Natural Language Processing Tools for the Digital Humanities

Christopher Manning

Stanford University

Digital Humanities 2011

http://nlp.stanford.edu/~manning/courses/DigitalHumanities/

Commencement 2010

My humanities qualifications

- B.A. (Hons), Australian National University
- Ph.D. Linguistics, Stanford University

- But:
  - I'm not sure I've ever taken a real humanities class (if you discount linguistics classes and high school English…)


So, feel free to ask questions!

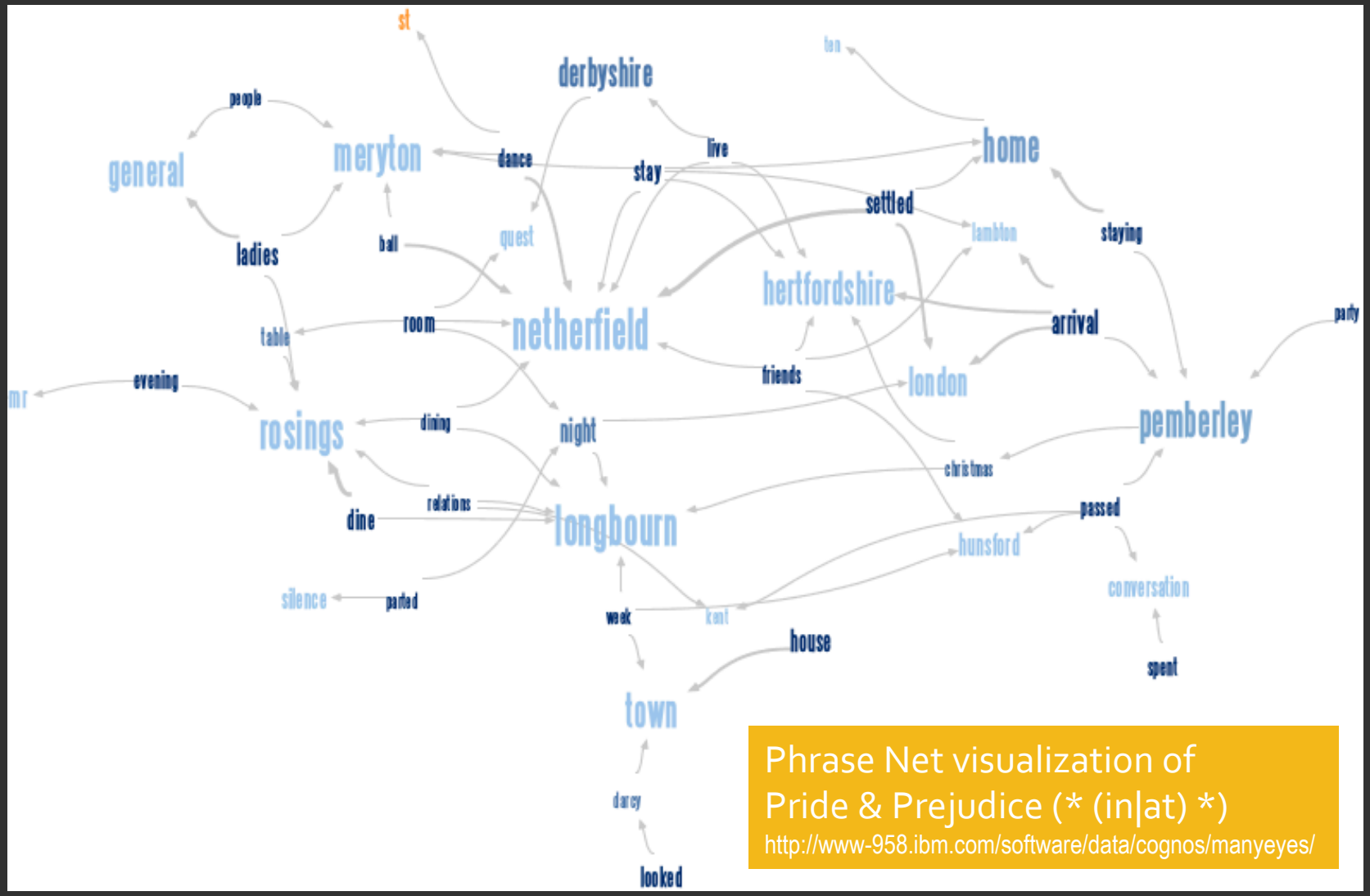
Text

legitimacy in the publishing field in those days, commercial novels were entirely excluded from hardcover release for decades.

Through the 1950's and 60's book publishing was a booming industry in the U.S., though its profit margins remained low: 4 – 6%, on average. Gone were the days of the patron publisher, as one by one, big publishers went public and became answerable to shareholders rather than their founders' sensibilities. The 1970's brought merger-mania to the publishing industry, with CEOs anxious to try out the successful business strategies of their peers in other industries. Mergers and acquisitions may have slashed costs and boosted the bottom line in manufacturing, finance and the service sectors, but publishing industry margins remained stubbornly narrow. The 1980's saw the birth of the media megaconglomerate: single, huge corporations with national or even global holdings in multiple media markets. Not content to dominate film, music, broadcast media, newspapers and periodicals, the media megas started buying up book publishers.

Thanks to over three decades of mergers, buyouts and consolidations, all of which were carried out with the motive of drastically improving profits, the U.S. publishing industry is now dominated by just six media megaconglomerates, Viacom, Time Warner and News Corp. among them. If these names sound familiar, it's because they also own and operate virtually every television, cable and radio broadcast network, as well as nearly all major magazines, newspapers, TV shows, movie studios, music labels, videogame franchises, cable channels and internet service providers in the United States. Media megas are bottom-line focused with a vengeance, and the blockbuster-centric mentality they've used to bring the mainstream film, music and TV industries to heel is now being forcibly applied to book publishing. Priorities have not just shifted, they've completely reversed. Sales forecasts are the primary driver when manuscripts are selected for publication; quality of the work is now the secondary, far distant consideration.

The promise



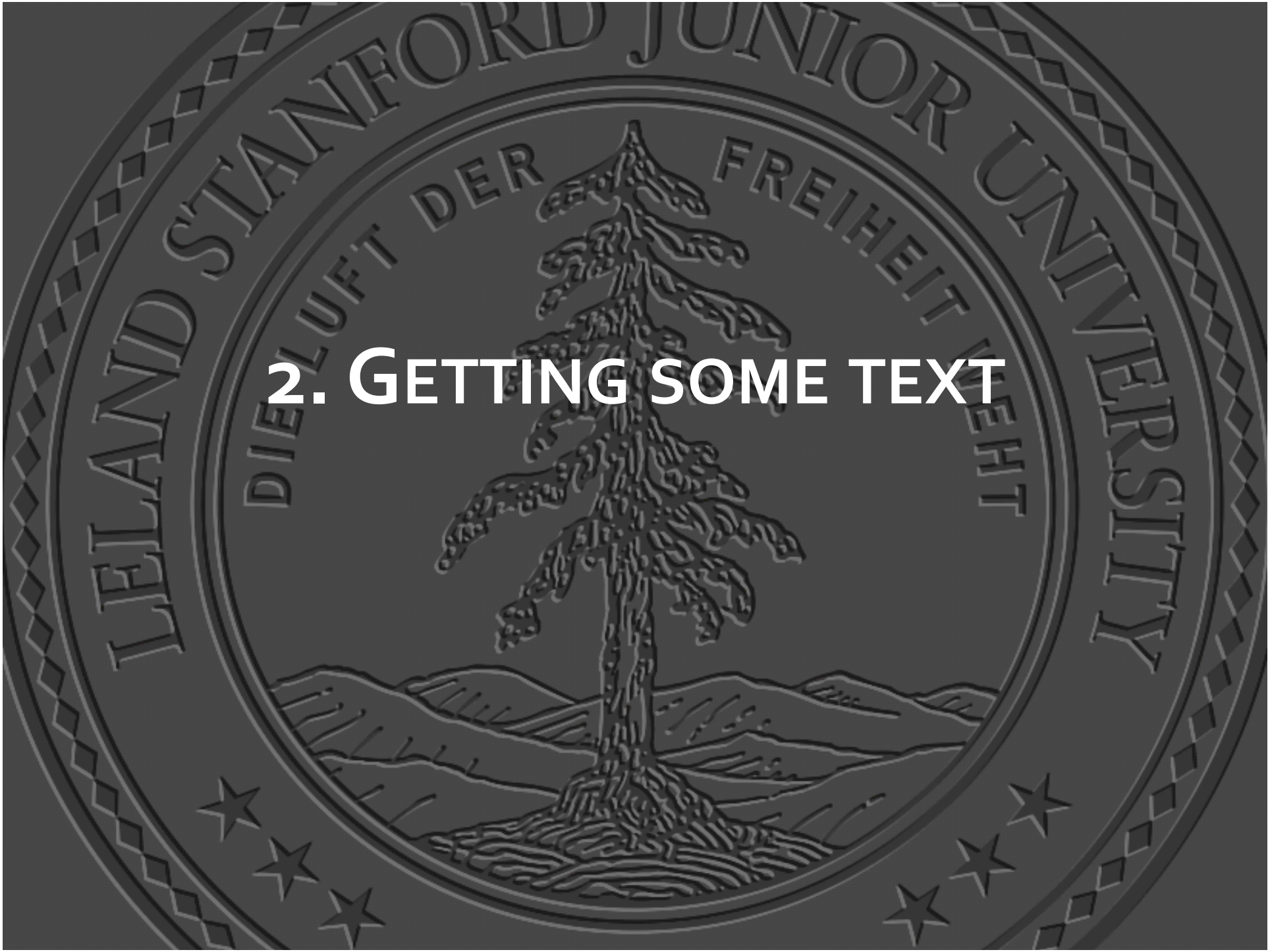
Phrase Net visualization of
Pride & Prejudice (* (in|at) *)
http://www-958.ibm.com/software/data/cognos/manyeyes/

“How I write” [code]

- I think you tend to get too much of people showing the glitzy output of something
- So, for this tutorial, at least in the slides I’m trying to include the low-level hacking and plumbing
- It’s a standard truism of data mining that more time goes into “data preparation” than anything else. Definitely goes for text processing.

Outline

1. Introduction
2. Getting some text
3. Words
4. Collocations, etc.
5. NLP Frameworks and tools
6. Part-of-speech tagging
7. Named entity recognition
8. Parsing
9. Coreference resolution
10. The rest of the languages of the world
11. Parting words



2. Getting some text

First step: Text

- To do anything, you need some texts!
  - Many sites give you various sorts of search-and-display interfaces
  - But, normally you just can't do what you want in NLP for the Digital Humanities unless you have a copy of the texts sitting on *your* computer
  - This may well change in the future: There is increasing use of cloud computing models where you might be able to upload code to run it on data on a server
    - or, conversely, upload data to be processed by code on a server

First step: Text

- People in the audience are probably more familiar with the state of play here than me, but my impression is:
  1. There are increasingly good supplies of critical texts in well-marked-up XML available commercially for license to university libraries
  2. There are various, more community efforts to produce good digitized collections, but most of those seem to be available to "friends" rather than to anybody with a web browser
  3. There's Project Gutenberg ☺
     - Plain text, or very simple HTML, which may or may not be automatically generated
     - Unicode utf-8 if you're lucky, US-ASCII if you're not

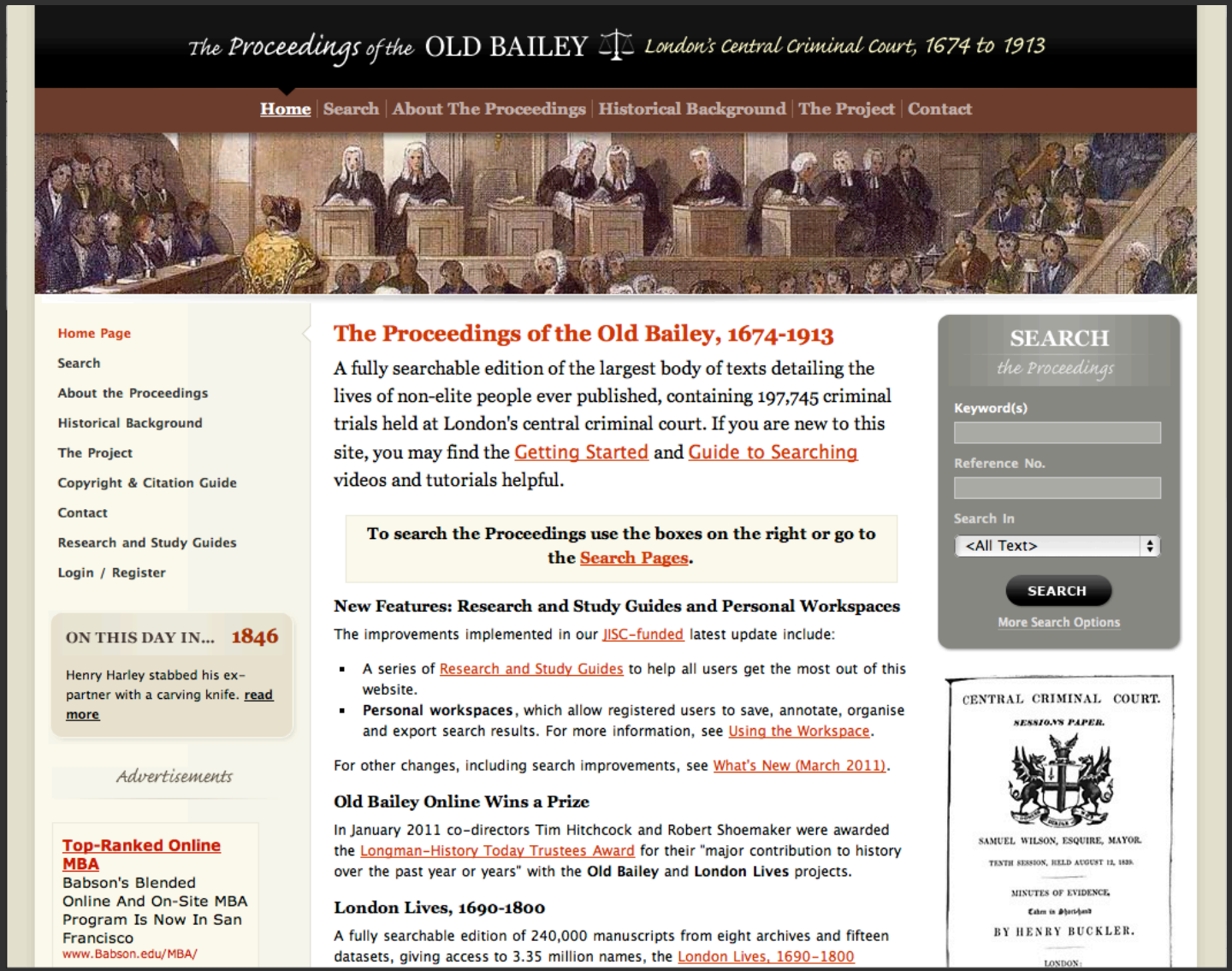
1. Early English Books Online

- TEI-compliant XML texts
- http://eebo.chadwyck.com/

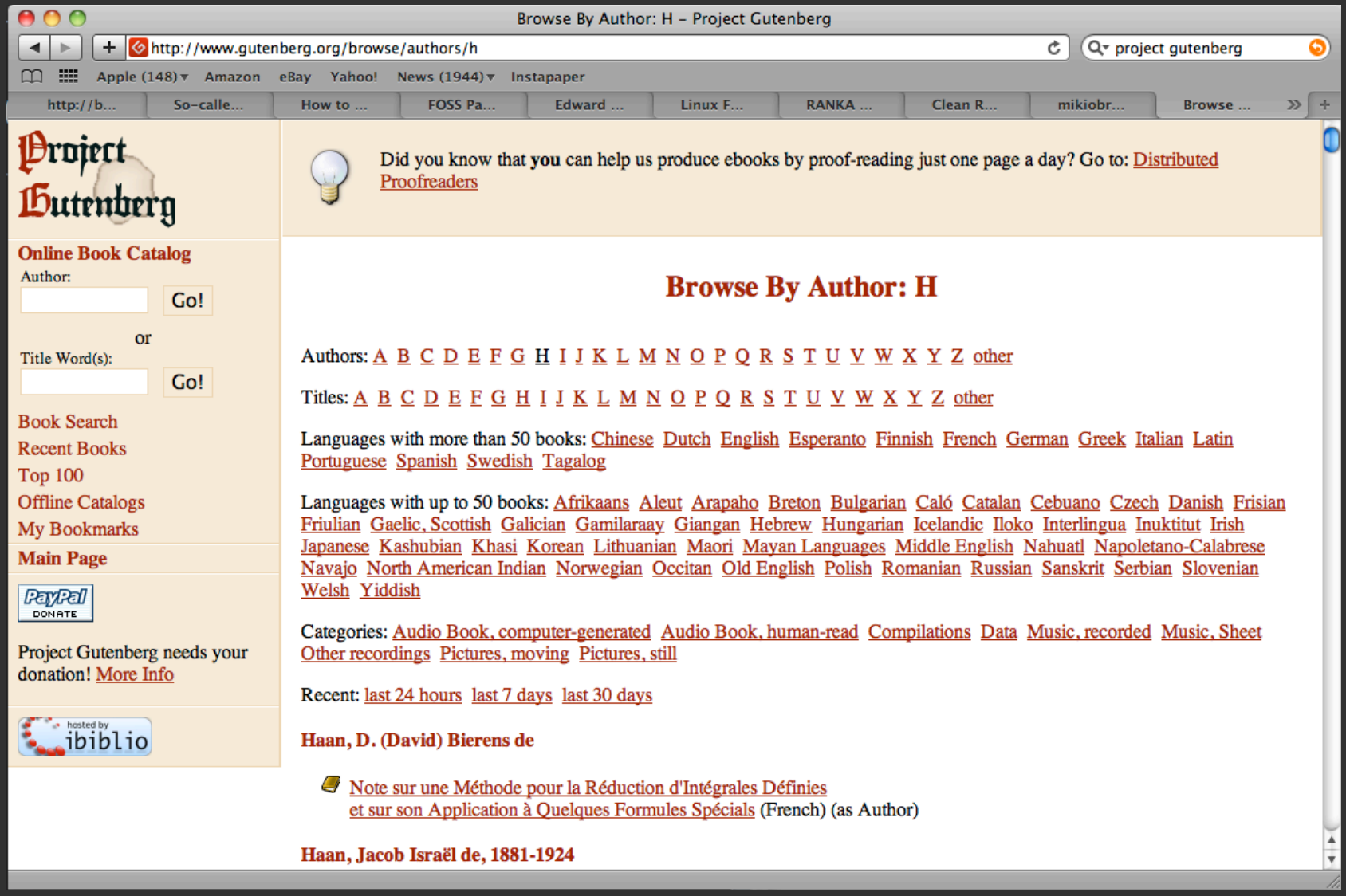
What is Early English Books Online?

From the first book published in English through the age of Spenser and Shakespeare, this incomparable collection now contains more than 125,000 titles listed in Pollard & Redgrave's *Short-Title Catalogue (1475-1640)* and Wing's *Short-Title Catalogue (1641-1700)* and their revised editions, as well as the *Thomason Tracts (1640-1661)* collection and the *Early English Books Tract Supplement*. Libraries possessing this collection find they are able to fulfill the most exhaustive research requirements of graduate scholars - from their desktop - in many subject areas, including English literature, history, philosophy, linguistics, theology, music, fine arts, education, mathematics, and science.

2. Old Bailey Online



3. Project Gutenberg



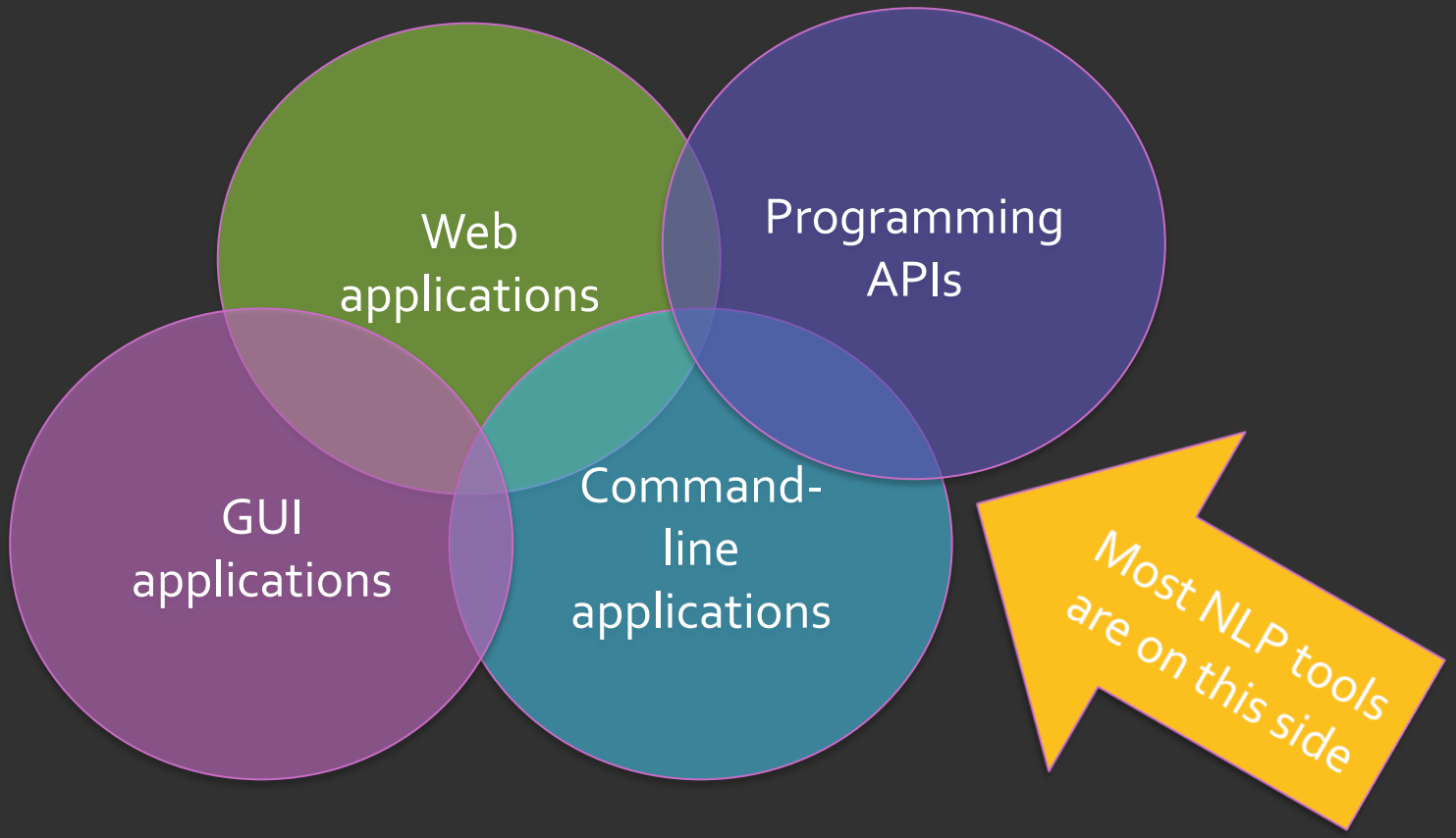
Running example: H. Rider Haggard

- The hugely popular *King Solomon's Mines* (1885) by H. Rider Haggard is sometimes considered the first of the "Lost World" or "Imperialist Romance" genres

- *Allan Quatermain* (1887)
- *She* (1887)
- *Nada the Lily* (1892)
- *Ayesha: The Return of She* (1905)
- *She and Allan* (1921)

- Zip file at: http://nlp.stanford.edu/~manning/courses/DigitalHumanities/

Interfaces to tools



You'll need to program

- Lisa Spiro, TAMU Digital Scholarship 2009:

  I'm a digital humanist with only limited programming skills (Perl & XSLT). Enhancing my programming skills would allow me to:

  - Avoid so much tedious, manual work
  - Do citation analysis
  - Pre-process texts (remove the junk)
  - Automatically download web pages
  - And much more…

You'll need to program

- Program in what?
  - Perl
    - Traditional seat-of-the-pants scripting language for text processing (it nailed flexible regex). I use it some below….
  - Python
    - Cleaner, more modern scripting language with a lot of energy, and the best-documented NLP framework, NLTK.
  - Java
    - There are more NLP tools for Java than any other language. And it's one of those most popular languages in general. Good regular expressions, Unicode, etc.

You'll need to program

- Program with what?
  - There are some general skills that you'll want the cut across programming languages
    - Regular expressions
    - XML, especially XPath and XSLT
    - Unicode
- But I'm wisely not going to try to teach programming or these skills in this tutorial ☺

Grabbing files from websites

- wget (Linux) or curl (Mac OS X, BSD)
  - wget http://www.gutenberg.org/browse/authors/h
  - curl -O http://www.gutenberg.org/browse/authors/h

- If you really want to use your browser, there are things you can get like this Firefox plug-in
  - DownThemAll http://www.downthemall.net/

  but then you just can't do things as flexibly

Grabbing files from websites

```
#!/usr/bin/perl
while (<>) { last if (m/Haggard/); }
while (<>) {
   last if (m/Hague/);
   if (m!pgdbetext\"><a href="/ebooks/(\d+)">(.*)</a> \(English\)!) {
      $title = $2;
      $num = $1;
      $title =~ s/<br>/ /g;
      $title =~ s/\r//g;
      print "curl -o \"$title $num.txt\" http://www.gutenberg.org/cache/epub/$num/pg$num.txt\n";
      # Expect only one of the html to exist
      print "curl -o \"$title $num.html\" http://www.gutenberg.org/files/$num/$num-h/$num-h.htm\n";
      print "curl -o \"$title $num-g.html\" http://www.gutenberg.org/cache/epub/$num/pg$num.html\n";
   }
}
```

Grabbing files from websites

```
wget http://www.gutenberg.org/browse/authors/h
perl getHaggard.pl < h > h.sh
chmod 755 h.sh
./h.sh
# and a bit of futzing by hand that I will leave out….
```

- Often you want the 90% solution: automating nothing would be slow and painful, but automating everything is more trouble than it's worth for a one-off process

Typical text problems

"Devilish strange!" thought he, chuckling to himself; "queer business! Capital trick of the cull in the cloak to make another person's brat stand the brunt for his own---capital! ha! ha! Won't do, though. He must be a sly fox to get out of the Mint without my

[Page 59]

knowledge. I've a shrewd guess where he's taken refuge; but I'll ferret him out. These bloods will pay well for his capture; if not, *he'll* pay well to get out of their hands; so I'm safe either way---ha! ha! Blueskin," he added aloud, and motioning that worthy, "follow me."

Upon which, he set off in the direction of the entry. His progress, however, was checked by loud acclamations, announcing the arrival of the Master of the Mint and his train.

Baptist Kettleby (for so was the Master named) was a "goodly portly man, and a corpulent," whose fair round paunch bespoke the affection he entertained for good liquor and good living. He had a quick, shrewd, merry eye, and a look in which duplicity was agreeably veiled by good humour. It was easy to discover that he was a knave, but equally easy to perceive that he was a pleasant fellow; a combination of qualities by no means of rare occurrence. So far as regards his attire, Baptist was not seen to advantage. No great lover of state or state costume at any time, he was

[Page 60]

generally, towards the close of an evening, completely in dishabille, and in this condition he now presented himself to his subjects. His shirt was unfastened, his vest unbuttoned, his hose ungartered; his feet were stuck into a pair of pantoufles, his arms into a greasy flannel dressing-gown, his head into a thrum-cap, the cap into a tie-periwig, and the wig into a gold-edged hat. A white apron was tied round his waist, and into the apron was thrust a short thick truncheon, which looked very much like a rolling-pin.

The Master of the Mint was accompanied by another gentleman almost as portly as himself, and quite as deliberate in his movements. The costume of this personage was somewhat singular, and might have passed for a masquerading habit, had not the imperturbable gravity of his demeanour forbidden any such supposition. It consisted of a close jerkin of brown frieze, ornamented with a triple row of brass buttons; loose Dutch slops, made very wide in the seat and very tight at the knees; red stockings with black clocks, and

[Page 61]

a fur cap. The owner of this dress had a broad weather-beaten face, small twinkling eyes, and a bushy, grizzled beard. Though he walked by the side of the governor, he seldom exchanged a word with him, but appeared wholly absorbed in the contemplations inspired by a broad-bowled Dutch pipe.

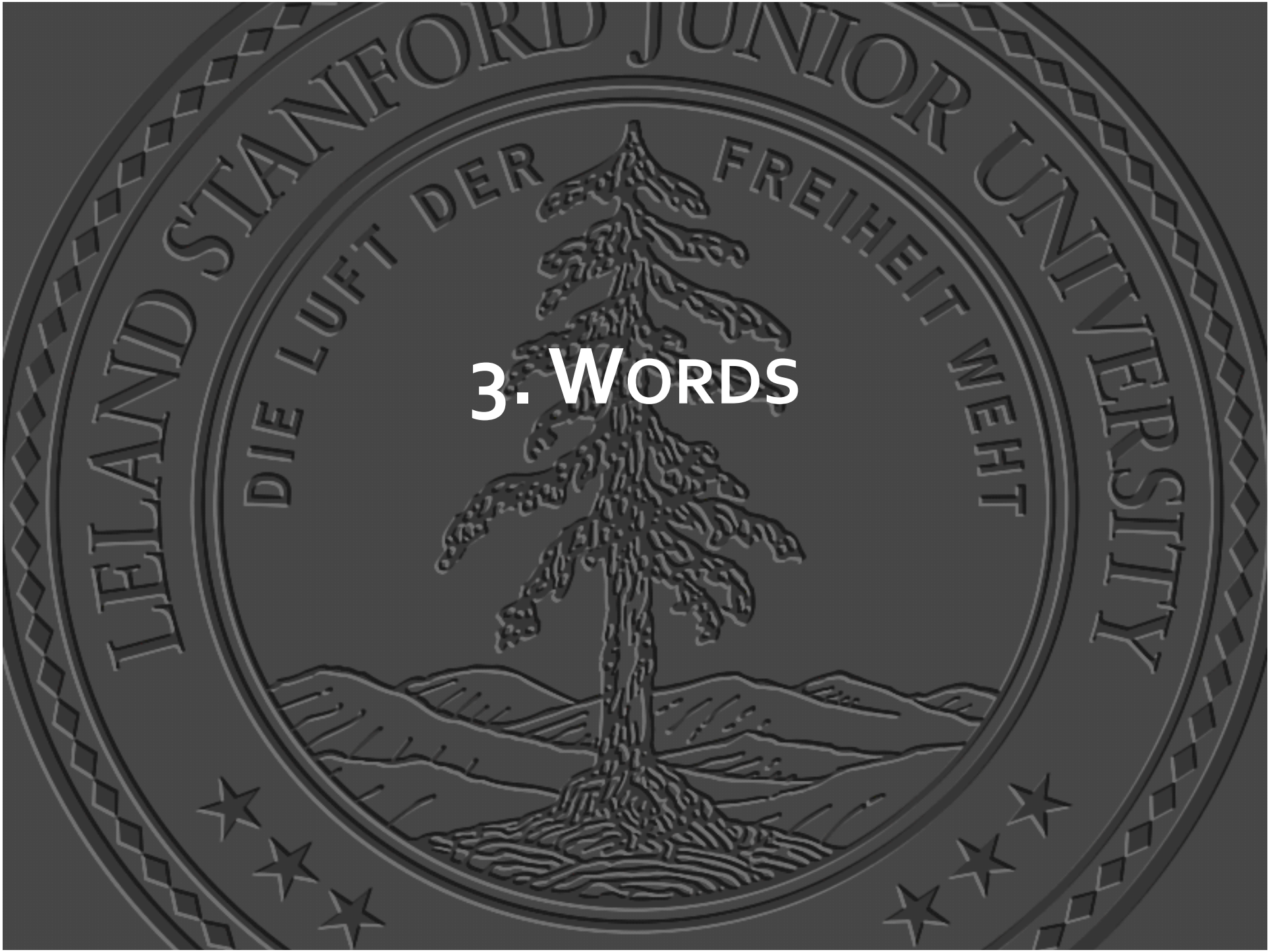
There are always text-processing gotchas …

- … and not dealing with them can badly degrade the quality of subsequent NLP processing.

1. The Gutenberg *.txt files frequently represent italics with _underscores_.
2. There may be file headers and footers
3. Elements like headings may be run together with following sentences if not demarcated or eliminated (example later).

There are always text-processing gotchas …

```
#!/usr/bin/perl
$finishedHeader = 0;
$startedFooter = 0;
while ($line = <>) {
  if ($line =~ /^\*\*\*\s*END/ && $finishedHeader) {
    $startedFooter = 1;
  }
  if ($finishedHeader && ! $startedFooter) {
    $line =~ s/_//g;  # minor cleanup of italics
    print $line;
  }
  if ($line =~ /^\*\*\*\s*START/ && ! $finishedHeader) {
    $finishedHeader = 1;
  }
}
if ( ! ($finishedHeader && $startedFooter)) {
  print STDERR "**** Probable book format problem!\n";
}
```

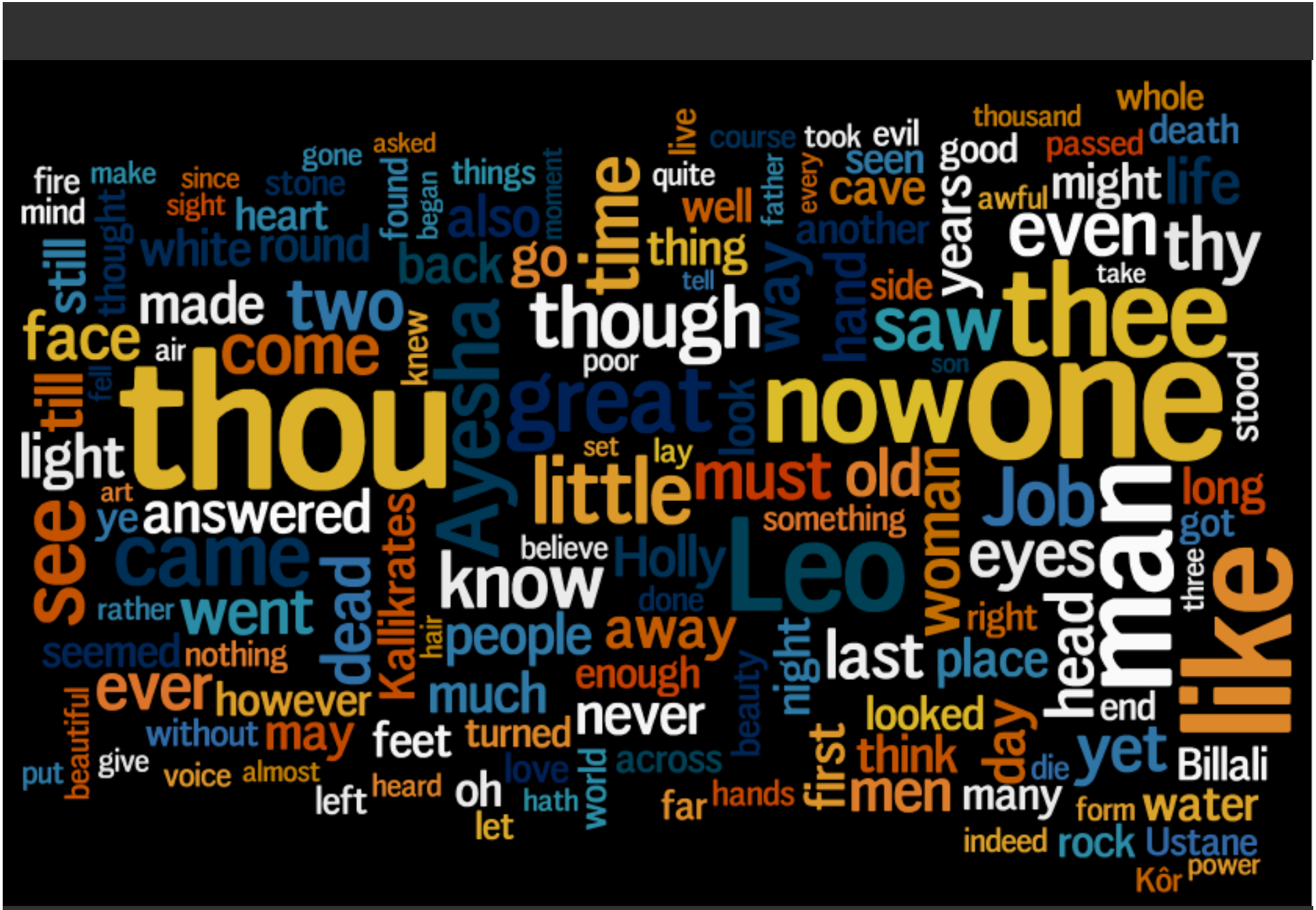

LELAND STANFORD JUNIOR UNIVERSITY
DIE LUFT DER FREIHEIT WEHT

3. Words

In the beginning was the word

- Word counts

- Word counts are the basis of all the simple, first order methods of text analysis
  - tag clouds, collocations, topic models
- Sometimes you can get a fair distance with word counts



She (1887)

http://wordle.net/ Jonathan Feinberg

Ayesha: The Return of She (1905)

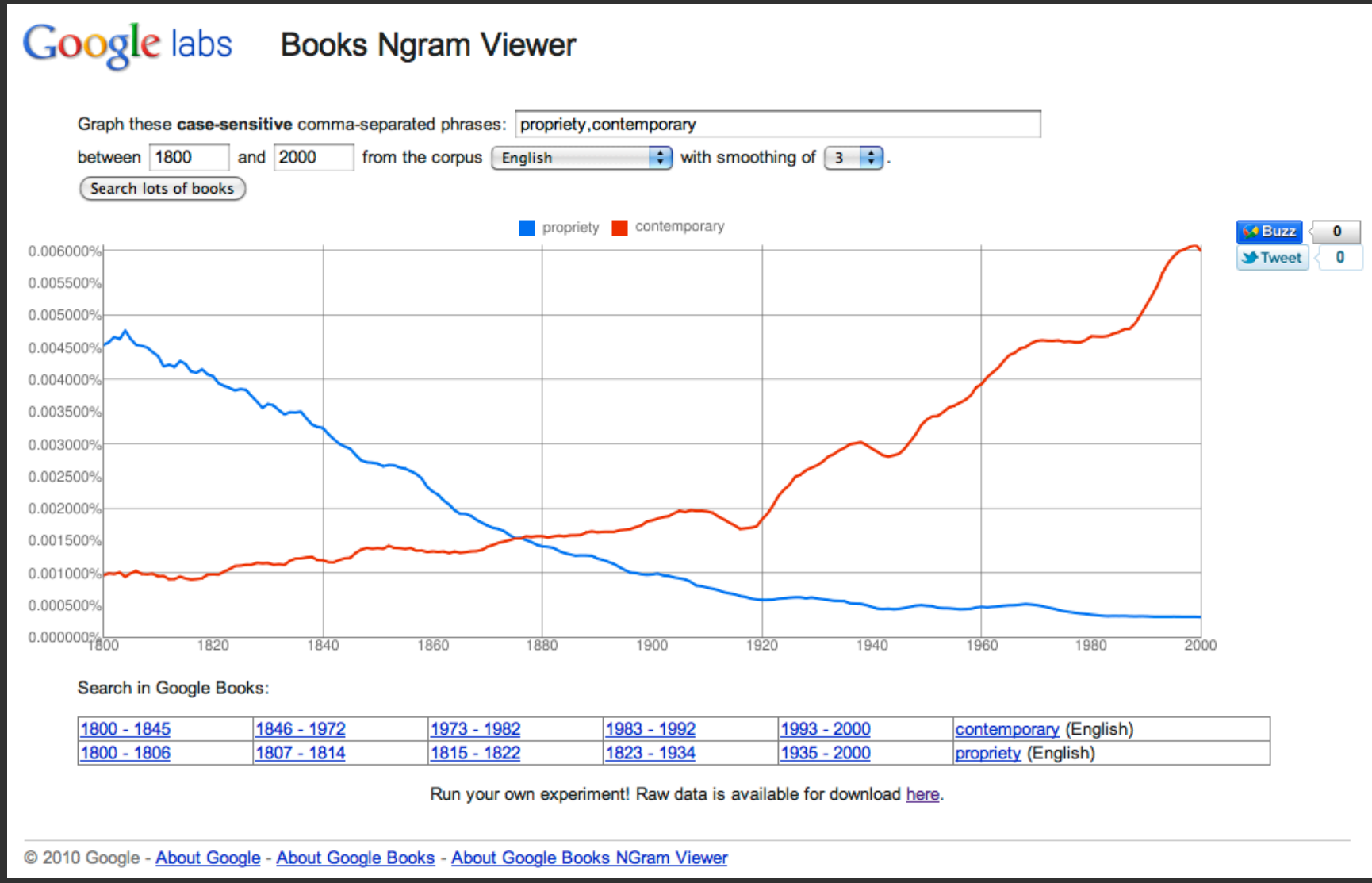
She and Allan (1921)

Wisdom's Daughter: The Life and Love Story of She-Who-Must-Be-Obeyed (1923)

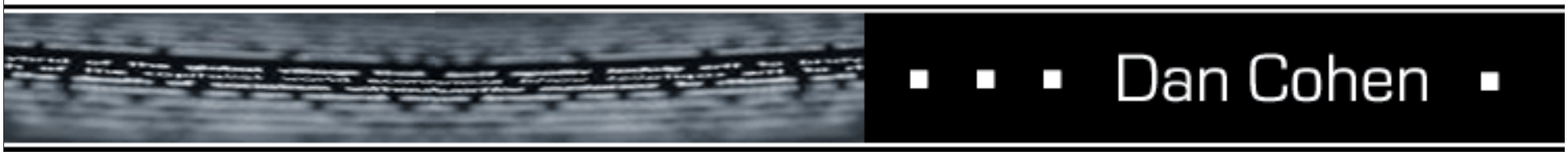
Wisdom's Daughter: The Life and Love Story of She-Who-Must-Be-Obeyed (1923)

Google Books Ngram Viewer

http://ngrams.googlelabs.com/



Google Books Ngram Viewer



- … you have to be the most jaded or cynical scholar not to be excited by the release of the Google Books Ngram Viewer … **Digital humanities needs gateway drugs.** … "Culturomics" sounds like an 80s new wave band. If we're going to coin neologisms, let's at least go with Sean Gillies' satirical alternative: ***Freakumanities***.… For me, the biggest problem with the viewer and the data is that you cannot seamlessly move from distant reading to close reading

Language change: *as least as*

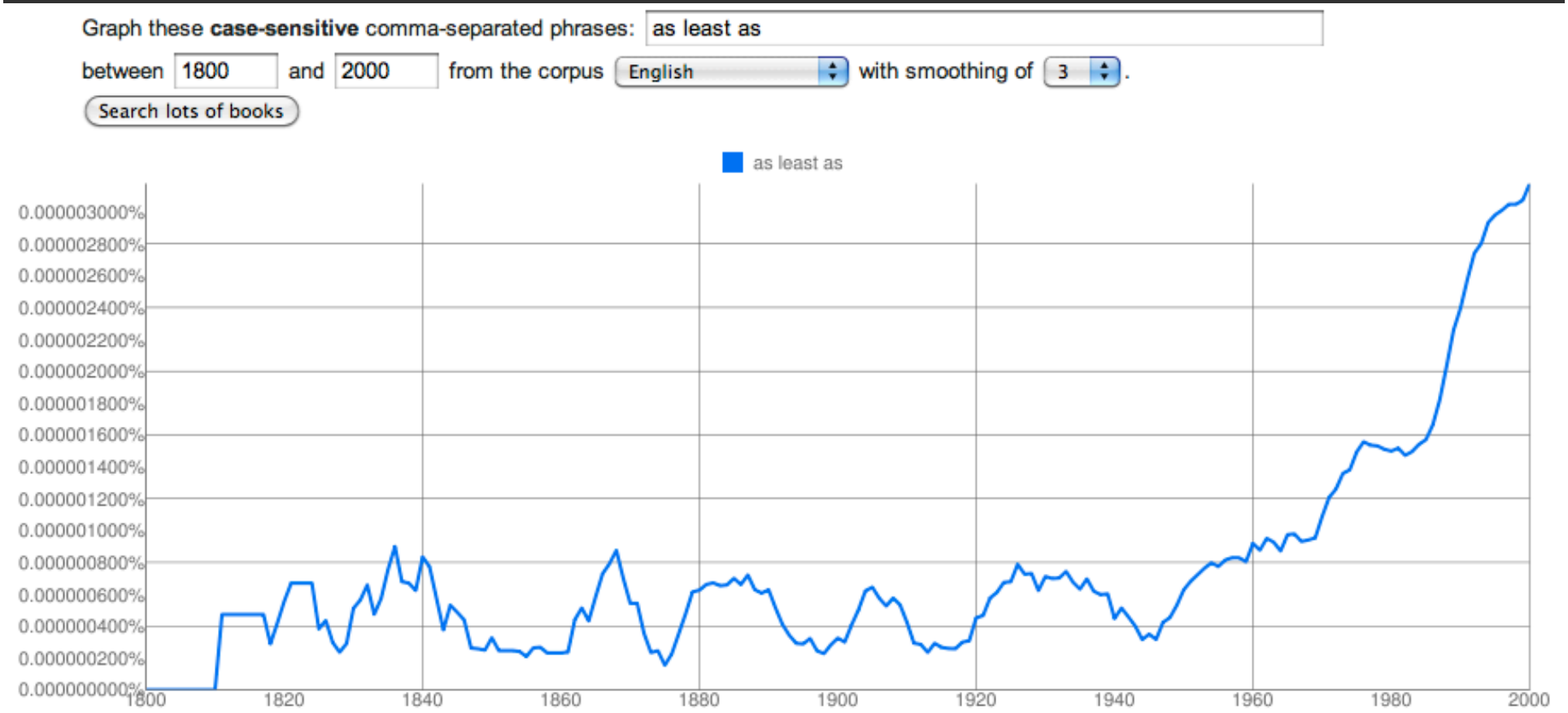
C. D. Manning. 2003. Probabilistic Syntax

- I found this example in Russo R., 2001, *Empire Falls* (on p.3!)*:*
  - By the time their son was born, though, Honus Whiting was beginning to understand and privately share his wife's opinion, as least as it pertained to Empire Falls.
- What's interesting about it?

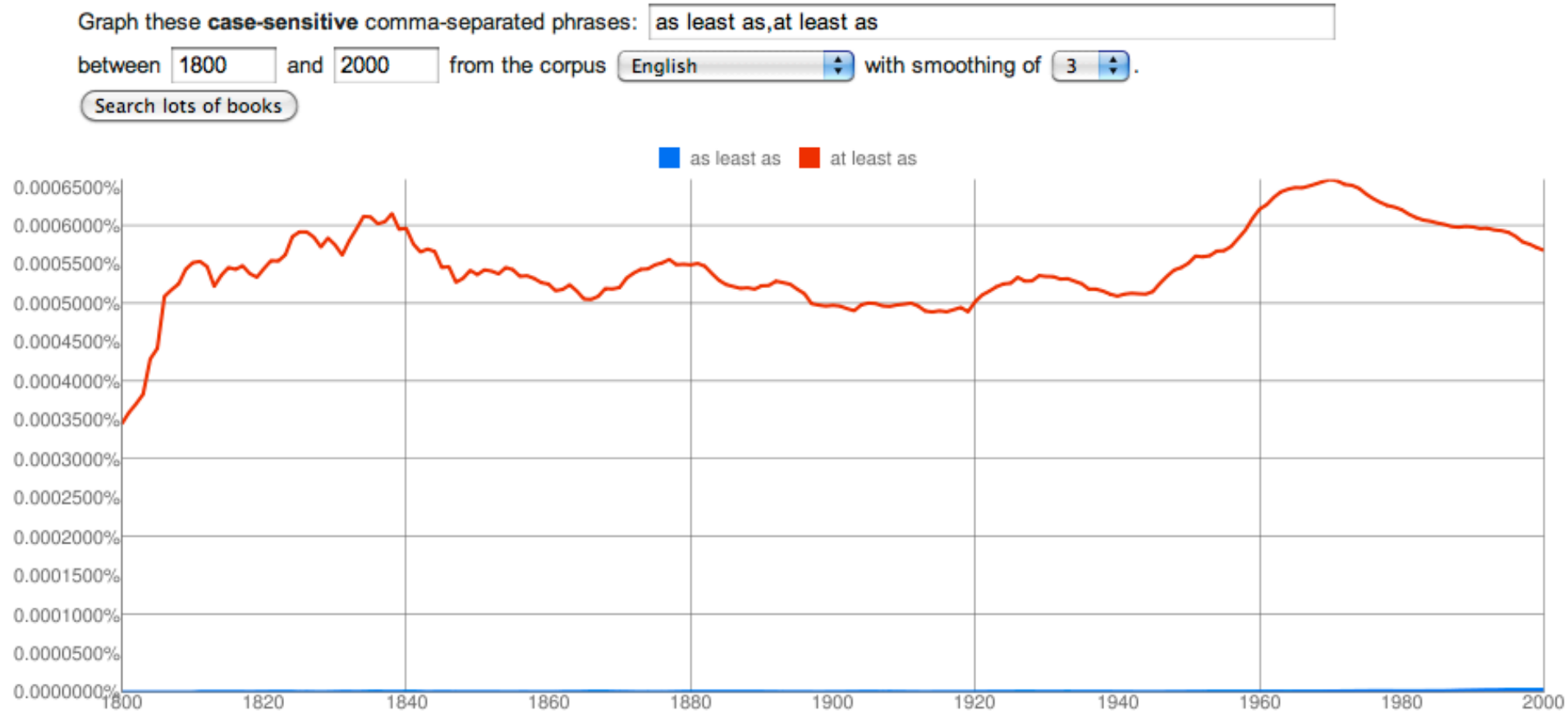
Language change: *as least as*

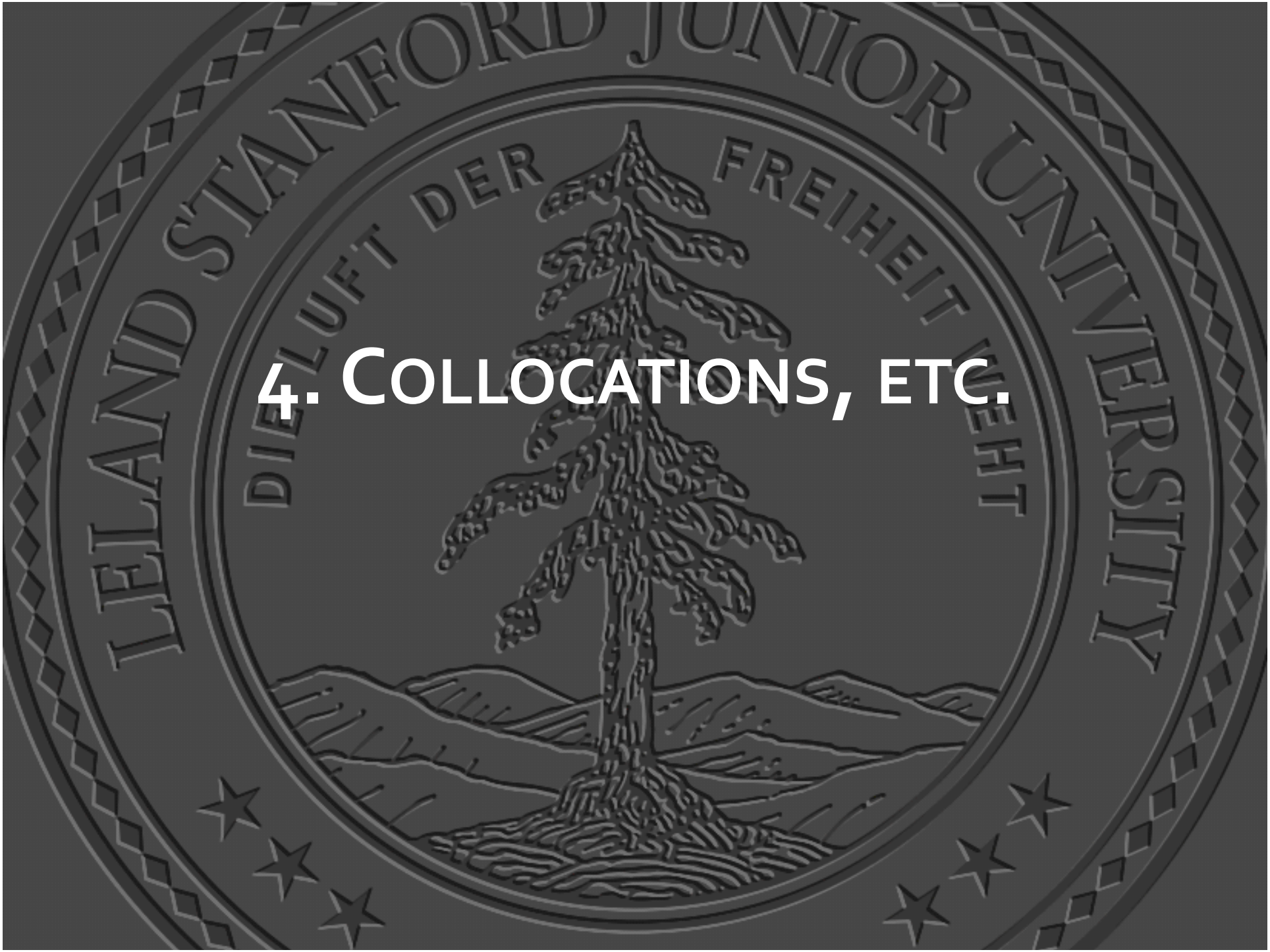
- A language change in progress? I found a bunch of other examples:
  - Indeed, the will and the means to follow through are *as least as* important *as* the initial commitment to deficit reduction.
  - As many of you know he had his boat built at the same time as mine and it's *as least as* well maintained and equipped.
- Apparently not a "dialect"
  - Second, if the required disclosures are made by on-screen notice, the disclosure of the vendor's legal name and address must appear on one of several specified screens on the vendor's electronic site and must be *at least as* legible and set in a font *as least as* large *as* the text of the offer itself.

Language change: *as least as*



Language change: *as least as*

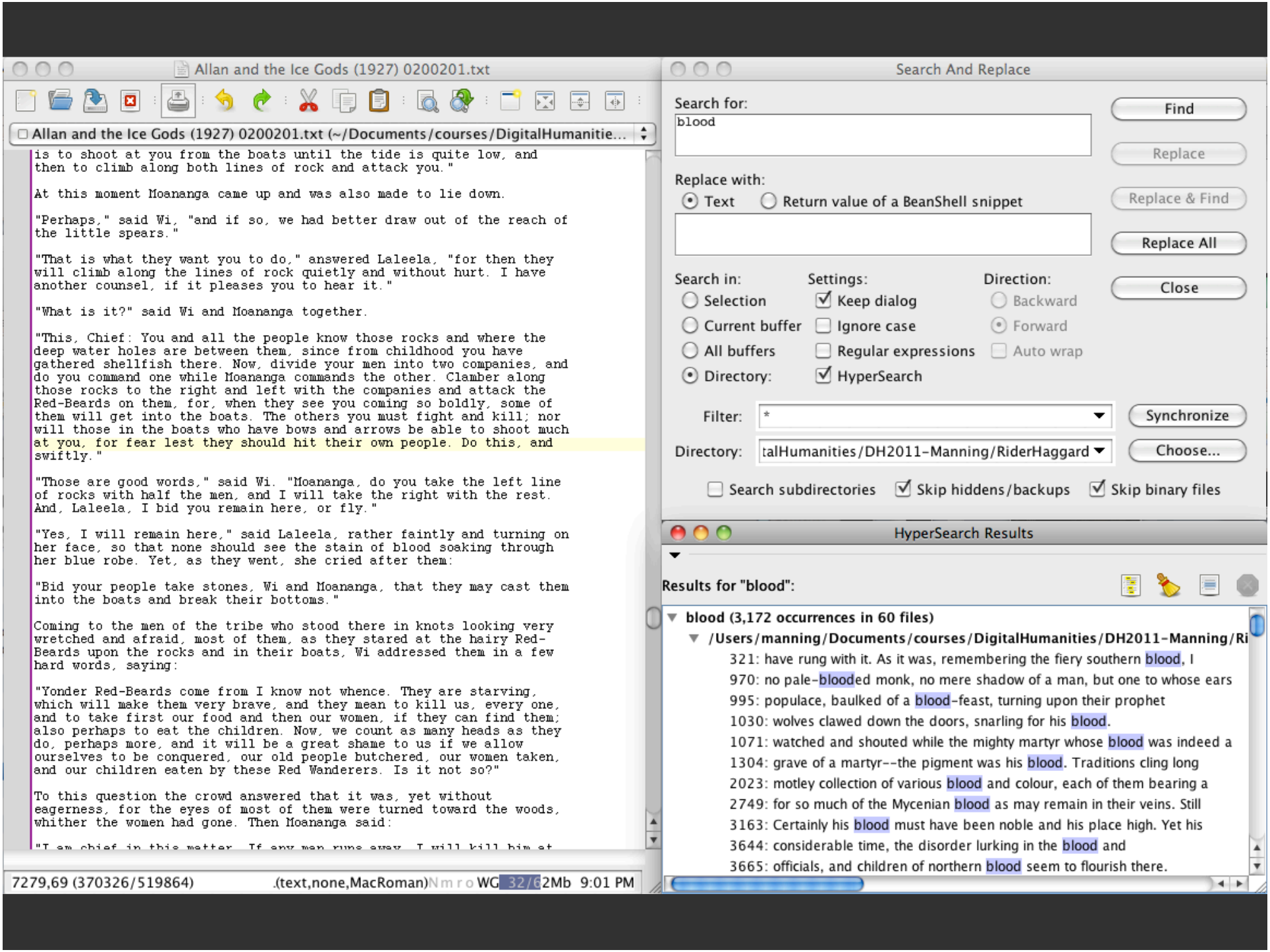




4. Collocations, etc.

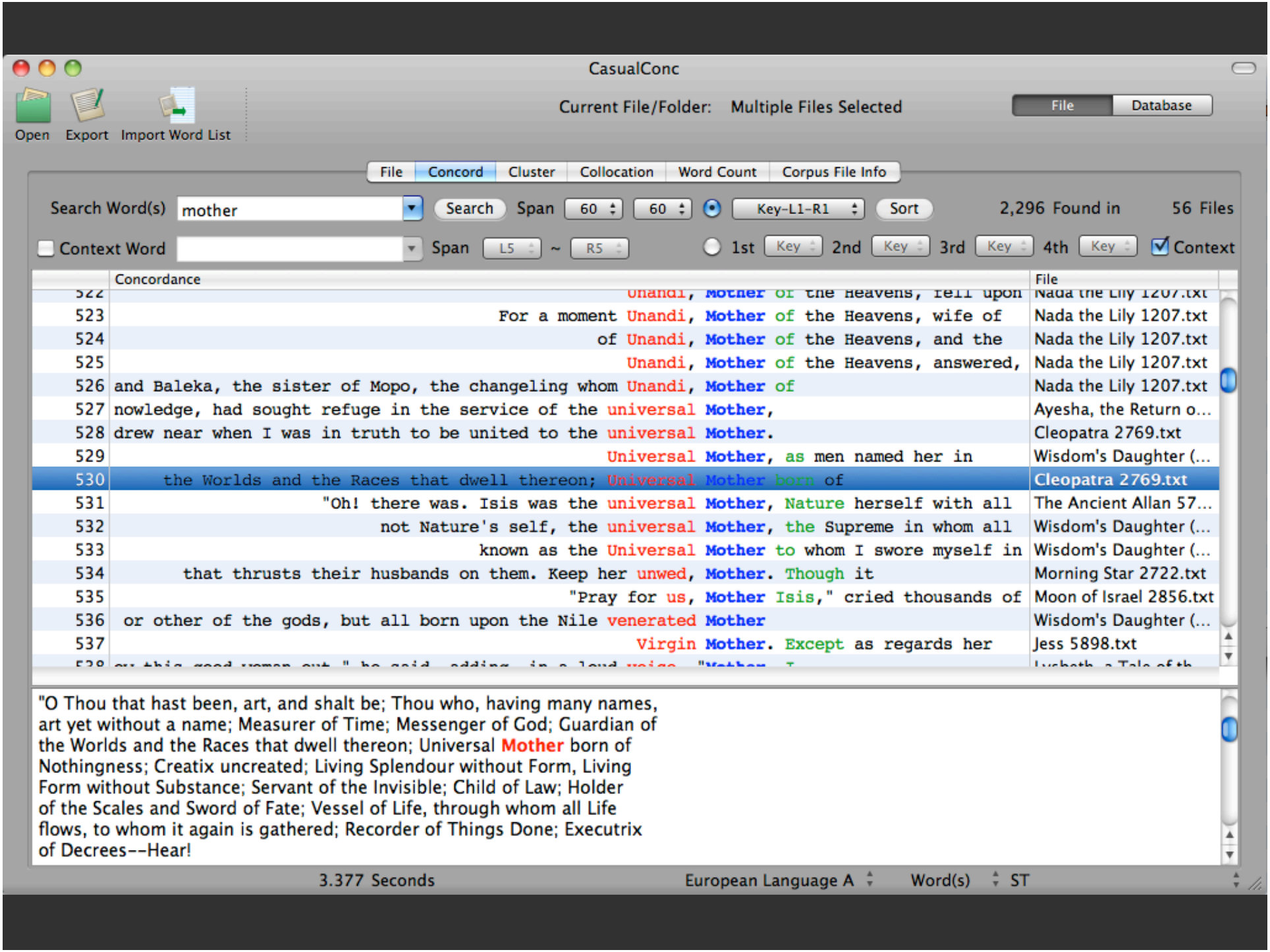
Using a text editor

- You can get a fair distance with a text editor that allows multi-file searches, regular expressions, etc.
  - It's like a little concordancer that's good for close reading
    - jEdit http://www.jedit.org/ 
    - BBedit on Windows

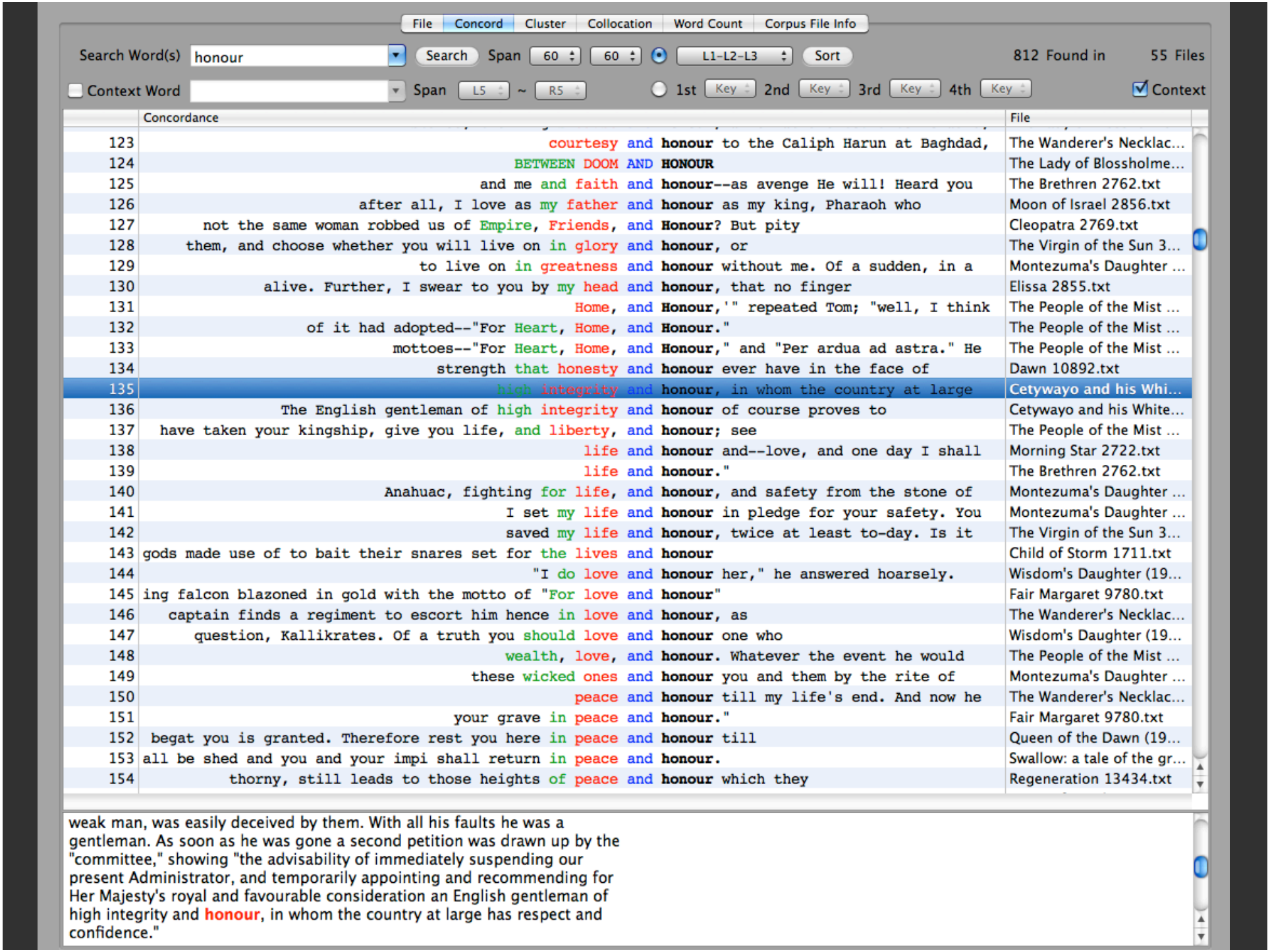

Allan and the Ice Gods (1927) 0200201.txt
Allan and the Ice Gods (1927) 0200201.txt (~/Documents/courses/DigitalHumanitie...
is to shoot at you from the boats until the tide is quite low, and
then to climb along both lines of rock and attack you."
At this moment Moananga came up and was also made to lie down.
"Perhaps," said Wi, "and if so, we had better draw out of the reach of
the little spears."
"That is what they want you to do," answered Laleela, "for then they
will climb along the lines of rock quietly and without hurt. I have
another counsel, if it pleases you to hear it."
"What is it?" said Wi and Moananga together.
"This, Chief: You and all the people know those rocks and where the
deep water holes are between them, since from childhood you have
gathered shellfish there. Now, divide your men into two companies, and
do you command one while Moananga commands the other. Clamber along
those rocks to the right and left with the companies and attack the
Red-Beards on them, for, when they see you coming so boldly, some of
them will get into the boats. The others you must fight and kill; nor
will those in the boats who have bows and arrows be able to shoot much
at you, for fear lest they should hit their own people. Do this, and
swiftly."
"Those are good words," said Wi. "Moananga, do you take the left line
of rocks with half the men, and I will take the right with the rest.
And, Laleela, I bid you remain here, or fly."
"Yes, I will remain here," said Laleela, rather faintly and turning on
her face, so that none should see the stain of blood soaking through
her blue robe. Yet, as they went, she cried after them:
"Bid your people take stones, Wi and Moananga, that they may cast them
into the boats and break their bottoms."
Coming to the men of the tribe who stood there in knots looking very
wretched and afraid, most of them, as they stared at the hairy Red-
Beards upon the rocks and in their boats, Wi addressed them in a few
hard words, saying:
"Yonder Red-Beards come from I know not whence. They are starving,
which will make them very brave, and they mean to kill us, every one,
and to take first our food and then our women, if they can find them;
also perhaps to eat the children. Now, we count as many heads as they
do, perhaps more, and it will be a great shame to us if we allow
ourselves to be conquered, our old people butchered, our women taken,
and our children eaten by these Red Wanderers. Is it not so?"
To this question the crowd answered that it was, yet without
eagerness, for the eyes of most of them were turned toward the woods,
whither the women had gone. Then Moananga said:
7279,69 (370326/519864)
(text,none,MacRoman)
Search And Replace
Search for:
blood
Find
Replace
Replace with:
Text
Return value of a BeanShell snippet
Replace & Find
Replace All
Close
Search in:
Selection
Current buffer
All buffers
Directory:
Settings:
Keep dialog
Ignore case
Regular expressions
HyperSearch
Direction:
Backward
Forward
Auto wrap
Filter:
*
Synchronize
Directory:
talHumanities/DH2011-Manning/RiderHaggard
Choose...
Search subdirectories
Skip hiddens/backups
Skip binary files
HyperSearch Results
Results for "blood":
blood (3,172 occurrences in 60 files)
/Users/manning/Documents/courses/DigitalHumanities/DH2011-Manning/Ri
321: have rung with it. As it was, remembering the fiery southern blood, I
970: no pale-blooded monk, no mere shadow of a man, but one to whose ears
995: populace, baulked of a blood-feast, turning upon their prophet
1030: wolves clawed down the doors, snarling for his blood.
1071: watched and shouted while the mighty martyr whose blood was indeed a
1304: grave of a martyr--the pigment was his blood. Traditions cling long
2023: motley collection of various blood and colour, each of them bearing a
2749: for so much of the Mycenian blood as may remain in their veins. Still
3163: Certainly his blood must have been noble and his place high. Yet his
3644: considerable time, the disorder lurking in the blood and
3665: officials, and children of northern blood seem to flourish there.

Traditional Concordancers

- WordSmith Tools Commercial; Windows
  - http://www.lexically.net/wordsmith/
- Concordance Commercial; Windows
  - http://www.concordancesoftware.co.uk/
- AntConc Free; Windows, Mac OS X (only under X11); Linux
  - http://www.antlab.sci.waseda.ac.jp/antconc_index.html
- CasualConc Free; Mac OS X
  - http://sites.google.com/site/casualconc/
    - by Yasu Imao


CasualConc
Open
Export
Import Word List
Current File/Folder: Multiple Files Selected
File
Database
File
Concord
Cluster
Collocation
Word Count
Corpus File Info
Search Word(s)
mother
Search
Span
60
60
Key-L1-R1
Sort
2,296 Found in 56 Files
Context Word
Span
L5
~
R5
1st
Key
2nd
Key
3rd
Key
4th
Key
Context
Concordance
File
523 For a moment Unandi, Mother of the Heavens, wife of Nada the Lily 1207.txt
524 of Unandi, Mother of the Heavens, and the Nada the Lily 1207.txt
525 Unandi, Mother of the Heavens, answered, Nada the Lily 1207.txt
526 and Baleka, the sister of Mopo, the changeling whom Unandi, Mother of Nada the Lily 1207.txt
527 nowledge, had sought refuge in the service of the universal Mother, Ayesha, the Return o...
528 drew near when I was in truth to be united to the universal Mother. Cleopatra 2769.txt
529 Universal Mother, as men named her in Wisdom's Daughter (...
530 the Worlds and the Races that dwell thereon; Universal Mother born of Cleopatra 2769.txt
531 "Oh! there was. Isis was the universal Mother, Nature herself with all The Ancient Allan 57...
532 not Nature's self, the universal Mother, the Supreme in whom all Wisdom's Daughter (...
533 known as the Universal Mother to whom I swore myself in Wisdom's Daughter (...
534 that thrusts their husbands on them. Keep her unwed, Mother. Though it Morning Star 2722.txt
535 "Pray for us, Mother Isis," cried thousands of Moon of Israel 2856.txt
536 or other of the gods, but all born upon the Nile venerated Mother Wisdom's Daughter (...
537 Virgin Mother. Except as regards her Jess 5898.txt
"O Thou that hast been, art, and shalt be; Thou who, having many names, art yet without a name; Measurer of Time; Messenger of God; Guardian of the Worlds and the Races that dwell thereon; Universal Mother born of Nothingness; Creatix uncreated; Living Splendour without Form, Living Form without Substance; Servant of the Invisible; Child of Law; Holder of the Scales and Sword of Fate; Vessel of Life, through whom all Life flows, to whom it again is gathered; Recorder of Things Done; Executrix of Decrees--Hear!
3.377 Seconds
European Language A
Word(s)
ST


CasualConc
Open
Export
Import Word List
Current File/Folder: Multiple Files Selected
File
Database
File
Concord
Cluster
Collocation
Word Count
Corpus File Info
Search Word(s)
black
Search
Span
L5
~
R5
Visual
2,671 Found in 58 Files
Frequency
Sort
L5 L4 L3 L2 L1 Key R1 R2 R3 R4 R5
the the the of the black and and the the the
and and and and a one of and and and
of of of the of kendah who a of of
a a a in and meg the was to a
in to was a his eyes which he a to
he was to with that as that to it in
to he with to great bow in with in his
was his i that was kloof i i i as
i that in his with hair was of that he
his it it was in horse with his was i
it in on white this with he that with that
with with that on her people a in white with
that for he by is heart white is had was
not i him from as man to it he is
white you as is long water as who you at
by is white it two men for as her it
for white you for some maned s had his on
is him there as were marble it you she for
on on were but are stone were white as not
as as his her coal rock at her him you
at black not like little windows from they black her
had had out into or dog we on heart had
like one who against these hearted has which at were
my s but at those beard lion she have from
Collocation
Cooccurrence
7.431 Seconds
European Language A
Word(s)
ST

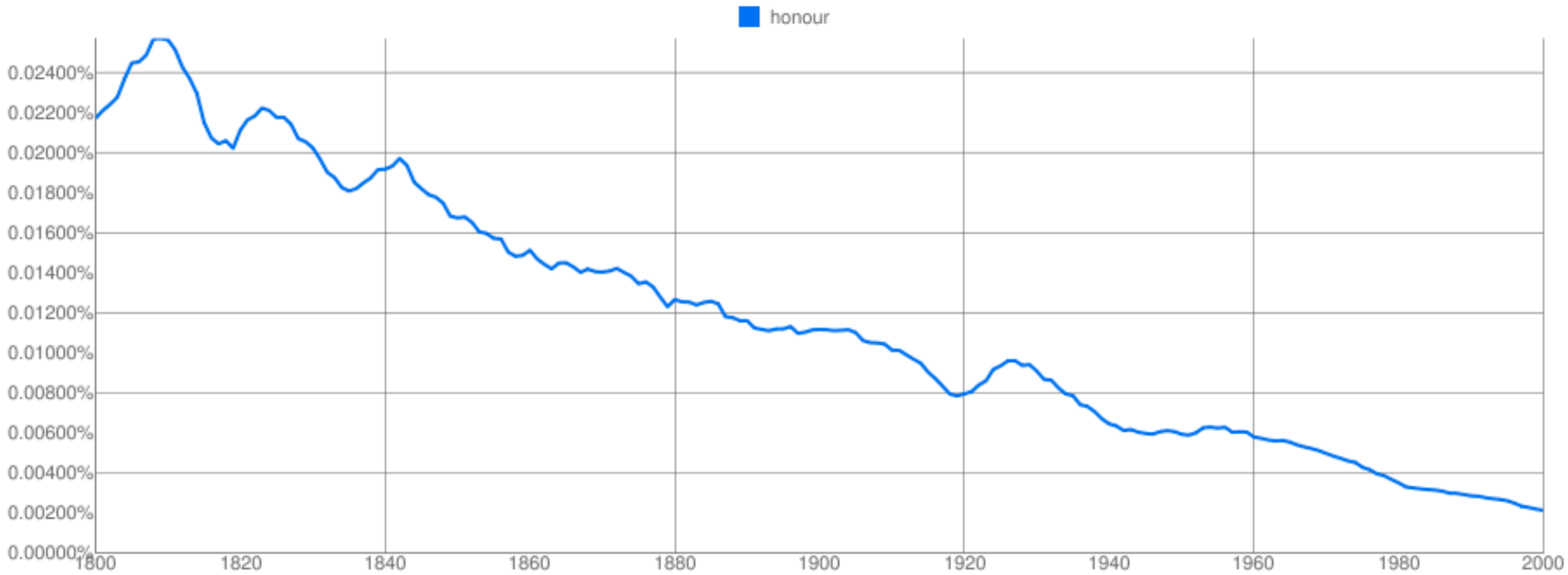

File
Concord
Cluster
Collocation
Word Count
Corpus File Info
Search Word(s)
honour
Search
Span
60
60
L1-L2-L3
Sort
812 Found in 55 Files
Context Word
Span
L5
~
R5
1st
Key
2nd
Key
3rd
Key
4th
Key
Context
Concordance
File
123 courtesy and honour to the Caliph Harun at Baghdad, The Wanderer's Necklac...
124 BETWEEN DOOM AND HONOUR The Lady of Blossholme...
125 and me and faith and honour--as avenge He will! Heard you The Brethren 2762.txt
126 after all, I love as my father and honour as my king, Pharaoh who Moon of Israel 2856.txt
127 not the same woman robbed us of Empire, Friends, and Honour? But pity Cleopatra 2769.txt
128 them, and choose whether you will live on in glory and honour, or The Virgin of the Sun 3...
129 to live on in greatness and honour without me. Of a sudden, in a Montezuma's Daughter ...
130 alive. Further, I swear to you by my head and honour, that no finger Elissa 2855.txt
131 Home, and Honour,'" repeated Tom; "well, I think The People of the Mist ...
132 of it had adopted--"For Heart, Home, and Honour." The People of the Mist ...
133 mottoes--"For Heart, Home, and Honour," and "Per ardua ad astra." He The People of the Mist ...
134 strength that honesty and honour ever have in the face of Dawn 10892.txt
135 high integrity and honour, in whom the country at large Cetywayo and his Whi...
136 The English gentleman of high integrity and honour of course proves to Cetywayo and his White...
137 have taken your kingship, give you life, and liberty, and honour; see The People of the Mist ...
138 life and honour and--love, and one day I shall Morning Star 2722.txt
139 life and honour." The Brethren 2762.txt
140 Anahuac, fighting for life, and honour, and safety from the stone of Montezuma's Daughter ...
141 I set my life and honour in pledge for your safety. You Montezuma's Daughter ...
142 saved my life and honour, twice at least to-day. Is it The Virgin of the Sun 3...
143 gods made use of to bait their snares set for the lives and honour Child of Storm 1711.txt
144 "I do love and honour her," he answered hoarsely. Wisdom's Daughter (19...
145 ing falcon blazoned in gold with the motto of "For love and honour" Fair Margaret 9780.txt
146 captain finds a regiment to escort him hence in love and honour, as The Wanderer's Necklac...
147 question, Kallikrates. Of a truth you should love and honour one who Wisdom's Daughter (19...
148 wealth, love, and honour. Whatever the event he would The People of the Mist ...
149 these wicked ones and honour you and them by the rite of Montezuma's Daughter ...
150 peace and honour till my life's end. And now he The Wanderer's Necklac...
151 your grave in peace and honour." Fair Margaret 9780.txt
152 begat you is granted. Therefore rest you here in peace and honour till Queen of the Dawn (19...
153 all be shed and you and your impi shall return in peace and honour. Swallow: a tale of the gr...
154 thorny, still leads to those heights of peace and honour which they Regeneration 13434.txt
weak man, was easily deceived by them. With all his faults he was a gentleman. As soon as he was gone a second petition was drawn up by the "committee," showing "the advisability of immediately suspending our present Administrator, and temporarily appointing and recommending for Her Majesty's royal and favourable consideration an English gentleman of high integrity and honour, in whom the country at large has respect and confidence."

The decline of honour

Graph these **case-sensitive** comma-separated phrases: honour

between 1800 and 2000 from the corpus British English with smoothing of 3.

Search lots of books



Search in Google Books:

1800 - 1809	1810 - 1820	1821 - 1831	1832 - 1949	1950 - 2000	honour (British English)

5. NLP Frameworks and Tools

The Big 3 NLP Frameworks

- GATE – General Architecture for Text Engineering (U. Sheffield)
  - http://gate.ac.uk/
  - Java, quite well maintained (now)
  - Includes tons of components
- UIMA – Unstructured Information Management Architecture. Originally IBM; now Apache project
  - http://uima.apache.org/
  - Professional, scalable, etc.
  - But, unless you're comfortable with Xml, Eclipse, Java or C++, etc., I think it's a non-starter
- NLTK – Natural Language Toolkit (started by Steven Bird)
  - http://www.nltk.org/
  - Big community; large Python package; corpora and *books* about it
  - But it's code modules and API, no GUI or command-line tools
  - Like R for NLP. But, hey, R's becoming very successful….

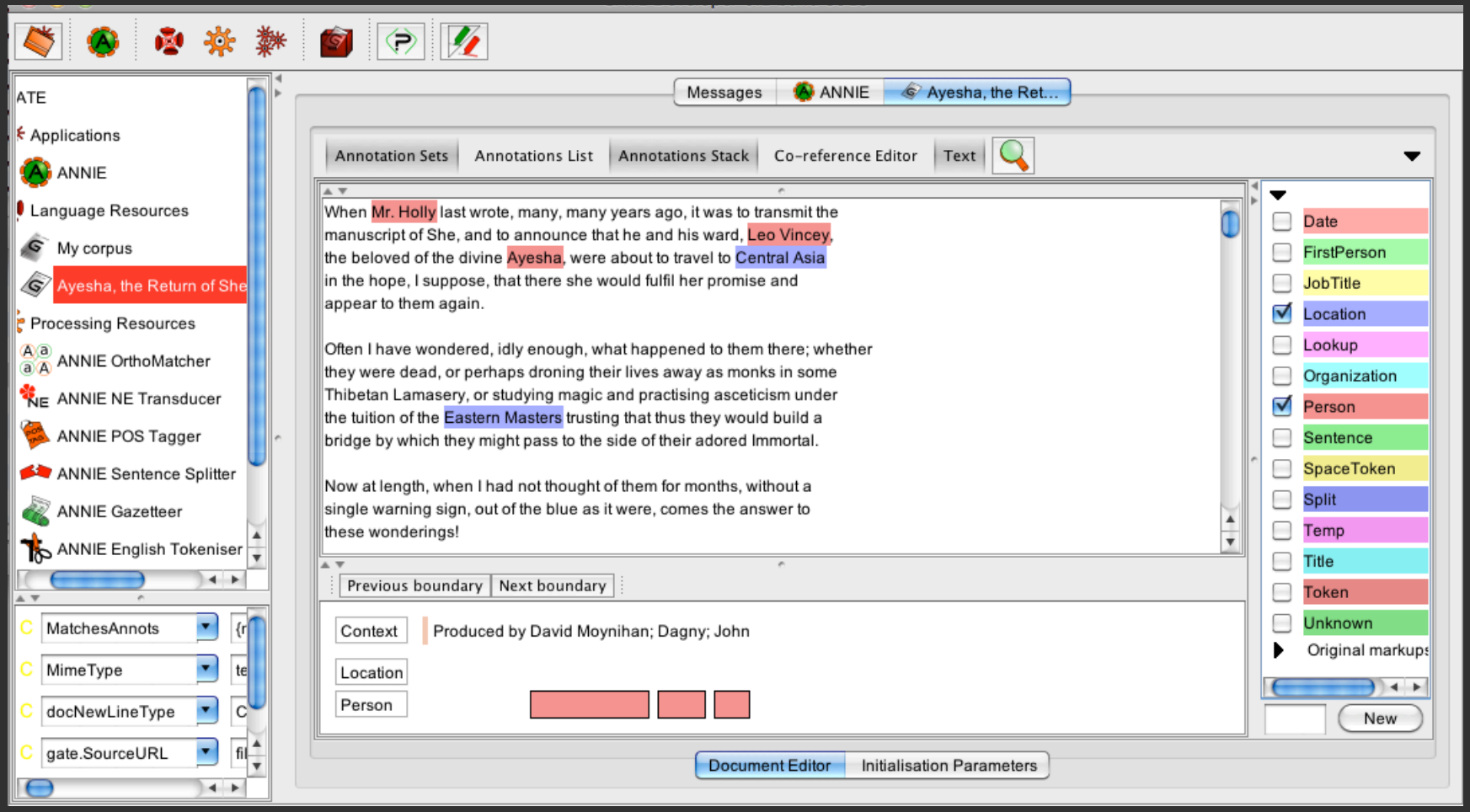
The main NLP Packages

- NLTK Python
  - http://www.nltk.org/
- OpenNLP
  - http://incubator.apache.org/opennlp/
- Stanford NLP
  - http://nlp.stanford.edu/software/
- LingPipe
  - http://alias-i.com/lingpipe/
- More one-off packages than I can fit on this slide
  - http://nlp.stanford.edu/links/statnlp.html

NLP tools: Rules of thumb for 2011

1. Unless you're unlucky, the tool you want to use will work with Unicode (at least BMP), so most any characters are okay
2. Unless you're lucky, the tool you want to use will work only on completely plain text, or *extremely* simple XML-style mark-up (e.g., <s> … </s> around sentences, recognized by regexp)
3. By default, you should assume that any tool for English was trained on American newswire

GATE


ATE
Applications
ANNIE
Language Resources
My corpus
Ayesha, the Return of She
Processing Resources
ANNIE OrthoMatcher
ANNIE NE Transducer
ANNIE POS Tagger
ANNIE Sentence Splitter
ANNIE Gazetteer
ANNIE English Tokeniser
MatchesAnnots
MimeType
docNewLineType
gate.SourceURL
Messages
ANNIE
Ayesha, the Ret...
Annotation Sets
Annotations List
Annotations Stack
Co-reference Editor
Text
When Mr. Holly last wrote, many, many years ago, it was to transmit the manuscript of She, and to announce that he and his ward, Leo Vincey, the beloved of the divine Ayesha, were about to travel to Central Asia in the hope, I suppose, that there she would fulfil her promise and appear to them again.
Often I have wondered, idly enough, what happened to them there; whether they were dead, or perhaps droning their lives away as monks in some Thibetan Lamasery, or studying magic and practising asceticism under the tuition of the Eastern Masters trusting that thus they would build a bridge by which they might pass to the side of their adored Immortal.
Now at length, when I had not thought of them for months, without a single warning sign, out of the blue as it were, comes the answer to these wonderings!
Previous boundary
Next boundary
Context
Produced by David Moynihan; Dagny; John
Location
Person
Date
FirstPerson
JobTitle
Location
Lookup
Organization
Person
Sentence
SpaceToken
Split
Temp
Title
Token
Unknown
Original markups
New
Document Editor
Initialisation Parameters

Rule-based NLP and Statistical/ Machine Learning NLP

- Most work on NLP in the 1960s, 70s and 80s was with hand-built grammars and morphological analyzers (finite state transducers), etc.
  - ANNIE in GATE is still in this space
- Most academic research work in NLP in the 1990s and 2000s use probabilistic or more generally machine learning methods ("Statistical NLP")
  - The Stanford NLP tools and MorphAdorner, which we will come to soon, are in this space

Rule-based NLP and Statistical/ Machine Learning NLP

- Hand-built grammars are fine for tasks in a closed space which do not involve reasoning about contexts
  - E.g., finding the possible morphological parses of a word
- In the old days they worked *really* badly on "real text"
  - They were always insufficiently tolerant of the variability of real language
  - But, built with modern, empirical approaches, they can do *reasonably* well
    - ANNIE is an example of this

Rule-based NLP and Statistical/ Machine Learning NLP

- In Statistical NLP:
  - You gather corpus data, and usually hand-annotate it with the kind of information you want to provide, such as part-of-speech
  - You then train (or "learn") a model that learns to try to predict annotations based on features of words and their contexts via numeric feature weights
  - You then apply the trained model to new text
- This tends to work much better on real text
  - It more flexibly handles contextual and other evidence
- But the technology is still far from perfect, it requires annotated data, and degrades (sometimes very badly) when there are mismatches between the training data and the runtime data

How much hardware do you need?

- NLP software often needs plenty of RAM (especially) and processing power
- But these days we have *really powerful* laptops!
- Some of the software I show you could run on a machine with 256 MB of RAM (e.g., Stanford Parser), but much of it requires more
- Stanford CoreNLP requires a machine with 4GB of RAM
- I ran everything in this tutorial on the laptop I'm presenting on … 4GB RAM, 2.8 GHz Core 2 Duo
- But it wasn't always pleasant writing the slides while software was running….

How much hardware do you need?

- Why do you need more hardware?
  - More speed
    - It took me 95 minutes to run *Ayesha, the Return of She* through Stanford CoreNLP on my laptop….
  - More scale
    - You'd like to be able to analyze 1 million books
- Order of magnitude rules of thumb:
  - POS tagging, NER, etc: 5–10,000 words/second
  - Parsing: 1–10 sentences per second

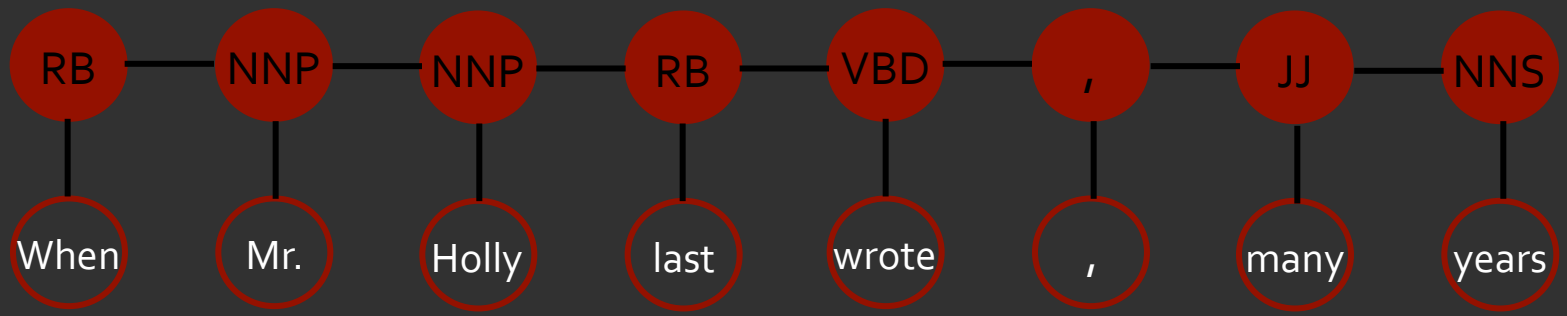
How much hardware do you need?

- Luckily, most of our problems are *trivially parallelizable*
  - Each book/chapter can be run separately, perhaps on a separate machine
- What do we actually use?
  - We do most of our computing on rack mounted Linux servers
    - Currently 4 x quad core Xeon processors with 24 GB of RAM seem about the sweet spot
    - About $3500 per machine … not like the old days

6. Part-of-speech Tagging

Part-of-Speech Tagging

- Part-of-speech tagging is normally done by a *sequence model* (acronyms: HMM, CRM, MEMM/CMM)
  - A POS tag is to be placed above each word
  - The model considers a local context of possible previous and following POS tags, the current word, neighboring words, and features of them (capitalized?, ends in *-ing*?)
  - Each such *feature* has a *weight*, and the evidence is combined, and the most likely sequence of tags (according to the model) is chosen



Stanford POS tagger

http://nlp.stanford.edu/software/tagger.shtml

```
$ java -mx1g -cp ../Software/stanford-postagger-full-2011-06-19/stanford-postagger.jar edu.stanford.nlp.tagger.maxent.MaxentTagger -model ../Software/stanford-postagger-full-2011-06-19/models/left3words-distsim-wsj-0-18.tagger -outputFormat tsv -tokenizerOptions untokenizable=allKeep -textFile She\ 3155.txt > She\ 3155.tsv

Loading default properties from trained tagger ../Software/stanford-postagger-full-2011-06-19/models/left3words-distsim-wsj-0-18.tagger

Reading POS tagger model from ../Software/stanford-postagger-full-2011-06-19/models/left3words-distsim-wsj-0-18.tagger ... done [2.2 sec].

Jun 15, 2011 8:17:15 PM edu.stanford.nlp.process.PTBLexer next

WARNING: Untokenizable: ? (U+1FBD, decimal: 8125)

Tagged 132377 words at 5559.72 words per second.
```

Greek stand-alone Koronis character (a little obscure?)

Stanford POS tagger

- For the second time you do it…

```
$ alias stanfordtag "java -mx1g -cp /Users/manning/Software/
stanford-postagger-full-2011-06-19/stanford-postagger.jar
edu.stanford.nlp.tagger.maxent.MaxentTagger -model /Users/
manning/Software/stanford-postagger-full-2011-06-19/models/
left3words-distsim-wsj-0-18.tagger -outputFormat tsv -
tokenizerOptions untokenizable=allKeep -textFile"

$ stanfordtag RiderHaggard/King\ Solomon\'s\ Mines\ 2166.txt >
tagged/King\ Solomon\'s\ Mines\ 2166.tsv

Reading POS tagger model from /Users/manning/Software/
stanford-postagger-full-2011-06-19/models/left3words-distsim-
wsj-0-18.tagger … done [2.1 sec].

Tagged 98178 words at 9807.99 words per second.
```

MorphAdorner

http://morphadorner.northwestern.edu/

- MorphAdorner is a set of NLP tools developed at Northwestern by Martin Mueller and colleagues *specifically for English language fiction*, over a long historical period from EME onwards
  - lemmatizer, named entity recognizer, POS tagger, spelling standardizer, etc.
- Aims to deal with variation in word breaking and spelling over this period
- Includes its own POS tag set: NUPOS

MorphAdorner

```
$ ./adornplaintext temp temp/3155.txt
2011-06-15 20:30:52,111 INFO  - MorphAdorner version 1.0
2011-06-15 20:30:52,111 INFO  - Initializing, please wait...
2011-06-15 20:30:52,318 INFO  - Using Trigram tagger.
2011-06-15 20:30:52,319 INFO  - Using I retagger.
2011-06-15 20:30:53,578 INFO  - Loaded word lexicon with 151,922 entries in 2 seconds.
2011-06-15 20:30:55,920 INFO  - Loaded suffix lexicon with 214,503 entries in 3 seconds.
2011-06-15 20:30:57,927 INFO  - Loaded transition matrix in 3 seconds.
2011-06-15 20:30:58,137 INFO  - Loaded 162,248 standard spellings in 1 second.
2011-06-15 20:30:58,697 INFO  - Loaded 5,434 alternative spellings in 1 second.
2011-06-15 20:30:58,703 INFO  - Loaded 349 more alternative spellings in 14 word classes in 1 second.
2011-06-15 20:30:58,713 INFO  - Loaded 0 names into name standardizer in < 1 second.
2011-06-15 20:30:58,779 INFO  - 1 file to process.
2011-06-15 20:30:58,789 INFO  - Before processing input texts: Free memory: 105,741,696, total memory: 480,694,272
2011-06-15 20:30:58,789 INFO  - Processing file 'temp/3155.txt' .
2011-06-15 20:30:58,789 INFO  - Adorning temp/3155.txt with parts of speech.
2011-06-15 20:30:58,832 INFO  - Loaded text from temp/3155.txt in 1 second.
2011-06-15 20:31:01,498 INFO  -    Extracted 131,875 words in 4,556 sentences in 3 seconds.
2011-06-15 20:31:03,860 INFO  -      lines: 1,000; words: 27,756
2011-06-15 20:31:04,364 INFO  -      lines: 2,000; words: 58,728
2011-06-15 20:31:04,676 INFO  -      lines: 3,000; words: 84,735
2011-06-15 20:31:04,990 INFO  -      lines: 4,000; words: 115,396
2011-06-15 20:31:05,152 INFO  -      lines: 4,556; words: 131,875
2011-06-15 20:31:05,152 INFO  -    Part of speech adornment completed in 4 seconds. 36,100 words adorned per second.
2011-06-15 20:31:05,152 INFO  -    Generating other adornments.
2011-06-15 20:31:13,840 INFO  -    Adornments written to temp/3155-005.txt in 9 seconds.
2011-06-15 20:31:13,840 INFO  - All files adorned in 16 seconds.
```

Ah, the old days!

```
$ ./adornplaintext temp temp/Hunter\ Quartermain.txt
2011-06-15 17:18:15,551 INFO  - MorphAdorner version 1.0
2011-06-15 17:18:15,552 INFO  - Initializing, please wait...
2011-06-15 17:18:15,730 INFO  - Using Trigram tagger.
2011-06-15 17:18:15,731 INFO  - Using I retagger.
2011-06-15 17:18:16,972 INFO  - Loaded word lexicon with 151,922 entries in 2
seconds.
2011-06-15 17:18:18,684 INFO  - Loaded suffix lexicon with 214,503 entries in 2
seconds.
2011-06-15 17:18:20,662 INFO  - Loaded transition matrix in 2 seconds.
2011-06-15 17:18:20,887 INFO  - Loaded 162,248 standard spellings in 1 second.
2011-06-15 17:18:21,300 INFO  - Loaded 5,434 alternative spellings in 1 second.
2011-06-15 17:18:21,303 INFO  - Loaded 349 more alternative spellings in 14 word
classes in 1 second.
2011-06-15 17:18:21,312 INFO  - Loaded 0 names into name standardizer in 1 second.
2011-06-15 17:18:21,381 INFO  - No files found to process.
```

- But it works better if you make sure the filename has no spaces in it ☺

Comparing taggers: Penn Treebank vs. NUPOS

Holly	NNP	Holly	n1
,	,	,	,
if	IN	if	cs
you	PRP	you	pn22
will	MD	will	vmb
accept	VB	accept	vvi
the	DT	the	dt
trust	NN	trust	n1
,	,	,	,
I	PRP	I	pns11
am	VBP	am	vbm

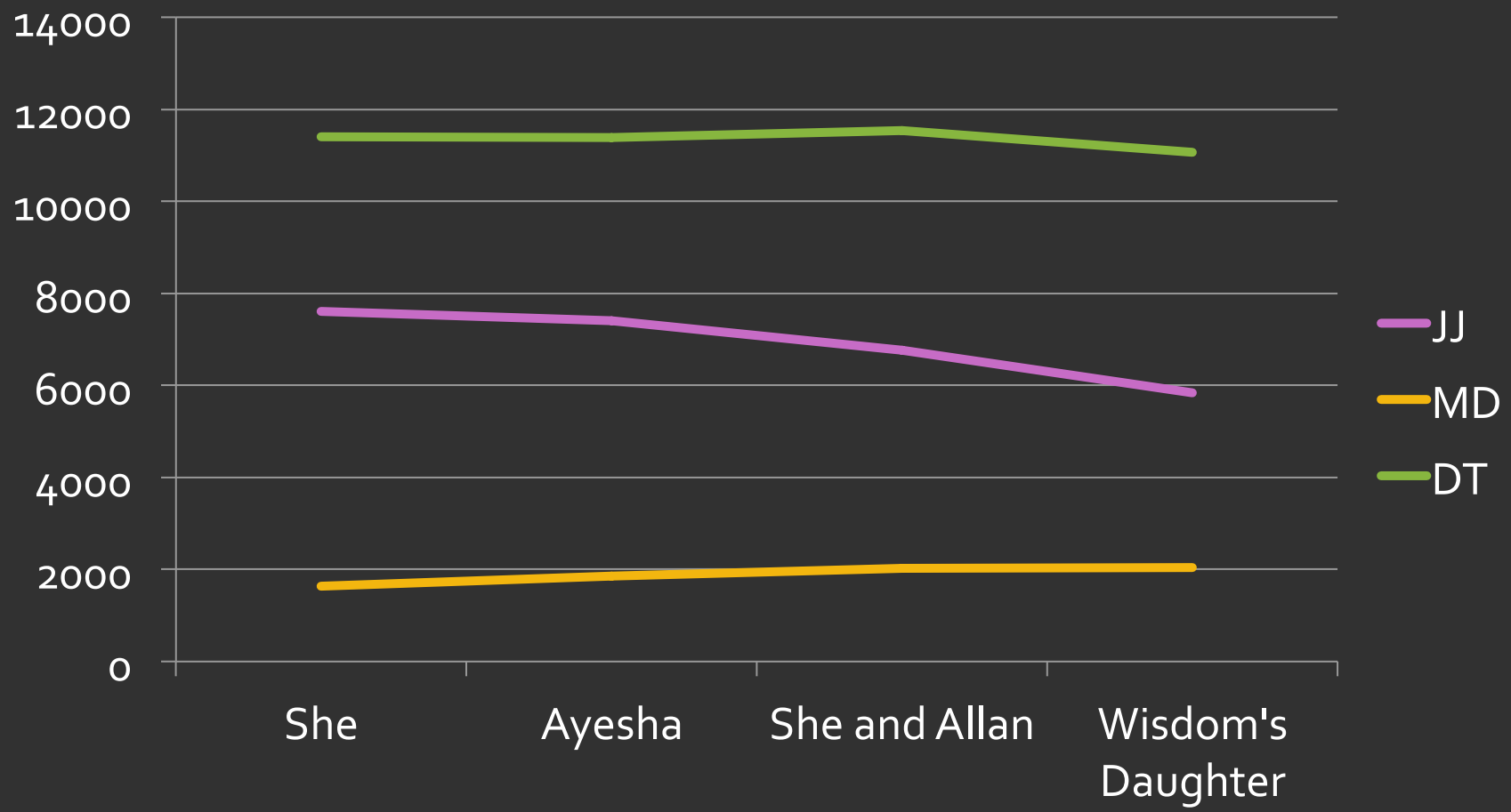
going	VBG	going	vvg
to	TO	to	pc-acp
leave	VB	leave	vvi
you	PRP	you	pn22
that	IN	that	d
boy	NN	boy's	ng1
's	POS		
sole	JJ	sole	j
guardian	NN	guardian	n1
.	.	.	.

Comparing taggers: Penn Treebank vs. NUPOS

Holly	NNP	Holly	n1
,	,	,	,
if	IN	if	cs
you	PRP	you	pn22
will	MD	will	vmb
accept	VB	accept	vvi
the	DT	the	dt
trust	NN	trust	n1
,	,	,	,
I	PRP	I	pns11
am	VBP	am	vbm

going	VBG	going	vvg
to	TO	to	pc-acp
leave	VB	leave	vvi
you	PRP	you	pn22
that	IN	that	d
boy	NN	boy's	ng1
's	POS		
sole	JJ	sole	j
guardian	NN	guardian	n1
.	.	.	.

Stylistic factors from POS



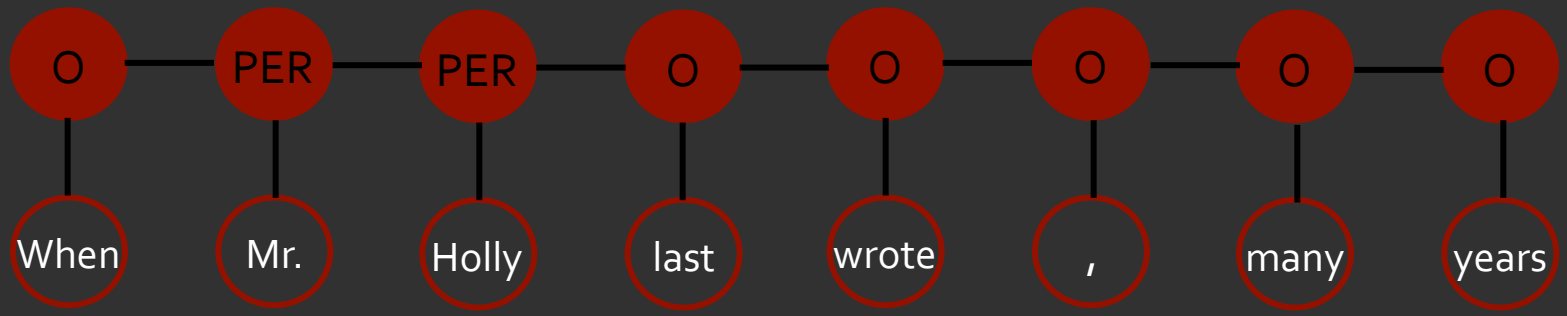
7. Named Entity Recognition (NER)

Named Entity Recognition – "the Chad problem"

Germany's representative to the European Union's veterinary committee Werner Zwingman said on Wednesday consumers should …

IL-2 gene expression and NF-kappa B activation through CD28 requires reactive oxygen production by 5-lipoxygenase.

Conditional Random Fields (CRFs)



- We again use a sequence model – different problem, but same technology
  - Indeed, sequence models are used for lots of tasks that can be construed as labeling tasks that require only local context (to do quite well)
- There is a background label – O – and labels for each class
- Entities are both *segmented* and *categorized*

Stanford NER Features

- Word features: current word, previous word, next word, a word is anywhere in a +/– 4 word window
- Orthographic features:
  - Jenny → Xxxx
  - IL-2 → XX-#
- Prefixes and Suffixes:
  - Jenny → <J, <Je, <Jen, …, nny>, ny>, y>
- Label sequences
- Lots of feature conjunctions

Stanford NER

http://nlp.stanford.edu/software/CRF-NER.shtml

```
$ java -mx500m -Dfile.encoding=utf-8 -cp Software/stanford-ner-2011-06-19/stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier Software/stanford-ner-2011-06-19/classifiers/all.3class.distsim.crf.ser.gz -textFile RiderHaggard/She\ 3155.txt > ner/She\ 3155.ner
```

For thou shalt rule this <LOCATION>England</LOCATION>----”

"But we have a queen already," broke in <LOCATION>Leo</LOCATION>, hastily.

"It is naught, it is naught," said <PERSON>Ayesha</PERSON>; "she can be overthrown.”

At this we both broke out into an exclamation of dismay, and explained that we should as soon think of overthrowing ourselves.

"But here is a strange thing," said <PERSON>Ayesha</PERSON>, in astonishment; "a queen whom her people love! Surely the world must have changed since I dwelt in <LOCATION>Kôr</LOCATION>."


LELAND STANFORD JUNIOR UNIVERSITY
DIE LUFT DER FREIHEIT WEHT

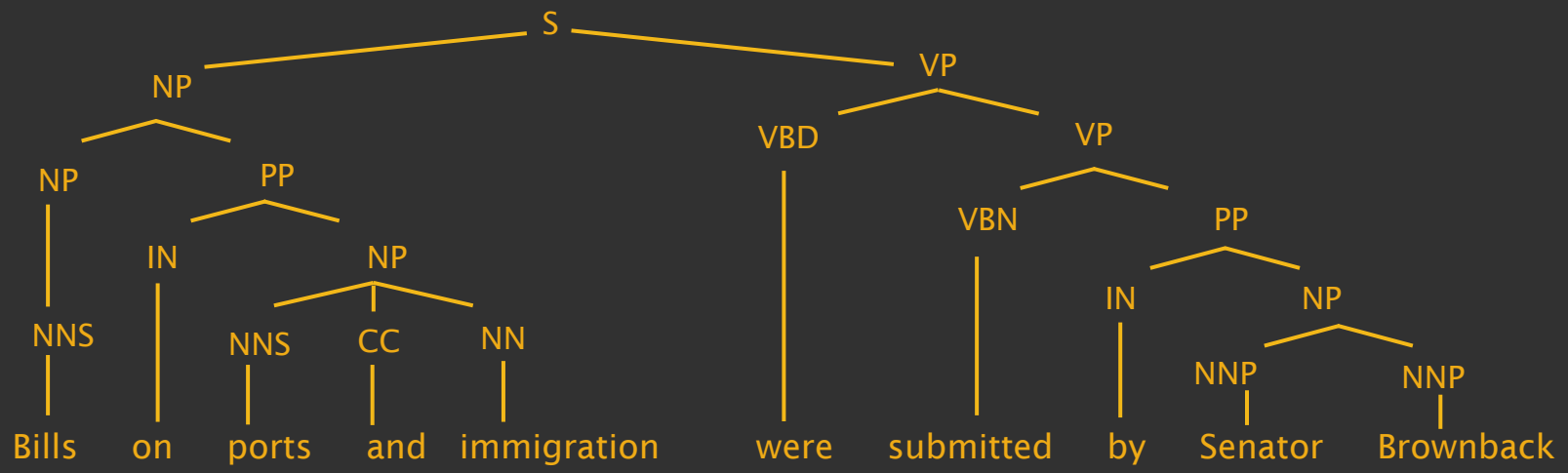
8. Parsing

Statistical parsing

- One of the big successes of 1990s statistical NLP was the development of statistical parsers
- These are trained from hand-parsed sentences ("treebanks"), and know statistics about phrase structure and word relationships, and use them to assign the most likely structure to a new sentence
- They will return a sentence parse for *any* sequence of words. And it will usually be *mostly* right
- There are many opportunities for exploiting this richer level of analysis, which have only been partly realized.

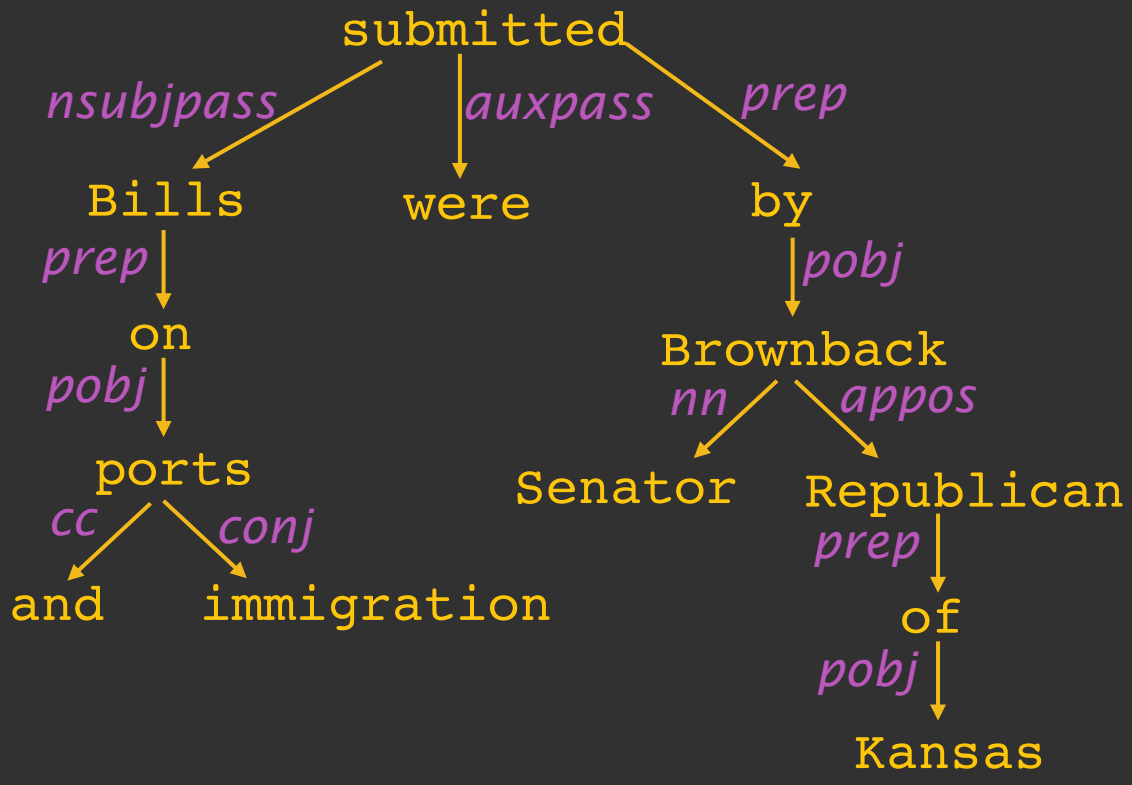
Phrase structure Parsing

- Phrase structure representations have dominated American linguistics since the 1930s
- They focus on showing words that go together to form natural groups (constituents) that behave alike
- They are good for showing and querying details of sentence structure and embedding



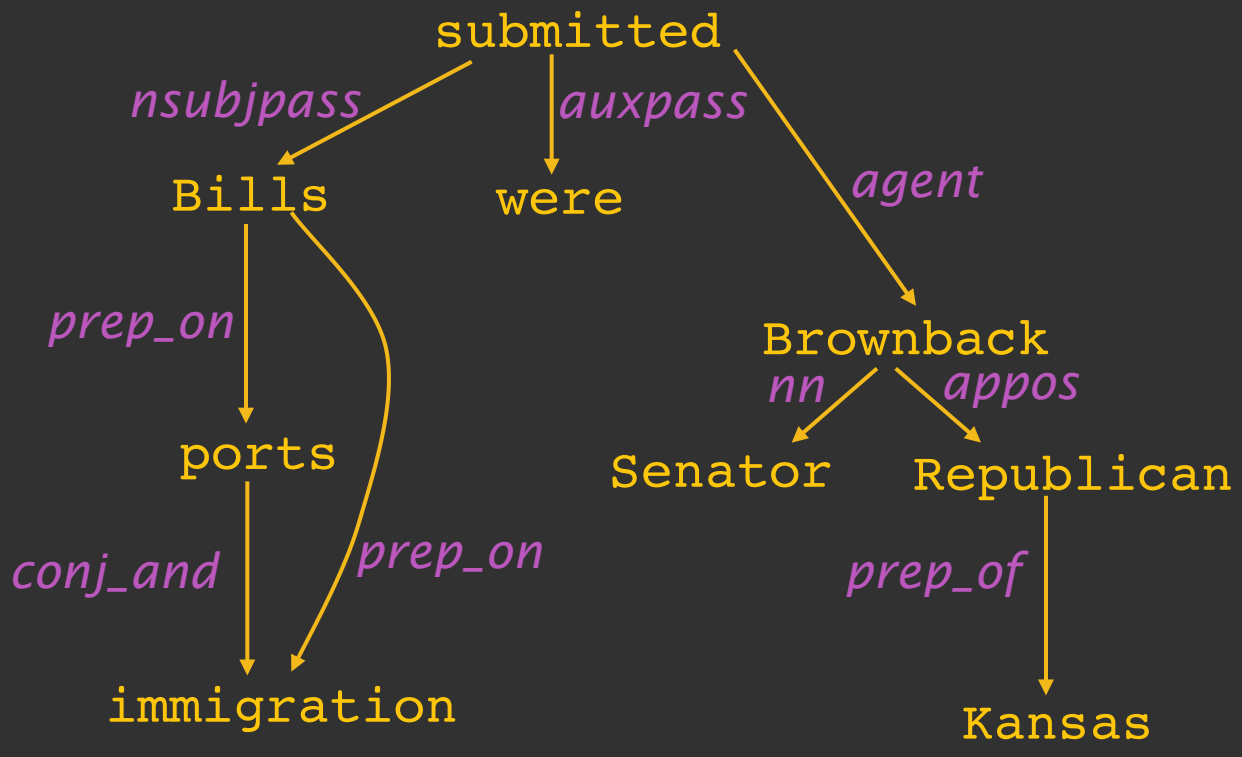
Dependency parsing

- A dependency parse shows which words in a sentence modify other words
- The key notion are governors with dependents
- Widespread use: Pāṇini, early Arabic grammarians, diagramming sentences, …



Stanford Dependencies

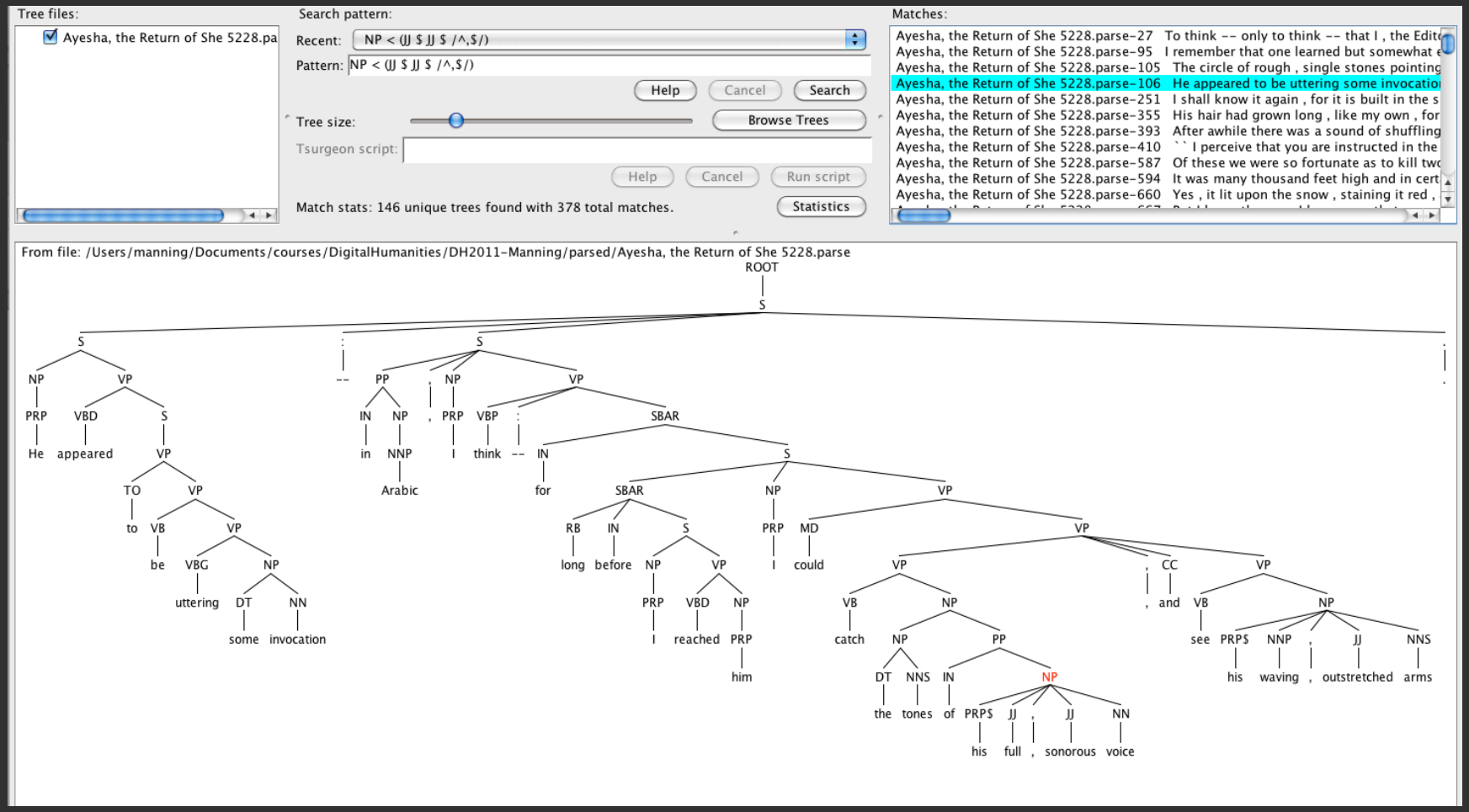
- SD is a particular dependency representation designed for easy extraction of meaning relationships [de Marneffe & Manning, 2008]
  - It's basic form in the last slide has each word as is
  - A "collapsed" form focuses on relations between main words



Statistical Parsers

- There are now *many* good statistical parsers that are freely downloadable
  - Constituency parsers
    - Collins/Bikel Parser
    - Berkeley Parser
    - BLLIP Parser = Charniak/Johnson Parser
  - Dependency parsers
    - MaltParser
    - MST Parser
- But I'll show the Stanford Parser ☺

Tregex/Tgrep2 – Tools for searching over syntax



dreadful things

She

amod(day-18, dreadful-17)
amod(day-45, dreadful-44)
amod(feast-33, dreadful-32)
amod(fits-51, dreadful-50)
amod(form-59, dreadful-58)
amod(laugh-9, dreadful-8)
amod(manifestation-9, dreadful-8)
amod(manner-29, dreadful-28)
amod(marshes-17, dreadful-16)
amod(people-12, dreadful-11)
amod(people-46, dreadful-45)
amod(place-16, dreadful-15)
amod(place-6, dreadful-5)
amod(sight-5, dreadful-4)
amod(spot-13, dreadful-12)
amod(thing-41, dreadful-40)
amod(thing-5, dreadful-4)
amod(tragedy-22, dreadful-21)
amod(wilderness-43, dreadful-42)

Ayesha

amod(clouds-5, dreadful-2)
amod(debt-26, dreadful-25)
amod(doom-21, dreadful-20)
amod(fashion-50, dreadful-47)
amod(form-10, dreadful-7)
amod(oath-42, dreadful-41)
amod(road-23, dreadful-22)
amod(silence-5, dreadful-4)
amod(threat-19, dreadful-18)

Making use of dependency structure

J. Engelberg *Costly Information Processing* (AFA, 2009):

- An efficient market should *immediately* incorporate all publicly available information.
- But many studies have shown there is a lag
  - And the lag is greater on Fridays (!)
- An explanation for this is that there is a cost to information processing
- Engelberg tests and shows that "soft" (textual) information takes longer to be absorbed than "hard" (numeric) information … it's higher cost information processing
- But "soft" information has value beyond "hard" information
  - It's especially valuable for predicting further out in time

Evidence from earnings announcements
[Engelberg AFA 2009]

- But how do you use the "soft" information?
- Simply using proportion of "negative" words (from the Harvard General Inquirer lexicon) is a useful predictive feature of future stock behavior

  Although sales remained steady, the firm continues to suffer from rising oil prices.
- "But this [or text categorization] is not enough. In order to refine my analysis, I need to know that the negative sentiment is *about* oil prices."
- He thus turns to use of the typed dependencies representation of the Stanford Parser.
  - Words that negative words relate to are grouped into 1 of 6 categories [5 word lists or "other"]

Evidence from earnings announcements
[Engelberg 2009]

- In a regression model with many standard quantitative predictors…
  - Just the negative word fraction is a significant predictor of 3 day or 80 day post earnings announcement abnormal returns (CAR)
    - Coefficient –0.173, $p < 0.05$ for 80 day CAR
  - Negative sentiment about different things has differential effects
    - Fundamentals: –0.198, $p < 0.01$ for 80 day CAR
    - Future: –0.356, $p < 0.05$ for 80 day CAR
    - Other: –0.023, $p < 0.01$ for 80 day CAR
  - Only some of which analysts pay attention to
    - Analyst forecast-for-quarter-ahead earnings is predicted by negative sentiment on Environment and Other but not Fundamentals or Future!

Syntactic Packaging and Implicit Sentiment
[Greene 2007; Greene and Resnik 2009]

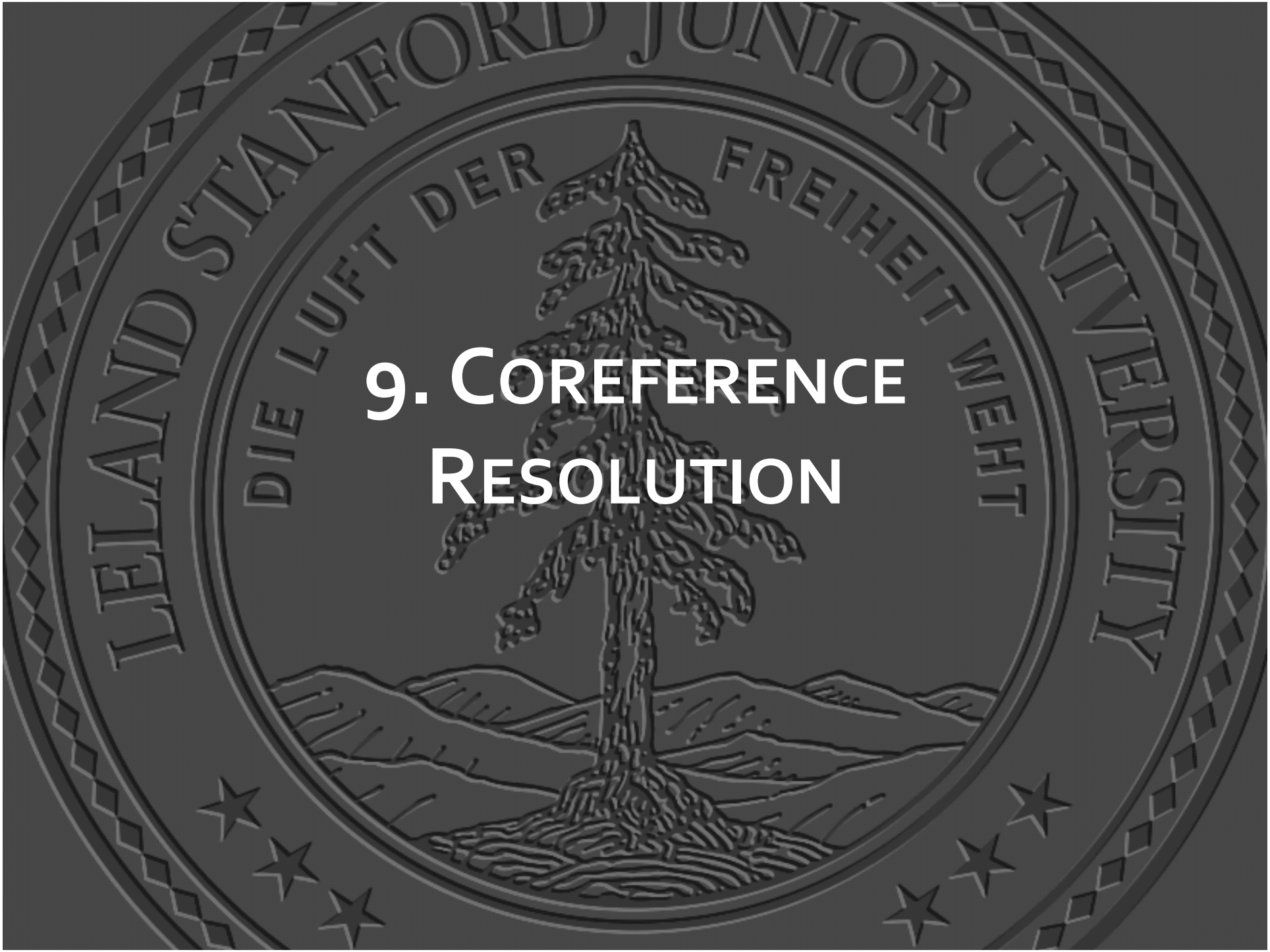
- Positive or negative sentiment can be carried by words (e.g., adjectives), but often it isn't....
  - These sentences differ in sentiment, even though the words aren't so different:
    - A soldier veered his jeep into a crowded market and killed three civilians
    - A soldier's jeep veered into a crowded market and three civilians were killed
- As a measurable version of such issues of linguistic perspective, they define OPUS features
  - For domain relevant terms, OPUS features pair the word with a syntactic Stanford Dependency:
    - killed:DOBJ NSUBJ:soldier killed:NSUBJ

Predicting Opinions of the Death Penalty
[Greene 2007; Greene and Resnik 2009]

- Collected pro- and anti- death penalty texts from websites with manual checking
- Training is cross-validation of training on some pro- and anti- sites and testing on documents from others [can't use site-specific nuances]
- Baseline is word and word bigram features in a support vector machine [SVM = good classifier]

Condition	SVM accuracy
Baseline	72.0%
With OPUS features	88.1%

- 58% error reduction!



9. Coreference Resolution

Coreference resolution

- The goal is to work out which (noun) phrases refer to the same entities in the world
  - Sarah asked her father to look at her. He appreciated that his eldest daughter wanted to speak frankly.
- ≈ anaphora resolution ≈ pronoun resolution ≈ entity resolution

Coreference resolution warnings

- Warning: The tools we have looked at so far work one sentence at a time – or use the whole document but ignore all structure and just count – but coreference uses the whole document
- The resources used will grow with the document size – you might want to try a chapter not a novel
- Coreference systems normally require processing with parsers, NER, etc. first, and use of lexicons

Coreference resolution warnings

- English-only for the moment….
- While there are some papers on coreference resolution in other languages, I am aware of no downloadable coreference systems for any language other than English
- For English, there are a good number of downloadable systems, but their performance remains modest. It's just not like POS tagging, NER or parsing

Coreference resolution warnings

Nevertheless, it's not yet known to the State of California to cause cancer, so let's continue….

Stanford CoreNLP

http://nlp.stanford.edu/software/corenlp.shtml

- Stanford CoreNLP is our new package that ties together a bunch of NLP tools
  - POS tagging
  - Named Entity Recognition
  - Parsing
  - *and* Coreference Resolution
- Output is an XML representation [only choice at present]
- Contains a state-of-the-art coreference system!

Stanford CoreNLP

```
$ java -mx3g -Dfile.encoding=utf-8 -cp "Software/
stanford-corenlp-2011-06-08/stanford-
corenlp-2011-06-08.jar:Software/stanford-
corenlp-2011-06-08/stanford-corenlp-
models-2011-06-08.jar:Software/stanford-
corenlp-2011-06-08/xom.jar:Software/stanford-
corenlp-2011-06-08/jgrapht.jar"
edu.stanford.nlp.pipeline.StanfordCoreNLP -file
RiderHaggard/Hunter\ Quatermain\'s\ Story\
2728.txt -outputDirectory corenlp
```

What Stanford CoreNLP gives

  - Sarah asked her father to look at her .
  - He appreciated that his eldest daughter wanted to speak frankly .
- Coreference resolution graph
  - sentence 1, headword 1 (gov)
  - sentence 1, headword 3
  - sentence 1, headword 4 (gov)
  - sentence 2, headword 1
  - sentence 2, headword 4

What Stanford CoreNLP gives

- Sarah asked her father to look at her .
- He appreciated that his eldest daughter wanted to speak frankly .

- Coreference resolution graph
  - sentence 1, headword 1 (gov)
  - sentence 1, headword 3

  - sentence 1, headword 4 (gov)
  - sentence 2, headword 1
  - sentence 2, headword 4

The rest of the languages of the world

English-only?

- There are a lot of languages out there in the world!
- But there are a lot more NLP tools for English than anything else
- However, there is starting to be fairly reasonable support (or the ability to build it) for most of the top 50 or so languages…
- I'll say a little about that, since some people are definitely interested, even if I've covered mainly English

POS taggers for many languages?

- Two choices:
  1. Find a tagger with an existing model for the language (and period) of interest
  2. Find POS-tagged training data for the language (and period) of interest and train your own tagger
     - Most downloadable taggers allow you to train new models – e.g., the Stanford POS tagger
       - But it may involve considerable data preparation work and understanding and not be for the faint-hearted

POS taggers for many languages?

- One tagger with good existing multi-lingual support
  - TreeTagger (Helmut Schmid)
    - http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
    - Bulgarian, Chinese, Dutch, English, Estonian, French, Old French, Galician, German, Greek, Italian, Latin, Portuguese, Russian, Spanish, Swahili
    - Free for non-commercial, not open source; Linux, Mac, Sparc (not Windows)
  - Stanford POS Tagger presently comes with:
    - English, Arabic, Chinese, German
- One place to look for more resources:
  - http://nlp.stanford.edu/links/statnlp.html
    - But it's always out of date, so also try a Google search ☺

Chinese example

- Chinese doesn't put spaces between words
  - Nor did Ancient Greek
- So almost all tools first require word segmentation
  - I demonstrate the Stanford Chinese Word Segmenter
  - http://nlp.stanford.edu/software/segmenter.shtml
- Even in English, words need some segmentation – often called tokenization
  - It was being implicitly done before further processing in the examples till now: "I'll go." → " I 'll go . "

Chinese example

- `$` ../Software/stanford-chinese-segmenter-2010-03-08/segment.sh ctb Xinhua.txt utf-8 0 > Xinhua.seg
- `$` java -mx300m -cp ../Software/stanford-postagger-full-2011-05-18/stanford-postagger.jar edu.stanford.nlp.tagger.maxent.MaxentTagger -model ../Software/stanford-postagger-full-2011-05-18/models/chinese.tagger -textFile Xinhua.seg > Xinhua.tag

Chinese example

space before 。below!

$ perl -pe 'if (! m/^\s*$/ && ! m/^.{100}/) { s/$/ 。/; }' < Xinhua.seg > Xinhua.seg.fixed

$ java -mx600m -cp ../Software/stanford-parser-2011-06-15/stford-parser.jar edu.stanford.nlp.parser.lexparser.LexicalizedParser -encoding utf-8 ../Software/stanford-parser-2011-04-17/chineseFactored.ser.gz Xinhua.seg.fixed > Xinhua.parsed

$ java -mx1g -cp ../Software/stanford-parser-2011-06-15/stanford-parser.jar edu.stanford.nlp.parser.lexparser.LexicalizedParser -encoding utf-8 -outputFormat typedDependencies ../Software/stanford-parser-2011-04-17/chineseFactored.ser.gz Xinhua.seg.fixed > Xinhua.sd

Other tools

- Dependency parsers are now available for many languages, especially via MaltParser:
  - http://maltparser.org/
- For instance, it's used to provide a Russian parser among the resources here:
  - http://corpus.leeds.ac.uk/mocky/
- The OPUS (Open Parallel Corpus) collects tools for various languages:
  - http://opus.lingfil.uu.se/trac/wiki/Tagging%20and%20Parsing
- Look around!

Data sources

- Parsers depend on annotated data (treebanks)
- You can use a parser trained on news articles, but better resources for humanities scholars will depend on community efforts to produce better data
- One effort is the construction of Greek and Latin dependency treebanks by the Perseus ProjectI:
  - http://nlp.perseus.tufts.edu/syntax/treebank/

Parting words

Applications? (beyond word counts)

- There are starting to be a few applications in the humanities using richer NLP methods:
- But only a few….

Applications? (beyond word counts)

- Cameron Blevins. 2011. Topic Modeling Historical Sources: Analyzing the Diary of Martha Ballard. *DH 2011.*
  - Uses (latent variable) *topic models* (LDA and friends)
    - Topic model are primarily used to find themes or topics running through a group of texts
    - But, here, also helpful for dealing with spelling variation (!)
    - Uses MALLET (http://mallet.cs.umass.edu/), a toolkit with a fair amount of stuff for text classification, sequence tagging and topic models
      - » We also have the Stanford Topic Modeling Toolbox
        - http://nlp.stanford.edu/software/tmt/tmt-0.3/
  - Examines change in diary entry topics over time

Applications? (beyond word counts)

- David K. Elson, Nicholas Dames, Kathleen R. McKeown. 2010. Extracting Social Networks from Literary Fiction. *ACL 2010.*
  - How size of community in novel or world relates to amount of conversation
    - (Stanford) NER tagger to identify people and organizations
    - Heuristically matching to name variants/shortenings
    - System for speech attribution (Elson & McKeown 2010)
    - Social network construction
  - Results showing that urban novel social networks are not richer than those in rural settings, etc.

Applications? (beyond word counts)

- Aditi Muralidharan. 2011. A Visual Interface for Exploring Language Use in Slave Narratives *DH 2011.* http://bebop.berkeley.edu/wordseer
  - A visualization and reading interface to American Slae Narratives
    - (Stanford) Parser used to allow searching of particular grammatical relationships: *grammatical search*
    - Visualization tools to show a word's distribution in text and to provide a "collapsed concordance" view – and for close reading
  - Example application is exploring relationship with God

Parting words

This talk has been about tools –
they're what I know

But you should focus on disciplinary insight –
not on building corpora and tools, but on using them as tools for producing disciplinary research


LELAND STANFORD JUNIOR UNIVERSITY
DIE LUFT DER FREIHEIT WEHT