DIGITAL CLASSICS ONLINE

Automatic Annotation of *Nomina Sacra*

Carina Geldhauser

Abstract: *Nomina sacra* are a specific kind of named entities appearing in biblical manuscripts. Due to the large amount of biblical manuscripts, many questions about *nomina sacra* could not be answered to the present time. In order to use the methods of Digital Humanities for research questions on *nomina sacra*, they need to be consistently and accurately annotated. We report on our recent efforts on combining Handwritten Text Recognition (HTR) with annotation for biblical manuscripts written in Greek majuscule script. We reflect on the lessons learned from this work, especially on the technical aspects such as the available NER algorithms for classical languages, the performance of machine-learning based tools in comparison to rule-based annotation algorithms. We also discuss the pro's and con's of the approach we chose in our work.

Nomina sacra

A big part of the work in creating digital editions of ancient text is the standardised annotation of relevant features in the text. Which features are 'relevant' depends on the scope of the edition, the nature of the text and its linguistic features. Often, proper names are a relevant category, and this work deals with the annotation of a very specific class of proper names, the so-called *nomina sacra*.

A *nomen sacrum* is an abbreviation of a specific name or word carrying a meaning (e.g. The Spirit) that appears systematically in manuscripts of the Bible, most prominently in Greek manuscripts, some as old as *Papyrus Bodmer* II (known as P66). It is characterised by an overline bar spanning two or more letters from the original word, for example ΘΣ for Θεός.

The origin of *nomina sacra* as a scribal practice is not totally clarified as of today. Ludwig Traube, who coined the usage of *nomina sacra* as a technical term in his 1907 monograph,[1] conjectured the origin of *nomina sacra* in Hellenistic Judaism and their usage of the Tetragramm *JHWH*, but the current hypothesis is that *nomina sacra* are a characteristic Christian phenomenon.[2]

Nomina sacra as a phenomenon in Biblical manuscripts has been described by many scholars, and a number of theories on their origin and usage were developed. Within the manuscripts that were investigated by theologians, abbreviation by contraction, meaning that the first and last letter (at least) of each word are used, prevails. However, especially in earlier manuscripts,[3] we also find the alternate practice of abbreviation by suspension, meaning that the initial two letters of the word are used. We may, at the moment, only speculate on the prevalence of abbreviation by contraction – it certainly was of advantage as it indicated the case of the abbreviated noun.

1 Traube (1907).

2 Hurtado (2017), 127 even uses the variations in spelling (contraction, suspension, or mixed abbreviation) as evidence for this hypothesis.

3 For example, Jesus Christ is abbreviated as *IH XP* in the opening verses of Revelation in P18.

The current hypothesis is that *nomina sacra* were not merely used to save space, but as an act of reverence,[4] evidenced by the observation that *nomina sacra* were sometimes used to distinguish between 'mundane' and 'sacred' usages of the same word (e.g., spirit vs The Spirit), and employed even when other common abbreviations were not used.

The picture is yet not so simple as the observations mentioned above do not hold for all manuscripts, and there are different abbreviated forms used for the same word in different manuscripts. Moreover, there were different ways that scribes used to deal with the Hebrew *tetragrammaton JHWH*, spanning from unabbreviated forms in the Greek text to the peculiar double zeta with a horizontal line through the middle in the *Septuagint* manuscript *Papyrus Oxyrhynchus* 1007.[5]

Methods of Digital Humanities could help to shed further light into the usage of *nomina sacra*, to bolster or to reject the above-mentioned hypotheses, and to help getting new insight. For this, we need a reliable way to mark/annotate *nomina sacra* within a (digitalised or) transcribed text – which motivated our work.

The Annotation Problem for Greek Manuscripts

Naively, one might believe that all questions regarding *nomina sacra*, as a particular case of Named Entity Recognition (NER), might be solved within minutes. Indeed, for modern high-resource languages, named entity recognition and other Natural Language Processing (NLP) tools have been successfully used to annotate texts, allowing authors to explore huge datasets through statistical methods.

However, scholars in Classics and Biblical Theology have a harder life here, as Ancient Greek is, from the Machine Learning point of view, a low-resource language.[6] Not because there would not exist enough literary texts, prosaic and poetic, in Greek language – far from it! However, there exist few and mainly[7] quite small annotated text corpora that can be used as training sets for data-hungry ML models.

Furthermore, there might not even exist reliable digitalisations of the texts that one intends to annotate. Sometimes, but not always, a digitalisation can be generated quickly through readily available software such as *transcribus* and *escriptorium*. In general, we need to be cautious when dealing with manuscripts[8], as often, the automatic post-processing step removes diacritics or hyphens, meaning a significant piece of information is lost.

Another cause of trouble arises when using readily digitised texts from different sources in one dataset: first, different OCR software supply different output formats, and unification might be difficult or prone to create errors in many instances of the dataset. Also, copyright or licensing restrictions are a big obstacle in gathering available datasets for training of NLP tools.

Second, even if technical obstacles are overcome, the ML model might still pick up on specific, technical cues such as encodings and output file characteristics, instead of textual cues. Some even have problems with punctuation signs.[9] Furthermore, the heterogeneity in encoding causes current ML tools

4 See, e.g. Hurtado (2017), 100, 104–106.

5 Wilkinson (2015), 55.

6 See e.g. Kostkan et al. (2023), 128.

7 The recent *Opera Graeca Adnotata* by Celano (2024) could be a gamechanger here for certain applications (not *nomina sacra*).

8 See Geldhauser / Malyshev (2024).

9 Vatri / McGillivray (2020), 189 reported that GLEM was unable to identify a word followed by a punctuation mark as equal to the same form in the middle of a sentence.

to fail. It is known[10] that features characteristic of a philological, palaeographical, or diplomatic edition, such as special characters, textual gaps, abbreviations, headlines, and orthographic redundancy may "confuse" the algorithm or at least present additional degrees of freedom that create the need for a much larger training dataset in order to be classified as noise in the training data. Clearing these features from the dataset, or to diminish it by automatic and manual cleaning, may help to resolve the issue, but may also remove important cues like the overline bar characterising a nomen sacrum. Specific to Greek texts, let us mention the apostrophe issue that greCy developer J. Myerston remarked in his readme file:

> "Unfortunately, there is no consensus among the different internet projects that offer ancient Greek texts about how to represent the Ancient Greek apostrophe. Modern Greek simply uses the regular apostrophe, but ancient texts available in *Perseus* and *Perseus* under Philologic use various unicode characters for the apostrophe. Instead of the apostrophe, we find the Greek *koronis*, modifier letter apostrophe, and right single quotation mark. Provisionally, I have opted to use modifier letter apostrophe in the corpus with which I trained the models. This means, that if you want the greCy models to properly handle the apostrophe you have to make sure that the Ancient Greek texts that you are processing use the modifier letter apostrophe ʼ (U+02BC). Otherwise the models will fail to lemmatize and tag some words in your texts that ends with an 'apostrophe'".[11]

Named Entity Recognition for Classical Languages: Progress and Challenges

Named Entity Recognition is a task that grew out of the more general task of information extraction since the 1990s. Initially, handcrafted rule-based algorithms were used, but later machine-learning techniques became more and more popular. Nowadays Natural Language Processing (NLP) models such as NER are everywhere, and advertised with amazing performance. However, users of these technology observe that as soon as an off-the-shelf algorithm is applied to 'the real world', performance drops.[12] The reason for this problem is the training of NLP algorithms on a very specific, standard dataset, or basically a limited set of canonical varieties of it, used as benchmark corpora. So, essentially all NLP models are trained on English newswire. Real world data differs radically from the benchmarks, and although "any domain can be reasonably supported, porting a system to a new domain or textual genre remains a major challenge".[13] Constructing a robust model requires finding a technical and intellectual compromise between collecting sufficiently varied data and avoiding extreme training on some similar structures.[14]

10 See, e.g., Chastang et al. (2021), paragraph 73.

11 See https://github.com/jmyerston/greCy?tab=readme-ov-file (last access 28.07.2025).

12 Plank (2016a), 1; Nadeau / Sekine (2007).

13 Nadeau / Sekine (2007), 2. They report that some tests reveal up to 20% to 40% of precision drop when applying an algorithm to a different genre.

14 Plank (2016a), 1.

Challenges in NER for Classical Languages

In order to properly understand and appreciate the developments in NER for classical languages, its specific tools and their performances, we need to take a closer look at the challenges that NER faces on historical documents in general. The challenges can be roughly divided into four types:[15]

1. The historical variety space.
2. Noisy input.
3. Language dynamics.
4. Lack of resources.

The *variety space* is a term coined by Barbara Plank, which she defined as an "unknown high-dimensional space, whose dimensions contain fuzzy aspects such as language (or dialect), topic or genre, and social factors (age, gender, personality etc.) amongst others".[16] The usual *terminus technicus* is 'domain', but Plank remarks that this term is ill-defined in the literature and overloaded. Thinking of the huge variety of historical texts, from administrative documents, correspondences, archives of all sorts, literary works, articles, reports, memoirs etc, spanned over thousands of years and countless languages, it is easy to see that no NER tool can be equally capable of handling all of these.

NER is classically trained on standardised datasets containing English-language data.[17] Some research, always on modern texts, has been dedicated to test the ability of these systems to generalise to a different genre, unseen mentions, another domain or document type. All studies reveal a 'NER transfer gap' already for modern texts, and given the needs of humanities research, which is much broader than the typical contemporary NLP applications (that are often motivated by commercial interests, narrowing the scope of the tool), we need to be very careful when using an off-the-shelf tool for our work.

The term *noisy input* is another way of saying OCR errors. Like for all efforts to annotate a digitised text, the accuracy of the annotation highly depends on the accuracy of the OCR or HTR step that created the text basis. While human understanding is quite robust in the sense that it does not bother about tokenisation problems or character misinterpretations (e.g. the characters 'e' and 'c' in printed Latin texts might be mixed up, especially for printed input), the 'OCR noise propagation' may cause a drastic decline in F-score of up to 30%,[18] making annotations as random as tossing a coin.

The term 'language dynamics' summarises the variations in spelling, naming conventions, and in general the differences between modern and historical languages, which affect the performance of NLP tools. A machine-learning based algorithm trained on Ionic Greek forms[19] as found in Herodot. There are some works which study the difference in the structure of entity names, e.g. between historical and contemporary news,[20] but further research is needed: even relatively simple sounding issues such as correctly linking, e.g., 'Madame Pierre Curie' to 'Marie Curie' seem to be left unaddressed for now.

A big challenge for NER of historical texts in general is the heterogeneity of typologies: The mainstream typologies for modern documents have a few 'universal' classes, e.g. *Person*, *Organisation*,

15 We follow here the systematisation of Ehrmann et al. (2023), section 4.

16 Plank (2016a).

17 Standard datasets include CoNLL-2003 (English news texts from Reuters in the year 1996) BERT was originally trained on the *BookCorpus* and English *Wikipedia*: see Devlin et al. (2019).

18 E.g., van Strien et al. (2020), 195 reported a drop from 87% to 63% for person identities.

19 E.g. the lemmatizer *GLEM* was trained on Ionic prose texts by Herodotus, see also the discussion in Vatri / McGillivray (2020).

20 See Rosset (2012).

Location, which could, in principle, be re-used for historical documents, but in general the entity types might require adaptations to fit to the application at hand. An off-the-shelf NER tool is unlikely to capture all entities of interests in a historical document, or certain classes like *Person* need to be divided into *Historical Person*, *Literary Person* or even finer into *Greek/Latin (Half-)God*, *Mythical Creature* etc. – depending on the research question. Also inside a topology, spelling variations are a problem for NER, and so-called 'historical normalisation' might be necessary.

Finally, a big challenge for NER for historical languages is the lack of financial resources, but also digitally usable lexica and annotated corpora. Most NER methods use supervised learning, i.e. they depend on already annotated corpora (= labelled data) to train their models. When annotated corpora are scattered over time and domains, the models can neither be improved nor their generalisation capacity increased. That 'state-of-the-art'-results on specific tasks can be achieved for classical languages if that dataset is large enough is shown by LatinBERT,[21] which was trained on the *Perseus Digital Library*, Latin *Wikipedia* and Latin texts from the Internet Archive.

We should, however, be aware of the "news bias"[22] in NER: Most annotated corpora for (standard) NER consist of news texts. As historical newspaper collections were massively digitised during the last years, it was logical to also base the annotated corpora for historical NER on those historical newspaper collections. These corpora also benefit from available word embedding technologies that are able to flag potential OCR misspellings and therefore allow for post-OCR correction. But then historical NER will run eventually in the same "news bias" as modern NER.

New Developments

In recent years there have been several approaches to improve NER for classical languages. Let us recall that common difficulties for historical NER are spelling variations, including punctuation, capitalisation and person name abbreviation, unknown names, and in general the complexity of entities. To give some very simple examples from a Latin language project:[23] Gaius Iulius Caesar should be marked as one entity, not as a "partial match". The algorithm may learn well compound entities of clear forms, e.g. name + de + toponym, such as *Bertrannus de Verziaco*, but already *Gariardus de loco Antimiano* may be difficult to resolve as one person name.

Chastang et al. (2021), who worked on a model for the automatic recognition of named entities in medieval Latin charters, describe very well the need for careful pre-processing to detect nested, overlapping or ill-formed annotations. Overlapping annotations stem from names that serve different functions: a saint's name can denominate the historical person, an abbey, a feudal territory, a festivity date, etc. As a standard machine learning classifier is not designed to attribute more than one class to each instance, the confusion between the multiple usage of a name (i.e., the *overlapping entities*) must be solved by creating designated classes for each entity. However, the more entities created, the more choice and potential ambiguity is created.

Moreover, Latin is a very versatile language, which makes NER a difficult task *in se*:

> The overgeneralization of very common particles (such as *de* in compound entities), as well as of location trigger words (such as *terra, serum, pars, domus, manus, apud*) and also of personal co-occurrent words (such as *episcopus, beatus, dominus, sacerdos, miles*, etc.) can lead to false positives when the model finds an entity different than that expected. (…) Latin phrase order is irregular, and exceptions in medieval variants are almost infinite; consequently, training taking

21 See Bamman / Burns (2020) and https://github.com/dbamman/latin-bert (last access 28.07.2025). LatinBERT achieved "state-of-the-art" performance on POS tagging.

22 Used as *terminus technicus* in Plank (2016a).

23 Chastang et al. (2021).

> into account grammatical rules, co-occurrences, and context can generate many false positives (…) The difficulty in recognizing entities (...) lies not so much in their quantity as in their extensive consequences. The percentage of complex entities does not exceed 11% of the total in our corpora, but the statistical impact on results due to bad recognition of such entities is more elevated [with growing complexity].[24]

With this in mind, it is understandable that the transformer-based NER for ancient Greek by Yousef et al. (2023b) performed poorly on multi-token entities: the available training data was composed of single-token entities.

The work of Berti (2019) started from large annotated datasets of specific sources and used semantic annotation platforms and Machine Learning. Vatri /McGillivray (2020) compared several lemmatizers and concluded that those based on large lexica are still producing better results than the ML-based lemmatizers.

The interesting work of Yousef et al. (2023a) suggests to use cross-lingual annotation projection to transfer NER annotations, done on translations, to Greek and Latin. The idea here is not to machine-translate detected name entities and to look them up in the "more difficult" language, but to employ automatic word alignment to find the equivalents of the detected entity in the parallel sentence and then to project the annotation.[25] The precondition for the success of such an approach is the alignment of the text corpora in both languages on the sentence or paragraph level. This makes the Bible as an obvious demonstration example. The perceived accuracy of the annotations was 86% for English-Ancient Greek.[26] Misclassifications appeared as a consequence of the translation, which adopted ad different type of entity, and multi-token entities such as 'Jesus Christ' or 'Pontius Pilate' were frequently misaligned or only partially aligned. Nevertheless, their result is a promising sign, if one happens to have a fine-tuned model for the languages and sentence-level alignment at hand.

Which Tool to Use?

Considering the above-mentioned challenges, what shall we do when we have a NER task to do? The answer is, as always 'it depends' – in particular, it depends what we want to annotate, and which factors are important to us.

First of all, what is our goal with the annotation? Which question do we wish to answer? Do we have a huge unexplored, but fully digital text corpus at hand, and we wish to get a rough overview? Are we aiming for a fine-grained annotation to answer delicate research questions? Is the goal a digitally enhanced edition, which is easily searchable?

Second, what is our raw data? In which documents and for which usecases do we intend to use our desired annotation? Of which century or which type of language are we concerned? Does there already exist a high-quality OCR of the desired texts? If not, which century and which writing style do the manuscripts-to-be-annotated have? Do we already have a robust OCR model that we can use?

Third, we need to decide which role does accuracy play for us: Do we need a highly accurate annotation within a digital edition, with (ideally) no false-positives or false-negatives? Or do we merely wish to get an overview, a visualization or another type of digital enhancement of the text?

To give a comprehensive decision 'algorithm' in answer to the above-mentioned questions is beyond the scope of this work. A lesson-learned from our work is the following rule of thumb: The more accuracy plays a role, the more specific a tool has to be, and the more likely we will end up with pro-

24 Chastang et al. (2021), paragraphs 72, 83, 84.

25 This approach was suggested in the computational linguistics community by Ni et al. (2018).

26 We use 'perceived accuracy' here as there was no standardised scoring employed, but qualitative inspection of random sample verses by humans. It is not known if partial matches were perceived as accurate or as not accurate.

gramming a rule-based algorithm ourselves, potentially including a large lexicon as feed-in data to it if necessary. But if our aim is to quickly get an overview or explore a huge corpus, the more likely it is that we are OK with a couple of false positives, as they do not skew a statistical analysis of a text too much.

Of course, with limited time, technical expertise and resources, we may often be inclined to use off-the-shelf NER tools for our data at hand. This is perfectly reasonable, but the question is whether we can reasonably expect accuracies[27] when our dataset is very different from the standard CoNLL-03 English corpus. Despite the advertisement, all-purpose general NER may have significant drop in accuracy in our application. Already an accuracy of 80% means that every fifth annotation is wrong. As we may, after hasty employment of an off-the-shelf tool, not expect much more than 70%, rather less, we need to ask ourselves if an error rate of 3 matches out of 10 is still giving us a meaningful output of our research question.

Recognition and Annotation of *Nomina Sacra*

Depending on the factors above, we may decide which tool to turn to. There are already good tools like *spaCy*, *greCy* or *odyCy*, that do give annotations of reasonable quality on certain Greek texts. But we need to keep in mind that Ancient Greek is a highly fragmented language with plenty of variants, both regional and temporal. Currently, each available model is limited to the particular dialect or morphology of their training dataset and generalization is rather poor.[28]

To our best knowledge, the current available NER tools do not recognize or evaluate *nomina sacra*. We guess that probably no available model has ever seen the original abbreviated form, as available datasets are made with a different scope than ours: We are interested (also) in the temporal development of scribal practice, hence, the comparison of manuscripts from different centuries is an important tool for our research question. The majority of scholars interested in texts are, however, dealing with questions on the content of the text, and therefore are not concerned with potential scribal mistakes or incomplete textual transmission, they merely need 'the text', by which they mean 'the original / true text'. Which makes total sense as there are not too many cases of complex transmission histories that are well-studied.

In case of biblical texts, 'the text of the Bible' is usually defined as Nestle-Aland's latest edition, which is also available in digital form. Nestle-Aland's edition is the output of decade-long research on the largest possible amount of available manuscripts[29] (papyri, majuscules, minuscules, lectionares, talismans), it is not 1:1 aligned to the text of any particular manuscript, but a carefully curated edition that aims to get as close as possible to the 'Ausgangstext', the best approximation to the original Bible text. While being an incredible piece of scholarly work, it is, unfortunately, not helpful for us.

The ideal (non-existing) machine-learning algorithm to recognize *nomina sacra* would scan the manuscript image for overline bars, check whether the letters under the overline bar are connected to a 'sacred word', in the sense that the letters form a reasonable abbreviation of a dictionary word that is a named entity in the Bible, and then would finally expand the abbreviation into that word.[30]

To train a model for such a goal, we would need a large set of manuscript-specific transcriptions of high-quality manuscript scans that contain *nomina sacra*. That led us to use available transcribed co-

27 It is hard to give reliable accuracy numbers, as the versions of commercial tools put out in the web change fast, and their underlying datasets are not always revealed. Just to give an idea, BiLSTM achieved an F-score of 90,1% in Huang et al. (2015) and, when improved with subword representations, of 91,2%, see Ma / Hovy (2016).

28 See Kostkan et al. (2023).

29 The editing institution, the Institute For New Testament Textual Research, has collected approximately 5800 manuscripts up to the present: see https://www.uni-muenster.de/INTF/ (last access 28.07.2025) for more information.

30 More technical details can be found in Geldhauser / Malyshev (2024).

dices such as *Codex Sinaiticus* as a ground truth: we started fine-tuning an *escriptorium* model with 50 pages of artificially created ground truth from *Codex Sinaiticus*, for which we provided a ground-truth annotation, including an annotation of *nomina sacra*.

We used this 'data augmentation' approach as it was the fastest way to ensure we had a lot of examples of *nomina sacra* correctly annotated. Despite 50 pages should have been sufficient to adapt an existing model, according to the *escriptorium* user community's 'rule of thumb', the model failed to recognize *nomina sacra* – indeed, it seems one needs much more data to preserve the overline bar as a cue.

Therefore, it became clear that our aim was far more difficult to achieve than we thought: after recognising the overline bar as a cue, an imaginary future algorithm also has to understand, given the heterogeneity of the abbreviations used, that two or more different sets of characters may correspond to the same expanded word form.

Moreover, to successfully annotate *nomina sacra* of transcriptions of biblical texts originating in a different century, a ML model has to be general enough to deal with a variety of different writing styles, page formats etc.

Hence, we subsequently turned to a sequential method, disentangling recognition and annotation. For annotation, we decided for a rule-based approach.[31] Here, we define rule-based approach in the pure sense of the word: the user creates a list of relevant abbreviations for annotation, using a tool of their choice, and the algorithm contains a function that finds and expands abbreviated *nomina sacra* from the transcribed text. Here, no 'learning' of the algorithm is required, it is a simple search script.

This 'disentangling'-approach is possible as Bruce Metzger compiled what seems to be a complete list of words treated as *nomina sacra* from Greek papyri: The Greek counterparts of God, Lord, Jesus, Christ, Son, Spirit, David, Cross, Father, Israel, Saviour, Man, Jerusalem, and Heaven[32] all occur as *nomina sacra* in Greek language biblical manuscripts of the 3rd century and often earlier, and, according to Comfort and Barrett, also Mother is consistently used as nomen sacrum from the 4th century onwards.[33] These 15 *nomina sacra*, together with their relevant forms for genitive and other cases, and both in contraction and suspension abbreviation, are used as rulebook for annotation.

The advantage of our method is its simplicity, fastness and precision. Of course, our method is not universal, but tailored to the problem: the working of our algorithm employs 'expert knowledge' in the sense that we use the special characteristics of *nomina sacra*, which were discovered and listed by humans. In case the OCR were flawless, we would get a perfect accuracy thanks to the rule-based approach that we take in the annotation step.

A tailored approach is by design not thought to generalise well to other situations, hence, we do not claim that our approach will be suitable for a vast amount of other purposes. However, our concept will be helpful also to scholars that look for a fine-grained annotation of named entities in other settings: Indeed, our approach can be used whenever a complete list of entities of a given class is readily available, independently of the century, the genre, the type of text etc. For example, reference works such as the *Genealogisches Handbuch des Adels* provide a list of noble people in the former *Sacrum Imperium Romanum*, and new insights on power structures, alliances and networks of the nobility of a certain time could be gained when annotating all appearances of noble people listed in this reference work.

---

31 Historically, the terminus rule-based approach was attributed to methods that exploit grammar and regularities in the data. In the 1990s, these were state-of-the-art algorithms that rely on linguistic pre-processing, such as morpho-syntactic tagging, tokenisation and sentence splitting, and that often require external resources such as trigger words in gazetteers.

32 Metzger (1981), 36.

33 Comfort / Barrett (2001), 34.

Summary and Conclusion

NER techniques are an excellent way to provide an overview of a corpus, and can provide added value to many datasets in general and digital editions in particular.

We reviewed a couple of existing methods and concluded that there is no all-purpose tool that always gives good results. A big obstacle for the application of the most widely known and used NER tools to texts and research questions of scholars in Classics is their unsuitable training databases, normally modern newspaper collections. Moreover, the vast variety of texts and the variability of the Latin and Greek language over the course of the centuries implies the need for a multitude of tools.

In our specific application, we were interested in the annotation of a specific named entity that is not among the standard list, and for which there is not yet a suitably large training dataset. Hence, there was no potential basis of success for a machine-learning based algorithm. Nevertheless, we could get quite promising results with a tailored rule-based approach.

In terms of high-level recommendations for users, one takeaway from our project is that NLP tools, at their current state-of-the-art, are not the right tool for a highest-accuracy project. This is currently an issue ‘by design’: An LLM tool was built to produce an output that is ‘natural‘, by whatever standard, it uses probabilities and predictions on words based on their underlying dataset. In this sense, it is designed to output a ‘good’ text, but it is not designed to perform a very precisely defined ‘algorithmic’ task with 100% accuracy. When deciding which tool to (further) develop, digital humanists should not let themselves be blinded by the current AI hype, but decide strategically which kind of working mechanism is suitable for their research question and the specific task to be performed algorithmically.

References

Agerri et al. (2018): R. Agerri / Y. Chung / I. Aldabe / N. Aranberri / G. Labaka / G. Rigau, Building named entity recognition taggers via parallel corpora, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki 2018.

Bamman / Burns (2009): D. Bamman / P. J. Burns, LatinBERT: A contextual language model for classical philology, in: arXiv:2009.10053.

Berti (2019): M. Berti, Named entity annotation for ancient Greek with INCEpTION, in: CLARIN Annual Conference Proceedings, Leipzig 2009, 1–4.

Celano (2024): G. G. A. Celano, Opera Graeca Adnotata: Building a 34M+ Token Multilayer Corpus for Ancient Greek, in: arXiv:2404.00739.

Chastang et al. (2021): P. Chastang / S. Torres Aguilar / X. Tannier, A Named Entity Recognition Model for Medieval Latin Charters, DHQ 15/4 (2021), https://www.digitalhumanities.org/dhq/vol/15/4/000574/000574.html (last access 28.07.2025).

Comfort / Barrett (2001): P. W. Comfort / D. Barrett, Text of the Earliest New Testament Greek Manuscripts (2nd ed.), Wheaton (IL) 2001, 34–35.

Devlin et al. (2019): J. Devlin / M.-W. Chang / K. Lee / K. Toutanova, BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, 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), Minneapolis (MN) 2019, 4171–4186.

Ehrmann et al. (2023): M. Ehrmann / A. Hamdi / E. L. Pontes / M. Romanello / A. Doucet, Named entity recognition and classification in historical documents: A survey, in: ACM Computing Surveys 56/2 (2023), 1–47.

Geldhauser / Malyshev (2024): C. Geldhauser / K.A. Malyshev, Semi-automatic annotation of Greek majuscule manuscripts: Steps towards integrated transcription and annotation, in: Annals of Computer Science and Information Systems 41 (2024), 37–44.

Huang et al. (2015): Huang Z. / Xu W. / Yu K., Bidirectional LSTM-CRF Models for Sequence Tagging, in: arXiv:1508.01991.

Hurtado (2017): L. W. Hurtado, Texts and Artefacts: Selected Essays on Textual Criticism and Early Christian Manuscripts, London 2017.

Kostkan et al. (2023): J. Kostkan / M. Kardos / J. P. B. Mortensen / K. L. Nielbo, OdyCy – A general-purpose NLP pipeline for Ancient Greek, in: Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Dubrovnik 2023, 128–134.

Ma / Hovy (2016): Ma X. / E. Hovy, End-to-End Sequence Labeling via Bi-Directional LSTM-CNNs-CRF, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin 2016, 1064–1074.

Metzger (1981): B. M. Metzger, Manuscripts of the Greek Bible: An Introduction to Palaeography. London 1981.

Nadeau / Sekine (2007): D. Nadeau / S. Sekine, A survey of named entity recognition and classification, in: Lingvisticae Investigationes 30 (2007), 3–26.

Ni et al. (2017): Ni J. / D. Georgiana / F. Radu, Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver 2017, 1470–1480.

Plank (2016a): B. Plank, What to do about non-standard (or non-canonical) language in NLP, in: arXiv:1608.07836.

Plank (2016b): B. Plank, What to Do about Non-Standard (or Non-Canonical) Language in NLP, in: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016), Bochum 2016, 13–20.

Rosset (2012): S. Rosset / C. Grouin / K. Fort / O. Galibert / J. Kahn / P. Zweigenbaum, Structured Named Entities in Two Distinct Press Corpora: Contemporary Broadcast News and Old Newspapers, in: 6th Linguistics Annotation Workshop (The LAW VI), Jeju 2012, 40–48.

van Strien et al. (2020): D. van Strien / K. Beelen / M. Coll Ardanuy / K. Hosseini / B. McGillivray / G. Colavizza, Assessing the Impact of OCR Quality on Downstream NLP Tasks, in: Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Valletta 2020, 484–496.

Traube (1907): L. Traube, Nomina sacra. Versuch einer Geschichte der christlichen Kürzung, München 1907.

Vatri / McGillivray (2020): A. Vatri / B. McGillivray, Lemmatization for ancient Greek: An experimental assessment of the state of the art, in: Journal of Greek linguistics 20/2 (2020), 179–196.

Wilkinson (2015): R. J. Wilkinson, Tetragrammaton: Western Christians and the Hebrew Name of God: From the Beginnings to the Seventeenth Century, Leiden 2015.

Yousef et al. (2023a): T. Yousef / C. Palladino / G. Heyer / S. Jänicke, Named entity annotation projection applied to classical languages, in: Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, 175–182.

Yousef et al. (2023b): T. Yousef / C. Palladino / S. Jänicke, Transformer-based named entity recognition for ancient Greek, in: Digital Humanities 2023, University of Graz 2023, 1–3.

Author Contact Information[34]

Dr. Carina Geldhauser
ETH Zürich
Rämistr. 101
8092 Zürich
Schweiz
E-mail: carina.geldhauser@math.ethz.ch

34