From Oral History to Structured Data: The MalachNER Dataset

Christopher Brückner, Karin Roginer Hofmeister, Jiří Kocián, Pavel Pecina
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Praha, Czech Republic
{bruckner,hofmeisterova,kocian,pecina}@ufal.mff.cuni.cz

Abstract

We present MalachNER, a new multilingual dataset for Named Entity Recognition (NER) in oral testimonies of Holocaust survivors. MalachNER has been sourced from different archives and annotated based on comprehensive domain-specific guidelines refined by a collaboration of international experts. Covering 10 European languages, differs significantly from previously released datasets: It is primarily based on noisy, verbatim transcribed speech, rather than on digitized written documents. These transcripts are characterized, among other challenges, by fillers, dialectal speech, and in-line annotations indicating incomprehensible words, which are not commonly encountered in other datasets. However, large volumes of yet unprocessed oral history make such a dataset a necessity. In addition to the description of the dataset and its annotation guidelines, we show with baseline experiments that MalachNER is complementary with previously released data, and the key to training domain-specific language models that generalize well to written and oral testimony alike, achieving state-of-the-art performance on both types of documents.

Keywords: Named Entity Recognition, Multilingual, Dataset, Speech, Holocaust Testimony

1. Introduction

While Named Entity Recognition (NER) is a well-known language processing task, it is still relatively unexplored in speech (Caubrière et al., 2020) and in the historical domain (Ehrmann et al., 2023). In particular, annotated language resources related to Nazi persecution are scarce (Carter et al., 2022; Anuradha Nanomi Arachchige et al., 2023) despite the existence of enormous archives such as the USC Shoah Foundation's Visual History Archive[1].

These archives are primarily based on oral testimony of Holocaust survivors. The nature of these documents introduces several challenges, since speech transcripts, regardless of whether they have been transcribed manually or automatically, include speech artifacts that are missing in most NER datasets. These artifacts include, for example, filler words and in-line annotations denoting inaudible words. Such noise cannot always be reliably removed automatically, which is amplified by the variety of artifacts increasing with the number of languages, source archives, and transcriptionists. The necessity of a dataset providing these challenges arises from the fact that oral history is an important resource for Holocaust studies (Vrzgulova, 2024), and manually curated transcriptions are often not available (Lehečka et al., 2023).

In the following, we present MalachNER, a multilingual NER dataset of manually transcribed Holocaust survivor testimonies sourced from different archives, annotated by domain experts who are speakers of Croatian, Czech, Danish, Dutch, English, German, Hungarian, Polish, Serbian, and Slovak. Finally, we compare different domain-specific NER models and show that the processing of speech-transcribed documents is challenging for language models trained exclusively on written testimony, and vice versa; but state-of-the-art performance on both types of documents can be achieved by a model fine-tuned simultaneously on written and oral history.

MalachNER is published under a closed license for academic usage on LINDAT[2]. The source code of the NER experiments is available on GitHub[3], and the best model produced by these experiments is available on Hugging Face[4] under the MIT license. The annotation guidelines are included in the dataset repository and additionally mirrored[5] for open access.

2. Related Work

The most notable multilingual NER dataset specific to Holocaust testimony is EHRI-NER (Dermentzi and Scheithauer, 2024). The documents in this dataset have been repurposed from the EHRI Online Editions[6], which consist of digitized written documents that were not originally annotated for

[1] https://vha.usc.edu

[2] http://hdl.handle.net/20.500.12800/1-6129

[3] https://github.com/chbridges/malach-ner

[4] https://huggingface.co/ufal/xlm-roberta-ehri-malach-ner

[5] https://ufallab.ms.mff.cuni.cz/~bruckner/htres2026/

[6] https://www.ehri-project.eu/ehri-online-editions/

Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026, pages 59–65
11 May 2026.

Source	Type	Example
Mlynář (2016)	pause	This was my first sibling. {{....}} You want to continue about my siblings?
	uncertain word	Ah... it was in in {{Kazinci}} street...
	background noise	My fathers name was {{Video stops for a moment}}, mother was Regina...
VHA	incomprehensible	We called them [NON-ENGLISH].
	long pause	And [PAUSES FOR 3 SECONDS] they went to Sweden.
	short pause	An ex– a– a– a Pole, a Christian Pole came back from the United States.
	uncertain word	I had [? otherwise ?] impression that the relationship was good.
FU Berlin	incomprehensible	Now, I think that living in this_. We still went to school.
	pauses, comments	(-) And so, (–) in 1944, uh, winter 43 [1943] my mother became very ill.
	video description	<end of tape 1>

Table 1: Examples of noise appearing in different transcripts.

NER, but for Entity Linking. As such, the EHRI-NER also tags generic non-named entities such as "barracks" or "family", given that these entities can be disambiguated within the context of the document they appear in. EHRI-NER extends the common tags Person, Organization, and Location, with the domain-specific tags Ghetto, Camp, and Date. Anuradha Nanomi Arachchige et al. (2023) proposed an iterative hybrid annotation approach for the annotation of Holocaust-specific name entities in digitized English-only Holocaust testimonies using a granular domain-specific entity type ontology, including specific tags such as Warships and Rivers. Ehrmann et al. (2023) point out the arising challenges in historical documents, including different types of noise introduced by digitized written documents and transcribed speech, as well as historical language and entities not accounted for by language models trained on modern text.

3. Dataset Description

3.1. Data Source and Collection

3.1.1. Testimony archives

Most of the testimonies appearing in MalachNER originate from the Visual History Archive (VHA) and have been manually transcribed and published either by the USC Shoah Foundation itself or by Freie Universität Berlin within the "Zeugen der Shoah" project[7]. Ten languages have been chosen based on the availability of domain experts who speak these languages.

The transcripts taken directly from the VHA include interviews with Mark Verstandig (English, longest VHA interview), Walter Guttmann (Dutch), Ruth Felix (Czech), and Halina Elczewska (Polish). The transcripts sourced from FU Berlin include interviews with Simon Wiesenthal (German, exceptional relevance), Lajos Erdélyi (Hungarian), Softic Sadrudina Gavrankapetanovic (Croatian), and Branislav Ackovic (Serbian). Additionally, the longest transcribed Danish interview, with Rosalin Christensen, has been provided by Mlynář (2016). Most of these testimonies have been selected by a domain expert according to different factors such as the length, relevance to the domain, and the density of named entities. For Croatian and Serbian, no other manual transcripts are available.

Three additional short interviews in Czech, with Karel Blahouš, Maria Kotrbáčková, and Drahomíra Blosgebrová, have been obtained from the United States Holocaust Memorial Museum (USHMM)[8] to match the proportion of Czech with the other languages and increase the variety of speakers. A Slovak interview with Rozália Guttmannová has been provided by the Milan Šimečka Foundation[9].

The dataset contains the transcription of 13 testimonies of approximately 37 hours of speech in total. Each testimony is split into "tapes", where each tape covers approximately 30 minutes of speech. Except for the significantly longer English and German interviews and the slightly smaller Croatian and Serbian data, the languages are proportional in terms of tokens.

3.1.2. Preprocessing

The transcripts mentioned in Section 3.1.1 have varying degrees of noise. The Slovak transcript has been carefully curated by the Milan Šimečka Foundation and does not, in fact, feature any speech artifacts or in-line annotations, and interruptions or inquiries by the interviewer are scarce. The Czech interviews from USHMM are more conversational, but speech artifacts in their transcripts are limited.

The remaining interviews, on the other hand, have been transcribed in much more detail, reproducing the original speech more faithfully, annotating pauses, and containing markers for incomprehensible utterances or background noise, as well as comments. Similar to language-dependent filler words, in-line annotations can switch languages and contain typos. This is particularly extreme in the FU Berlin transcripts, where not all noise can be

[7] https://transcripts.vha.fu-berlin.de

[8] https://collections.ushmm.org

[9] https://nadaciamilanasimecku.sk

removed safely. A comprehensive list of noise examples is given in Table 1. Generally, all noise that can be removed safely has been removed or substituted with simple regular expressions. However, sufficient speech artifacts, such as filler words and repetitions, remain in the text to reproduce speech more accurately and make the task significantly more challenging than EHRI-NER (Dermentzi and Scheithauer, 2024), which is primarily based on digitized written testimony and protocols.

3.2. Annotation Guidelines

After preprocessing, domain experts have annotated the testimony transcripts in LabelStudio (Tkachenko et al., 2020-2022) with a combination of the general-domain entity types defined by CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) and additional domain-specific entity types defined by EHRI-NER (Dermentzi and Scheithauer, 2024). The resulting dataset contains the following types:

- **PER (Person)** Names of identifiable individuals and families, not including titles.

  E.g., ~~Dr~~ György Neuhauser, Führer, Novákovi (family), Petrův ("Petr's", possessive adjective)

- **ORG (Organization)** Named groups, institutions, or firms.

  E.g., SS, Nazis, Czechoslovak Army, Germans

- **LOC (Location)** Geographic locations, including states, cities, names of temples, synagogues, and rivers.

  E.g., Vienna, Kazinczy Street, Danube, Austro-Hungarian Empire

- **CAMP** Nazi Camps (e.g., concentration camps, extermination camps, transit camps). Used as a more specific subtype of LOC.

  E.g., Birkenau, Auschwitz, Theresienstadt family camp (section of Auschwitz II-Birkenau)

- **GHETTO** Jewish Ghettos. Used as a more specific subtype of LOC.

  E.g., Theresienstadt, Budapest (if referring to one of the ghettos in the city)

- **DATE** Calendar dates and years mentioned in the text. Does not include underspecified times such as "February" or "Monday".

  E.g., 11th February 1940, 13.4.44, '42

- **MISC** Entities that cannot be assigned to any other tag, such as historical events, ethnicities, religious groups, ideologies, or languages.

  E.g., Reichspogromnacht, Mein Kampf, Germans, Jewish, Holocaust, Zyklon B

Additionally, a TERM category for Holocaust-specific terminology such as "gas chambers" has been defined. This category has not been used for annotations, but helped annotators disambiguate between MISC entities and generic terms that do not refer to specific named entities. The EHRI vocabularies[10] have been used as a help to disambiguate LOC, CAMP, and GHETTO.

The annotation guidelines are largely based on guidelines composed for NER and Entity Linking annotations of testimonies from the Wiener Holocaust Library and the Hungarian Jewish Museum and Archives[11] (see Dermentzi and Scheithauer, 2024), but have been adapted to exclusively NER and actively expanded in collaboration with the annotators to resolve as many ambiguities as possible.

3.3. Dataset Statistics

10/23 English and 4/7 Czech tapes have been annotated by additional annotators. As suggested by recent literature (Mayhew et al., 2024), we measure the inter-annotator agreement by computing the F_1 scores of annotated spans using the `seqeval` framework (Nakayama, 2018). Czech has a high overall agreement of 95%, with perfect agreement on Person and Ghetto entities. The agreement on English is lower (81%), since the additional annotator processed the documents when the guidelines were still in an early state and many ambiguities were unresolved. In particular, the lowest agreements on ORG (68%) and MISC (78%) stem from the ambiguity of terms such as "Germans", which can refer to the German military or ethnicity.

The produced dataset has been tokenized with UDPipe (Straka, 2018), since by default, LabelStudio tokenizes the documents in a non-standard way, without splitting punctuation marks from words and thus tagging them as parts of entity spans. Table 2 shows the distribution of tokens and entities for each language. MalachNER is slightly smaller than EHRI-NER and has a smaller entity density stemming from the noise present in the speech transcripts. The distribution of entity types is similar, although MalachNER has fewer tagged dates and a significant lack of ghettos, as they are rarely mentioned by name. This discrepancy is possibly due to the fact that in EHRI-NER, also non-named entities are tagged. Secondly, different geographically conditioned trajectories of the Holocaust might have this effect.

The annotated dataset is split into training and test splits for each language. These splits have not

[10] https://portal.ehri-project.eu/vocabularies

[11] https://www.ehri-project.eu/call-for-applications-unlocking-holocaust-testimony-ehri-clarin-datathon-workshop/

Language	Tokens	PER	ORG	LOC	CAMP	GHETTO	DATE	MISC	Total
Croatian	17,405	137	60	211	0	3	62	171	644
Czech	27,751	235	50	263	18	18	44	306	934
Danish	26,434	24	83	113	102	7	27	159	515
Dutch	23,369	157	119	338	80	2	114	188	1,028
English	96,279	621	381	879	12	0	170	1,230	3,293
German	75,132	558	342	638	123	0	169	695	2,525
Hungarian	25,402	205	58	311	27	0	70	327	998
Polish	23,403	187	76	93	23	15	27	57	478
Serbian	15,615	53	47	124	3	0	26	44	297
Slovak	24,710	110	105	152	87	3	66	241	764
Total	355,500	2,287	1,321	3,122	475	48	805	3,418	11,476

Table 2: The number of tokens and annotated entities per language. The last column denotes the total number of entities.

been sampled randomly: First mentions of camps and ghettos commonly appear in the second or third tape of a testimony. Thus, every fifth tape, starting with tape 2, serves as test data, resulting in test ratios between 20% and 30% per language. For instance, from an interview with 23 available tapes, tapes 2, 6, 11, 16, and 21 are selected as the test split, resulting in a 22% test ratio.

4. Baseline Experiments

4.1. Experimental Setup

Two architectures are considered as baseline models: `XLM-RoBERTa-large` (Conneau et al., 2020) and `XLM-RoBERTa-ehri-ner-all`[12], which is `XLM-RoBERTa-large` fine-tuned on the EHRI-NER dataset (Dermentzi and Scheithauer, 2024). The models are evaluated in two experiments:

1. `XLM-RoBERTa-ehri-ner-all` is evaluated on the EHRI-NER and MalachNER test sets. Then, it fine-tuned further on the MalachNER training set, and re-evaluated on both test sets.
2. `XLM-RoBERTa-large` is fine-tuned from scratch on the training splits of EHRI-NER, MalachNER, and on both datasets at once. Each time, it is evaluated on the test splits of both datasets.
3. The best model fine-tuned on both datasets is further evaluated on each individual language in both datasets.

Hyperparameters are the same as those used by Dermentzi and Scheithauer (2024): Models are trained for 3 epochs using a learning rate of 3e-5 with weight decay at 0.01 and a batch size of 16. In addition to a seed of 42, the training is repeated with seeds 0 and 1234 to report 95% confidence intervals of the mean F_1 scores. When fine-tuning on MalachNER, 20% of the sentences in the training set are sampled with a fixed seed of 42 to create a held-out development set. When evaluating on EHRI-NER, the predictions of the additional MISC entity type are removed.

[12] https://huggingface.co/ehri-ner/xlm-roberta-large-ehri-ner-all

4.2. Results

For the first two experiments, the 95% confidence intervals of tag-wise and overall F_1 scores are reported in Table 3. The F_1 scores of the per-language evaluation of the best model are reported in Table 4. The different nature of the two datasets becomes obvious in Table 3: Models trained only on EHRI-NER perform badly on the speech transcripts of MalachNER, whereas models trained only on MalachNER perform badly on EHRI-NER. While the continued fine-tuning of `XLM-RoBERTa-large-ehri-ner` on MalachNER leads to the best models on MalachNER, in particular for the recognition of ghettos, the F_1 scores on EHRI-NER drop significantly, indicating the forgetting of earlier learned concepts.

The best possible solution to handle written and oral testimony at once is to combine both datasets and sample from both during fine-tuning: The resulting models achieve F_1 scores comparable with the best models fine-tuned only on one dataset for all entity types except for organizations and, in the case of MalachNER, ghettos. However, the confidence intervals suggest that the improvements are not significant in most cases. The lower score for organizations can be explained by the ambiguity with non-organizational groups of people, such as ethnicities and religious groups, which are covered by the added MISC type in MalachNER.

In addition to ORG, the GHETTO tag stands out in the MalachNER experiments with low F_1 scores and high variance. Here, the results benefit more from fine-tuning on the EHRI-NER training set than on the MalachNER training set. This can be explained by the lack of ghettos mentioned by name in the sourced testimonies, which was also shown

		XLM-R-large-ehri-ner		XLM-RoBERTa-large		
		Frozen	**Fine-tuned**$_M$	**Fine-tuned**$_E$	**Fine-tuned**$_M$	**Fine-tuned**$_{EM}$
EHRI-NER	PER	87.00	82.00 ± 0.00	86.00 ± 2.48	*74.33 ± 1.43*	**86.67 ± 1.43**
	ORG	63.00	41.33 ± 5.17	**65.33 ± 1.43**	*33.67 ± 1.43*	62.33 ± 1.43
	LOC	82.00	67.67 ± 1.43	**82.00 ± 2.48**	*59.67 ± 3.79*	**82.00 ± 2.48**
	CAMP	70.00	62.33 ± 3.79	**73.33 ± 1.43**	*41.00 ± 13.14*	72.00 ± 4.30
	GHETTO	80.00	77.67 ± 2.87	84.33 ± 2.87	*76.00 ± 17.39*	**85.00 ± 2.48**
	DATE	84.00	66.67 ± 14.34	**86.67 ± 1.43**	*49.67 ± 23.61*	81.33 ± 5.17
	Overall	81.00	67.67 ± 3.79	**82.00 ± 0.00**	*58.00 ± 4.97*	81.00 ± 2.48
MalachNER	PER	79.00	**90.67 ± 1.43**	*83.00 ± 2.48*	**90.67 ± 1.43**	**90.67 ± 1.43**
	ORG	29.00	**72.67 ± 5.17**	*25.00 ± 2.48*	72.33 ± 5.17	70.00 ± 4.30
	LOC	71.00	**90.00 ± 0.00**	*66.67 ± 2.87*	89.67 ± 1.43	89.00 ± 0.00
	CAMP	66.00	74.00 ± 2.48	*69.33 ± 14.56*	75.33 ± 7.17	**77.67 ± 6.25**
	GHETTO	67.00	67.00 ± 23.96	**71.00 ± 6.57**	*55.67 ± 1.43*	68.00 ± 0.00
	DATE	50.00	83.67 ± 7.99	*53.33 ± 5.17*	81.33 ± 6.25	**84.33 ± 3.79**
	MISC	–	–	–	**84.67 ± 1.43**	*83.00 ± 0.00*
	Overall	66.00	**85.67 ± 1.43**	*64.33 ± 1.43*	84.67 ± 1.43	84.33 ± 1.43

Table 3: Mean F_1 scores and 95% confidence intervals of different models fine-tuned and evaluated three times on EHRI-NER and MalachNER. The subscripts *E* and *M* denote fine-tuning on EHRI-NER and MalachNER, respectively, whereas *EM* denotes that the model samples from both datasets during fine-tuning. Note that the column **Fine-tuned**$_E$ reproduces the frozen XLM-RoBERTa-large-ehri-ner model with different initial seeds. **Best** and *worst* results of the newly fine-tuned models are marked.

Test Split		PER	ORG	LOC	CAMP	GHETTO	DATE	MISC	Overall	Support
EHRI-NER	cs	0.93	0.43	0.81	0.72	0.87	0.89	–	0.82	536
	de	0.85	0.61	0.82	0.79	0.87	0.80	–	0.80	802
	en	–	–	1.00	–	–	–	*0.00*	0.67	1
	fr	–	–	1.00	–	–	–	–	1.00	1
	hu	0.94	0.67	0.71	0.79	–	0.82	*0.00*	0.75	79
	nl	1.00	0.89	1.00	–	–	–	–	–	9
	pl	0.84	0.80	0.79	0.73	0.67	0.62	–	0.77	117
	sk	1.00	1.00	0.95	–	–	–	–	0.97	17
	yi	0.73	0.47	0.77	0.19	0.33	0.00	–	0.71	217
MalachNER	cs	0.87	0.82	0.64	0.96	1.00	0.77	0.92	0.86	287
	da	0.86	0.43	0.90	0.90	0.00	0.88	0.66	0.74	186
	de	0.93	0.72	0.87	0.76	–	0.82	0.79	0.83	554
	en	0.92	0.67	0.97	*0.00*	–	0.97	0.91	0.91	716
	hr	0.94	1.00	1.00	–	1.00	0.75	0.81	0.87	76
	hu	0.95	0.82	0.95	1.00	–	0.80	0.78	0.85	134
	nl	0.93	0.82	0.85	0.63	0.00	0.76	0.65	0.80	288
	pl	0.85	0.60	0.65	–	0.57	0.60	0.20	0.65	90
	sk	1.00	0.98	0.81	0.70	1.00	0.92	0.88	0.89	231
	sr	0.82	0.36	0.96	0.00	–	–	0.74	0.81	71

Table 4: F_1 scores of the best NER model for all languages in both datasets. The last column shows the total number of annotated entities in the test split. Note that the published EHRI-NER test splits are not representative for all languages. Scores of *0.00* in *italics* indicate that the model predicted an entity type not present in the test split, which is generally the case for the MISC type in EHRI-NER. In all other cases where the score is 0.00, the test split contains only up to 3 instances of the corresponding entity type.

by Dermentzi and Scheithauer (2024) in the EHRI-NER dataset. As can be seen in Table 4, most of the test data lacks this entity type. Special attention should be given to this tag during model selection, and this problem can likely only be solved with additional data containing more annotated named ghettos. The surprisingly low scores for Polish in MalachNER indicate inconsistencies in the annotations either within MalachNER or with the Polish EHRI-NER data.

5. Conclusions

We presented MalachNER, a new Named Entity Recognition dataset based on transcribed oral testimonies in 10 languages. Expanding on entity ontologies defined by previous NER datasets, we compiled comprehensive domain-specific annotation guidelines and processed approximately 37 hours of speech with the help of domain experts. Baseline experiments show that MalachNER is complemen-

tary to existing datasets, and a model fine-tuned simultaneously on written and oral history can bridge the gap between these types of documents, achieving comparable state-of-the-art results on both despite the added challenge of noise.

On the other hand, NE-annotated Holocaust-related language resources are still scarce, which is reflected in the models' performance in tagging organizations and ghettos. This can be tackled in the future by acquiring more annotated data, but also by fine-tuning models that were pre-trained on large amounts of unannotated domain-specific data (Brückner et al., 2026). Furthermore, noise removal techniques can be explored to improve the data quality.

The MalachNER dataset, its annotation guidelines, experimental code, and the best model resulting from the experiments are published and can be accessed via the hyperlinks found in the introduction.

6. Limitations

While aiming to cover as many languages as possible, the number of languages has been limited by the number of available speakers of these languages. As a result, only one of the ten languages is represented by more than one speaker and thus exhibits more variance in speech. Furthermore, the amount of text per language has been limited by the availability of manual transcripts, as well as the time budget of the annotators, leading to a language imbalance in the created dataset. Due to licensing, the data is published under a closed license.

7. Acknowledgements

This project is funded by the European Union's Horizon Europe research and innovation programme under grant agreement No. 101061016.

Disclaimer: Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

The work described herein has been using tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

This research was partially supported by Charles University, project GA UK No. 380126 and SVV project number 260 821.

8. Bibliographical References

Isuri Anuradha Nanomi Arachchige, Le Ha, Ruslan Mitkov, and Johannes-Dieter Steinert. 2023. Enhancing named entity recognition for holocaust testimonies through pseudo labelling and transformer-based models. In *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, HIP '23, page 85–90, New York, NY, USA. Association for Computing Machinery.

Christopher Brückner, Jan Lehečka, Jan Švec, and Pavel Pecina. 2026. Modeling the language of holocaust survivors' testimony with domain-adapted transformers. In *Proceedings of the Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, Palma de Mallorca, Spain. ELRA.

Kirsten Strigel Carter, Abby Gondek, William Underwood, Teddy Randby, and Richard Marciano. 2022. Using ai and ml to optimize information discovery in under-utilized, holocaust-related records. *AI & SOCIETY*, 37(3):837–858.

Antoine Caubrière, Sophie Rosset, Yannick Estève, Antoine Laurent, and Emmanuel Morin. 2020. Where are we in named entity recognition from speech? In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4514–4520, Marseille, France. European Language Resources Association.

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.

Maria Dermentzi and Hugo Scheithauer. 2024. Repurposing holocaust-related digital scholarly editions to develop multilingual domain-specific named entity recognition tools. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 18–28, Torino, Italia. ELRA and ICCL.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2).

Jan Lehečka, Jan Švec, Josef V. Psutka, and Pavel Ircing. 2023. Transformer-based Speech Recognition Models for Oral History Archives in English, German, and Czech. In *Interspeech 2023*, pages 201–205.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Šuppa, Hila Gonen, Joseph Marvin Imperial, Börje F. Karlsson, Peiqin Lin, Nikola Ljubešić, LJ Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337, Mexico City, Mexico. Association for Computational Linguistics.

Jakub Mlynář. 2016. Pluralita identit v autobiografickém vyprávění československých Židů žijících v zahraničí. *HISTORICKÁ SOCIOLOGIE*, 2016:33–51.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

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.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Monika Vrzgulova. 2024. Open forum: Oral history in holocaust research. *Eastern European Holocaust Studies*, 2(1):151–157.