Daidalos:
NER for Literary Studies on Latin and Ancient Greek Texts

Nomina Omina: Detecting and Preserving Ancient Greek and Latin Proper Names in the Age of Artificial Intelligence, Leipzig, 27/06/2024

Dr. Andrea Beyer (Humboldt-Universität zu Berlin)

daidalos
— Digital Research for All —

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Named Entity Recognition for Literary Studies on Latin and Ancient Greek Texts

01 | Daidalos

02 | NER in Research: Standalone Method

03 | NER in Research: Part of a Pipeline

04 | NER in Teaching

daidalos

— Digital Research for All —

Gefördert durch

DFG Deutsche Forschungsgemeinschaft

01 | Daidalos

Project

Infrastructure

Goals

daidalos — Digital Research for All —

Gefördert durch DFG Deutsche Forschungsgemeinschaft

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Why Call a Project “Daidalos”?

We …

- develop an NLP infrastructure
- that will enable researchers in Classical Philology and related disciplines
- to apply various methods of natural language processing
- which are uncommon in the German speaking philological community.

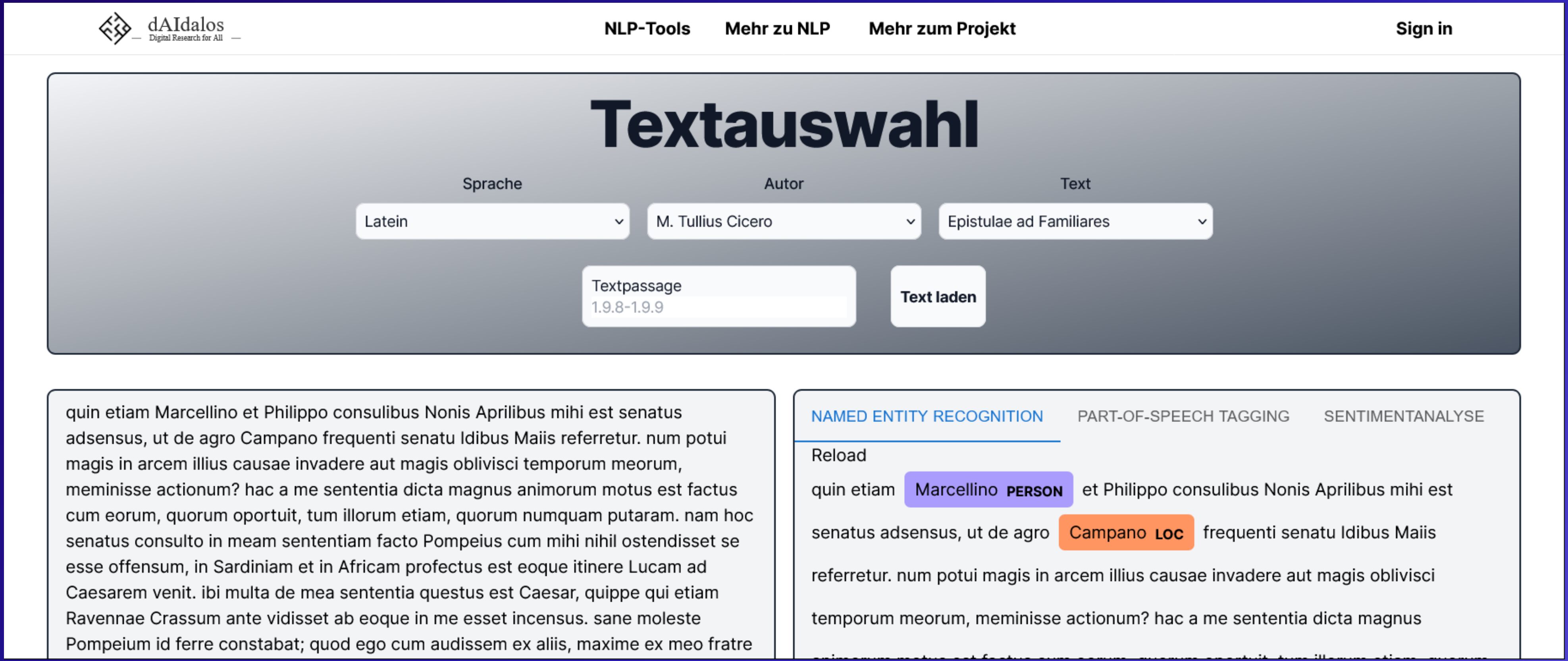
I was the most famous **inventor, craftsman, and builder** in **antiquity** – forget my human failures.

daidalos-projekt.de

daidalos — Digital Research for All —

Gefördert durch DFG Deutsche Forschungsgemeinschaft

daidalos-projekt.de



Textauswahl

Sprache: Latein

Autor: M. Tullius Cicero

Text: Epistulae ad Familiares

Textpassage: 1.9.8-1.9.9

Text laden

quin etiam Marcellino et Philippo consulibus Nonis Aprilibus mihi est senatus adsensus, ut de agro Campano frequenti senatu Idibus Maiis referretur. num potui magis in arcem illius causae invadere aut magis oblivisci temporum meorum, meminisse actionum? hac a me sententia dicta magnus animorum motus est factus cum eorum, quorum oportuit, tum illorum etiam, quorum numquam putaram. nam hoc senatus consulto in meam sententiam facto Pompeius cum mihi nihil ostendisset se esse offensum, in Sardiniam et in Africam profectus est eoque itinere Lucam ad Caesarem venit. ibi multa de mea sententia questus est Caesar, quippe qui etiam Ravennae Crassum ante vidisset ab eoque in me esset incensus. sane moleste Pompeium id ferre constabat; quod ego cum audissem ex aliis, maxime ex meo fratre

NAMED ENTITY RECOGNITION | PART-OF-SPEECH TAGGING | SENTIMENTANALYSE

Reload

quin etiam Marcellino PERSON et Philippo consulibus Nonis Aprilibus mihi est senatus adsensus, ut de agro Campano LOC frequenti senatu Idibus Maiis referretur. num potui magis in arcem illius causae invadere aut magis oblivisci temporum meorum, meminisse actionum? hac a me sententia dicta magnus animorum motus est factus cum eorum, quorum oportuit, tum illorum etiam, quorum

Menu: NLP-Tools

☑ Select: language, author, work, text passage

☑ Run

☑ Choose between NLP methods NER, POS, Sentiment Analysis

Daidalos Platform

daidalos
— Digital Research for All —

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Goals

Infrastructure

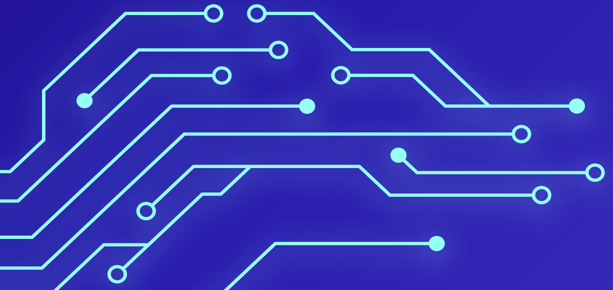
Multiple NLP methods and corpora, adjustable settings, pipelines for literary research questions, Identity & Access Management

Community of Practice

OA-Publication with research tandems, learning opportunities (Jupyter Notebooks, H5P), data bases on tools and literature, workshops

Interpretable AI

Transparency & sustainability by using model cards, data sheets, and well documented evaluations of methods



Gefördert durch
DFG Deutsche Forschungsgemeinschaft

02 | NER in Research: Standalone Method

Example

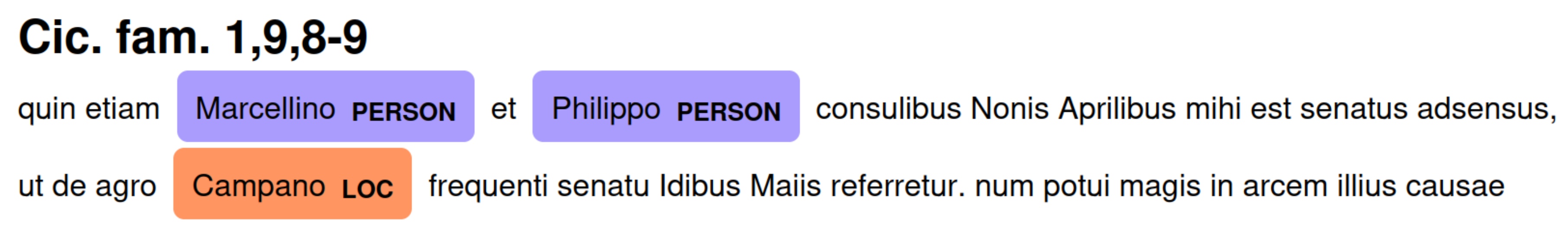
Tagger: Quality & Applications

Challenges & Solutions

daidalos
— Digital Research for All —

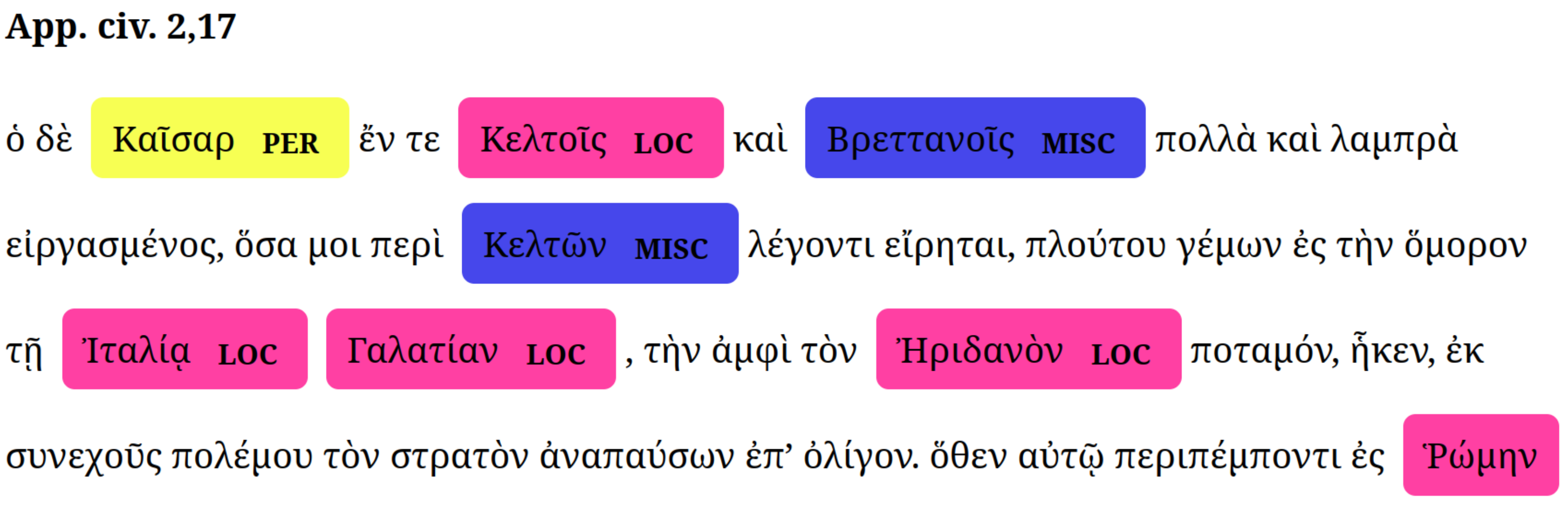
Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Example



Cic. fam. 1,9,8-9

quin etiam Marcellino **PERSON** et Philippo **PERSON** consulibus Nonis Aprilibus mihi est senatus adsensus, ut de agro Campano **LOC** frequenti senatu Idibus Maiis referretur. num potui magis in arcem illius causae



App. civ. 2,17

ὁ δὲ Καῖσαρ **PER** ἔν τε Κελτοῖς **LOC** καὶ Βρεττανοῖς **MISC** πολλὰ καὶ λαμπρὰ εἰργασμένος, ὅσα μοι περὶ Κελτῶν **MISC** λέγοντι εἴρηται, πλούτου γέμων ἐς τὴν ὅμορον τῇ Ἰταλίᾳ **LOC** Γαλατίαν **LOC** , τὴν ἀμφὶ τὸν Ἠριδανὸν **LOC** ποταμόν, ἧκεν, ἐκ συνεχοῦς πολέμου τὸν στρατὸν ἀναπαύσων ἐπ’ ὀλίγον. ὅθεν αὐτῷ περιπέμποντι ἐς Ῥώμην

daidalos — Digital Research for All —

Gefördert durch DFG Deutsche Forschungsgemeinschaft

Tagger: Quality & Applications

	Latin	Ancient Greek
Model Name	la_core_web_lg	UGARIT/flair_grc_bert_ner
Publication	Burns 2023	Yousef et al. 2023
NLP Software	spaCy	Flair NLP
Architecture	floret vectors Transition-based Parser	BERT (Transformer) vectors Long Short-Term Memory network Conditional Random Field
Training Data	Caesar, Ovid, Pliny (Elder & Younger)	Homer, Herodotus, Athenaeus
Tagset	persons, locations	persons, locations, peoples

daidalos
— Digital Research for All —

Gefördert durch
DFG
Deutsche Forschungsgemeinschaft

Challenges & Solutions

- Existing problems
  - Discontinuous, nested or overlapping annotation spans, such as "*[monasterio] Sancto Petro Cluniacensis [Ecclesiae]*"
  - Ambiguity, underspecification
  - Coordination, ellipsis, metonymy, multi-word expressions
- Possible countermeasures
  - Multi-layer annotation
  - Explicit annotations for uncertainty
  - Distinction in complexity between manual and automatic annotation

Chastang et al. 2021

03 | NER in Research: Part of a Pipeline

Example

Tagger: Quality & Applications

Challenges & Solutions

Example

General Research Question

How do you find something in the corpus that is not mentioned explicitly?

Field of Research

Omissions in Latin & Ancient Greek Historiography

Historians do not mention certain events, although they should refer to them due to their relevance, e.g. Cassius Dio does not mention the conference of Luca 56 BC.

Detailed Research Question

Is there a canonical way (place, person, topic) of mentioning this conference? Which contexts speak in favour of a mention, which against?

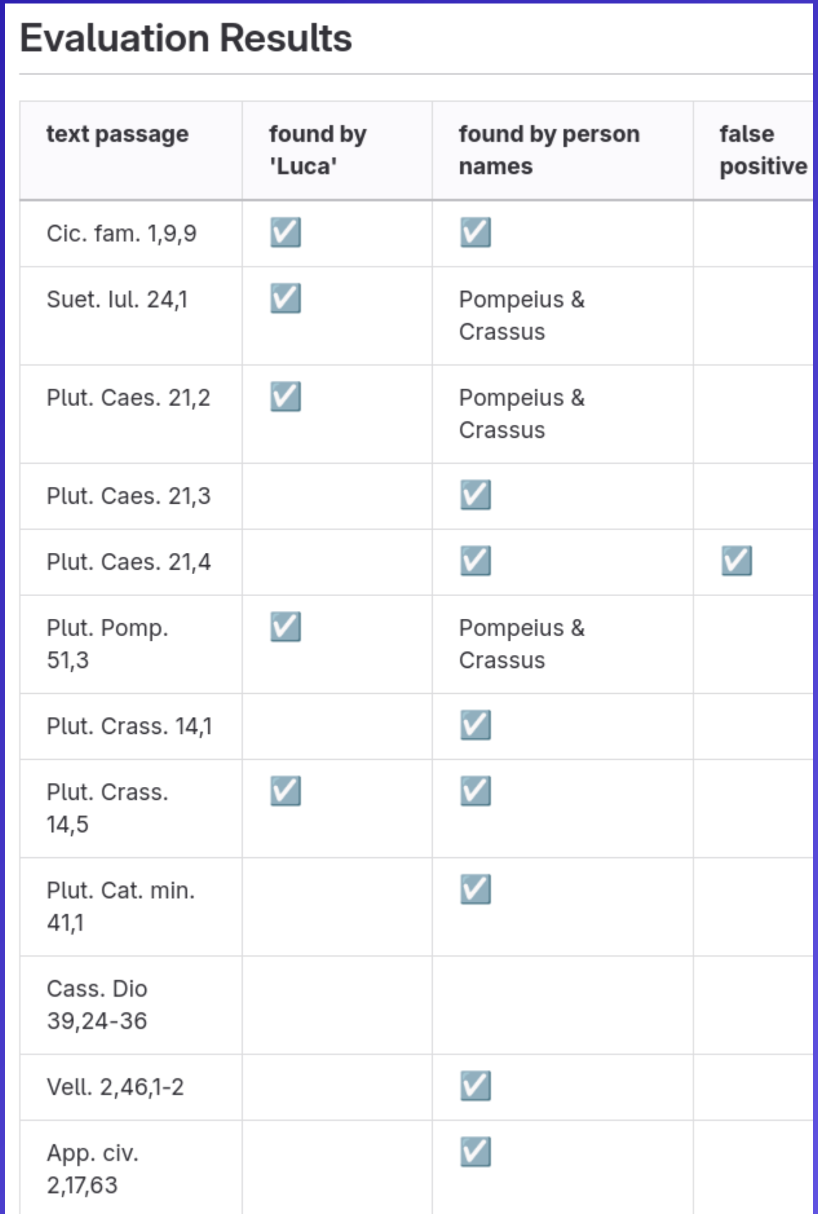
Pipeline

for passage retrieval:

- NER for mentions of places
- lemmatisation for mentions of Caesar, Pompeius and Crassus in close proximity

Pipelines: Quality & Applications

- Combination of …
  - Latin and Ancient Greek
  - NER and lemmatisation
  - Rule-based search and manual inspection
- Additional tools
  - Lemmatisers
    - Ancient Greek: greCy (grc_proiel_trf)
    - Latin: LatinCy (= same as for NER)
  - Corpus search engine: ANNIS



Evaluation Results

text passage	found by 'Luca'	found by person names	false positive
Cic. fam. 1,9,9	☑	☑	
Suet. Iul. 24,1	☑	Pompeius & Crassus	
Plut. Caes. 21,2	☑	Pompeius & Crassus	
Plut. Caes. 21,3		☑	
Plut. Caes. 21,4		☑	☑
Plut. Pomp. 51,3	☑	Pompeius & Crassus	
Plut. Crass. 14,1		☑	
Plut. Crass. 14,5	☑	☑	
Plut. Cat. min. 41,1		☑	
Cass. Dio 39,24-36			
Vell. 2,46,1-2		☑	
App. civ. 2,17,63		☑	

daidalos — Digital Research for All —

Gefördert durch DFG Deutsche Forschungsgemeinschaft

Challenges & Solutions

- Modelling 'context' as contiguous sequence of 20 words
- Conditions for search match:
  - Mention of Luca
  - Mention of Caesar AND Pompeius AND Crassus
- Identification of false positives
  - Through Close Reading for automatically retrieved text passages
- Few errors in automatic lemmatisation and NER
  - Negligible for our use case
- How to estimate false negatives?

daidalos — Digital Research for All —

Gefördert durch DFG Deutsche Forschungsgemeinschaft

04 | NER in Teaching

Model cards

Datasheets

Jupyter Notebooks

Digital Literacies

HUMBOLDT-UNIVERSITÄT ZU BERLIN

dAIdalos
— Digital Research for All —

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Teaching is About What, How, and Why

Which NER Tagger Should You Use?

Model Cards & Datasheets offer an overview

How Do You Learn to Use NER?

Curated Jupyter Notebooks provide an introduction

Why Should You Learn to Use NER?

Understanding NER is part of improving one‘s own Digital Literacies

daidalos
— Digital Research for All —

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Model Cards …

… accompany the models and provide handy information

… can be Markdown files with additional metadata

… are essential for discoverability, reproducibility, and sharing

But model cards are difficult …

… to understand by average researchers who lack the necessary digital literacies

… to compare with each other for selecting the most suitable tagger

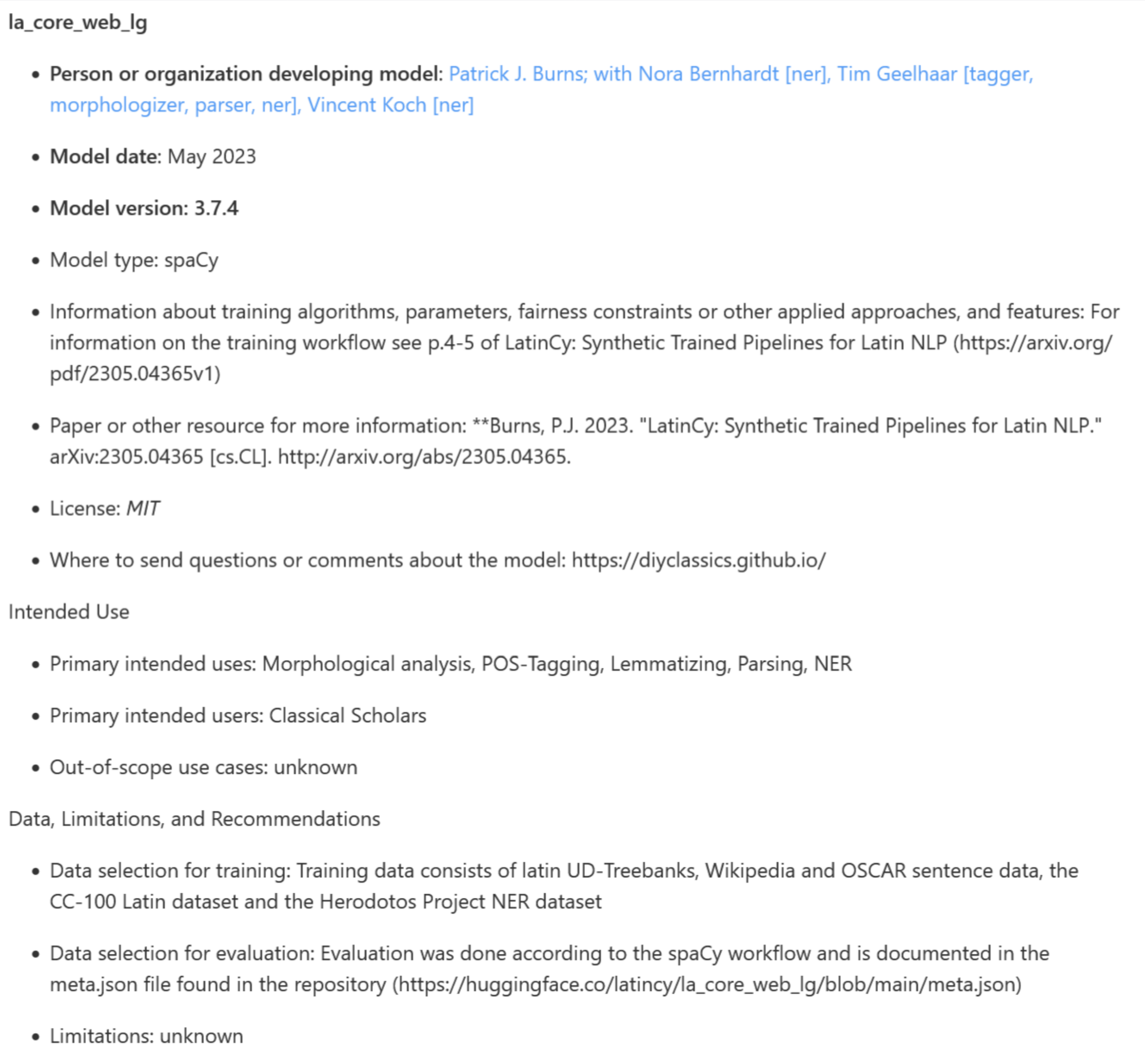
Model cards should describe …

… the model, its intended use, potential limitations, including biases and ethical considerations, the data, selection for training and evaluation, possible limitations, and recommendations, if necessary

daidalos — Digital Research for All —

Gefördert durch DFG Deutsche Forschungsgemeinschaft

Model Card



la_core_web_lg

- **Person or organization developing model**: Patrick J. Burns; with Nora Bernhardt [ner], Tim Geelhaar [tagger, morphologizer, parser, ner], Vincent Koch [ner]
- **Model date**: May 2023
- **Model version: 3.7.4**
- Model type: spaCy
- Information about training algorithms, parameters, fairness constraints or other applied approaches, and features: For information on the training workflow see p.4-5 of LatinCy: Synthetic Trained Pipelines for Latin NLP (https://arxiv.org/pdf/2305.04365v1)
- Paper or other resource for more information: **Burns, P.J. 2023. "LatinCy: Synthetic Trained Pipelines for Latin NLP." arXiv:2305.04365 [cs.CL]. http://arxiv.org/abs/2305.04365.
- License: *MIT*
- Where to send questions or comments about the model: https://diyclassics.github.io/

Intended Use

- Primary intended uses: Morphological analysis, POS-Tagging, Lemmatizing, Parsing, NER
- Primary intended users: Classical Scholars
- Out-of-scope use cases: unknown

Data, Limitations, and Recommendations

- Data selection for training: Training data consists of latin UD-Treebanks, Wikipedia and OSCAR sentence data, the CC-100 Latin dataset and the Herodotos Project NER dataset
- Data selection for evaluation: Evaluation was done according to the spaCy workflow and is documented in the meta.json file found in the repository (https://huggingface.co/latincy/la_core_web_lg/blob/main/meta.json)
- Limitations: unknown

https://anonymous.4open.science/r/seflag-DC3B/documentation/model_cards/latincy.md

Gefördert durch DFG Deutsche Forschungsgemeinschaft

Datasheets …

… offer question-driven information about the dataset of a model

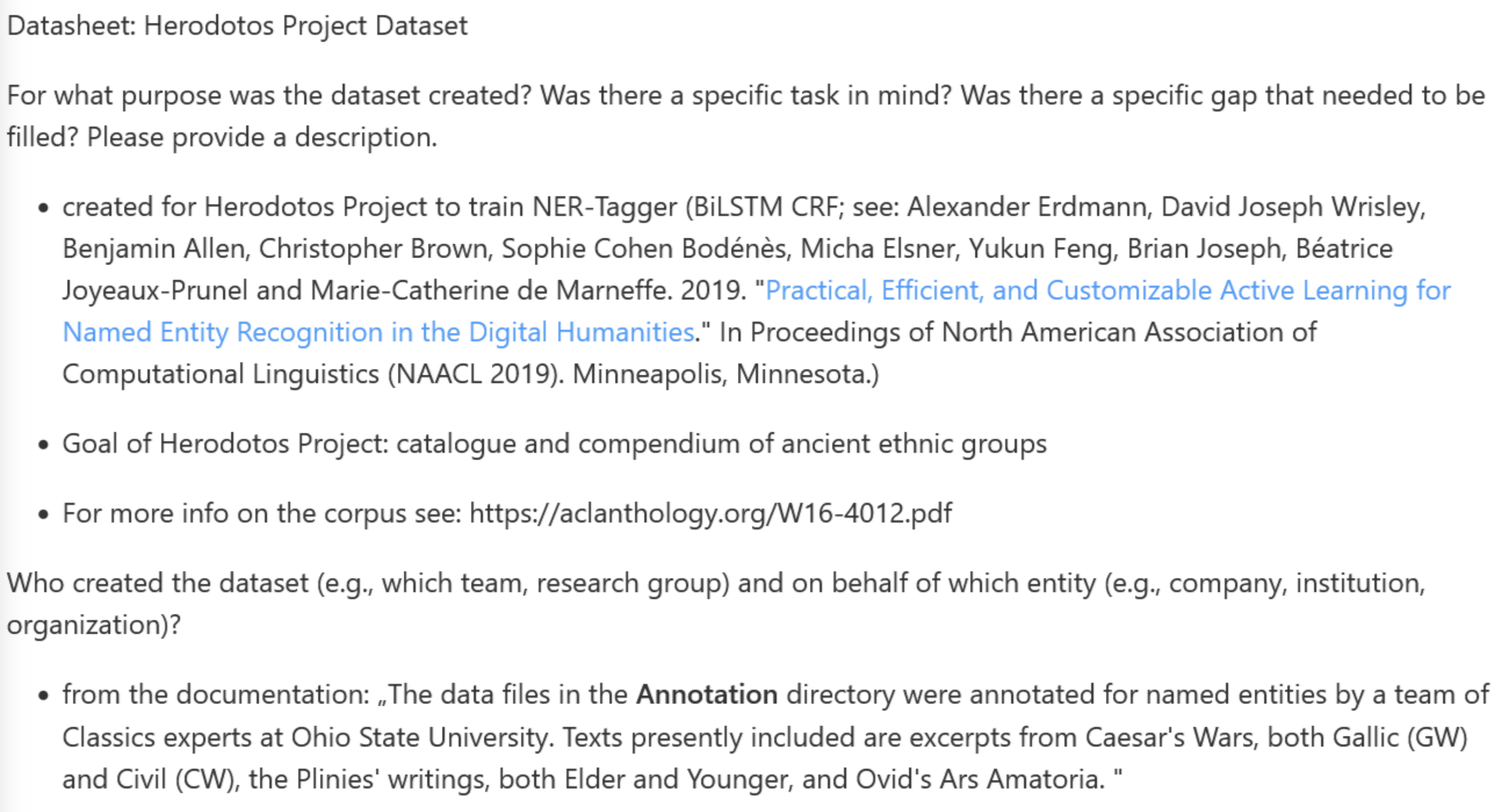
… include questions on possible sensitive data

But datasheets might contain too much information that is not structured enough for unexperienced users / researchers.

daidalos
— Digital Research for All —

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Datasheet



Datasheet: Herodotos Project Dataset

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

- created for Herodotos Project to train NER-Tagger (BiLSTM CRF; see: Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeaux-Prunel and Marie-Catherine de Marneffe. 2019. "Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities." In Proceedings of North American Association of Computational Linguistics (NAACL 2019). Minneapolis, Minnesota.)
- Goal of Herodotos Project: catalogue and compendium of ancient ethnic groups
- For more info on the corpus see: https://aclanthology.org/W16-4012.pdf

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

- from the documentation: „The data files in the **Annotation** directory were annotated for named entities by a team of Classics experts at Ohio State University. Texts presently included are excerpts from Caesar's Wars, both Gallic (GW) and Civil (CW), the Plinies' writings, both Elder and Younger, and Ovid's Ars Amatoria. "

https://anonymous.4open.science/r/seflag-DC3B/documentation/datasheet_latin.md (excerpt: only first paragraph)

daidalos — Digital Research for All —

Gefördert durch DFG Deutsche Forschungsgemeinschaft

Jupyter Notebooks as Interactive Worksheets

- Jupyter Notebooks are files that contain interactive worksheets
- Code can be supplemented with
  1. Text
  2. Coloured boxes
  3. Table of contents
  4. Integration of graphics or videos
  5. …
- Aim: acquisition of new learning content, more in-depth study or repetition, easy access to digital methods

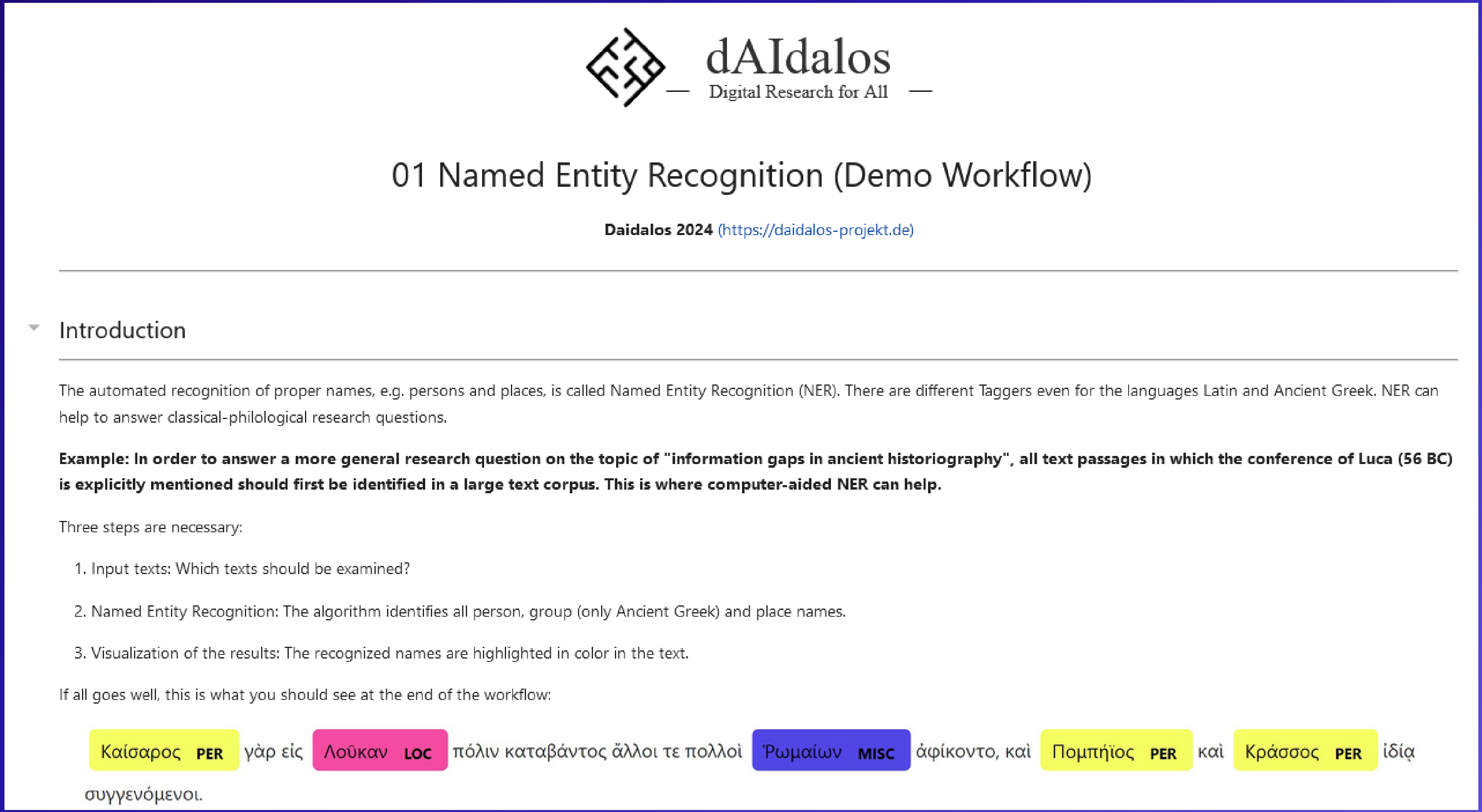
But working with Jupyter Notebooks is much more demanding than it may seem at first …

daidalos
— Digital Research for All —

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Overview

- Short method definition
- Embedding in research topic
- Approach
- Expected result



dAIdalos

— Digital Research for All —

01 Named Entity Recognition (Demo Workflow)

Daidalos 2024 (https://daidalos-projekt.de)

Introduction

The automated recognition of proper names, e.g. persons and places, is called Named Entity Recognition (NER). There are different Taggers even for the languages Latin and Ancient Greek. NER can help to answer classical-philological research questions.

Example: In order to answer a more general research question on the topic of "information gaps in ancient historiography", all text passages in which the conference of Luca (56 BC) is explicitly mentioned should first be identified in a large text corpus. This is where computer-aided NER can help.

Three steps are necessary:

1. Input texts: Which texts should be examined?
2. Named Entity Recognition: The algorithm identifies all person, group (only Ancient Greek) and place names.
3. Visualization of the results: The recognized names are highlighted in color in the text.

If all goes well, this is what you should see at the end of the workflow:

Καίσαρος PER γὰρ εἰς Λοῦκαν LOC πόλιν καταβάντος ἄλλοι τε πολλοὶ Ῥωμαίων MISC ἀφίκοντο, καὶ Πομπήϊος PER καὶ Κράσσος PER ἰδίᾳ συγγενόμενοι.

Level 1 AI Literacy

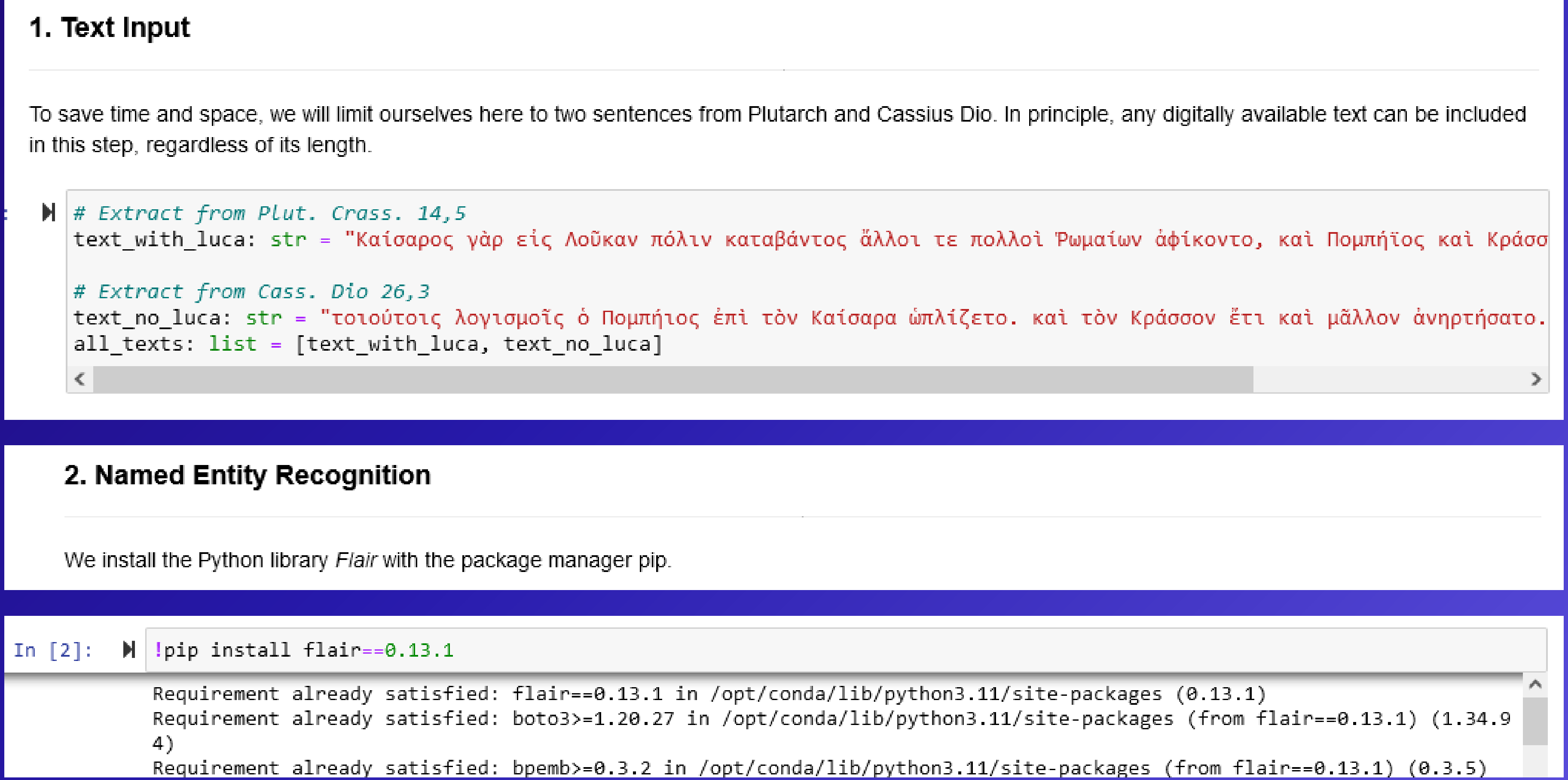
- Understand the method
- Fully guided
- Use given example

dAIdalos — Digital Research for All —

DFG Gefördert durch Deutsche Forschungsgemeinschaft

Challenges

- Using Jupyter Notebooks
- Generalisation unclear (e.g. any text)
- Technical vocabulary (e.g. library)
- Running code and dealing with potential error messages (software dependencies)



1. Text Input

To save time and space, we will limit ourselves here to two sentences from Plutarch and Cassius Dio. In principle, any digitally available text can be included in this step, regardless of its length.

```
# Extract from Plut. Crass. 14,5
text_with_luca: str = "Καίσαρος γὰρ εἰς Λοῦκαν πόλιν καταβάντος ἄλλοι τε πολλοὶ Ῥωμαίων ἀφίκοντο, καὶ Πομπήϊος καὶ Κράσσ

# Extract from Cass. Dio 26,3
text_no_luca: str = "τοιούτοις λογισμοῖς ὁ Πομπήιος ἐπὶ τὸν Καίσαρα ὡπλίζετο. καὶ τὸν Κράσσον ἔτι καὶ μᾶλλον ἀνηρτήσατο.
all_texts: list = [text_with_luca, text_no_luca]
```

2. Named Entity Recognition

We install the Python library *Flair* with the package manager pip.

```
In [2]:  !pip install flair==0.13.1

         Requirement already satisfied: flair==0.13.1 in /opt/conda/lib/python3.11/site-packages (0.13.1)
         Requirement already satisfied: boto3>=1.20.27 in /opt/conda/lib/python3.11/site-packages (from flair==0.13.1) (1.34.9
         4)
         Requirement already satisfied: bpemb>=0.3.2 in /opt/conda/lib/python3.11/site-packages (from flair==0.13.1) (0.3.5)
```

— Digital Research for All —

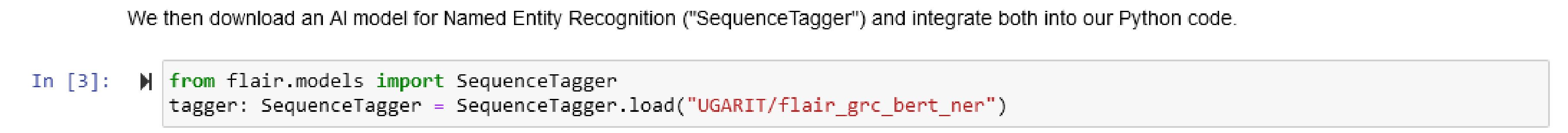
daldalos

Gefördert durch

DFG Deutsche Forschungsgemeinschaft

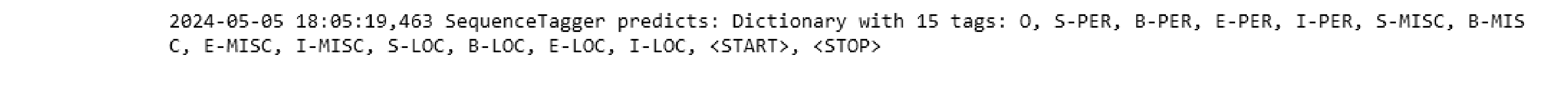
Challenges

- Connect explanation with code snippets
- Comprehend technical outputs
- Understand and interpret results (e.g. result accuracy for each entity)

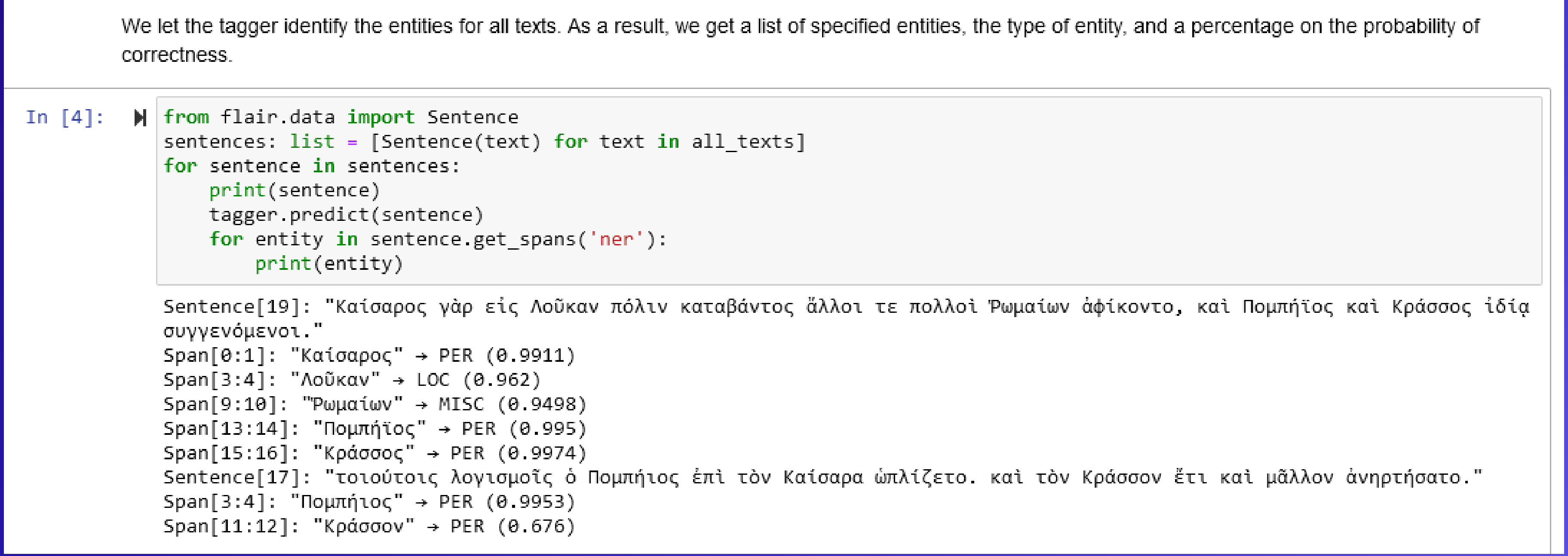


We then download an AI model for Named Entity Recognition ("SequenceTagger") and integrate both into our Python code.

```
In [3]:   from flair.models import SequenceTagger
          tagger: SequenceTagger = SequenceTagger.load("UGARIT/flair_grc_bert_ner")
```



```
2024-05-05 18:05:19,463 SequenceTagger predicts: Dictionary with 15 tags: O, S-PER, B-PER, E-PER, I-PER, S-MISC, B-MIS
C, E-MISC, I-MISC, S-LOC, B-LOC, E-LOC, I-LOC, <START>, <STOP>
```



We let the tagger identify the entities for all texts. As a result, we get a list of specified entities, the type of entity, and a percentage on the probability of correctness.

```
In [4]:   from flair.data import Sentence
          sentences: list = [Sentence(text) for text in all_texts]
          for sentence in sentences:
              print(sentence)
              tagger.predict(sentence)
              for entity in sentence.get_spans('ner'):
                  print(entity)
```

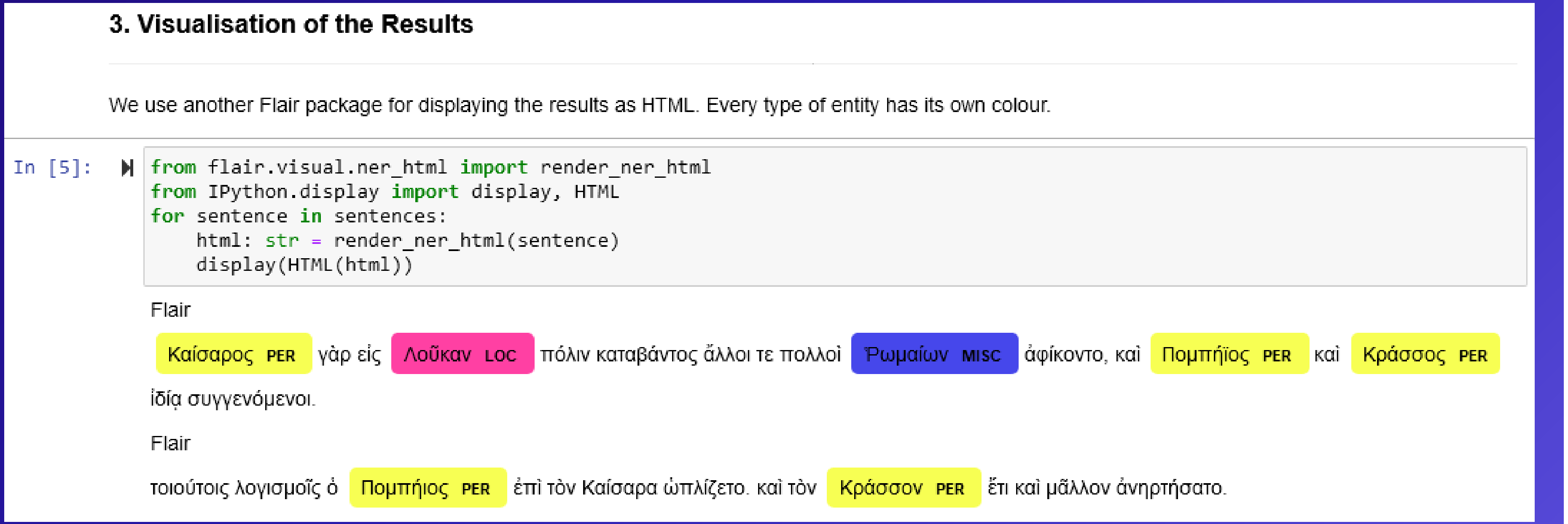
```
Sentence[19]: "Καίσαρος γὰρ εἰς Λοῦκαν πόλιν καταβάντος ἄλλοι τε πολλοὶ Ῥωμαίων ἀφίκοντο, καὶ Πομπήϊος καὶ Κράσσος ἰδίᾳ
συγγενόμενοι."
Span[0:1]: "Καίσαρος" → PER (0.9911)
Span[3:4]: "Λοῦκαν" → LOC (0.962)
Span[9:10]: "Ῥωμαίων" → MISC (0.9498)
Span[13:14]: "Πομπήϊος" → PER (0.995)
Span[15:16]: "Κράσσος" → PER (0.9974)
Sentence[17]: "τοιούτοις λογισμοῖς ὁ Πομπήιος ἐπὶ τὸν Καίσαρα ὡπλίζετο. καὶ τὸν Κράσσον ἔτι καὶ μᾶλλον ἀνηρτήσατο."
Span[3:4]: "Πομπήιος" → PER (0.9953)
Span[11:12]: "Κράσσον" → PER (0.676)
```

daidalos
— Digital Research for All —

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Challenges

- HTML
- Dealing with incorrect results
- Understanding limits and opportunities of this method



3. Visualisation of the Results

We use another Flair package for displaying the results as HTML. Every type of entity has its own colour.

```
In [5]:  from flair.visual.ner_html import render_ner_html
         from IPython.display import display, HTML
         for sentence in sentences:
             html: str = render_ner_html(sentence)
             display(HTML(html))
```

Flair

Καίσαρος PER γὰρ εἰς Λοῦκαν LOC πόλιν καταβάντος ἄλλοι τε πολλοὶ Ῥωμαίων MISC ἀφίκοντο, καὶ Πομπήϊος PER καὶ Κράσσος PER ἰδίᾳ συγγενόμενοι.

Flair

τοιούτοις λογισμοῖς ὁ Πομπήϊος PER ἐπὶ τὸν Καίσαρα ὡπλίζετο. καὶ τὸν Κράσσον PER ἔτι καὶ μᾶλλον ἀνηρτήσατο.

Generative AI and the Future of NLP in Classics:

Will we use specific taggers?

Do we need to learn about digital methods, if one multi-modal LLM could answer our research questions with similar quality?

○ ● ● ● ○ ● ● Wang et al. 2023

dAIdalos — Digital Research for All —

Gefördert durch DFG Deutsche Forschungsgemeinschaft

Bibliography

- Bada, Michael, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner, et al. 2012. “Concept Annotation in the CRAFT Corpus.” BMC Bioinformatics 13 (1): 161. https://doi.org/10.1186/1471-2105-13-161
- Beersmans, M., de Graaf, E., Van de Cruys, T., & Fantoli, M. (2023). Training and Evaluation of Named Entity Recognition Models for Classical Latin. Proceedings of the Ancient Language Processing Workshop, 1–12. https://aclanthology.org/2023.alp-1.1/
- Burns, Patrick J. 2023. “LatinCy: Synthetic Trained Pipelines for Latin NLP.” arXiv Preprint arXiv:2305.04365. https://arxiv.org/pdf/2305.04365.pdf
- Chastang, Pierre, Sergio Octavio Torres Aguilar, and Xavier Tannier. 2021. “A Named Entity Recognition Model for Medieval Latin Charters.” Digital Humanities Quarterly 15 (4).
- Ehrmann, Watter, Romanello, Clematide, and Flückiger. 2020. “Impresso Named Entity Annotation Guidelines,” January. https://doi.org/10.5281/zenodo.3604227
- Erdmann, A., Brown, C., Joseph, B., Janse, M., Ajaka, P., Elsner, M., & De Marneffe, M.-C. (2016). Challenges and Solutions for Latin Named Entity Recognition. Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), 85–93.
- Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. “Datasheets for Datasets.” Communications of the ACM 64 (12): 86–92. https://dl.acm.org/doi/10.1145/3458723.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.3287596
- Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., & Wang, G. (2023). GPT-NER: Named Entity Recognition via Large Language Models (arXiv:2304.10428). arXiv. https://doi.org/10.48550/arXiv.2304.10428
- Yousef, T., Palladino, C., & Jänicke, S. (2023). Transformer-Based Named Entity Recognition for Ancient Greek. Book of Abstracts, 420–422. https://doi.org/10.5281/zenodo.8107629

daidalos — Digital Research for All —

Gefördert durch DFG Deutsche Forschungsgemeinschaft