Literature studies in *Literateca*: between digital humanities and corpus linguistics

Diana Santos
Professor
Institutt for litteratur, områdestudier og europeiske språk,
Universitetet i Oslo & Linguateca

1. Introduction

The availability of more and more literary texts in electronic form has opened up the possibility to do research on large quantities of data using computer-intensive methods, leading to proposals variously entitled Macroanalysis, Computational Literary Genre Stylistics, or more poetically "tackling the Great Unread". Their proponents are literary scholars, who (as a class) fell under the spell of computers rather late.[1]

Apparently, the early defenders of using computers for text understanding, traditionally called corpus linguists, have recently been moved aside by "Big Data" fans (ranging from literature scholars through sociologists to media professionals), who use much larger quantities of text and quite shallower tools or resources (such as Google N-grams). Big Data supporters seem to prefer quantity over quality (or, put in a different way, they had to surrender to quantity).[2]

1. This is a statement in need of qualification: In fact, the first area of humanities scholarship that embraced statistics was literature – or the first non-scientific area that statisticians looked into, a fact which never ceases to surprise students of statistics for humanities. Disputed authorship, the computation of style... see Yule (1944), Kenny (1982). But somehow it was not – never – part of mainstream literature studies.
2. It is probably unnecessary, in the present context, to remind the readers that Google books – or any other project of massive digitization – includes a large number of badly encoded or wrongly meta-tagged material, and that there are anyway several different levels of encoding,

This is an obvious paradigm shift. Or is it? I will try to challenge this view.

In the early days of corpus linguistics, corpus preparation and choice of materials to include was key. But now a new breed of scholars has emerged interested in trends only perceivable at a much larger scale – which no longer can be humanly revised (such as the web, or the shallowly processed web). A vocal defender of this perspective is Kilgarriff et al. (2007), as his title "BNC Design Model Past its Sell-by" eloquently conveys.

While this could be swept under the carpet as yet another instantiation of the "scruffies vs. neats" debate, which occurs in every area of knowledge (see Goble and Wroe (2004) for an inspiring presentation), it can also be understood as different traditions fighting for the same space. Or fighting, at least, for the primacy of their methods for the investigation of the same target – literature, in this case.

While there's more than one way in any branch of knowledge, some suggest a middle way (or hybrid approach). Big data literary studies are no exception: a back-and-forth travel between the two extremes (of distant[3] and close reading) has been proposed by macroscope theorists (Tangherlini, 2013), inspired by parallel proposals in the natural sciences (Börner, 2011). I believe that, to play in this field, one needs to be able to integrate insights and tools from different areas, or at least be aware of them. But not everything can be united, and it is important to be aware of possibly irreducible basic disagreements.

Or, one might ask, is this conflict just apparent, and the only thing that really changed is the size of the material? Consider Biber (1985), a pioneer in corpus linguistics, who was one of the first to apply dimensionality reduction methods to language data: he used small corpora. True, he started as an outsider, but began to gain more and more acceptance when times changed (and volumes of data increased). While in the '90s most corpus linguists ignored statistical methods, now it is almost impossible to get published without applying them. True, one can take the *dejá vu* stance and invoke the pendular fashion alternating between empiricism and rationalism, but considerable work using complex statistical techniques has been advanced lately, and that should not be ignored.

Coming from another discipline altogether, literary scholars adhering to digital humanities (DH) have happily embraced big data techniques from the

as well as a number of crucial decisions to take in any digital project.

3. As coined and argued for in Moretti (2000).

start. Could there be ways to join forces and interact? In this paper, I attempt to do so, inspired among others by the discussion in Mambrini et al. (2012). I will describe some work done in *Literateca* to study literature in Portuguese, drawing from knowledge in both fields. It was Christian-Emil Ore who brought me to digital humanities processing of literature, therefore I dedicate to him this report of my initial attempts. But first I will sketch briefly the state of the art as I see it.

2. A brief overview of literary questions approached with the help of computers

Many of the techniques currently being used for pure DH studies concerning literature[4] (that is, dealing with literary questions and not only with literary material) have migrated from other fields. "Topic modelling" comes from information retrieval, as Jockers (2013: 22) so candidly reports; zooming in and out, from hard science visualization needs, as acknowledged by Tangherlini (2013); and other approaches through even more indirect routes such as digital history (Hoof, 2013) and gender studies (Smith et al., 2014).

The scientific foundation for the digital techniques themselves, techniques that deal with large amounts of something (text, or other entities), now called "big data", comes from statistics, a field which has had surprisingly little appeal for corpus linguists until recently (but see Oakes (1998) for an early overview of statistics in corpus linguistics). But what will mostly concern me in this overview is the questions that interested scholars, not specifically the technologies they happened to use.

2.1 Genre

Let us start with the question of genre. Electronic corpora have almost since their beginning incorporated some (external) genre characterization, see e.g.

4. In this paper I will not cover new kinds of literature, for example digital literature. I will limit myself to the study of "traditional" literature, written and "frozen" in a published text, and usually by a single author.

the Brown corpus,[5] or Sinclair's definition of a corpus.[6] However, we may say, these are linguist's definitions of genre.

Biber for one (see Biber (1985, 1988) and Biber & Gray (2010)) has been influential in applying statistical methods to understand genre. But his interests did not concern specifically or even mostly fiction: he was looking at oral vs. written, familiar vs. distant, medium, etc. When later writing a corpus-based grammar with Stig Johansson and others (Biber et al., 1999), fiction was used as one of the four registers (conversation, fiction, newspaper language, and academic prose) with which to describe English grammar, but that was that as far as literature was his target. So, and to the extent that "genre" is mentioned at all in corpus studies, it has mainly been used to address different kinds of writing or speaking, most of it far removed from fiction.

On the other hand, the bulk of work on genre in the digital humanities is concerned with literary genre, thanks to the growing availability of electronic literary text collections; cf. the overview of Irish American literature of Jockers (2013), or the analysis of French Classical and Enlightenment drama of Schöch (2017), done with the help of topic modelling. Also, Italian contemporary literature (Cortelazzo et al., 2012), or Scandinavian late nineteenth century (Broadwell and Tangherlini, 2017) have been analysed for different purposes.[7]

2.2 Authorship

One of the most difficult problems so far is the influence of the author – how important are the writers themselves? How can one distinguish writers' style from genre itself? How much of a genre is created by one or a set of influential writers? This has been acknowledged, e.g. by Jockers (2013: 70, and 99–104) and in Schöch's (2017) conclusions.

5. II. Imaginative Prose (126 samples) – K. General Fiction; L. Mystery and Detective Fiction; M. Science Fiction; N. Adventure and Western Fiction; P. Romance and Love Story; R. Humor (from http://clu.uni.no/icame/brown/bcm.html accessed 27 May 2017)
6. "[…] selected according to external criteria to represent, as far as possible"... (Sinclair, 2005).
7. Using, respectively, 106 novels from the nineteenth century (Jockers), 391 plays published between 1630 and 1789, amounting to 5.6 million tokens (Schöch), 92 narrative works by 33 authors published between 1941 and 209, amounting to 7.8 million words (Cortelazzo) and 85 works (Broadwell).

The kind of questions in genre categorization could be asked about one author or group of authors. How can we get at their style, and how does this differ from the themes they write about?

There is and has been a huge number of scholars interested in the (sub)field of authorship attribution and dating – see Oakes (2014) for a recent overview, and Zhao and Zobel (2007) for a classification study with 634 works in English. One can also find many researchers concerned with literary style (Mahlberg, 2015) for reasons other than attribution. However, most studies have dealt with corpora of one (possibly two) author(s) and not with literary corpora comprising many authors. For example, Steinberg (1973) studied different characters in James Joyce, and Schmidt (1980) looked at adjective use in direct speech by feminine and masculine characters in Jane Austen's novels.

No criticism is implied: it is hard enough to process and deal with one author's material, not least in corpus studies. But it is surely rewarding to have a bird's eye view of a set of possibly stylistically relevant properties for a large set of authors, as distant reading proponents claim.

Let us present, for English, the CLIC Dickens project, whose goal is to do corpus stylistics, "lead[ing] to new insights into how readers perceive fictional characters",[8] in particular by developing a web interface at http://clic.bham.ac.uk/ to access annotation of 15 novels, ca. 3.9 million words. One of the research questions addressed by the project is the characterization of suspended quotations: whether they are correlated with body language, how they reflect authorial position, and how the reporting verbs are distributed in those quotations (Mahlberg et al., 2013).

This project is a good example of the use of more detailed corpus linguistics techniques and analysis for literary purposes, but it is also good evidence that the size of the materials involved is several orders of magnitude smaller than what big data supporters require.

While such a project has obvious similarities to *Literateca* (which will be described presently), not least by giving access to annotated corpora over the web, it is worth mentioning that the specific subject matter and the way it is operationalized is very much specific to the English language. The way direct speech is encoded in English is not universal, quite the contrary, as discussed in Santos (1998) and Freitas et al. (2016).

8. http://www.nottingham.ac.uk/research/groups/cral/projects/clic.aspx (accessed 26 May 2017).

This could be a healthy reminder that not only do national literatures differ, the linguistic matter that authors can mould to develop their own style can also be very different. Thus, the argument could proceed, some stylistic devices and singularities in one language cannot necessarily be generalised to world literature.[9]

2.3 Plot

One can also study and investigate how each work works, which characters appear when, and how they are connected. Moretti (2011) makes the case for plot networks in a literary context (see also Ardanuy & Sporleder, 2014), while others before him were concerned with "standard" divisions in the building of particular genres, for example Propp's (1928) work on folklore tales and more modern attempts using computational linguistics techniques (Volkova et al., 2010).

From a different perspective, topic modelling has also been suggested as a probe to understanding both the thematic structure of literary works and the underlying author concerns, as Jockers (2013) attempted with his inspired example of Melville and Austen's literary cocktail.

Using sentiment analysis, Nalisnick & Baird (2013) tried to automatically model character-to-character sentiment, going therefore further than just capturing themes and temporal lines. One could call this automatic close reading, contrasting with the distant reading of Smith et al. (2014), who used corpus analysis to show sexism in films, counting the lines of male versus female characters.

Other kinds of features that can be considered attributes of the plot or the characters can be recalled: Steinberg analysed the number of omniscient narrator sentences in Joyce's *Ulysses* in the chapters concerned with each character (and in those which displayed stream-of-consciousness), in order to give "the reader different impressions of the personalities of Stephen and Bloom" (Steinberg, 1973: 103). Halliday (1971) in turn showed how marked differences in transitivity in two different parts of one novel by Golding conveyed meaning

9. This is an interesting literary argument that goes against Moretti's agenda of world literature as a paradigm shift. A similar case can be made of writers like Soseki in early 20th century Japan, trying to innovate Japanese literature by importing and adapting formal features of the English novel into Japanese (Auestad, 2017).

at several levels. Finally, Preminger and Fludal (2016), in their automatic recommendation experiments, studied pace as a property of books.

2.4 Intertextuality, type of author and presence of the reader

Surprisingly, the apparently easiest thing to measure, namely the repetition of names, themes, and lines in posterior works influenced by others, which we may call obvious marks of intertextuality, has so far received less attention. Conversely, the position and conceptualization of the author of the text, the narrator, and the (imagined) reader have been much more studied (see McMurry (2015) for an interesting take on these issues) and have in fact been identified in corpus studies. The already cited Mahlberg et al. (2013), Steinberg (1973) and Stubbs (2005) are examples of this.

But although repeated patterns should be easy to ascertain, and text reuse and plagiarism detection have been hot subjects for a long time in other quarters, there are few works on detecting influence that I know of. I am only aware of a few. Stubbs (2005) corpus linguistic study on Joseph Conrad's *Heart of Darkness* points out several lexical items that stem from the King James Bible, Dickens and Jules Verne. Oppenheim (1988), on the other hand, attempts to ascertain Joyce's influence in Hemingway as far as short stories go.

Maybe the explanation is simple: to investigate intertextuality one needs a large corpus (distant reading), whereas to look at the presence of their own authors, narrators and readers, closely reading one or a few texts is enough.

3. Background: Gramateca

Gramateca is a Web environment to provide scholars around the world with access and tools for studying Portuguese grammar, whose philosophy and purpose has been described elsewhere (Santos, 2014b). It is a subproject of Linguateca, a network for fostering progress in the computational processing of the Portuguese language. One of the resources that Linguateca offers researchers and developers is a large variety of annotated corpora. A subset of the material consists of lusophone literature, that is, works written in Portuguese by native speakers from Angola, Brazil, Mozambique, Portugal, etc. –

although Gramateca gives access to a much wider range of different genres and text types.

As far as the "literature subcorpus" goes, the basic material comes from disparate sources: on the one hand, the Vercial, Obras, Tycho Brahe (Galves & Faria, 2010) and Colonia (Zampieri and Becker, 2013) projects, which made full texts (in the public domain) available; on the other, some parallel corpora developed in Linguateca, composed of excerpts from (usually) modern texts.[10] For more information on the individual corpora, please see the links in the references.

All corpora have been syntactically parsed with PALAVRAS (Bick, 2000), while semantic annotation was done in-house in a two-step rule-based process. This work has spanned several years and several subprojects and will not be described here for lack of space, but the interested reader is directed to Santos (2014a) for an overview.

Table 1 provides a quantitative summary of the literary data as of December 2017 (note that the total is not the sum of the different rows because authors and works may appear in more than one corpus). It should also be noted that the texts were chosen (by the respective corpus compilers) because they were considered either canonical texts or well written or because they had been translated. They are not necessarily fictional. Sermons, edited letters, travel reports,

Table 1. Quantitative data on the literary works originally written in Portuguese available from Linguateca as of 12 January 2018. Literateca has only one version of each work. "Tokens" refers to words and punctuation marks.

Corpus	Authors	Works	Works in *Literateca*	Size in million tokens
Vercial	54	326	326	14.7
OBras	19	57	57	2.7
Tycho Brahe	51	62	57	2.7
Colonia	53	91	45	5.0
PANTERA	28	51	44	0.25
Total (unique)	159		529	

10. There is a significantly higher quantity of literary text in corpora available from Gramateca, such as the NILC corpus, the ECI-EBR corpus or the Corpus Brasileiro, but they do not have associated metadata and so they will not be used for the present article. Foreign literature (in English and Norwegian) translated into Portuguese is not being used either.

history treatises and even political speeches are included. A more fine-grained genre description will have to be discussed in detail elsewhere (see Santos (2016)); however, every corpus has genre metadata, so it is possible if necessary to reduce the number of works to consider for analysis.

Figure 1 displays a chronological view of the texts written after 1800. I show it here because it vindicates the "distant reading" approach: works from the period 1930-1970 are conspicuously missing. Not because there was a shortage of writers in Portuguese in those years, but because of copyright issues, coupled with different interests of different corpus compilers. This is obviously something that has to be solved for Portuguese corpus linguistics for the study of literature, as the graphical visualization shows.

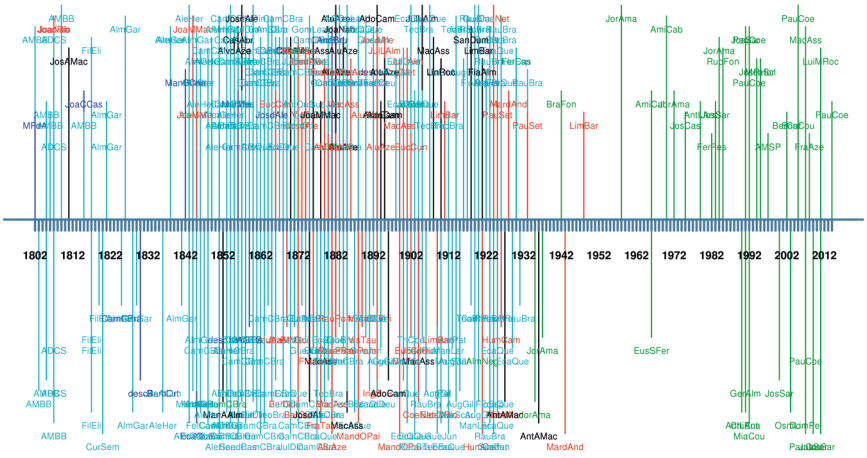

Figure 1. Chronology of the texts published after 1800, indexed by the names of their authors. Colour coding indicates corpus provenience, green indicating PANTERA.

4. Exploring the space of possible features

First I explored the annotation of the corpora, in order to see whether those features would cluster the texts into something meaningful. I tried several clustering techniques, as well as investigated how the features were related, inspired by Baayen (2008) and using R (R development team, 2008).

Details of which categories were annotated and how have been reported elsewhere in detail (see Silva and Santos, 2012; Santos et al. 2011; Freitas, 2015; Mota and Santos, 2015; Freitas et al., 2016). Here it suffices to say that one could (more or less reliably) count how many words existed for a particular colour space, how many words were employed to describe a particular body region, or a particular emotion, or a particular kind of clothes. Also, reporting verbs could be counted, as well as specific syntactic features such as the number of relative clauses, *that*-clauses, and finite clauses, or the use of negation, long dash, first person singular, feminine personal pronouns, etc.

The first version of the cluster analysis (not shown here) presented a sensible grouping of most features, but I was surprised to see that the emotion group labeled FURIA "wrath" had been assigned a class of its own. Closer investigation identified the consistent mis-annotation of the word *bravo*.[11] This shows that these "distant" explorations may be useful also to correct the (semi-automatic) annotation by requesting specific cases of "close" reading. After solving this problem, we got a more sensible, although not necessarily fully predictable, clustering of the different features shown in Figure 2. Closeness of the features indicates that they are correlated in the material. I have selected six clusters, from left to right: the first contains mainly clothing, with two internal body parts and some rare colours. The second contains almost all syntactic properties, together with desire and hope. This makes sense, since most of the cases of hope or desire are expressed in subbordinate clauses. The third cluster includes most emotions, the few remaining ones being found in a nearby cluster, namely the fourth. The fifth cluster includes the most important colours, and the sixth yet another group of less important ones (golden, other colours, rose and purple). The most surprising result in this clustering attempt is the fact that ingratitude is joined with two tense features typical of oral speech, namely the progressive and the present perfect (PPC). I have no explanation for this yet.

Doing a correspondence analysis with all 114 features, there was a clear outlier: the minutes of slave meetings in Brasil.[12] Removing them, we are left

11. The word *bravo* has several meanings. In addition to a compliment exclamation, conveying admiration, it means courageous (brave), and wild (for plants or animals). It also means, in the Brazilian variety of Portuguese, in modern language only, angry. By correcting the annotation, wrath stopped being a weird feature and clustered nicely with the other emotions.
12. One might claim that this was obvious from the beginning, but in fact one of the advantages of automatic methods is to confirm "obvious" things. On the other hand, it is important to stress that other minutes in the material were not classified as outliers, and the same was true of many other history texts.

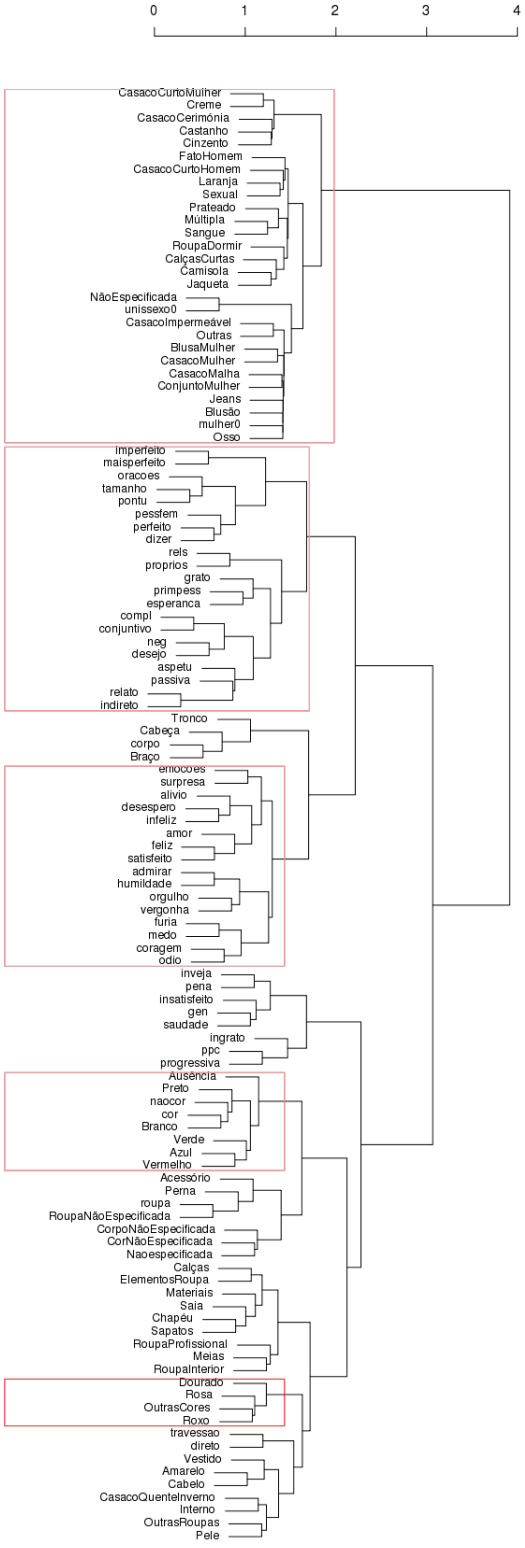

Figure 2. Clustering the 114 linguistic features that were used for the present paper, using Kendall correlation and complete divisive hierarchical clustering.

with the situation depicted in Figure 3. In red, we show only the most discriminative features, namely the emotions gratitude and ingratitude (grato and ingrato), the use of proper nouns, relative clauses, reference to sexual body parts, the golden colour (Dourado), use of first person, of a particulat tense (PPC), and of indirect speech, among others. They have to be studied further in order for us to understand why they stand out.

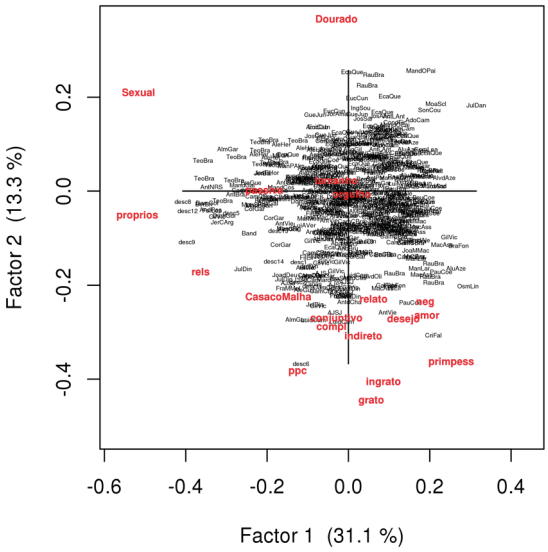

Figure 3. Correspondence analysis of 528 works with the most discriminative features in red: every work is marked with its author in grey.

Figure 4 shows a factor analysis with three factors, in which we see that the features used appear at different places in the plane.

Finally, performing a linear discriminant analysis (LDA), two new works stand out. A treaty on virtue and vice by Matias Aires in 1705, atypical in the abundant references to emotions; and *Peregrinação* by Fernão Mendes Pinto, an adventure travel book comparable to Marco Polo's, and which is unique in Portuguese literary history (written in 1569–1570, but first published in 1614). I removed both before (re)creating figure 5. If you are not able to make sense of the authors' names, at least note that works by the same author stand reassuringly near each other.

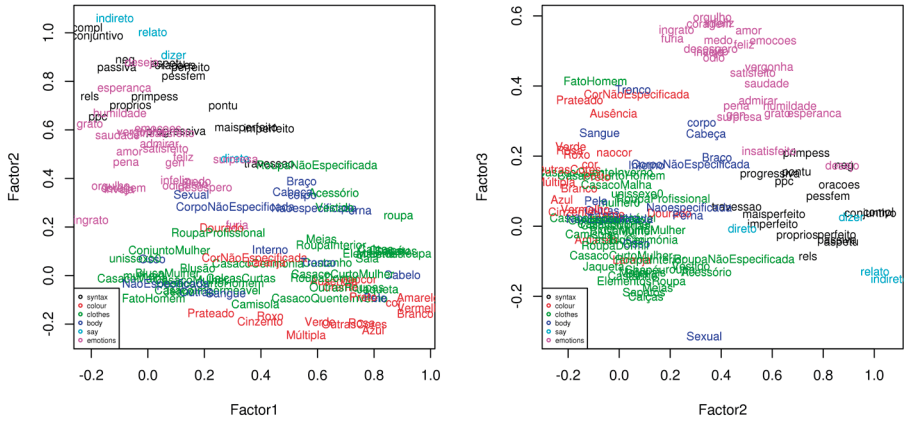

Figure 4. Factor analysis with promax rotation.

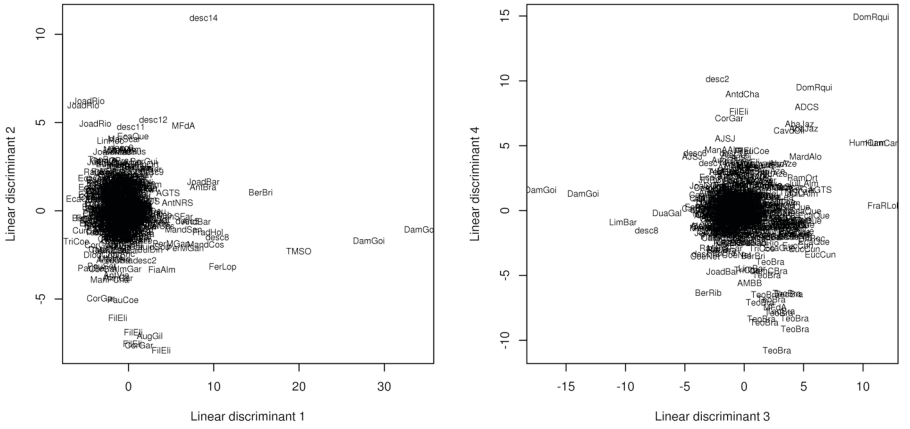

Figure 5. Texts across the four first dimensions of an LDA-analysis of 526 texts.

5. Addressing the reader

A relatively simple study done with basic corpus linguistics methods is looking at how different writers employ the word *reader* (that, incidentally, in Portuguese is marked for genre: *leitor* and *leitora*, which can be pluralized as well). From an initial quantitative overview, that showed that use of these words was mainly in the 19th century novel – although some older poetry authors used them copiously in their introductions or prefaces, it was necessary to do a close

reading to separate characters reading in the text as part of the plot from readers of the text, addressed by the author. Not unexpectedly, the most reader-communicative authors are 19th century novelists, namely Machado de Assis, Camilo, Júlio Dinis, Alexandre Herculano, Raul Pompéia and Almeida Garrett. Only the three first frequently addressed their female readers, though.

6. Topic modelling

Since topic modelling is such a hot topic in current digital humanities approaches to literature, I installed the Mallet package (McCallum, 2002) to assess its results for Portuguese. The first thing I realized was that a better modelling would be achieved by using only nouns and adjectives, so I converted my PALAVRAS-parsed files (from the Vercial and Obras corpus, encompassing therefore only 383 works) into sequences of noun and adjective lemmas, divided them into chunks so that each chunk belonged to only one work and author, and tried my luck.

Chang et al. (2009) note that topic modelling often yields best results when topics themselves are not humanly interpretable. Jockers, on the other hand, claims that most topics are easily identifiable after taking some time creating stopword lists. My impression is that few topics are really obvious, but I am a newcomer to the field. I therefore only present some topics (four out of the 100) as word clouds. I would name them respectively as "Catholic church", "animal realm", "festivities", and "politics".

Figure 6-1, 6-2.

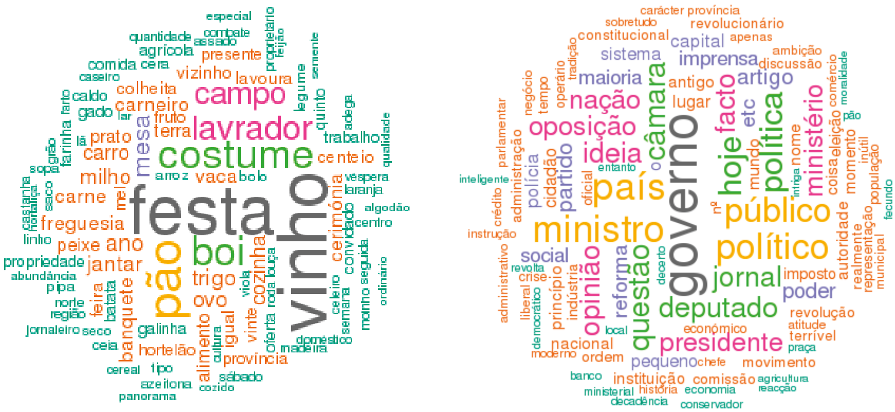

Figure 6-3, 6-4.

Figure 6. Four topics produced by MALLET for the Vercial and OBras corpora. In topic 1, the main words correspond to bishop, pope and prelate; in topic 2, dog, wolf, donkey and animal; for topic 3, party, wine, day, night and people; and for topic 4, government, minister, politics, public, and country. The original numbers of the topics were 2, 37, 64 and 80.

7. Concluding remarks

This paper had two goals: First, to present *Literateca* as a resource for all those interested in lusophone literature, most especially those who have not yet studied it with digital humanities techniques due to a lack of digital resources. Presenting a new environment, one must also illustrate the application of current methods, to make its use appealing.[13]

Secondly, I wanted to test the use of resources and techniques from two different research communities: corpus linguistics and literary digital humanities, complementarily instead of alternatively. I thus employed linguistic features for distant reading, and for improving topic modelling, while pointing out that clustering could improve annotation if one did both analyses in tandem.

13. Literateca is, like Gramateca, work in progress. Hence prospective users are directed to http://www.linguateca.pt/Literateca/, where they can find details, data and statements of intentions for the future.

Now that there are data publicly available and methods already tested for other purposes and other literatures, a myriad of interesting subjects springs to mind in the study of lusophone literature. I end this text discussing briefly two: emotional signatures and the import of tense in literature.

Emotions are obviously something intrinsically associated with literature, and are strongly social and cultural as well, so studying how different authors and works mention them can go a long way in understanding the import and originality of the specific authors and works.

Also, given that Portuguese has a very rich tense system, studying how different authors apply the tense palette should provide good insight into both the language and the works themselves. Both studies have to be postponed for a later occasion, but I hope to entice some readers of this paper to engage in them, as well as receive comments and feedback from Christian-Emil on these issues.

References

Ardanuy, Mariona Coll and Caroline Sporleder. 2014. Structure-based Clustering of Novels. *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL) @ EACL 2014*, Gothenburg, Sweden, April 27, ACL, 2014, pp. 31–39.

Auestad, Reiko Abe. 2017. Translating Japanese Narratives into English and Norwegian: Challenges of "Free Indirect Discourse" in Japanese. Presentation at the *World literature in the periphery* workshop, University of Oslo, 11–12 May 2017.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A practical introduction to Statistics using R*. Cambridge University Press, 2008.

Biber, Douglas. 1985. Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics* 23 (2), pp. 337–360.

—.1988. *Variation across speech and writing*. Cambridge University Press.

Biber, Douglas and Bethany Gray. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes* 9(1), pp. 1–82.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and E. Finegan. 1999. *The Longman grammar of spoken and written English*. Longman.

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. thesis, Aarhus University, Denmark.

Börner, Katy. 2011. Plug and Play Macroscopes. *Communications of the ACM* 54, 3, pp. 60–69.

Broadwell, Peter and Timothy R. Tangherlini. 2017. Confusing the modern breakthrough: Naïve bayes classification of authors and works. In *Digital Humanities in the Nordic countries, second conference*, http://dhn2017.eu/abstracts/#_Toc475332525

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta (eds), *Advances in Neural Information Processing Systems*, pp. 288–296.

Cortelazzo, Michele, Paolo Nadalutti and Arjuna Tuzzi. 2012. Una versione iterativa della distanza intertestuale applicata a un corpus di opere della letteratura italiana contemporanea. In Anne Dister, D. Longre and G. Purnelle (eds.), *JADT 2012 Actes des 11es Journes internationals d'analyse statistique des donnes textuelles*, LASLA SESLA, pp. 295–307.

Freitas, Cláudia. 2015. Esqueleto: anotação das palavras do corpo humano. Linguateca. http://www.linguateca.pt/acesso/Esqueleto/ Esqueleto.html

Freitas, Cláudia, Bianca Freitas and Diana Santos. 2016. QUEMDISSE?: Reported speech in Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 4410–4416.

Galves, Charlotte and Pablo Faria. 2010. Tycho Brahe Parsed Corpus of Historical Portuguese. URL: http://www.tycho.iel. unicamp.br/~tycho/ corpus/en/index.html

Goble, Carole and Chris Wroe. 2004. The Montagues and the Capulets. In *Comparative and Functional Genomics* 5 (8), December 2004, pp. 618–622.

Halliday, M.A.K. 1971. Linguistic function and literary style: An inquiry into the language of William Golding's The Inheritors. In Seymour Chatman (ed.), *Literary Style: A symposium*, 1971. Reprinted in Weber, J. J. (ed.), *The Stylistics Reader: From Roman Jakobson to the Present*. London: Arnold, pp. 56–86.

Hoof, Lieve Van. 2013. Dead languages and digital humanities: Social network analysis in the ancient world. What are Digital Humanities? Presentation at UiO, June 14–15, 2013.

Jockers, Matthew L. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Kenny, Anthony. 1982. *The computation of style: an introduction to statistics for students of literature and humanitie*s. Pergamon Press.

Kilgarriff, Adam, Sue Atkins and Michael Rundell. 2007. BNC Design Model Past its Sell-by. *Corpus Linguistics Conference*, Birmingham, UK, 2007.

Mahlberg, Michaela, Catherine Smith and Simon Preston. 2013. Phrases in literary contexts: Patterns and distributions of suspensions in Dickens's novels. *International Journal of Corpus Linguistics* 18(1), pp. 35–56.

Mahlberg, Michaela. 2015. Literary Style. In Douglas Biber and Randi Reppen (eds.), *The Handbook of Corpus Linguistics*, Cambridge University Press, pp. 346–361.

Mambrini, Francesco, Marco Passarotti and Caroline Sporleder. 2012. Annotation of Corpora for Research in the Humanities. Proceedings of the ACRH Workshop, Heidelberg, 5 Jan. 2012. *Journal for Language Technology and Computational Linguistics* 26 (2), pp. 7–10.

McCallum, Andrew Kachites. 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu. 2002.

McMurry, Margarida. 2015. *The Role of Assumptions in Author-Text-Audience Relationships: An Analysis of the Creative and Reading Processes in Narrative Fiction*. Ph.D. thesis, University of Oslo, Faculty of Humanities.

Moretti, Franco. 2000. Conjectures on world literature. *New Left review* 1, Jan-Feb 2000, pp. 54–68.

—.2011. Network theory, plot analysis. *New Left review* 68, Mar-Apr 2011, pp. 80–102.

Mota, Cristina & Diana Santos. 2015. "Emotions in natural language: a broad-coverage perspective". Linguateca. http://www. linguateca.pt/ acesso/EmotionsBC.pdf

Nalisnick, Eric T. and Henry S. Baird. 2013. Extracting sentiment networks from Shakespeare's plays. *IEEE 12th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 758–762.

Oakes, Michael P. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1998.

—.2014. *Literary Detective Work on the Computer*. John Benjamins Publishing Co., 2014.

Oppenheim, Rosa. 1988. The mathematical analysis of style: A correlation-based approach. *Computers and the Humanities* 22 (4), pp. 241–52.

Preminger, Michael and Gjertrud Fludal. 2016. OAUC at CLEF2016 SBS Lab: Using Appeal Elements to Improve Automatic Book Recommendation – Proof of Concept. In Krisztian Balog, Linda Cappellato, Nicola Ferro and Craig Macdonald (eds.), *Working Notes of CLEF 2016 (Conference and Labs of the Evaluation forum), Évora, Portugal, 5–8 September, 2016, CEUR Workshop Proceedings*. Vol. 1609, 2016, pp. 1145–1154.

Propp, Vladimir. 1928. *Morphology of the tale*, Leningrad.

R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2008.

Santos, Diana. 1998. Punctuation and multilinguality: Reflections from a language engineering perspective. In Jo Terje Ydstie and Anne C. Wollebæk (eds.), *Working Papers in Applied Linguistics*, University of Oslo, pp. 138–160.

—.2014a. Corpora at Linguateca: Vision and roads taken. In Tony Berber Sardinha and Telma de Lurdes São Bento Ferreira (eds.), *Working with Portuguese Corpora*, Bloomsbury, 2014, pp. 219–236.

—.2014b. Gramateca: corpus-based grammar of Portuguese. In Jorge Baptista, Nuno Mamede, Sara Candeias, Ivandré Paraboni, Thiago A.S. Pardo, and Maria das Graças Volpe Nunes (eds.), *International Conference on Computational Processing of Portuguese (PROPOR'2014)*, Springer, pp. 214–219.

—.2016. Comparando corpos orais (transcritos) e escritos no âmbito da Gramateca. In Camilla Bardel and Anna De Meo (eds.), *Proceedings from the conference Parler les langues romanes/Parlare le lingue romanze/Hablar las lenguas romances/Falando línguas românicas (The ninth GSCP International Conference)*. Universitá di Napoli L'Orientale, Il Torcoliere, pp. 127–142.

Santos, Diana, Cristina Mota and Augusto Soares da Silva. 2011. Guarda-fatos: notas sobre a anotação do campo semântica do vestuário nos corpos do AC/DC. Linguateca. http://www.linguateca.pt/ acesso/GuardaFatos.pdf

Schöch, Christof. 2017. Topic modeling genre: An exploration of French classical and enlightenment drama. *Digital Humanities Quarterly* 11 (2), 2017.

Schmidt, Kari Anne Rand. 1980. Male and female language in Jane Austen's novels. In Stig Johansson and Bjørn Tysdahl (eds.), *Papers from the First Nordic Conference for English Studies (Oslo, 19–19 September, 1980)*, pp. 198–210.

Silva, Rosário and Diana Santos. 2012. Arco-íris: notas sobre a anotação do campo semântico da cor em português. Linguateca: http://www.linguateca.pt/acesso/ArcoIris.pdf

Sinclair, John. 2005. Corpus and text – basic principles. In Martin Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, Oxbow Books, pp. 1–16.

Smith, Stacy L., Marc Choueiti and Katherine Pieper. 2014. Gender bias without borders: An investigation of female characters in popular films across 11 countries. https://seejane.org/wp-content/ uploads/gender-bias-without-borders-full-report.pdf

Steinberg, Erwin R. 1973. *The stream-of-consciousness and beyond in Ulysses*. University of Pittsburgh, 1973.

Stubbs, Michael. 2005. Conrad in the computer: examples of quantitative stylistic matters. *Language and literature* 14(1), 2005, pp. 5–24.

Tangherlini, Timothy R. 2013. The Folklore Macroscope: Challenges for a Computational Folkloristics. The 34th Archer Taylor Memorial Lecture. *Western Folklore* 72(1), pp. 7–27.

Volkova, Ekaterina P., Betty J. Mohler, Detmar Meurers, Dale Gerdemann and Heinrich H. Bülthoff. 2010. Emotional perception of fairy tales: achieving agreement in emotion annotation of text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (CAAGET '10)*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 98–106.

Yule, George Udny. 1944. *The statistical study of literary vocabulary*. Cambridge University Press, 1944.

Zampieri, Marcos and Martin Becker. 2013. Colonia: Corpus of Historical Portuguese. In Marcos Zampieri and Sascha Diwersy (eds.), *Non-standard Data Sources in Corpus-based Research*, Volume 5, ZSM-Studien, Schriften des Zentrums Sprachenvielfalt und Mehrsprachigkeit der Universität zu Köln Shaker, pp. 77–84.

Zhao, Ying and Justin Zobel. 2007. Searching with Style: Authorship Attribution in Classic Literature. In Dobbie, G. (ed.), *Proc. Thirtieth Australasian*

Computer Science Conference (ACSC2007), Ballarat, Australia.ACS. 2007, pp. 59–68.

Corpora

Colonia	http://corporavm.uni-koeln.de/colonia/
OBras	http://www.linguateca.pt/OBRAS/OBRAS.html
PANTERA	http://www.linguateca.pt/PANTERA/
Tycho Brahe	http://www.tycho.iel.unicamp.br/corpus/index.html
Vercial	http://www.linguateca.pt/acesso/corpus.php?corpus=VERCIAL

Acknowledgements:
I am grateful to FCCN for hosting Linguateca and thus Literateca, to Linguateca's team for the resources, and to UNINETT Sigma2, the National Infrastructure for High Performance Computing and Data Storage in Norway, for usage of the abel cluster.