ZEFYS2025: A German Historical Newspaper Dataset for Named Entity Recognition and Entity Linking

Sophie Schneider, Ulrike Förstel, Kai Labusch, Jörg Lehmann, Clemens Neudecker
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, Berlin, Germany
Correspondence: clemens.neudecker@sbb.spk-berlin.de

Abstract

We present ZEFYS2025, a dataset for named entity recognition and entity linking that comprises 100 annotated German historical newspaper pages covering the time period from 1837 to 1940. We describe the data selection and annotation processes and evaluate various models fine-tuned on the NER downstream task using ZEFYS2025 as well as additional historical and contemporary German NER datasets. Our findings show that the recognition performance on ZEFYS2025 is on par or better than on other historical newspaper datasets while pretraining on ZEFYS2025 improves NER performance in most cases.

1 Introduction

The concept of a named entity (NE) can be defined by drawing upon the notion of a rigid designator (Kripke, 1980). It implies that many of the NE instances in texts equal proper names, although theoretically other units are encompassed in this definition as well. While named entity recognition (NER) aims at detecting and classifying NEs within a text into predefined classes (e.g., person, location and organisation), entity linking (EL) concerns their disambiguation and linking to corresponding entries in a database. Information extraction techniques such as NER and EL can support the semantic indexing and enrichment of historical documents (Ehrmann et al., 2019). Digital cultural heritage collections are often queried for NEs (Sumikawa et al., 2019), and information retrieval applications such as semantic search, cross-document contextualisation or automatic text summarisation can be improved through NER and EL.

Due to numerous mentions of people, places and events contained in them, historical newspapers are a rich source for NEs. Furthermore, historical newspapers are frequently used by scholars in the digital humanities, as their broad coverage and mix of content supports diverse research use cases, often in combination with NEs. This makes historical newspapers particularly valuable for NER and EL processing and large historical newspaper corpora with annotated NEs highly desirable.

The introduction of the Transformer (Vaswani et al., 2017) led to the emergence of numerous pretrained models, including various downstream language modeling tasks, one of which is NER (Wolf et al., 2020). However, historical sources present these models with a wide variety of specific challenges (Ehrmann et al., 2023). Texts obtained by optical character recognition (OCR) of historical documents inevitably exhibit noise leading to a drop in NER accuracy (Hamdi et al., 2020, Hamdi et al., 2019), and spelling conventions have undergone significant changes over time (Provatorova et al., 2024). Variations in orthography and language usage are complemented by changes in how persons, organisations, or places are addressed. Thus, the application of generic NE processing tools to historical sources does not achieve the same quality as for contemporary texts; resources derived from modern perspectives such as benchmark datasets or knowledge bases do not include domain-specific entities or information (Dereza et al., 2023). Establishing large, annotated corpora is a cumbersome and costly endeavour, although they are indispensable to facilitate sound performances of NER and EL systems or to precisely measure the performance of NER and EL methods. With ZEFYS2025, we contribute a consistently annotated, comprehensive dataset for NE processing of historical German texts.

2 Related Work

Historical German-language datasets for NER and EL with more than 100,000 tokens are only sparsely available. In our experiments, we focused on contemporary (CoNLL-2003, GermEval 2014) as well as historical corpora (e.g., HIPE 2020, Europeana) and also included smaller datasets (such as NEISS-Sturm and HisGermaNER). We considered

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 48–58, 2025

adding the German subset of the NewsEye dataset (Hamdi et al., 2021) which comprises 527,756 tokens derived from historical newspapers. However, as our initial test results indicated that prediction on this specific subset is challenging (in line with those of previous examinations, (Schweter et al., 2022)), we excluded it from further experiments and evaluations. We still included the NewsEye dataset statistics in our dataset overview for comparison (see tables 1 and 2). Except for HIPE 2020 and NewsEye, the datasets discussed here contain NE tags, but no EL information.

2.1 CoNLL-2003

The CoNLL-2003 shared task (Tjong Kim Sang and De Meulder, 2003) introduced datasets for contemporary English and German NER. The German part of CoNLL-2003 contains 310,318 tokens and was extracted from articles published in the German newspaper *Frankfurter Rundschau* in August 1992. The CoNLL-2003 data includes miscellaneous names (MISC) as an additional entity type besides persons (PER), locations (LOC) and organizations (ORG). Nested entities are not considered, only the surface level is annotated.

2.2 GermEval 2014

GermEval 2014 (Benikova et al., 2014) is a contemporary German NER corpus consisting of 591,006 tokens. It was sampled from German Wikipedia articles and online newspapers. NEs are tagged with four main types: PER, LOC, ORG and other (OTH). GermEval 2014 annotations can occur as nested, partly or derived NEs. The annotation guidelines aim at capturing semantically relevant information within a text rather than restricting it to its syntactic functionality.

2.3 HIPE 2020

The HIPE-2020 shared task (Ehrmann et al., 2020a) resulted in a dataset for evaluating NE processing tasks on French, German and English historical newspapers. Its German subset includes 153,875 tokens from historical newspapers up to 1950. HIPE annotation guidelines (Ehrmann et al., 2020b) contain several entity types and subtypes/components, while nested annotations were only carried out for PER, LOC and ORG entities. The dataset includes links to Wikidata identifiers (QIDs) as well.

2.4 NEISS

The NER datasets developed in the NEISS project (Zöllner et al., 2021) are built from existing digital editions. More precisely, the NEISS corpus consists of two different German datasets. The "Essays from H. Arendt" dataset is constructed from 23 documents published between 1932 and 1976 and comprises 157,637 tokens. NEs have been tagged intellectually and entity types were primarily revised/mapped, for instance to resolve ambiguities between tag type definitions. Besides PER, LOC and ORG, ethnicity, event and language are kept as NE types. By contrast, the "Sturm Edition" contains 174 letters from 1914 to 1922. It includes only the entity types PER, LOC and date, but not ORG. Its size is limited to 35,953 tokens.

2.5 HisGermaNER

The HisGermaNER dataset[1] is compiled from 100 newspaper pages from the Austrian National Library (ÖNB), covering a time span from 1710 to 1840. NE types are identical to those in our ZEFYS2025 dataset and the dataset format is derived from the HIPE-2022 shared task. According to our calculations performed after preprocessing this dataset, it comprises roughly 72,000 tokens.

2.6 Europeana

In the context of the Europeana Newspapers project, a multilingual resource for NER on the basis of Dutch, French and German newspapers has been created (Neudecker, 2016). The German part of this dataset consists of 100 newspaper pages and 96,735 tokens in total, compiled from the Dr Friedrich Teßmann Library (LFT) and the Austrian National Library (ÖNB). A later addition to the dataset included newspaper pages by the Berlin State Library (SBB).[2] However, we refrained from including the SBB material, since a significantly modified version of it is already contained in the ZEFYS2025 dataset. The classes used for annotating NEs are PER, LOC and ORG.

3 The ZEFYS2025 dataset

3.1 Data selection

With ZEFYS2025, we present a German dataset for historical NER and EL consisting of 100 an-

[1] https://huggingface.co/datasets/stefan-it/HisGermaNER

[2] https://github.com/EuropeanaNewspapers/ner-corpora

notated newspaper pages, totaling approximately 350,000 tokens. All newspaper pages were extracted from the Berlin State Library's newspaper system ZEFYS (an abbreviation of ZEitungsinFormationssYStem),[3] preprocessed and annotated with a predefined set of NE types and links to Wikidata identifiers. The pages included in ZEFYS2025 stem from 20 newspaper titles, all of which were published between 1837 and 1940, such as the *Berliner Börsen-Zeitung*, the *Berliner Tageblatt und Handels-Zeitung* and *Berliner Volkszeitung*.

A key objective in establishing ZEFYS2025 was the creation of a dataset that is both sufficiently sized and homogeneously annotated. Frequently, datasets for historical NE processing are tailored to a particular project or task, resulting in incompatible and segmented resources due to differences in annotation. Accordingly, HIPE-2022 included a task specifically addressing heterogeneous annotation tag sets and guidelines (Ehrmann et al., 2022). For ZEFYS2025, we combine pre-existing NE datasets, originally produced in two distinct projects over a time period of 10 years, with additional, newly annotated newspaper pages into a single coherent resource. Existing annotations were harmonized to a common standard (cf. 3.2) and annotations for all pages were carefully checked in iterations to ensure consistency. In order to facilitate the comparison and analysis of the effects of OCR errors, the final dataset comprises full texts which were produced automatically by OCR (84 pages) and also 16 pages with manually transcribed text (i.e. ground truth).

3.2 Annotation guidelines and NE tagset

We refer to the Impresso guidelines (Ehrmann et al., 2020b) as a starting point, from which an adapted version was developed. Impresso defines five named entity types (persons, locations, organisations, human products and time) and includes 23 fine-grained subtypes. We restrict our NE tag set to the three entity types PER, LOC and ORG originally included in the ENAMEX tag types defined for the MUC-7 Named Entity Task (Chinchor and Robinson, 1998). Furthermore, we generally did not consider components such as the title or function of a person to be part of a NE, with the exception of them being regarded as inseparable from the remaining entity token sequence. Most other datasets discussed here describe additional NE types, for instance a class for other or miscellaneous entities. In some cases also derived or partial entities were tagged, and even instructions themselves can lead to NEs being treated differently. To achieve as much interoperability as possible, we deliberately focused on a minimalist tag set and preferably annotated the literal instead of a metonymic or similar figurative meaning when in doubt. By discarding the individual tag classes, all specialized NE tag sets from the presented datasets can be mapped to our broader tag set and therefore merged with the ZEFYS2025 dataset (Palladino and Yousef, 2024). We also emphasize the demand for more generalizable and compatible historical NER/EL datasets (Ehrmann et al., 2023, Hamdi et al., 2021) and point towards the limitations of such mapping approaches (e.g., loss of very specific information incorporated in annotations). Nonetheless, from a pragmatic point of view we consider this a useful strategy for dealing with differently annotated datasets.

Concerning the annotation of NEs, we allow nesting up to a depth of one, resulting in NEs on a higher (NE-TAG) and second, lower level (NE-EMB). Only the higher level NEs were considered for the EL annotations. Ideally, each recognized NE is linked to one definite Wikidata identifier (QID[4]) in our dataset. This is challenging due to innumerable factors, for instance the evolution of entities and their meaning (entity drift) or domain and time shifts between historical resources and the modern knowledge bases used for linking (Munnelly and Lawless, 2018, Linhares Pontes et al., 2020). The links added should describe a certain entity as specifically as possible – in particular with regard to the (spatio-)temporal dimension and historical correctness. This can result in multiple entries of an entity in the database being (correctly) linked to the same token sequence.

How NEs are tagged and linked to entries in a contemporary knowledge base does not only depend on the context in which they appear, but also requires attentiveness to the rhetorical use of language. We linked immediately recognizable entities on a literal basis and did not strive for resolving any intended meaning. An exception to this rule are the numerous company names listed within historical financial newspapers. These companies are tagged as ORG at the highest level in spite of them being often abbreviated to the persons or places

[3] https://zefys.staatsbibliothek-berlin.de/?lang=en

[4] https://www.wikidata.org/wiki/Q43649390

after which they are named (e.g. *Baltimore and Ohio* instead of *Baltimore and Ohio Railroad*).

3.3 Annotation tool

A browser-based tool[5] has been developed for the manual annotation and correction of NEs, tokens and segmentation. The tool is a simple HTML-/JS-based web application and can be run locally. It accepts tab-separated value (TSV) files in the data format described in 3.5 as input and returns similar TSV files when saving the results after editing. Several functionalities are implemented for ease of use when annotating NEs. In additional columns, a IIIF Image API URL and image region coordinates can be specified to display image snippets from a scanned newspaper page on hovering. Shortcuts for all NE tagging and further options were included to speed up the annotation and correction process. OCR corrections are facilitated by the integration of a virtual keyboard offering a variety of specific Unicode characters occurring in historical documents.

3.4 Annotation process

Automatic annotation. As a starting point to the entire processing chain, OCR has been applied to 84 out of 100 newspaper pages, with 16 pages being manually transcribed to ground truth, resulting in either ALTO-XML or PAGE-XML files. The OCR output and ground truth were then converted into our TSV format with the help of TSV tools[6]. This data conversion included preprocessing steps such as tokenization and sentence splitting with SoMaJo.[7] The resulting files were automatically enriched with NER by applying a BERT-based system presented at KONVENS 2019 (Labusch et al., 2019). In a second step, recognized NEs have been disambiguated and linked to Wikidata entries using an EL system presented at HIPE-2020 (Labusch and Neudecker, 2020) and HIPE-2022 (Labusch and Neudecker, 2022).

Intellectual revision. Using the annotation tool, all automatically generated annotations have been intellectually checked and revised by a group of German native speakers. This included the verification of the NE tags assigned previously as well as the addition or deletion of tags if needed. For NEs confirmed in this step, existing links were checked for the most precise and correct option for linking, and missing links to entities were inserted when available. To flag ambiguous or uncertain cases arising during annotation for later discussion, an extra NE-TAG class *TODO* was used during this process. Consensus about challenging cases collected throughout the annotation was reached in regular discussion meetings, and the set of instructions was expanded iteratively, adding further rules or examples when necessary. Since the revision was carried out by one expert per page only, it was not possible to calculate inter-annotator agreement or similar measures to assess consistency between multiple annotators. Instead, we employed computational methods to assist in localizing and reducing inconsistencies within extensively annotated datasets (Rücker and Akbik, 2023, Reiss et al., 2020). The automated analyses identified reappearing tokens throughout our dataset with a high divergence in how they were transcribed, tagged or linked. In additional correction loops, these inconsistencies were systematically reviewed once again by the annotators.

3.5 Dataset characteristics and format

The resulting ZEFYS2025 dataset presented here comprises 100 newspaper pages totaling 348,307 tokens, 16,727 sentences, 14,072 entities (∼3% of which are nested) and 10,341 links. Locations are not only the most frequently occurring entity type within our dataset, they also seem to be by far the most straightforward to disambiguate and link to Wikidata (less than 5% NIL fraction for LOC, whilst ∼31% and ∼47% NIL fractions for ORG and PER entities). Using the train_test_split function provided by the Hugging Face datasets library (Lhoest et al., 2021), the ZEFYS2025 dataset was sampled and divided into a 80%/10%/10% train-validation-test split. See tables 1 (overall dataset statistics) and 2 (PER, LOC and ORG statistics) for more detailed information on ZEFYS2025 in comparison to other datasets described here. In these tables, we present the numbers as stated in the corresponding original publications referenced here. For the sake of completeness, we also calculated missing values afterwards. The results of these calculations might slightly deviate from the statistics taken from a publication, as we might have used a newer dataset version than the one decribed originally, and file preprocessing was performed prior to the calculations.

We release ZEFYS2025 in a file format based on the GermEval TSV format (Benikova et al.,

[5] https://github.com/qurator-spk/neat

[6] https://github.com/qurator-spk/page2tsv

[7] https://github.com/tsproisl/SoMaJo

dataset	sentences	tokens
ZEFYS2025	16,727	348,307
CoNLL-2003	18,933	310,318
GermEval 2014	31,300	591,006
HIPE 2020	5,887*	153,875
NewsEye	26,826*	527,756
NEISS-Arendt	10,731*	157,647
NEISS-Sturm	3,250*	35,953
HisGermaNER	2,345	72,430*
Europeana-LFT	2,768*	70,254*
Europeana-ÖNB	976*	28,012*
	entities	**links**
ZEFYS2025	14,072	10,341
CoNLL-2003	20,357	-
GermEval2014	41,005	-
HIPE 2020	6,584	5,066
NewsEye	13,786	2,279
NEISS-Arendt	4,424	-
NEISS-Sturm	2,144	-
HisGermaNER	2,700	-
Europeana-LFT	7,350*	-
Europeana-ÖNB	3,302*	-

Table 1: Description of the ZEFYS2025 dataset in comparison to other German NER datasets. *These data have been calculated on our final, preprocessed versions.

dataset	PER	LOC simple	ORG
ZEFYS2025	4,389	6,049	3,223
train	3,476	4,771	2,592
validation	432	616	310
test	481	662	321
CoNLL-2003	5,369	6,579	4,441
GermEval 2014	10,500	12,165	7,175
HIPE 2020	1,910	3,006	660
NewsEye	3,500	5,904	3,370
NEISS-Arendt	1,702	1,087	455
NEISS-Sturm	930	492	-
HisGermaNER	1,488	1,182	30
Europeana-LFT	2,797*	3,433*	1,120*
Europeana-ÖNB	1,589*	1,618*	95*
		nested	
ZEFYS2025	207	203	1
GermEval 2014	488	1,452	281
HIPE 2020	29	209	61
NewsEye	246	544	176
		links	
ZEFYS2025	2,329	5,775	2,237
HIPE 2020	1,332	2,883	458
NewsEye	685	610	967

Table 2: Amount of named entity annotations per class across different datasets. *These data have been calculated on our final, preprocessed version.

2014), which in turn is derived from the widely used CoNLL data format (Tjong Kim Sang and De Meulder, 2003). Further columns were added containing links to Wikidata entries and token coordinates from the corresponding scanned page images. Annotations comply with the IOB scheme (Ramshaw and Marcus, 1999), more specifically IOB2, in which the position of the tags at the beginning (B-) or inside (I-) an entity is determined more precisely by means of a prefix. Tokens that do not constitute a NE (chunk) receive an outside (O) tag.

4 Experiments

4.1 Data preprocessing

The CoNLL-2003[8] and GermEval 2014[9] datasets have been retrieved directly via the Hugging Face API as they did not require extensive preprocessing from our side. We downloaded the HIPE 2020,[10] NEISS,[11] HisGermaNER[12] and Europeana[13] datasets directly from the corresponding repositories. Comments and other metadata, mostly indicating the beginning of a document and its provenance, were removed at this stage. Train-validation-test splits and sentence splitting have been reused for all historical and contemporary datasets from the original publications. For Europeana, we refer to the German subset in the latest dataset version, from which we extracted the ÖNB/LFT parts and added train-validation-test plus sentence splits in the same way as in ZEFYS2025. With regard to our second and third experiment, all NE tags were matched with their tag names from ZEFYS2025 (*B-pers* was changed to *B-PER* etc.). Similarly, subtypes and components were ignored to fit the broader class names from our selection. We created an new version of each dataset on the basis of the updated labels from the coarse (upper level) NE annotations.

4.2 Model training

Building upon the Hugging Face Transformers library (Wolf et al., 2020), we implemented a pipeline to read and transform datasets, load models from Hugging Face and train and evaluate var-

[8] https://huggingface.co/datasets/eriktks/conll2003

[9] https://huggingface.co/datasets/GermanEval/germeval_14

[10] https://github.com/hipe-eval/HIPE-2022-data/tree/main/data/v2.1/hipe2020/de

[11] https://github.com/NEISSproject

[12] https://huggingface.co/datasets/stefan-it/HisGermaNER

[13] https://github.com/cneud/ner-corpora

ious dataset-model combinations.[14] Prior to the actual training process, we filtered the NER labels of the corresponding datasets to only contain the classes PER, LOC and ORG that are also present in ZEFYS2025. In other words, NE tags describing entities other than persons, places or organisations were replaced with an outside (O) tag. All models were trained until the F_1-score on the validation split did not improve further for two consecutive epochs. We optimized over different batch sizes and learning rates, selected the results with the best F_1-score on the validation split and then tested on the test split. All reported global F_1-scores summarize the class-specific F_1-scores for PER, LOC and ORG by weighing them according to their support in the dataset.

4.3 Model selection

In our choice of models for the experiments, we focus on BERT-based models (Devlin et al., 2019) that have been obtained using different pretraining strategies (GermanBERT, RoBERTa, Electra, GBERT), are either optimized for contemporary (GermanBERT, GBERT) or historical German (hmBERT, Europeana ELECTRA), or are multilingual (hmBERT, Roberta, XLM-Roberta). We consider distilled or smaller model variants (DistilBERT, hmBERT-mini, hmBERT-tiny, comp. Zöllner et al., 2021) as well as larger architectures (GermanBERT, GBERT, hmBERT, Europeana-ELECTRA, RoBERTa, XLM-RoBERTa).

- DistilBERT (Sanh et al., 2019) constitutes a version of the original BERT reduced through knowledge distillation. It is 40% lighter and 60% smaller while sacrificing only little performance in general language understanding as well as on several downstream tasks.

- RoBERTa (Liu et al., 2019) originates from a replication study of BERT and introduces an optimized approach with several changes in hyperparameter settings and strategies employed during pretraining procedure.

- XLM-RoBERTa (Conneau et al., 2020) is a multilingual masked language model, pretrained with texts in 100 different languages. Besides its multilingual approach, XLM-RoBERTa still performs well on monolingual and low-resource language modeling tasks.

model	$\mathbf{F_1^{TE}}$	$\mathbf{F_1^{VAL}}$	$\mathbf{F_1^{PER}}$	$\mathbf{F_1^{LOC}}$	$\mathbf{F_1^{ORG}}$
		CoNLL-2003			
RoBERTa	0.94	0.97	0.96	0.93	0.92
XLM-RoBERTa	0.93	0.96	0.96	0.92	0.90
GermanBERT	0.90	0.94	0.91	0.91	0.88
DistilBERT	0.92	0.94	0.96	0.91	0.88
GBERT	0.90	0.94	0.92	0.90	0.88
hmBERT	0.89	0.93	0.92	0.90	0.84
Europ-ELECTRA	0.88	0.91	0.90	0.90	0.84
hmBERT-mini	0.84	0.89	0.86	0.86	0.80
hmBERT-tiny	0.77	0.84	0.80	0.79	0.72
		GermEval			
GermanBERT	0.88	0.89	0.93	0.90	0.79
GBERT	0.88	0.89	0.94	0.89	0.78
XLM-RoBERTa	0.86	0.87	0.91	0.88	0.76
Europ-ELECTRA	0.84	0.84	0.90	0.87	0.73
hmBERT	0.82	0.83	0.88	0.86	0.69
RoBERTa	0.81	0.83	0.87	0.83	0.71
DistilBERT	0.76	0.77	0.80	0.80	0.66
hmBERT-mini	0.72	0.75	0.78	0.76	0.59
hmBERT-tiny	0.63	0.66	0.66	0.69	0.50

Table 3: Results on contemporary datasets. F_1^{TE}: F_1-score on test split. F_1^{VAL}: early stopping F_1-score on validation split. F_1^{PER},F_1^{LOC},F_1^{ORG}: Class-specific F_1-scores on test split.

- GBERT (Chan et al., 2020) is a BERT-based model, pretrained with different German datasets and applying Whole Word Masking (WWM). WWM ensures that all subword tokens of a word are masked simultaneously.

- hmBERT (Schweter et al., 2022) is a historical multilingual BERT-based model which is pretrained on historical German, English, French, Swedish and Finnish corpora from Europeana and the British Library. By filtering these data based on OCR confidence scores and creating domain-specific vocabulary, the particular challenges associated with historical language were considered.

- Additionally, we included the GermanBERT[15] and Europeana-ELECTRA[16] models published by DBMDZ on Hugging Face. GermanBERT has been pretrained on contemporary German text. ELECTRA is an alternative BERT pretraining approach that relies on a classification task rather than on token prediction (Clark et al., 2020).

[14] https://github.com/qurator-spk/sbb_ner_hf

[15] https://huggingface.co/dbmdz/bert-base-german-cased

[16] https://huggingface.co/dbmdz/electra-base-german-europeana-cased-discriminator

model	$\mathbf{F_1^{TE}}$	$\mathbf{F_{1_{PH}}^{TE}}$	$\mathbf{F_1^{PER}}$	$\mathbf{F_1^{LOC}}$	$\mathbf{F_1^{ORG}}$
Europ-ELECTRA	0.83	**0.84**	0.84	0.89	0.70
hmBERT	**0.83**	**0.83**	0.85	0.88	0.71
GermanBERT	0.82	**0.83**	0.84	0.89	0.70
GBERT	**0.82**	0.81	0.83	0.88	0.67
XLM-RoBERTa	0.80	**0.81**	0.82	0.87	0.67
RoBERTa	0.78	**0.80**	0.80	0.87	0.69
DistilBERT	**0.75**	**0.75**	0.74	0.82	0.64
hmBERT-mini	**0.72**	**0.72**	0.70	0.80	0.57
hmBERT-tiny	0.60	**0.64**	0.61	0.74	0.46

Table 4: Results on the new ZEFYS2025 dataset. F_1^{TE}: F_1-score on test split. $F_{1_{PH}}^{TE}$: F_1-score on test split when model is pretrained on the joined training splits of all historical NER datasets. F_1^{PER},F_1^{LOC},F_1^{ORG}: Class-specific F_1-scores of best configuration on test split, i.e., either with pretraining on joined historical NER datasets or without.

4.4 Baseline results

In a first experiment, we establish baseline results for the contemporary and historical German NER ground truth datasets in section 2. We fine-tune the state-of-the-art Transformer models listed in 4.3 with respect to the NER task. Table 3 shows the results for the contemporary datasets CoNLL-2003 and GermEval 2014. Column F_1^{TE} of table 4 presents the respective results on our ZEFYS2025 dataset for the same models. Column F_1^{TE} of table 5 lists the respective results for the other historical NER ground truth datasets listed in section 2.

4.5 Two-stage fine-tuning

In a second and third experiment, we investigate to which degree ZEFYS2025 is compatible with other historical datasets. We study whether the results improve when models are either previously fine-tuned on the ZEFYS2025 dataset or on the joined historical datasets mentioned in section 2 including ZEFYS2025. Final training is performed in a second stage with respect to the target dataset.

In the second experiment, we use the ZEFYS2025 training split for the first of the two consecutive NER fine-tunings. The training iteration was stopped if the F_1-score on the ZEFYS2025 validation split did not increase any longer for two consecutive epochs. Then we perform the second fine-tuning with the training split of the final target dataset, for instance HIPE 2020, until the F_1-score on the validation split of the final target dataset did not increase any more for two consecutive epochs. We report the resulting F_1-scores on the test split of the target dataset in table 5 in the $F_{1_{PZ}}^{TE}$ column.

In the third experiment we use the same setup as in the second one, but for the first fine-tuning stage we train on the joined training splits of all the historical datasets listed in section 2 including ZEFYS2025. We use the F_1-score on joined validation splits of all these datasets as early stopping criterion. The second fine-tuning stage is performed exactly as in the second experiment. The columns $F_{1_{PH}}^{TE}$ in table 4 and table 5 report the resulting F_1-score on the test splits of the respective datasets.

For all experiments we also provide the class-specific F_1-scores F_1^{PER}, F_1^{LOC} and F_1^{ORG} in table 3, table 4 and table 5. The class-specific scores presented in these tables relate in each row to the best-performing run, i.e., baseline, pretraining on ZEFYS2025 (PZ) or pretraining on joined historical ground truth (PH).

5 Discussion

Generally speaking, models pretrained with contemporary data perform best on contemporary datasets (RoBERTa, XLM-RoBERTa, GBERT, GermanBERT), whereas models pretrained with historical German perform better on the historical datasets (Europeana-ELECTRA, hmBERT). With respect to the GermEval 2014 dataset, German models (GermanBERT, GBERT) perform best, while multi-lingual models (RoBERTa, XLM-RoBERTa) perform best on the CoNLL-2003 dataset. On all tested datasets, smaller models (DistilBERT, hmBERT-mini, hmBERT-tiny) perform worse than larger models.

Throughout all experiments, the recognition performance for PER and LOC is significantly better than for ORG. This might be due to PER and LOC being relatively unambiguous entity classes, whereas ORG always has the smallest support and overlaps with LOC depending on the annotation guidelines. It might therefore be the most difficult class to learn. The only exception to the general rule is CoNLL-2003, where recognition performance for ORG is similar to PER and LOC. CoNLL-2003 has a strong support of the ORG class and is not affected by OCR errors, which might explain the observed high recognition performance of many models.

model	$\mathbf{F_1^{TE}}$	$\mathbf{F_{1_{PZ}}^{TE}}$	$\mathbf{F_{1_{PH}}^{TE}}$	$\mathbf{F_1^{PER}}$	$\mathbf{F_1^{LOC}}$	$\mathbf{F_1^{ORG}}$
HisGermanNER						
hmBERT	0.80	0.82	**0.84**	0.88	0.84	0.21
Europ-ELECTRA	0.80	**0.84**	0.81	0.86	0.86	0.15
GBERT	0.78	0.80	**0.81**	0.84	0.84	0.00
GermanBERT	0.77	0.79	**0.82**	0.81	0.87	0.16
XLM-RoBERTa	0.74	0.79	**0.82**	0.82	0.85	0.17
RoBERTa	0.71	0.76	**0.79**	0.81	0.82	0.00
DistilBERT	0.61	0.67	**0.74**	0.72	0.78	0.10
hmBERT-mini	0.60	0.68	**0.72**	0.73	0.76	0.00
hmBERT-tiny	0.30	0.51	**0.58**	0.51	0.67	0.00
Europeana-LFT						
Europ-ELECTRA	0.79	**0.80**	0.79	0.83	0.81	0.67
XLM-RoBERTa	**0.78**	**0.78**	**0.78**	0.80	0.80	0.68
hmBERT	0.76	0.76	**0.77**	0.79	0.79	0.68
RoBERTa	**0.76**	0.75	**0.76**	0.76	0.78	0.69
GermanBERT	0.73	0.73	**0.74**	0.77	0.75	0.62
GBERT	**0.74**	0.73	0.73	0.77	0.75	0.63
DistilBERT	0.65	**0.67**	0.66	0.63	0.73	0.59
hmBERT-mini	0.63	**0.67**	0.65	0.66	0.72	0.53
hmBERT-tiny	0.50	0.55	**0.58**	0.57	0.64	0.44
Europeana-ÖNB						
Europ-ELECTRA	0.84	0.86	**0.88**	0.82	0.94	0.44
XLM-RoBERTa	0.84	**0.88**	0.86	0.85	0.92	0.13
hmBERT	0.84	**0.86**	**0.86**	0.78	0.92	0.50
RoBERTa	0.81	0.84	**0.87**	0.81	0.92	0.20
GBERT	0.83	**0.84**	0.83	0.75	0.92	0.40
GermanBERT	0.82	0.83	**0.84**	0.73	0.92	0.53
hmBERT-mini	0.70	0.76	**0.78**	0.68	0.86	0.28
DistilBERT	0.70	0.74	**0.75**	0.61	0.86	0.15
hmBERT-tiny	0.50	0.63	**0.65**	0.48	0.78	0.38
HIPE 2020						
Europ-ELECTRA	0.76	**0.78**	0.75	0.75	0.84	0.57
GermanBERT	0.73	**0.76**	0.74	0.75	0.81	0.60
hmBERT	0.73	**0.75**	**0.75**	0.72	0.82	0.54
GBERT	**0.74**	**0.74**	**0.74**	0.72	0.81	0.48
XLM-RoBERTa	0.73	**0.74**	**0.74**	0.71	0.82	0.50
RoBERTa	0.67	**0.70**	**0.70**	0.68	0.77	0.50
DistilBERT	0.60	**0.62**	**0.62**	0.52	0.74	0.33
hmBERT-mini	0.57	0.58	**0.62**	0.50	0.74	0.38
hmBERT-tiny	0.40	0.48	**0.51**	0.36	0.64	0.31
Neiss-Arendt						
GermanBERT	**0.88**	**0.88**	**0.88**	0.94	0.83	0.74
hmBERT	0.87	**0.88**	0.87	0.94	0.85	0.71
GBERT	0.87	0.87	**0.88**	0.95	0.81	0.71
RoBERTa	0.86	**0.87**	**0.87**	0.92	0.83	0.72
DistilBERT	0.85	**0.87**	**0.87**	0.92	0.80	0.72
XLM-RoBERTa	0.86	0.86	**0.87**	0.93	0.85	0.69
Europ-ELECTRA	**0.86**	**0.86**	**0.86**	0.93	0.80	0.75
hmBERT-mini	0.85	**0.86**	0.84	0.91	0.84	0.70
hmBERT-tiny	0.79	**0.83**	**0.83**	0.89	0.82	0.64
Neiss-Sturm						
XLM-RoBERTa	0.90	**0.93**	**0.93**	0.95	0.88	-
hmBERT	**0.91**	**0.91**	**0.91**	0.94	0.83	-
Europ-ELECTRA	0.89	**0.92**	**0.92**	0.92	0.90	-
GermanBERT	**0.91**	**0.91**	0.90	0.93	0.86	-
GBERT	0.89	**0.91**	**0.91**	0.95	0.84	-
RoBERTa	0.88	**0.91**	**0.91**	0.95	0.86	-
DistilBERT	0.85	0.88	**0.90**	0.92	0.86	-
hmBERT-mini	0.82	**0.86**	**0.86**	0.90	0.79	-
hmBERT-tiny	0.71	0.75	**0.78**	0.79	0.74	-

Table 5: Comparison on different historical NER datasets. Column descriptions see table 4. PZ: Pretraining on ZEFYS2025. PH: Pretraining on joined historical NER datasets.

In the ZEFYS2025, Europeana-ÖNB and HIPE 2020 datasets, performance for LOC is significantly better than for PER, while performance for PER and LOC is similar in HisGermanNER and Europeana-LFT. NEISS-Arendt, NEISS-Sturm, CoNLL-2003 and GermEval 2014 exhibit a better performance for PER than for LOC. NEISS-Sturm is the only dataset that does not contain ORG labels, hence we did not report any results for this class there.

In most cases pretraining on ZEFYS2025 improves the results in comparison to a training that is only performed on a single dataset. The impact of pretraining with ZEFYS2025 is especially significant for smaller models, i.e., hmBERT-mini, hmBERT-tiny, DistilBERT. The datasets that benefit most from pretraining with ZEFYS2025 are HisGermanNER, Europeana-ÖNB and HIPE 2020. On Europeana-LFT, NEISS-Arendt and NEISS-Sturm, there are only minor improvements observable; the latter two have not been derived from historical newspapers.

The impact of pretraining on joined historical datasets including ZEFYS2025 is again most significant for smaller models, i.e., hmBERT-mini, hmBERT-tiny, DistilBERT. For all models, it is most beneficial on the HisGermanNER, Europeana-ÖNB and HIPE 2020 datasets. In most cases, the improvement is similar when training is performed on a single, combined dataset as compared to pretraining with ZEFYS2025. In some cases the results are even better than pretraining only with ZEFYS2025, in particular for small models (hmBERT-mini, hmBERT-tiny, DistilBERT).

6 Conclusion

We present a comprehensively annotated German NER and EL dataset created from historical newspapers. The ZEFYS2025 dataset exceeds the size of previously published historical German NER/EL datasets concerning the number of entities and links included. The contemporary datasets CoNLL-2003 and GermEval 2014 are larger, but do not contain EL information. We described the dataset creation and annotation processes in depth and explained how we reached annotation quantity *and* quality through a combination of computational methods and iterative intellectual revisions. By focusing on a core tag set supported by clear guidelines, we aimed at achieving maximum interoperability with similar resources.

In several experiments and dataset evaluations we were able to show the following contributions:

- NER results on ZEFYS2025 indicate data consistency for a broad range of models that is on par with or higher than that of the other historical datasets.
- ZEFYS2025 can in many cases successfully be used as pretraining data to improve the recognition performance for historical NER on other datasets.
- ZEFYS2025 is particularly useful for NER and EL in historical source material that is affected by OCR artefacts.

We believe that this dataset constitutes a valuable resource for a wide variety of information extraction tasks on German-language historical material, concerning not only NER, but also EL due to the large number of manually verified links to Wikidata. As a future research direction, ZEFYS2025 will allow us to conduct further analyses and evaluations of the previously developed system for historical EL, e.g. in a similar way to the HIPE Shared Tasks of 2020 and 2022. It is probably also of use to others in terms of benchmarking comparable systems in this specific domain, since the inadequacy or simply absence of resources has constituted an obstacle so far.

The ZEFYS2025 dataset and associated resources have been published under a CC-BY-4.0 licence on Zenodo[17] and GitHub.[18] We also release an open source tool for fine-tuning and evaluating NER models for German historical newspaper contents with the help of the Hugging Face Transformers library.[19]

Acknowledgments

The dataset was annotated by all authors; we acknowledge contributions by Markus Bierkoch (annotations) and Knut Lohse (provision of ground truth), both Staatsbibliothek zu Berlin – Berlin State Library.

Funding statement

This work was partially supported by the German Federal Government Commissioner for Culture and the Media (BKM), project grant no. 2522DIG002.

[17] https://doi.org/10.5281/zenodo.15771823
[18] https://github.com/qurator-spk/ZEFYS2025
[19] https://github.com/qurator-spk/sbb_ner_hf

References

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA).

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German‘s next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nancy Chinchor and Patricia Robinson. 1998. Appendix E: MUC-7 Named Entity Task Definition (version 3.5). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 – May 1, 1998.*

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. *Preprint*, arXiv:2003.10555.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Oksana Dereza, Theodorus Fransen, and John P. Mccrae. 2023. Do not Trust the Experts - How the Lack of Standard Complicates NLP for Historical Irish. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 82–87, Dubrovnik, Croatia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maud Ehrmann, Estelle Bunout, and Marten Düring. 2019. Historical Newspaper User Interfaces: A Review. In *Proceedings of the 85th IFLA General Conference and Assembly*, pages 1–24.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2):1–47.

Maud Ehrmann, Matteo Romanello, Stefan Bircher, and Simon Clematide. 2020a. Introducing the CLEF 2020 HIPE Shared Task: Named Entity Recognition and Linking on Historical Newspapers. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, volume 12036, pages 524–532. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Extended Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. Publisher: CEUR-WS.

Maud Ehrmann, Camille Watter, Matteo Romanello, Simon Clematide, and Alex Flückiger. 2020b. Impresso Named Entity Annotation Guidelines. Publisher: Zenodo Version Number: 2.2.0.

Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidere, Mickael Coustaty, and Antoine Doucet. 2019. An Analysis of the Performance of Named Entity Recognition over OCRed Documents. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 333–334, Champaign, IL, USA. IEEE.

Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. 2020. Assessing and Minimizing the Impact of OCR Quality on Named Entity Recognition. In *Digital Libraries for Open Knowledge*, pages 87–101, Cham. Springer International Publishing.

Ahmed Hamdi, Elvys Linhares Pontes, Emanuela Boros, Thi Tuyet Hai Nguyen, Günter Hackl, Jose G. Moreno, and Antoine Doucet. 2021. A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2328–2334, Virtual Event Canada. ACM.

Saul A. Kripke. 1980. *Naming and necessity*, rev. and enl. edition. Library of philosophy and logic. Blackwell, Oxford.

Kai Labusch and Clemens Neudecker. 2020. Named Entity Disambiguation and Linking on Historic Newspaper OCR with BERT. In *CLEF 2020*, pages 1–14.

Kai Labusch and Clemens Neudecker. 2022. Entity Linking in Multilingual Newspapers and Classical Commentaries with BERT. In *CLEF 2022*, pages 1–11.

Kai Labusch, Clemens Neudecker, and David Zellhöfer. 2019. BERT for Named Entity Recognition in Contemporary and Historical German. In *KONVENS. Conference on Natural Language Processing*, pages 1–9.

Quentin Lhoest, Albert Villanova Del Moral, Yacine Jernite, Abhishek Thakur, Patrick Von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Emanuela Boros, Ahmed Hamdi, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. 2020. Entity Linking for Historical Documents: Challenges and Solutions. In Emi Ishita, Natalie Lee San Pang, and Lihong Zhou, editors, *Digital Libraries at Times of Massive Societal Transition*, volume 12504, pages 215–231. Springer International Publishing, Cham.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Gary Munnelly and Seamus Lawless. 2018. Investigating entity linking in early english legal documents. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, page 59–68, New York, NY, USA. Association for Computing Machinery.

Clemens Neudecker. 2016. An open corpus for named entity recognition in historic newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4348–4352, Paris, France. European Language Resources Association (ELRA).

Chiara Palladino and Tariq Yousef. 2024. Development of robust NER models and named entity tagsets for Ancient Greek. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 89–97, Torino, Italia. ELRA and ICCL.

Vera Provatorova, Marieke van Erp, and Evangelos Kanoulas. 2024. Too Young to NER: Improving Entity Recognition on Dutch Historical Documents. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 30–35, Torino, Italia. ELRA and ICCL.

Lance A. Ramshaw and Mitchell P. Marcus. 1999. Text Chunking Using Transformation-Based Learning. In Nancy Ide, Jean Véronis, Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, volume 11, pages 157–176. Springer Netherlands, Dordrecht. Series Title: Text, Speech and Language Technology.

Frederick Reiss, Hong Xu, Bryan Cutler, Karthik Muthuraman, and Zachary Eichenberger. 2020. Identifying Incorrect Labels in the CoNLL-2003 Corpus. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 215–226, Online. Association for Computational Linguistics.

Susanna Rücker and Alan Akbik. 2023. CleanCoNLL: A Nearly Noise-Free Named Entity Recognition Dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8628–8645, Singapore. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing – NeurIPS 2019*, pages 1–5, Vancouver BC, Canada.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmbert: Historical multilingual language models for named entity recognition. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 1109–1129. CEUR-WS.org.

Yasunobu Sumikawa, Adam Jatowt, Antoine Doucet, and Jean-Philippe Moreux. 2019. Large Scale Analysis of Semantic and Temporal Aspects in Cultural Heritage Collection's Search. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 77–86, Champaign, IL, USA. IEEE.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, pages 1–11. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick Von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jochen Zöllner, Konrad Sperfeld, Christoph Wick, and Roger Labahn. 2021. Optimizing Small BERTs Trained for German NER. *Information*, 12(11):443.