Prototyping an Atlas of Early Modern English Drama: An Experiment on DraCor Data

Luca Giovannini[1], Andreas Wagner[2]
[1] University of Potsdam, Germany – giovannini@uni-potsdam.de
[2] University of Erlangen-Nuremberg, Germany - andreas.w.wagner@fau.de

ABSTRACT (ENGLISH)
This short contribution leverages the recently released *English Drama Corpus* from the *DraCor* project (Fischer et al. 2019) to create an interactive map of geographical references in early modern English drama by means of Named Entity Recognition (NER) tools. The prototype allows to visualise interesting patterns in the playwrights' use of locations while underscoring the need of improved NLP tools to effectively tackle geospatial research questions about early modern literature.

Keywords: early modern drama; spatial humanities; geocoding; digital cartography

ABSTRACT (ITALIANO)
Un prototipo di atlante per il dramma inglese della prima modernità. Questo contributo presenta una mappa interattiva dei riferimenti geografici contenuti all'interno dell'*English Drama Corpus*, recentemente pubblicato all'interno della piattaforma *DraCor* (Fischer et al. 2019). Il prototipo mostra alcuni interessanti prospettive nell'uso dei luoghi da parte dei drammaturghi, ma sottolinea al contempo la necessità di sviluppare nuovi e più avanzati modelli per il trattamento automatico di testi della prima modernità.

Parole chiave: letteratura drammatica; geocodifica; cartografia digitale

1. BACKGROUND
Best exemplified by works like the *Atlas of Literature* (Bradbury 1996) or the *Atlas of the European Novel, 1800-1900* (Moretti 1999), the 'spatial turn' the humanities experienced in the last 20 years has long invested literary studies as well, resulting in multiple projects that utilise cartographic tools to explore texts in their relations with real or fictional places. Furthermore, the large availability of digitised corpora and the constant improvements in Named Entity Recognition (NER) techniques (cf. Keraghel, Morbieu and Nadif, 2024) have naturally given many of these projects a digital edge (cf. Perenič 2014, Mitchell 2017). While the geographical dimension of early modern English drama has often been investigated with respects to its performance locations, [1] quantitative approaches to the plays' contents from this angle are still less developed. This contribution aims to constitute a first exploratory step in that direction by prototyping and showcasing a map of places mentioned in a large number of early modern English texts. In doing so, this proof of concept also represents one of the first research applications of the recently released *English Drama Corpus* (*EngDraCor*). This collection situates itself within the wider *DraCor* ecosystem (Fischer et al. 2019), i.e., a large collection of TEI-encoded, machine-actionable corpora of dramatic texts from different literary traditions. *EngDracor* currently features 434 English plays from the 1550s to the 1650s, mostly onboarded from the *Early Print* initiative (Mueller 2012), and can serve as a valuable resource for computational literary scholars, enabling the exploration of diverse research questions related to the structure, themes, and stylistic elements of the works.

2. DATA MODELLING
In our experiment, we first utilised the *DraCor* API [2] to iteratively extract and store all spoken text from the plays. [3] Performing NER on early modern texts, however, proved to be quite challenging: the procedures surveyed in Humbel et al. (2021) were often tailored to specific use cases and rather difficult to generalise, even though we ended up implementing a pipeline somehow similar to Gregory and Hardie (2021). Directly deploying appropriate linguistic models was also hard: MacBERTh (Manjavacas and

[1] See e.g. the well-established *Records of Early English Drama* (*REED*) project (https://ereed.org).
[2] https://dracor.org/doc/api.
[3] This is the format of the API call used: https://dracor.org/api/v1/corpora/eng/plays/{slug}/spoken-text. The slug is the unique identifier for each DraCor play, usually made up by the author's surname and the shortened title (e.g. *marlowe-doctor-faustus*).

Fonteyn 2022), a BERT model fine-tuned on early modern texts, lacks a module devoted to NER, while MonadGPT (Langlais 2023) is a conversational agent quite unsuited to our task. [4] On the other hand, modernisation of spelling to improve NER performance, e.g. through software like VARD (Baron and Rayson 2008), seemed not a sustainable option in our case.

Given the lack of satisfactory out-of-the-box solutions, we opted for experimenting with modern standard named entity recognition pipelines, testing the performances of the packages *spaCy* and *stanza*. [5] Despite some preprocessing steps to improve general reliability, the first package performed quite poorly, with close reading of results revealing a very low precision rate (only about 20% of the items retrieved were actual geographical entities [6]). Results with *stanza* were more satisfactory, but still required extensive post-correction. To this aim, we experimented with rule-based solutions, involving dictionaries of geographical entities and gazetteers (e.g. the *World Historical Gazetteer*, cf. Jakacki 2024), but without satisfying results. Eventually, after an extended trial-and-error phase, we settled for the most time- and result-efficient pipeline, which involved fixing most recurrent errors through the 'clustering' and 'reconciliation' functions of the OpenRefine software [7] and then performing a thorough LLM-assisted clean-up (via GPT4, Achiam et al. 2023). In this phase, we tried to lay the groundwork for the later geocoding by normalising spellings (e.g. *Brytaine* to *Britain*), retracing back adjectives to nouns (e.g. *Tartarian* to *Tartary*), and removing common and proper names which slipped in.

We then geocoded the list of locations for each play via the Nominatim geocoder, based on OpenStreetMap data, [8] and performed an extensive manual clean-up of the results, correcting further errors and rectifying wrong coordinates. Ultimately, we assigned every item to one of these four categories: 'city' (including elements such as roads, squares, and neighbourhoods), 'country' or 'region', 'continent', and 'natural feature' (like mountains or rivers).

3. VISUALISATION AND DISCUSSION

To present our results, a React web application was been built upon the database and is now available at https://atlas-emed.github.io. The website displays two main components: a globe map view on the left and a content view on the right (Figure 1). The map view hosts locally loaded *geojson* files that are built whilst deployment from the database file, with one *geojson* file created and loaded for each layer. When a user clicks on one of the more than 500 locations featured in the map, the app displays a list of plays associated to it in the content view (Figure 2). This takes the form of a series of "DramaCard" components, which in turn fetch and display some metadata (author name, play genre, year) from the *DraCor* API. By selecting a card, the map is updated to show only locations tied to the that text, allowing a more granular exploration of geographical references.

On a methodological level, it should be noted that the extraction conducted here is based on the spoken text of the plays, and not on the <set> element from the TEI markup. Consequently, the map does not reflect the actual setting of the works, as done e.g. in the *Database of German-Language One-Act Plays 1740–1850* (Çakir and Fischer 2022), [9] but rather attempts to register all geographic mentions within them.

Even a quick glance at this map provides some initial insights into the "cartographic imagination" (Smith 2016) of early modern English playwrights. As expected, the British peninsula – and in particular the metropolitan area of London, with its iconic alleys, squares, and roads – remains central. Two other main areas of interest coincide with the Italian and Hellenic peninsulas: this can be easily explained by the constant influence of classical (ancient Greek and Roman) motifs and tropes, and by the Italian context being a net exporter of narrative materials towards England. Accordingly, cities most often mentioned – after European powerhouses such as London and Paris – are Renaissance centres like Venice, Naples, and Florence. Smaller clusters of often-mentioned locations can be recognised also in the main western states

---

[4] From the official documentation: "MonadGPT is a finetune of Mistral-Hermes 2 on 11,000 early modern texts in English, French and Latin, mostly coming from EEBO and Gallica".

[5] https://spacy.io; https://stanfordnlp.github.io/stanza. We also wanted to test the state-of-art *BookNLP* package (https://github.com/booknlp/booknlp), whose entity tagging model is trained on a large dataset of literary works (Bamman et al., 2019), but it is currently not working due to a upstream dependency issue which has not yet been fixed (see https://github.com/booknlp/booknlp/pull/25).

[6] We employed here the smallest English-language model, *en_core_web_sm* (https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-3.7.1).

[7] https://openrefine.org.

[8] https://nominatim.org.

[9] https://einakter.dracor.org/locations.

(France, Spain, Germany) and in the Middle East, due to the presence of Biblical-themed contents. Nonetheless, the boundaries of the world as imagined by the *EngDraCor* authors extend way beyond Europe, stretching from the recently encountered West Indies to the near-mythological steppes of Scythia and Tartary and as far as Japan.

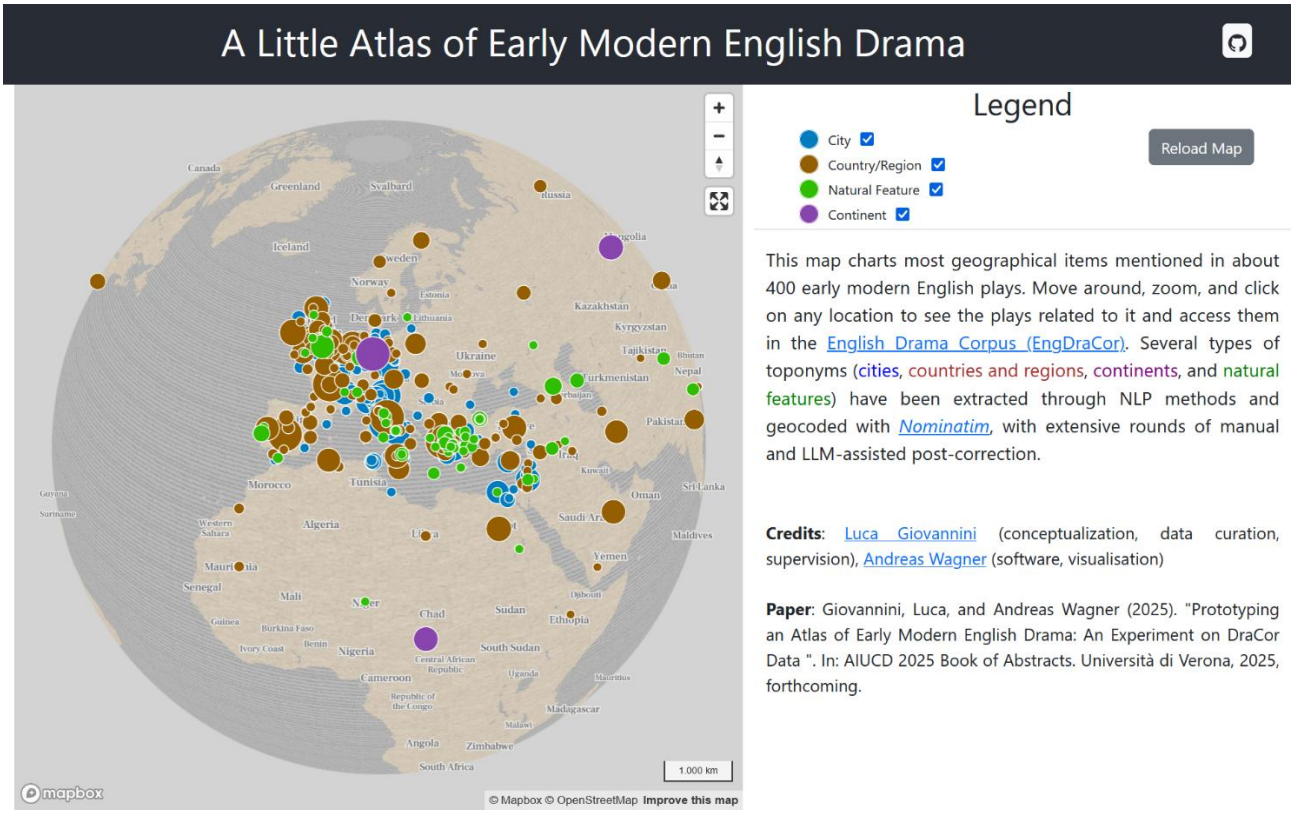

Figure 1. Homepage of the *Little Atlas of Early Modern English Drama* (atlas-emed.github.io).

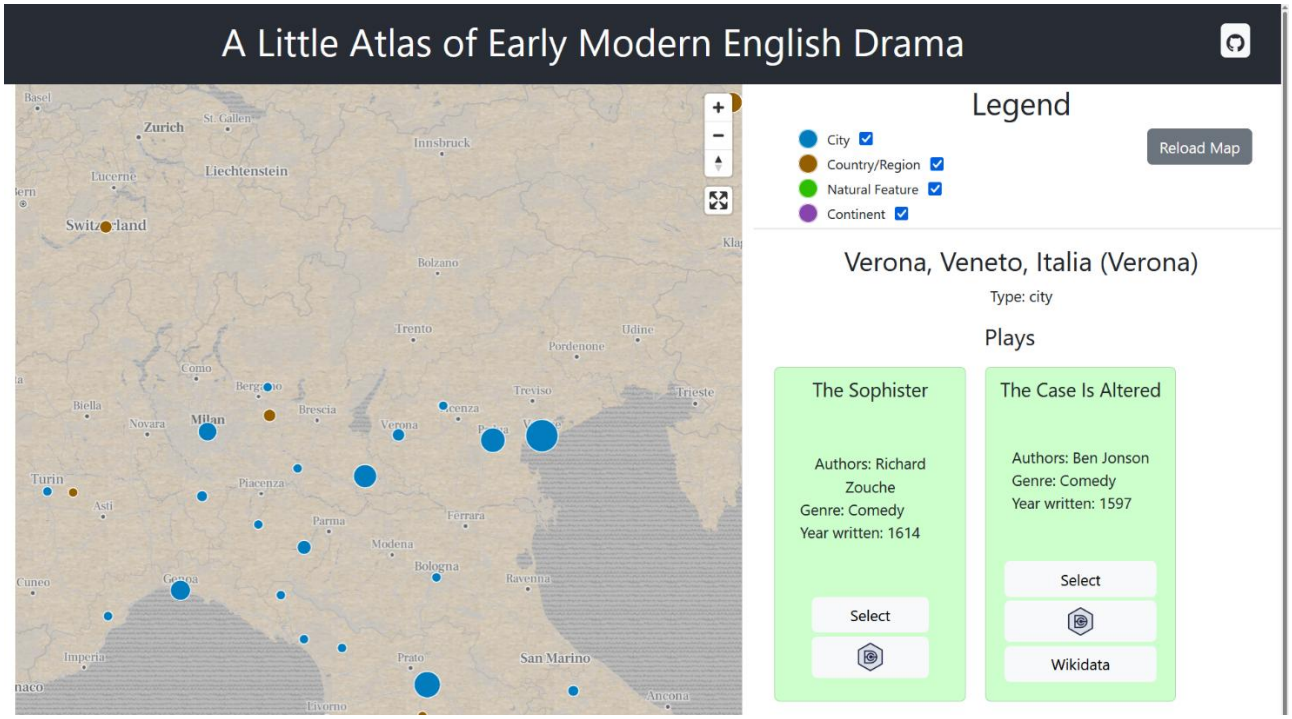

Figure 2. The *Little Atlas* displaying cards for *EngDraCor* plays where Verona is mentioned.

While the map might prove useful for digital English studies, and especially in teaching contexts, this contribution is however mostly meant to start a broader conversation about the need for tailored NLP tools in early modern literary studies. The pipeline followed here, while ultimately delivering acceptable results, was indeed too long and cumbersome to be easily redeployed on other early modern texts. The lack of more reliable and scalable solutions for NER tasks hindered a swift extraction of geographic items, and the results after the extensive manual and AI-powered post-correction were still not on par with the accuracy of entity taggers for modern languages. Given the importance of early modern sources within computational literary studies and digital humanities at large, it seems thus necessary to work towards improving NER tools for such texts. At the same time, one should also accept that, when dealing with historical geographical names, a certain degree of fuzziness is unavoidable, since one often works with entities whose boundaries have changed, evolved, or shifted over time.
From a research-oriented, *DraCor*-centric perspective, instead, the next step would naturally involve the reproduction of this experiment on the other *DraCor* corpora, which feature more recent texts that modern NLP tools may handle more effectively. In this way, it might be possible to begin investigating geographical discourse from a transnational perspective, e.g. by exploring the significance of locations which occur frequently in different dramatic traditions and at different times.

REFERENCES

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). GPT-4 Technical Report. arXiv Preprint. DOI: https://doi.org/arXiv:2303.08774.

Bamman, D., Popat, S., & Shen, S. (2019, June). An Annotated Dataset of Literary Entities. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 2138–2144).

Baron, A., & Rayson, P. (2008). VARD 2: A Tool for Dealing with Spelling Variation in Historical Corpora. In Proceedings of the Postgraduate Conference in Corpus Linguistics. Aston University, Birmingham, UK. URL: https://ucrel.lancs.ac.uk/people/paul/publications/BaronRaysonAston2008.pdf.

Bradbury, M. (Ed.). (1996). The Atlas of Literature. De Agostini.

Fischer, F., Börner, I., Göbel, M., Hechtl, A., Kittel, C., Milling, C., & Trilcke, P. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In Proceedings of DH2019: "Complexities", Utrecht, July 9–12, 2019. Utrecht University. DOI: https://doi.org/10.5281/zenodo.4284002.

Gregory, I. N., & Hardie, A. (2011). Visual GISting: Bringing Together Corpus Linguistics and Geographical Information Systems. Literary and Linguistic Computing, 26(3), 297–314. DOI: https://doi.org/10.1093/llc/fqr022.

Humbel, M., Nyhan, J., Vlachidis, A., Sloan, K., & Ortolja-Baird, A. (2021). Named-Entity Recognition for Early Modern Textual Documents: A Review of Capabilities and Challenges with Strategies for the Future. Journal of Documentation, 77, 1223–1247. https://doi.org/10.1108/JD-02-2021-0032.

Jakacki, D. K. (2024). Review: World Historical Gazetteer. Reviews in Digital Humanities, 5(5). https://doi.org/10.21428/3e88f64f.6bceb2bf.

Keraghel, I., Morbieu, S., & Nadif, M. (2024). A survey on recent advances in named entity recognition. arXiv preprint. DOI: https://doi.org/10.48550/arXiv.2401.10825.

Langlais, P. C. (2023). MonadGPT. URL: https://huggingface.co/Pclanglais/MonadGPT.

Manjavacas, E., & Fonteyn, L. (2022). Adapting vs. Pre-Training Language Models for Historical Languages. Journal of Data Mining & Digital Humanities NLP4DH. DOI: https://doi.org/10.46298/jdmdh.9152.

Mitchell, P. (2017). Literary Geography and the Digital: The Emergence of Neogeography. In The Routledge Handbook of Literature and Space (pp. 85–94). Routledge. https://doi.org/10.4324/9781315745978-8.

Moretti, F. (1999). Atlas of the European Novel: 1800–1900. Verso.

Mueller, M., and the Early Print team. (2012). EarlyPrint. URL: https://earlyprint.org.

Perenič, U. (2014). An Overview of Literary Mapping Projects on Cities: Literary Spaces, Literary Maps and Sociological (Re)Conceptualisations of Space. Neohelicon, 41(1), 13–25. DOI: https://doi.org/10.1007/s11059-013-0226-5.

Smith, D. K. (2016). The Cartographic Imagination in Early Modern England: Re-Writing the World in Marlowe, Spenser, Raleigh and Marvell. Routledge.