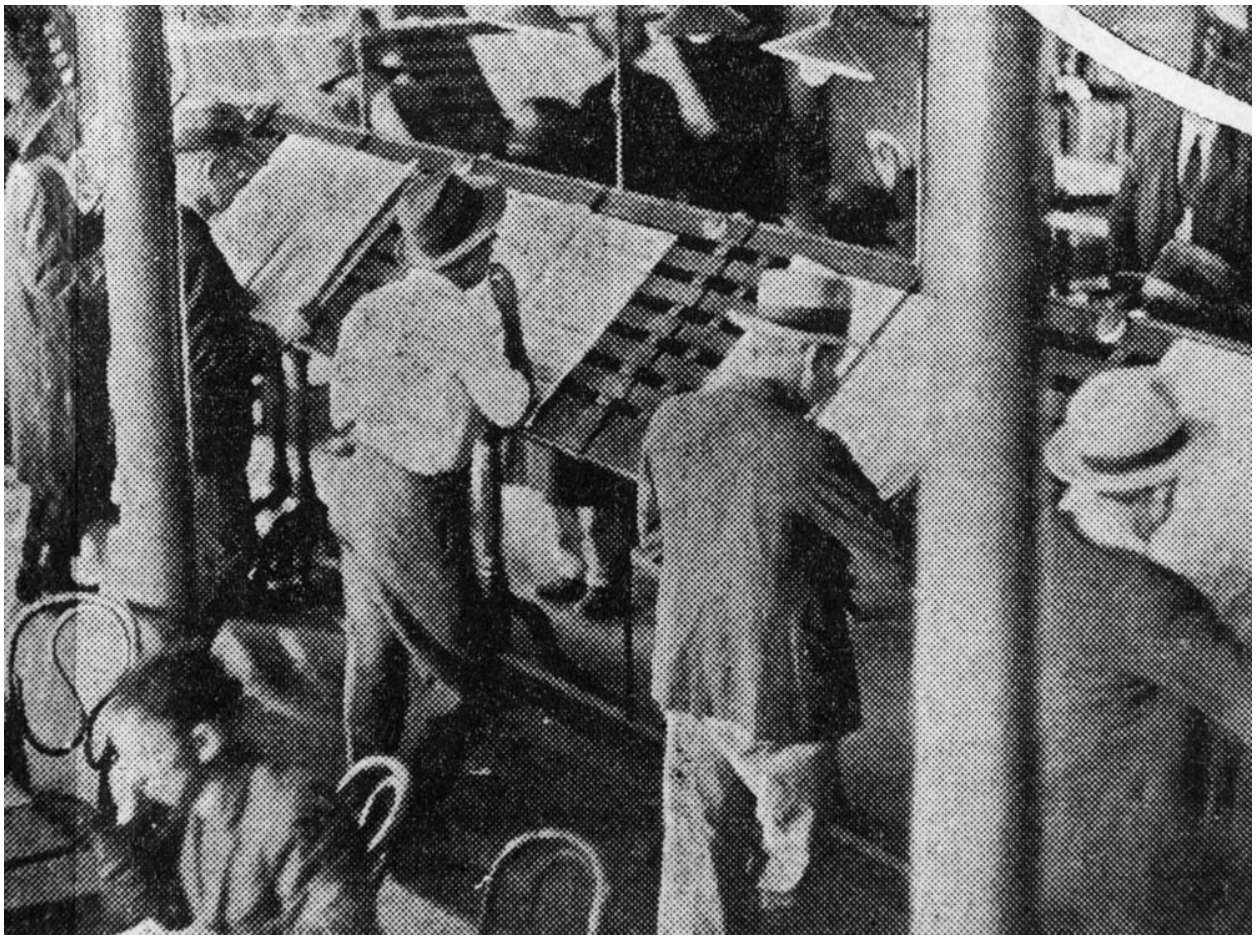
HIPE - Shared Task Participation Guidelines

Identifying Historical People, Places and other Entities



A CLEF 2020 Evaluation Lab organized by the *impresso* project.

20 February 2020 - Version: 1.1

Authors: *Maud Ehrmann, Matteo Romanello, Simon Clematide, Alex Flueckiger.*

CLEF 2020 Thessaloniki

impresso Media Monitoring of the Past

EDIT LOGS

20 February 2020 - version 1.1
(additions are in green font)

- new flag in MISC column
- modality of bundle 5: new possibility to run for pure NEL, with provided mentions (p.6, '*EL task settings*')
- additional specifications on evaluation metrics
- new EL setting (p6)

10 January 2020 - version 1.0
First version of participation guidelines.

HIPE 2020 PARTICIPATION GUIDELINES

1. INTRODUCTION

1.1 Motivation

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. If NE processing tools are increasingly being used in the context of historical documents, performances are below the ones on contemporary data and are hardly comparable.
In this context, the impresso project proposes the CLEF 2020 Evaluation Lab HIPE (Identifying Historical People, Places and other Entities) on named entity recognition and linking on diachronic historical newspaper material in French, German and English. The objective is this shared task is threefold:

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 fostering efficient semantic indexing of historical documents in order to support scholarship on digital cultural heritage collections

1.2 Overview

The HIPE shared task consists of two main named entity processing tasks, namely mention recognition and classification (with two levels of difficulty), and entity linking. The shared task corpora are composed of newspaper articles sampled among several Swiss, Luxembourgish and American historical newspapers on a diachronic basis. Registered teams can participate in some or all of the tasks.

1.3 Organisation

HIPE is a CLEF 2020 Evaluation Lab organized by the '*impresso* - Media Monitoring of the Past' project[1]. *Impresso* is an interdisciplinary research project involving a team of computational linguists, designers and historians who collaborate on the semantic indexing of a large-scale, multilingual corpus of digitized historical newspapers. The project is supported by the Swiss National Science Foundation (SNSF) under grant number CR-SII5_173719.

1.4 Information

For contact, registration, data download, important dates and updates, please refer to the HIPE website: https://impresso.github.io/CLEF-HIPE-2020/

[1] https://impresso-project.ch

2. TASKS

2.1 Task 1 - Named entity recognition and classification (NERC)

Overview

Subtask 1.1 - NERC Coarse-grained: this task includes the recognition and classification of entity mentions according to coarse-grained types (cf. column 1 in Table 2).

Subtask 1.2 - NERC Fine-grained: this task includes the recognition and classification of entity mentions according to fine-grained types (cf. column 2 in Table 2), plus the detection and classification of nested entities of depth 1, as well as entity mention components (title, function, etc.).

Table 1 summarizes which annotation types systems are expected to produce for Task 1.

	Task 1.1 'Coarse'	**Task 1.2 'Fine'**
NE mentions with coarse types	yes	yes
NE mentions with fine types	no	yes
Consideration of metonymic sense	yes	yes
NE components	no	yes
Nested entities of depth 1	no	yes

Table 1. Expected annotation types for Task 1.

NERC system annotation guidelines

Table 2 lists the entity types to consider for Task 1 (NERC). For more information about system annotation rules, please refer to Appendix A.

2.1 Task 2 Entity linking (EL)

Overview

This task corresponds to the linking of named entity mentions to unique referents in a knowledge base (KB), or to a NIL node if the mention does not have a referent in the KB. The

chosen KB is Wikidata, and the present task uses a frozen dump of 13 November 2019[2] (latest-all.nt.bz2).

Coarse-grained tag set	Fine-grained tag set	Metonymy applies	Entity nesting applies	Linking applies
pers	pers.ind pers.coll pers.ind.articleauthor	yes	yes	yes
org	org.adm org.ent org.ent.pressagency	yes	yes	yes
prod	prod.media prod.doctr	yes	no	yes
time	time.date.abs	no	no	no
loc	loc.adm.town loc.adm.reg loc.adm.nat loc.adm.sup	yes	yes	yes
	loc.phys.geo loc.phys.hydro loc.phys.astro	yes	yes	yes
	loc.oro	yes	yes	yes
	loc.fac	yes	yes	yes
	loc.add.phys loc.add.elec	yes	yes	yes
	loc.unk	no	no	no

Table 2. Entity types to annotate (Task 1) and link (Task 2)

EL system annotation guidelines

1. Systems are required to link mentions of types PERS, ORG, PROD and LOC, by giving their corresponding wikidata id (the 'Q' id);
2. In case the referred entity does not exist in the knowledge base, systems should indicate 'NIL'; please note that it is not allowed to annotate with wikipedia disambiguation pages.
3. Entity links must be set with respect to both literal and metonymic (when present) mention annotations. Both classification/recognition and linking of metonymic mentions are difficult tasks for humans and machines alike. Since metonymy is not the main focus of HIPE, metonymy linking is seen as a 'nice-to-have' for participant systems, and there will be flexible evaluation scenarios. For specific rules about metonymy linking, please refer to Appendix A.

[2] https://files.ifi.uzh.ch/cl/siclemat/impresso/clef-hipe-2020/wikidata-2019-11-13.nt.bz2

4. Entity components and nested entities are excluded from linking.

EL task settings

The entity linking task includes two settings: with and without prior knowledge of mention boundaries. Concretely speaking, the evaluation period will consist of two consecutive, separated rounds, where a first NEL task without prior information on mentions will be evaluated during round 1 (i.e. bundles 1, 2, see Table 4 in section 5.1), and a second one with information on mention boundaries (but no named entity type information) during the second round (bundle 5).

3. DATA

3.1 Terminology

Content item (sometimes abbreviated to 'item') refers to newspaper segments below the page level; examples include advertisements, images, tables, weather forecasts, obituaries, etc. In the present shared task, content items correspond to newspaper *articles* only, and both terms are used interchangeably.

Time bucket in HIPE corresponds to a decade. Selected items in a time bucket always belong to the first year of the bucket.

3.2 Corpus

Evaluation corpora are composed of articles sampled among several Swiss, Luxembourgish and American historical newspapers on a diachronic basis.

A. Corpus selection - The corpus was composed based on systematic and purposive sampling. For each newspaper, articles were randomly sampled among articles that 1) belong to the first years of a set of predefined decades covering the life-span of the newspaper, and 2) have a title, have more than 50 characters, and belong to any page (no restriction to front pages only). 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.

B. Corpus characteristics - OCR quality corresponds to real-life setting, *i.e.* it varies according to digitization time and archival material. We do not provide different OCR versions of the same texts, but will provide an OCR quality assessment measure alongside each article, as well as links towards line segment images. Corpus and annotation statistics will be published with the full data release in February 2020. The time-span of the whole corpus goes from 1798 until 2018. Table 3 gives a first overview of the corpus.

C. Corpus annotation - The corpus has been manually annotated by native speakers using the INCEpTION annotation platform [1] and according to HIPE annotation guidelines. Before

annotating, collaborators are first trained on a 'mini-reference' corpus – consisting of 10 content items per language – in order to ensure their understanding of the guidelines, and to check their inter-annotator agreement (IAA) with the mini-ref corpus. Some items of the test set will be double-annotated and adjudicated, as well as randomly sampled items among the training set (train) and the development set (dev).

Anonymized title	Language	Digitized time span	HIPE time span	Nb of time buckets
CH newspaper 1	fr	1798-1999	1798-1988	19
CH newspaper 2	fr	1881-2019	1888-2018	13
CH newspaper 3	fr	1738-2018	1738-2008	27
LU newspaper 1	fr	1871-1934	1878-1928	6
CH newspaper 4	de	1780-2019	1798-1950	16
LU newspaper 3	de	1848-1950	1848-1948	11
LU newspaper 4	de	1913-1950	1918-1948	4
CH newspaper 5	de	1846-1891	1848-1888	5
ca. 20 US newspapers	en	info not yet available	info not yet available	info not yet available

Table 3. Broad overview of the corpus.

3.3 Data sets

For each task, the following data sets will be released:

	French (fr)	German (de)	English (en)
sample	January 2019	January 2019	-
train	14.02.2020	14.02.2020	-
dev	14.02.2020	14.02.2020	14.02.2020
test	after the evaluation	after the evaluation	after the evaluation

These data sets can only be used for personal and/or academic purposes. Data dumps contain terms of use statements to which each participant is liable for.

3.4 Formats

Data is released in IOB format (inside-outside-beginning format)[3], in a similar fashion to that of

[3] https://en.wikipedia.org/wiki/Inside–outside–beginning_(tagging)

the CoNLL-U format[4].

A. Directory Structure

Sample, training and development data sets consist of UTF-8, tab-separated-values files. These files contain information needed for all tasks (NERC coarse, NERC fine, and linking) for each language. In concrete terms, there is one .tsv file per language and data set split. For example, the file `HIPE-data-v01-dev-de.tsv' contains all content items of the German part of the corpus which are meant as development set. The Figure below gives an overview of the data release folder structure:

```
|--data/
|    |--release-v01/
|    |   |--de/
|    |   |    `--HIPE-data-v01-dev-de.tsv
|    |   |    `--HIPE-data-v01-sample-de.tsv
|    |   |    `--HIPE-data-v01-train-de.tsv
|    |   |--en/
|    |   |    `--HIPE-data-v01-dev-en.tsv
|    |   |--fr/
|    |   |    `--HIPE-data-v01-dev-fr.tsv
|    |   |    `--HIPE-data-v01-sample-fr.tsv
|    |   |    `--HIPE-data-v01-train-fr.tsv
```

Data release folder structure.

B. File contents

As mentioned, files encode annotations needed for all tasks (NERC coarse, NERC fine and linking) and contain the following information:

- Empty lines, which mark the boundaries between content items;
- Comment lines, which give further information and start with the character '#';
- Annotated lines, which contain a token followed by tab-separated annotations.

A file contains all the content items of one language. Content items are separated with empty lines and are preceded with the following comment lines:

  - `#language`: the language of the content item: de, en, fr;
  - `#newspaper`: the newspaper id (an acronym) the content item comes from;
  - `#date`: the publication date of the content item (YYYY-MM-DD):
  - `#document_id`: the document id, composed of: the newspaper id, the date, the character *p* (page) or *i* (item) indicating the legacy type of the item, and the 4-zero-padded id number. Example: NZZ-1798-01-17-a-p0002.

Beside content item divisions, we also indicate original layout line breaks as they are printed in the newspaper. A comment line containing the IIIF link to the image of the line segment is inserted: `#segment_iif_link`. These IIIF comment lines can be used to retrieve the image of the line, or deleted (the line break information is also present in the MISC column).

[4] https://universaldependencies.org/format.html

Each line consists of 9 columns:

1. TOKEN: the annotated token.
2. NE-COARSE-LIT: the coarse type (IOB-type) of the entity mention token, according to the literal sense.
3. NE-COARSE-METO: the coarse type (IOB-type) of the entity mention token, according to the metonymic sense.
4. NE-FINE-LIT: the fine-grained type (IOB-type.subtype.subtype) of the entity mention token, according to the literal sense.
5. NE-FINE-METO: the fine-grained type (IOB-type.subtype.subtype) of the entity mention token, according to the metonymic sense.
6. NE-FINE-COMP: the component type of the entity mention token.
7. NE-NESTED: the coarse type of the nested entity (if any).
8. NEL-LIT: the Wikidata Q id of the literal sense, or `NIL'.
9. NEL-METO: the Wikidata Q id of the metonymic sense, or `NIL'.
10. MISC: a flag which can take the following values:
    - `NoSpaceAfter', to indicate the absence of white space after the token.
    - `EndOfLine', to indicate the end of a layout line.
    - `Partial-START:END', to indicate the character on/offsets of mentions that do not cover the full token (esp. for German compounds).

Non-specified values are marked by the underscore character `_'.

Figure 1 hereafter shows an example of an annotated IOB file for the French sentence "*L'emper * ur de Russie a quitté Varsovie le} o avril et est parti pour la Crimée.*" ("The Emperor of Russia left Warsaw on April x and left for Crimea").

```
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 = 1818-05-22
# document_id = GDL-1818-05-22-a-i0007
# segment_iiif_link = _
Nouvelles O O O O O O _ _ _
diverses  O O O O O O _ _ NoSpaceAfter
. O O O O O O _ _ EndOfLine
# segment_iiif_link = _
L'  O O O O O O _ _ NoSpaceAfter
emper O O O O O O _ _ _
* O O O O O O _ _ _
ur  O O O O O O _ _ _
de  O O O O O O _ _ _
Russie  B-loc O B-loc.adm.nat O O O Q159  _ _
a O O O O O O _ _ _
quitté  O O O O O O _ _ _
Varsovie  B-loc O B-loc.adm.town  O O O Q270  _ _
le  O O O O O O _ _ NoSpaceAfter
} O O O O O O _ _ _
o O O O O O O _ _ _
avril O O O O O O _ _ _
et  O O O O O O _ _ _
est O O O O O O _ _ _
parti O O O O O O _ _ EndOfLine
# segment_iiif_link = _
pour  O O O O O O _ _ _
la  O O O O O O _ _ _
Crimée  B-loc O B-loc.adm.reg O O O Q7835 _ NoSpaceAfter
....
```

Figure 1. Example of HIPE IOB file for French

Figure 2 shows a similar example with metonymic annotations for the French sentence "*H. C. Lausanne - H. C. Chaux-de-Fonds Lausanne et La Chaux-de-Fonds, deux équipes romandes qui peinent….*" ("H. C. Lausanne - H. C. Chaux-de-Fonds Lausanne and La Chaux-de-Fonds, two teams from French-speaking Switzerland who are struggling…")
"H. C." stands for Hockey Club. Here the phrase *H. C. Lausanne* is annotated as an organisation (literal sense), with a link towards the corresponding Wikidata entry Q675245. The nested entity *Lausanne* is annotated as a location, and is not linked since entity linking is not required for nested entities (see annotation guidelines and Appendix hereafter). The next occurrence of *Lausanne* is metonymic, and annotated accordingly, *i.e.* with different types and different Wikidata ids.

Finally, Figure 3 shows similar examples for German.

About tokenization:
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. Participants can choose to apply their own sentence splitting and tokenization. The tokenization applied to produce the IOB files is based on simple white space splitting, leaving all punctuation signs (including apostrophes) as separate tokens. The flag 'NoSpaceAfter' provides information about how to reconstruct the text.

About nested entities:
Annotations step from the outer to the innermost entity. NE-COARSE and NE-FINE correspond to the outermost entity mentions, and NE-NESTED corresponds to the first nested level. Only one level of nested entities is to be annotated. Please refer to system annotation guidelines in Appendix A.

D. Additional resources

The HIPE Evaluation lab provides additional lexical resources, please refer to the HIPE website.

E. *Visualization interface*

HIPE is supported by the *impresso* project, which develops an exploration interface for newspapers. All French and German training material from HIPE is visible through this interface, which can be helpful to see the image and text of an article. Full access to the material is provided upon acceptance of terms of use, which can be asked following the information online.

Interface URL: https://impresso-project.ch/app
Shorthand to visualize an article:
https://impresso-project.ch/app/article/IMP-1958-01-25-a-i0161 (then change the article ID).

```
TOKEN NE-COARSE-LIT NE-COARSE-METO  NE-FINE-LIT NE-FINE-METO  NE-FINE-COMP  NE-NESTED NEL-
# language = fr
# newspaper = IMP
# date = 1958-01-25
# document_id = IMP-1958-01-25-a-i0161
# segment_iiif_link = _
ÇHOCKEY O O O O O O _ _ _
SUR O O O O O O _ _ _
GLACE O O O O O O _ _ _
J O O O O O O _ _ EndOfLine
# segment_iiif_link = _
Derby O O O O O O _ _ _
romand  O O O O O O O _ _ EndOfLine
# segment_iiif_link = _
H B-org O B-org.ent O O O Q675245 _ NoSpaceAfter
. I-org O I-org.ent O O O Q675245 _ _
C I-org O I-org.ent O O O Q675245 _ NoSpaceAfter
. I-org O I-org.ent O O O Q675245 _ _
Lausanne  I-org O I-org.ent O O B-loc.adm.town  Q675245 _ NoSpaceAfter
- O O O O O O _ _ EndOfLine|NoSpaceAfter
# segment_iiif_link = _
H B-org O B-org.ent O O O Q680502 _ NoSpaceAfter
. I-org O I-org.ent O O O Q680502 _ _
C I-org O I-org.ent O O O Q680502 _ NoSpaceAfter
. I-org O I-org.ent O O O Q680502 _ _
Chaux-de-Fonds  I-org O I-org.ent O O B-loc.adm.town  Q680502 _ EndOfLine
# segment_iiif_link = _
Lausanne  B-loc B-org B-loc.adm.town  B-org.ent O O Q807  Q675245 _
et  O O O O O O O _ _ _
La  B-loc B-org B-loc.adm.town  B-org.ent O O Q68124  Q680502 _
Chaux-de-Fonds  I-loc I-org I-loc.adm.town  I-org.ent O O Q68124  Q680502 NoSpaceAfter
, O O O O O O _ _ EndOfLine
# segment_iiif_link = _
deux  O O O O O O _ _ _
équipes O O O O O O _ _ _
romandes  O O O O O O O _ _ _
qui O O O O O O _ _ _
peinent O O O O O O O _ _ EndOfLine
```

Figure 2. Example of metonymy annotations for French.

4. EVALUATION

NERC is evaluated in terms of macro and micro Precision, Recall, F1-measure. Two evaluation scenarios are considered: strict (exact boundary matching) and relaxed (fuzzy boundary matching).

Each column is evaluated independently, according to the following metrics:

- **Micro average P, R, F1** at entity level (not at token level), i.e. consideration of all true positives, false positives, true negatives and false negative over all documents.
    - strict (exact boundary matching) and fuzzy (at least 1 token overlap).

    - separately per type and cumulative for all types.
- **Document-level macro average** P, R, F1 at entity level (not on token level). i.e. average of separate micro evaluation on each individual document.
    - strict and fuzzy
    - separately per type and cumulative for all types

Our definition of macro differs from the usual one, and macro measures are computed as aggregates on document-level instead of entity-type level. Specifically, macro measures average the corresponding micro scores across all the documents, accounting for (historical) variance in document length and not for class imbalances.

Note that in the strict scenario, predicting wrong boundaries leads to severe punishment of one false negative (entity present in the gold standard but not predicted by the system) and one false positive (predicted entity by the system but not present in the gold standard). Although this may be severe, we keep this metric in line with CoNLL and refer to the fuzzy scenario if the boundaries of an entity are considered as less important.

The Slot Error Rate (SER) is dropped for the shared task evaluation.

The evaluation for NEL works similarly as for NERC. The link of an entity is interpreted as a label. As there is no IOB-tagging, a consecutive row of identical links is considered as a single entity. In terms of boundaries, NEL is only evaluated according to the fuzzy scenario. Thus, to get counted as correct, the system response needs only one overlapping link label with the gold standard.

With respect to the linking of metonymic mentions, two evaluation scenarios will be considered: strict, where only the metonymic link will be taken into account, and relaxed, where the union of literal and metonymic annotations will be taken into account. This is not implemented yet in the scorer, it will be done with the next release.

```
Das O O O O O O _ _ _
aus O O O O O O _ _ _
3 O O O O O O _ _ _
Infanterie  O O O O O O _ _ NoSpaceAfter
- O O O O O O _ _ _
und O O O O O O _ _ _
2 O O O O O O _ _ _
Kavalle O O O O O O _ _ NoSpaceAfter
¬ O O O O O O _ _ EndOfLine|NoSpaceAfter
# segment_iiif_link = _
rie-Regimentern O O O O O O _ _ _
bestehende  O O O O O O _ _ _
Corps O O O O O O _ _ _
des O O O O O O _ _ _
Prinzen B-pers  O B-pers.ind  O B-comp.function O Q430775 _ _
von I-pers  O I-pers.ind  O B-comp.name O Q430775 _ EndOfLine|NoSpaceAfter
# segment_iiif_link = _
Condé I-pers  O I-pers.ind  O I-comp.name O Q430775 _ NoSpaceAfter
, O O O O O O _ _ _
welches O O O O O O _ _ _
in  O O O O O O _ _ _
kaiserliche O O O O O O _ _ _
Dienste O O O O O O _ _ _
genommen  O O O O O O _ _ EndOfLine|NoSpaceAfter
# segment_iiif_link = _
worden  O O O O O O _ _ NoSpaceAfter
, O O O O O O _ _ _
ist O O O O O O _ _ _
nun O O O O O O _ _ _
nach  O O O O O O _ _ _
Wladimir  B-loc O B-loc.adm.town  O O B-loc.adm.town  NIL _ NoSpaceAfter
, O O O O O O _ _ _
Luzk  B-loc O B-loc.adm.town  O O O Q7550 _ _
und O O O O O O _ _ _
Kowel B-loc O B-loc.adm.town  O O O Q156704 _ EndOfLine|NoSpaceAfter
```

Figure 3. Example of HIPE IOB file for German.

5. SYSTEM RESPONSES

5.1 General rules

- Registration is open until 26 April 2020. Please refer to the HIPE website for more information.
- Teams can participate in one task bundle per language, as listed in Table 4 hereafter.
- Teams should submit at least one, and up to three runs for their chosen task bundle.
- Teams can use any external resources (e.g. additional language resources provided by HIPE, available elsewhere or homemade, and other annotated data).
- Teams are highly encouraged to share the additional resources they use, either during or after the evaluation.

Bundle id	Associated tasks	Relevant columns in the IOB response file
bundle1	NERC-coarse and NERC-fine and NEL	TOKEN, NE-COARSE, NE-FINE1, NE-FINE1-COMP, NE-FINE2, NEL
bundle2	NERC-coarse and NEL	TOKEN, NE-COARSE, NEL
bundle3	NERC-coarse and NERC-fine	TOKEN, NE-COARSE, NE-FINE1, NE-FINE1-COMP, NE-FINE2
bundle4	NERC-coarse	TOKEN, NE-COARSE
bundle5	NEL	TOKEN, NEL

Table 4. List of task bundles that a system can participate in.

5.2 Evaluation period

Please check important dates on HIPE website.
At the end of each evaluation period, participants will send their system responses via email to maud.ehrmann@epfl.ch, which will be evaluated by the HIPE team using the scorers. Gold Standard data will be distributed after the publication of the evaluation results.

5.3 System response guidelines

Input test data will consist of diachronic historical newspaper content items in French, German and English. Data will be encoded similarly as the train and dev data, but without the annotations: one token per line, with each content item layout line separated with a blank line.

Rules for system response files

- files must be in UTF-8, tsv encoded (.tsv extension), with annotations in IOB format.
- files need to contain all content item lines and empty lines in the order of the original input file.
- files must comply with the following naming convention:
  TEAMNAME_TASKBUNDLEID_LANG_RUNNUMBER.tsv
  where:
  TEAMNAME: is the name of the team such as registered via the CLEF portal
  TASKBUNDLEID: is one of the bundle ids as indicated in Table 4.
  LANG: is de,fr,en
  RUNNUMBER: is 1,2,3
  Example: dreamteam_bundle1_de_2.tsv

```
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 = 1818-05-22
# document_id = GDL-1818-05-22-a-i0007
# segment_iiif_link = _
Nouvelles _ _ _ _ _ _ _ _ _ _
diverses  _ _ _ _ _ _ _ _ _ NoSpaceAfter
. O O O O O _ EndOfLine
# segment_iiif_link = _
L _ _ _ _ _ _ _ _ _ NoSpaceAfter
' _ _ _ _ _ _ _ _ _ NoSpaceAfter
emper _ _ _ _ _ _ _ _ _ _
* _ _ _ _ _ _ _ _ _ _
ur  _ _ _ _ _ _ _ _ _ _
de  _ _ _ _ _ _ _ _ _ _
Russie  _ _ _ _ _ _ _ _ _ _
a _ _ _ _ _ _ _ _ _ _
quitté  _ _ _ _ _ _ _ _ _ _
Varsovie  _ _ _ _ _ _ _ _ _ _
le  _ _ _ _ _ _ _ _ _ NoSpaceAfter
} _ _ _ _ _ _ _ _ _ _
o _ _ _ _ _ _ _ _ _ _
avril _ _ _ _ _ _ _ _ _ _
et  _ _ _ _ _ _ _ _ _ _
est _ _ _ _ _ _ _ _ _ _
parti _ _ _ _ _ _ _ _ _ EndOfLine
# segment_iiif_link = _
pour  _ _ _ _ _ _ _ _ _ _
la  _ _ _ _ _ _ _ _ _ _
Crimée  _ _ _ _ _ _ _ _ _ NoSpaceAfter
...
```

Figure 4. Example for input test file.

- files must include all columns and instantiate the unspecified values in the required columns according to the chosen task bundle.
- at present files must include the document (#document_id comment) and line separator (#segment_iiif_link comment) information. They may include the other comment lines. The evaluation script will be updated so that no comment line is required at all.

System response file submission - System response files must be sent in a .zip archive via email to the task organizers (maud.ehrmann@epfl.ch) by the submission deadline indicated on the HIPE website. An acknowledgement of receipt will be sent upon reception.

6. WORKSHOP and WORKING NOTE PAPER

Participants will submit a 'Working Note' paper to be presented during the final workshop co-located with the CLEF conference (September 2020) and to be published online via the CEUR Workshop Proceeding open access publication service. Please check submission instructions and important dates on the HIPE website.

7. REFERENCES

[1] J.-C. Klie, M. Bugert, B. Boullosa, R. E. de Castilho, and I. Gurevych. 2018 *The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation* in Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, 2018, pp. 5–9.

[2] N. S. Moosavi and M. Strube (2016) *Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric* in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 632–642.

APPENDIX A - NERC System Annotation Guidelines.

We give hereafter the main annotation rules to consider while designing a NERC system. Please note that this is a summary of the annotation guide. For more information one should refer to the Impresso Annotation Guidelines.

A.1 Entity types and components

We reproduce Table 2 about the entity types to consider. An exact definition of each type is given in the annotator guidelines, and a brief one is given in Table 5 hereafter.

Coarse-grained tag set	Fine-grained tag set	Metonymy applies	Entity nesting applies	Linking applies
pers	pers.ind pers.coll pers.ind.articleauthor	yes	yes	yes
org	org.adm org.ent org.ent.pressagency	yes	yes	yes
prod	prod.media prod.doctr	yes	no	yes
time	time.date.abs		no	no
loc	loc.adm.town loc.adm.reg loc.adm.nat loc.adm.sup	yes	yes	yes
	loc.phys.geo loc.phys.hydro loc.phys.astro	yes	yes	yes
	loc.oro	yes	yes	yes
	loc.fac	yes	yes	yes
	loc.add.phys loc.add.elec	yes	yes	yes
	loc.unk	no	no	no

A.2 Lexical characteristics

Linguistic units considered as named entities must include a proper name, or a definite description having the status of a proper name, i.e. definite descriptions with a nominative function and a certain referential stability (see section 2.2.A on p. 3 of annotation guidelines).

Phrases such as

- *Die präkolumbianische Zivilisation, la civilisation précolombienne*
- *l'armée bavaroise*
- *les forces tchadiennes*
- *le gouvernement français*

are *not* annotated because they do not contain proper names.

Phrases such as *le gouvernement Franco* are annotated:

```
le <org.adm> gouvernement
      <comp.name> <pers.ind> Franco </pers.ind> </comp.name>
</org.adm>
```

A.3 Named entity boundaries

Each token is either completely part of a named entity or not at all. Named entity mentions exclude subordinate clauses, incidental clauses and determiners. They include pre- and post-modifiers (see section 2.2.B of annotation guidelines).

A.4 About very noisy OCR entities

Such entities were annotated including the garbage characters which the annotator – while looking at the article facsimile – thought they should be part of the mention.

A.5 About nested entities

System should annotate nested entities of depth one only.

A.6 About metonymy

NERC annotation: When it applies, entities of type PERS, ORG, LOC and PROD are annotated according to their metonymic sense in both coarse and fine NERC settings.

NE Linking: Please refer to Section 4 of annotation guidelines.

A.7 About coordinated entities

Refer to Section 2.4.B of the annotation guidelines p6.

A.8 About components

Components are to be annotated for Task 1.2 (fine-grained) are the following:
For the type PERS:

- `comp.func`
- `comp.title`
- `comp.name`
- `comp.qualifier`
- `comp.demonym`

For all other types, except DATE

- `name`, used to mark the name of the entity.

The component `name` is optional when the mention contains only one name.

A.9 Quick guide (also present in the annotation guidelines)

Entity types and subtypes	
`pers.ind`	A single person (*Roger Federer*)
`pers.ind.articleauthor`	A single person who is the author of an article.
`pers.coll`	A named group of people including musical groups (die *Beatles*, La *Mano Negra*). (note: die *Schweizer*, Les *français* are **not** annotated.)
`org.ent`	Organization that markets products or provides services (Die *Peugeot Gesellschaft,* Die *Waid;* La *société Peugeot*, la *Pitié-Salpêtrière*). (note: Die *schweizer* Polizei; la police francaise ist **not** annotated)
`org.ent.pressagency`	Special type related to newspaper to spot press agencies.
`org.adm`	Organization that plays a mainly administrative role (Die *Stadtverwaltung Bern*; la *mairie de Paris*). (note: Das *Departement für auswärtige Angelegenheiten*; Le *Ministère des Affaires Étrangères* is **not** annotated)
`loc.adm.town`	District, locality, hamlet, village, city, etc. (*Paris*, *Val de Crüye*).
`loc.adm.reg`	Cantons, communities of municipalities, departments, regions, etc. (*Autonome Gemeinschaft Baskenland;* les *Bouches du Rhône*, Le *Pays-Basque espagnol*).
`loc.adm.nat`	Countries (*Schweiz*; *France*).
`loc.adm.sup`	World regions, continent (*Maghreb; Pays-Basque*).
`loc.phys.geo`	Mountains, plains, plateaus, caves, volcanoes, canyons (Die *Alpen,* Der *Vesuv; gouffre de Padirac*, Le *mont Ventoux*).
`loc.phys.hydro`	Oceans, seas, rivers, streams, ponds, marshes (Der *Atlantik,* Der *Golfstrom;* La *Seine*, Le *Lac Paladru*).
`loc.phys.astro`	Planets, stars, galaxies and their parts (Der *Mond*, Die *Milchstrasse;* La *terre,* la *mer de la Tranquillité*).

`loc.oro`	Refers to roads, highways, streets, avenues, squares, etc. (Die *Autobahn A6;* L'*autoroute A6*).
`loc.fac`	Refers to the buildings (Der *Prime Tower;* Le *Palais de l'Élysée*).
`loc.add.phys`	Refers to physical addresses (*LIMSI-CNRS*, *Bâtiment 508, BP133, 91403 Orsay Cedex*).
`loc.add.elec`	Refers to electronic contact information (telephone and fax numbers, URL, e-mail address, identification of social network or Internet communication tools, etc., *http://www.limsi.fr/, 01-69-85-80-00*)
`Loc.unk`	Type used when it is not possible to choose among other location types.
`prod.media`	Newspapers, magazines, broadcasts, sales catalogues, etc. (*Die Zeit;* Le *Figaro*, *Le sept à huit, La ferme célébrités*).
`prod.doctr`	Political, philosophical, religious, sectarian doctrines. (Der *Sozialismus, Theravada Buddhismus; Zeugen Jehovas;* Le *socialism*, le *bouddhisme theravâda*,le *structuralism*, la *scientology*).
`time.date.abs`	An absolute date (*Sonntag der 13. November 2016; lundi 25 janvier 2010*)
Component	
`name`	is the only transversal component and is applied to any class except `time`. (Die *Peugeot* Gesellschaft; la société *Peugeot*)
`comp.name`	The component includes first, middle and last names as well as nickname and initials of individuals (*Samuel L. Jackson, S.L.J.*)
`comp.title`	Title or designator of a person. (*Herr* Chirac, *Ihre Hoheit* Rainier*; M.* Chirac, *Son Altesse* le prince Rainier).
`comp.qualifier`	A qualifier specifies a person in the form of a qualifying adjective. (Der *konservative* Christoph Blocher; le *socialiste* Bertrand Delanoë)
`comp.function`	A function or job of a named person. (*Bürgermeister* Ann Hidalgo von Paris; *maire* de Paris Anne Hidalgo).
`comp.demonym`	The geographical origin of a person (Le *français* Alain Vigneron).