Mining for characterising patterns in literature using Correspondence Analysis
An experiment on French novels

Francesca Frontini
Istituto di Linguistica Computazionale “A. Zampolli” - Pisa
Mohamed Amine Boukhaled, Jean-Gabriel Ganascia
LIP6 UPMC / Labex OBVIL Paris

Göttingen Dialog in Digital Humanities, Tuesday 14th July 2015

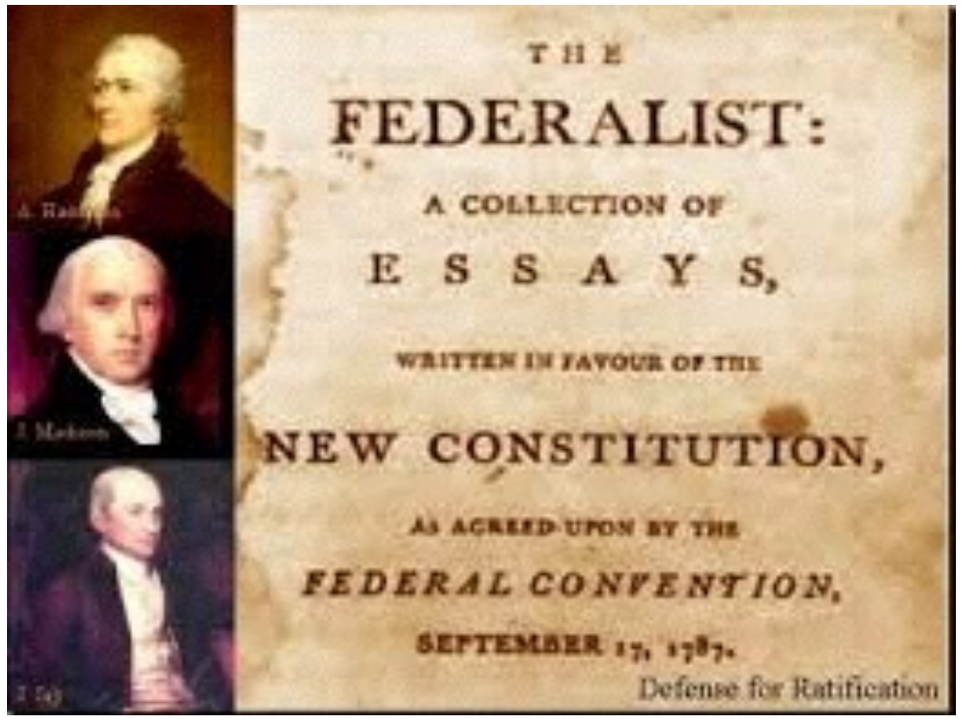
Outline

- Small excursus on the study of literature using computational / algorithmic means
- Where does the present study stand
- Our methodology
- The case study on novels
- What does this tell us

Modern day stylometry

- Mosteller and Wallace's (1964) analysis of the distribution of function words (e.g., prepositions, conjunctions, articles) in a corpus of the 85 'Federalist Papers'.

- "In summary, we can say with better foundation than ever before that Madison was the author of the 12 disputed papers"



The Federalist Papers were written in 1787-1788 by Alexander Hamilton, John Jay and James Madison, under the pseudonym of Publius

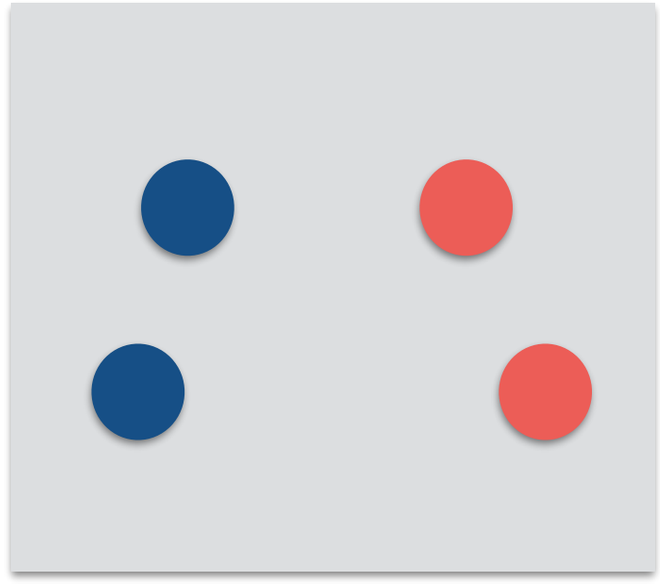
Modern day stylometry

- Used for authorship attribution
- Writer's invariance; stylistic fingerprints of the author
- Statistical analysis, classification, multivariate analysis, MDS, PCA, …
- Linguistic features of variable degree of complexity (average word length, … , function words distribution)

	T1	T2	T3	T4
f1	1	2	2	3
f2	3	4	0	2
f3	0	1	0	6
f4	1	6	1	9

Texts are represented as vectors of features

distances between vectors represent distances between texts



Computational stylistics

- Stylometric methods applied not only to authorship attribution, but also to **computational analysis of style**

- Useful tools such as Stylo for R, allow for basic stylometric analysis

- No complex linguistic features (word or character ngrams)

- Burrows 1987 (on Austen), and many others…

Critique

- Willie van Peer 1989 *Quantitative Studies of Literature. A Critique and an Outlook.*

- Hugh Craig 1999. *Authorial attribution and computational stylistics: if you can tell authors apart, have you learned anything about them?*

- What features for computational stylistics? **Subconscious fingerprints vs author's choice**

Corpus stylistics

- Geoffry Leech; Semino & Short 2004; Mahlberg 2012

- corpus linguistics, discourse analysis

- rich linguistic analysis

- attention to the form - function relationship, and to linguistic bases of literary style

- Douglas Biber's work on the linguistic traits defining Genre, Register, Style

Concordance Results

KWIC Plot

Searched for **poor** within **whole text**.

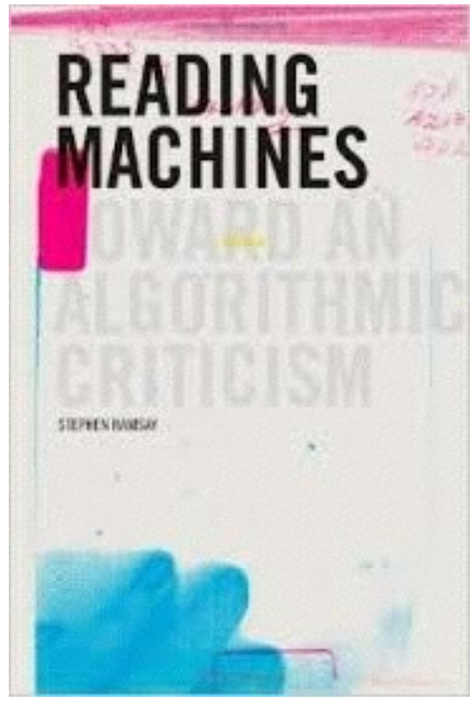
1 to 1000 of 2326 entries | CSV | Print | Toggle metadata | Filter concordance:

	Left	Node	Right	Book	Ch	Par	Sen	In bk
1	…differenbetween a child and a woman; I felt so	poor,	so trifling, and so far off that I never could	Bleak House	3	3	19	
2	…froremained unsoftened. On the day after my	poor	good godmother was buried, the gentleman in …	Bleak House	3	28	104	
3	…tlittle school, and the unexpected sight of the	poor	children outside waving their hats and bonnets…	Bleak House	3	107	270	
4	Lord High Chancellor, at his best, appeared so	poor	a substitute for the love and pride of parents. "…	Bleak House	3	135	336	
5	Ada was a little frightened, I said, to humour the	poor	old lady, that we were much obliged to her. "Y…	Bleak House	3	169	400	
6	leading the way back. "By no means," said the	poor	old lady, keeping up with Ada and me. "Anythi…	Bleak House	3	172	408	
7	…agot his head through the area railings!" "Oh,	poor	child," said I; "let me out, if you please!" "Pray	Bleak House	4	19	39	
8	…somethisaid Mr. Guppy. I made my way to the	poor	child, who was one of the dirtiest little unfortun…	Bleak House	4	21	42	
9	… came into Mrs. Jellyby's presence, one of the	poor	little things fell downstairs-- --down a whole fli…	Bleak House	4	22	48	

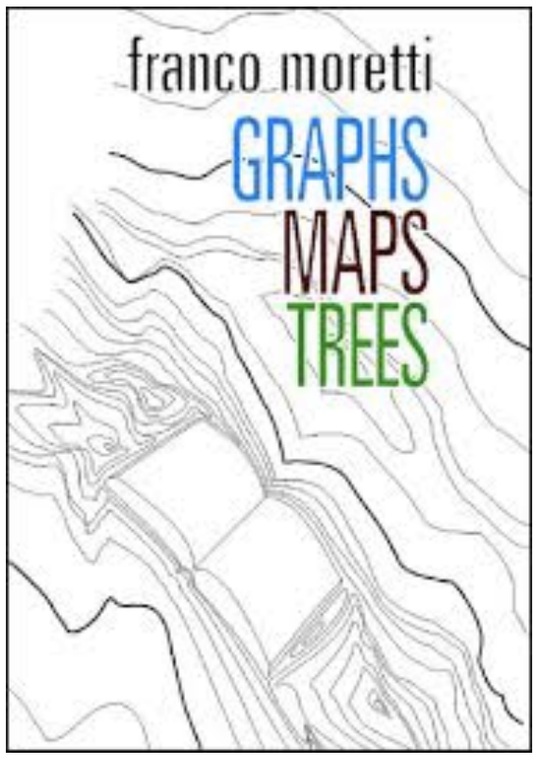
Concordancers used to extract lexical bundles
Interpretative work has greater weight

Algorithmic criticism

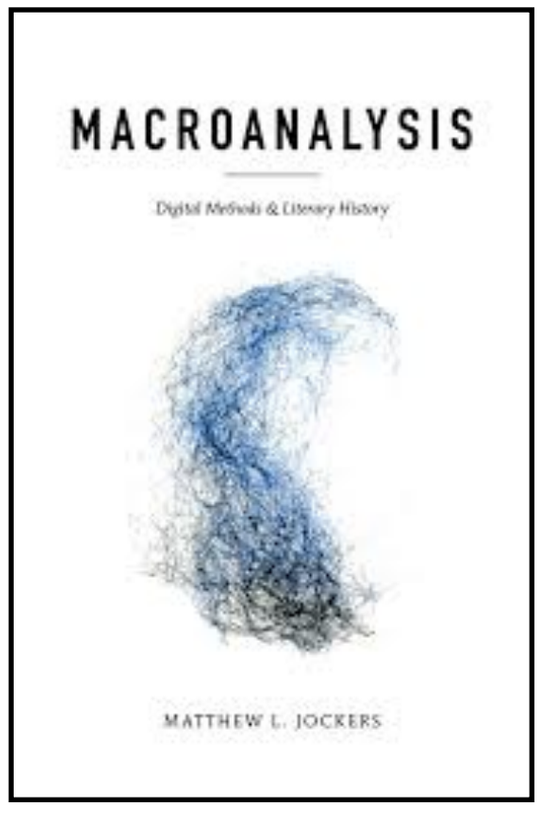
- Stephen Ramsay 2011 *Reading machines*
- **hermeneutical vs experimental methods**
  - *“This work takes the contrary view and proposes that scientific method …. is, for the most part, incompatible with the terms of humanistic endeavor”*
  - <u>investigate/manipulate the text</u> with computational methods vs <u>prove</u> with computational methods



Distant reading & macro analysis



Moretti, 2005



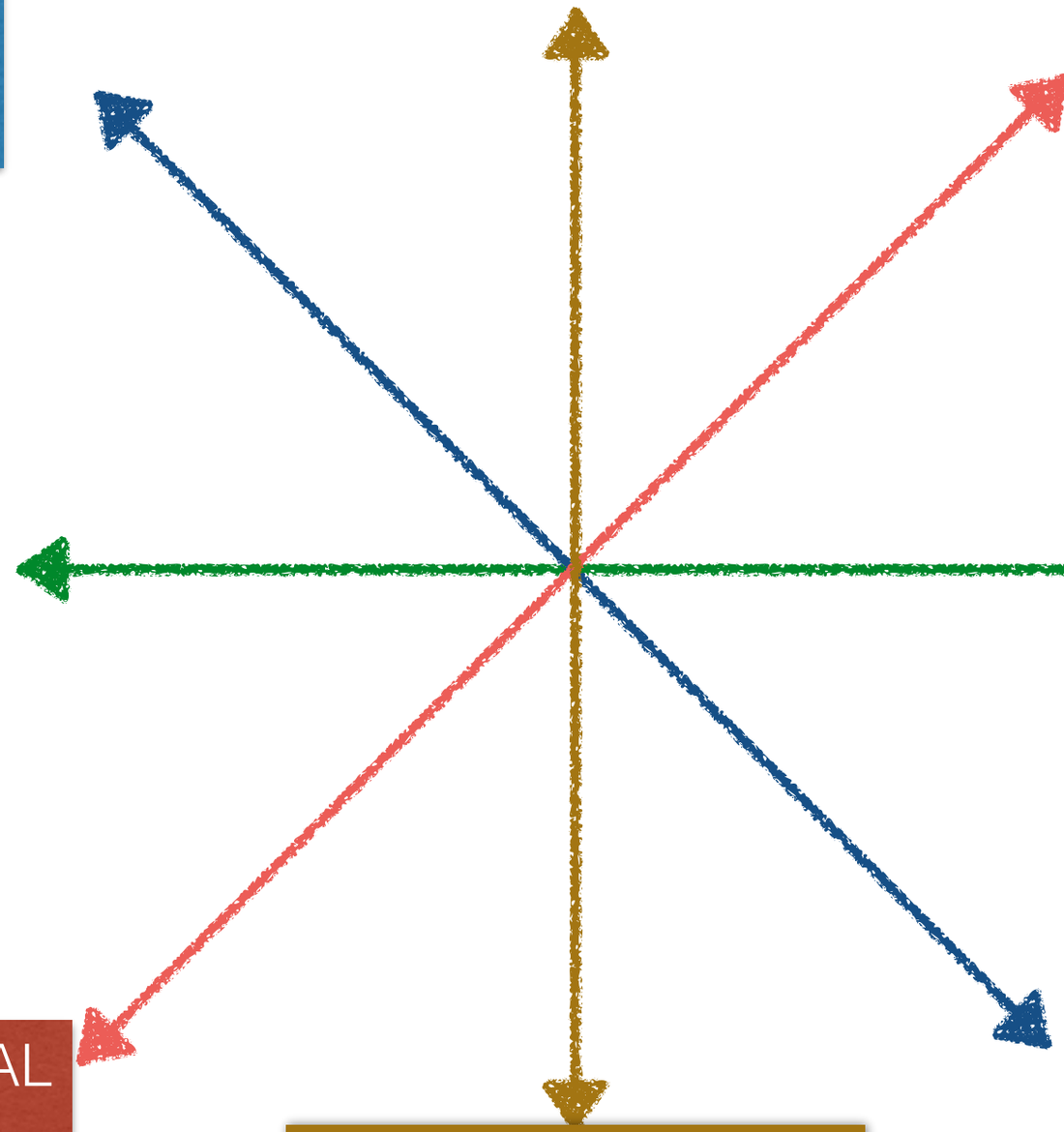
Jockers, 2013

Large quantities of texts, diachronic approach, metadata, visualisations

COUNT FEATURES IN TEXT
statistics, plots

MICRO
focus on few works/authors

HERMENEUTIC APPROACH (EXPLAIN)



SIMPLE LINGUISTIC FEATURES

COMPLEX LINGUISTIC FEATURES

EXPERIMENTAL APPROACH (PROVE)

MACRO
many books
historical perspective

MANIPULATE TEXT
search, visualise

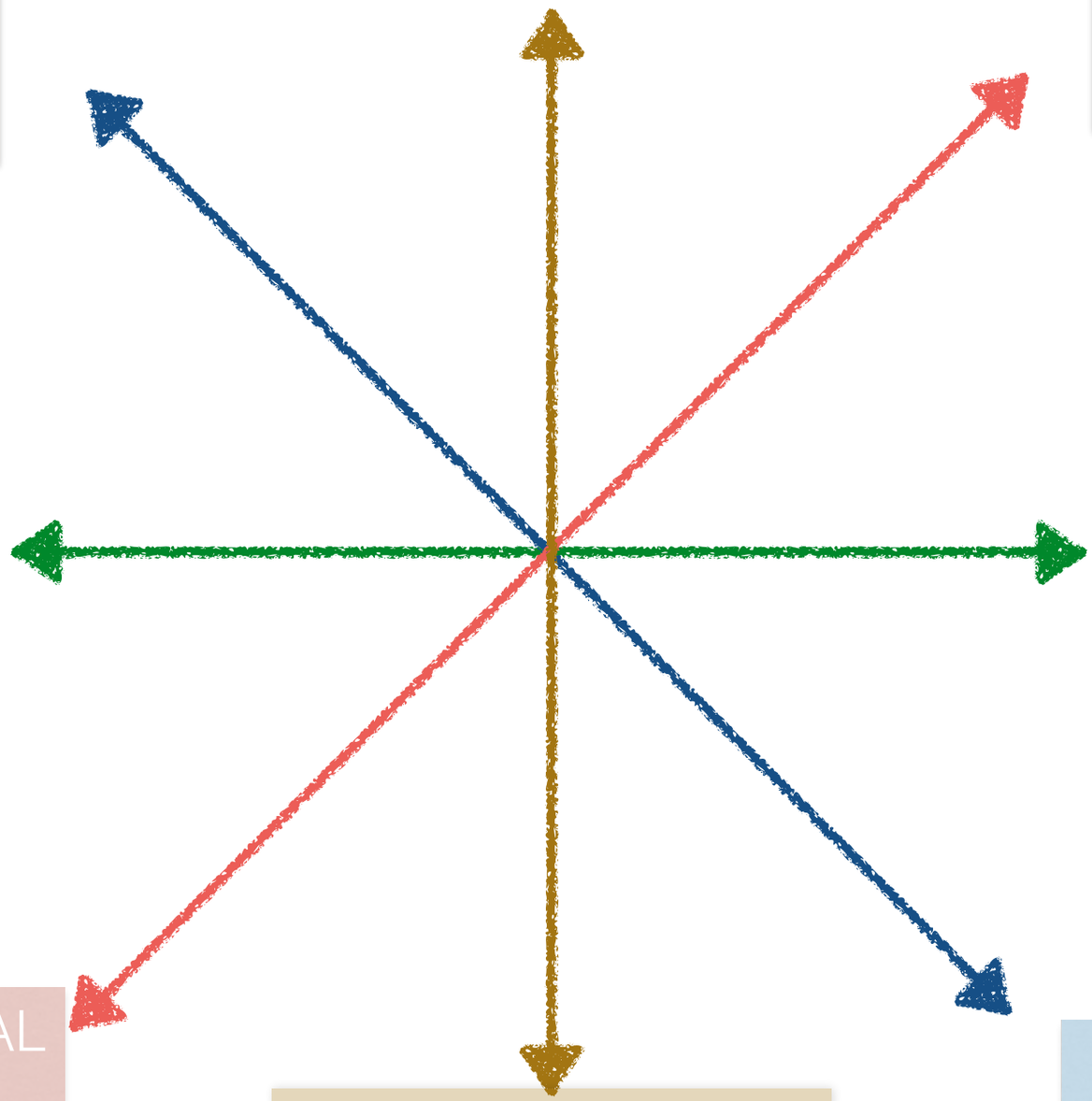
COUNT FEATURES IN TEXT
statistics, plots

MICRO
focus on few works/authors

HERMENEUTIC APPROACH (EXPLAIN)

SIMPLE LINGUISTIC FEATURES

COMPLEX LINGUISTIC FEATURES

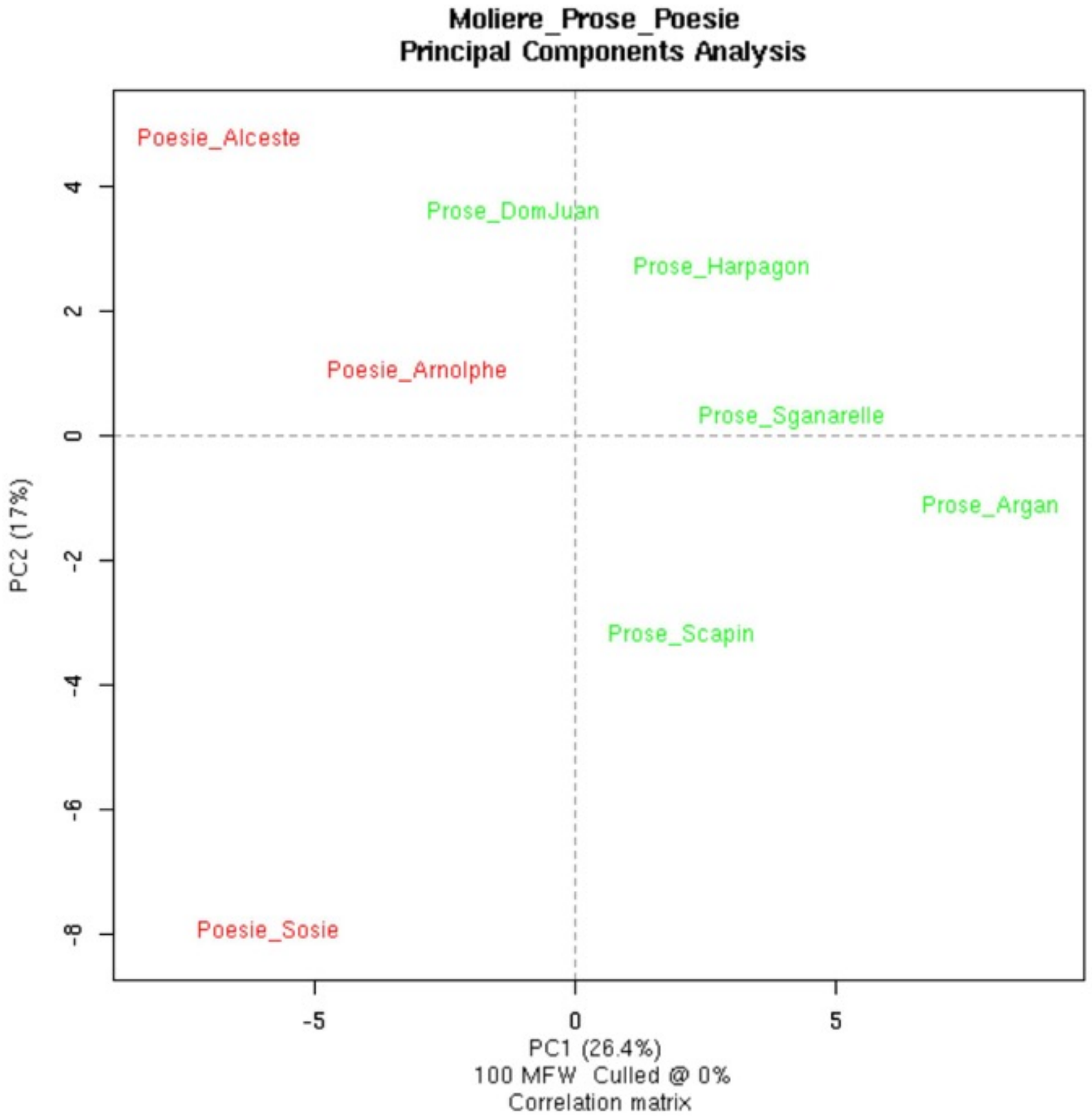


EXPERIMENTAL APPROACH (PROVE)

MACRO
many books
historical perspective

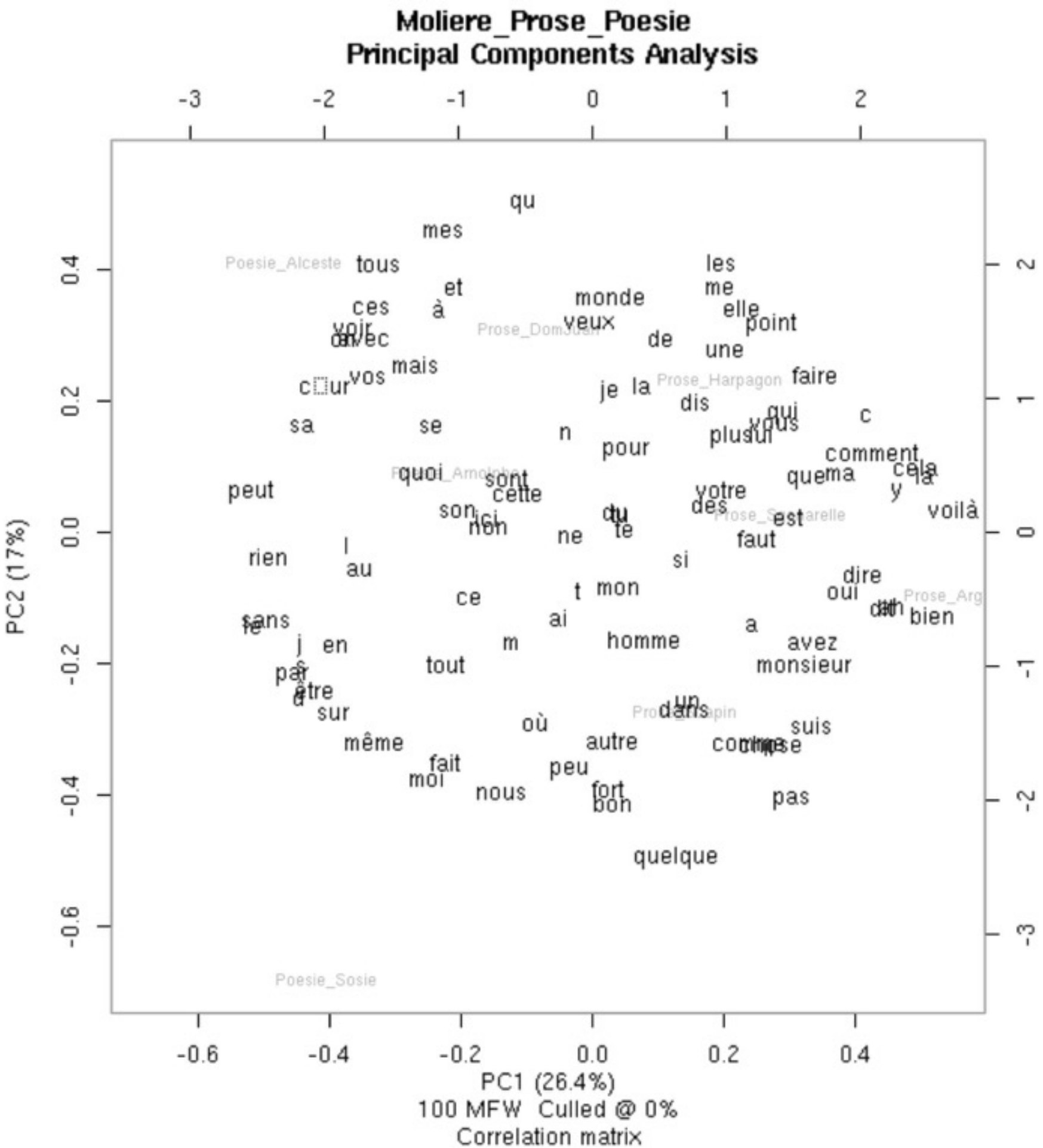
MANIPULATE TEXT
search, visualise

Multivariate analysis


Moliere_Prose_Poesie
Principal Components Analysis
Poesie_Alceste
Prose_DomJuan
Prose_Harpagon
Poesie_Arnolphe
Prose_Sganarelle
Prose_Argan
Prose_Scapin
Poesie_Sosie
PC2 (17%)
4
2
0
-2
-4
-6
-8
-5
0
5
PC1 (26.4%)
100 MFW Culled @ 0%
Correlation matrix

Issue with multivariate analysis

- Why this distribution?
- What patterns make it happen?
- What do they have in common?
- Real motivations for these similarities, based on the text!


Moliere_Prose_Poesie
Principal Components Analysis
PC1 (26.4%)
PC2 (17%)
100 MFW Culled @ 0%
Correlation matrix
Poesie_Alceste
Prose_DomJuan
Prose_Harpagon
Prose_Arnolphe
Prose_Scapin
Prose_Arg
Poesie_Sosie
qu
mes
tous
les
me
et
à
monde
ces
elle
point
veux
de
une
mais
vos
faire
je
la
dis
qui
c
sa
se
n
pour
plus
comment
quoi
sont
cela
peut
cette
votre
y
voilà
son
ici
non
ne
te
des
est
faut
rien
au
si
dire
oui
ce
t
mon
bien
ai
a
sans
j
en
m
homme
avez
tout
monsieur
sur
où
suis
même
autre
peu
fait
moi
nous
fort
bon
pas
quelque

Our work

- Focus on syntax
- EREMOS: a tool to extract all possible **<u>syntactic patterns</u>** for a given text
- Find **interestingness measures** to automatically rank and filter features
- Similar works on theatre, poetry, …

Methodology 1

- tokenise, lemmatise, PoS tag (TreeTagger) each novel

- **sequential pattern mining** (Agrawal 1993, Fournier-Viger et al 2014) extracts sequences of PoS, with possible gaps

- describe each novel with a vector of pattern frequencies

Methodology 2

- “Je suis perdu, je suis assassiné, on m'a coupé la gorge, on m'a dérobé mon argent. “

- Je_je_PRO suis_être_VER…

- (PRO) (VER) (VER) (PUN) **f. 2**
  (PRO)(PRO)(VER)(VER)(DET)(NOM) **f. 2**

Methodology 3

- Extract **thousands of patterns** for each text; no previous feature selection. **Bottom up approach**.

- How to filter out the irrelevant patterns?

- Perform **Correspondence Analysis** (Benzécri, 1977)

- Possible to **filter for contribution, extracting patterns that contribute the most to the final plot**

- instance retrieval with EREMOS for significant patterns **(interpretation!)**

Style in novels

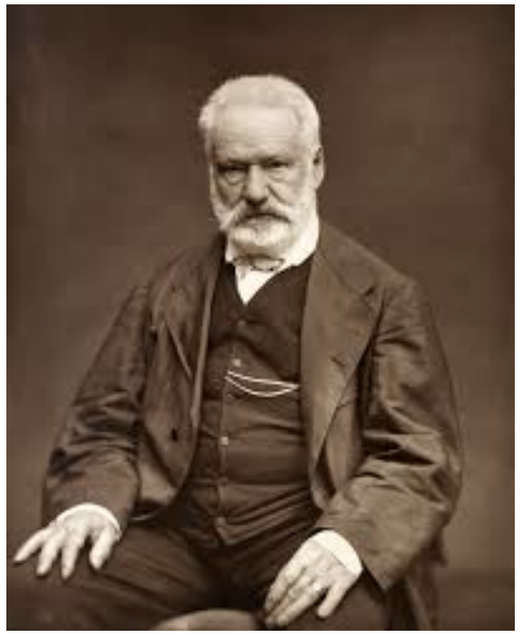
- Analyse novels of same author

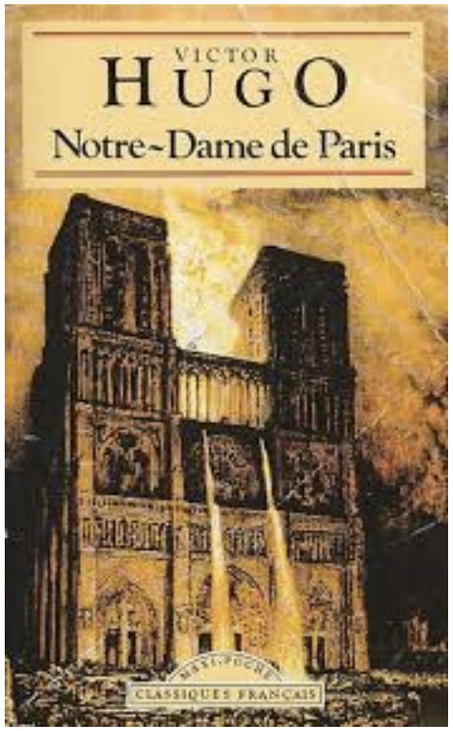
- Analyse **novels of different authors** (style)

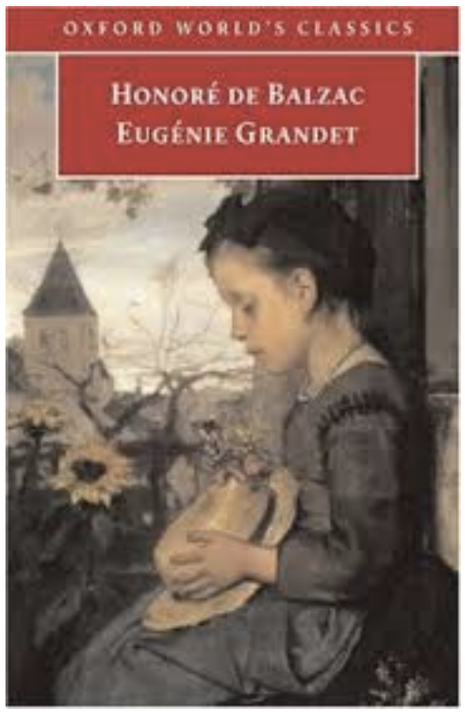
Approach

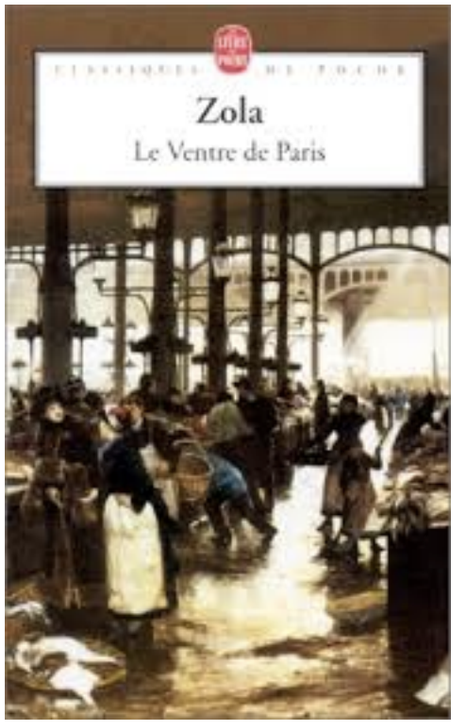
- Compare syntactic style of four memorable novels
- Bottom up feature selection and filtering
    - PoS ngrams 3-5
- Instance retrieval
- See if the emerging patterns match with expectations

Four classic French Novels





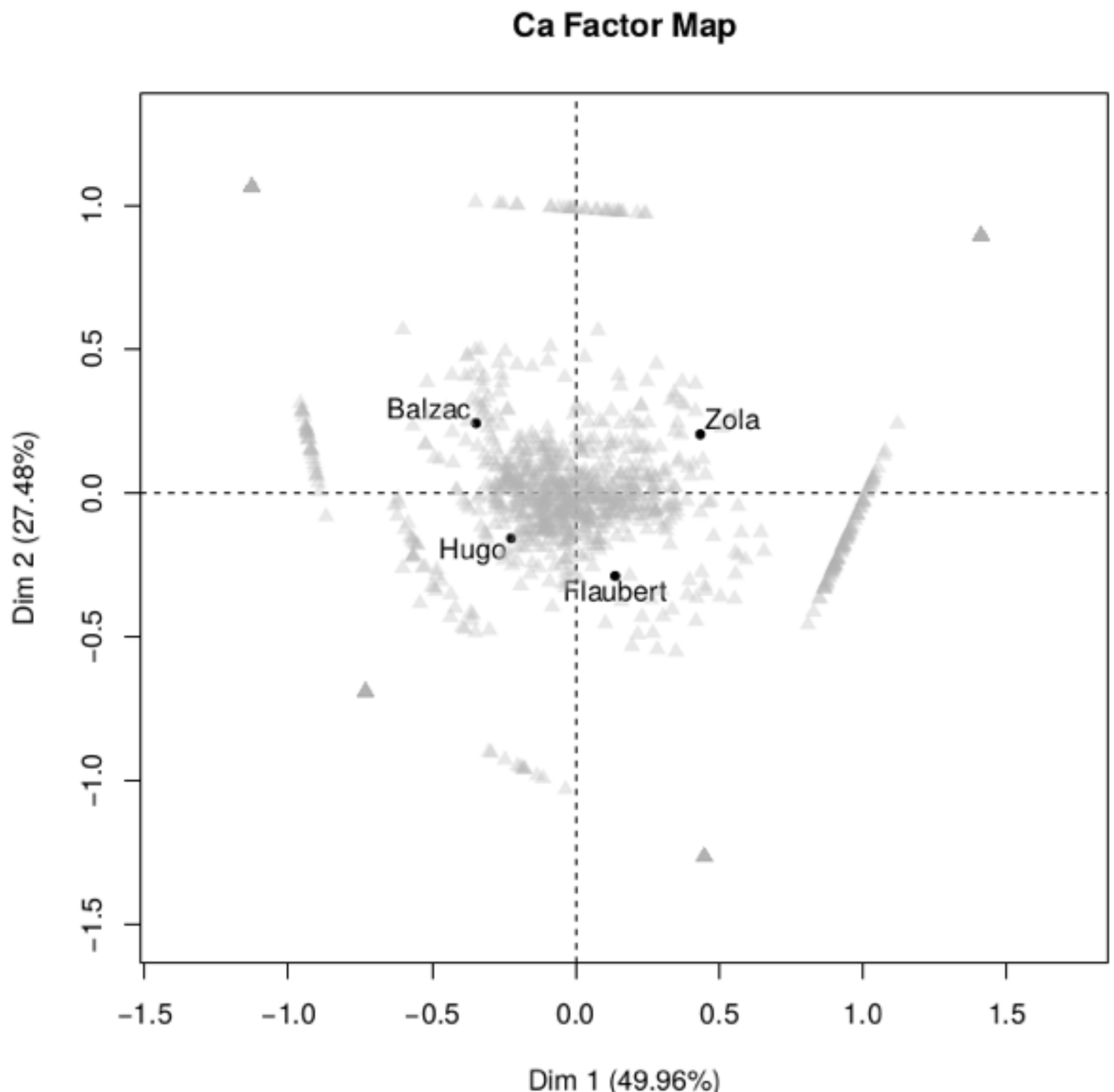




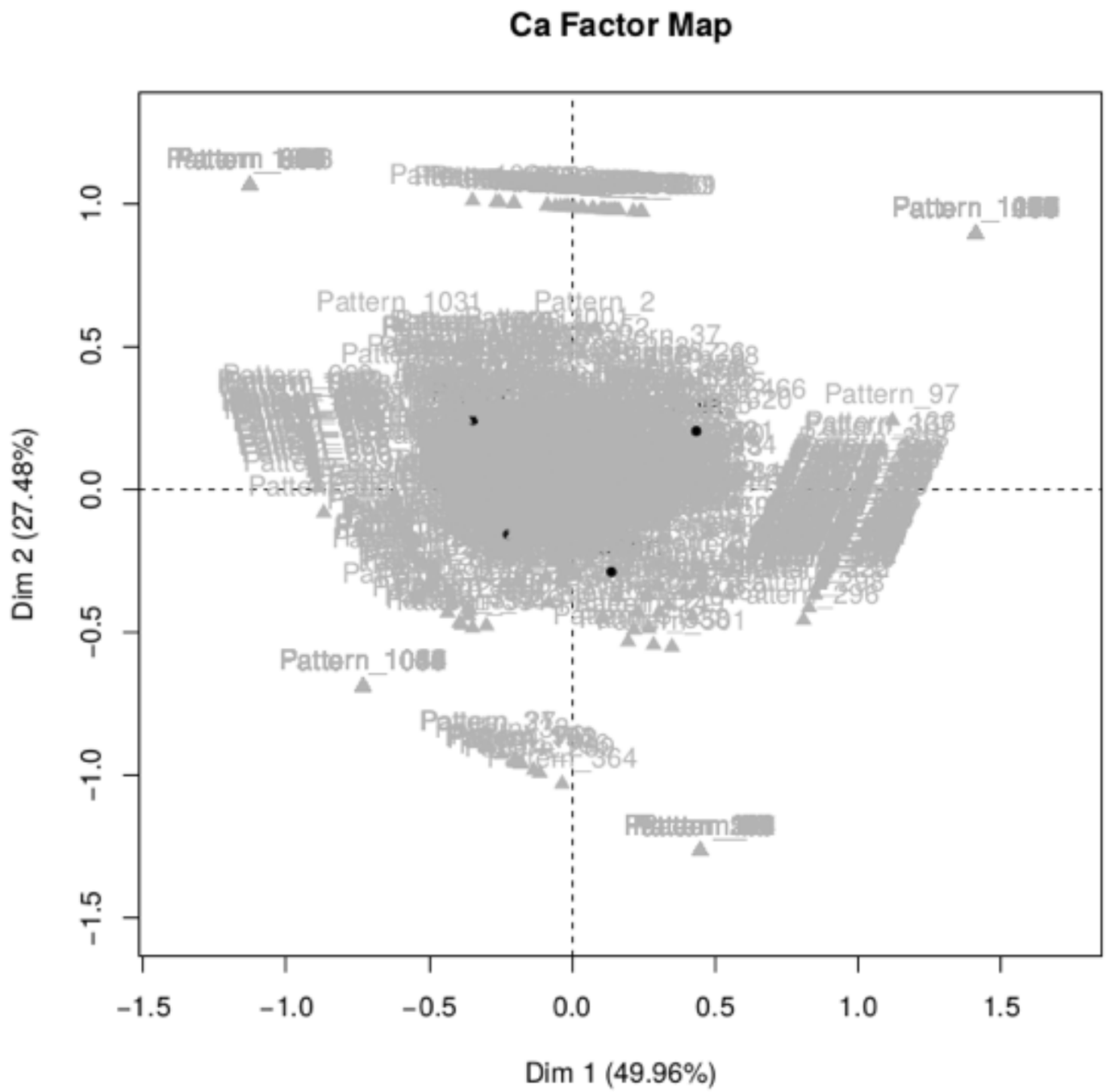
4 Novels

Hugo	**Notre Dame de Paris**	**1831**
Balzac	**Eugenie Grandet**	**1833**
Flaubert	**Madame Bovary**	**1856**
Zola	**Ventre de Paris**	**1873**

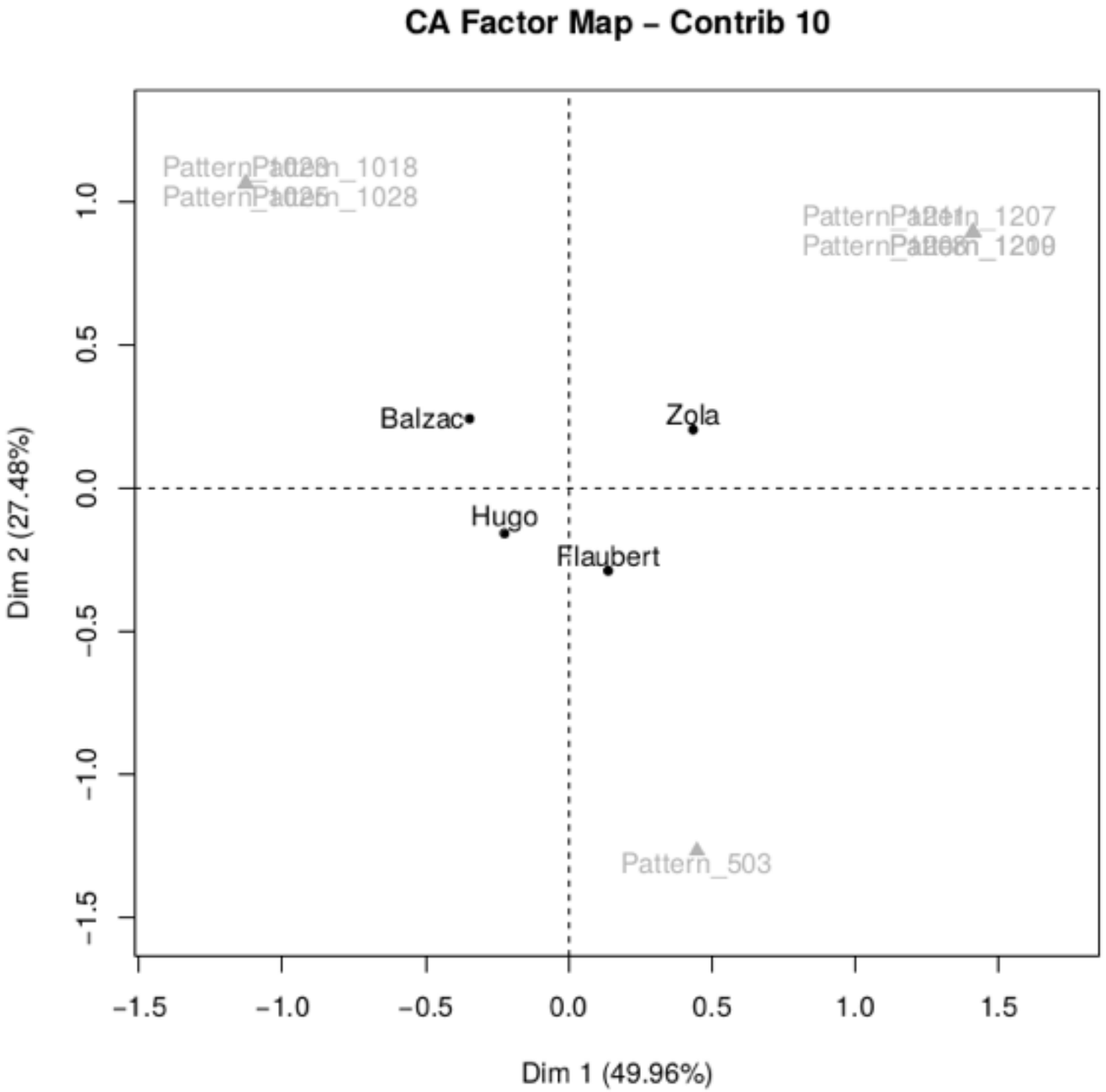
Correspondence Analysis



Bi-plot without filtering



Filtering by contribution



Top 10 patterns

Contri	Pattern ID	Pattern	Novel/Author
1	Pattern_1211	NOM - ADJ - PUN - DET - NOM	Zola
2	Pattern_1028	VER – NAM - PRP	Balzac
3	Pattern_1025	NOM – PRP – VER – DET - NOM	Balzac
4	Pattern_1023	PRP – NAM – VER	Balzac
5	Pattern_1210	DET – ADJ - NAM	Zola
6	Pattern_503	NOM – PUN – KON - PUN	Flaubert
7	Pattern_1209	PUN – DET – NOM – ADJ - PUN	Zola
8	Pattern_1208	DET – NOM – ADJ – PUN - DET	Zola
9	Pattern_1018	VER – DET – NOM – PRP - VER	Balzac
10	Pattern_1207	VER – PUN – VER - PRP	Zola

Top 10 patterns

Contri	Pattern ID	Pattern	Novel/Author
1	Pattern_1211	NOM - ADJ - PUN - DET - NOM	Zola
2	Pattern_1028	VER – NAM - PRP	Balzac
3	Pattern_1025	NOM – PRP – VER – DET - NOM	Balzac
10	Pattern_1207	VER – PUN – VER - PRP	Zola

Zola and Balzac are more “syntactically marked”

Top 5 per novel

```
for a in authors:

    for p in listOfPatternsOrderedByDecreasingContribution:

        n = getNearestNovel(p)

        add p to listOfPatterns[a]

        if listOfPatterns[a] has length = 5:

            exit
```

Zola

Contrib. rank	Pattern ID	Pattern
1	Pattern_1211	NOM - ADJ - PUN - DET - NOM
5	Pattern_1210	DET – ADJ - NAM
7	Pattern_1209	PUN – DET – NOM – ADJ - PUN
8	Pattern_1208	DET – NOM – ADJ – PUN - DET
10	Pattern_1207	VER – PUN – VER - PRP

Zola

[1211_A] Elle parut l' âme , la **clarté vivante , l' idole** saine et solide de la charcuterie ; et on ne la nomma pl

[1211 B] avec des ra le poussaie ' il longeait s de crampes et de souleurs , le ventre plié , la **vue troublée , les pieds** comme tirés , sans qu' il en eût conscience , par cette image de Paris , au loin , très -loin , derrière l' horizon , qui l' appelait , qui l' attendait .

Vivid descriptions, enumerations

Zola

Contrib. rank	Pattern ID	Pattern
1	Pattern_1211	NOM - ADJ - PUN - DET - NOM
5	Pattern_1210	DET – ADJ - NAM
7	Pattern_1209	PUN – DET – NOM – ADJ - PUN
8	Pattern_1208	DET – NOM – ADJ – PUN - DET
10	Pattern_1207	VER – PUN – VER - PRP

Zola

[1209_A] le ventre plié **<u>, la vue troublée ,</u>** les pieds comme tirés , ...

[1208_A] le ventre pieds comme tirés , ...

Verbless constructions

Zola

Contrib. rank	Pattern ID	Pattern
1	Pattern_1211	NOM - ADJ - PUN - DET - NOM
5	Pattern_1210	DET – ADJ - NAM
7	Pattern_1209	PUN – DET – NOM – ADJ - PUN
8	Pattern_1208	DET – NOM – ADJ – PUN - DET
10	Pattern_1207	VER – PUN – VER - PRP

Zola

[1207_A] Il **marchait , dormant à** demi , dodelinant des oreilles , lorsque , à la hauteur de la rue de Longcha es quatre

Implicit, parenthetical constructions

Zola

Contrib. rank	Pattern ID	Pattern
1	Pattern_1211	NOM - ADJ - PUN - DET - NOM
5	Pattern_1210	DET – ADJ - NAM
7	Pattern_1209	PUN – DET – NOM – ADJ - PUN
8	Pattern_1208	DET – NOM – ADJ – PUN - DET
10	Pattern_1207	VER – PUN – VER - PRP

Zola

[1210_A] la petite Pauline ...
[1210_B] la belle Normande ...

Modifiers of proper nouns

Stylistic traits

- Rich descriptions and enumerations (typical of Zola's early style)
- Strong preference for implicit clauses
- Mimic jargon of Parisian populace

Balzac

Contrib. rank	Pattern ID	Pattern
2	Pattern_1028	VER – NAM - PRP
3	Pattern_1025	NOM – PRP – VER – DET - NOM
4	Pattern_1023	PRP – NAM – VER
9	Pattern_1018	VER – DET – NOM – PRP - VER
11	Pattern_1016	PUN – VER – PRO - PRP

Balzac

```
[1028_A] dit Grandet en ..
[1028_B] reprit Charles en ..
[1028_C] dit Eugé
```

Dialogue markers

Balzac

Contrib. rank	Pattern ID	Pattern
2	Pattern_1028	VER – NAM - PRP
3	Pattern_1025	NOM – PRP – VER – DET - NOM
4	Pattern_1023	PRP – NAM – VER
9	Pattern_1018	VER – DET – NOM – PRP - VER
11	Pattern_1016	PUN – VER – PRO - PRP

Balzac

[1016_A] Bonjour , Grandet **<u>, dit -il au</u>** vigneron
[1016_B] Mademoise

Dialogue markers

Balzac

Contrib. rank	Pattern ID	Pattern
2	Pattern_1028	VER – NAM - PRP
3	Pattern_1025	NOM – PRP – VER – DET - NOM
4	Pattern_1023	PRP – NAM – VER
9	Pattern_1018	VER – DET – NOM – PRP - VER
11	Pattern_1016	PUN – VER – PRO - PRP

Balzac

[1025_A] Depuis le classement de ses différents clos , ses vignes , grâce à des soins constants , étaient devenues la tête du pays , mot technique en **usage pour indiquer les vignobles** qui produisent la première qualité de vin .

[1025_B] et le président eux ; mais l' a résuma leurs **pen** , et offrant s x que madame , dit -il , pourrait faire à monsieur les honneurs de Saumur ?

explicit subordination markers

Balzac

Contrib. rank	Pattern ID	Pattern
2	Pattern_1028	VER – NAM - PRP
3	Pattern_1025	NOM – PRP – VER – DET - NOM
4	Pattern_1023	PRP – NAM – VER
9	Pattern_1018	VER – DET – NOM – PRP - VER
11	Pattern_1016	PUN – VER – PRO - PRP

Balzac

[1023_A] L' Histoire **de France est** là tout entière .
[1023_A] Les habitants **de Saumur étant** peu révolutionnaires ,....

toponyms

Balzac

Contrib. rank	Pattern ID	Pattern
2	Pattern_1028	VER – NAM - PRP
3	Pattern_1025	NOM – PRP – VER – DET - NOM
4	Pattern_1023	PRP – NAM – VER
9	Pattern_1018	VER – DET – NOM – PRP - VER
11	Pattern_1016	PUN – VER – PRO - PRP

Balzac

[1018_A] Charles **tendit la main en défaisant** son anneau

[1018_B] G[illegible] mot à dire .

explicit subordination markers

Stylistic traits

- Frequent dialogues
- Preference for explicit subordination: style more “accessible” than Zola’s?
    - *Grandet regarda sa fille sans trouver un mot à dire . > Grandet, la bouche fermée, ….*

Flaubert

Contrib. rank	Pattern ID	Pattern
6	Pattern_503	NOM – PUN – KON - PUN
20	Pattern_364	NOM – PUN – KON - ADV
31	Pattern_ 362	PUN – ADV – PRO - VER
35	Pattern_ 289	PUN – KON – DET – NOM - PRP
42	Pattern_ 327	PUN – KON – PUN - PRP

Flaubert

Le soir , quand Charles rentrait , elle sortait de dessous ses draps ses longs bras maigres , les lui passait **autour du cou , et , l' ayant fait** asseoir au bord du lit , se metta de ses chagrins : il l' oubliait , il tre !

punctuation

Stylistic trait

- Mangiapane (2012): rythmic rather than logic role of punctuation in Flaubert

Hugo

Contrib. rank	Pattern ID	Pattern
80	Pattern_ 31	NOM – KON – DET – NOM - PRP
85	Pattern_ 27	KON – PRO - NOM
184	P	
185	P	
339	P	

First pattern has rank 80!
Less marked style

Hugo

Contrib. rank	Pattern ID	Pattern
80	Pattern_ 31	NOM – KON – DET – NOM - PRP
85	Pattern_ 27	KON – PRO - NOM
184	Pattern_ 520	PUN – KON – VER
185	Pattern_ 833	NOM – PUN - KON
339	Pattern_ 190	ADV – ADJ - KON

Hugo

[31_A]Au centre de la haute façade gothique du Palais , le grand escalier , sans relâche remonté et descendu par un double courant qui , après s' être brisé sous le perron intermédiaire , s' épandait à larges vag , le grand esca mment dans la **pla**

descriptions, mostly of places

Hugo

Contrib. rank	Pattern ID	Pattern
80	Pattern_ 31	NOM – KON – DET – NOM - PRP
85	Pattern_ 27	KON – PRO - NOM
184	Pattern_ 520	PUN – KON – VER
185	Pattern_ 833	NOM – PUN - KON
339	Pattern_ 190	ADV – ADJ - KON

Hugo

[27_A] Ajoutons que Coppenole était du peuple , et **que ce public** qui l' entourait était du peuple .
[27_B] Et songer **que ce peuple** avait été sur le point de se rebel , par impatience d'

repetitions, demonstratives

Hugo

Contrib. rank	Pattern ID	Pattern
80	Pattern_ 31	NOM – KON – DET – NOM - PRP
85	Pattern_ 27	KON – PRO - NOM
184	Pattern_ 520	PUN – KON – VER
185	Pattern_ 833	NOM – PUN - KON
339	Pattern_ 190	ADV – ADJ - KON

Hugo

[520_A] Quasimodo se plaça devant le prêtre , fit jouer les muscles de ses poings athlétiques **, et regarda** les assaillant[illegible]ncement de dents d' un tigre fâché .

punctuation

Hugo

Contrib. rank	Pattern ID	Pattern
80	Pattern_ 31	NOM – KON – DET – NOM – PRP
85	Pattern_ 27	KON – PRO – NOM
184	Pattern_ 520	PUN – KON – VER
185	Pattern_ 833	NOM – PUN – KON
339	Pattern_ 190	ADV – ADJ – KON

Hugo

[190_A]Qui est aussi fraîche et **aussi gaie que** si elle était veuve .

comparisons

Stylistic trait

- Popular, narrative style. Not very marked.

- Rich descriptions that help us to visualise medieval Paris

Recap

- basic NLP pre-processing
- sequential pattern mining (computationally complex technique that avoids feature pre-selection)
- Correspondence Analysis and filtering for interesting patterns
- Instance retrieval for textual analysis

Conclusions

- Confirm prior knowledge about authors by using a completely bottom up methodology
- Compare authors and find out peculiarities in style
- Rank them in terms of more or less “marked” style

http://obvil.paris-sorbonne.fr

Thanks for your attention!