Notes on NER@NLP for medievalists school

4 blocks of 90 mins

1. Block:
   1. general intro 15 mins
   2. WHAT 60 mins
      1. theoretical part (definitions of NE, categories, nested vs. flat, graph-based)
      2. Group discussion on annotation guidelines of the corpora (how is a NE defined? what classes? nested? boundaries? is the goal of the NE-Annotation formulated?)
         1. Medieval Charters (fr, la, sp)
         2. Herodot corpus (la)
         3. BeNASch (de) (<https://zenodo.org/records/11500543>)
         4. TODO SELINA (arabic? otherwise maybe the dutch one or the other german one)
            1. Suggestion: see corpora → CANER? (Classical Arabic)
   3. Showcase: MOM-NER (Sandy) 15 mins
2. Block:
   1. WHY (Sandy) 30 mins
   2. What annotation toolkit exists? (Sandy) => Transkribus, Inception (Doccano, Prodigy) (10 mins)
   3. How to create Data (Hands-On) (35 mins) (Sandy) → Ismail will prepare it (probably on the train ride)
      1. Think of annotation you’re interest in
      2. Getting familiar with toolkit + Annotate
   4. Showcase: Active Learning + NER + French (Selina) (15 min)

Lunch Break

1. Block (early afternoon) (Ismail)
   1. Showcase: HGB and Benasch (30 min)
   2. HOW (45 mins)
      1. including historical NER challenges
   3. start explaining hands-on (15 min)
2. Block (middle afternoon)
   1. incl. hands on (45 min)
   2. application to own data (45 min)

**Plan:**

* General Introduction (Ismail)
  + bit of history? => rule-based, gazetteer-based, unsupervised
* WHAT: What's a Named Entity? (Selina)
  + Assumptions:
    - flat vs nested
    - categories
    - all vs specific
  + graph-based
  + group discussion
* WHY: Why get Named Entities? (Sandy)
  + Index-building
  + NE-Linking
  + Relation- & Event-Extraction => information extraction in general, network analysis
  + => Target of NER!
* HOW: What's Named Entity Recognition? (Ismail)
  + Specific challenges in historical ner
  + Supervised Sequence-to-Sequence (Flair, Spacy, Transformer)
  + By LLMs (Text-to-Text; QA)
  + Advantages / Disadvantages?
  + Evaluation of Results?
    - More than just Recall & Precision!
* Group Discussion (include into the theoretical part)
  + Different historical datasets
    - how is a NE defined?
    - what classes?
    - nested? boundaries?
    - is the goal of the NE-Annotation formulated?
  + Format e.g. “[World Cafe](https://en.wikipedia.org/wiki/World_caf%C3%A9_(conversation))”, “Re-Shuffle group members”
    - 6, 6, 5, 5
  + Which corpora? **ToDo**
    - at least 2 latin corpora
    - 1 Historical German
    - 1 Historical French
    - 1 Ottoman Turkish, Arabic, Hungarian?

? <https://kitab-project.org/> , <https://corpus.quran.com/> ?

* Showcases (include partially into Why)
  + Historical Land Records of Basel - Nested Named Entity Rec with Flair
  + Graz Projects?
    - MOM-NER (Sandy)
    - French? (Selina)
  + Invite someone from the audience (to be checked who could be asked)
* Hands-On (All in Google Colab?)
  + How to create data?
    - Writing annotation guidelines or adopting others?
    - What frameworks exist? => Transkribus, Inception (Doccano, Prodigy), TEI (TEI-Publisher)?
  + use an established framework with pretrained models with example data in **Flair** / Spacy
    - (BERT)
  + Finetune a model (or train yourself) - Flair (\w pretrained LM) ?
    - Resources: <https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.36>
    - <https://zenodo.org/record/5600884>
  + Use this with prepared data set
  + Use a NER model on their own data
* Flair is using Sentences, has not more than a simple tokenizer
* uses IOB for annotation and training; Inception does exports CONLL => convert to IOB

Some bibliography:

<https://drive.google.com/file/d/1Aync-RTm3M7sh2uWoQgu7mCeD4OE2VGM/view?usp=drive_link>

=> Zotero-Folder: <https://www.zotero.org/groups/5526789/summerschoolgraz_ner/collections/6K3WFI6F>

# ToDo:

Set up Zotero => Ismail

Look for corpora: arabic => Sandy

* <https://github.com/RamziSalah/Classical-Arabic-Named-Entity-Recognition-Corpus>

Latin => Selina, Ismail

only guidelines, not corpus: [Guidelines for the Annotation of Named Entities in the Domain of Classics](https://zenodo.org/records/6368101)

<https://github.com/Herodotos-Project/Herodotos-Project-Latin-NER-Tagger-Annotation>

German => Ismail

BeNASch (German): <https://dhbern.github.io/BeNASch/>

DROC (English): <https://d-nb.info/1161488316/34>

Königsfelden (German): <https://www.koenigsfelden.uzh.ch/exist/apps/ssrq/resources/images/TRL_Web.pdf>

French => Selina

Multilingual Named Entity Recognition for Medieval Charters Using Stacked Embeddings and Bert-based Models.?

<https://gitlab.com/magistermilitum/ner_medieval_multilingual/>

<https://gitlab.com/magistermilitum/ner_medieval_multilingual/-/blob/main/datasets/french_ner_annotated_08_03_2022.json?ref_type=heads>

Hands-On:

Flair => Ismail

Inception => Sandy

Collab-Environment => Sandy to check with Tamás → no common folder set up yet

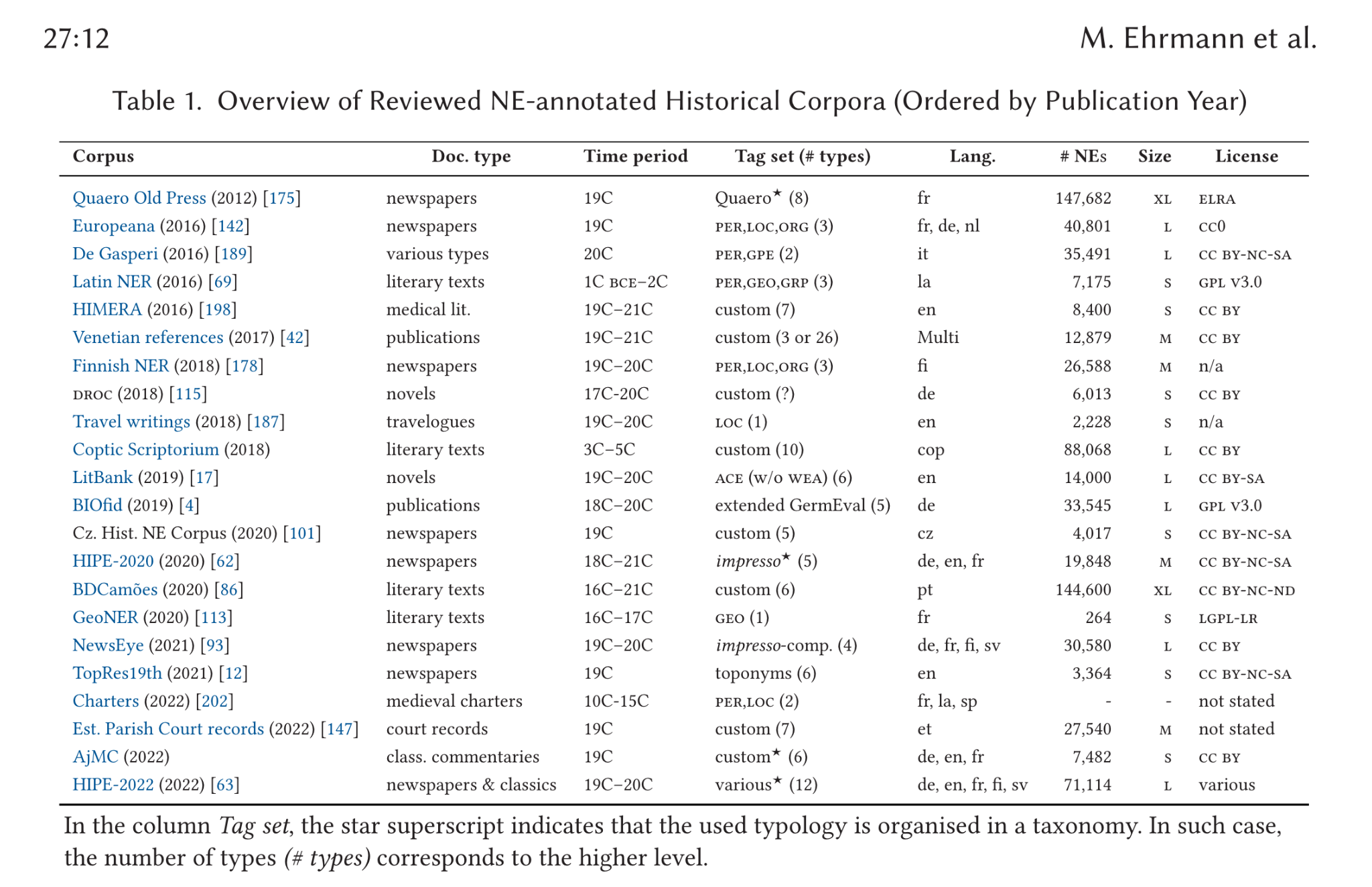
Suggestion: write your own collab notebooks for now and eventually share it?

Next meeting: 10.6. 10:00

## Corpora

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named Entity Recognition and Classification in Historical Documents: A Survey. ACM Comput. Surv. 56, 2, Article 27 (February 2024), 47 pages. <https://doi.org/10.1145/3604931>

p. 22 overview of historical NE corpora:



* **Medieval Charters (2022, fr, la sp)**
  + French/Latin/Spanisch
    - <https://gitlab.com/magistermilitum/ner_medieval_multilingual/>
    - <https://gitlab.com/magistermilitum/ner_medieval_multilingual/-/blob/main/datasets/french_ner_annotated_08_03_2022.json?ref_type=heads>
    - medieval charters in Latin, French and Spanish
    - goal: address the problem of multilingual writing practices in the Late Middle Ages
    - they have a good description of the AG in their paper
    - focus on named entities as rigid designators
    - annotated NE: persons and places
    - also nested entities (“In addition, we annotate the nested entities which are detected in charters since the early 11th century and whose use became the norm since the late-12th century”)
    - they dont include honorifics for example or the dignity title (« don Suero Pérez , obispo de Çamora » we annotate « Suero Pérez » (PERS) and « Çamora » (LOC) but we do not include the honorific prefix: « don » (Lord, dominus) and the dignity title: « obispo » (bishop)
* **Latin:**
  + <https://github.com/Herodotos-Project/Herodotos-Project-Latin-NER-Tagger-Annotation>
  + active-learning NER pipeline for low-resource Latin
  + Names of peoples are annotated as GRP; names of persons are annotated as PRS; and names of geographical places are annotated as GEO in the [BIO](https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging)) scheme.
  + to check: boundaries? nested?
  + Further information on the corpus, including splits for training and testing, can be found in Erdmann et al. (2016), "[Challenges and Solutions for Latin Named Entity Recognition](http://www.aclweb.org/anthology/W16-4012)."
  + where is the annotated corpus?
* **French: GeoNER**
  + <https://github.com/PhilippeGambette/GeoNER-corpus>
  + literary texts
  + 16C–17C
  + LOC
  + Annotation guidelines?
* **de, en, fr newspapers 18th-20th century: HIPE-2020**
  + https://github.com/impresso/CLEF-HIPE-2020/tree/master/data
* **Early Modern Dutch:**
  + <https://2024.dhbenelux.org/wp-content/uploads/2024/05/DHB24_paper_Koolen_et-al_accessing-republic.pdf> (the dataset is not published yet, but the conference paper has some rough description)
* CANER - Classical Arabic
  + <https://github.com/RamziSalah/Classical-Arabic-Named-Entity-Recognition-Corpus?tab=readme-ov-file>
  + Paper: <https://www.researchgate.net/publication/330075080_BUILDING_THE_CLASSICAL_ARABIC_NAMED_ENTITY_RECOGNITION_CORPUS_CANERCORPUS>