Named Entity Recognition for Distant Reading in ELTeC

Francesca Frontini
Praxiling CNRS
Université Paul-Valéry Montpellier 3
francesca.frontini@univ-montp3.fr

Carmen Brando
CRH,
EHESS, Paris
carmen.brando@ehess.fr

Joanna Byszuk
Institute of Polish Language
Polish Academy of Sciences
joanna.byszuk@ijp.pan.pl

Ioana Galleron
LaTTiCe CNRS
Université Sorbonne Nouvelle - Paris 3
ioana.galleron@sorbonne-nouvelle.fr

Diana Santos
Linguateca &
University of Oslo
d.s.m.santos@ilos.uio.no

Ranka Stanković
University of Belgrade
ranka.stankovic@rgf.bg.ac.rs

Abstract

The "Distant Reading for European Literary History" COST Action, which started in 2017, has among its main objectives the creation of an open source, multilingual European Literary Text Collection (ELTeC). In this paper we present the work carried out to manually annotate a selection of the ELTeC collection for Named Entities, as well as to evaluate existing NER tools as to their capacity to reproduce such annotation. In the final paragraph, points of contact between this initiative and CLARIN are discussed.

1 Introduction

The Distant Reading for European Literary History (COST Action CA16204) kicked off in 2017 with the goal of using computational methods of analysis for large collections of literary texts. The objective is to establish shared practices in the application of innovative computational methods, while at the same time reflecting on the impact that such methods have on our capacity to raise and answer new questions about literary history and theory. More details are to be found on the Action's website[1].

One of the most ambitious deliverables of this COST Action is the creation of a multilingual open source collection, named European Literary Text Collection (ELTeC). In its final version, the corpus will contain at least 10 linguistically annotated subcollections of 100 novels per language (1840-1920[2]), that is at least 1,000 full-text novels. To make subcollections representing particular languages as comparable as possible, the novels are selected to represent a) various types of novels as to their length: short stories, epic novels, b) five twenty-year time periods within the examined time span, c) text of various levels of canonicity, as judged by the number of reprints, and d) as equal as possible ratio of female and male authors. As of now, a first version of the ELTeC corpus has been collected and published with a light TEI encoding[3], with more language collections to be included in the further releases throughout the duration of the Action. Since obtaining more works for some language collections is possible, the Action also plans the publication of the extended ELTeC (including more texts per language or ones published slightly before the assumed time span) estimated to take the total number of full-text novels to at least 2,500.

A case study on Named Entity annotation was carried out in the Working Group 2 "Methods and Tools" (WG2), with an aim of establishing common annotation guidelines – which are specifically oriented to answering a set of scholarly research questions identified by the Action participants – and testing a selection of NER tools, in order to assess their capacity of automatic reproduction of such annotation, a task crucial for annotating corpora of this and bigger sizes.

In this abstract we shall introduce the desiderata for the NE annotation, the current state of the multilingual NE corpus and of the annotation, describe the evaluation framework and provide preliminary results for a selection of NE systems. Finally, we will discuss the relevance of these activities with respect

[1] https://www.distant-reading.net

[2] Chronological limits are due to constraints related to copyright and availability of quality full texts.

[3] https://distantreading.github.io/ELTeC/index.html

to the CLARIN community as well as to the CLARIN services and tools, with some ideas for possible collaboration.

2 Developing the NE layer of the ELTeC corpus

2.1 Desiderata and annotation set

NER is a well known task in NLP, and there are several sets of guidelines for performing NER annotation (see among others (Nadeau and Sekine, 2007; Chinchor, 1998; LDC, 2008; Santos et al., 2006; Zupan et al., 2017)), establishing a number of categories and rules for creating annotated corpora, so that automatic algorithms can try to replicate such annotation. However, NER is a preliminary module for other tasks, such as information extraction. The philosophy guiding the definition of the best known NE guidelines is mostly geared towards very specific scenarios, such as extraction of events in news or scientific texts. Their application to literary texts is therefore not straightforward, despite current attempts at producing new annotation guidelines and datasets (Bamman et al., 2019).

Within the context of the COST action, WG3[4] came up with a set of research questions in order to help with the definition of a more targeted set of annotations. A first set of desiderata emanates from the idea that a novel is an epic set in the private space of a bourgeois home, something which demands researchers to be able to detect indicators of social structure and roles, such as honorifics, names of professions, etc. Another set of research topics touches upon questions about identity, otherness, but also the distinction between urban and rural spaces, which require the annotation of demonyms, as well as a higher granularity in the annotation of toponyms, to facilitate detecting different types of locations (cities vs villages and countryside)[5]. Finally, questions about cultural references, role models and cosmopolitanism can only be answered if references to works of art, authors, folklore and periodical publications are detected.

Such considerations led to the inclusion of categories such as demonyms (DEMO), professions and titles (ROLE), works of art (WORK) in the tagset, alongside more canonical categories, such as person names (PERS), places (LOC), events (EVENT), organisations (ORG) (see Table 1). However, this annotation was later simplified to avoid nested annotations and overlap.

	DEMO	EVENT	LOC	ORG	OTHER	PERS	ROLE	WORK
cze	163	5	275	0	0	1150	454	0
deu	66	2	323	12	0	973	458	4
eng	56	7	198	37	0	1184	203	25
fra	77	3	262	22	128	900	244	18
hun	29	7	152	20	0	1091	367	7
nor	4	8	83	25	3	990	201	10
por1	17	9	351	19	0	940	490	54
por2	34	1	256	30	7	1059	347	7
slv	133	54	336	37	0	1230	620	2
srp	121	18	185	11	0	985	301	4
	700	114	2425	213	138	10514	3685	131

Table 1: Data on the manually NE-annotated corpus.

2.2 Current state of the corpus

The NE annotation of the corpus is part of the plan for the so called level 2 annotation, which will also include morpho-syntactic and direct speech annotation. At this moment, WG2 carries out the NE annotation for a subset of languages: Czech (cze), German (deu), English (eng), French (fra), Hungarian (hun), Norwegian (nor), Portuguese[6] (por1, por2), Slovene (slv), and Serbian (srp). For each language

[4]Working Group 3, dedicated to Literary Theory and History `https://www.distant-reading.net/wg-3/`

[5]At this stage Entity Linking is not envisaged for our corpus, and in any case it would not easily solve the problem of distinguishing between types of populated places for past or fictional locations in novels.

[6]For Portuguese, two sets of novel excerpts are available: por1 comprises canonical novels, in modern ortography, while por2 was created from novels with old ortography, non-canonical.

involved, a sample collection was prepared from the novels already available in ELTeC. Every language-specific collection contains 20 files, each of which is composed of excerpts from one novel. Each file is made up of five ~400-word passages randomly selected from the novel.

The annotation teams worked on the same set of guidelines, using the BRAT[7] annotation tool, but of course some adaptations were necessary due to differences between languages. When two persons worked on the same collection, we performed cross-checking. The latest version of the annotated NE corpus is available online, together with the latest version of the annotation guidelines[8]. For more information we refer also a recent paper (Stankovic et al., 2019) presented at the DH Budapest 2019 conference. The results of the latest round of annotations (May 2019) are presented in Table 1.

2.3 Testing automatic NER

Another important activity of WG2 is testing existing tools to assess their capacity to reproduce the envisaged annotation. At this stage testing is performed without previous domain adaptation, and with a preference for tools that are easy to install and use. The rationale behind this choice is to evaluate whether literary scholars without advanced technical skills will be able to use existing NLP technologies in their research.

In this study, we focused on four collections only, in English, French, Portuguese and Serbian, these being the languages the authors of this paper are most familiar with. This first evaluation round allowed us to develop a common evaluation set up. For each collection, we tested two tools: one common for all (spaCy[9]), and another one language specific (Stanford-NER for English[10], SEM for French[11], PALAVRAS-NER for Portuguese (Bick, 2006) and SrpNER for Serbian). BRAT outputs were compared to annotations produced by these tools by using a shared evaluation script, thus guaranteeing the consistency with the agreed upon strategy for identifying hits and misses in terms of entity detection.

In this first round the evaluation of string detection was strict (segments must match exactly); but we are planning to perform a second round with relaxed evaluation. Only the PERS and LOC tags were mapped to similar categories of NER tools, and evaluated. The performance of the tools was evaluated separately for each tag.

2.4 NER results and analysis

Current evaluation results are presented in Table 2. These preliminary results show the difficulty of the task of NE-annotation of literary novels. A strict evaluation of detection is often penalising for PERS, because of honorifics which we chose to include in our annotation, and is further complicated by the fact that the XML annotated input was processed as such by tools which often expect plain text[12]. In most cases, LOC seems to be less problematic for the pre-trained models. We follow with some remarks for each specific language.

Portuguese: It was expected that off-the-shelf Named Entity recognisers that were developed for modern Portuguese would perform significantly worse in the *por2* collection. This was confirmed for PALAVRAS-NER, which showed lower performance for *por2*, but not for spaCy. This NER system had considerably lower results than PALAVRAS-NER, but no significant performance drop between the two sets (in fact, for PERS it fares even better for *por2*). We believe this is because only the easy cases are catered for by spaCy, and those cases do not depend too much on ortography.

English: In spaCy, tokenization issues due to TEI XML tags included in tokens account for 16% of PERS and 38% for LOC errors. It also misses a lot of PERS (21%) due to undetected honorifics (Mr/Mrs/Miss). Stanford NER has better precision for LOC than spaCy and has no problems with TEI XML tags. For both models, some additional training and fine tuning would be needed for better performances.

[7] `https://brat.nlplab.org/` see also (Stenetorp et al., 2012).

[8] `http://brat.jerteh.rs/#/eltec-simplified/`

[9] We used the out-of-the-box model for all languages except for Serbian, for which none was available.

[10] `https://nlp.stanford.edu/software/CRF-NER.shtml`

[11] `https://www.lattice.cnrs.fr/sites/itellier/SEM.html`

[12] This evaluation scenario is realistic in the context of Digital Literary Studies, where digital editions with a minimal TEI encoding are to be further enriched using NLP tools, to be then analysed by literary scholars.

French: The manual corpus contains many PERS and fewer LOC annotations. However, spaCy-fra annotates too many LOC hence the low precision for this category, and SEM-fra annotates too few PERS, hence the low recall for this category. In general, there are many detection issues in particular with entities including determiners. Despite the better performance, spaCy-fra has an odd behaviour due to parsing, tokenising, presence of XML tags, capital letters in the beginning of sentences, and it recognises entities composed of more than 4 tokens, which are actually rare.
Serbian: SrpNER is a rule based system and the best NER tool for Serbian. 11% of missing annotations are related to PERS multiword units with honorifics "g./gđa." (Mr/Mrs). For LOC, missing annotations are celestial bodies, names of the streets and facilites. With the Serbian spaCy model, about 7% of errors are related to TEI XML tags, much fewer than with the English model, as the training set included TEI annotated text. Further improvements are foreseen.

	Cat	Correct	Missing	Spurious	Precision	Recall	Excess
SEM-fra	LOC	73	112	84	0.465	0.395	0.535
	PERS	82	512	115	0.416	0.138	0.584
SPACY-fra	LOC	103	78	468	0.180	0.569	0.820
	PERS	329	194	297	0.526	0.629	0.474
PALAVRAS-por1	LOC	223	63	44	0.835	0.780	0.165
	PERS	816	90	86	0.905	0.901	0.095
SPACY-por1	LOC	225	84	440	0.338	0.728	0.662
	PERS	465	256	374	0.554	0.645	0.446
PALAVRAS-por2	LOC	151	67	91	0.624	0.693	0.376
	PERS	857	133	285	0.750	0.866	0.250
SPACY-por2	LOC	157	57	396	0.284	0.734	0.716
	PERS	569	236	393	0.591	0.707	0.409
Stanford-eng	LOC	98	100	126	0.438	0.495	0.563
	PERS	649	535	399	0.619	0.548	0.381
SPACY-eng	LOC	98	100	170	0.366	0.495	0.634
	PERS	536	648	240	0.691	0.453	0.309
SrpNER-srp	LOC	107	78	19	0.849	0.578	0.151
	PERS	718	267	158	0.820	0.729	0.180
SPACY-srp	LOC	57	128	104	0.354	0.308	0.646
	PERS	553	432	315	0.637	0.561	0.363

Table 2: Results of the strict evaluation, per language and category.

3 Relationship to CLARIN and conclusion

The creation, annotation and publication of the ELTeC corpus falls clearly within the scope of CLARIN activities, and ties with CLARIN are already assured by individual participants in the COST action (such as Tomaž Erjavec, Slovenian National Coordinator). Nevertheless, the presentation of the corpus at the CLARIN conference represents an opportunity to discuss further avenues for collaboration. Following the FAIR principles and CLARIN best practices, we identify the following points.

The final version of the manually checked NE subset of the ELTeC corpus, as described in this paper, represents a new reference resource in the domain of literary NE annotation, especially for the broad spectrum of languages[13] included. As such, its preservation and visibility should be ensured. When it comes to the publication and preservation of the whole richly annotated ELTeC collection once it is completed (a goal set to be achieved with the end of this Action in 2021), the final decision is yet to be made, with current interest in Textgrid[14] and GAMS[15]. Collaboration with CLARIN should ensure the visibility of the resource, by means of metadata harvesting to the CLARIN Virtual Language Observatory and by listing the corpus in initiatives such as the CLARIN Resource Families.

The usability of the ELTeC collection could be further enhanced by making it a part of the Federated Content Search initiative (especially with the further addition of the morpho-syntactic and semantic

[13]Not all languages in ELTeC have been made the object of NE annotation so far, but future additions are envisaged.

[14]Textgrid [https://textgrid.de/], a repository supported by the Göttingen State and University Library (SUB) and the Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen mbH (GWDG).

[15]`http://gams.uni-graz.at/`

layer). This should be facilitated by the fact that our COST action is planning to publish the corpus also on the TXM platform[16], which already implements CQP queries.

Last but not least, the collaboration could touch upon the question of NLP tools and services, their usability and domain adaptation. Given the fact that for most languages only parts of the ELTeC corpora can be manually annotated, what can CLARIN do to achieve quality automatic processing for the rest of the corpus? And more generally, how can this help in creating ad hoc models for similar texts which are not yet included in the ELTeC collection? Right now CLARIN offers easy to use services to process texts, such as the CLARIN Switchboard and Weblicht. However, NER is not yet available for most languages, and when it is, it generally does not support the processing of TEI-XML texts, something which constitutes a major issue in processing as shown in our results. Moreover, the NER models that are currently available are not necessarily adapted for literary studies and do not allow for the annotation of all of the NER categories demanded by ELTeC design.

The presentation will describe the current state of the ELTeC corpus, with a focus on the NE manually annotated subset and on the tests carried out with various NER modules, and will constitute an opportunity to discuss the aforementioned points with experts from CLARIN ERIC and the various national consortia.

References

Bamman, D., Popat, S., and Shen, S. 2019. 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)*, pages 2138–2144, Minneapolis, Minnesota, June. Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1220.

Bick, E. 2006. Functional Aspects in Portuguese NER. In Vieira, R., Quaresma, P., da Graça Volpes Nunes, M., Mamede, N. J., Oliveira, C., and Dias, M. C., editors, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006 (PROPOR'2006)*, pages 80–89. Springer Verlag. https://link.springer.com/chapter/10.1007/11751984_9.

Chinchor, N. A. 1998. Overview of MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. https://www.aclweb.org/anthology/M98-1001.

LDC. 2008. ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, Version 6.6. Technical report, Linguistic Data Consortium. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf.

Nadeau, D. and Sekine, S. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January. Publisher: John Benjamins Publishing Company, http://www.ingentaconnect.com/content/jbp/li/2007/00000030/00000001/art00002.

Santos, D., Seco, N., Cardoso, N., and Vilela, R. 2006. HAREM: An Advanced NER Evaluation Contest for Portuguese. In Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odjik, J., and Tapias, D., editors, *Proceedings of LREC 2006 (LREC'2006)*, pages 1986–1991. http://www.lrec-conf.org/proceedings/lrec2006/pdf/59_pdf.pdf.

Stankovic, R., Santos, D., Frontini, F., Erjavec, T., and Brando, C. 2019. Named Entity Recognition for Distant Reading in Several European Literatures. In *DH Budapest 2019*, Budapest. http://elte-dh.hu/wp-content/uploads/2019/09/DH_BP_2019-Abstract-Booklet.pdf.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April. Association for Computational Linguistics. https://www.aclweb.org/anthology/E12-2021.

Zupan, K., Ljubešić, N., and Erjavec, T. 2017. Annotation guidelines for Slovenian named entities: Janes-NER. Technical report, Jožef Stefan Institute, September. https://www.clarin.si/repository/xmlui/bitstream/handle/11356/1123/SlovenianNER-eng-v1.1.pdf.

[16] http://textometrie.ens-lyon.fr/