, which
12 5 1 sewer. the
13 4 1 sewer and
14 3 1 belt-sewer
15 3 1 collecting sewer
Search Term
Words
Case
Regex
N-Grams
Cluster Size
sewer
Advanced
Min. 2
Max. 2
Start
Stop
Sort
Min. Freq.
Min. Range
Sort by
Invert Order
Search Term Position
1
1
Sort by Freq
On Left
On Right
Clone Results

Screenshot


AntConc 3.4.3m (Macintosh OS X) 2014
Corpus Files
pg135.txt
Total No.
1
Files Processed
Concordance
Concordance Plo
File View
Clusters/N-Grams
Collocates
Word Lis
Keyword Lis
Total No. of N-Gram Types 220998
Total No. of N-Gram Tokens 576732
Rank Freq Range N-gram
1 5499 1 of the
2 3770 1 in the
3 2014 1 on the
4 1852 1 to the
5 1506 1 he had
6 1371 1 at the
7 1275 1 it was
8 1203 1 it is
9 1197 1 of a
10 1168 1 and the
11 1148 1 he was
12 1114 1 jean valjean
13 877 1 in a
14 857 1 from the
15 826 1 to he
Search Term
Words
Case
Regex
N-Grams
N-Gram Size
Advanced
Min. 2
Max. 2
Start
Stop
Sort
Min. Freq.
Min. Range
Sort by
Invert Order
Search Term Position
Sort by Freq
On Left
On Right
1
1
Clone Results

Screenshot


AntConc 3.4.3m (Macintosh OS X) 2014
Corpus Files
pg135.txt
Total No.
1
Files Processed
Concordance
Concordance Plo
File View
Clusters/N-Grams
Collocates
Word Lis
Keyword Lis
Total No. of Collocate Types: 605
Total No. of Collocate Tokens: 1650
Rank Freq Freq(L) Freq(R) Stat Collocate
1 2 1 1 12.77122 poisson
2 2 1 1 12.77122 picareria
3 2 1 1 12.77122 joins
4 1 0 1 11.77122 washes
5 1 1 0 11.77122 tournon
6 1 0 1 11.77122 surveys
7 1 0 1 11.77122 sonnerie
8 1 1 0 11.77122 solaces
9 1 1 0 11.77122 sardinia
10 1 1 0 11.77122 roquette
11 1 1 0 11.77122 retrospectively
12 1 1 0 11.77122 residuum
13 1 1 0 11.77122 resemblances
14 1 1 0 11.77122 perforating
15 1 0 1 11.77122 outlets
16 1 1 0 11.77122 orbi
Search Term
Words
Case
Regex
sewer
Advanced
Start
Stop
Sort
Sort by
Invert Order
Sort by Stat
Window Span
Same
From... 5L
To... 5R
Min. Collocate Frequency
1
Clone Results

AntConc

Word & Keyword lists

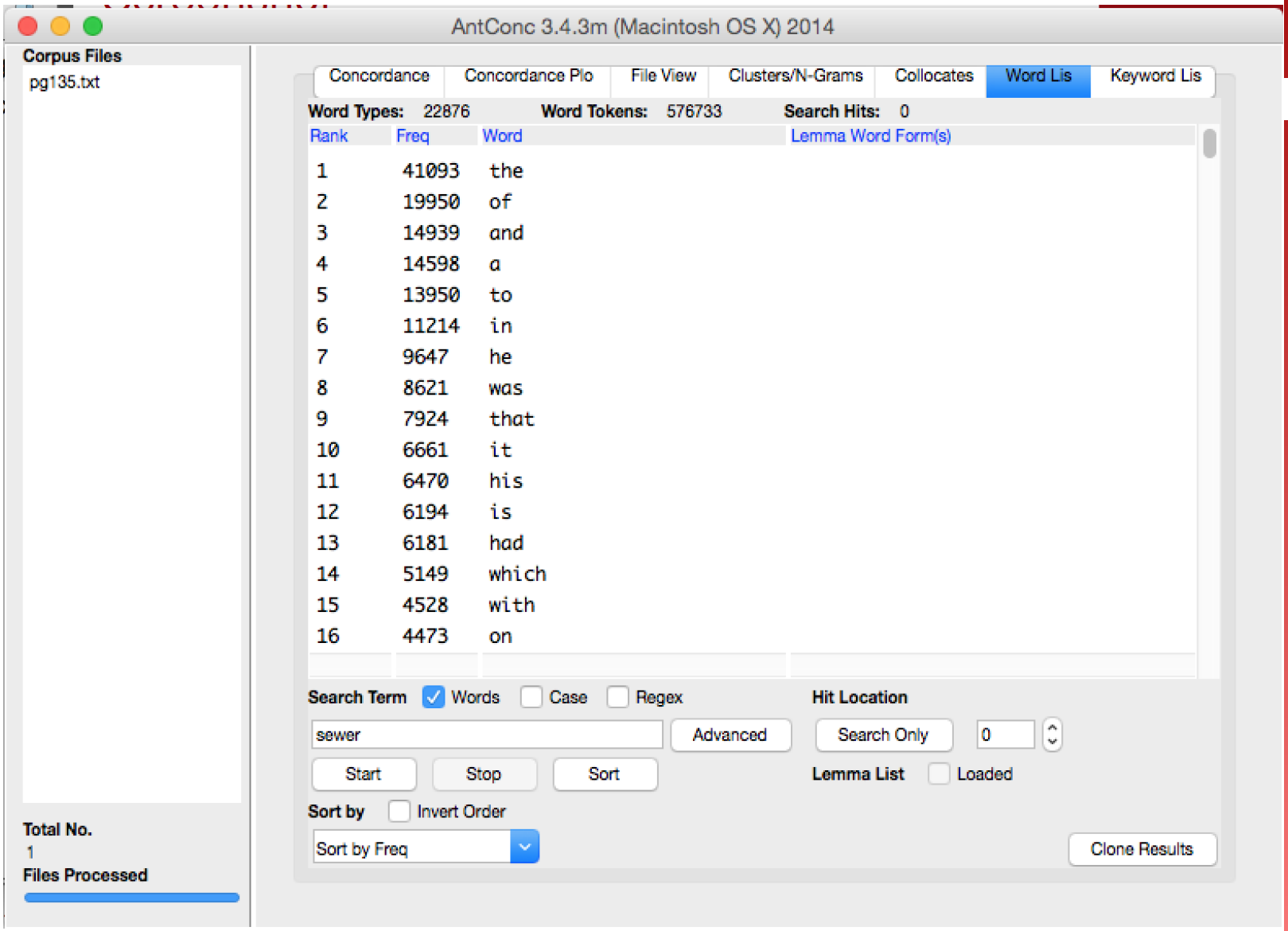
Word lists are easy to generate

- This is an easy way to get word frequency counts

You can also generate Keyword Lists, which show unusually (in)frequent words as compared to some reference corpus

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

Screenshot


AntConc 3.4.3m (Macintosh OS X) 2014
Corpus Files
pg135.txt
Total No.
1
Files Processed
Concordance
Concordance Plo
File View
Clusters/N-Grams
Collocates
Word Lis
Keyword Lis
Word Types: 22876
Word Tokens: 576733
Search Hits: 0
Rank Freq Word Lemma Word Form(s)
1 41093 the
2 19950 of
3 14939 and
4 14598 a
5 13950 to
6 11214 in
7 9647 he
8 8621 was
9 7924 that
10 6661 it
11 6470 his
12 6194 is
13 6181 had
14 5149 which
15 4528 with
16 4473 on
Search Term
Words
Case
Regex
sewer
Advanced
Hit Location
Search Only
0
Start
Stop
Sort
Lemma List
Loaded
Sort by
Invert Order
Sort by Freq
Clone Results

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

Exercise

Pick two categories from Project Gutenberg

- Hypothesize testable differences between the categories
- Try to control for non-category-related factors
  - e.g., a *travelogue* about India and a *novel* in the US will have several non-category-related differences
- Use AntConc to test your hypotheses

Categorization

If you do basic *n*-gram analysis, you can start looking into software that categorizes texts on this basis, e.g.,

- libTextCat: http://software.wise-guys.nl/libtextcat/
- TCatNG: http://tcatng.sourceforge.net

Poking around in this type of software may help you research what's going on in your data

- You can also poke around in software for clustering documents, calculating document similiarity, etc.