Edinburgh University Press
12 (2f) Jackson's Entry
Edinburgh EH8 8PJ
www.euppublishing.com

Dear Author,

Here is a proof of your article to appear in the forthcoming issue of *IJHAC: a Journal of Digital Humanities*.

Please check the proof carefully and send any corrections (quoting page and line references) to Joana Vieira Paulino, the journal managing editor, at the following email addresses no later than 5 days after receipt: ijhac@fcsh.unl.pt

Please remember that you are responsible for correcting your proofs. The proof is sent to you for correction of typographical errors only. Revision of the substance of the text is not permitted, unless discussed with the journal editors. Please answer any queries raised by the typesetter.

EUP Journals Blog
EUP now has a blog where you can post comments about your article and your research (see EUP Journals Blog at http://euppublishingblog.com)

- EUP will tweet each post upon publication and on occasions when it links to news, events or anniversaries
- A link on the journal or book webpage to the blog post will be made live
- Posts will be tagged with key words and themes and used in online search engines though AdWords, adverts and social media campaigns
- Posts will be highlighted in relevant email campaigns and/or printed publicity items
- Authors will be given the permalink to their post and encouraged to share and reference the post among personal networks and through their own blogs and websites.
- Contact the Journals Marketing Manager, Teri Williams (Teri.Williams@eup.ed.ac.uk) for more information.

Open Access
EUP offers green (default) and gold Open Access publication options for your article. More information can be found here: https://www.euppublishing.com/customer-services/open-access

ALCS
In order to ensure that you receive income for secondary uses of your work, including photocopying and digital reproduction, we recommend that you join the Authors' Licensing and Collecting Society (ALCS). Details on how to join the Society can be found at: https://www.alcs.co.uk/What-we-do/Membership-of-ALCS.

EUP Book and Journal Discounts
As a thank you for publishing with Edinburgh University Press, you are entitled to a 40% discount on all EUP books, journal subscriptions and the journal issue containing your paper. Please visit https://www.euppublishing.com/customer-services/authors/author-and-editor-discounts for more information.

Postage: Please note that postage costs are additional and will be charged at current rates.

Ordering: Please contact marketing@eup.ed.ac.uk to place book orders, and journals@eup.ed.ac.uk for journal orders.

We hope all is in order with your proof, but please get in touch if you have any queries or concerns.

Best wishes,

Ann Vinnicombe
Managing Production Editor
Journals Production Department
Edinburgh University Press
Email: Ann.Vinnicombe@eup.ed.ac.uk

Contributor Discount Order Form

EUP book and journal contributors recieve 40% discount on EUP books and journals when ordering directly from us. Journal contributors recieve a 40% discount on additional copies of the issue to which they contributed.

If you would like to order journals and books, please fill out two separate forms.

EDINBURGH University Press

Please deliver the following titles to:

PLEASE PRINT CLEARLY

Customer Details

Name:.....

Email Address: .....

Shipping Address: .....

Postal Code:.....Country:.....

Telephone: .....

Billing Address (if different from Shipping): .....

Postal Code:.....Country:.....

Telephone: .....

Books: edinburghuniversitypress.com

QTY	ISBN	AUTHOR/TITLE	FULL PRICE	DISCOUNT PRICE
			Books total	
			P&P	
			Total	

*Books P&P UK: £2.50 for the first book, 50p per additional book; EU: £3.00 per book; ROW: £5.00 per book

Journal subscriptions: euppublishing.com

ISSN	TITLE	PRINT / ONLINE / P&O	FULL PRICE	DISCOUNT PRICE
			Journal subs total	

Journal single issues

ISSN	TITLE	PRINT / ONLINE / P&O	FULL PRICE	DISCOUNT PRICE
			Journal single issue total	

Payment method

Please do not send cash through the post. To pay by card payment, we will set up a secure payment link. For your own protection, do not send your card details by email. Return book order forms to: **marketing@eup.ed.ac.uk** and journal order forms to: **journals@eup.ed.ac.uk**

Edinburgh University Press Limited
Registered in Scotland no. 139240 Registered address: The Tun - Holyrood Road, 12 (2f) Jackson's Entry, Edinburgh, EH8 8PJ, UK
Registered at Companies House Edinburgh on 9th day of July 1992 Company Registration No. SC139240 Charities No. SC035813

EDINBURGH University Press

Library subscription recommendation form

Complete this section and give to your librarian

Dear Librarian

I recommend _____ as a valuable addition to the Library collection

ISSN _____
eISSN_____
ISBN_____

Name _____

Department _____
Email _____

Reason(s)

☐ USAGE - I will regularly use this journal and recommend articles to colleagues and students

☐ ENHANCEMENT - This will enhance the Library's scholarly collection and benefit research, learning and teaching in the field

☐ CONNECTION - I am a member of the editorial/advisory board and/or a regular contributing author

Additional Comments_____
_____
_____

For more information, to request a quote or to place an order for a journal subscription or collection, email: journals@eup.ed.ac.uk, call us on +44 (0) 131 650 4220 or visit:

www.euppublishing.com

Edinburgh University Press, The Tun - Holyrood Road, 12 (2f) Jackson's Entry, Edinburgh, EH8 8PJ

PLACING GIS AND NLP IN LITERARY GEOGRAPHY: EXPERIMENTS WITH LITERATURE IN PORTUGUESE

DIANA SANTOS and DANIEL ALVES

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, the* Atlas of Literary Landscapes of Mainland Portugal *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 are not 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, such as Fernando Pessoa, for instance,[1] or projects that use literary corpora to study the language and its uses, not the literary features of the works assembled.[2] We can, however, mention recent works on automatic periodization as a promising first step in this direction.[3]

Regarding the use of geographic information systems (GIS) to study Portuguese literature, currently to our knowledge we are only aware of Canosa's study of *Peregrinação*[4] and the studies developed by Alves and Queiroz

International Journal of Humanities and Arts Computing 17.1 (2023): 47–64
DOI: 10.3366/ijhac.2023.0299

www.euppublishing.com/ijhac

about literary landscapes.[5] Although there are many more English-language projects, for instance, that combine computational linguistics with literature,[6] and that use cartography and GIS for literary studies[7] in what has been called 'literary cartography', 'literary geography' or 'literary GIS',[8] we believe that this comparison of two disparate methodologies can still be interesting to a wider audience. Regarding named-entity recognition (NER), we could also mention the Pelagios project and the *Recogito* tool this developed, an environment where NER and close reading came together to enhance the recognition and annotation of place names.[9]

Our aim with the work we describe in this article, developed in the scope of the BILLIG project (Bilateral Lusophone Literature Initiative using GIS and Linguistics), financed by European Economic Area grants,[10] was to assess and produce some first steps in a cross-disciplinary endeavour 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 the *Atlas*) and *Literateca*, the goal was to provide better systems for both projects and contribute to the improvement of knowledge about ways to use natural language processing (NLP) and GIS in literary studies.

1.1 The Atlas

The *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,[11] 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 them.[12]

In order to facilitate the identification of the geographical references contained in this corpus, each literary representation of the 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',[13] 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.[14]

1.2 Literateca

Literateca is an environment for studying literature in Portuguese, started in 2018, with a web interface to literary corpora annotated with linguistic and literary information. It is a direct descendant of *Gramateca*, an environment in which to do corpus-based studies of Portuguese grammar,[15] enhanced with metadata regarding literary works (author, data, literary school, genre, etc.) and with an interface to R-based statistical procedures common in digital literary studies (topic models, correspondence analysis, principal components analysis, among others).[16] Some of the annotation of *Literateca* has in addition been motivated by literary questions.

Although Portuguese is the fifth or sixth most spoken language in the world, and has a huge literature spawning hundreds of years, most digitization projects have focused on canonical and better-known works – that is, quality – and not on breadth and width – that is, quantity. The few projects that have taken a wide grasp of material, such as those done by Google, have produced digital objects of very poor quality. So, almost unbelievably, it proved hard to obtain 100 novels by 80 different authors from Portugal covering the period 1840–1920, which was one of the goals of the COST project 'Distant Reading for European Literary History'.[17] 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 by Brazilian authors.

1.3 Aim and Scope of the Study

In this study we will begin, in Section 2, by comparing generically the two projects and their different assumptions, strengths and weaknesses. Then, in Section 3, we will 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[18] 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 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 the *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 9.7 (5 September 2022), some 40 million words, corresponding to 933 works, ranging from the *Chronicle of the King D. João I* from 1380 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 the *Atlas*, 377 works from 184 writers, published between 1849 and 2019, are included. The main criterion for inclusion of 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 7,751 excerpts in the database, classified in 27 thematic categories and including more than 1.7 million words. All excerpts were then geographically annotated at two levels: the NUTS III (nomenclature of territorial units for statistics) where the landscape description is included; and, where possible, the specific locations that are mentioned or can be identified in the excerpt. In the comparisons discussed below we use only 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.[19]

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 the *Atlas* does), most of them by canonical 19th-century writers such as Eça de Queirós (11 works) and Camilo Castelo Branco (5/6 works). The remaining works were two by Abel Botelho, one by Florbela Espanca, one by Carlos Malheiro Dias and one by Conde de Ficalho. While this undoubtedly constitutes

a very small and skewed sample, 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 amounted to 2,277,577 tokens in *Literateca* (full texts) but just 379,611 tokens in the *Atlas* (excerpts). It is to these that 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 the *Atlas* and to 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. The final goal was to assess the possibility of an automatic location assignment. We will address each of these in turn. However, our 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 between 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 the *Atlas* with geographical coordinates. This relatively large discrepancy (1,517 cases) can be put down to many (independent) factors:

1. Often, the description of the geographical place in *Atlas* is written using its full name, while in context a much shorter name is used – see, for example, *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. Some locations were wrongly classified as such by the automatic parser, such as *D. João V* (a king's name) or *Fado do Bairro Alto* (a name of a song that mentions a neighbourhood in Lisbon) or *Estado* (a general reference to the state, as in the Portuguese state);
3. Some generic locations, like *Portugal*, were not geolocated 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, such as rivers, seas or mountains, they may not be geolocated in the *Atlas*. Others, like (the

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

Kind of difference	**Number of cases**	**%**
Automatic error	116	33
Missing from the *Atlas*	96	28
Different (often more encompassing) description	70	20
Outside Portugal	47	14
Other cases	18	5
Total	347	

river) *Tejo*, were, but this of course raises problems either way – because the literary excerpt may describe the Tejo near Lisbon or the Tejo near Santarém (nearly 80 kilometres to the north), and both have a unique geo-reference 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;[20]

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

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

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 mismatch 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 the *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 1,642 occurrences, improving from 44.7 to 59.8 per cent.

4. ANALYSING 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.

Table 2. The number of place names per excerpt in the *Atlas* according to PALAVRAS-NER.

Number of place names	Number of excerpts
No place name	1,557
1	1,399
2	753
3	394
4	264
5	131
6	68
7	51
8	26
9	19
10	13
10+	34

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 had not, using PALAVRAS-NER, more specifically the features *top* (toponym) and *civ* (city).[21] Table 2 shows an overview of how many different place names were identified per excerpt:

These figures show that a sizeable percentage of text excerpts, classified by the *Atlas* as describing a region or place, do not have named entities (at least as 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:

> From a slightly open window, Vasco da Gama, still a little dizzy, observes the servants hurrying back and forth, 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 theatre plays continue, golden lights are lit. [...] Vasco leaves the party sometime later, pretending tiredness, but with his face lit by hope. The orange trees, under moonlight, in that square of an absolute white, are covered with gold.[22] (Urbano Tavares Rodrigues, *Os campos da promessa*) (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)

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	2,538
1	2,379
2	1,051
3	645
4	373
5	230
6	139
7	91
8	61
9	56
10	32
10+	112

> Between the house and the faraway 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.[23] (Sophia de Mello Breyner Andresen, *Histórias da Terra e do Mar*) (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)

Looking now from the perspective of the *Atlas*, 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 (e.g. *Campos do Mondego* / Mondego Fields, a vague territory between the two Portuguese cities of 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.

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 2,538 with no place name, because it was possible to figure out the larger region the excerpt mentioned or referred to (e.g. 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, who did not register a mentioned place name or location or was unable to identify it.

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

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 causes: there were no place names entities in the excerpt (as discussed in the previous subsection), which consequently required human interpretation to be located, or the named entity recognizer failed.[24]

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 analysed a random set of 50 excerpts where *Atlas* readers and PALAVRAS-NER disagree.

In addition to errors from both sides – owing to different designations and readers' 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 mentioned are *Viana do Castelo/Santa Luzia* and *Lisboa/Estrela*. Estrela is a location in Lisbon, 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 of cities, already registered in the first level of annotation (NUTS III).

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

'Other cases' corresponds to those 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, for example:

> 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 only 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 of an innovative style, Troy falls, Carmo and Trindade fall [. . .])

Carmo and Trindade are two places in Lisbon that were deeply affected by the 1755 earthquake and whose names are popularly used in Portugal to refer to any catastrophic event.

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 interpolating that data with upper-level administrative units where the coordinates were included;[25] comparing lists of place names with already constructed gazetteers where the locations and their respective coordinates are associated with several levels of administrative or other type of territorial units (e.g. from parish to continent level).[26]

Given that we have already uncovered several cases where the two approaches have considerably different results (namely a large number 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 two cases (e.g. 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?

Table 5. Distribution among the most common kinds of places.

Kind of place	**Amount**
City	35,279
Country	27,159
Territory or region	5,180
Street	5,041
Town (*vila*)	4,385
Continent	4,120
Parish (*freguesia*)	3,716
Province	2,457
Religious building	2,418
Mmunicipality	1,832
Public building	1,218
Quarter	1,083
Planet	1,018
Village (*aldeia*)	649

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.[27]

In order to try to answer these questions, we began to undertake a human revision of *Literatec*a's place names, most specifically in the two corpora that include exclusively Portuguese works, namely *Vercial* and *NOBRE*. This revision, still under way, 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 article, 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 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 those that were vague, falling between, for example a city or a municipality, only unambiguous cases).

We tried to automatically geocode the above-mentioned curated/revised (1,952) cases. For the geocoding process (assigning latitude/longitude coordinates to places and addresses), after analysing 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 (the *Atlas* and the Military Charter perform well here) but which would also be able to identify places outside Portuguese territory, and for that the community-driven project *OpenStreetMap* was the most balanced solution. Using *QGIS* we initially found coordinates for 85.7 per cent to place names on our list, which is very high. However, by going through every case, we could certify only 905 cases, resulting in 46.4 per cent.

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

1. in several cases the old orthography, most notably the already-mentioned Peregrinação from the 16th century;
2. the existence of many identical place names in Brazil and Portugal;
3. the existence of many places in the USA 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. any names of streets in Portugal have changed, as is in fact acknowledged in many historical novels, which present both old and newer names.

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 list of places mentioned in only one work that we expect to find several different imaginary places, or others so small and irrelevant (from 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 to directly reference physical places, but offer a wide spectrum of uses often only loosely connected with the place, as illustrated by the following random examples of the occurrence of the word *Lisboa* (Lisbon) 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 notícias 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 it 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 Municipal Registry)

While this is well known by those who work in the field of named entity recognition,[28] 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 9.7 of *Literateca*, 5 September 2022) only 1,902 different place names georeferenced, corresponding to 86,190 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.[29]

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.[30]

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

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 library and is invoked by *Literateca*'s interface.[31] Figure 1 shows two maps of which Portuguese cities or towns are mentioned in the works by two well-known 19th-century authors, Camilo Castelo Branco and Eça de Queirós.

We chose these two authors because the majority of works by them are not 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 different 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?[32]

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.

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 landscape 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 of being described in language), as overviewed by Santos and Chaves in 2006.[33] All these things must be taken into consideration if we want to understand the place of places in literature.

In this case study we have taken some of these challenges seriously: first, we produced a detailed comparison of the results from two computational technologies (developed by the 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 taking 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, to Inês Lucas for her thorough revision of place names in *Literateca*, and to Marcin Wlodek for the geocoding work in *Literateca*. This work was partially supported by the Fund for Bilateral Relations (EEA Grants) of the Financial Mechanism Programme (2014–21) 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, and ILOS, at the University of Oslo, for a small grant for geocoding activities.

ORCID

Diana Santos https://orcid.org/0000-0002-3108-7706
Daniel Alves https://orcid.org/0000-0002-3541-8197

END NOTES

[1] M. Portela, '"Nenhum problema tem solução": Um arquivo digital do livro do desassossego', *MATLIT: Materialidades da Literatura*, 1.1 (2013), 9–33.

[2] E. dos S. Rodrigues, C. Freitas and V. Quental, 'Análise de inteligibilidade textual por meio de ferramentas de processamento automático do português: Avaliação da Coleção Literatura para Todos', *Letras de Hoje*, 48.1 (2013), 91–9.

[3] D. Santos, E. Pires, J. M. Lopes, R. S. Fuão and C. Freitas, 'Periodização automática: estudos linguístico-estatísticos de literatura lusófona', *Linguamática*, 12.1 (2020), 81–95, https://doi.org/10.21814/lm.12.1.314.

[4] A. X. Canosa, 'Referentes por coordenadas e georreferências relativas das entidades geográficas mencionadas na Peregrinação', in C. Pazos-Alonso et al., ed., *De Oriente a Ocidente: Estudos da Associação Internacional de Lusitanistas*, vol. I (Coimbra, 2019), 11–34.

[5] 'Studying urban space and literary representations using GIS: Lisbon, Portugal, 1852–2009', *Social Science History*, 37.4 (2013), 457–81, https://doi.org/10.1215/01455532-2346861; 'Exploring literary landscapes: From texts to spatiotemporal analysis through collaborative work and GIS', *International Journal of Humanities and Arts Computing*, 9.1 (2015), 57–73, https://doi.org/10.3366/ijhac.2015.0138.

[6] See, e.g., M. Mahlberg, C. Smith and S. Preston, 'Phrases in literary contexts: Patterns and distributions of suspensions in Dickens's novels', *International Journal of Corpus Linguistics*, 18.1 (2013), 35–56; M. C. Ardanuy and C. Sporleder, '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, 2014), 31–9; H. Vala et al., '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', *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, 769–74; D. Bamman, S. Popat and S. Shen, 'An annotated dataset of literary entities', *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1 (Long and Short Papers), 2019, 2138–44.

[7] See R. Heuser, F. Moretti and E. Steiner, 'The emotions of London', *Literary Lab Pamphlet*, 13 (2016).

[8] R. T. Tally, 'Literary cartography: Space, representation, and Narrative', *Faculty Publications –English*, 2008, https://digital.library.txstate.edu/handle/10877/3932; B. Piatti, A. Reuschel and L. Hurni, 'Literary geography – or how cartographers open up a new dimension for literary studies', *Proceedings of the 24th International Cartography Conference* (Santiago, 2009), http://icaci.org/files/documents/ICC_proceedings/ICC2009/html/nonref/24_1.pdf; D. Cooper and I. N. Gregory, 'Mapping the English Lake District: A literary GIS', *Transactions of the Institute of British Geographers*, 36.1 (2011), 89–108.

[9] V. Vitale, P. d. S. Cañamares, R. Simon, E. Barker, L. Isaksen and R. Kahn, 'Pelagios – connecting histories of place. Part I: Methods and tools', *International Journal of Humanities and Arts Computing*, 15.1–2 (2021), 5–32.

[10] *BILLIG*, https://billig.fcsh.unl.pt/, accessed on 5 September 2022.

[11] Tally, 'Literary cartography'.

[12] Alves and Queiroz, 'Studying urban space and literary representations using GIS: Lisbon, Portugal, 1852–2009'; Alves and Queiroz, 'Exploring literary landscapes: From texts to spatiotemporal analysis through collaborative work and GIS'.

[13] M. L. Jockers, *Macroanalysis: Digital methods and literary history* (Illinois, 2013); F. Moretti, *Distant Reading* (London, 2013).

[14] See, e.g., *A Literary Atlas of Europe*, http://www.literaturatlas.eu/en/; *Digital Literary Atlas of Ireland, 1922–1949*, http://cehresearch.org/DLAI/; *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 5 September 2022.

[15] D. Santos, 'Gramateca: Corpus-based grammar of Portuguese', in J. Baptista et al., ed., *Computational Processing of the Portuguese Language, 11th International Conference, PROPOR 2014, São Carlos/SP, Brazil, 6–8 October 2014, Proceedings* (Heidelberg, 2014), 214–19.

[16] See, for more details, D. Santos, 'Literature studies in *Literateca*: Between digital humanities and corpus linguistics', in M. Doerr et al., ed., *Humanists and the digital toolbox: In honour of Christian-Emil Smith Ore* (Oslo, 2019), 89–109.

[17] See D. Santos, E. Bick and M. Wlodek, 'Avaliando entidades mencionadas na coleção ELTeC-por', *Linguamática*, 12.2 (2020), 29–39; C. Schöch et al., 'Creating the European Literary Text Collection (ELTeC): Challenges and perspectives', *Modern Languages*, 2021.

[18] E. Bick, 'The parsing system "Palavras": Automatic grammatical analysis of Portuguese in a constraint grammar framework', PhD thesis, Aarhus University, 2000; E. Bick, 'Functional aspects in Portuguese NER', in R. Vieira et al., ed., *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006 (PROPOR'2006)* (Berlin, 2006), 80–9.

[19] LITESCAPE.PT, https://litescape.ielt.fcsh.unl.pt/, accessed on 5 September 2022.

[20] We are obviously aware of other forms of geocoding and its difficulties. 'Mapping [. . .] enforces a location in a space, while text descriptions allow flexible, fluid, and fuzzy associations of features with locations' (M. Yuan, 'Spatializing text for deep mapping', in D. J. Bodenhamer, J. Corrigan and T. M. Harris, eds, *Making deep maps: Foundations, approaches, and methods* (New York, 2022), 50–64, here cited at 62). None of the existing geocoding systems can solve the question of where the River Tejo in Lisbon is versus where the River Tejo in Santarém is. How finely grained should a literary annotation of rivers or mountains be?

[21] E. Bick, 'Automatic semantic role annotation for Portuguese', *TIL, V Workshop em Tecnologia da Informação e da Linguagem Humana* (Rio de Janeiro, 2007), 1715–19.

[22] 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.

[23] 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.

[24] We do not imply that PALAVRAS-NER would be the only possible system to use here. However, it was the best system in HAREM, and also performed well in a recent comparison focused on literary texts. See F. Frontini, C. Brando, J. Byszuk, I. Galleron, D. Santos and R. Stanković, 'Named entity recognition for distant reading in ELTeC', in C. Navarretta and M. Eskevich, eds, *CLARIN Annual Conference 2020, Proceedings*, 5–7 October 2020, 37–41.

[25] D. Alves, 'Using a GIS to reconstruct the nineteenth century Lisbon parishes', *Humanities, Computers and Cultural Heritage. Proceedings of the XVIth International Conference of the Association for History and Computing* (Amsterdam, 2005), 12–17; D. Alves, '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 (2017), 139–58, https://doi.org/10.1080/2373518X.2017.1329194; Alves and Queiroz, 'Studying urban space and literary representations using GIS: Lisbon, Portugal, 1852–2009'.

[26] R. Mostern, H. Southall and M. L. Berman, eds, *Placing names: Enriching and integrating gazetteers* (Indianapolis, 2016).

[27] See http://lotrproject.com/map/#zoom=3\&lat=-1315.5\&lon=1500\&layers=BTTTTT for visualization of a fictional universe, last accessed 5 September 2022.

[28] See the HAREM evaluation contest for Portuguese. D. Santos et al., 'HAREM: an advanced NER evaluation contest for Portuguese', in N. Calzolari et al., eds, *Proceedings of LREC 2006* (ELRA, 2006), 1986–91.

[29] By interacting with the *Literateca* corpus through the AC/DC interface – see the following note.

[30] https://www.linguateca.pt/acesso/corpus.php?corpus=LITERATECA, last accessed 5 September 2022.

[31] https://leafletjs.com/, last accessed 5 September 2022.

[32] A similar hypothesis was advanced regarding the city of Lisbon and its literary representations in Alves and Queiroz, 'Studying urban space and literary representations using GIS'.

[33] D. Santos and M. Chaves, 'The place of place in geographical IR', *Proceedings of GIR06, the 3rd Workshop on Geographic Information Retrieval, SIGIR 2006* (Seattle, 2006), 5–8.