A Comprehensive Study of Open-Source Libraries for Named Entity Recognition on Handwritten Historical Documents

Claire Bizon Monroc[1,2](✉), Blanche Miret[1], Marie-Laurence Bonhomme[1], and Christopher Kermorvant[1,3]

[1] TEKLIA, Paris, France
claire.bizonm@gmail.com
[2] DYOGENE-INRIA, Paris, France
[3] Normandie Université, LITIS, Rouen, France

Abstract. In this paper, we propose an evaluation of several state-of-the-art open-source natural language processing (NLP) libraries for named entity recognition (NER) on handwritten historical documents: spaCy, Stanza and Flair. The comparison is carried out on three low-resource multilingual datasets of handwritten historical documents: HOME (a multilingual corpus of medieval charters), Balsac (a corpus of parish records from Quebec), and Esposalles (a corpus of marriage records in Catalan). We study the impact of the document recognition processes (text line detection and handwriting recognition) on the performance of the NER. We show that current off-the-shelf NER libraries yield state-of-the-art results, even on low-resource languages or multilingual documents using multilingual models. We show, in an end-to-end evaluation, that text line detection errors have a greater impact than handwriting recognition errors. Finally, we also report state-of-the-art results on the public Esposalles dataset.

Keywords: Named entity recognition · Text line detection · Handwritten historical documents

1 Introduction

As more and more historical documents are digitized, large collections of document images are constituted in patrimonial institutions. Automatic natural language processing (NLP) can be a valuable tool for researchers in the humanities who analyze and interpret large masses of unstructured documents [19]. For historical document images, in particular, the development of efficient document layout analysis (DLA) and handwritten text recognition (HTR) systems make their textual content available for downstream information extraction tasks with classical NLP tools. Named entity recognition (NER), in particular, has a potential to improve the user experience of humanities researchers by providing search and exploration facilities in large corpora. Moreover, recent years have

S. Uchida et al. (Eds.): DAS 2022, LNCS 13237, pp. 429–444, 2022.
https://doi.org/10.1007/978-3-031-06555-2_29

seen the rise of NLP as a major element of computational humanities, leading to the development of efficient neural models for NER tasks, to the point that NER on standard contemporary large-resource languages has been called a solved problem [19]. These developments triggered the birth of several libraries, making state-of-the-art neural architectures easily available for both industrial development and research. Yet, the application of these state-of-the-art tools to historical documents still gives rise to specific challenges.

NER systems performance are highly dependent on the target language and domain. As previously noted [5], applying standard tools on textual data from a different field without adaptation data can lead to a dramatic drop of performance. Yet, out-of-the-box information extraction models in libraries are generally trained and evaluated on contemporary language data extracted from Wikipedia or modern newspapers. Although notable efforts have been made to develop corpora and language models for historical German [18] for example, textual data is still mostly extracted from XIX$^{\text{th}}$ century documents.

Medieval documents, such as charters, are generally written in *dead languages*, such as Latin or old versions of European languages, for which there is no native speaker and no production of new documents. Very few of the massive annotated language corpora that drove the success of neural NLP tools in English and German are available for historical languages [18], and model performance on historical versions of their languages can suffer due to diachronic changes and lack of standardization in naming entities in old texts. Moreover, classical NLP tools applied to historical documents can only be the last step in a pipeline composed of at least a line recognition step followed by HTR. The quality of the HTR plays a large part in the decreased performances of NER models on automatically processed historical documents, and designing meaningful metrics to evaluate those results has proved challenging [14]. Another challenge is the existence of nested entities. Examples in the literature have mostly focused on flat entities, but medieval name identification often embed location names (e.g., *Fridericus, Dei gracia dux Austrie et Styrie*, a *name* entity containing the *location* entities *Austrie* and *Styrie*). Furthermore, nested entities raise challenges in terms of modeling and evaluation.

In this paper, we explore these different aspects through our experiments using three state-of-the art natural language processing libraries for named entity recognition (NER) on handwritten historical documents (spaCy [16], Stanza [22], and Flair [3]) using three different datasets of handwritten historical documents: a multilingual corpus of medieval charters (HOME), a corpus of Quebec parish records in French (Balsac), and a public corpus of marriage records in Catalan (Esposalles). The contributions of this paper are the following: **(1)** We provide a comprehensive study of open-source NER tools and metrics on low-resource, multilingual handwritten documents , highlight the different impact levels on this task considering the preceding steps in real use case scenarios (e.g., text line detection and HTR). and show that they can be considered as a sustainable option for this particular type of documents. **(3)** We propose a different use of nested entities compared to [9], with a better exploitation of the data. **(4)** We obtain state-of-the-art performance on the Esposalles dataset outperforming the best published results. These results in particular point out how current off-the-shelf NER libraries offer state-of-the art tools that are able to reach and even surpass tailor-made models.

Next, we continue with the presentation of related work in Sect. 2. Section 3 describes the three historical corpora and the challenges they raise. The open-source libraries are presented in Sect. 4. The hyperparameters and the different metrics used to evaluate our models are detailed in Sect. 5. Finally, we draw conclusions and highlight some future perspectives in Sect. 6.

2 Related Work

The amount of previous research on NER for handwritten historical document is relatively limited in comparison to the vast amount of published work on contemporary documents. Some authors proposed to extract named entities without explicit HTR [1], by extracting specific features from the images and predicting the named entity (NE) with a bidirectional long short-term memory (BLSTM) neural network. Their method was tested on relatively simple documents (Georges Washington letters, Queensland State Archives dataset, IAM dataset) but they claimed it can be used on any language in Latin script, in which an NE usually starts with a capital letter.

Several types of HTR and NER pipelines on the Esposalles dataset were explored [21], using different types of graphical and neural-based models for NER. This research showed that when the HTR error rate is as low as 5%, the NER performance is as good as on recognized text as on the groundtruth. On the same dataset, another study proposed to train a single convolutional-sequential model that jointly performs handwriting recognition and named entity extraction [10], with the objective to reduce the error propagation between the two steps.

Other researchers trained CRF-based NER models using linguistic features on a large corpus of medieval cartularies and charters in Medio Latin [2]. The authors obtained a score greater than 90% for persons and locations, but using manually transcribed text. A comparison of two different approaches to NER with a newly produced fine-grained multilingual dataset of medieval charters, was recently proposed [9]. One approach was a traditional sequential approach (HTR and NER) whereas the second one was an integrated approach in which the model transcribes and predicts the NEs simultaneously. The integrated approach yielded the best results, but the experiments were carried on manually segmented lines. Two-stage (HTR and afterward NER) and end-to-end architectures (NER directly at image level) were also compared [23]. Manually annotated the known IAM and George Washington datasets with NE labels and their experiments showed, contrary to previous research [9], that a two-stage model can achieve higher scores than an end-to-end one.

Another comparison of five different types of NER systems (statistical, rule-based and neural) was explored [25], on two collections of letters in English from the 17th and 18th centuries, but without retraining the models, and showed that ensemble methods can improve NER results.

Table 1. Number of pages, lines, words and entities in the different corpora.

	HOME				Esposalles	Balsac
No. of	Czech	German	Latin	All	Catalan	French
Page	202	173	126	501	125	896
Line	3,591	3,199	1,971	8,761	3,827	45,479
Word	66,257	77,086	35,759	179,102	39,527	205,165
Entity	4,117	4,419	3,315	11,851	16,782	25,564

After massive digitization campaign of historical newspapers in most western libraries, many studies have been conducted on the performance of NER on digitized texts. A previous study tackled NER on old Finnish newspapers [17], showing that with an OCR error rate around 25% to 30%, the NER performance is highly degraded for systems based on linguistic methods. Moreover, the performance of a NER was proved [14] that it can drop with at least 30% points, when the character error rate (CER) of the OCR process increases from 1% up to 7% by adding synthetic noise to the documents. To this scope, the CLEF HIPE 2020 evaluation [11,12] gathered 13 teams on shared tasks dedicated to the evaluation of named entity processing on historical newspapers in French, German and English. This evaluation showed that NER on historical newspapers are capable of dealing with historical inputs and reach performance comparable to those obtained on contemporary texts when enough training data is available but confirm an important degradation of the performance in presence of OCR noise.

3 Handwritten Historical Document Corpora

We conducted experiments on three different corpora of historical documents annotated with named entities: HOME, Balsac, and Esposalles. Detailed statistics on the corpora can be found in Table 1. Also, Table 2 presents for each corpus the following statistics per language: type of entity, the average number of tokens and characters in an entity, the total number of entities and the number of nested entities[1].

HOME is a corpus of annotated images of handwritten medieval charters, 499 charters from the Archives of the Bohemian Crown, as well as archives of several monasteries, chosen for their historic and linguistic significance to the history of Central Europe. A corrected and augmented version of the corpus has been made available by the National Archives of the Czech Republic and served as the basis for the work presented in this paper. A HOME Czech charter example is presented in Fig. 1 (1). Contrary to previous research [9], we take into account the fact that entities can be split across two lines, and therefore consider such groups as constituting one single entity.

[1] Note that Table 2 presents a count for *person* entities, which do not appear as such in the Esposalles corpus. Therefore, the number of *person* entities in this table is actually the sum of *name* and *surname* entities in the dataset.

Table 2. For each type of entity, average entity lengths in number of tokens (words), characters (chars), and entities counts for each corpus and language. **PER** = Person, **LOC** = Location, **DAT** = Date, **OCC** = Occupation, **ST** = Civil state.

			HOME				Esposalles	Balsac
			Czech	German	Latin	All	Catalan	French
PER	Length	Token	3.0	5.1	2.9	3.6	1.61	2.33
		Char	20.4	32.9	21.3	24.2	9.62	14.27
	Count	Entity	1,997	1,356	1,374	4,727	9,167	15,810
		Nested	3	1	5	9	–	–
LOC	Length	Token	1.2	1.0	1.1	1.1	1.89	2.08
		Char	9.4	7.2	8.3	8.1	9.35	12.33
	Count	Entity	1,956	2,909	1,826	6,691	2,959	2,823
		Nested	1,054	1,035	840	2,929	–	–
DAT	Length	Token	8.5	10.7	7.4	9.0	–	5.21
		Char	58.0	66.5	56.0	60.4	–	23.28
	Count	Entity	164	154	115	433	–	4,551
		Nested	1	0	0	1	–	–
OCC	Length	Token	–	–	–	–	1.25	1.31
		Char	–	–	–	–	7.15	10.44
	Count	Entity	–	–	–	–	3,207	2,380
		Nested	–	–	–	–	–	–
ST	Length	Token	–	–	–	–	1.01	–
		Char	–	–	–	–	6.82	–
	Count	Entity	–	–	–	–	1,449	–
		Nested	–	–	–	–	–	–

Balsac corpus contains 896 images of parish records from Quebec regions, from 1850 to 1916. The records contain baptism, marriage and death records, and are an important data source for social, historical and genealogical research. A Balsac parish record example is presented in Fig. 1 (2). The pages are organized in two columns, with a small margin paragraph for each act, specifying a reference number and the name of the act's main subject, and a main column with the proper content of the act. Both margins and main paragraphs were annotated and used for training.

Esposalles dataset [13] is a subset of 125 pages from the Esposalles database, which contains marriage license records from the Archives of the Cathedral of Barcelona, written in Catalan by a single writer from the XVII$^{\text{th}}$ century. An Esposalles record example is presented in Fig. 1 (3). This dataset was used in the 2017 ICDAR Information Extraction in Historical Handwritten Records campaign, and it is still possible to submit results for its tasks online[2]. For the purpose of comparing results on this dataset to those we got on HOME and Balsac,

[2] https://rrc.cvc.uab.es/?ch=10&com=introduction.

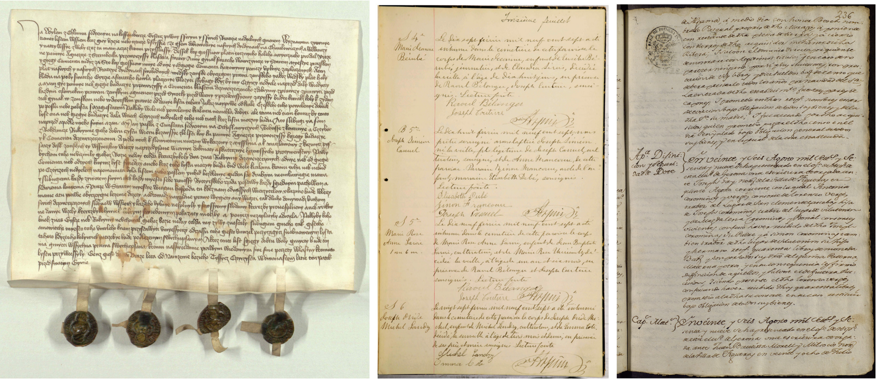

Fig. 1. Page examples from the three datasets: (1) HOME Czech charter (2) Balsac parish record (3) Esposalles record.

we performed the basic track NER task of the competition (where the entities to identify are name, surname, occupation, location and state).

3.1 Nested Entities in HOME Corpus

Previous research [9] ignored nested entities and continuations (situations where an entity overlaps two lines in a charter). Their solution was useful to reduce the complexity of the NER task and create a benchmark to compare sequential and combined approaches. Nevertheless, disregarding nested entities is suboptimal. First, nested entities actually are a large part of the annotated entities, up to half of them for entities of type *location*. Ignoring them, therefore, could mean disregarding an important volume of valuable information. Secondly, removing nested annotations increases ambiguity in the labels, as nested entities are then annotated with the labels of their parent entities. For example, the following entity extracted from a Latin charter refers to a person:

Nos Fridericus, Dei gracia dux Austrie et Styrie [...] *profitemur et recognoscimus*
PER: Fridericus, Dei gracia dux Austrie et Styrie

We refer to this PER entity as *parent*. When ignoring nested entities, the tokens *Austrie* and *Styrie* are labelled as *person*, but when met outside a nested entity, the same tokens are annotated as *location*. We refer to these entities as *nested*.

regni Hungarie et ducatus Austrie
LOC: Hungarie - Austrie

To avoid these ambiguities, we decided to incorporate the present nested entities. There are several ways to perform this: (1) giving such entities two labels (their own and the label of their parent). This would create a multilabel classification problem, which is not supported by most of the out-of-the-box NER tools that our work aims to compare. (2) training two models, one with the labels of the parent entities and the other with the labels of the nested entities, to predict the overlapping entities separately. A lighter way to solve the issue is to "*flatten*" nested entities, that is to systematically use nested labels when they exist, and reduce parent entities to their core to avoid double labelling. The following example illustrates this method.

> *Nos Fridericus, Dei gracia dux Austrie et Styrie* [...] *profitemur et recognoscimus*
> **PER**: Fridericus, Dei gracia dux Austrie et Styrie

becomes

> *Nos Fridericus, Dei gracia dux Austrie et Styrie* [...] *profitemur et recognoscimus*
> **PER**: Fridericus, Dei gracia dux
> **LOC**: Austrie - Styrie

This solution is also relevant to one of the main purposes of applying entity extraction to historical documents: to link those entities to indexes of people or locations, or use the recognized entities to fill databases, which also requires entity linking (EL), so that if one record concerns a person already in the database, for example, it gets correctly linked to that individual. The loss of specific information concerning the relation between parent and nested entities can be compensated by providing neighborhood context as an input to an EL model. The identification of the core part of an entity was done following a simple rule: after extracting nested entities, all parts of the entity with more than a defined threshold of characters are kept. To avoid the potential creation of meaningless entities from various linking and stopwords, we filtered out for each language a list of non-semantically significant words when they were located at the edge of entities, i.e., "et cetera" in Latin, "und" in German or "z", "od" in Czech.

4 Named Entity Recognition Libraries

We compared the performance of three open-source and widely used NLP Python libraries on the NER task: Stanza, Flair and spaCy.

Stanza (v1.1.1)[3] [22] was created and maintained by the Stanford NLP group. It provides different NLP tools that can be used in a pipeline (e.g., tokenization, NER) and pre-trained neural models supporting 66 different languages, including German, Czech, Latin, Catalan and French.

Flair (v0.7)[4] [3] provides, besides state-of-the-art NLP models, contextual word and character embeddings, a feature that is now also supported by Stanza.

[3] https://stanfordnlp.github.io/stanza.
[4] https://github.com/flairNLP/flair.

Moreover, Flair allows stacking embeddings, a technique that allowed the authors to exceed state-of-the-art NER results on German, architecture that we consider in this study.

spaCy (v2.3.5)[5] [16] is explicitly designed to ease deploying models into production. The library comes with out-of-the-box support of multiple languages, including Czech, German, Catalan and French. We built on pre-trained models to train a NER pipeline whenever they were made available by the library. When they were not, we trained the NER on the available basic language defaults.

5 Experiments

In this section, we describe our experimental setup: the data, the metrics and the different model configurations and training procedures. First, for the HOME and Balsac datasets, the documents were randomly split in three training, validation and test sets with an 80/10/10 ratio. NER models were then evaluated on the test corpus of their language or on all languages for multilingual models. The files in IOB (short for inside, outside, beginning) required by the NER tools and that were used for training, validation and test, were created from the XML documents provided by Transkribus[6], which was the tool used for transcription and annotation. For Balsac, there was originally no *person* annotation, but two distinct *name* and *surname* annotations. As the end-goal was to identify individuals, they were reunited in a unique tag during pre-processing. Regarding Esposalles, for training and testing, files in the IOB format were created from the text files containing the text line transcriptions, and the text files containing the entity categories for each word of the text line, available after downloading the Esposalles dataset[7]. We used the train/validation/test sets defined for the ICDAR 2017 competition [13], both for training the HTR and the NER models.

5.1 Evaluation Metrics

Basic Page-Level Metric. This first metric compares the entities found in the prediction with the ones from the groundtruth, at page level, disregarding the positions and number of occurrences of the entities, and ignoring punctuation. If an entity from the groundtruth is predicted with the true label and text, then it is considered as a true positive, regardless of where in the page and how many times it appeared. This metric returns the precision, recall and F1 scores.

Nerval for Automatic Transcription. The Nerval[8] metric is used for the evaluation of the NER models on the automatic transcriptions. These are first aligned with the groundtruth at character level, by minimizing the Levenshtein distance

[5] https://spacy.io.

[6] https://readcoop.eu/transkribus/.

[7] https://rrc.cvc.uab.es/?ch=10&com=tasks.

[8] https://gitlab.com/teklia/nerval.

between them. Each entity in the groundtruth is then matched with a corresponding entity in the aligned transcription, with the same entity label, or an empty character string if no match is found. If the distance between the two entities is less than threshold (by default 30%) of the groundtruth entity length, the predicted entity is considered as recognized. For the purpose of matching detected entities to existing datasets, we estimated that a 70% match between the entity texts was an acceptable threshold.

Esposalles Metric. This metric is described in [13]. For each marriage record, the text and the label of the entity are compared to the entities from the groundtruth transcriptions and annotations. The two transcriptions are aligned and the score for each entity can be: 0 if the entity label is wrong, 1 if both the transcription and the label are exact, and $1 - CER$, between the automatic transcription and the groundtruth for that entity if the label is correct, but the transcription is different. This metric does not return precision, recall and F1 scores but a global accuracy score for each entity category and all entities.

Text line detection metric. The metric used to evaluate the line detection is the average precision at a 0.75 intersection-over-union (IoU) threshold (AP@0.75), which means that a detected line is matched to an annotated text line if they intersect by at least 75% of their surfaces [6]. The AP@0.75 is the proportion of correctly predicted text lines based on this IoU threshold.

Text Recognition Metric. Character error rate (CER) and word error rate (WER) are used to evaluate the text recognition models. CER corresponds to the Levenshtein distance between two strings, that is, the minimum number of character editions to be made to transform one string to the other. WER uses the same principle at word level.

5.2 Hyperparameters and Model Training

After an optimization using grid search, the chosen hyperparameters were (hidden layer dimension, learning rate, dropout): Flair (256, 0.1, -), spaCy (64, 0.001, -), and for Stanza (256, 0.01, 0.3).

We trained two Stanza models for each corpus: *Stanza Basic*, the default Stanza configuration with word2vec embeddings [20], pre-trained for CoNLL 2018, default configuration[9]. We also explored *Stanza Fasttext*, the default Stanza configuration, but with FastText embeddings [7].

For Flair, the models were trained with a combination of word2vec word embeddings and contextualized character embeddings [3]. For Czech, French, and German, we used pre-trained Flair embeddings (contextual character embeddings provided by the library). For Latin, we used Fasttext pre-trained embeddings and we also trained our own Flair embeddings on a large corpus of

[9] https://github.com/stanfordnlp/stanza-train.

Table 3. Micro-averaged page-level percent of precision, recall and F1 scores by language and (monolingual) NER model, on manual transcriptions, with the *Basic page-level metric*.

	HOME			Esposalles	Balsac	
	Czech	German	Latin	Catalan	French	All
ICFHR 2020 [9] (different strategy for nested entities and different metric)						
Precision	70.8	85.2	59.7	–	–	–
Recall	50.7	67.9	44.4	–	–	–
F1	59.1	75.6	50.9	–	–	–
Stanza Basic						
Precision	83.5	92.8	87.3	97.5	77.3	87.7
Recall	83.5	92.6	85.5	98.2	86.3	89.2
F1	83.5	92.7	**86.4**	97.8	81.6	88.4
Stanza Fasttext						
Precision	84.3	93.8	84.3	96.8	78.7	87.6
Recall	86.0	92.5	85.2	97.8	86.6	89.6
F1	**85.1**	**93.1**	84.7	97.3	82.5	**88.5**
Flair						
Precision	80.8	90.0	82.6	96.1	94.1	88.7
Recall	77.3	87.0	76.5	96.5	93.5	82.3
F1	79.0	88.5	79.4	96.3	**93.8**	85.1
spaCy v2						
Precision	81.1	90.2	n/a	97.5	84.3	88.3
Recall	76.3	89.8	n/a	98.3	84.3	87.2
F1	78.6	90.0	n/a	**97.9**	84.3	87.7

10 millions tokens available in the public domain and collected by The Latin Library[10]. There were no Flair embeddings for Catalan either, thus we chose French embeddings for the Esposalles dataset. The multilingual Flair model for HOME was trained by feeding all the charter transcriptions in the train set to Flair pre-trained (on 300+ languages) multilingual character embeddings. Regarding spaCy, for German and French, for which three different models are available, we used the medium-sized model `de_core_news_md`, and `fr_core_news_sm`, respectively. For Czech and Catalan, there are no pre-trained models. Thus, we trained our NER models with the default linguistic features support. Latin is, also, not supported by this library, thus, we used the multilingual spaCy model. Finally, the text line detection were performed with Doc-UFCN [6] and the automatic text recognition by Kaldi [4].

[10] https://github.com/cltk/lat_text_latin_library.

Table 4. Performance comparison of multilingual vs monolingual NER on the HOME corpus (manual transcriptions) with the *basic page-level metric*.

	HOME					
	Czech	German	Latin	Czech	German	Latin
	Flair			spaCy v2		
Precision	80.8	90.0	82.6	81.1	90.2	n/a
Recall	77.3	87.0	76.5	76.3	89.8	n/a
F1	*79.0*	*88.5*	*79.4*	78.6	90.0	n/a
	Flair multilingual			spaCy v2 multilingual		
Precision	75.1	88.8	84.1	81.6	91.0	84.5
Recall	73.7	84.8	79.5	80.6	90.6	82.1
F1	74.4	86.8	81.8	**81.0**	**90.8**	**83.3**

Table 5. Performance comparison when applying NER to manually annotated text (Ann.) and automatically recognized text (Rec.) with nerval metric. "–": no multilingual models for Balsac and Esposalles, "n/a": for spaCy in Latin, since there is no Latin base model.

	HOME						Esposalles		Balsac	
	Czech		German		Latin		Catalan		French	
	Ann	Rec	Ann	Rec	Ann	Rec	Ann	Rec	Ann	Rec
Stanza Basic										
Precision	79.1	60.8	93.8	84.0	86.9	77.8	97.6	96.5	77.9	69.7
Recall	86.8	40.0	93.3	67.5	88.9	67.5	98.0	96.6	85.5	78.3
F1	82.8	48.3	93.5	74.8	**87.9**	**72.3**	97.8	96.5	81.5	73.7
Stanza Fasttext										
Precision	79.1	63.1	95.2	85.1	84.7	75.6	97.1	96.3	79.2	71.9
Recall	88.4	42.4	93.7	67.5	87.9	67.0	97.5	96.2	86.3	78.8
F1	**83.5**	**50.7**	**94.4**	75.3	86.3	71.0	97.3	96.2	82.6	75.2
Flair										
Precision	77.7	57.2	92.1	86.5	84.3	76.6	97.9	96.8	93.7	86.0
Recall	81.8	34.5	88.1	65.6	82.1	67.3	98.2	97.1	93.1	85.6
F1	79.7	43.0	90.1	**75.7**	83.2	71.6	**98.0**	**96.9**	**93.4**	**85.8**
Flair multilingual										
Precision	73.9	53.4	89.9	79.7	83.0	73.0	–	–	–	–
Recall	77.6	33.4	85.9	62.8	81.0	64.1	–	–	–	–
F1	75.7	41.1	87.9	70.3	82.0	68.3	–	–	–	–
spaCy										
Precision	76.4	61.5	91.9	83.8	n/a	n/a	97.1	95.8	83.1	74.5
Recall	80.3	37.9	91.0	67.5	n/a	n/a	97.8	96.3	83.6	76.7
F1	78.3	46.9	91.5	74.7	n/a	n/a	97.5	96.1	83.4	75.6
spaCy multilingual										
Precision	77.5	58.2	91.5	83.5	85.6	75.3	–	–	–	–
Recall	83.4	38.4	90.6	68.3	86.3	66.0	–	–	–	–
F1	80.4	46.3	91.1	75.1	85.9	70.3	–	–	–	–

Table 6. Evaluation of the text line detection using the *average precision @0.75* for Doc-UFCN model and evaluation of the HTR process using character error rate (CER) and word error rate (WER) on the test sets of the different corpora.

	HOME				Balsac	Esposalles
AP@0.75	48.57				91.13	–
	Kaldi HOME (multilingual)				Kaldi Balsac	Kaldi Esposalles
	Czech	German	Latin	all	French	Catalan
CER	8.70	7.48	10.37	8.93	6.41	1.32
WER	29.71	26.40	35.59	29.26	17.41	3.51

6 Results

Results on the Manual Transcriptions. Table 3 presents the results of our experiments using the different NER models on the manually transcribed HOME, Balsac, and Esposalles datasets. Table 3 does not allow us to pick a clear winner out of the three NER libraries. Averaging the scores over all the datasets, *Stanza FastText* performs better on two of the HOME languages. However, spaCy and Flair obtain the best results on the Esposalles and Balsac datasets. Therefore, it is unclear which library will perform best on a given dataset, thus, practically, testing the three libraries is our recommendation. Table 4 offers a comparison of the results obtained by a multilingual NER model and a monolingual one, on the manually transcribed HOME dataset. This table shows that Flair performs better when it is specialized for one language, but spaCy benefits from the larger amount of training data of the multilingual model, outperforming both a monolingual one and Flair, on all the HOME languages. These results show that using multilingual NER models is a sustainable option for low-resource languages or when the documents contain multiple languages. Both tables reported the results with the *Basic page-level metric*.

Results on the Manual and Automatic Transcriptions. Table 5 presents how NER results are impacted by text line detection and handwriting text recognition (HTR), both on manual and HTR transcriptions, with the *Metric for aligned automatic transcriptions*. We observe that there is only a marginal decrease in accuracy on the Esposalles dataset between the NER on the manual transcription and the one on the HTR text, and that can be explained by how there is no text line detection step for Esposalles as the line images are given, and the HTR model for Esposalles has a 3.51% WER as shown in Table 6, which is very low. We can draw a similar (and expected) conclusion looking at the results on Balsac, where the accuracy always drops by less than ten percentage points between manual and transcribed text, and where the text line detection works well and the HTR model has a WER of 17.41%, which is moderate. On the other end of the spectrum, we see that the text line detection does not perform well on the HOME dataset (AP@0.75 of 48.57), and the HTR model has

Table 7. Results from various NER libraries on the Esposalles dataset, manual and automatic (Kaldi) transcription, using the *Esposalles metric*.

	PER (name)	PER (surname)	LOC	OCC	ST	all
Naver Labs [21]						
auto						0.95
CITlab-ARGUS-2 [13]						
auto						0.92
spaCy v2						
manual	0.98	0.97	0.98	0.98	0.98	0.98
auto	0.97	0.95	0.96	0.97	0.98	0.96
spaCy v3						
manual	0.97	0.96	0.95	0.97	0.98	0.97
auto	0.96	0.94	0.93	0.97	0.98	0.95
Flair						
manual	0.99	0.99	0.98	0.98	0.98	0.98
auto	0.98	0.95	0.96	0.98	0.98	**0.97**
Stanza Basic						
manual	0.96	0.98	0.94	0.97	0.97	0.97
auto	0.96	0.95	0.92	0.97	0.97	0.96
Stanza + FastText						
manual	0.96	0.97	0.94	0.97	0.97	0.96
auto	0.96	0.95	0.92	0.97	0.97	0.95

a WER between 25% and 35%, depending on the language. It should be noted that the test sets for HOME are rather small: 17 pages for the Czech language, 14 pages for German, and 11 pages for Latin. Therefore, one page on which the lines are badly detected can heavily impact the final results. Even on the manual text, the Czech language is the one on which the NER models perform the most poorly. However, the dramatic drop in recall on the recognized text may be in part explained by the quality of the line detection. The average difference in number of lines between the groundtruth and the detected lines in Czech are three lines, with one page with only 8.9% of lines detected, one page with only 75% of lines detected. This is worse for German, for which the average difference in line count is of 1.3 lines. For Latin, the number of detected lines is consistent with the groundtruth. Thus, the recall drops less between manual and automatic transcription in Latin than it does for German, although the WER of the HTR engine is almost ten points high for Latin. It seems that the quality of the text line detection heavily impacts the recall, while the performance of the HTR models have a more effected precision. Regarding the performance of the multilingual models, they offer a viable solution on the HOME dataset, with performance scores close to the monolingual models. It is worth noting that the results presented for multilingual Flair or multilingual spaCy are based on a multilingual HTR, which opens perspectives for a fully multilingual processing workflow.

Results on the Esposalles Dataset. Table 7 compares the performance scores on the manual and HTR transcriptions of the Esposalles dataset, with the previous best results on this task, using the *Esposalles metric*. Using the dedicated Esposalles scorer, Flair outperforms the best published result on this dataset [21]. All the tested models yield similar and competitive results.

7 Conclusions and Future Work

In this study, we presented a comparison of three off-the-shelf open-source NER libraries on three historical datasets which include five different languages: Czech, German, Latin, Catalan and French. The NER models were evaluated on both manual and automatic transcriptions of the handwritten documents. We showed that, in the case of low or moderate error rates, while the drop in performance of the NER on the recognized text is limited, the quality of text line detection has a large impact on the results. We also showed that multilingual models, for both HTR and NER, demonstrated competitive performance and represent a viable option in case of scarce training data. None of the three compared library outperforms the other, thus, we recommend testing the three of them (with default hyperparameters). Finally, using Flair NER, we established state-of-the-art results on the Esposalles dataset. As future work, we envision a study of other off-the-shelf models for NER, Transformer-based [24], recently provided by spaCy and Stanza. We are also considering in exploring some multitask methods, especially curated for nested entities in digitized documents [8,15].

Acknowledgement. This work is part of the *HOME History of Medieval Europe* research project, supported by the French National Research Agency (Grant agreements No. ANR-17-JPCH-0006) through the European JPI Cultural Heritage and Global Change. The authors wish to thank Emanuela Boroş for the fruitful discussions and her careful reading of the manuscript.

References

1. Adak, C., Chaudhuri, B.B., Blumenstein, M.: Named entity recognition from unstructured handwritten document images. In: Workshop on Document Analysis Systems, pp. 375–380 (2016)
2. Aguilar, S.T., Tannier, X., Chastang, P.: Named entity recognition applied on a data base of medieval latin charters. The case of chartae burgundiae. In: International Workshop on Computational History (2016)
3. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: International Conference on Computational Linguistics (2018)
4. Arora, A., et al.: Using ASR methods for OCR. In: International Conference on Document Analysis and Recognition, pp. 663–668 (2019)

5. Bamman, D.: Natural language processing for the long tail. In: Digital Humanities (2017)
6. Boillet, M., Kermorvant, C., Paquet, T.: Robust text line detection in historical documents: learning and evaluation methods. IJDAR (2022). https://doi.org/10.1007/s10032-022-00395-7
7. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
8. Boroş, E., et al.: Alleviating digitization errors in named entity recognition for historical documents. In: Conference on Computational Natural Language Learning, pp. 431–441 (2020)
9. Boros, E., et al.: A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters. In: International Conference on Frontiers in Handwriting Recognition, pp. 79–84 (2020)
10. Carbonell, M., Villegas, M., Fornés, A., Lladós, J.: Joint recognition of handwritten text and named entities with a neural end-to-end model. In: International Workshop on Document Analysis Systems (2018)
11. Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Extended overview of clef hipe 2020: named entity processing on historical newspapers. In: CEUR Workshop Proceedings. No. 2696, CEUR-WS (2020)
12. Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Overview of CLEF HIPE 2020: named entity recognition and linking on historical newspapers. In: Arampatzis, A., et al. (eds.) CLEF 2020. LNCS, vol. 12260, pp. 288–310. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_21
13. Fornés, A., et al.: Icdar 2017 competition on information extraction in historical handwritten records. In: International Conference on Document Analysis and Recognition (2017)
14. Hamdi, A., Jean-Caurant, A., Sidere, N., Coustaty, M., Doucet, A.: An analysis of the performance of named entity recognition over ocred documents. In: Joint Conference on Digital Libraries (2019)
15. Hamdi, A., Carel, E., Joseph, A., Coustaty, M., Doucet, A.: Information extraction from invoices. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) ICDAR 2021. LNCS, vol. 12822, pp. 699–714. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86331-9_45
16. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020)
17. Kettunen, K., Ruokolainen, T.: Names, right or wrong: named entities in an ocred historical finnish newspaper collection. In: International Conference on Digital Access to Textual Cultural Heritage (2017)
18. Labusch, K., Zu, S., Kulturbesitz, B., Neudecker, C., Zellhöfer, D.: Bert for named entity recognition in contemporary and historical german. In: Conference on Natural Language Processing (2019)
19. McGillivray, B., Poibeau, T., Ruiz, P.: Digital humanities and natural language processing: "Je t'aime... Moi non plus". Digit. Humanit. Q. **14**(2) (2020)
20. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: International Conference on Learning Representations (2013)
21. Prasad, A., Déjean, H., Meunier, J., Weidemann, M., Michael, J., Leifert, G.: Bench-marking information extraction in semi-structured historical handwritten records. CoRR (2018)

22. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A Python natural language processing toolkit for many human languages. In: Annual Meeting of the Association for Computational Linguistics: System Demonstrations (2020)
23. Tüselmann, O., Wolf, F., Fink, G.A.: Are end-to-end systems really necessary for NER on handwritten document images? In: Lladós, J., Lopresti, D., Uchida, S. (eds.) ICDAR 2021. LNCS, vol. 12822, pp. 808–822. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86331-9_52
24. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
25. Won, M., Murrieta-Flores, P., Martins, B.: Frontiers in Digital Humanities **5** (2018)