3.06pt

Introduction to statistical methods for language and literature

Diana Santos

ILOS

d.s.m.santos@ilos.uio.no

Spring 2017, literature

Summary

Different ways of using statistics in literature

- Authorship attribution
- Stylometrics
- Gender studies
- Distant reading
- Topic modelling

Chronology of statistics in literature

Following Kenny's book *The computaton of style: an introduction to statistics for students of literatur and humanities* (1982), this is the history before computers:

1851 Augustus de Morgan: authorship attribution

1867 Lewis Campbell: relative dating of Plato's works

1887 Mendenhall: word spectrum as a function of word length

1888 Ritter: relative dating

, 1892 Sherman: language evolution, from average sentence length

1897 Lutoslawski: 500 features for stylometry

, 1944 Yule: sentence length to distinguish among authors, and vocabulary richness for author attribution

Chronology of statistics in literature (II)

After the arrival of computers:

, 1957 Wake, Morton: Paulus' epistles (are they written by Paulus?)

1962 Ellegård: *Junius brev*, 20 possible authors

1964 Mosteller & Wallace: *The Federalist papers*, 2 possible authors

197- Kenny: 3 chapters in two books of Aristotle on ethics

Suitbert Ertel: dogmatism ratio in philosophy (or other texts): How mant (in)certainty words in six different domains (frequency, degree, etc.)

NB! A large part of statistical application to literature is about poetry: meter, rythm, aliteration, kinds of sounds, etc.), for example Zylinski's chronology in 1906 of Euripides plays, based on the number of "resolved feet".

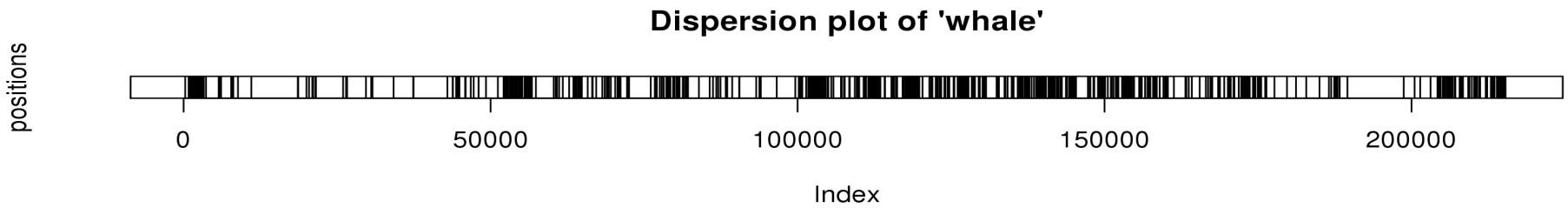
Simple example

For two books attributed to Aristotle, with the following distribution og part of speech of the last word in the sentence, for the first 100 sentences. (O: other, N: noun, V: verb)
Assuming that pos of the last word in a sentence is a good attribution feature, what can we conclude from the table?

Part of speech	N	V	O
Rhetorics	28	32	40
R. to Aleksander	27	52	21

Studies of particular authors or works

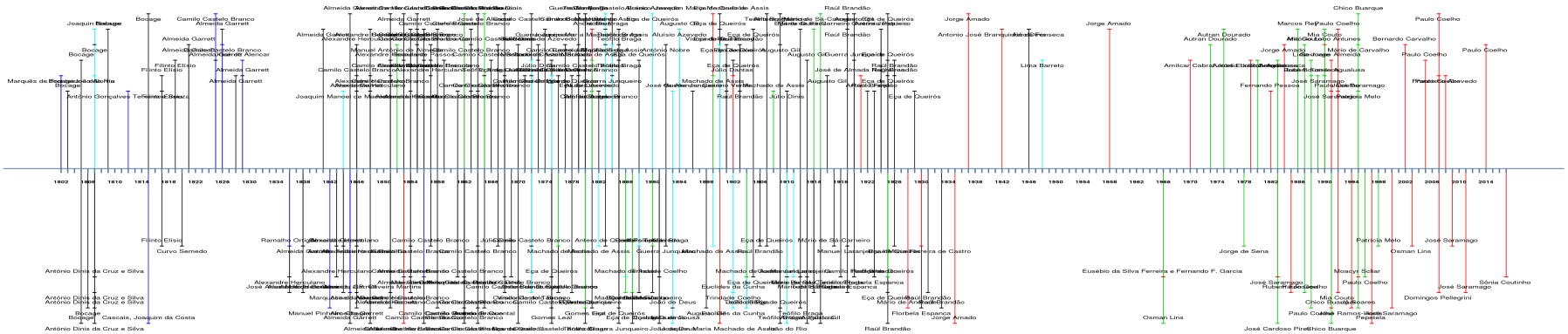
- the distribution of a word in a book



- the comparison of different chapters, or of different books
- do Jane Austen characters differ in adjective use according to gender? And is adjective use different when women talk among them or with men?
- authorial position in Dickens indicated by suspended quotations and body language within them

The notion of distant reading

You cannot read everything, “distant reading” is a way to look at large bodies of literary works. (Franco Moretti, conjectures on world literature, and the “Literary Lab”)
Hence the notion of the “macroscope”, to see things from a distance.



Looking at whole literatures

Jockers's examples based on metadata

- Chronological plotting of lexical richness in British novel titles
- Chronological plotting of "love" in British novel titles

using the content (words)

- classification into genre
- locative prepositions are more frequen in Gothic novels because these are place-oriented
- less frequent use of "the" in British vs. American novels, but correlated over time
- word clusters per nationality

Topic modelling

Automatic discovering of topics (or themes) (collocations of collocations): topics are ranked sets of words, some more interpretable and coherent than others

- Sharon Block discovers topics from a newspaper collection
- Cameron Blevin analyses a diary
- Mathew Jockers uses 3,346 nineteenth century books in English

Topic modelling: the procedure with MALLET

- choose stopwords to ignore
- use only nouns
- segment into 1,000 word segments
- extract 500 topics

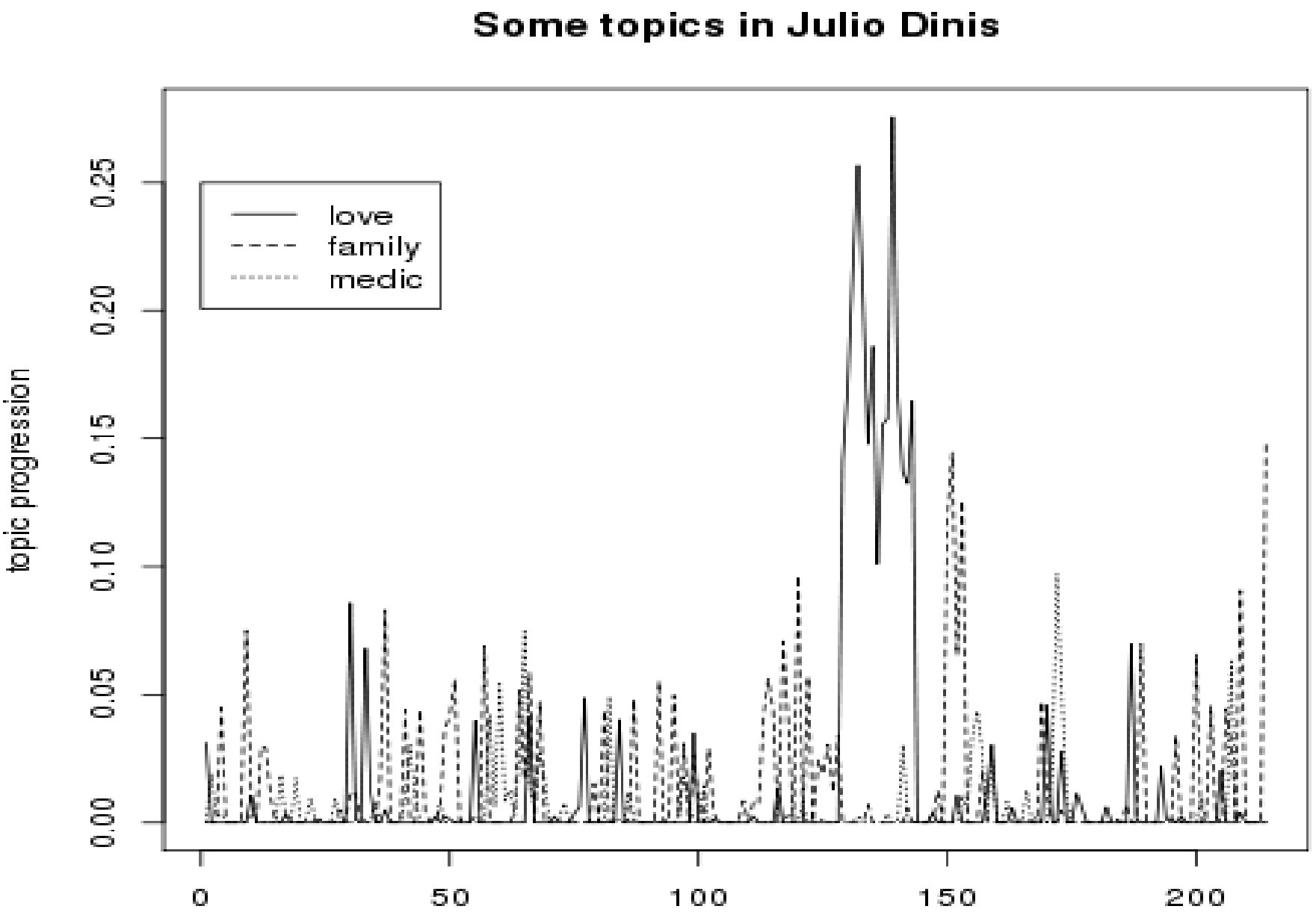
Each topic is a list of words (can be seen as a word cloud); and for each text there is a proportion of topics it includes.

One example: topics in Portuguese literature

100 topics from 176 works

0	0.05	amor olho coração triste alma peito doce flor mal bem amante be
1	0.05	doutor médico doente doença amigo medicina enfermo cirurgião
2	0.05	ali árabe largo água grande deserto pequeno rua cheio luz cidade
3	0.05	olho mil mão luz ouro onda rosto vivo asa nuvem mar alto raio v
4	0.05	romano lusitano catão senado povo cidade cônsul exército sempre
5	0.05	filha senhora mulher marido carta casa menina mãe pai bom bem
6	0.05	homem bem grande bom jornal coisa apenas nunca ora dia alto n
7	0.05	pastor doce campo dia monte flor verso verde gado canto vale fo

One example: topics in Portuguese literature (II)



Looking at whole literatures (II)

- Schöch looks at French classical and Enligthment drama (391 plays)
- Cortelazzo et al at Italian contemporary literature (92 books)
- Broadwell & Tangherlini at Scandinavian late nineteenth century

Some final examples from Baayen

- Zipf's law in Moby Dick (p. 42, exercise 3)
- Investigate Alice in Wonderland in the exercises pages 65-67
- Prose vs. poetry in Old French (section 5.1.3)
- Discussion of lexical richness (section 6.5)

References used for this presentation

- Jockers, Matthew L. *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.
- Jockers, Matthew L. *Text analysis with R for Students of Literature*. Springer, 2014.
- Kenny, Anthony. *The computaton of style: an introduction to statistics for students of literatur and humanities*, Pergamon Press, 1982.
- Schmidt, Kari Anne Rand. "Male and female language in Jane Austen's novels", in Stig Johansson & Bjørn Tysdahl (eds.), *Papers from the First Nordic Conference for English Studies (Oslo, 19-19 September, 1980)*, pp. 198-210.
- Schöch, Christof. "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama", *Digital Humanities Quarterly*.
- Tangherlini, Timothy R. "The Folklore Macroscope: Challenges for a Computational Folkloristics." The 34th Archer Taylor Memorial Lecture. *Western Folklore* 72(1), 2013, pp. 7-27.

Other references

- Allison, Sarah, Ryan Heuser, Matthew Jockers, Franco Moretti & Michael Witmore. "Quantitative Formalism: an Experiment". Literary Lab, Pamphlet 1, january 15, 2011. http://litlab.stanford.edu/LiteraryLabPamphlet1.pdf
- Moretti, Franco. "Conjectures on world literature", *New Left review* 1, Jan-Feb 2000, pp. 54-68.
- Moretti, Franco. "Network theory, plot analysis", *New Left review* 68, Mar-Apr 2011, pp. 80-102. https://newleftreview.org/II/68/franco-moretti-network-theory-plot-analysis
- MALLET: MAchine Learning for LanguagE Toolkit, `http://mallet.cs.umass.edu/`