Placing GIS and NLP in literary geography: experiments with literature in Portuguese

Diana Santos[1,2] and Daniel Alves[3]

[1] ILOS-UiO, Postboks 1003 Blindern, NO-0315 Oslo, Norway
[2] Linguateca
[3] IHC, NOVA-FCSH, Portugal
d.s.m.santos@ilos.uio, dra@fcsh.unl.pt

Abstract. In this case study we discuss different approaches to the study of literature in digital humanities and try to join two methodologies, namely distant reading and spatial analysis. We first describe shortly the two projects involved, Atlas and Literateca, highlighting and quantifying the different ways to deal with place in literature in Portuguese. Then we describe some different paths to compare and harmonize the two approaches, focusing on annotation, extraction and geocoding of place names.

Keywords: Digital Humanities, Portuguese, Literature, Geographic information systems, Distant reading.

1 Introduction

As far as we know there aren't many projects that use computational linguistics for the study of literature in Portuguese. Most of the existing work can be classified rather as digital archives of literary works, in particular dedicated to a specific author (Portela, 2013), or projects that use literary corpora to study the language and their uses, not the literary features of the works assembled (Rodrigues, Freitas and Quental, 2013). We can however mention Santos et al. (2020) as a promising first step in this direction.

Regarding the use of geographic information systems (GIS) to study Portuguese literature, currently to our knowledge we are only aware of the study of *Peregrinação* by Canosa (2019) and the studies developed by Alves and Queiroz about literary landscapes (2013, 2015). Although there are many more English language projects, for instance, that combine computational linguistics with literature (see e.g. Mahlberg, Smith and Preston, 2013; Ardanuy and Sporleder, 2014; Valaa *et al.*, 2015) and that use cartography and GIS for literary studies (see Heuser, Moretti and Steiner, 2016) in what has been called "literary cartography", "literary geography" or "literary GIS" (Tally, 2008; Piatti, Reuschel and Hurni, 2009; Cooper and Gregory, 2011), we believe that this comparison of two disparate methodologies can still be interesting to a wider audience.

Our aim with the work we describe in this paper, developed in the scope of the project *BILLIG - Bilateral Lusophone Literature Initiative using GIS and Linguistics* financed by EEA grants[4], was to assess and produce some first steps in a cross-disciplinary endeavor that used methods and techniques from computational linguistics and spatial analysis to compare and harmonize a close and distant reading approach to the annotation, extraction and geocoding of place names in Portuguese literature.

Using material from two different projects, the Atlas of Literary Landscapes of Mainland Portugal (hereafter just Atlas) and Literateca, the goal was to provide better systems to both projects and contribute to improve knowledge about ways to use natural language processing (NLP) and GIS in literary studies.

1.1 Atlas

Atlas is an interdisciplinary project, existing since 2010, configuring itself with a markedly digital methodology for academic analysis. Its main objective translates into a reading about the environment and landscapes of the territory of mainland Portugal configured in literary texts. Admitting the idea that writers are also cartographers (Tally, 2008), one of the main purposes of this project is embodied in the mapping of literary texts. At the root of its methodology is the possibility of extracting, categorizing and mapping the various representations that Portuguese and foreign writers of the last century and a half have produced on the Portuguese mainland and on the natural, cultural and social heritage that inhabits and interacts in it (Alves and Queiroz, 2013, 2015).

In order to facilitate the identification of the geographical references contained in this corpus, each literary representation of landscapes of mainland Portugal was recorded as a single excerpt, in a shared database. These excerpts are distinct passages, which can be read and understood independently and, above all, give us a clear sense of the aesthetic aspects of the works from which they derive. After surveying and identifying these extracts, the readers that collaborate in the project classify them into categories (corresponding to geographic, ecological, socioeconomic, cultural and / or historical themes), and also assign geographic coordinates to the locations identified.

The project uses a hybrid methodology: it combines the traditional methods of "close reading", with a perspective of "distant reading" (Jockers, 2013; Moretti, 2013) embodied in the use of a shared database created in PostgreSQL, a geographic information system (GIS) and quantitative methods. At this level, it is a singular project in Portugal but has similarities with other digital literary mapping projects existing in several countries[5].

[4] *BILLIG*, https://billig.fcsh.unl.pt/, accessed on December 19, 2020.

[5] See, for instance, *A Literary Atlas of Europe*, http://www.literaturatlas.eu/en/; *Digital Literary Atlas of Ireland, 1922-1949*, http://cehresearch.org/DLAI/; *GéoCulture, Le Limousin Vu Par Les Artistes*, https://geoculture.fr/; *Mapping Lake District Literature* https://www.lancaster.ac.uk/fass/projects/spatialhum.wordpress/; or *The Space of Slovenian Literary Culture*, http://pslk.zrc-sazu.si/en/. All accessed on December 19, 2020.

1.2 Literateca

Literateca is an environment for studying literature in Portuguese, created in 2018, with a Web interface to literary corpora annotated with linguistic and literary information. It is a direct descendent of Gramateca, an environment to do corpus-based studies of Portuguese grammar (Santos, 2014), enhanced with metadata regarding literary works (author, data, literary school, genre) and with an interface to R-based statistical procedures common in digital literary studies (topic models, correspondence analysis, principal components analysis). See Santos (2019) for more details. Some of the annotation of Literateca has in addition been motivated by literary questions.

Although Portuguese is the 5th or 6th most spoken language in the world, and has a huge literature spawning hundreds of years, most digitization projects have focused on canonical and best works, that is, quality, and not on breadth and width, that is, quantity. The few that took a wide grasp such as those done by Google, produced digital objects of very bad quality. So although this may sound almost unbelievable, it was hard to get 100 novels by 80 different authors from Portugal from 1840-1920, which was one of the goals of the COST project "Distant reading for European literary history" (see Santos, Bick and Wlodek, 2020; Schöch *et al.*, 2021). This collection is a subset of Literateca, which also includes older and more modern texts, other genres (drama and poetry, short stories, etc.) and especially works from Brazilian authors.

1.3 Aim and scope of the study

In this study we will begin by comparing generically the two projects and their different assumptions, strengths and weaknesses in Section 2. Then in Section 3 we identify the textual intersection and compare the results obtained by automatic NER performed in Literateca with the data manually annotated in the Atlas to identify the differences. In Section 4 we look at the whole Atlas as annotated by PALAVRAS-NER (Bick, 2000, 2006) and give a more encompassing description of the differences between the two projects and methods, looking at excerpts without named entities and discussing automatic scope identification. In Section 5 we use Atlas geographical coordinates in Literateca and attempt to give some initial measures of the geographic reading of Portuguese literature.

The main aim is eventually to answer the two related questions: What kinds of places are found in literary text, and how many of them can be geolocated?

2 Are the two projects comparable?

At a very high level of abstraction, both Atlas and Literateca use literary works to learn about the culture they represent and annotate them with information that can also be used to reflect on the works themselves. However, if we look in more detail, there are significant differences in method and assumptions, as we will show here.

2.1 Statistics about the two projects

Literateca, which is a project constantly increasing, includes in version 8.2, 9 November 2022, ca. 37 million words, corresponding to 926 works, ranging from 1380's chronicles of King D. João I to Luísa Marques da Silva's novel *mISTério@Tagus* from 2020. It should be noted that, except for some few modern cases, due to copyright, all other works are included in their entirety. There were no other reasons for inclusion in Literateca apart from availability.

As to Atlas, 377 works from 184 writers, published between 1778 and 2019, are included. The main criteria to include a literary work was the presence of landscape descriptions, and only excerpts that included geographical information or from where locations could be inferred were digitized. As of May 2022, there are 7751 excerpts in the database, classified in 27 thematic categories and including more than 1.7 million words. All excerpts were then geographically annotated in two levels: the NUTS III (nomenclature of territorial units for statistics) where the landscape description is included, and, when possible, the specific locations that are mentioned or can be identified in the excerpt. In the comparisons ahead we only use the material from this second level of annotation. Since most works were still under copyright, even the digitized excerpts are not always available publicly on the project Web services[6].

2.2 How many works are included in both projects?

The first and most naive comparison would seek to establish which works appear in both projects, so that we may directly enrich them with information from the other one.

By simply browsing the lists available for both projects, we found that only 21 works were common (22, if one counted different volumes as different works, as Atlas does), most of them by canonical nineteenth century writers such as Eça de Queirós (11 works) and Camilo Castelo Branco (5/6). The remaining works were two by Abel Botelho, one by Florbela Espanca, one by Carlos Malheiro Dias and one by Conde de Ficalho.

This definitely constitutes a very small and skewed sample, but we looked at the data both projects bring in order to assess the comparability of the information on the same material.

In connection with the different data gathering methods, the works (full text) amounted to 2,277,577 tokens in Literateca, but just to 379,611 tokens in Atlas (excerpts). It is to these we turn to in section 3, to identify geographical named entities.

3 First intersection: experiments in the Atlas

Our first experiment was to apply the NLP tools to the whole of Atlas and use it to check the coverage and the correctness of the manual annotation, as well as identify the cases where the automatic annotation tool used in Literateca needed improvement.

[6] LITESCAPE.PT, http://litescape.ielt.fcsh.unl.pt/, accessed on December 19, 2020.

The final goal is to assess the possibility of an automatic location assignment. We will do this in turn. But the first move was to look at the textual intersection of the two projects, both because it is more manageable to deal with in its entirety, and because eventual changes or discoveries would benefit both projects.

3.1 Can we compare the place names obtained by the two methods?

The automatic annotation using Literateca's tools of the intersection of the two projects, that is, the Atlas excerpts which were also included in Literateca, amounting to 167,088 words, yielded 2,287 occurrences of geographical entities, which correspond to 1,227 annotations in Atlas with geographical coordinates. This relatively large discrepancy (1,517 cases) can be due to many (independent) factors:

1. the fact that often the description of the geographical place in Atlas is written in its full name, while in context a much shorter name is used, see e.g *Café/Pastelaria Benard* (mentioned in the excerpts usually only as *Benard*), *Museu Nacional dos Coches* (usually referenced in the texts as *Museu dos Coches*), or *Rotunda do Marquês* (sometimes referenced only as *Rotunda*).

2. the fact that some locations were wrongly classified by the automatic parser, like *D. João V* (a king's name) or *Fado do Bairro Alto* (a name of a song that mentions a neighborhood in Lisbon) or *Estado* (a general reference to state, as in Portuguese state).

3. the fact that some generic locations, like *Portugal*, were not geo-located by the Atlas users, the same happening to those referring to foreign countries or regions, like *Moçambique*, *Guiné* or *Norte de África.*

4. when the locations were of a physical character, like rivers, seas, or mountains, they may not be geo-located in the Atlas. Others, like (river) *Tejo*, were, but this of course raises problems either way. Because the literary excerpt may describe Tejo near Lisbon, or Tejo near Santarém (nearly 80 km north), and both have a unique georeference but do not refer to the whole river. In the Atlas, those cases resulted in two, or more, different coordinates related to the same geographical feature. We did not consider this problem in our comparison, assuming there was only one coordinate for any river.

5. some locations may have not been annotated in the Atlas by mistake, or because they corresponded to fictional entities.

We investigated the overlap and the differences in Table 1, for the 1,517 cases, that correspond to 347 distinct (automatically found) place names.

Table 1. Classifying the differences between Atlas human classification and Literateca's automatic one for the common subset

Kind of difference	**Number of cases**	**%**
Automatic error	116	33
Missing from Atlas	96	28
Different (often more encompassing) description	70	20
Outside Portugal	47	14

Other cases	18	5
Total	347	

More than half of the cases were due to errors of the automatic analysis or missing locations in the Atlas, but there were also some other kinds of mismatches that we had not foreseen when starting error analysis, such as popular (as opposed to official) names for Portuguese regions, or actual changes in the naming of streets.

Given that a significant number of cases was formally distinct but substantively the same (70 cases), we created a new version of our gazetteer where we "translated" the larger, normalized description of Atlas into smaller, more informal, names, for the (60) cases where the shorter name was not ambiguous between several locations. We managed then to geolocate 230 distinct named entities covering 1642 occurrences, improving from 44.7% to 59.8%.

4 Analyzing the whole Atlas

Even though in the previous section we looked only at the textual intersection of the two projects, it is more interesting to look at the whole Atlas to have an idea of the differences.

4.1 How many excerpts have named entities (place names)?

We list in what follows an overview of NE density in the whole Atlas: which cases had geographical named entities, and which ones hadn't, using PALAVRAS-NER, more specifically the features *top* (toponym) and *civ* (city) (Bick, 2007). Table 2 shows an overview of how many different place names were identified per excerpt:

Table 2. According to PALAVRAS-NER, how many place names per excerpt in Atlas

Number of place names	Number of excerpts
no place name	1,557
one	1,399
two	753
three	394
four	264
five	131
six	68
seven	51
eight	26
nine	19
ten	13
more than ten	34

These figures show that a sizeable percentage of text excerpts, classified by Atlas as describing a region or place, do not have named entities (at least automatically recognized by PALAVRAS-NER). In fact, 2,786 excerpts are in this condition, and they do seem to be assigned geographical entities based on the human knowledge. We could not look at all the material in detail, but a cursory examination does seem to confirm that the textual excerpts do not refer to a named place, as the following excerpts illustrate.

"De uma janela entreaberta, Vasco da Gama, ainda um pouco atordoado, avista lá fora os criados numa dobadoira de idas e vindas, retirando, das tendas de damasco branco, bacias ainda cobertas de iguarias: manjares, conservas, frutos; a boda está dentro e fora do palácio; continuam os momos e entremeses, acendem-se velas de cera dourada. […] Vasco abandona a festa pouco depois, pretextando cansaço, mas com a face banhada de esperança. As laranjeiras, sob o luar, naquele largo branco de cal e absoluto, estão cobertas de ouro." (Urbano Tavares Rodrigues, *Os campos da promessa*, 1998, p. 68) (Places annotated by the Atlas readers: Paço Real de Évora / Palácio de D. Manuel - https://maps.google.pt/maps?hl=pt-PT&ll=38.567778,-7.909167&spn=0,0) [Our translation: "From a slightly open window, Vasco da Gama, still little dizzy, observes the servants back and forward removing, from white damask tents, bowls still full of treats: delicacies, preserves, fruits; the party is inside and outside the palace; the clowns and the theater plays continue, golden lights are lighted. […] Vasco leaves the party sometime later, pretending tiredness, but with his face lighted by hope. The orange trees, under moonlight, in that square of an absolute white, are covered with gold."]

"Entre a casa e a cidade longínqua estendem-se as dunas como um grande jardim deserto, inculto e transparente onde o vento que curva as ervas altas, secas e finas faz voar em frente dos olhos o loiro dos cabelos. Ali crescem também os lírios selvagens cujo intenso perfume, pesado e opaco como o perfume de um nardo, corta o perfume árido e vítreo das areias." (Sophia de Mello Breyner Andresen, *Histórias da Terra e do Mar*, 1984, p. 69) (Places annotated by the Atlas readers: Praia da Granja - https://maps.google.pt/maps?hl=pt-PT&ll=41.0326829,-8.6463397&spn=0,0) [Our translation: "Between the house and the far away town the dunes stretch as a large deserted garden, uncultivated and transparent where the wind which curves the high, dry and thin grass makes the blond hair fly before the eyes. There grow also the wild lilies whose strong scent, heavy and opaque as a nard's perfume, cuts the arid and glassy smell of the sand."]

Looking now from the Atlas perspective, it includes 3,158 distinct places, of which 2,986 have geographical coordinates. The ones that miss coordinates are mostly references to small land properties (*quintas* [farms], *herdades* [homesteads]) difficult to locate, old streets that disappear or change their name, fictional places or vague locations (ex: *Campos do Mondego* [Mondego Fields], a vague territory between two Portuguese cities, Coimbra and Montemor).

Table 3 shows a comparable account to that of Table 2, describing how many named locations were identified per excerpt by the Atlas contributors.

Table 3. How many place names per excerpt were classified by the readers in Atlas

Number of place names	Number of excerpts
no place name	2538
one	2379
two	1051
three	645
four	373
five	230
six	139
seven	91
eight	61
nine	56
ten	32
more than ten	112

As mentioned, the Atlas approach was twofold. Every excerpt was georeferenced to a region (NUTS III), and to a more precise location or locations. That is why we have 2538 with no place name, because it was possible to figure out the larger region the excerpt was mentioned or refer to (for instance Alentejo, Douro or Algarve) but no specific location was mentioned or could be inferred. Of course, some of these can be the result of an error from the reader, that didn't register a mentioned place name or location, or was unable to identify it.

4.2 Which locations were not detected by the NER system?

We want to identify the locations that were manually annotated but were not identified by the NER system. This may be due to two different cases: there were no place names entities in the excerpt (as discussed in the previous subsection), which required consequently human interpretation to be located, or the named entity recognizer failed.

This is however very difficult to identify automatically, given that, as expounded in Section 3.1, the names used are often different. We have anyway tried to establish in how many cases there was agreement and have analyzed a random set of 50 excerpts where Atlas readers and PALAVRAS-NER disagrees.

In addition to errors from both sides, different designations, and reader's knowledge, we also identified another reason for disagreement related to the specific Atlas methodology for place names annotation mentioned above: Atlas readers would only indicate the region in the first level of annotation (not as a specific place) even if the region is present in the text, while the automatic NE recognizer flagged all place names. Some examples that can be mention are *Viana do Castelo/Santa Luzia* and

Lisboa/Estrela. Estrela is a location in Lisboa, and Santa Luzia is a church in Viana do Castelo, so Atlas readers only marked *Estrela* and *Santa Luzia*, because they are more specific than the mentions to the cities, already registered in the first level of annotation (NUTS III).

The result from our examination of 50 excerpts is in Table 4.

Table 4. Differences between the location names identified in 50 random excerpts, containing 135 putative place names

Type	Number
PALAVRAS-error	44
Not present in the text	17
Atlas-error	15
Missing on purpose	14
Different designation for the same location	12
Other cases	10

The other cases correspond to place names that were mentioned but did not refer to the main location that the excerpt describes, or place names used in a non-geographical way, see respectively:

"Esta vila adormecida estava a cem léguas do Porto e da vida." (Raul Brandão, *Os Pescadores*) [Our translation: "This sleepy village was 100 miles away from Porto and from life."] (Porto is not described in this excerpt, which is about a village. It is just mentioned as a comparison point.)

"Sempre que em Lisboa se constrói um prédio de estilo, com prosápia inovadora, cai Tróia, caem o Carmo e a Trindade […]" (Mário de Carvalho, *Era Bom que Trocássemos umas Ideias sobre o Assunto*) [Our translation: "Whenever one creates in Lisbon a building with style, innovative, Troy falls, Carmo and Trindade falls down […]"] (Carmo and Trindade are two places in Lisbon deeply affected by the 1755 Earthquake and whose names are popularly used meaning a huge polemics, a terrible war)

4.3 Could the geographical scope be automatically detected?

We might want to do the opposite investigation: given Literateca's NLP tools, how often would one be able to automatically identify a particular region talked about in an excerpt, and give these classifications for humans to check?

This would require having some algorithm that from named entities would select an excerpt and attribute some geographical tags. Several approaches for doing this can be mentioned: using geocoding algorithms inserted in GIS software to give geographic coordinates to addresses or locations and then interpolate that data with upper level administrative units where the coordinates were included (Alves, 2005, 2017; Alves and Queiroz, 2013); comparing lists of place names with already built gazetteers where the locations and their respective coordinates are associated with

several levels of administrative or other type of territorial units (from parish to continent level, for instance) (Mostern, Southall and Berman, 2016).

Given that we have already uncovered several cases where the two approaches have considerably different results (namely a large amount of cases with no named entities, and overgeneration of spurious place names by PALAVRAS-NER), this cannot obviously be done in general. But we used the manually investigated 50 cases of the previous section to provide an estimate of in how many cases (with more than one place name) one could try to attempt this scope suggestion.

Out of 20 cases with more than one place name found by PALAVRAS-NER, only 13 could possibly lead to a scope. Of these, this scope was already mentioned/part of the set in 2 cases (Examples: Guadiana, Lisboa, Portugal, Santa Catarina → scope: Portugal; Alentejo, Lisboa, Praça da Alegria, Praça das Flores, São Bento, Tejo → scope: Lisboa)

5. A statistical characterization of places in Portuguese literature

As already stated, our ultimate question, to learn more about places in literature, was: what kinds of places are found in literary text, and how many of them can be geolocated?

It is expected that the overall density of place names in literary texts is much lower than the one found in the Atlas excerpts, which, as explained above, were chosen precisely because they were about places.

We also want to know about the typology of places: what kind of granularity do they show? Likewise, the issue of fictive versus real places is important in literature. The latter cannot be geolocated, except if the author has provided a fictive map as well.[7]

In order to try to answer these questions, we started human revision of Literateca's place names, most specifically in the two corpora that include exclusively Portuguese works, namely Vercial and NOBRE. This revision, underway, proceeds from the most common lemma automatically annotated as place and adds the right subclassification (country, region, city, street, etc.). At the time of writing this paper, we had revised 2521 distinct cases of place names, covering 109,413 occurrences in the corpus subset we used, which features 24 million words. (For the record, before human revision 18,512 places had been proposed by NER, corresponding to 203,032 running words in the corpus; the current estimate, adding up the revised cases to the ones not yet revised, comprises 8,365 possible place names, corresponding to 180,367 running words.)

Table 5 shows the current distribution among the most common kinds of places (we did not include here the cases which were vague between for example a city or a municipality, only unambiguous cases).

Table 5. Distribution among the most common kinds of places

[7] See http://lotrproject.com/map/#zoom=3&lat=-1315.5&lon=1500&layers=BTTTTT for visualization of a fictional universe. Last access 6 January 2021.

Kind of place	Amount
city	35279
country	27159
territory or region	5180
street	5041
town (*vila*)	4385
continent	4120
parish (*freguesia*)	3716
province	2457
religious building	2418
municipality	1832
public building	1218
quarter	1083
planet	1018
village (*aldeia*)	649

We tried to automatically geocode the above mentioned curated/revised (1952) cases. For the geocoding process (assigning latitude / longitude coordinates to places and addresses), after analyzing several lists of places or toponyms (that of the Atlas itself and another created from the toponyms of the Portuguese Military Charter) and different tools (ZeeMaps, BatchGeo, Edinburgh Geoparser and QGIS WebService Geocode), we decided to use the QGIS WebService Geocode, with the OpenStreetMap gazetteer, as it presents itself as the most simple and intuitive to use and also the one that generated more satisfactory results globally. We needed a process that could deal with the Portuguese names (Atlas and the Military Charter perform well here) but which would also be able to identify places outside the Portuguese territory, and for that the community driven project OpenStreetMap was the most balanced solution. Using QGIS we found initially 85.7% coordinates to place names in our list, which is a very high percentage. However, by going through every case, we could only certify 905 cases, resulting in 46.4%.

The major reasons for lack of precision of the geocoding procedure are the following:

1) old orthography in several cases, most notably the already mentioned Peregrinação from the 16th century

2) the existence of many identical places in Brazil and Portugal, while this concerned Portuguese literature

3) the existence of many places in the United States of America that have identical names with biblical (and some European) places

4) the fact that many street names in Portugal are not unique, and therefore the street geocoding procedure, to be precise, would have to indicate the city where the novels occur

5) many names of streets in Portugal have changed, as is in fact acknowledged in many historical novels, which present both

In fact, going through the whole list we could but note that fictional places were very few and mainly referred to shared fictions like Paradise or Olympus. This may be the case because fictional places that occur in only one novel/work are necessarily less frequent. It is in the long tail of places mentioned in only one work that we expect to find several different imaginary places, or others so small and irrelevant (in a global or popular perspective) that escape the gazetteers used in the geocoding tools.

It should be stressed that the most interesting and challenging issue is the fact that place names are often not used in its place meaning, but offer a wide spectrum of uses often only loosely connected with the place, as illustrated by the following random examples of the ocurrence of the word *Lisboa* in Literateca:

"[…] é o vice-rei nas províncias do norte... o nosso bom padre Luís de Sousa, que pelos modos está nomeado patriarca de Lisboa…" (Camilo Castelo Branco, *A Brasileira de Prazins*) [Our translation: "[…] it is the vice-king of the North provinces… our good old Father Luís de Sousa, who apparently was named Lisbon's patriarch…"]

"*Nenhum dos viajantes recebera noticias de Lisboa"* (António Augusto Teixeira de Vasconcelos, *A ermida de Castromino*) [Our translation: "None of the travellers had received news from Lisbon"]

"*À herança de D. Antónia Joaquina Xavier concorreram três famílias de Lisboa, Évora e Tavira que se apelidavam Nobres"* (Camilo Castelo Branco, *A Caveira da Mártir*) [Our translation: "As heirs of D. Antónia Joaquina Xavier came three families from Lisbon, Evora and Tavira with the surname Nobre"]

"*O abade dissertava gravemente sobre os caminhos de ferro e suas vantagens, relembrando as antigas jornadas do seu tempo, a cavalo ou de liteira, quando para ir do Porto a Lisboa era preciso fazer testamento"* (Luís Magalhães, *O Brasileiro Soares*) [Our translation: "The abbot was speaking gravely about railways and their advantages, remembering the old journeys of his time, on horseback or litter, when in order to travel from Oporto to Lisbon one had to write a will"]

"*– E como Lisboa se não entregava, estou a adivinhar que maiores foram para ela as antipatias da corte"* (António de Campos Júnior, *A Ala dos Namorados*) [Our translation: "Since Lisbon did not surrender, I can guess that she raised higher antipathies from the court"]

"*Livro dos Pregos, f. 3, no Cartório da Câmara Municipal de Lisboa"* (Alexandre Herculano, *História de Portugal IV*) [Our translation: "Book of Nails, page 3, Lisbon Registry"]

While this is well known by those who work in the field of named entity recognition (see e.g. the HAREM evaluation contest for Portuguese, Santos *et al.*, 2006), it means that we cannot create maps for every occurrence of a place name. We must be able to annotate only those cases that refer to physical locations, which we did here by human revision. Therefore, we have so far (version 8.2 of Literateca, 10 May 2022) only 1965 different place names georeferenced, corresponding to 109,413 words.

In any case, this is publicly available on the Web to everyone interested in the subject, and we may say that this is a resource which is undoubtedly useful for literary studies of place.

6. Using GIS to improve interaction with Literateca

The BILLIG project also allowed us to directly apply GIS technology to the Literateca Web interface, which provided mainly concordances, lists or tables.[8] Some queries to Literateca could present the results in a map form, as others also produce a word cloud.

We have therefore experimented with the following: we used the 763 locations for which we had coordinates from the previous experiments and coded all cases which were mentioned in the corpus as a location, and, if the query results contained at least one case with coordinates, we gave the user the possibility to create a map.

As a proof of concept, we thus developed a PHP application that interacts with the Leaflet[9] library and is invoked by Literateca's interface. In Figure 1 we show two maps of which Portuguese cities or towns are mentioned in the works by two well-known authors, Camilo Castelo Branco and Eça de Queirós.

We chose these two authors because they did not write historical novels, where the identification of place names is especially problematic: many of the locations no longer exist or have changed name, which means that a good map-drawing application should also take in consideration the time of the reference to the location, or at least require the users to pose queries that make sense within a specific temporal stamp. This has to be addressed in further work.

In any case, given the different literary profile and the difference place of origin of these two renowned authors, it was surprising to see the striking similarity between the two maps: is literary Portugal fixed, independent of the author? We suspect that considerable differences might emerge if the frequency of occurrence of the places – and not only mere occurrence – were present in the map.

[8] https://www.linguateca.pt/acesso/corpus.php?corpus=LITERATECA Last access 6 January 2021.

[9] https://leafletjs.com/ Last access 6 January 2021.

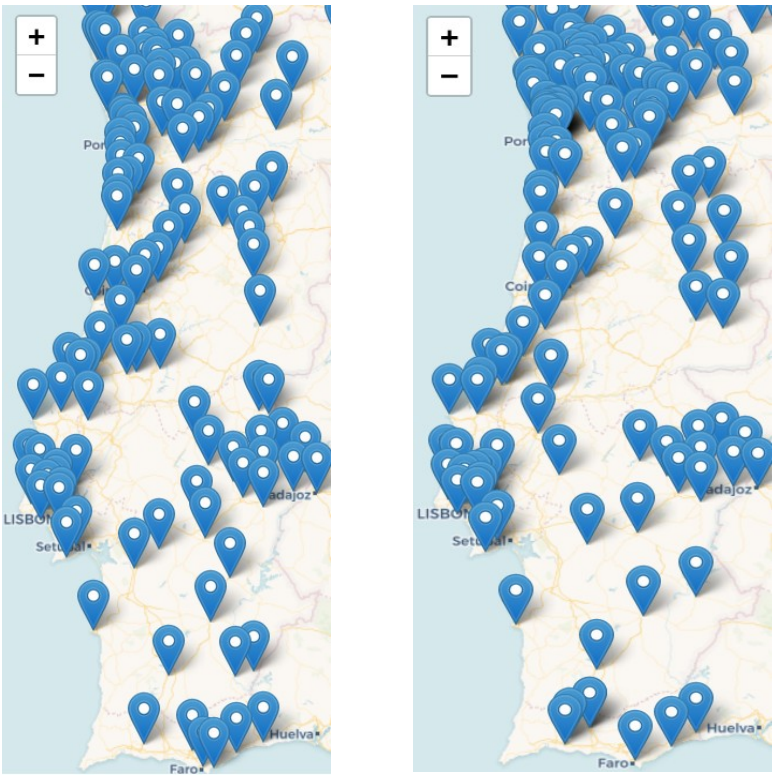

Figure 1. To the left, the cities, towns and villages named in Eça de Queirós' works, to the right the ones named in Camilo Castelo Branco's

7. Concluding remarks

The experiments we have done here show – or rather, confirm – that there are many thorny issues to be taken in consideration in literary GIS: not only natural language processing cannot replace a close reader when facing a scenery description without place names, but also place names are used in many other ways than just referencing a place. Furthermore, place names are time-dependent, ambiguous (the same name can indicate different places), vague (one denomination can be used for many kinds of geographic entities or for different levels of granularity), and varying (the same place has often many ways to be described in language), as overviewed in Santos and Chaves (2006). All these things must be taken into consideration if we want to understand the place of places in literature.

In this case study we took some of these challenges seriously: first, we produced a detailed comparison of the results from two computational technologies (developed by independent projects, *Atlas* and *Literateca*), clearly showing that named entity recognition, and human interpretation produce fairly disparate results. Then, we attempted to merge techniques from both approaches in an experimental system that, in

addition to take into account the several meanings of place names, also produced maps based on that information, as a pilot Web service for distant reading of places in Portuguese literature.

Acknowledgements

We are grateful to Paulo Alves for developing the Leaflet interface, and to Inês Lucas for her thorough revision of place names in Literateca. This work was partially supported by the Fund for Bilateral Relations (EEA Grants) of the Financial Mechanism Programme (2014-2021) with the grant number FBR_OC1_13. We also thank UNINETT Sigma2 – the National Infrastructure for High Performance Computing and Data Storage in Norway – for use of their computational resources, as well as FCCN for hosting Linguateca in their servers.

Works cited

Alves, D. (2005) 'Using a GIS to reconstruct the nineteenth century Lisbon parishes', in *Humanities, Computers and Cultural Heritage. Proceedings of the XVIth international conference of the Association for History and Computing*. Amsterdam: Royal Netherlands Academy of Arts and Sciences, pp. 12–17.

Alves, D. (2017) 'Shopkeepers and the city: the spatial economy of the retail trade in a European capital city (Lisbon, 1890–1910)', *History of Retailing and Consumption*, 3(2), pp. 139–158. doi: 10.1080/2373518X.2017.1329194.

Alves, D. and Queiroz, A. I. (2013) 'Studying urban space and literary representations using GIS: Lisbon, Portugal, 1852-2009', *Social Science History*, 37(4), pp. 457–481. doi: 10.1215/01455532-2346861.

Alves, D. and Queiroz, A. I. (2015) 'Exploring Literary Landscapes: From Texts to Spatiotemporal Analysis through Collaborative Work and GIS', *International Journal of Humanities and Arts Computing*, 9(1), pp. 57–73. doi: 10.3366/ijhac.2015.0138.

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

Bick, E. (2000) *The Parsing System 'Palavras': Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph. D. Aarhus University Press.

Bick, E. (2006) 'Functional Aspects in Portuguese NER', in Vieira, R. et al. (eds) *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006 (PROPOR'2006)*. Berlin: Springer Verlag, pp. 80–89.

Bick, E. (2007) ‘Automatic Semantic Role Annotation for Portuguese’, in *TIL, V Workshop em Tecnologia da Informação e da Linguagem Humana*. Rio de Janeiro, pp. 1715–1719.

Canosa, A. X. (2019) ‘Referentes por coordenadas e georreferências relativas das entidades geográficas mencionadas na Peregrinação’, in Pazos-Alonso, C. et al. (eds) *De Oriente a Ocidente: Estudos da Associação Internacional de Lusitanistas*. Coimbra: Angelus Novus, pp. 11–34.

Cooper, D. and Gregory, I. N. (2011) ‘Mapping the English Lake District: A literary GIS’, *Transactions of the Institute of British Geographers*, 36(1), pp. 89–108.

Heuser, R., Moretti, F. and Steiner, E. (2016) ‘The emotions of London’, *Literary Lab Pamphlet*, (13).

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

Mahlberg, M., Smith, C. and Preston, S. (2013) ‘Phrases in literary contexts: Patterns and distributions of suspensions in Dickens’s novels’, *International Journal of Corpus Linguistics*, 18(1), pp. 35–56.

Moretti, F. (2013) *Distant Reading*. London: Verso Books.

Mostern, R., Southall, H. and Berman, M. L. (eds) (2016) *Placing Names: Enriching and Integrating Gazetteers*. Indiana: Indiana University Press.

Piatti, B., Reuschel, A.-K. and Hurni, L. (2009) ‘Literary geography – or how cartographers open up a new dimension for literary studies’, in *Proceedings of the 24th International Cartography Conference*. Santiago: International Cartographic Association. Available at: http://icaci.org/files/documents/ICC_proceedings/ICC2009/html/nonref/24_1.pdf.

Portela, M. (2013) ‘“Nenhum Problema Tem Solução”: Um Arquivo Digital do Livro do Desassossego’, *MATLIT: Materialidades da Literatura*, 1(1), pp. 9–33.

Rodrigues, E. dos S., Freitas, C. and Quental, V. (2013) ‘Análise de inteligibilidade textual por meio de ferramentas de processamento automático do português: avaliação da Coleção Literatura para Todos’, *Letras de Hoje*, 48(1), pp. 91–99.

Santos, D. *et al.* (2006) ‘HAREM: An Advanced NER Evaluation Contest for Portuguese’, in Calzolari, N. et al. (eds) *Proceedings of LREC 2006*. ELRA, pp. 1986–1991.

Santos, D. (2014) ‘Gramateca: corpus-based grammar of Portuguese’, in Baptista, J. et al. (eds) *Computational Processing of the Portuguese Language, 11th International Conference, PROPOR 2014, São Carlos/SP, Brazil, October 6-8, 2014, Proceedings*. Heidelberg: Springer, pp. 214–219.

Santos, D. (2019) ‘Literature studies in Literateca: between digital humanities and corpus linguistics’, in Doerr, M. et al. (eds) *Humanists and the digital toolbox: In honour of Christian-Emil Smith Ore*. Oslo: Novus forlag, pp. 89–109.

Santos, D. *et al.* (2020) ‘Periodização automática: Estudos linguístico-estatísticos de literatura lusófona’, *Linguamática*, 12(1), pp. 81–95. doi: 10.21814/lm.12.1.314.

Santos, D., Bick, E. and Wlodek, M. (2020) ‘Avaliando entidades mencionadas na coleção ELTeC-por’, *Linguamática*, 12(2), pp. 29–39.

Santos, D. and Chaves, M. S. (2006) ‘The place of place in geographical IR’, in *Proceedings of GIR06, the 3rd Workshop on Geographic Information Retrieval, SIGIR 2006*. Seatle, pp. 5–8.

Schöch, C. *et al.* (2021) ‘Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives’.

Tally, R. T. (2008) ‘Literary cartography: Space, representation, and narrative’, *Faculty Publications—English*. Available at: https://digital.library.txstate.edu/handle/10877/3932 (Accessed: 24 May 2011).

Valaa, H. *et al.* (2015) ‘Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On the Difficulty of Detecting Characters in Literary Texts’, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 769–774.