CLEF-HIPE-2020
Named Entity Recognition and Linking on Historical Newspapers

Maud Ehrmann (EPFL)
Matteo Romanello (EPFL)
Alex Flückiger (UZH)
Simon Clematide (UZH)

CLEF 2020 (online), 23.09.2020

EPFL

impresso Media Monitoring of the Past

Universität Zürich[UZH]

impresso Media Monitoring of the Past

CLEF-HIPE-2020 in a nutshell

- **HIPE**: Identifying **Hi**storical **P**eople, **P**laces and other **E**ntities

- 1st NE processing shared task on historical documents

- Tasks:
    - NE recognition and classification
    - NE linking

- Participating teams: 13

CLEF ASSOCIATION

CLEF 2020 Thessaloniki

Why HIPE?

New data:
Emergence of large-scale archives of digitized contents

New needs:
Content retrieval by humanities scholars

Challenge: NLP on historical texts is hard

- Spelling variations
- Noisy OCR
- Multilingualism
- Data sparsity
- Limited resources or KB coverage

➔ Objectives

1. strengthen the **robustness** of approaches;
2. enable **performance comparison**;
3. foster **efficient semantic indexing** of digitized cultural heritage collections.

Background

impresso project
mining 200 years of historical newspapers

EPFL
Digital Humanities Laboratory (DHLAB)
Ecole Polytechnique Federale de Lausanne, Switzerland

University of Zurich[UZH]
Institute of Computational Linguistics Zurich University, Switzerland

C²DH
Centre for Contemporary and Digital History (C2DH) Luxembourg University, Luxembourg.

Swiss National Library, SNL
National Library of Luxembourg, BnL.
State Archives of Valais, AEV.
University of Basel
Swiss Economic Archives, SWA.
LE TEMPS
Le Temps
NZZ
Neue Zürcher Zeitung, NZZ.
UNIL | Université de Lausanne
History department, University of Lausanne, UNIL.
infoclio.ch
infoclio

FNSNF
Swiss National Science Foundation

Project: https://impresso-project.ch
Interface: https://impresso-project.ch/app/

Semantic indexation of historical newspapers

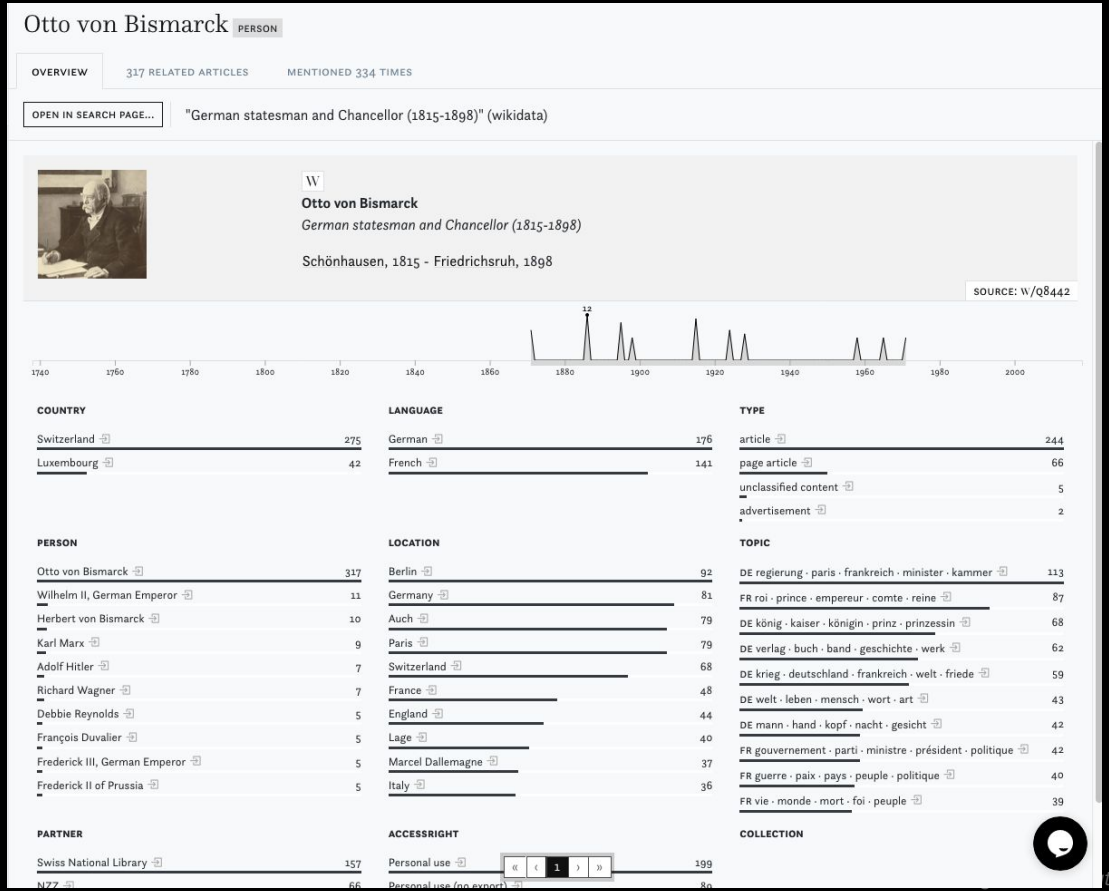
Search NEs
(among others)
over 47M articles

Semantic indexation of historical newspapers

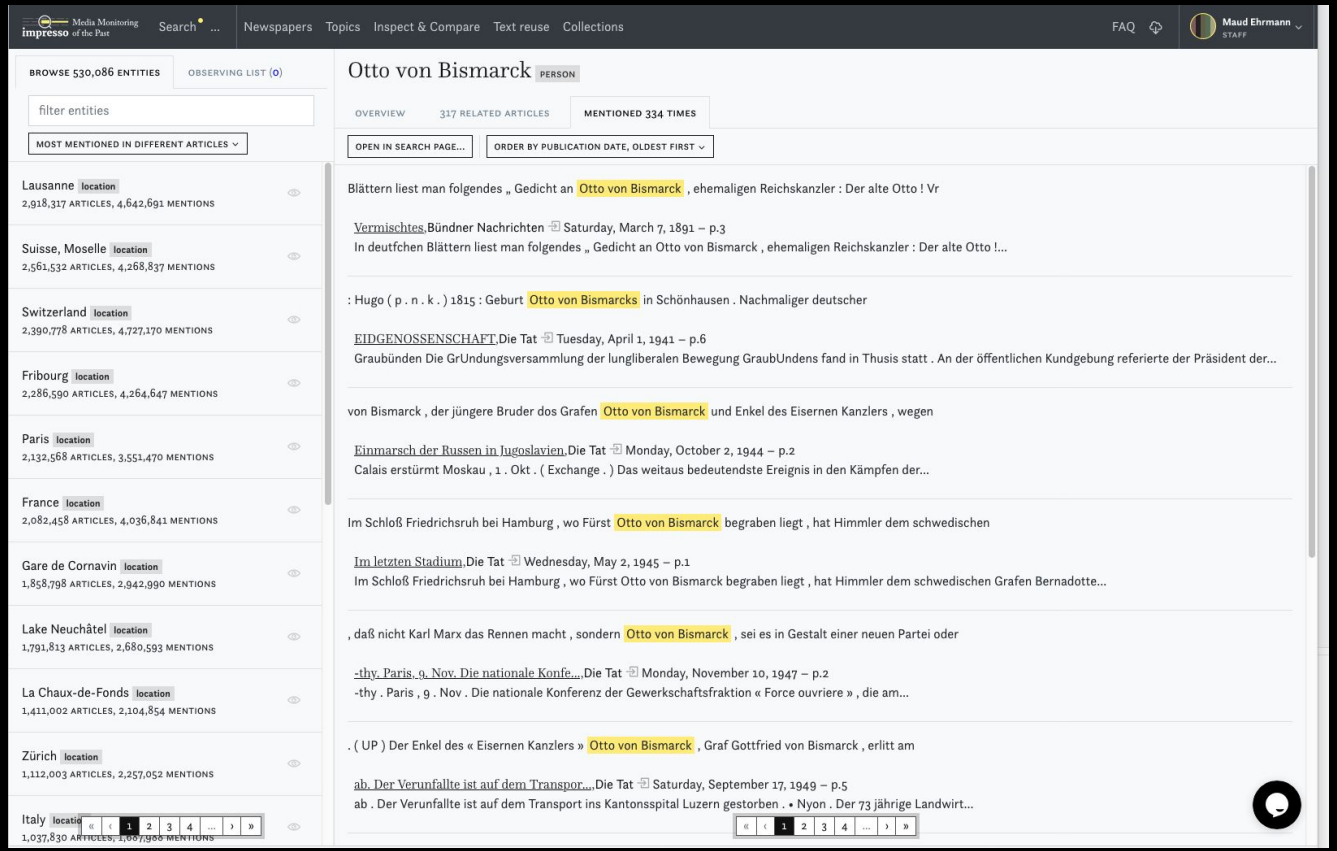
Visualize facsimile, OCR and entity mentions

Semantic indexation of historical newspapers

Overview of named entities



Semantic indexation of historical newspapers



Tasks

1. NERC

Recognition and classification of entity mentions with
- subtask 1: coarse types
- subtask 2: fine-grained types.

NERC Coarse

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

NERC Fine
+ entity components
+ metonymy
+ nested entities

Table 1: Entity types used for NERC tasks.

Tasks

2. Entity Linking

Towards Wikidata QID or NIL

- end-to-end EL: w/o mention boundaries
- EL-only: with mention boundaries

Participation bundles:

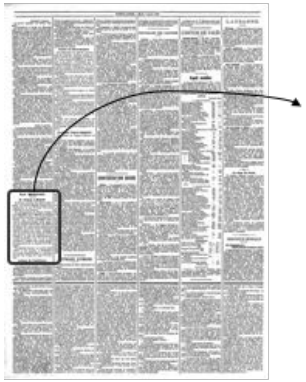
Bundle id	Associated tasks
bundle1	NERC-coarse and NERC-fine and NEL
bundle2	NERC-coarse and NEL
bundle3	NERC-coarse and NERC-fine
bundle4	NERC-coarse
bundle5	NEL

Participation guidelines:

10.5281/zenodo.3677171

Corpus selection

- Digitized newspaper archives (CH, LU, US)
- Diachronic: from 1738 to 2019
- Multilingual: fr, de, en
- Sampling and manual triage:
  - journalistic content
  - no feuilleton, cross-words, meteo, etc.
  - exclusion of extreme OCR noise
  - no provision of different OCR ➜ real-life setting



Le Maroc.

M. Pichon à Madrid.

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

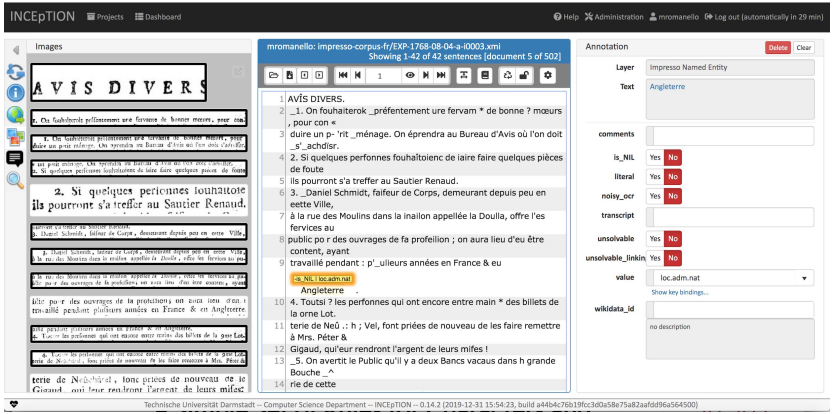
Le correspondant du *Temps* en Espagne télégraphie qu'à Madrid on attache une grande importance au voyage de M. Pichon. Les cercles politiques, financiers et militaires croient qu'on approche du moment décisif dans les affaires du Maroc. On estime qu'en présence de l'agitation marocaine contre l'organisation 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.

3 L. e A ££ uroo.

M. Piclaon à ffltadrid.
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'organisation 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 combinée plus activé **de la France et de l'Espagne.**

Corpus annotation

- Trilingual annotators, trained on a mini-ref
- INCEpTION platform



- NERC annotation difficulties:
  - NE mention boundaries
  - consideration of multiple languages
  - what is to be annotated or not
  - definition at time *x*
  - metonymy

M. Curtoys d' Anduaga, doyen du corps diplojt□elfsue espagnol, et ministre plenipotentiaire pendant 50 ans

Z□urichputsch, Baslerpropaganda

Commission imperiale, Die franz□osische Regierung

Is *Savoie or Moldavia* a region or a country?

Annotation guidelines
10.5281/zenodo.3604227

Corpus annotation

- Trilingual annotators, trained on a mini-ref
- INCEpTION platform
- NERC annotation difficulties:
  - NE mention boundaries
  - consideration of multiple languages
  - what is to be annotated or not
  - definition at time *x*
  - metonymy

Germany, Q183
962-1813: Holy Roman Empire, Q12548
1806-1813: Confederation of the Rhine, Q154741
1815-1866: German Confederation, Q151624
1867-1870: North German Confederation,Q150981
1871-1918: German Empire, Q43287
1918-1933: Weimar Republic, Q41304
1933-1945: Nazi Germany, Q7318
1949-1990: West Germany, Q713750
1949-1990: East Germany, Q16957

- EL annotation difficulties:
  - Requires historical knowledge + Sherlock Holmes skills
  - Historical statuses of entities unequally represented in KB

Annotation guidelines
10.5281/zenodo.3604227

Corpus characteristics

newspaper articles	563
tokens	444,596
(linked) mentions	18,962
metonymy	1252
components	6,219
noisy mentions (test set)	10%
NIL	25.72%

mentions: 10,923 (Fr), 6584 (De), 1455 (En)

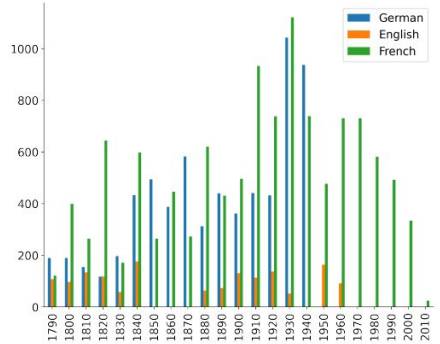


Fig. 2: Diachronic distribution of mentions across languages.

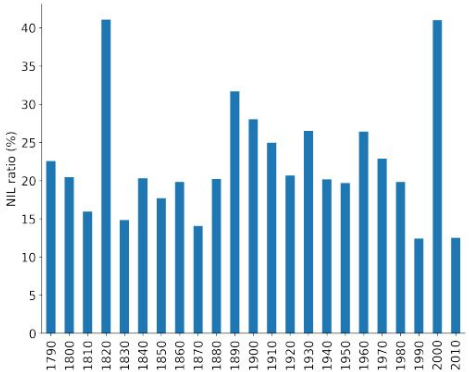


Fig. 3: Diachronic ratio of NIL entities.

Corpus release

- train/dev/test (70/15/15)
- no train set for English
- no sentence segmentation
- no sophisticated tokenization
- document metadata

```
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-p0002/296,3651,514,64/full/0/default.jpg
M   B-pers  0   B-pers.ind  0   B-comp.title    0   Q1069292    _   NoSpaceAfter
.   I-pers  0   I-pers.ind  0   I-comp.title    0   Q1069292    _   _
PIclaon I-pers  0   I-pers.ind  0   B-comp.name 0   Q1069292    _   _
à   0   0   0   0   0   0   _   _   _
ffltadrid   B-loc   0   B-loc.adm.town  0   0   0   Q2807   _   NoSpaceAfter
.   0   0   0   0   0   0   _   _   EndOfLine
# 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  0   B-pers.ind  0   B-comp.title    0   Q1069292    _   NoSpaceAfter
.   I-pers  0   I-pers.ind  0   I-comp.title    0   Q1069292    _   _
Pichon  I-pers  0   I-pers.ind  0   B-comp.name 0   Q1069292    _   NoSpaceAfter
,   I-pers  0   I-pers.ind  0   0   0   Q1069292    _   _
ministre    I-pers  0   I-pers.ind  0   B-comp.function 0   Q1069292    _   _
des I-pers  0   I-pers.ind  0   I-comp.function 0   Q1069292    _   _
affaires    I-pers  0   I-pers.ind  0   I-comp.function 0   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  0   I-pers.ind  0   I-comp.function 0   Q1069292    _   NoSpaceAfter
,   0   0   0   0   0   0   _   _   _
esl 0   0   0   0   0   0   _   _   _
arrive  0   0   0   0   0   0   _   _   NoSpaceAfter
'   0   0   0   0   0   0   _   _   _
à   0   0   0   0   0   0   _   _   _
Madrid  B-loc   0   B-loc.adm.town  0   0   0   Q2807   _   NoSpaceAfter
.   0   0   0   0   0   0   _   _   _
Aujourd 0   0   0   0   0   0   _   _   NoSpaceAfter
'   0   0   0   0   0   0   _   _   NoSpaceAfter
hui 0   0   0   0   0   0   _   _   NoSpaceAfter
,   0   0   0   0   0   0   _   _   EndOfLine
```


https://github.com/impresso/CLEF-HIPE-2020/tree/master/data

10.5281/zenodo.3706857

Auxiliary resources

In-domain Fr, De, and En embeddings:

- fastText word embeddings (with and w/o subwords)
- flair character embeddings (now integrated into the flair framework)

https://files.ifi.uzh.ch/cl/siclemat/impresso/clef-hipe-2020/
10.5281/zenodo.3706808

Evaluation

- Entities (not tokens) as the unit of reference
- Macro & **Micro** Precision, Recall and F1 measure
- Evaluation scenarios:

	NERC	**EL**
Strict	exact mention boundaries	consideration of the top link only, (overlapping mention boundaries)
Fuzzy	overlapping boundaries	historical mapping, cut-offs @3 and @5 (overlapping mention boundaries)

HIPE Scorer: https://github.com/impresso/CLEF-HIPE-2020-scorer

HIPE Eval Toolkit: https://github.com/impresso/CLEF-HIPE-2020-eval

Participation

40
registrations

75 runs
42% French
31% German
26% English
6 teams work on all languages

13
participating teams
All participated to NERC-Coarse
3 to NERC-Fine
5 to EL-only and end-to-end EL

11
Working Notes

Participating systems' main features

- 11 teams applied **neural approaches** for NERC;
- Most of them worked with **contextualized embeddings**, esp. **BERT**;
- Experimentation with **various input embeddings** (char, subword, word, historical or contemporary, type-level or contextualized)
- Some attempted to improve the **newspaper line-based input format** with proper sentence segmentation and tokenization;

Results overview (NERC)

- Neural system with strong embedding resource prevail;
- Performances correlates with amount of train/dev data;
- BERT-based systems > Bi-LSTM;
- Great performances diversity, but results are better than expected (6 teams > .8);
- NERC fine with 12 classes more difficult;
- NE components show reasonable performances.

F1 scores	French		German		English	
	Strict	Fuzzy	Strict	Fuzzy	Strict	Fuzzy
NERC-Coarse literal						
Baseline	.646	.769	.476	.585	.405	.562
Median	.677	.808	.636	.766	.463	.645
Best system	.840	.921	.797	.878	.632	.806
NERC-Coarse metonymic						
Best system	.783	.783	.634	.694	-	-
NERC-Fine						
Best system	.784	856	.668	.771	-	-
NE components						
Best system	.657	.751	.642	.707	-	-

Results overview (EL)

- EL performances are lower, and as diverse;
- NERC error propagation in end-to-end setting, but EL-only not a lot better;
- Performance increase with cut-offs @3 and @5.

F1 scores	French		German		English	
	Strict	Fuzzy	Strict	Fuzzy	Strict	Fuzzy
End-to-end Entity Linking (literal)						
Baseline	.257	.270	.180	.195	239	.239
Best system	.598	.617	.534	.557	.531	.531
End-to-end Entity Linking (metonymic)						
Best system	.297	.462	.396	.469	-	-
Entity linking only (with mentions provided)						
Baseline	.498	.512	.418	.437	.506	.506
Best system	.639	.659	.582	.602	.658	.658

Overall, what helps:

- BERT;
- actively tackling the problems of OCR noise, word hyphenation and sentence segmentation;
- in-domain resources.

Time-based observations

Analysis of F1 score as a function of time.

Hypothesis: the older, the more difficult.

Observation: no strong correlation between article publication date and performance.

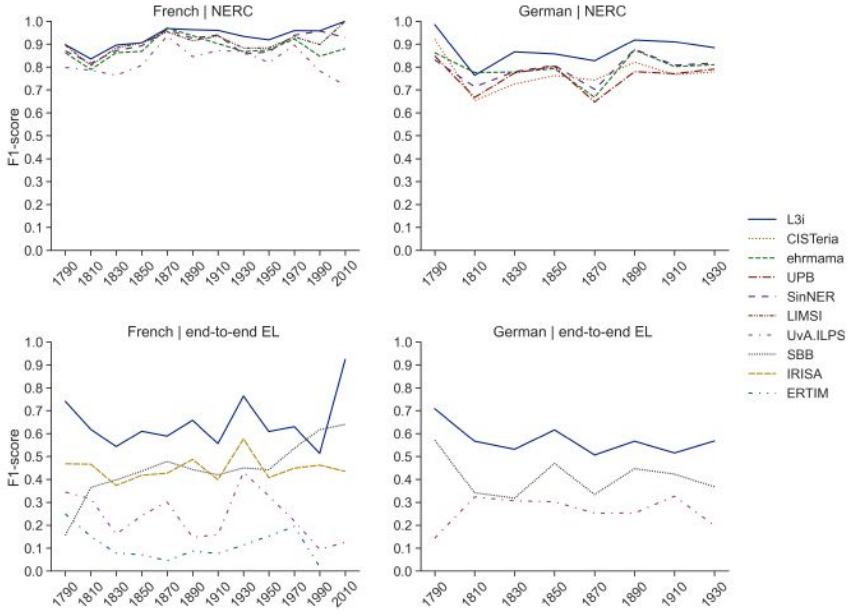
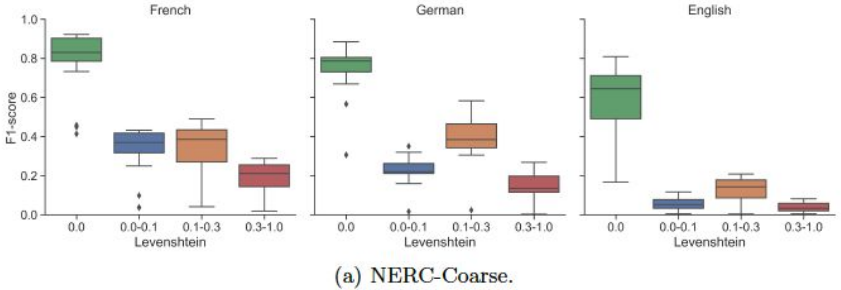


Fig. 4: F1 score as a function of time for the 5 best systems for NERC (top) and end-to-end EL (bottom) for the languages French (left) and German (right). The x-axis shows 20-years time buckets (e.g. 1790 = 1790-1809).

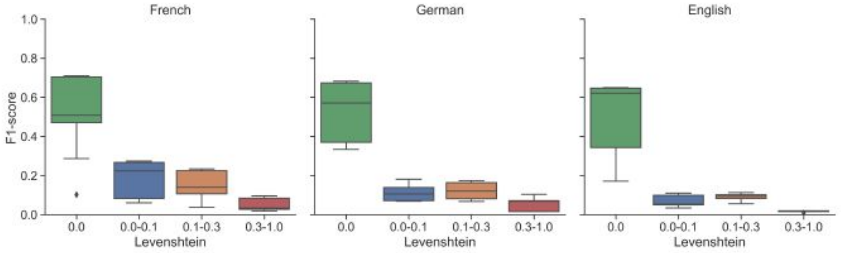
Impact of OCR noise

Evaluation on various noise levels

- noise: length-normalized Levenshtein distance between surface form and manual transcription;
- noisy vs non-noisy have remarkable differences on both NERC and EL;
- greatest performance variation at medium noise level



(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).

Conclusion (1/2)

- **Robustness test** for NERC and EL approaches on challenging historical material;
- New insights in **domain and language adaptation**;
- **Neural-based systems** with strong resources and proper segmentation are capable of dealing with historical and noisy inputs;
- EL, nested entities, entity components remain challenging;
- Performances are affected by OCR noise, but not by document publication date;

Discover more about systems this afternoon 3-6:30pm!

Conclusion (2/2)

Main Outcomes:

- Contribution to the advance of SoTA for NE processing on historical texts;
- Datasets;
- Scorer;
- Annotation guidelines;
- Step towards efficient semantic indexing of historical material.

Future Directions:

- Potential HIPE2 in 2021 with additional document types and languages.

Many thanks to

- CLEF organizers
- Participating teams (kudos!)
- *NZZ*, Le *Temps*, and the Swiss and Luxembourg national libraries
- Richard Eckart de Castillo, Clemens Neudecker, Sophie Rosset and David Smith
- INCEpTION project team
- Camille Watter, Gerold Schneider, Emmanuel Decker and Ilaria Comes
- SNSF (grant number CR-SII5 173719)

Thank you for your attention

HIPE : impresso.github.io/CLEF-HIPE-2020/

Scorer : impresso.github.io/CLEF-HIPE-2020/

Evaluation toolkit : impresso.github.io/CLEF-HIPE-2020/

Impresso project : impresso.github.io/CLEF-HIPE-2020/

@ImpressoProject