HAL
open science

Mining for characterising patterns in literature using correspondence analysis: an experiment on French novels

Francesca Frontini, Mohamed Amine Boukhaled, Jean-Gabriel Ganascia

To cite this version:

Francesca Frontini, Mohamed Amine Boukhaled, Jean-Gabriel Ganascia. Mining for characterising patterns in literature using correspondence analysis: an experiment on French novels. Digital Humanities Quarterly, 2017, Göttingen Dialog in Digital Humanities 2015, 11 (2). hal-01527780

HAL Id: hal-01527780

https://hal.science/hal-01527780v1

Submitted on 25 May 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

DHQ: Digital Humanities Quarterly
Preview
2017
Volume 11 Number 2

Mining for characterising patterns in literature using correspondence analysis: an experiment on French novels

Francesca Frontini <francesca_dot_frontini_at_univ-montp3_dot_fr>, Université Paul-Valéry Montpellier 3 - Praxiling UMR 5267 CNRS - UPVM3
Mohamed Amine Boukhaled <mohamed_dot_boukhaled_at_lip6_dot_fr>, Laboratoire d'Informatique de Paris 6 (LIP6 UPMC) / Labex OBVIL
Jean-Gabriel Ganascia <Jean-Gabriel_dot_Ganascia_at_lip6_dot_fr>, Laboratoire d'Informatique de Paris 6 (LIP6 UPMC) / Labex OBVIL

Abstract

This paper presents and describes a bottom-up methodology for the detection of stylistic traits in the syntax of literary texts. The extraction of syntactic patterns is performed blindly by a sequential pattern mining algorithm, while the identification of significant and interesting features is performed at a later stage by using correspondence analysis and by ranking patterns by contribution.

Computational stylistics

Computational stylistics is a form of computer aided literary analysis, that aims to extract significant stylistic traits characterising a literary work, an author, a genre, a period… Computational stylistics shares similarities with computer aided authorship attribution; indeed stylometric methods have begun to be developed in order to identify the most likely author of a text of unknown attribution (see [Craig 2004] for a discussion). 1

The term *stylometry* ([Grzybek 2014] for a history of the concept) could and has been used as a hypernym for both disciplines; the commonalities lie in the method. Having chosen a set of texts, some measurable properties are identified, and texts are mined for such properties or features. Features can be very diverse. 2

Very basic features are traits such as the number of sentences in a text, the number of words in a text, the average number of words in a sentence, average word length, punctuation frequency. Other features are more properly linguistic, relating to vocabulary size of a text (lexical richness), frequency of function words, frequency of Part Of Speech (PoS) tags or PoS n-grams[1], syntactic complexity[2], etc. Obviously some of the higher level features require either manual counting or a reliable NLP tool to perform the annotation. 3

Once features have been selected and counted, each text can be seen as a vector that contains for each feature the counts of such feature in the text itself. The different texts to be compared constitute a matrix such as the one represented in Table 1. Counts can be in absolute or relative terms. Normalization is recommended when the sizes of the texts are different. Nevertheless one should consider that smaller texts have generally a higher internal variability; thus it is recommended not to compare texts of too different sizes and sometimes experiments are carried out on selected samples of equal size. 4

	Feature 1	Feature 2	Feature 3
Text 1	30	15	6
Text 2	403	30	515
Text 3	305	149	58

Table 1. Example of a matrix containing the counts for three possible countable features.

Several methods can then be used to measure and compare the texts based on the frequencies of the features. 5

They generally rely on vector distance measurements that allow one to identify whether two texts show the same behaviour with respect to the selected features. In Table 1 for instance it is evident that Texts 1 and 3 show similar distributions, despite the difference in scale, as they have roughly the same proportion of Features 1, 2 and 3, while Text 3 shows a totally different behaviour, with a higher count for Feature 3 than for Feature 1 and 2.

The similarity of methods notwithstanding the purpose of computational stylistics is profoundly different from the one of authorship attribution. Indeed attribution methods aim to identify unconscious traits in the work of a given author, that give him/her away and that are for this reasons normally defined as *fingerprints*. The basic features listed above (such as word or sentence length), together with function word distribution, have so far proved to be very efficient fingerprints. It is possible that such traits persist in a single author somewhat independently of the kind of text he/she is writing, even aside from literary production in a strict sense (so for instance they could be traced in his/her personal letters). 6

On the other hand literary style is something that an author masters in a more conscious way. It is possible that different works by the same author may show different stylistic traits, although others may be found in all of his/her works. Generally speaking, we can assume that more complex linguistic features are used in a more conscious and controlled way and thus when some of them are strongly over-used or under-used in an author with respect to others, this may be taken as a possible stylistic trait. 7

Moreover, authorship attribution can be clearly framed as a classification problem (who is the most likely author of text A given a bunch of candidates) and indeed it is applied as such not only to literature but also in forensic contexts. Computational stylistics is an open-ended problem [Craig 2004] that consists in identifying such traits as are most distinctive of a set of texts, with respect to other ones. Thus from the computational point of view, computational stylistic methods are framed as algorithms that rank linguistic features in a given text based on measures of *interestingness*. 8

Clearly such measures are more difficult to evaluate in terms of the accuracy measures commonly used in information retrieval. A debate is currently on going on whether computational stylistic methods should be a way to radically change the methodology of literary criticism and make it more "scientific". The influential book by Ramsay, *Reading Machines. Toward an algorithmic criticism* [Ramsay 2011] tells us that it may but needn't be the case. 9

The method presented in this paper extracts and ranks *interesting* differences in the use of syntactic structures from a given set of texts, and will be illustrated by comparing four novels by four different authors. It is therefore an exploratory algorithm - based on a set of freely available computational tools - that aims to provide an aid to literary scholars without fundamentally changing their normal method of analysis. For this reason it should not be applied to discover completely new facts about literary style, but rather to substantiate (or disprove) known facts[3]. More specifically, the qualitative analysis of results that is presented in the second part of this paper constitutes an evaluatory test bed as well as a help for computer scientists and NLP specialists to fine-tune their methods. 10

This is not an uncommon scenario in today's computational stylistics research; it can be compared to the early stages of historical linguistics, when researchers established their comparative method of genetic reconstruction on Romance languages, for which in fact the antecedent (Latin) was available. Only the possibility of independently verifying their methods on an attested source could establish the correctness of comparative methods and allowed researchers to subsequently reconstruct other proto-languages for which no attestation was present (Proto-Germanic, Indo-european). 11

This caveat notwithstanding, we believe that our methodology, alongside other similar approaches, can already be useful for specialists, who can find confirmation of known facts and thus substantiate their claims with more data. With the added value that such algorithms are able to easily process large quantities of text, and can thus be applied to that part of literature that Franco Moretti [Moretti 2005] calls the *archive* (as opposed to the *canon*), notably to works whose lower literary prestige and high number make computational methods more attractive. 12

Computational stylistics

Works investigating the lexical differences between authors using stylometric techniques are common in the literature; generally, studies count and compare individual lexical elements (see some examples on Shakespeare and other playwrights in [Craig 2009], among many others) or lexical patterns (or bundles, as in [Mahlberg 2013]). Similar techniques are applied to the study of genres and of literary trends in a historical perspective, in works that often go under the headwords of *Distant Reading* [Moretti 2005] or *Macroanalysis* [Jockers 2013]. Finally, the 13

sampling of the most frequent words in a set of texts is at the basis of some of the most widespread techniques for authorship attribution, such as Burrows's Delta [Burrows 2002] and its later re-implementations [Rybicki 2011].

Sometimes collocations are extracted and analysed with concordancing tools [Mahlberg 2013] and then compared to a norm corpus (this is the approach normally followed in the corpus stylistics tradition since the seminal works by Geoffrey Leech). In other cases dimensionality reduction methods are used to graphically represent the differences between texts by projecting on a bi-dimensional space the distances between the vectors representing the texts. Multi-Dimensional Scaling (MDS), Principal Component Analysis (PCA) and Correspondence Analysis (CA) are often used for this purpose [Burrows 2012], and are implemented in several tools for stylistic analysis[4]. Such methods are interesting because they allow the researcher to verify a priori assumptions on the similarity/distance of different texts by observing the clusters that appear on the graph. Moreover some of these techniques (PCA and CA) allow for the visual representation on a graph of the features that are more associated with one text than to the other ones. 14

When working with a small number of pre-selected features, a commonly used technique which gives the user insight into the decision process that the algorithm used to produce the visualisation is the so called bi-plot (see Figure 1 for an example). 15

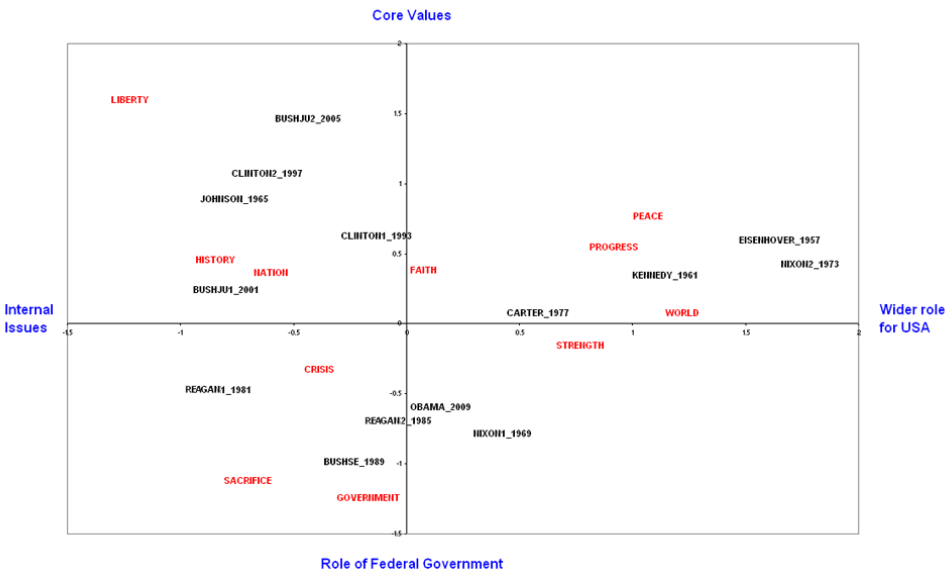

Figure 1. Exploring US Presidential Inaugural Addresses. A piece of whimsy about text and context. An example taken from the documentation of the software TLAB[5]

As you can see from the plot, texts are represented in the bi-dimensional space together with the selected features. The visualisation places the features in the space near to the texts they are more strongly associated with. This means that the researcher can easily identify which features are more responsible for the differences between texts. Also, the relative positioning of the features with respect to each other is indicative of the possible meaning that the representation takes along the two axes. The labels in blue outside the plot on Figure 1 are the interpretation that a human may give of the bi-dimensional distribution of the features. 16

Such visualization methods play a very important role in that they allow the researcher not only to confirm or discard a given a priori classification, but also to explain the reason why this is the case. In this sense, alongside hypothesis driven studies, they can also provide a tool for investigation and analysis that is more in line with common practices within literary criticism. 17

Clearly such techniques can be used not only to study the authors' lexicon, but also to try and detect syntactic differences in style. Reliable parsing is not always available, and may work less well on literary texts even for English; nevertheless syntactic patterns in the form of PoS n-grams or constructions can be easily obtained, as PoS Tagging is nowadays available in many languages and domains. 18

The two main options when trying to port multivariate analysis to syntax are the following: 19

- either a top down approach to feature selection is chosen thus identifying a priori a limited set of

syntactic structures (for an example see, among others, [Dell'Orletta 2011]);
- or bottom-up pattern extraction algorithms are used, which allows for an open ended set of features.

These two scenarios mirror a distinction introduced by [Quiniou 2012] between *paradigmatic* approaches, focussing on the counting of categories, and *syntagmatic* approaches targeting the combinatorial properties of language. 20

Clearly the second option is more in line with the idea of an exploratory tool and gives us some hope of being able to use such techniques in the future to discover new facts about literary texts. Nevertheless pattern extraction algorithms (such as sequential pattern mining, which we shall present in the next paragraph) are known to produce a huge amount of patterns, and thus large vectors, even for small portions of text. Thus bi-plots, that are so useful for exploration, become unreadable due to the projection of thousands of features. In the following sections of this paper a feature ranking method is presented that aims to overcome such impediments, thus allowing researchers to conjugate a bottom-up feature selection procedure and an exploratory visualization of the results. 21

The methodology

The proposed methodology is based on the exploitation of two different techniques, sequential pattern mining and correspondence analysis, followed by an interpretation of the results, and is subdivided into 5 steps: 22

1. data annotation
2. pattern extraction
3. pattern filtering
4. correspondence analysis and visualization
5. pattern extraction

Data is first segmented into sentences, then syntactical categories are annotated by using a freely available tool, TreeTagger [Schmid 1994], [Schmid 1995]; [Stein 2003] for the French tagset used here). Example 1 shows how a French sentence is annotated: 23

Le livre est sur la table.
DET:art - NOM - VER:pres - PRP - DET:art - NOM - SENT

example 1.

On the PoS tagged text sequential pattern mining is applied, namely a data mining technique introduced by [Agrawal 1995] in order to extract interesting characteristics and patterns in sequential databases. [Quiniou 2012] in their study have shown the value of using sequential data mining methods for the stylistic analysis of large texts. Our extraction tool, EREMOS[6] (Extraction et REcherche de MOtifs Syntaxique), allows for the extraction of: 24

- - sequences of full or simplified PoS tag sequences (DET:art - NOM - VER:pres vs DET - NOM - VER)
- - with or without the insertion of lexical elements (Le - NOM - VER:pres vs. DET:art - NOM - est)
- - with or without gaps (DET:art - NOM - VER:pres vs. DET:art - [*] - est)
- - of any given length (normally up to 5 positions including gaps)

Patterns are extracted with their counts. Three types of filtering are applied; one is based on threshold settings. Users can set an absolute threshold, filtering away, for instance, patterns with frequency less than, say, 5 in a text; or a relative one filtering away, for instance, patterns that do not occur in at least 1% of the sentences. Finally automatic filtering is applied in order to eliminate patterns that are included in another one. So, for instance, if we find the following results: 25

- Pattern 1: DET - NOM - VER - PRP - DET = f. 50
- Pattern 2: DET - NOM - VER - PRP = f. 50

we can deduce that all instances the shorter pattern only occur in the context of the longer one and Pattern 2 can be thus be filtered away.

This extraction method has been tested on several corpora such as theatrical plays, poems and novels. According to the size of the corpora and the settings, this extraction method can produce up to 10,000 patterns that can be seen globally as a syntactic description of the text. This method therefore is meant to bypass the feature-selecting phase. The researcher doesn’t need to pre-compile a list of possible syntactic sequences that may differentiate one 26

text from the others. Patterns are extracted bottom-up and blindly. Obviously a large quantity of such patterns will be insignificant for stylistic differentiation as they have probably the same frequency in all texts.

Thus correspondence analysis is then performed as follows: 27

1. pattern vectors from all texts under analysis are imported into one big matrix
2. patterns that are not present in one text are assigned zero by default, (smoothing is also possible)
3. the matrix may or may not be normalised, transforming absolute frequencies into relative ones.
4. correspondence analysis is performed using the FactoMiner tool for R [Husson 2011]
5. contributions are extracted for each pattern
6. filtering of pattern is performed
7. plots and result tables are printed

The last 3 points are crucial and require further description.

Correspondence analysis (CA) is a dimensionality reduction technique developed by Jean Paul Benzécri ([Benzécri 1977], [Greenacre 2007]) that is well known and often used in digital humanities and textual analysis [Lebart 1998]. The main advantage of using CA with respect to, say PCA, and in using an advanced tool such as FactoMiner to perform it rather than some of the commonly used stylometry GUIs, is that the complete results of the analysis are available in a series data structure. Two tables contain the coordinates to project both the texts and the patterns into the plot, respectively. These allow for a selective printing of a subset of patterns on the plot; moreover the proximity of a pattern to any of the texts can be easily calculated by Euclidean distance, thus allowing for the automatic identification of patterns more strongly associated with one text than to the others. 28

The third is the most important result table for our methodology and contains the *contribution* of each pattern on the two axes; contribution is defined as the actual contribution of that pattern to the overall displacement of the position of texts in the resulting plot. If a pattern is strongly overrepresented in a text with respect to the others, it will contribute greatly to the displacement of the text in the bi-dimensional space. Thus, the average contribution on the two axes of this pattern will be higher than the one of other patterns that have more or less the same frequencies in all texts. Subsequently, contribution can be used as an interestingness measure to rank patterns. 29

Finally the extraction tool EREMOS is also equipped with an instance retrieving method; that allows researchers to see all instances in the text corresponding to any given pattern. This latter feature is also very important as experts can verify the evidence in texts and map the automatically identified patterns to the actual linguistic structures that such patterns are in fact mirroring. 30

We shall see with an example how this works practically. The current discussion is not intended to be a thorough critical analysis of the chosen texts, but it aims only to show what possible uses experts may make of the data. 31

An experiment: four classic French novels

In order to show how the methodology works in practice, four classic French novels were chosen: 32

- Victor Hugo, *Notre Dame de Paris*
- Honoré de Balzac, *Eugenie Grandet*
- Gustave Flaubert, *Madame Bovary*
- Emile Zola, *Le ventre de Paris*.

The idea is to compare these four 19th-century works of fiction in order to extract differentiating stylistic traits, without any a-priori targeted structure. In the present experiment, for simplicity, a basic configuration is chosen for EREMOS, extracting patterns 33

- from 3 to 5 positions
- of simplified PoS tags
- without gaps
- without lexical elements.

In this configuration, EREMOS is basically working as a 3-4-5grams extractor.[7] Thus possible patterns from example 1 may be:

- Pattern_1[8]: DET - NOM - VER

- Pattern_2: VER - PRP - DET - NOM
- Pattern_3: DET - NOM - VER - PRP – DET

When performing CA with the basic settings, Figure 2 is what results.

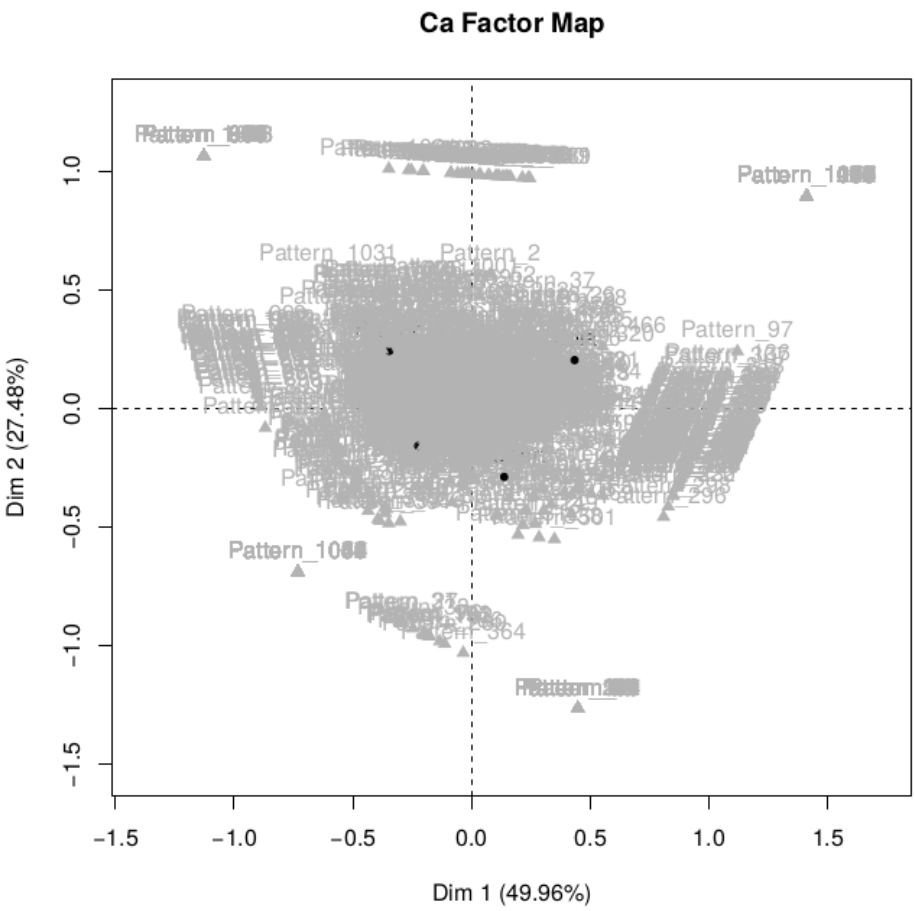

Figure 2. Corresponding analysis on the four selected novels with basic setting.

As you can see, the large amount of patterns makes the plot unusable and the position of the tests themselves in the plot invisible. 34

Using one of the settings provided by FactoMiner we are able to produce a more readable Figure 3, where patterns are unlabelled and printed in grey with partial tran sparency. 35

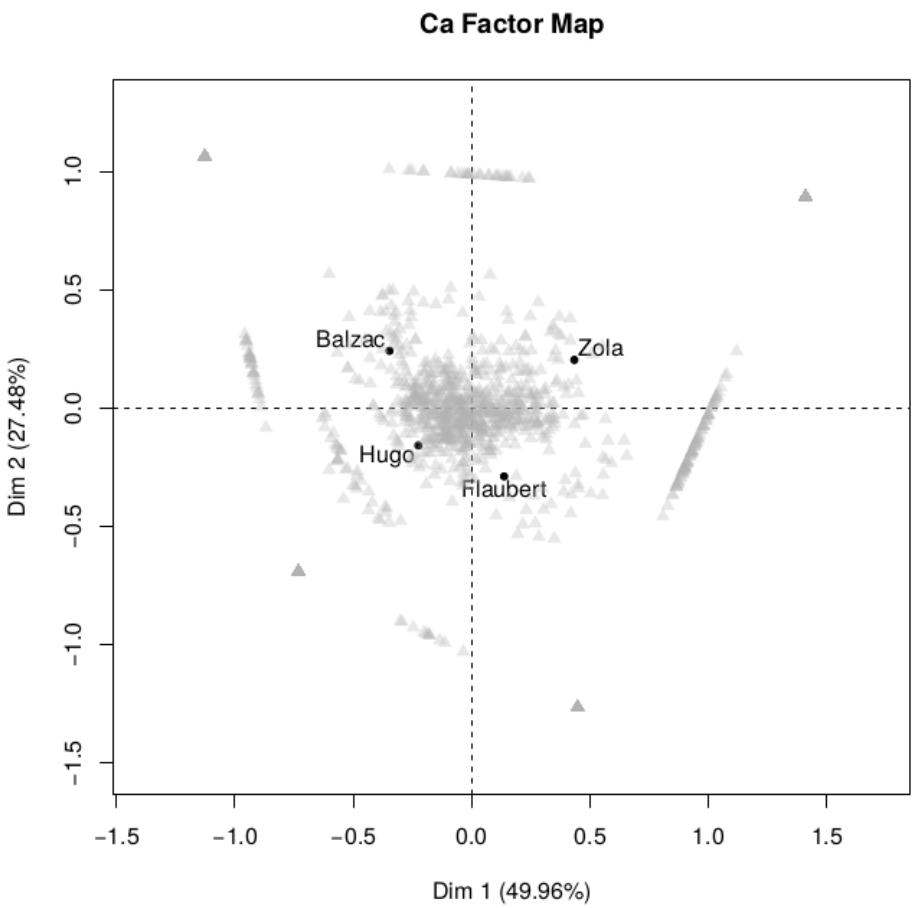

Figure 3. Plot with partially out-shadowed patterns. In the image patterns are unlabelled and represented in grey, while texts are represented in black.

Figure 3 shows us a clearer picture. Novels (here labelled with the names of their authors) are diverging on the two axes. 36

To understand the positioning of patterns and texts in a bi-plot the metaphor of a magnetic field can be used. The majority of patterns are concentrated in the centre, because they are equally attracted (represented) in all texts. On the other hand some patterns are strongly attracted (over-represented) by just one text and are repulsed by the others, positioning themselves at the extremity. Others are equally attracted by two texts only, positioning themselves somewhat in between. Moreover the force of attraction is not the same. Some patterns seem to be stronger in "pulling" a text towards them: so for instance in our case Balzac and Zola are less central. This can be interpreted in the sense that such texts have stronger characterizing features than the other two. 37

By using the figures produced by FactoMineR, we can go further, and actually remove the cloud of central patterns, while retaining for further analysis those patterns that are most contributive in terms of the displacement of the texts over the two axes. Figure 4 shows for instance a plot displaying only the 10 most contributive patterns. Moreover, by combining contribution and proximity, it is possible to select, among the patterns with high contribution, those that are nearer to one text than to the other three. 38

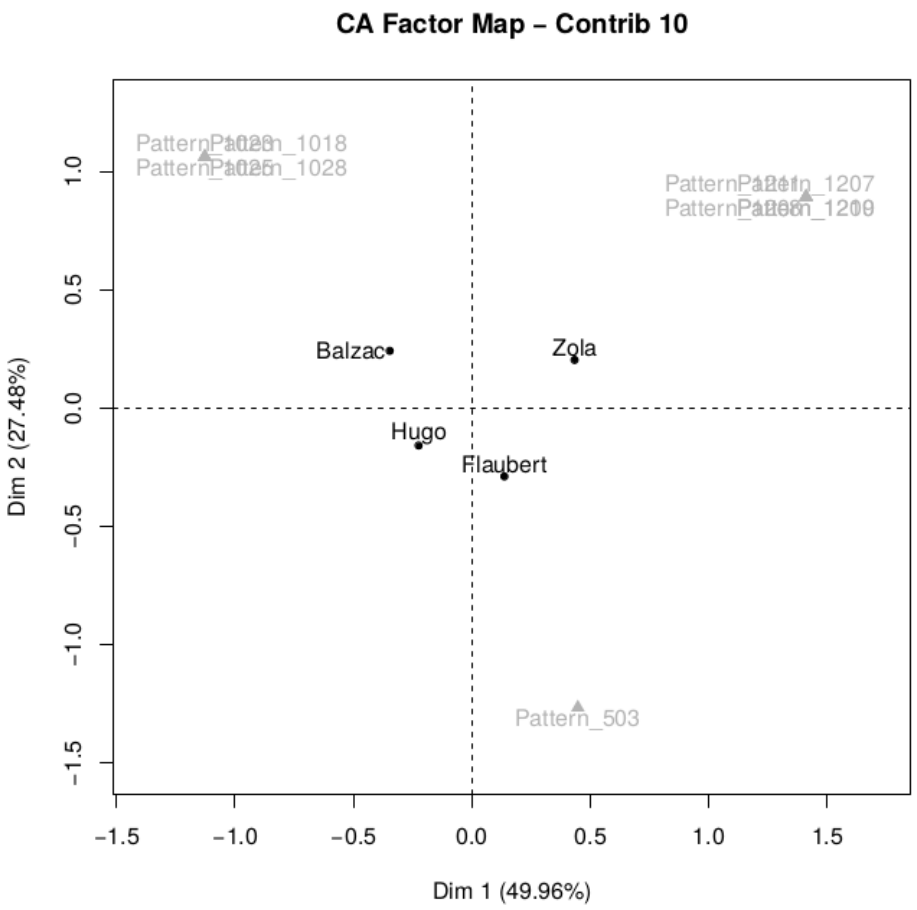

Figure 4. Top 10 most contributive patterns resulting from CA.

First of all let us see the aggregated resulting table, which provides the results of CA in a textual way. In Table 2 the ten most contributive patterns of the analysis are printed. For each one, the author they are mostly associated with is indicated. This is calculated by measuring the Euclidean distance between the position of each text and the feature, and choosing the nearest text. 39

Contrib. rank	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

Table 2. Top 10 most contributive patterns. Compare with Figure 4.

Moreover, a simple algorithm can be used to extract the top 5 most contributive patterns for each novel, namely the top 5 patterns that are more associated with each text. 40

```
for a in authors:
    for p in listOfPatternsOrderedByDecreasingContribution:
n = getNearestNovel(p)
```

```
                    add p to listOfPatterns[a]
                    if listOfPatterns[a] has length = 5:
                        exit
```

Algorithm 1: procedure to extract 5 top contributive patterns for each author. 41

By running Algorithm 1 we can extract the following lists of patterns, which we analyse in the following paragraph in detail. Notice again how the ranks of the first 5 patterns of Hugo are much higher than the others. This means that such patterns have a lower contribution. 42

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

Table 3. Top 5 most contributive patterns for Zola's Le ventre de Paris.

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

Table 4. Top 5 most contributive patterns for Balzac's Eugenie Grandet.

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

Table 5. Top 5 most contributive patterns for Flaubert's Madame Bovary.

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

Table 6. Top 5 most contributive patterns for Hugo's Notre Dame de Paris.

Analysis

As we can see from Tables 3 to 6, the analytic results confirm the intuition derived from the plot. Zola and Balzac are associated with the most contributive patterns, namely with patterns that are strongly over-used in their respective novels. Among the top 10 only one is associated with Flaubert and none with Hugo. In fact, the first five patterns in order of contribution that stands closer to Hugo (Table 6) are ranked 80 to 339, while all other novels have associated patterns in the first 10 positions. 43

Is it possible to say that Balzac and Flaubert show a more syntactically marked language? In order to do that, we need to analyse more closely the instances for each pattern, and see if the differences in the pattern frequencies are due to stylistic reasons or to other more epiphenomenal facts. The phase of analysis is very important because superficial formatting differences in the text may sometime cause errors of tagging or simply push the frequency of some insignificant patterns[9]. 44

In what follows some individual patterns among those extracted for each novel are discussed. 45

Zola

Pattern_1211 is very distinctive of Zola's *Ventre de Paris*. It occurs 188 times in this novel and close to none in the other ones. Typical instances of this pattern are sentences like: 46

> *[1211_A] Elle parut l' âme , la **clarté vivante , l' idole** saine et solide de la charcuterie ; et on ne la nomma plus que la belle Lisa .*
> *[1211_B] Il était venu de Vernon sans manger , avec des rages et des désespoirs brusques qui le poussaient à mâcher les feuilles des haies qu' il longeait ; et il continuait à marcher , pris 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 .*
>
> **example 2.**

From the analysis of such instances it becomes clear that this pattern is used in descriptions and enumerations [1211_A]; it is also very used in parenthetical phrases [1211_B] with the function of free adjuncts with adverbial function, namely modifying the verb (here *marcher*, walk). Such phrases could be rewritten as normal prepositional phrases introduced by *avec* (with), but the author shows a strong preference for this structure. 47

The same can be said of Pattern_1208 and Pattern_1209, which is often a concatenation of 1211: 48

> *[1209_A] le ventre plié **, la vue troublée ,** les pieds comme tirés , ...*
> *[1208_A] le ventre plié **, la vue troublée , les** pieds comme tirés , ...*
>
> **example 3.**

Pattern_1207 also seems to be an expression of the same preference of Zola's for implicit clauses to modify the verb and express manner. 49

Notice how all these patterns contain punctuation elements, often commas. The style of Zola is dry, effective, with frequent use of parentheticals rather than explicit forms. 50

> *[1207_A] Il **marchait , dormant à** demi , dodelinant des oreilles , lorsque , à la hauteur de la rue de Longchamp , un sursaut de peur le planta net sur ses quatre pieds .*
>
> **example 4.**

Instead the second most important pattern for *Le ventre de Paris* - Pattern_1210 - is associated with a very specific linguistic structure, namely with the modification of proper names, mostly those of women. Here the feature identified seems more lexical than syntactical, probably Zola is trying to recreate the jargon of the Parisian populace, with people often being called with nicknames. 51

> *[1210_A] la petite Pauline ...*
> *[1210_B] la belle Normande ...*
>
> **example 5.**

The analysis thus seems to be in line with the received idea of Zola's intentionally choosing a realist style that should represent the reality of with authenticity and in an objective way. 52

Balzac

A first look at *Eugenie Grandet*'s patterns tells us that Balzac has a somewhat different style, with a preference for 53

verbal structures and preposition, thus of explicit structures rather than implicit ones.

The first pattern is strongly associated with dialogical structures, which are very frequent in this work: 54

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

example 6.

The same can be said of Pattern_1016, which is used mostly to (post-) introduce direct speech: 55

*[1016_A] Bonjour , Grandet **, dit -il au** vigneron*
*[1016_B] Mademoiselle **, dit -il** à Eugénie*

example 7.

Pattern_1025 is associated with two structures, both subordinate infinitives, with an explicative value (1025_A) or to describe co-occurring events (1025_B). 56

*[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] A cette observation , le notaire et le président dirent des mots plus ou moins malicieux ; mais l' abbé les regarda d' un air fin et résuma leurs pensées **en prenant une pincée de tabac** , et offrant sa tabatière à la ronde : Qui mieux que madame , dit -il , pourrait faire à monsieur les honneurs de Saumur ?*

example 8.

Pattern_1023 is used in phrases containing proper names, often place names in the function of modifiers of a noun. 57

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

example 9.

Pattern_1018 shows a main transitive verb with its object and an implicit subordinate phrase. Like Pattern_1025, it is used to better specify actions or events. Notice that basically this type of pattern constitutes the counterpart to those used by Zola, who prefers the verbless forms of predicate modification. With a little imagination we could write Zola's version of [1018_B] as something like "Grandet, la bouche fermée, regarda sa fille". 58

*[1018_A] Charles **tendit la main en défaisant** son anneau*
*[1018_B] Grandet **regarda sa fille sans trouver** un mot à dire .*

example 10.

Thus Balzac's style is more verbose, more explicit. The use of preposition to introduce phrases or clauses is important to highlight the relationship between head and modifier. It makes sentences less difficult to interpret. Balzac is considered the father of realism, but he aimed at a broader and more popular audience than Zola's, (for financial reasons as well as for artistic ones). His style reflects possibly this necessity, as well as the time constraints of his immense production. 59

Flaubert

All of *Madame Bovary*'s patterns contain punctuation. The five patterns all capture the same phenomenon, notably the fact that Flaubert's punctuation allows the comma to intervene before the conjunction as in: 60

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 mettait à lui parler de ses chagrins : il l' oubliait , il en aimait une autre !*

example 11.

Patterns indicating a certain style in punctuation should always be taken with a grain of salt, since punctuation in the edited version does not always reflect the choice of the author, but may be submitted to editorial guidelines. Nevertheless Mangiapane [Mangiapane 2012]highlights the rhythmical rather than functional role that punctuation has in Flaubert. Indeed, from the rhythmical point of view, in the given example the commas mark the breathing pause that is present before as well as after the conjunction “et”. 61

Hugo

As was presented before, Hugo’s work is less syntactically marked than that of the others. The patterns that do show some overrepresentation in *Notre Dame de Paris* are simple syntactical structures rather than complex ones. 62

Two of these patterns (Pattern_31, Pattern_27) are absent in Zola and Balzac, but are shared with Flaubert. Pattern_31 is the longest. It seems to be used mostly in descriptions of places, which are very rich in the historical novel of Hugo, and help the reader to enter into the world of medieval Paris. 63

[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 vagues sur ses deux pentes latérales , le grand escalier , dis -je , ruisselait incessamment dans la ***place comme une cascade dans*** *un lac ..*

example 12.

Pattern_27 is often used in structure subordinate clauses that show a preference for demonstrative pronouns to underline situations. 64

[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 rebeller contre monsieur le bailli , par impatience d' entendre son ouvrage !*

example 13.

Pattern_520 and Pattern_833 are shared with other authors, though slightly overrepresented in Flaubert. Here too the punctuation variant found in Flaubert emerges, though not as strongly. 65

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

example 14.

Pattern_190 finally is used in comparisons and descriptions. 66

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

example 15.

By this analysis, the style of Hugo seems to emerge as full of lively descriptions, simple, personal, engaging, and popular just as we know it from literary tradition. 67

Conclusion

We presented a detailed outline of a methodology for the extraction of syntactic patterns in texts and for the measurement of their interestingness in a corpus. The results suggest that this methodology bears substantial promise as a hermeneutical instrument offered to experts in the literary domain to investigate style in texts and to extract interesting stylistic features. For this reason both EREMOS, namely the tool required to perform the pattern extractions, and the R scripts used to perform correspondence analysis on such extractions have been released 68

and made available to the community[10].

69 Other interestingness measures besides correspondence analysis have been tested, which are based on the distribution of the patterns in different parts of the same text [Boukhaled 2015a], as well as on the comparison of each text to a reference corpus [Boukhaled 2015b]. Such measures give partially comparable results to the method presented here, which is based on the cross-comparison of texts by means of correspondence analysis and on the ranking of interesting patterns by contribution. However the latter is particularly useful in that it not only extracts the most interesting patterns for each text, but also provides information on shared patterns and on the fact that some texts are less marked, or less characterised than others.

70 New experiments using this methodology on a number of other texts have been carried out; in particular parallel researches have studied the syntactical aspects of characterization in Molière's plays ([Frontini 2015a], [Frontini 2015b], [Frontini Benard 2015]). Furthermore, we have tested its application to tasks of authorship attribution, by contrasting a collection of original and apocryphal short stories attributed to Robert Challe [Frontini 2015]. The development of the tool came about through close contact with experts in literary criticism who were asked to give advice on the extracted patterns. Results show how features that have been independently identified by literary scholars as characterising a text or a group of texts can be automatically extracted with the proposed method.

71 A further interesting domain of application could be the study of *pastiches*, namely texts that imitate and/or satirise the style of an author or a genre. Finally and most importantly, further research will be necessary to compare the kind of stylistic traits emerging from the use of syntactic patterns for correspondence analysis with those emerging when using lexical features, in order to see what differences these two linguistic dimensions can capture on the same group of texts.

Acknowledgments

72 This research was supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference the ANR-11-IDEX-0004-02 and by an IFER “Fernand Braudel” Scholarship awarded by the Fondation Maison Sciences de l'Homme. We thank the anonymous reviewers of GDDH 2015 and DHQ for their helpful comments.

Notes

[1]An n-gram is a sequence of n contiguous elements, in this case of n contiguous part of speech tags.

[2]For a list of syntactic features that can be extracted from parse trees see [Dell'Orletta 2011].

[3]The same consideration holds for other similar exploratory approaches, as was pointed out by Franco Moretti in a recent lecture at the Sorbonne, Paris, 2015.

[4]Most well known stylometry tools offer methods for dimensionality reduction and plotting (Signature, JGAAP, TLAB, stylo for R, ...) but they often work on word count matrixes only, or they do not allow for finer result manipulation, which is the advantage of the method presented in this paper. For a general introduction on the use of PCA to stylistics see [Binongo 1999]; for an example among many others of the application of MDS to authorship attribution see [López-Escobedo 2013].

[5]http://tlab.it/en/allegati/esempi/inaugural.htm. We thank the TLAB developers for granting permission to use their plot in this publication.

[6]EREMOS [Boukhaled 2016] was developed by the ACASA team at LIP6 in Paris, and is available via a web interface at http://eremos.lip6.fr.

[7]Current research is ongoing with domain experts to identify which is the most useful configuration for EREMOS.

[8]Patterns are assigned an identifier, as it would be difficult to deal with their full description in CA plots.

[9]Thus in the first phases of an analysis, CA can help even in checking that corpora are correctly prepared and normalised.

[10]Please visit http://eremos.lip6.fr and https://github.com/francescafrontini/CAforEREMOS for further information.

Works Cited

Agrawal 1995 Agrawal, R., Srikant, R. "Mining sequential patterns." In: *Proceedings of the Eleventh International*

Conference on Data Engineering. Presented at the Proceedings of the Eleventh International Conference on Data Engineering, (1995), pp. 3–14.

Benzécri 1977 Benzécri, J.-P. "Histoire et préhistoire de l'analyse des données. Partie V: l'analyse des correspondances." *Cahiers de L'analyse Des Données*, 2(1) (1977): 9–40.

Binongo 1999 Binongo, J.N.G., Smith, M.W.A. "The application of principal component analysis to stylometry". *Literary and Linguistic Computing* 14 (1999): 445–466.

Boukhaled 2015a Boukhaled, M. A., Frontini, F., Bourgne, G., Ganascia, J.-G., 2015. "Computational Study of Stylistics: A Clustering-Based Interestingness Measure for Extracting Relevant Syntactic Patterns." *International Journal of Computational Linguistics and Applications* 6 (2015): 45–62.

Boukhaled 2015b Boukhaled, M. A., Frontini, F., Ganascia, J.-G. "Une mesure d'intérêt à base de surreprésentation pour léxtraction des motifs syntaxiques stylistiques." In: *Actes de La 22e Conférence Sur Le Traitement Automatique Des Langues Naturelles (TALN2015)*, (2015), pp. 391–396.

Boukhaled 2016 Boukhaled, M.A., *On computational stylistics: mining literary texts for the extraction of characterizing stylistic patterns*. PhD thesis. Université Pierre et Marie Curie (UPMC), (2016).

Burrows 2002 Burrows, J.F. "Delta: a Measure of Stylistic Difference and a Guide to Likely Authorship". *Literary and Linguistic Computing* 17 (2002): 267–287.

Burrows 2012 Burrows, J.F., Craig, H. "Authors and Characters." *English Studies* 93 (2012): 292–309.

Craig 2004 Craig, H. "Stylistic analysis and authorship studies". In: Schreibman, S., Ray, S., Unsworth, J. (eds.), *A Companion to Digital Humanities*. Blackwell, Oxford, (2004), pp. 273–288.

Craig 2009 Craig, D.H., Kinney, A.F., 2009. *Shakespeare, computers, and the mystery of authorship*. Cambridge University Press (2009).

Dell'Orletta 2011 Dell'Orletta, F., Montemagni, S., and Venturi, G. "Read-it: Assessing readability of Italian texts with a view to text simplification." In *Proceedings of the second workshop on speech and language processing for assistive technologies*, Association for Computational Linguistics (2011), pp. 73-83.

Frontini 2015 Frontini, F., Boukhaled, M. A., and Ganascia, J. G., 2015. "Analyse et extraction des motifs syntaxiques dans la prose de Robert Challe et de ses apocryphes". Presented at the *Robert Challe: approches numériques des questions d'auctorialité* Workshop in Paris, 28/03/2015. http://obvil.paris-sorbonne.fr/sites/default/files/projets /analyse_motifs_syntaxiques_if_et_apocryphes.pdf (last visited on 25/09/2016)

Frontini 2015a Frontini, F., Boukhaled, M. A., and Ganascia, J. G. "Linguistic Pattern Extraction and Analysis for Classic French Plays". Presented at the *ConSciLA* Workshop, Paris 16/01/2015 http://lipn.univ-paris13.fr/ charnois/conscilaGenres/resumes/frontini.pdf(last visited on 25/09/2016)

Frontini 2015b Frontini, F., Boukhaled, M.A., Ganascia, J.G. "Molière's Raisonneurs: a quantitative study of distinctive linguistic patterns." In: *Corpus Linguistics 2015 - Abstract Book*. Presented at the Corpus Linguistics, Lancaster, UK, (2015), pp. 114-117.

Frontini Benard 2015 Frontini, F., Benard, E. "The Syntax of Stage. Studying Linguistic Patterns in Molière". Presented at the Göttinger philologisches Forum, Göttingen 03/12/2015 https://www.uni-goettingen.de /de/empfehlung-»the-syntax-of-stage«-vortrag-von-francesca-frontini--elodie-bénard-am-3-dezember-2015-crc-texstrukturen/525494.html (last visited on 25/09/2016)

Greenacre 2007 Greenacre, M. *Correspondence analysis in practice*. CRC press (2007).

Grzybek 2014 Grzybek, P. "The Emergence of Stylometry: Prolegomena to the History of Term and Concept" In: Kroó, Katalin; Torop, Peeter (eds.), *Text within Text - Culture within Culture*. Budapest, Tartu: L'Harmattan, (2014), pp. 58-75.

Husson 2011 Husson, F., Josse, J., Le, S., and Mazet, J. "FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R", R package version 1.24 (2011).

Jockers 2013 Jockers, M.L. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press (2013).

Lebart 1998 Lebart, L., Salem, A. and Berry, L. *Exploring textual data* (Vol. 4). Springer Science and Business Media (1998).

López-Escobedo 2013 López-Escobedo, F., Méndez-Cruz, C.-F., Sierra, G., Solórzano-Soto, J. "Analysis of Stylometric Variables in Long and Short Texts". *Procedia - Social and Behavioral Sciences*. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013) 95 (2013): 604–611.

Mahlberg 2013 Mahlberg, M. *Corpus Stylistics and Dickens's Fiction*, Routledge advances in corpus linguistics.

Routledge, New York (2012).

Mangiapane 2012 Mangiapane, S. "Ponctuation et mise en page dans Madame Bovary : les interventions de Flaubert sur le manuscrit du copiste." *Flaubert. Revue critique et génétique*. 8 (2012) http://flaubert.revues.org/1883?lang=en (last visited on 25/09/2016)

Moretti 2005 Moretti, F. *Graphs, maps, trees: abstract models for a literary history*. Verso, London/New York (2005).

Quiniou 2012 Quiniou, S., Cellier, P., Charnois, T., Legallois, D. "What about Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics?" in: Gelbukh, A. (ed.), *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science. Springer, Berlin Heidelberg, (2012), pp. 166–177.

Ramsay 2011 Ramsay, S. *Reading machines: Toward an algorithmic criticism*. University of Illinois Press (2011).

Rybicki 2011 Rybicki, J., Eder, M. "Deeper Delta across genres and languages: do we really need the most frequent words?" *Literary and Linguistic Computing* 26:3 (2011): 315–321.

Schmid 1994 Schmid, H. "Probabilistic Part-of-Speech Tagging Using Decision Trees." In: *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK (1994).

Schmid 1995 Schmid, H. (1995). "Improvements in Part-of-Speech Tagging with an Application to German". *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland (1995).

Stein 2003 Stein, A. "French TreeTagger part-of-speech tags". (2003) http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html (Last accessed: 2015-03-25).