Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers

Maud Ehrmann[1] [0000-0001-9900-2193], Matteo Romanello[1] [0000-0002-1890-2577], Alex Flückiger[2], and Simon Clematide[2] [0000-0003-1365-0662]

[1] Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{maud.ehrmann,matteo.romanello}@epfl.ch
[2] University of Zurich, Zurich, Switzerland
{alex.flueckiger,simon.clematide}@uzh.ch

Abstract. This paper presents an extended overview of the first edition of HIPE (Identifying Historical People, Places and other Entities), a pioneering shared task dedicated to the evaluation of named entity processing on historical newspapers in French, German and English. Since its introduction some twenty years ago, named entity (NE) processing has become an essential component of virtually any text mining application and has undergone major changes. Recently, two main trends characterise its developments: the adoption of deep learning architectures and the consideration of textual material originating from historical and cultural heritage collections. While the former opens up new opportunities, the latter introduces new challenges with heterogeneous, historical and noisy inputs. In this context, the objective of HIPE, run as part of the CLEF 2020 conference, is threefold: strengthening the robustness of existing approaches on non-standard inputs, enabling performance comparison of NE processing on historical texts, and, in the long run, fostering efficient semantic indexing of historical documents. Tasks, corpora, and results of 13 participating teams are presented. Compared to the condensed overview [31], this paper includes further details about data generation and statistics, additional information on participating systems, and the presentation of complementary results.

Keywords: Named entity recognition and classification · Entity linking · Historical texts · Information extraction · Digitized newspapers · Digital humanities

[⋆] The present shared task is organized as part of **"*impresso* - Media Monitoring of the Past"**, a project which tackles information extraction and exploration of large-scale historical newspapers (https://impresso-project.ch/).

1 Introduction

Recognition and identification of real-world entities is at the core of virtually any text mining application. As a matter of fact, referential units such as names of persons, locations and organizations underlie the semantics of texts and guide their interpretation. Around since the seminal Message Understanding Conference (MUC) evaluation cycle in the 1990s [44], named entity-related tasks have undergone major evolutions until now, from entity recognition and classification to entity disambiguation and linking [71, 87].

Context. Recently, two main trends characterise developments in NE processing. First, at the technical level, the adoption of deep learning architectures and the usage of embedded language representations greatly reshapes the field and opens up new research directions [6, 63, 62]. Second, with respect to application domain and language spectrum, NE processing has been called upon to contribute to the field of Digital Humanities (DH), where massive digitization of historical documents is producing huge amounts of texts [105]. Thanks to large-scale digitization projects driven by cultural institutions, millions of images are being acquired and, when it comes to text, their content is transcribed, either manually via dedicated interfaces, or automatically via Optical Character Recognition (OCR). Beyond this great achievement in terms of document preservation and accessibility, the next crucial step is to adapt and develop appropriate language technologies to search and retrieve the contents of this 'Big Data from the Past' [53]. In this regard, information extraction techniques, and particularly NE recognition and linking, can certainly be regarded among the first and most crucial processing steps.

Motivation. Admittedly, NE processing tools are increasingly being used in the context of historical documents. Research activities in this domain target texts of different nature (e.g., museum records, state-related documents, genealogical data, historical newspapers) and different tasks (NE recognition and classification, entity linking, or both). Experiments involve different time periods, focus on different domains, and use different typologies. This great diversity demonstrates how many and varied the needs—and the challenges—are, but also makes performance comparison difficult, if not impossible.

Furthermore, it appears that historical texts pose new challenges to the application of NE processing [25, 83], as they do for language technologies in general [101]. First, inputs can be extremely noisy, with errors which do not resemble tweet misspellings or speech transcription hesitations, for which adapted approaches have already been devised [65, 17, 100]. Second, the language under study is mostly of earlier stage(s), which renders usual external and internal evidences less effective (e.g., the usage of different naming conventions and presence of historical spelling variations) [12, 11]. Further, beside historical VIPs, texts from the past contain rare entities which have undergone significant changes (esp. locations) or do no longer exist, and for which adequate linguistic resources and knowledge bases are missing [48]. Finally, archives and texts from the past are not as anglophone as in today's information society, making multilingual resources and processing capacities even more essential [26, 72].

Overall, and as demonstrated by Vilain et al. [109], the transfer of NE tools from one domain to another is not straightforward, and the performance of NE tools initially developed for homogeneous texts of the immediate past are affected when applied on historical materials [104]. This echoes the proposition of Plank [85], according to whom what is considered as standard data (i.e. contemporary news genre) is more a historical coincidence than a reality: in NLP non-canonical, heterogeneous, biased and noisy data is rather the norm than the exception.

Objectives. In this context of new needs and materials emerging from the humanities, the HIPE shared task[3] puts forward for the first time the systematic evaluation of NE recognition and linking on diachronic historical newspaper material in French, German and English. In addition to the release of a multilingual, historical NE-annotated corpus, the objective of this shared task is threefold:

1. strengthening the robustness of existing approaches on non-standard inputs;
2. enabling performance comparison of NE processing on historical texts;
3. fostering efficient semantic indexing of historical documents in order to support scholarship on digitized cultural heritage collections.

The remainder of this paper is organized as follows. Section 2 briefly presents previous work on NE processing, particularly for cultural heritage domains. Sections 3 and 4 present the tasks and the material used for the evaluation. Section 5 details the evaluation metrics and the organisation of system submissions. Section 6 introduces the 13 participating systems while Section 7 presents and discusses their results. Finally, Section 8 summarizes the benefits of the task and concludes.

2 Related Work

This section briefly summarizes the evolution of NE processing and the adaptation of main approaches to named entity recognition and classification (NERC) and entity linking (EL) to the cultural heritage domain.

NE processing overview. Since the seminal *Message Understanding Conference* series where NE recognition and classification was defined for the first time [45], numerous research work and evaluation campaigns subsequently developed. They reflect the complexification and diversification of NE-related tasks, as well as the evolution of information extraction from a document-oriented to a more entity-centric perspective. First, NER setting itself evolved, with the extension of typologies [97, 38], the enlargement of the scope of linguistic units to take into account (i.e. not only proper names) [23, 1, 2], and the consideration of

[3] `https://impresso.github.io/CLEF-HIPE-2020/`

languages other than English, with e.g. CoNLL, ESTER, HAREM, Evalita and Germeval [106, 39, 94, 66, 9]. Next, tasks diversified, with the introduction of relation extraction, metonymy resolution [68] and entity coreference [7], later on framed as entity linking with the emergence of large-scale knowledge bases [87]. Finally, besides the general domain of well-written newswire data, named entity processing was also applied to specific domains, particularly bio-medical [55, 42], and on more noisy inputs such as speech transcriptions [37], tweets in various languages [89, 84, 8], and historical texts. Regarding the latter, research work multiplied significantly during the last decade, and besides a modest evaluation campaign on French historical texts [36], no wide-ranging and systematic evaluation was organized. To the best of our knowledge, the CLEF HIPE 2020 shared task is the first to address NE processing for multilingual, diachronic and historical material.

NERC. Approaches to NERC over historical documents have grown with the evolution of techniques, from symbolic systems to traditional machine learning and, more recently, deep neural network-based approaches. Early approaches include the crafting of rule-based systems based of finite-state grammars and gazetteers, applied on e.g. American and Finish newspapers [52, 54], Swedish literary classics [12], or British parliamentary proceedings [46]. Main reported difficulties relate to OCR noise, often tackled via normalization rules based on string similarity. Then, following the (relative) greater availability of raw and annotated historical texts, research moved away in favor of machine learning approaches. Experiments first consisted in applying existing models (the most widely-spread being the Conditional Random Field-based Stanford NER system [34]) in order to assess their potential and compare their performances on OCRized vs. corrected texts [90], or on a diachronic basis [25]. Thereafter, work focused on training new (CRF) models on custom material, such as Australian or European newspapers [56, 73], or medieval charters [3]. Again, most work report difficulties with bad OCR, and strategies to cope with it include pre-processing (i.e. better sentence segmentation or word tokenization, OCR post-correction) or string normalisation. Finally, in line with the development of deep neural network approaches in NLP [19], new performances were attained for NERC on well-known contemporary data sets, first with CNN-BiLSTM [18], then with Bi-LSTM-CRF networks [63][4]. The later was widely adopted in the processing of a variety historical texts in e.g. English [102], French [27], German [88, 4] and Czech [50]. Neural approaches evolved further with the introduction of models able to better learn context, namely contextualized string embeddings [6] and Bidirectional Encoder Representations from Transformers (BERT) [22]. Recently applied on historical texts, such models—when trained on in-domain material—proved their capacity to better deal with OCR and to improve performances [95, 62, 27].

[4] CNN: Convolutional neural networks; Bi-LSTM: Bi-directional Long Short Term Memory.

Entity Linking. Appeared most recently, the task of linking entity mentions to their corresponding referents in a KB has received much attention since the pioneering experiments of [15] and [20] with English Wikipedia. Given its many applications in e.g. information retrieval, content analysis, data integration and knowledge base population, numerous works on EL were published during the last decade, and we refer the reader to [87] and [99] for an overview and analysis of main approaches to EL up to the neural wave, and to [98] for an overview of neural approaches. Main challenges in EL are name variation (several surface forms for one entity), name ambiguity (one surface form for several entities), and absence of the entity in the KB (NIL). Up to the apparition of neural approaches, EL methods traditionally belonged to two families: text similarity-based approaches (computing the similarity between the context of the mention to link and the entity candidate description in the KB), and graph-based approaches (computing the closeness between the mention and the candidate in a graph representing information on these objects). In both cases, main EL steps include: mention detection, candidate selection, and candidate ranking. In the digital humanities context, first experiments made use of existing, 'off-the-shelf' EL systems, such as [48] working on Dutch museum documents, or [93] on Italian WWII partisans' memoirs. [35] built their own graph-based system (REDEN) for the disambiguation of authors' mentions in a corpus of French literary criticism, and demonstrated the gain of complementing a generic KB (here DBpedia) with a domain-specific one (authority files from the French national library). Recent neural approaches in 'mainstream' (i.e. non-historical) EL outperformed state-of-the-art results, highlighting the role and importance of contextual word and entity embeddings, together with neural similarity function [58]. To date, and to the best of our knowledge, neural-based approaches were not yet applied on historical texts and, if much remains to be done, the HIPE shared task started paving the way in this direction.

Overall, experiments with NERC and EL on historical material were so far carried out on different types of documents, following diverse guidelines, and evaluated in isolation. The HIPE shared task allows, for the first time on such material, to compare performances and approaches in a systematic way.

3 Task Description

The HIPE shared task includes two NE processing tasks with sub-tasks of increasing level of difficulty.

Task 1: Named Entity Recognition and Classification (NERC)

- **Subtask 1.1 - NERC coarse-grained** (NERC-Coarse): this task includes the recognition and classification of entity mentions according to high-level entity types.
- **Subtask 1.2 - NERC fine-grained** (NERC-Fine): this task includes the recognition and classification of mentions according to finer-grained entity

Types	Sub-types	
pers	pers.ind	pers.ind.articleauthor
	pers.coll	
org	org.ent	org.ent.pressagency
	org.adm	
prod	prod.media	
	prod.doctr	
date	time.date.abs	
loc	loc.adm	loc.adm.town
		loc.adm.reg
		loc.adm.nat
		loc.adm.sup
	loc.phys	loc.geo
		loc.hydro
		loc.astro
	loc.oro	
	loc.fac	
	loc.add	loc.add.phys
		loc.add.elec

Table 1: Entity types used for NERC tasks.

types, as well as of nested entities and entity mention components (e.g. function, title, name).

Task 2: Named Entity Linking (EL). This task requires the linking of named entity mentions to a unique referent in a knowledge base – here Wikidata – or to a NIL node if the mention's referent is not present in the base. The entity linking task applies to non-nested mentions only and includes two settings: without and with prior knowledge of mention types and boundaries, referred to as end-to-end EL and EL only respectively.

4 Data

4.1 Corpus

The shared task corpus is composed of digitized and OCRized articles originating from Swiss, Luxembourgish and American historical newspaper collections and selected on a diachronic basis.[5]

[5] From the Swiss National Library, the Luxembourgish National Library, and the Library of Congress (Chronicling America project), respectively. Original collections correspond to 4 Swiss and Luxembourgish titles, and a dozen for English. More details on original sources can be found in [28].

Corpus selection. The corpus was compiled based on systematic and purposive sampling. For each newspaper and language, articles were randomly sampled among articles that a) belong to the first years of a set of predefined decades covering the life-span of the newspaper (longest duration spans ca. 200 years), and b) have a title, have more than 50 characters, and belong to any page. For each decade, the set of selected articles was additionally manually triaged in order to keep journalistic content only. Items corresponding to feuilleton, tabular data, cross-words, weather forecasts, time-schedules, obituaries, and those with contents that a human could not even read because of extreme OCR noise were therefore removed. Different OCR versions of same texts are not provided, and the OCR quality of the corpus therefore corresponds to real-life setting, with variations according to digitization time and preservation state of original documents. Figure 1 hereafter shows an example of a newspaper page facsimile, a selected article thereof, and its corresponding OCR. The corpus features an overall time span of ca. 200 years, from 1798 to 2018.

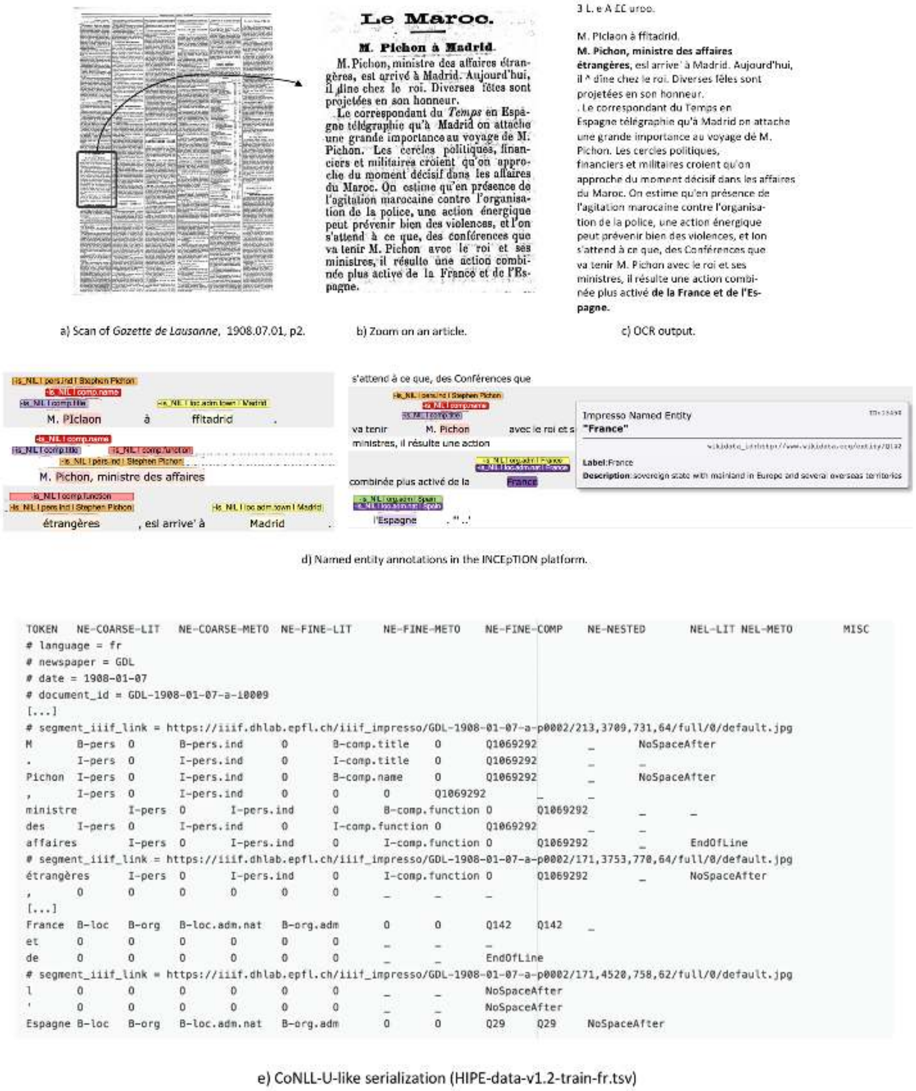
Named entity tagset and guidelines. The corpus was manually annotated according to the HIPE annotation guidelines [30]. Those guidelines were derived from the Quaero annotation guide, originally designed for the annotation of named entities in French speech transcriptions and already used on historical press corpora [92, 91]. HIPE slightly recast and simplified this guide, considering only a subset of entity types and components, as well as of linguistic units eligible as named entities. HIPE guidelines were iteratively consolidated via the annotation of a 'mini-reference' corpus—consisting of 10 content items per language—where annotation decisions were tested and difficult cases discussed[6]. Despite these adaptations, the HIPE corpus mostly remains compatible with Quaero-annotated data, as well as with the NewsEye project's NE data sets[7], annotated with guidelines derived from HIPE.

Table 1 presents the entity types and sub-types used for annotation, which participant systems had to recognize for NERC-Coarse (types) and NERC-Fine (most fine-grained sub-types). Named entity components, annotated for the type `Person` only, correspond to `name`, `title`, `function`, `qualifier` and `demonym`. Nested entities were annotated for `Person`, `Organization` and `Location` (a depth of 1 was considered during the evaluation), as well as metonymic senses, producing double tags for those entities referring to something intimately associated (metonymic sense) to the concept usually associated with their name (literal sense). As per entity linking, links correspond to Wikidata QIDs[8].

[6] The mini-reference corpus was released during the initial phase of the shared task as sample data and is available at `https://github.com/impresso/CLEF-HIPE-2020/tree/master/data/sample-v1.0`.

[7] `https://www.newseye.eu/`

[8] The November 2019 dump used for annotation is available at `https://files.ifi.uzh.ch/cl/impresso/clef-hipe`.



a) Scan of *Gazette de Lausanne*, 1908.07.01, p2.

Le Maroc.

M. Pichon à Madrid.

M. Pichon, ministre des affaires étrangères, est arrivé à Madrid. Aujourd'hui, il dine chez le roi. Diverses fêtes sont projetées en son honneur.

Le correspondant du *Temps* en Espagne télégraphie qu'à Madrid on attache une grande importance au voyage de M. Pichon. Les cercles politiques, financiers et militaires croient qu'on approche du moment décisif dans les affaires du Maroc. On estime qu'en présence de l'agitation marocaine contre l'organisation de la police, une action énergique peut prévenir bien des violences, et l'on s'attend à ce que, des conférences que va tenir M. Pichon avec le roi et ses ministres, il résulte une action combinée plus active de la France et de l'Espagne.

b) Zoom on an article.

3 L. e A ££ uroo.

M. PIclaon à ffitadrid.
M. Pichon, ministre des affaires
étrangères, esl arrive' à Madrid. Aujourd'hui,
il ^ dîne chez le roi. Diverses fêles sont
projetées en son honneur.
. Le correspondant du Temps en
Espagne télégraphie qu'à Madrid on attache
une grande importance au voyage dé M.
Pichon. Les cercles politiques,
financiers et militaires croient qu'on
approche du moment décisif dans les affaires
du Maroc. On estime qu'en présence de
l'agitation marocaine contre l'organisa-
tion de la police, une action énergique
peut prévenir bien des violences, et lon
s'attend à ce que, des Conférences que
va tenir M. Pichon avec le roi et ses
ministres, il résulte une action combi-
née plus activé **de la France et de l'Es-**
pagne.

c) OCR output.

is_NIL | pers.ind | Stephen Pichon
is_NIL | comp.name
is_NIL | comp.title
is_NIL | loc.adm.town | Madrid
M. PIclaon
à
ffitadrid
is_NIL | comp.name
is_NIL | comp.title
is_NIL | comp.function
is_NIL | pers.ind | Stephen Pichon
M. Pichon, ministre des affaires
is_NIL | comp.function
is_NIL | pers.ind | Stephen Pichon
is_NIL | loc.adm.town | Madrid
étrangères
, esl arrive' à
Madrid
s'attend à ce que, des Conférences que
is_NIL | pers.ind | Stephen Pichon
is_NIL | comp.name
is_NIL | comp.title
va tenir
M. Pichon
avec le roi et s
ministres, il résulte une action
is_NIL | org.adm | France
is_NIL | loc.adm.nat | France
combinée plus activé de la
France
is_NIL | org.adm | Spain
is_NIL | loc.adm.nat | Spain
l'Espagne
. " ..'

Impresso Named Entity
"France"
wikidata_id:http://www.wikidata.org/entity/Q142
Label: France
Description: sovereign state with mainland in Europe and several overseas territories

d) Named entity annotations in the INCEpTION platform.

```
TOKEN   NE-COARSE-LIT   NE-COARSE-METO  NE-FINE-LIT     NE-FINE-METO    NE-FINE-COMP    NE-NESTED       NEL-LIT NEL-METO        MISC
# language = fr
# newspaper = GDL
# date = 1908-01-07
# document_id = GDL-1908-01-07-a-i0009
[...]
# segment_iiif_link = https://iiif.dhlab.epfl.ch/iiif_impresso/GDL-1908-01-07-a-p0002/213,3709,731,64/full/0/default.jpg
M       B-pers  O       B-pers.ind      O       B-comp.title    O       Q1069292        _       NoSpaceAfter
.       I-pers  O       I-pers.ind      O       I-comp.title    O       Q1069292        _       _
Pichon  I-pers  O       I-pers.ind      O       B-comp.name     O       Q1069292        _       NoSpaceAfter
,       I-pers  O       I-pers.ind      O       O       O       Q1069292        _       _
ministre        I-pers  O       I-pers.ind      O       B-comp.function O       Q1069292        _       _
des     I-pers  O       I-pers.ind      O       I-comp.function O       Q1069292        _       _
affaires        I-pers  O       I-pers.ind      O       I-comp.function O       Q1069292        _       EndOfLine
# segment_iiif_link = https://iiif.dhlab.epfl.ch/iiif_impresso/GDL-1908-01-07-a-p0002/171,3753,770,64/full/0/default.jpg
étrangères      I-pers  O       I-pers.ind      O       I-comp.function O       Q1069292        _       NoSpaceAfter
,       O       O       O       O       O       O       _       _       _
[...]
France  B-loc   B-org   B-loc.adm.nat   B-org.adm       O       O       Q142    Q142    _
et      O       O       O       O       O       O       _       _       _
de      O       O       O       O       O       O       _       _       EndOfLine
# segment_iiif_link = https://iiif.dhlab.epfl.ch/iiif_impresso/GDL-1908-01-07-a-p0002/171,4520,758,62/full/0/default.jpg
l       O       O       O       O       O       O       _       _       NoSpaceAfter
'       O       O       O       O       O       O       _       _       NoSpaceAfter
Espagne B-loc   B-org   B-loc.adm.nat   B-org.adm       O       O       Q29     Q29     NoSpaceAfter
```

e) CoNLL-U-like serialization (HIPE-data-v1.2-train-fr.tsv)

Fig. 1: Illustration of HIPE data, from scanned image to NE annotation and CoNLL-U-liked released material. Example taken from the *Gazette de Lausanne* 1908.07.01, page 2.

Annotation framework. We used INCEpTION, a web-based platform for text annotation and knowledge management [57]. Segment (d) of Figure 1 offers two screenshots of the annotation platform, with the annotation of a person mention and its function (left), and of location mentions, with their metonymic senses (right). The functionalities of INCEpTION that proved particularly useful during the annotation campaign were: a) the support for querying against very large knowledge bases (e.g. Wikidata) with regards to EL annotation; b) the possibility of enabling the use of recommenders, which can considerably speed up the annotation process (e.g. when the very same named entity occurs multiple times within the same document); and c) the provision of an Abstract Programming Interface (API) that allows for automating certain operations, such as the bulk import/export of annotated documents. Moreover, since the shared task was one of the official use cases of the INCEpTION project,[9] some of the annotation platform's features were developed to accommodate specific needs of the HIPE annotation campaign, most notably the ability of displaying image segments alongside OCR transcriptions. Nevertheless, some aspects of our annotation process did not perfectly fit the generic workflows implemented in INCEpTION. First, the annotation by multiple annotators of different layers (i.e. mentions and entity links) within the same document and, second, the validation of annotated data so as to verify, for example, that every annotated mention has either a Wikidata link or the NIL flag. Both limitations were overcome by means of scripts based on the API.

Overall, INCEpTION proved to be a mature, stable and highly configurable annotation platform, able to support the complex workflows required by a collaborative annotation campaign such as the one undertaken for HIPE, as well as to deal with the specifities of historical newspaper data.

Annotation difficulties. The annotation campaign was carried out by the task organizers with the contribution of trilingual collaborators. Before starting annotating, each annotator was first trained on the mini-reference corpus in order to ensure a good understanding of the guidelines. This workflow proved to be valuable in resolving instruction's imprecisions and annotator's doubts, however some unclear points persisted and new difficulties appeared throughout the annotation campaign. As per NERC, major complications included, among others: a) the determination of entity boundaries in case of long functions or titles in apposition (e.g. *M. Curtoys d'Anduaga, doyen du corps diplojtëlfsue espagnol, et ministre plénipotentiaire pendant 50 ans*)[10]; b) the determination of what is to be considered (or not) as an `Organization`: despite clear specification, the definition of this class is not clear-cut and there are always groupings of some sort which prompt an interpretation as `Organization`, while they are not[11] (e.g. *Commission impériale*, *les gouvernements de l'Entente*, *Die französische Regierung*); c) the qualification of a location name as being of a region (`loc.adm.reg`) or of a nation (`loc.adm.nat`), particularly in a historical context (e.g. *Savoie, Moldavia*); d) the entanglement of entities, some of which have to be identified as nested or as components (e.g. the mix of `Person`, `Function` and `Organisation` in *Chez Manguet et Cherbuliez imprimaires-libraires à Genève*); e) the harmoniza-

[9] `https://inception-project.github.io/use-cases/impresso/`

[10] In these cases, we found that the annotation of components was greatly supporting the definition of entities' scope.

[11] According to our guidelines.

tion of rules across languages, e.g. with German compounds (e.g. *Zürichputsch*, *Baslerpropaganda*); f) the attempt to avoid country-related biases, such as the importance and role of canton councils in Switzerland vs. in other countries; and f) the annotation of metonymy, whose interpretation is rather subjective and may differ between annotators. We had no difficulties related to unreadable OCR since extremely noisy articles were filtered out beforehand, and since annotators could see original facsimiles while annotating.

With respect to Entity Linking, difficulties naturally related to the historical nature of the material. If it is highly preferable to have some historical background knowledge related to the collection—which most HIPE annotators had—, it appears that this is not in itself any guarantee of a swift resolution of mention referents. As a matter of fact, most person mentions in newspapers correspond to people who enjoyed a certain popularity at a specific time, but who are now medium- or little-known, except for experts in this or that Spanish dynasty, Swiss mountain tunnel, or local football club. As a result, historical background knowledge was mainly helpful for cases involving VIPs (e.g. *Wilhelm II*, *Jean Jaures*), and the linking of person mentions often proved to be comparable to detective work where one has to first understand who could be the person (by cross-referencing clues), before finding its ID in Wikidata (which, to our surprise, existed quite often). As a lesson learned, curiosity, persistence and investigation skills are as important as historical knowledge. Besides the mere identification of who's who, another difficulty was the choice of the relevant Wikidata ID for 'changing' entities, often locations, whose geographical and/or administrative realities evolved through time. Here the main issue turned out to be the unequal 'tracking' in the KB of the various historical statuses an entity could have had: while some countries might have an entry for each of their geopolitical phase (e.g. all French political regimes or German-related states throughout 19 and 20C), others have only a generic entry. This posed the problem of the coherency of annotation granularity, which, despite consistency checks, is not fully guaranteed in our data set, since none of the annotation—specific or generic—is entirely wrong, and annotators did take different decisions in the heat of annotation. The fuzzy setting in EL evaluation mitigates this aspect (see Section 5).

Overall, besides being time consuming, the annotation of multilingual historical texts proved to be rather challenging compared to our experience on contemporary data. As future improvements we took note of detailing further some points of the guidelines, specifically with respect to metonymy annotation and to entity linking.

Annotators' agreement. The inter-annotator agreement rates between two annotators were computed on a selection of documents (test set) using Krippendorf's α [59], as provided by INCEpTION version 0.15.2. Scores correspond to, for Fr, De and En respectively: .81, .79 and .80 for NERC, .73, .69 and .78 for linking towards a QID, and .95, .94 and .90 for linking towards NIL. NERC and linking towards NIL show a good agreement between annotators. The lower scores on entity linking confirm the difficulty of the task, especially in the con-

	Lg.	Docs	Tokens	Mentions	Literal			Metonymic		Nested	Comp.
					M	%noisy	%NIL	M	%NIL		
Train	fr	158	166218	7376	6925	9.27	28.16	451	3.33	473	3051
	de	104	86961	3832	3506	8.56	18.40	326	1.23	159	1441
	en	-	-	-	-	-	-	-	-	-	-
Total		262	253179	11208	10431	9.03	24.88	777	2.45	632	4492
Dev	fr	43	37953	1835	1727	10.65	22.52	108	0.00	91	724
	de	40	36176	1487	1390	18.92	23.88	97	3.09	75	563
	en	80	29060	981	966	1.35	45.86	15	0.00	-	-
Total		163	103189	4303	4083	11.27	28.51	220	1.36	166	1287
Test	fr	43	40855	1712	1600	13.13	22.81	112	1.79	82	709
	de	49	30738	1265	1147	15.34	20.92	118	0.00	73	431
	en	46	16635	474	449	7.13	42.54	25	0.00	-	-
Total		138	88228	3451	3196	13.08	24.91	255	0.78	155	1140
All	fr	244	245026	10923	10252	33.05	73.50	671	5.11	646	4484
	de	193	153875	6584	6043	42.82	63.21	541	4.32	307	2435
	en	126	45695	1455	1415	8.47	88.40	40	0.00	-	-
Total		563	444596	18962	17710	10.28	25.72	1252	1.92	953	6919

Table 2: Overview of corpus statistics (v1.3). **M** stands for number of mentions, **%noisy** stands for the percentage of mentions with at least one OCR error, and **%NIL** stands for the percentage of mentions linked to NIL.

text of historical documents. The low score observed on German (.69) is due to annotation discrepancies with respect to the linking of metonymic entities.

Corpus characteristics. For each task and language—with the exception of English—the HIPE corpus was divided into training, dev and test data sets (70/15/15). English was included later in the shared task and only dev and test sets were released for this language. The overall corpus consists of 563 annotated documents, for a total of 444,596 tokens and 18,962 (linked) mentions (see Table 2 for detailed overview statistics[12]). With 10,923 and 6,584 mentions, French and German corpora are larger than the English one (1,455). Despite our efforts to devise a balanced sampling strategy, the diachronic distribution of mentions is not entirely uniform across languages (see Fig. 2). This is mainly due to the following factors: the temporal boundaries of data to sample from (the German corpus stops at 1950, and the English one shortly afterwards); the varying content of newspaper articles; and, finally, the difficulty of sampling enough materials for certain decades due to OCR noise, such is the case with years 1850-1879 in the English corpus.

[12] These statistics are slightly different than those presented in [31] but, after thorough double checked, are to be considered as the reference ones.

An important aspect of the HIPE corpus, and of historical newspaper data in general, is the noise generated by OCR. Annotators were asked to transcribe the surface forms of noisy mentions so as to enable studying the impact of noisy mentions on NERC and EL tasks. In the test set—where we manually verified the consistency of annotators' transcriptions—about 10% of all mentions contain OCR mistakes.

Together with OCR, the limited coverage of knowledge bases such as Wikidata tends to have an impact on historical NE processing, and especially on linking. In our corpus, 25.72% of all literal mentions could not be linked to a Wikidata entry (NIL entities). Interestingly, and contrary to our initial assumption, NIL entities are uniformly distributed across time periods (see Fig 3). The NIL ratio is higher for `Person`, `Media` and `Organisation` entities, whereas for geographic places (`Location`) Wikidata shows a substantial coverage (see Table 3). `Date` mentions were not linked as per HIPE annotation guidelines.

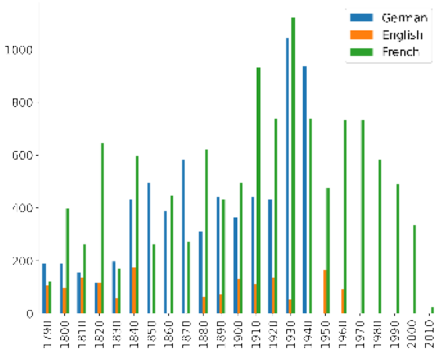


Fig. 2: Diachronic distribution of mentions across languages.

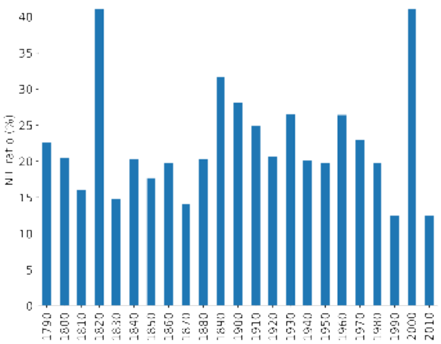


Fig. 3: Diachronic ratio of NIL entities.

Corpus release. Data sets were released in IOB format with hierarchical information, in a similar fashion to CoNLL-U[13], and consist of UTF-8, tab-separated-values files containing the necessary information for all tasks (NERC-Coarse, NERC-Fine, and EL).

Given the noisy quality of the material at hand, we chose not to apply sentence splitting nor sophisticated tokenization but, instead, to provide all necessary information to rebuild the OCR text. The tokenization applied to produce the IOB files is based on simple white space splitting, leaving all punctuation signs (including apostrophes) as separate tokens.[14] Participants could choose to apply their own sentence splitting and tokenization. Alongside each article, metadata (journal, date, title, page number, image region coordinates) and IIIF links

[13] `https://universaldependencies.org/format.html`

[14] The flag 'NoSpaceAfter' provides information about how to reconstruct the text.

Type	**Lg.**	**Literal**			**Metonymic**			**Lit.+Meto.**			**Nested**
		M	**L**	**%NIL**	**M**	**L**	**%NIL**	**M**	**L**	**%NIL**	
Location	fr	4716	4487	4.86	9	7	22.22	4725	4494	4.89	523
	de	3006	2883	4.09	30	30	0.00	3036	2913	4.05	209
	en	565	487	13.81	9	9	0.00	574	496	13.59	-
Total		8287	7857	5.19	48	46	4.17	8335	7903	5.18	732
Person	fr	3704	2051	44.63	0	0	0.00	3704	2051	44.63	33
	de	1910	1332	30.26	3	3	0.00	1913	1335	30.21	29
	en	558	154	72.40	0	0	0.00	558	154	72.40	-
Total		6172	3537	42.69	3	3	0.00	6175	3540	42.67	62
Organization	fr	1124	779	30.69	661	652	1.36	1785	1431	19.83	87
	de	660	458	30.61	507	505	0.39	1167	963	17.48	61
	en	194	120	38.14	31	31	0.00	225	151	32.89	-
Total		1978	1357	31.40	1199	1188	0.92	3177	2545	19.89	148
Date	fr	398	398	0.00	1	1	0.00	399	399	0.00	0
	de	240	240	0.00	0	0	0.00	240	240	0.00	6
	en	46	46	0.00	0	0	0.00	46	46	0.00	-
Total		684	684	0.00	1	1	0.00	685	685	0.00	6
Media	fr	310	231	25.48	0	0	0.00	310	231	25.48	3
	de	227	153	32.60	1	1	0.00	228	154	32.46	2
	en	52	20	61.54	0	0	0.00	52	20	61.54	-
Total		589	404	31.41	1	1	0.00	590	405	31.36	5
Grand Total		17710	13839	21.86	1252	1239	1.04	18962	15078	20.48	953

Table 3: Statistics per coarse entity type over all data sets, divided per language and per reading (literal and metonymic) and annotation depth (nested) types. **M** stands for 'mentions' (i.e. number of mentions), **L** stands for 'linked mentions' (i.e. number of mentions linked to Wikidata), and **%NIL** stands for the percentage of mentions linked to NIL.

to original page images are additionally provided when available [29]. Segment (e) of Figure 1 corresponds to an excerpt of the IOB HIPE data.

The HIPE corpus, comprising several versions of each data set for the 3 languages, is released under a CC BY-NC 4.0 license[15] and is available on Zenodo[16] as well as on the HIPE GitHub repository[17].

4.2 Auxiliary Resources

In order to support participants in their system design and experiments, we provided auxiliary resources in the form of 'in-domain' word and character-level

[15] `https://creativecommons.org/licenses/by-nc/4.0/legalcode`

[16] `https://zenodo.org/deposit/3706857`

[17] `https://github.com/impresso/CLEF-HIPE-2020/tree/master/data`

embeddings acquired from the same *impresso* newspapers titles and time periods from which HIPE training and development sets were extracted. Those embeddings correspond to fastText word embeddings [10] and flair contextualized string embeddings [5], both for French, German and English.

More specifically, fastText embeddings came in two versions, with subword 3-6 character n-grams and without, and were computed after a basic pre-processing (i.e., lower-casing, replacement of digits by 0 and deletion of all tokens and punctuation signs of length 1) that also tried to imitate the tokenization of the shared task data. Flair character embeddings were computed using flair 0.4.5[18] with a context of 250 characters, a batch size of 400-600 (depending on the GPU's memory), 1 hidden layer (size 2048), and a dropout of 0.1. Input was normalized with lower-casing, replacement of digits by 0, and of newlines by spaces; everything else was kept as in the original text (e.g. tokens of length 1). It is to be noted that the amount of training material greatly differed between languages (20G for French and 8.5G for German taken from Swiss and Luxembourgish newspapers; 1.1G for English taken from Chronicling America material).

These embeddings are released under a CC BY-SA 4.0 license[19] and are available for download.[20] Contextualized character embeddings were also integrated into the flair framework[21].

5 Evaluation Framework

5.1 Evaluation Measures

NERC and EL tasks are evaluated in terms of Precision, Recall and F-measure (F1) [67]. Evaluation is done at entity level according to two metrics: micro average, with the consideration of all TP, FP, and FN[22] over all documents, and macro average, with the average of document's micro figures. Our definition of macro differs from the usual one: averaging is done at document-level and not across entity-types, and allows to account for (historical) variance in document length and entity distribution within documents instead of overall class imbalances.

Both NERC and EL benefit from strict and fuzzy evaluation regimes. For NERC (Coarse and Fine), the strict regime corresponds to exact boundary matching and the fuzzy to overlapping boundaries. It is to be noted that in the strict regime, predicting wrong boundaries leads to a 'double' punishment of one false negative (entity present in the gold standard but not predicted by the system) and one false positive (entity predicted by the system but not present in the gold standard). Although it punishes harshly, we keep this metric to be in line with CoNLL and refer to the fuzzy regime when boundaries are of less importance.

[18] `https://github.com/flairNLP/flair`

[19] `https://creativecommons.org/licenses/by-sa/4.0/legalcode`

[20] `https://files.ifi.uzh.ch/cl/siclemat/impresso/clef-hipe-2020/flair/`

[21] `https://github.com/flairNLP/flair`

[22] True positive, False positive, False negative.

The definition of strict and fuzzy regimes differs for entity linking. In terms of boundaries, EL is always evaluated according to overlapping boundaries in both regimes (what is of interest is the capacity to provide the correct link rather than the correct boundaries). EL strict regime considers only the system's top link prediction (NIL or QID), while the fuzzy regime expands system predictions with a set of historically related entity QIDs. For example, "Germany" QID is complemented with the QID of the more specific "Confederation of the Rhine" entity and both are considered as valid answers. The resource allowing for such historical normalization was compiled by the task organizers for the entities of the test data sets, and is released as part of the HIPE scorer. For this regime, participants were invited to submit more than one link, and F-measure is additionally computed with cut-offs @3 and @5.

The HIPE scorer[23] was provided to the participants early on and is published under the MIT license. After the evaluation phase, a complete HIPE evaluation toolkit was also released, including the data used for evaluation (HIPE corpus v1.3), the anonymized system runs submitted by participating teams, and all the recipes and resources (e.g. historical mappings) required to replicate the present evaluation[24].

5.2 Task Bundles

In order to allow the greatest flexibility to participating teams as to which tasks to compete for while keeping a manageable evaluation frame, we introduced a system of task bundles offering different task combinations (see Table 4). Teams were allowed to choose only one bundle per language and to submit up to 3 runs per language. Only Bundle 5 (EL only) could be selected in addition to another one; this exception was motivated by the intrinsic difference between end-to-end linking and linking of already extracted entity mentions. Detailed information on system submission can be found in the HIPE Participation Guidelines [29].

Bundle	**Tasks**	**# teams**	**# runs**
1	NERC coarse, NERC fine and EL	2	10
2	NERC coarse and EL	3	10
3	NERC coarse and NERC fine	1	8
4	NERC coarse	7	27
5	EL only	5	20

Table 4: Task bundles.

[23] https://github.com/impresso/CLEF-HIPE-2020-scorer

[24] https://github.com/impresso/CLEF-HIPE-2020-eval

6 System Descriptions

In this first HIPE edition, 13 participating teams submitted a total of 75 system runs. All teams participated to NERC-Coarse, 3 to NERC-Fine, and 5 to end-to-end EL and EL only. The distribution of runs per language reflects the data, with 35 runs for French (42%), 26 for German (31%), and 22 for English (26%). Besides, six teams worked on all 3 languages. For NERC, all but 2 teams applied neural approaches, and most of them also worked with contextualized embeddings, in particular with BERT embeddings [108].

6.1 Baselines

As a baseline for NERC-Coarse, we trained a traditional CRF sequence classifier [77] using basic spelling features such as a token's character prefix and suffix, the casing of the initial character, and whether it is a digit. The model, released to participating teams as part of the HIPE scorer, dismisses the segmentation structure and treats any document as a single, long sentence. No baseline is provided for the NERC-Fine sub-task.

The baseline for entity linking (end-to-end EL and EL only) corresponds to AIDA-light [74], which implements the collective mapping algorithm by [47]. The wikimapper[25] tool was used to map Wikipedia URLs onto Wikidata QIDs, and the end-to-end EL baseline run relied on the CRF-based NERC baseline. Given the multilingual nature of the HIPE shared task, it is worth noting that AIDA-light was trained on a 2014 dump of the English Wikipedia, therefore accounting for a generous baseline.

6.2 Participating Systems

The following system descriptions are compiled from information provided by the participants. More accurate implementation details for most of the systems are available in the participants' papers [16]. As preliminary remarks, it can be noted that for NERC many teams experimented with different input embeddings, often testing character, sub-word and word-level representations trained on contemporary or historical material, and often combining classical type-level word embeddings (fastText) with contextualized embeddings (BERT, Flair, ELMo). Several teams also tried to improve the (newspaper) line-based input format by reconstructing linguistically motivated sentences and uniting hyphenated words. This preprocessing step turned out to be helpful.

Cisteria, a collaboration of the *Ludwig-Maximilians Universität* and the *Bayerische Staatsbibliothek München* from Germany, focused on NERC-coarse for German [96]. They experimented with external and HIPE character and word embeddings as well as several transformer-based BERT-style language models (e.g., German Europeana BERT[26]), all integrated by the neural flair NER tagging framework [5]. They used a state-of-the-art bidirectional LSTM with a

[25] `https://github.com/jcklie/wikimapper`

[26] `https://huggingface.co/dbmdz`

Conditional Random Field (CRF) layer as proposed by [49]. As a result of their experiments with a variety of pre-trained monolingual and multilingual word representations, they finally used different embeddings for literal and metonymic NERC models. No additional NER training material was used.

Ehrmama, affiliated with the University of Amsterdam, tackled coarse and fine-grained NERC for all languages [107]. They build on the bidirectional LSTM-CRF architecture of [63] and introduce a multi-task approach by splitting the top layers for each entity type. Their general embedding layer combines a multitude of embeddings, on the level of characters, sub-words and words; some newly trained by the team, as well as pre-trained BERT and HIPE's in-domain fastText embeddings. They also vary the segmentation of the input: line segmentation, document segmentation as well as sub-document segmentation for long documents. Their results suggest that splitting the top layers for each entity type is not beneficial. However, the addition of various embeddings improves the performance. Using (sub-)document segmentation clearly improved results when compared to the line segmentation found in newspapers. No additional NER training material was used for German and French; for English, the Groningen Meaning Bank [14] was adapted for training.

Ertim, affiliated with *Inalco*, Paris, applied their legacy (2010-13) NER system mXS[27] [75] for contemporary texts on the historical French HIPE data without any adaptation or training [76][28]. The system uses pattern mining and non-neural machine learning for NERC and their model is based on the QUAERO standard [92], which is the basis for the HIPE annotation guidelines. For EL, only the type `Person` was considered. The resolution is done in two steps, first an approximate string match retrieves French Wikipedia pages, second the Wikidata item is selected whose Wikipedia article has the highest cosine similarity with the HIPE newspaper article containing the mention.

Inria, by the *ALMAnaCH* project team affiliated at *Inria*, Paris, used DeLFT (Deep Learning Framework for Text)[29] for NERC tagging of English and French [60]. For English, the pre-trained Ontonotes 5.0 CoNLL-2012 model was used with a BiLSTM-CRF architecture. For EL-only, the off-the-shelf named entity recognition and linking system *entity-fishing*[30] was run on the HIPE data for predicting links for the literal meaning. For English, it achieved the best performance overall, for French, it ranked second best in F1 score. This Wikipedia-based system specifically stands out with its high recall.

Irisa, by a team from *IRISA*, Rennes, France, focused on French NERC and EL [33]. For NERC, they improved the non-neural CRF baseline system with additional features such as context tokens, date regex match, ASCII normalization of the focus token, and the 100 most similar words from the HIPE fastText word embeddings provided by the organizers. For EL, a knowledge-base driven approach was applied to disambiguate and link the mentions of their NERC

[27] `https://github.com/eldams/mXS`

[28] The final paper contains post-submission experiments.

[29] `https://github.com/kermitt2/delft`

[30] `https://github.com/kermitt2/entity-fishing`

systems and the gold oracle NERC mentions [32]. Their experiments with the HIPE data revealed that collective entity linking is also beneficial for this type of texts—in contrast to linking mentions separately.

L3i, by the L3i laboratory team affiliated with *La Rochelle University*, France, tackled all prediction tasks of HIPE for all languages and achieved almost everywhere the best results [13]. For NERC, they used a hierarchical transformer-based model [108] built upon BERT [22] in a multi-task learning setting. On top of the pre-trained BERT blocks (multilingual BERT for all languages, additionally Europeana BERT for German[31] and CamemBERT for French [69]), two task-specific transformer layers were optionally added to alleviate data sparsity issues, for instance out-of-vocabulary words, spelling variations, or OCR errors in the HIPE dataset. A state-of-the-art CRF layer was added on top in order to model the context dependencies between entity tags. For fine-tuning, relatively small batch sizes were used: 4 for German and English, 2 for French. For base BERT with a limited context of 512 sub-tokens, documents are too long and newspaper lines are too short for proper contextualization. Therefore, an important pre-processing step consisted in the reconstruction of hyphenated words and in sentence segmentation with Freeling [80]. The team submitted several runs based on different configurations of their model and resources. For the two languages with in-domain training data (French and German), the results of run 1 on literal NERC-coarse without the two transformer layers were slightly lower (roughly 1 percentage point in F score) than run 2 with transformer layers. For English without in-domain training data, two options for fine-tuning were tested: a) training on monolingual CoNLL 2003 data, and b) transfer learning by training on the French and German HIPE data. Both options worked better without transformer layers, (a) was slightly better on strict boundary evaluation, and (b) on fuzzy boundary evaluation. For their EL approach, based on [58], the team built a Wikipedia/Wikidata knowledge base per language and trained entity embeddings for the most frequent entries [40]. Based on Wikipedia co-occurrence counts, a probabilistic mapping table was computed for linking mentions with entities—taking several mention variations (e.g. lowercase, Levenshtein distance) into account to improve the matching. The candidates were filtered using DBpedia and Wikidata by prioritizing those that corresponded to the named entity type. For persons, they analysed the date of birth to discard anachronistic entities. Finally, the five best matching candidates were predicted.

Limsi, affiliated with *LIMSI, CNRS*, Paris, France, focused on NERC-coarse for French and achieved second best results there [41]. They submitted runs from 3 model variations: a) A model based on CamemBERT [69] that jointly predicts the literal and metonymic entities by feeding into two different softmax layers. This model performed best on the dev set for metonymic entities. b) The model (a) with a CRF layer on top, which achieved their best results on literal tags (F1=.814 strict). c) A standard CamemBERT model that predicts concatenated literal and metonymic labels directly as a combined tag (resulting

[31] `https://github.com/stefan-it/europeana-bert`

in a larger prediction tagset). This model performed best (within Limsi's runs) on the test set for metonymic entities (F1=.667 strict).

Nlp-uqam, affiliated with *Université du Quebec*, Montréal, Canada, focused on coarse NERC for French [21]. Their architecture involves a BiLSTM layer for word-level feature extraction with a CRF layer on top for capturing label dependencies [63], and an attention layer in between for relating different positions of a sequence [108]. For their rich word representation, they integrate a character-based CNN approach [18] and contextualized character-based flair embeddings [6] as provided by the HIPE organizers.

Sbb, affiliated with the Berlin State Library, Berlin, focused on NERC-coarse and EL for all languages [61]. For NERC, they applied a model based on multilingual BERT embeddings, which were additionally pre-trained on OCRed historical German documents from the SBB collection and subsequently fine-tuned on various multilingual NER data sets [62]. For EL, they constructed a multilingual knowledge base from Wikipedia (WP) articles roughly resembling the categories `Person`, `Location`, and `Organization`. The title words of these pages were embedded by BERT and stored in a nearest neighbor lookup index. A lookup applied to a mention returns a set of linked entity candidates. The historical text segment containing *the mention* and sentences from WP containing *a candidate* are then scored by a BERT sentence comparison model. This model was trained to predict for arbitrary WP sentence pairs whether they talk about the same entity or not. A random forest classifier finally ranks the candidates based on their BERT sentence comparison scores.

SinNer, affiliated with *INRIA* and Paris-Sorbonne University, Paris, France, focused on literal NERC-Coarse for French and German [78] and ranked third on both languages with their best neural run. The team preprocessed the line-based format into sentence-split segments. They provided two runs based on a BiLSTM ELMo architecture [82]. Run 1 is based on the classical ELMo architecture (without a CRF layer), combining type-level CNN word representations with a contextualized two-layer ELMo representation. For run 2, which performs better than their run 1 and is the one reported here, they combined modern Common Crawl-based fastText [43, 10] and pre-trained contextualized ELMo embeddings[32] in a modern BiLSTM-CRF architecture [103]. They optimized hyperparameters by training each variant three times and by selecting on F1 score performance on the dev set. For run 3, they retrained SEM[33] with the official HIPE data sets and applied entity propagation on the document level. For German, they augmented SEM's gazetteers with location lexicons crawled from Wikipedia. The considerably lower performance of run 3 illustrates the advantage of embedding-based neural NER tagging. Ablation experiments on sentence splitting showed an improvement of 3.5 F1 percentage points on French data for their neural system of run 1.

[32] [79] for French, [70] for German.

[33] SEM [24] is a CRF-based tool using Wapiti [64] as its linear-chain CRF implementation.

Upb, affiliated with the *Politehnica University of Bucharest*, Bucarest, Bulgaria, focused on literal NERC-coarse for all languages. Their BERT-based model centers around the ideas of transfer and multi-task learning as well as multilingual word embeddings. Their best performing runs combine multilingual BERT embeddings with a BiLSTM layer followed by a dense layer with local SoftMax predictions or alternatively, by adding a CRF layer on top of the BiLSTM.

Uva-ilps, affiliated with the University of Amsterdam and Radboud University, The Netherlands, worked on NERC-coarse and end-to-end EL for literal senses in all languages, as well as on literal and metonymic EL-only for English [86]. They fine-tuned BERT models for token-level NERC prediction using Huggingface's transformer framework [110], using the cased multilingual BERT base model for French and German and the cased monolingual BERT base model for English. For training their English model, they used the CoNLL-03 data [106]. Their end-to-end EL approach was implemented by searching for each entity mention in the English Wikidata dump indexed by ElasticSearch[34], an approach that outperformed the baseline system. The main problem there was the lack of German and French entities, although person names still could be found. For run 1 and 2 of EL-only on English, they improved the candidate entity ranking by calculating cosine similarities between the contextual embeddings of a sentence containing the target entity mention and a modified sentence where the mention was replaced with a candidate entity description from Wikidata. The semantic similarity scores were multiplied by relative Levenshtein similarity scores between target mention and candidate labels to prefer precise character-level matches. Run 2 added historical spelling variations, however, this resulted in more false positives. Run 3 used REL [51], a completely different neural NERC and EL system. Candidate selection in REL is twofold, 4 candidates are selected by a probabilistic model predicting entities given a mention, and 3 candidates are proposed by a model predicting entities given the context of the mention. Candidate disambiguation combines local compatibility (prior importance, contextual similarity) and global coherence with other document-level entity linking decisions. Their REL-based run 3 outperformed their runs 1 and 2 clearly.

Webis, by the *Webis* group affiliated with the *Bauhaus University Weimar*, Germany, focused on NERC-coarse for all languages. For each language, they trained a flair NERC sequence tagger [5] with a CRF layer using a stack of four embeddings: Glove embeddings [81], contextual character-based flair embeddings, and the forward and backward HIPE character-based flair embeddings. Their pre-processing included sentence reconstruction (by splitting the token sequence on all periods, except after titles, month abbreviations or numbers), and dehyphenation of tokens at the end of lines. For German, they experimented with data augmentation techniques by duplicating training set sentences and replacing the contained entities by randomly chosen new entities of the same type retrieved from Wikidata. A post-processing step resolved IOB tag sequence inconsistencies and applied a pattern-based tagging for time expressions. Although

[34] `https://www.elastic.co/`

internal dev set validation F1-scores looked promising, their official results on the test set had a bias towards precision. This could be due to format conversion issues.

7 Results and Discussion

We report results for the best run of each team and consider micro Precision, Recall and F1 scores exclusively. Results for NERC-Coarse and NERC-Fine for the three languages, both evaluation regimes and the literal and metonymic senses are presented in Table 5 and 6 respectively, while results for nested entities and entity components are presented in Table 7. Table 8 reports performances for end-to-end EL and EL only, with a cut-off @1 and Table 9 for EL only with cut-offs @3 and @5.

General observations. Neural systems with strong embedding resources clearly prevailed in HIPE NERC, beating symbolic CRF or pattern-matching based approaches by a large margin (e.g., compare baseline performance in Table 5). However, we also notice performance differences between neural systems that rely on BiLSTMs or BERT, the latter generally performing better.

In general and not unexpectedly, we observe that the amount of available training and development data correlates with system performances. French with the largest amount of training data has better results than German, and English is worse than German (see median numbers in Table 5). The one exception is EL only where English, as a well-resourced language, seems to have the necessary tooling to also excel on non-standard, historical text material (cf. Inria results). NERC-Coarse performances show a great diversity but top results are better than expected, specifically for French where they are almost on a par with performances on contemporary texts. Here, six teams have fuzzy F1 scores higher than .8, suggesting good prospects for entity extraction systems on historical texts, when trained with appropriate and sufficient data. Fine-grained NERC with more than 12 classes is obviously more difficult than predicting only 5 categories. However, the performance drop of the best performing system L3i is relatively mild for French, 6.5 percentage points on fuzzy F1, and a little stronger for German (10.7).

The recognition of entity components shows reasonable performances and suggests that knowledge base population and/or biography reconstruction from historical texts is feasible. The same cannot be said of nested entities.

Finally, EL performances are, as expected, lower than for NERC (best F1 score in the range of .58 to .63 for EL only strict across languages), and systems' performances are as diverse (cf. Table 8). The propagation of NERC mistakes in the end-to-end setting induces lower performances, however the provision of mention boundaries does not drastically improve results (e.g. 4 percentage points for the best system on French), suggesting that being able to deal with OCR noise (provided mentions are not OCR-corrected) and NIL entities is as important as exact mention recognition. When given the possibility to provide a list of

	French						German						English					
(a) Literal	Strict			Fuzzy			Strict			Fuzzy			Strict			Fuzzy		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
CISTERIA	-	-	-	-	-	-	.745	.578	.651	**.880**	.683	.769	-	-	-	-	-	-
EHRMAMA	.793	.764	.778	.893	.861	.877	.697	.659	.678	.814	.765	.789	.249	.439	.318	.405	.633	.494
ERTIM	.435	.248	.316	.604	.344	.439	-	-	-	-	-	-	-	-	-	-	-	-
INRIA	.605	.675	.638	.755	.842	.796	-	-	-	-	-	-	.461	.606	.524	.568	.746	.645
IRISA	.705	.634	.668	.828	.744	.784	-	-	-	-	-	-	-	-	-	-	-	-
L3I	**.831**	**.849**	**.840**	**.912**	**.931**	**.921**	**.790**	**.805**	**.797**	.870	**.886**	**.878**	**.623**	**.641**	**.632**	.794	**.817**	**.806**
LIMSI	.799	.829	.814	.887	.909	.898	-	-	-	-	-	-	-	-	-	-	-	-
NLP-UQAM	.705	.634	.668	.828	.744	.784	-	-	-	-	-	-	-	-	-	-	-	-
SBB	.530	.477	.502	.765	.689	.725	.499	.484	.491	.730	.708	.719	.347	.310	.327	.642	.572	.605
SINNER	.788	.802	.795	.886	.902	.894	.658	.658	.658	.775	.819	.796	-	-	-	-	-	-
UPB	.693	.686	.689	.825	.817	.821	.677	.575	.621	.788	.740	.763	.522	.416	.463	.743	.592	.659
UVA-ILPS	.656	.719	.686	.794	.869	.830	.499	.556	.526	.689	.768	.726	.443	.508	.473	.635	.728	.678
WEBIS	.731	.228	.347	.876	.273	.416	.695	.337	.454	.833	.405	.545	.476	.067	.117	**.873**	.122	.215
Baseline	.693	.606	.646	.825	.721	.769	.643	.378	.476	.790	.464	.585	.531	.327	.405	.736	.454	.562
Median	.705	.680	.677	.828	.829	.808	.686	.576	.636	.801	.752	.766	.461	.439	.463	.642	.633	.645
(b) Meto.	Strict			Fuzzy			Strict			Fuzzy			Strict			Fuzzy		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
CISTERIA	-	-	-	-	-	-	.738	.500	.596	.787	.534	.636	-	-	-	-	-	-
EHRMAMA	.697	.554	.617	.708	.562	.627	.696	.542	.610	.707	.551	.619	-	-	-	-	-	-
L3I	**.734**	**.839**	**.783**	**.734**	**.839**	**.783**	.571	**.712**	**.634**	.626	**.780**	**.694**	.667	**.080**	**.143**	**1.00**	**.120**	**.214**
LIMSI	.647	.688	.667	.655	.696	.675	-	-	-	-	-	-	-	-	-	-	-	-
NLP-UQAM	.423	.420	.422	.468	.464	.466	-	-	-	-	-	-	-	-	-	-	-	-
Baseline	.541	.179	.268	.541	.179	.268	**.814**	.297	.435	**.814**	.297	.435	**1.00**	.040	.077	**1.00**	.040	.077
Median	.647	.554	.617	.655	.562	.627	-	-	-	-	-	-	-	-	-	-	-	-

Table 5: Results for NERC-Coarse (micro P, R and F-measure). Bold font indicates the highest, and underlined font the second-highest value.

	French						German					
(a) Literal	Strict			Fuzzy			Strict			Fuzzy		
	P	R	F	P	R	F	P	R	F	P	R	F
EHRMAMA	.696	.724	.710	.776	.807	.791	**.650**	.592	.620	**.754**	.687	.719
ERTIM	.418	.238	.303	.568	.324	.412	-	-	-	-	-	-
L3I	**.772**	**.797**	**.784**	**.843**	**.869**	**.856**	.628	**.712**	**.668**	.734	**.813**	**.771**
(b) Metonymic												
EHRMAMA	.667	.554	.605	.667	.554	.605	**.707**	.551	.619	**.717**	.559	.629
L3I	**.718**	**.661**	**.688**	**.738**	**.679**	**.707**	.601	**.703**	**.648**	.659	**.771**	**.711**

Table 6: Results for NERC-Fine.

	French						German					
(a) Comp.	Strict			Fuzzy			Strict			Fuzzy		
	P	R	F	P	R	F	P	R	F	P	R	F
Ehrmama	**.695**	.632	**.657**	**.801**	.707	.751	**.681**	.494	.573	**.735**	.534	.618
Ertim	.042	.045	.043	.074	.080	.077	-	-	-	-	-	-
L3i	.680	**.732**	**.657**	.773	**.832**	**.801**	.595	**.698**	**.642**	.654	**.768**	**.707**
(b) Nested												
Ehrmama	**.397**	.280	.329	**.448**	.317	.371	-	-	-	-	-	-
L3i	.337	**.402**	**.367**	.357	**.427**	**.389**	.471	.562	.513	.517	.616	.562

Table 7: Results for nested entities and entity components.

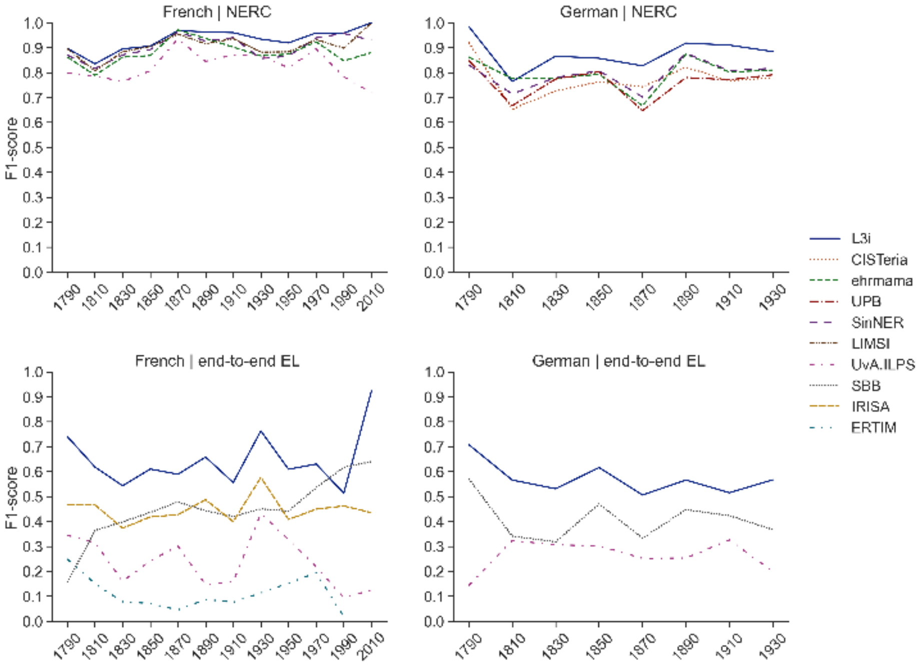
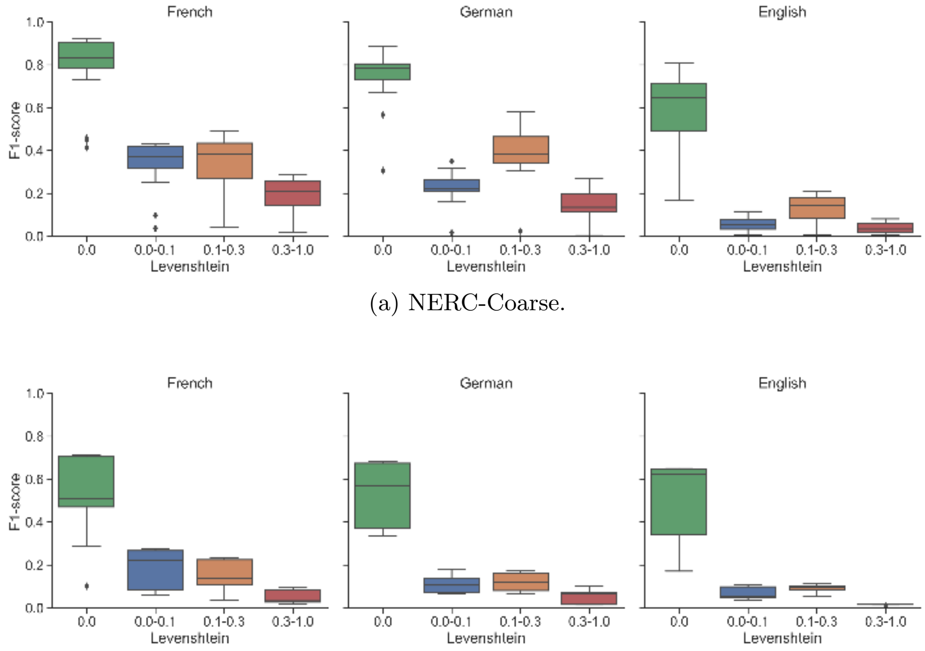


Fig. 4: F1 score as a function of time for the 5 best systems for NERC (top) and end-to-end EL (bottom) for the languages French (left) and German (right). The x-axis shows 20-years time buckets (e.g. 1790 = 1790-1809).

results and not only the top one, performances of all systems increase by about 1 (cut-off @3) to 2 (cut-off @5) points, showing the importance of candidate ranking (cf. Table 9).

System-based observations. With L3i, the HIPE 2020 campaign has a clear overall winner on NERC coarse and fine, literal and metonymic entities, components, as well as EL. The one exception is EL only for English, where Inria's entity-fishing system outperforms L3i. L3i is particularly convincing in terms of F1, as it consistently keeps precision and recall in good balance (even trending toward recall many times). Other systems, e.g. Inria, Ehrmama, or the baseline, typically suffer from a bias towards precision. It seems that actively tackling the problems of OCR noise, word hyphenation and sentence segmentation helps to achieve better recall.

Time-based observations. In order to gauge the impact of the article's publication date on system performances, we analyze the variation of F1 scores as a function of time (see Fig. 4). The initial hypothesis here was that the older the article, the more difficult it is to extract and link the mentions it contains. In general, there does not seem to be a strong correlation between the article's publication date and F1 scores. In the specific case of EL, this finding is in line with the uniform distribution of NIL entities across time (see Section 4).



(a) NERC-Coarse.

(b) End-to-end EL with the relaxed evaluation regime and a cutoff @3.

Fig. 5: Impact of OCR noise: distribution of performances across systems on entities with different noise level severity for NERC (a) and end-to-end EL (b).

End-to-end EL	**French**						**German**						**English**					
(a) Literal	Strict			Fuzzy			Strict			Fuzzy			Strict			Fuzzy		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Ertim	.150	.084	.108	.150	.084	.108	-	-	-	-	-	-	-	-	-	-	-	-
Irisa	.446	.399	.421	.465	.417	.439	-	-	-	-	-	-	-	-	-	-	-	-
L3i	**.594**	**.602**	**.598**	.613	**.622**	**.617**	.531	**.538**	**.534**	.553	**.561**	**.557**	**.523**	**.539**	**.531**	**.523**	**.539**	**.531**
Sbb	**.594**	.310	.407	**.616**	.321	.422	**.540**	.304	.389	**.561**	.315	.403	.257	.097	.141	.257	.097	.141
Uva-ilps	.352	.195	.251	.353	.196	.252	.245	.272	.258	.255	.283	.268	.249	.375	.300	.249	.375	.300
Baseline	.206	.342	.257	.257	.358	.270	.173	.187	.180	.188	.203	.195	.220	.263	.239	.220	.263	.239
(b) Meto.																		
Irisa	.023	.295	.043	.041	.527	.076	-	-	-	-	-	-	-	-	-	-	-	-
L3i	**.236**	**.402**	**.297**	**.366**	**.625**	**.462**	.324	.508	.396	.384	.602	.469	.172	.200	.185	.172	.200	.185
Baseline	.002	.027	.004	.008	.098	.015	.025	.136	.042	.026	.144	.044	.004	.040	.007	.004	.040	.007

EL only	**French**						**German**						**English**					
	Strict			Fuzzy			Strict			Fuzzy			Strict			Fuzzy		
(a) Literal	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Inria	.585	**.650**	.616	.604	**.670**	.635	-	-	-	-	-	-	**.633**	**.685**	**.658**	**.633**	**.685**	**.658**
Irisa	.475	.473	.474	.492	.491	.492	-	-	-	-	-	-	-	-	-	-	-	-
L3i	.640	.638	**.639**	.660	.657	**.659**	.581	**.582**	**.582**	.601	**.602**	**.602**	.593	.593	.593	.593	.593	.593
Sbb	**.677**	.371	.480	**.699**	.383	.495	**.615**	.349	.445	**.636**	.361	.461	.344	.119	.177	.344	.119	.177
Uva.ilps	-	-	-	-	-	-	-	-	-	-	-	-	.607	.580	.593	.607	.580	.593
Baseline	.502	.495	.498	.516	.508	.512	.420	.416	.418	.440	.435	.437	.506	.506	.506	.506	.506	.506
(b) Meto.																		
Irisa	.025	.357	.047	.041	.580	.076	-	-	-	-	-	-	-	-	-	-	-	-
L3i	**.303**	**.446**	**.361**	**.461**	**.679**	**.549**	**.443**	**.627**	**.519**	**.515**	**.729**	**.604**	**.286**	.480	**.358**	**.286**	.480	**.358**
Uva.ilps	-	-	-	-	-	-	-	-	-	-	-	-	.031	.058	.031	.031	.058	.031
Baseline	.213	.312	.254	.323	.473	.384	.265	.373	.310	.331	.466	.387	.219	.280	.246	.219	.280	.246

Table 8: Results for end-to-end EL (top) and EL only (bottom) with P, R and F1 @1.

	French			**German**			**English**		
@3	P	R	F	P	R	F	P	R	F
Irisa	.530	.463	.494	-	-	-	-	-	-
L3i	.676	.686	.681	.621	.630	.626	.627	.649	.638
Sbb	.624	.325	.428	.590	.332	.425	.299	.112	.163
Uva.ilps	.393	.218	.281	.300	.332	.315	.285	.429	.343
@5	P	R	F	P	R	F	P	R	F
Irisa	.554	.497	.524	-	-	-	-	-	-
L3i	.695	.705	.700	.627	.636	.632	.651	.674	.662
Sbb	.629	.328	.431	.601	.338	.432	.299	.112	.163
Uva.ilps	.397	.220	.283	.311	.345	.327	.304	.458	.366

Table 9: Results for EL-only with *fuzzy* P, R and F1 @3 and @5.

Impact of OCR noise. To assess the impact of noisy entities on the task of NERC and EL, we evaluated systems' performances on various noise levels (see Fig. 5). The level of noise is defined as the length-normalized Levenshtein distance between the surface form of an entity and its manual transcription. There is a remarkable difference between the performances for noisy and non-noisy mentions on both NERC and EL. Already as little noise as 0.1 severely hurts systems' abilities to predict an entity and may cut their performance by half. Interestingly, EL also suffers badly from little noise (norm. lev. dist. $>$ 0.0 and $<$ 0.1), even when provided with gold NERC annotations (EL only, not shown in the plot). Slightly and medium noisy mentions (norm. lev. dist. $>$ 0.0 and $<$ 0.3) show a similar impact, while for highly noisy mentions, the performance deteriorates further. We can observe the greatest variations between systems at the medium noise level, suggesting that the most robust systems get their competitive advantage when dealing with medium noisiness. On the effect of OCR noise on NERC, [104] claim that OCR errors impact more geo-political (GPE) mentions than persons or dates; in our breakdown of OCR noise impact by type, we can confirm that claim for little noise only (norm. lev. dist. $>$ 0.0 and $<$ 0.1), while this trend turns into the opposite for highly noisy entities.

8 Conclusion and Perspectives

From the perspective of natural language processing, the HIPE evaluation lab provided the opportunity to test the robustness of NERC and EL approaches against challenging historical material and to gain new insights with respect to domain and language adaptation. With regard to NERC, results show that it is possible to design systems capable of dealing with historical and noisy inputs, whose performances compete with those obtained on contemporary texts. Entity linking, as well as the processing of metonymy and nested entities remain challenging aspects of historical NE processing (the latter two probably due to the limited amount of annotated material). The results across the three languages present in the HIPE 2020 campaign suggest that performances mainly depend on the amount of the available in-domain training material. The evaluation study on influence of OCR noisy on performance confirmed the expectation of degraded quality for NERC and EL if more OCR errors are present. More surprising is the fact that neither NERC nor EL performance seem to correlate with the date of publication.

From the perspective of digital humanities, the lab's outcomes will help DH practitioners in mapping state-of-the-art solutions for NE processing on historical texts, and in getting a better understanding of what is already possible as opposed to what is still challenging. Most importantly, digital scholars are in need of support to explore the large quantities of digitized text they currently have at hand, and NE processing is high on the agenda. Such processing can support research questions in various domains (e.g. history, political science, literature, historical linguistics) and knowing about their performance is crucial in order to make an informed use of the processed data.

Overall, HIPE has contributed to advance the state of the art in semantic indexing of historical newspapers and, more generally, of historical material. As future work, we intend to explore several directions for a potential second edition of HIPE: expanding the language spectrum, strengthening the already covered languages by providing more training data, considering other types of historical documents, and exploring to what extent the improvements shown in HIPE can be transferred to similar tasks in other domains, or to linking problems that require knowledge bases other than Wikidata.

Acknowledgements

This HIPE evaluation lab would not have been possible without the interest and commitment of many. We express our warmest thanks to: the Swiss newspapers *NZZ* and *Le Temps*, and the Swiss and Luxembourg national libraries for sharing part of their data in the frame of the *impresso* project; Camille Watter, Gerold Schneider, Emmanuel Decker and Ilaria Comes for their commitment and hard work with the construction of the data set; the INCEpTION project team for its valuable and efficient support with the annotation tool; Richard Eckart de Castillo, Clemens Neudecker, Sophie Rosset and David Smith for their encouragement and guidance as part of the HIPE advisory board; and, finally, the 13 teams who embarked in this first HIPE edition, for their patience and scientific involvement. HIPE is part of the research activities of the project "*impresso* – Media Monitoring of the Past", for which we also gratefully acknowledge the financial support of the Swiss National Science Foundation under grant number CR-SII5_173719.

Bibliography

[1] ACE05: The ACE 2005 (ACE05) Evaluation Plan. Tech. rep., NIST ACE (10 2005), http://www.itl.nist.gov/iad/mig/tests/ace/2005/doc/ace05-evalplan.v3.pdf

[2] ACE08: ACE08 Evaluation Plan v1.2d. Tech. rep., NIST ACE (04 2008), http://www.itl.nist.gov/iad/mig//tests/ace/2008/doc/ace08-evalplan.v1.2d.pdf

[3] 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: 3rd International Workshop on Computational History (HistoInformatics 2016) (2016)

[4] Ahmed, S., Stoeckel, M., Driller, C., Pachzelt, A., Mehler, A.: BIOfid Dataset: Publishing a German Gold Standard for Named Entity Recognition in Historical Biodiversity Literature. In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). pp. 871–880. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/K19-1081, https://www.aclweb.org/anthology/K19-1081

[5] Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: An easy-to-use framework for state-of-the-art NLP. In: Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics (demonstrations). pp. 54–59. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://www.aclweb.org/anthology/N19-4010

[6] Akbik, A., Blythe, D., Vollgraf, R.: Contextual String Embeddings for Sequence Labeling. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1638–1649. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018), http://www.aclweb.org/anthology/C18-1139

[7] Artiles, J., Gonzalo, J., Sekine, S.: WePS 2 evaluation campaign: overview of the web people search clustering task. In: Proceedings of the 2nd Web People Search evaluation workshop (WePS 2009), collocated to the WWW conference (2009)

[8] Baldwin, T., de Marneffe, M.C., Han, B., Kim, Y.B., Ritter, A., Xu, W.: Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In: Proceedings of the Workshop on Noisy User-generated Text. pp. 126–135. Association for Computational Linguistics, Beijing, China (Jul 2015). https://doi.org/10.18653/v1/W15-4319, https://www.aclweb.org/anthology/W15-4319

[9] Benikova, D., Biemann, C., Kisselew, M., Pado, S.: Germeval 2014 named entity recognition shared task: Companion paper (2014), http://nbn-resolving.de/urn:nbn:de:gbv:hil2-opus-3006

[10] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017), https://www.aclweb.org/anthology/Q17-1010

[11] Bollmann, M.: A Large-Scale Comparison of Historical Text Normalization Systems. 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. 3885–3898. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1389

[12] Borin, L., Kokkinakis, D., Olsson, L.J.: Naming the past: Named entity and animacy recognition in 19th century Swedish literature. In: Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007). pp. 1–8 (2007)

[13] Boros, E., Linhares Pontes, E., Cabrera-Diego, L.A., Hamdi, A., Moreno, J.G., Sidère, N., Doucet, A.: Robust Named Entity Recognition and Linking on Historical Multilingual Documents. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)

[14] Bos, J., Basile, V., Evang, K., Venhuizen, N.J., Bjerva, J.: The Groningen meaning bank. In: Handbook of Linguistic Annotation, pp. 463–496. Springer Netherlands (2017)

[15] Bunescu, R., Paşca, M.: Using encyclopedic knowledge for named entity disambiguation. In: 11th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Trento, Italy (Apr 2006), https://www.aclweb.org/anthology/E06-1002

[16] Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.): CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings (2020)

[17] Chiron, G., Doucet, A., Coustaty, M., Visani, M., Moreux, J.P.: Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information. In: Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries. pp. 249–252. JCDL '17, IEEE Press, Piscataway, NJ, USA (2017), http://dl.acm.org/citation.cfm?id=3200334.3200364

[18] Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics **4**, 357–370 (2016). https://doi.org/10.1162/tacl_a_00104

[19] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research **12**(Aug), 2493–2537 (2011)

[20] Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). pp. 708–716. Association for Computational

Linguistics, Prague, Czech Republic (Jun 2007), https://www.aclweb.org/anthology/D07-1074

[21] Dekhili, G., Sadat, F.: Hybrid Statistical and Attentive Deep Neural Approach for Named Entity Recognition in Historical Newspapers. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)

[22] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805

[23] Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The automatic content extraction (ACE) program, tasks, data, and evaluation. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004). European Language Resources Association (ELRA), Lisbon, Portugal (May 2004), http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf

[24] Dupont, Y., Dinarelli, M., Tellier, I., Lautier, C.: Structured Named Entity Recognition by Cascading CRFs. In: Intelligent Text Processing and Computational Linguistics (CICling) (2017)

[25] Ehrmann, M., Colavizza, G., Rochat, Y., Kaplan, F.: Diachronic Evaluation of NER Systems on Old Newspapers. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)). pp. 97–107. Bochumer Linguistische Arbeitsberichte (2016), https://infoscience.epfl.ch/record/221391?ln=en

[26] Ehrmann, M., Nouvel, D., Rosset, S.: Named Entity Resources - Overview and Outlook. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)

[27] Ehrmann, M., Romanello, M., Bircher, S., Clematide, S.: Introducing the CLEF 2020 HIPE Shared Task: Named Entity Recognition and Linking on Historical Newspapers. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in Information Retrieval. pp. 524–532. Lecture Notes in Computer Science, Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_68

[28] Ehrmann, M., Romanello, M., Clematide, S., Ströbel, P.B., Barman, R.: Language Resources for Historical Newspapers: the *Impresso* Collection. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 958–968. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.lrec-1.121

[29] Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: HIPE - Shared Task Participation Guidelines (v1.1) (2020). https://doi.org/10.5281/zenodo.3677171

[30] Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Impresso Named Entity Annotation Guidelines (Jan 2020). https://doi.org/10.5281/zenodo.3604227

[31] Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11^{th} International Conference of the CLEF Association (CLEF 2020). Lecture Notes in Computer Science (LNCS), vol. 12260. Springer (2020)

[32] El Vaigh, C.B., Goasdoué, F., Gravier, G., Sébillot, P.: Using Knowledge Base Semantics in Context-Aware Entity Linking. In: Proceedings of the ACM Symposium on Document Engineering 2019. pp. 1–10. DocEng '19, Association for Computing Machinery, Berlin, Germany (Sep 2019). https://doi.org/10.1145/3342558.3345393

[33] El Vaigh, C.B., Le Noé-Bienvenu, G., Gravier, G., Sébillot, P.: IRISA System for Entity Detection and Linking at HIPE'20. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)

[34] Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05). pp. 363–370. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005). https://doi.org/10.3115/1219840.1219885, https://www.aclweb.org/anthology/P05-1045

[35] Frontini, F., Brando, C., Ganascia, J.G.: Semantic Web Based Named Entity Linking for Digital Humanities and Heritage Texts. In: Zucker, A., Draelants, I., Zucker, C.F., Monnin, A. (eds.) First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference. Portorož, Slovenia (2015), https://hal.archives-ouvertes.fr/hal-01203358

[36] Galibert, O., Rosset, S., Grouin, C., Zweigenbaum, P., Quintard, L.: Extended Named Entity Annotation on OCRed Documents : From Corpus Constitution to Evaluation Campaign. In: Proceedings of the Eighth conference on International Language Resources and Evaluation. pp. 3126–3131. Istanbul, Turkey (2012)

[37] Galibert, O., Leixa, J., Adda, G., Choukri, K., Gravier, G.: The ETAPE speech processing evaluation. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 3995–3999. European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014), http://www.lrec-conf.org/proceedings/lrec2014/pdf/1027_Paper.pdf

[38] Galibert, O., Rosset, S., Grouin, C., Zweigenbaum, P., Quintard, L.: Structured and extended named entity evaluation in automatic speech transcriptions. In: IJCNLP. pp. 518–526 (2011)

[39] Galliano, S., Geoffrois, E., Mostefa, M., Choukri, K., Bonastre, J.f., Gravier, G.: The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In: Proceedings of the 9^{th} European Conference on Speech Communication and Technology (INTERSPEECH'05. pp. 1149–1152 (2005)

[40] Ganea, O.E., Hofmann, T.: Deep Joint Entity Disambiguation with Local Neural Attention. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2619–2629 (2017)

[41] Ghannay, S., Grouin, C., Lavergne, T.: Experiments from LIMSI at the French Named Entity Recognition Coarse-grained task. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)

[42] Goulart, R.R.V., Strube de Lima, V.L., Xavier, C.C.: A systematic review of named entity recognition in biomedical texts. Journal of the Brazilian Computer Society **17**(2), 103–116 (Jun 2011). https://doi.org/10.1007/s13173-011-0031-9, https://doi.org/10.1007/s13173-011-0031-9

[43] Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://www.aclweb.org/anthology/L18-1550

[44] Grishman, R., Sundheim, B.: Design of the MUC-6 evaluation. In: Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland (1995)

[45] Grishman, R., Sundheim, B.: Message Understanding Conference - 6: A Brief History. In: Proceedings of the 16th International Conference on Computational Linguistics - Volume 1. pp. 466–471. COLING'96, Association for Computational Linguistics, Stroudsburg, PA, USA (1996). https://doi.org/10.3115/992628.992709, event-place: Copenhagen, Denmark

[46] Grover, C., Givon, S., Tobin, R., Ball, J.: Named Entity Recognition for Digitised Historical Texts. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco (May 2008), http://www.lrec-conf.org/proceedings/lrec2008/pdf/342_paper.pdf

[47] Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: EMNLP (2011)

[48] Hooland, S.V., De Wilde, M., Verborgh, R., Steiner, T., Van de Walle, R.: Exploring entity recognition and disambiguation for cultural heritage

collections. Digital Scholarship in the Humanities **30**(2), 262–279 (2015). https://doi.org/10.1093/llc/fqt067

[49] Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. CoRR **abs/1508.01991** (2015), `http://arxiv.org/abs/1508.01991`

[50] Hubková, H.: Named-entity recognition in Czech historical texts: Using a CNN-BiLSTM neural network model. Ph.D. thesis (2019)

[51] van Hulst, J.M., Hasibi, F., Dercksen, K., Balog, K., de Vries, A.P.: REL: An entity linker standing on the shoulders of giants. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '20, ACM (2020)

[52] Jones, A., Crane, G.: The challenge of virginia banks: an evaluation of named entity analysis in a 19th-century newspaper collection. In: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'06). pp. 31–40. IEEE (2006)

[53] Kaplan, F., di Lenardo, I.: Big Data of the Past. Frontiers in Digital Humanities **4** (2017). https://doi.org/10.3389/fdigh.2017.00012

[54] Kettunen, K., Ruokolainen, T.: Names, Right or Wrong: Named Entities in an OCRed Historical Finnish Newspaper Collection. In: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2017. pp. 181–186. ACM Press (2017). https://doi.org/10.1145/3078081.3078084

[55] Kim, J.D., Ohta, T., Tateisi, Y., Tsujii, J.: Genia corpus—a semantically annotated corpus for bio-textmining. Bioinformatics **19**(suppl 1), i180–i182 (2003)

[56] Kim, S.M., Cassidy, S.: Finding Names in Trove: Named Entity Recognition for Australian Historical Newspapers. In: Proceedings of the Australasian Language Technology Association Workshop 2015. pp. 57–65. Parramatta, Australia (Dec 2015)

[57] Klie, J.C., Bugert, M., Boullosa, B., de Castilho, R.E., Gurevych, I.: The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. pp. 5–9 (2018)

[58] Kolitsas, N., Ganea, O.E., Hofmann, T.: End-to-End Neural Entity Linking. In: Proceedings of the 22nd Conference on Computational Natural Language Learning. pp. 519–529. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/K18-1050

[59] Krippendorff, K.: Content analysis: An introduction to its methodology. Sage publications (1980)

[60] Kristanti, T., Romary, L.: DeLFT and entity-fishing: Tools for CLEF HIPE 2020 Shared Task. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)

[61] Labusch, K., Neudecker, C.: Named Entity Disambiguation and Linking Historic Newspaper OCR with BERT. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Working Notes. Working Notes of

CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)

[62] Labusch, K., Neudecker, C., Zellhöfer, D.: BERT for Named Entity Recognition in Contemporary and Historic German. In: Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers. pp. 1–9. German Society for Computational Linguistics & Language Technology, Erlangen, Germany (2019)

[63] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural Architectures for Named Entity Recognition. arXiv:1603.01360 [cs] (Mar 2016), http://arxiv.org/abs/1603.01360

[64] Lavergne, T., Cappé, O., Yvon, F.: Practical very large scale CRFs. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 504–513. Association for Computational Linguistics (2010)

[65] Linhares Pontes, E., Hamdi, A., Sidere, N., Doucet, A.: Impact of OCR Quality on Named Entity Linking. In: Jatowt, A., Maeda, A., Syn, S.Y. (eds.) Digital Libraries at the Crossroads of Digital Information for the Future. pp. 102–115. Lecture Notes in Computer Science, Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-34058-2_11

[66] Magnini, B., Cappelli, A., Tamburini, F., Bosco, C., et al., A.M.: Evaluation of natural language tools for italian: Evalita 2007. In: Proc. of the 6^{th} International Conference on Language Resources and Evaluation (LREC'08). Marrakech, Morocco (2008)

[67] Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R.: Performance measures for information extraction. In: In Proceedings of DARPA Broadcast News Workshop. pp. 249–252 (1999)

[68] Markert, K., Nissim, M.: SemEval-2007 task 08: Metonymy resolution at SemEval-2007. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). pp. 36–41. Association for Computational Linguistics, Prague, Czech Republic (Jun 2007), https://www.aclweb.org/anthology/S07-1007

[69] Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B.: CamemBERT: a tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7203–7219. Association for Computational Linguistics, Online (Jul 2020), https://www.aclweb.org/anthology/2020.acl-main.645

[70] May, P.: German ELMo Model (2019), https://github.com/t-systems-on-site-services-gmbh/german-elmo-model

[71] Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)

[72] Neudecker, C., Antonacopoulos, A.: Making Europe's Historical Newspapers Searchable. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS). pp. 405–410. IEEE, Santorini, Greece (Apr 2016). https://doi.org/10.1109/DAS.2016.83

[73] Neudecker, C., Wilms, L., Faber, W.J., van Veen, T.: Large-scale refinement of digital historic newspapers with named entity recognition. In: Proc IFLA Newspapers/GENLOC Pre-Conference Satellite Meeting (2014)

[74] Nguyen, D.B., Hoffart, J., Theobald, M., Weikum, G.: Aida-light: High-throughput named-entity disambiguation. In: LDOW (2014)

[75] Nouvel, D., Antoine, J.Y., Friburger, N.: Pattern Mining for Named Entity Recognition. LNCS/LNAI Series **8387i (post-proceedings LTC 2011)** (2014)

[76] Nouvel, D., Zagabe Seruti, J.C.: Adapting a pre-neural ML NER System to Historical Data. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)

[77] Okazaki, N.: CRFsuite: a fast implementation of Conditional Random Fields (CRFs) (2007), `http://www.chokkan.org/software/crfsuite/`

[78] Ortiz Suárez, P.J., Dupont, Y., Lejeune, G., Tian, T.: SinNer@CLEF-HIPE2020: Sinful adaptation of SotA models for Named Entity Recognition in historical French and German newspapers. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)

[79] Ortiz Suárez, P.J., Dupont, Y., Muller, B., Romary, L., Sagot, B.: Establishing a new state-of-the-art for French named entity recognition. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 4631–4638. European Language Resources Association, Marseille, France (May 2020), `https://www.aclweb.org/anthology/2020.lrec-1.569`

[80] Padró, L., Stanilovsky, E.: FreeLing 3.0: Towards wider multilinguality. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 2473–2479. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012), `http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf`

[81] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. vol. 14, pp. 1532–43 (2014)

[82] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep Contextualized Word Representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (2018). https://doi.org/10.18653/v1/N18-1202

[83] Piotrowski, M.: Natural language processing for historical texts. Synthesis Lectures on Human Language Technologies **5**(2), 1–157 (2012)

[84] Piskorski, J., Ehrmann, M.: On Named Entity Recognition in Targeted Twitter Streams in Polish. In: Proceedings of the 4^{th} Biennial International Workshop on Balto-Slavic Natural Language Processing, co-located with ACL (BSNLP 2013). pp. 84–93. Sofia, Bulgaria (2013)

[85] Plank, B.: What to do about non-standard (or non-canonical) language in NLP. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)). Bochumer Linguistische Arbeitsberichte (2016)

[86] Provatorova, V., Vakulenko, S., Kanoulas, E., Dercksen, K., van Hulst, J.M.: Named Entity Recognition and Linking on Historical Newspapers: UvA.ILPS & REL at CLEF HIPE 2020. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)

[87] Rao, D., McNamee, P., Dredze, M.: Entity linking: Finding extracted entities in a knowledge base. In: Poibeau, T., Saggion, H., Piskorski, J., Yangarber, R. (eds.) Multi-source, multilingual information extraction and summarization, pp. 93–115. Springer Berlin Heidelberg, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-28569-1, https://doi.org/10.1007/978-3-642-28569-1

[88] Riedl, M., Padó, S.: A named entity recognition shootout for german. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 120–125 (2018)

[89] Ritter, A., Clark, S., Etzioni, O., et al.: Named entity recognition in tweets: an experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1524–1534 (2011)

[90] Rodriquez, K.J., Bryant, M., Blanke, T., Luszczynska, M.: Comparison of named entity recognition tools for raw OCR text. In: Jancsary, J. (ed.) 11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, September 19-21, 2012. Scientific series of the ÖGAI, vol. 5, pp. 410–414. ÖGAI, Wien, Österreich (2012), http://www.oegai.at/konvens2012/proceedings/60_rodriquez12w/

[91] Rosset, S., Grouin, C., Fort, K., Galibert, O., Kahn, J., Zweigenbaum, P.: Structured named entities in two distinct press corpora: Contemporary broadcast news and old newspapers. In: Proceedings of the 6th Linguistic Annotation Workshop. pp. 40–48. Association for Computational Linguistics (2012)

[92] Rosset, Sophie, Grouin, Cyril, Zweigenbaum, Pierre: Entités nommées structurées : guide d'annotation Quaero. NOTES et DOCUMENTS 2011-04, LIMSI-CNRS (2011)

[93] Rovera, M., Nanni, F., Ponzetto, S.P., Goy, A.: Domain-specific named entity disambiguation in historical memoirs. In: CEUR Workshop Proceedings. vol. 2006. RWTH (2017)

[94] Santos, D., Seco, N., Cardoso, N., Vilela, R.: HAREM: An Advanced NER Evaluation Contest for Portuguese. In: Proceedings of the 5^{th} International Conference on Language Resources and Evaluation (LREC'06). pp. 1640–1643. Genoa (2006)

[95] Schweter, S., Baiter, J.: Towards robust named entity recognition for historic german. arXiv preprint arXiv:1906.07592 (2019)

[96] Schweter, S., März, L.: Triple E - Effective ensembling of embeddings and language models for NER of historical German. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)

[97] Sekine, S., Sudo, K., Nobata, C.: Extended Named Entity Hierarchy. In: Proceedings of The Third International Conference on Language Resources and Evaluation (LREC). Iles Canaries , Espagne (2002)

[98] Sevgili, O., Shelmanov, A., Arkhipov, M., Panchenko, A., Biemann, C.: Neural Entity Linking: A Survey of Models based on Deep Learning. arXiv:2006.00575 [cs] (May 2020), http://arxiv.org/abs/2006.00575, arXiv: 2006.00575

[99] Shen, W., Wang, J., Han, J.: Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. IEEE Transactions on Knowledge and Data Engineering **27**(2), 443–460 (Feb 2015). https://doi.org/10.1109/TKDE.2014.2327028, http://ieeexplore.ieee.org/document/6823700/

[100] Smith, D.A., Cordell, R.: A Research Agenda for Historical and Multilingual Optical Character Recognition. Tech. rep. (2018), http://hdl.handle.net/2047/D20297452

[101] Sporleder, C.: Natural Language Processing for Cultural Heritage Domains. Language and Linguistics Compass **4**(9), 750–768 (2010). https://doi.org/10.1111/j.1749-818X.2010.00230.x

[102] Sprugnoli, R.: Arretium or Arezzo? A Neural Approach to the Identification of Place Names in Historical Texts. In: Cabrio, E., Mazzei, A., Tamburini, F. (eds.) Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018. CEUR Workshop Proceedings, vol. 2253. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2253/paper26.pdf

[103] Straková, J., Straka, M., Hajic, J.: Neural architectures for nested NER through linearization. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5326–5331. Association for Computational Linguistics, Florence, Italy (Jul 2019), https://www.aclweb.org/anthology/P19-1527

[104] van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., McGillivray, B., Colavizza, G.: Assessing the Impact of OCR Quality on Downstream NLP Tasks. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence. pp. 484–496. SCITEPRESS - Science and Technology Publications, Valletta, Malta (2020). https://doi.org/10.5220/0009169004840496

[105] Terras, M.: The Rise of Digitization. In: Rikowski, R. (ed.) Digitisation Perspectives, pp. 3–20. SensePublishers, Rotterdam (2011). https://doi.org/10.1007/978-94-6091-299-3_1

[106] Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proceedings of the seventh conference on Natural language learning at HLT-

NAACL 2003-Volume 4. pp. 142–147. Association for Computational Linguistics (2003)

[107] Todorov, K., Colavizza, G.: Transfer Learning for Named Entity Recognition in Historical Corpora. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)

[108] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017), `http://arxiv.org/abs/1706.03762`

[109] Vilain, M., Su, J., Lubar, S.: Entity Extraction is a Boring Solved Problem: Or is It? In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers. pp. 181–184. NAACL-Short '07, Association for Computational Linguistics (2007), `http://dl.acm.org/citation.cfm?id=1614108.1614154`, event-place: Rochester, New York

[110] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv e-prints arXiv:1910.03771 (Oct 2019)