Exploratory Analysis of the Applicability of Formalised Knowledge to Personal Experience Narration

Victor Mireles, Stephanie Billib, Artem Revenko, Stefan Jänicke, Frank Uiterwaal, and Pavel Pecina

Abstract Some of the victims of Nazi prosecution have consigned their personal experiences in the form of diaries of their internment in concentration camps. Such human-centric texts may contrast with the organisation of knowledge about such events that, for example, historians and archivists make. In this work, we analyse six such narrations with the use of Entity Extraction and Named Entity Recognition techniques, present the results of the corresponding exploration, and discuss the suitability of such tools on this corpus. We show that knowledge tools, that have been successfully used to organise documents, can be lacking when describing personal accounts, and we suggest ways to alleviate this.

Keywords Named entity recognition · Entity extraction · Holocaust studies · Knowledge graphs · Digital humanities

1 Introduction

The preservation and study of the memory of crimes against humanity are necessary to maintain a societal consciousness of past atrocities, their organisation, and their preceding developments. Only in this way can future generations identify promptly and counteract any repetition of such tragedies. It is thus understandable and fortunate that the study of such events remains the focus of professionals and the general public.

In the particular case of the victims of Nazi prosecution, this study can start from firsthand accounts of the events, both retrospective and contemporary. These accounts are, for the most part, centred on the human experience of the tragedy and the everyday lives of the participants. This is in contrast with the historiographic tradition, which often focuses on longer periods of time or on events involving large groups of people.

Historiography relies heavily on the organisation of information, the clear definition of concepts, and the abstraction of events, people, and places into categories. With the advent of digital technologies, this organisation has taken on new forms and has been aided by the methods of Natural Language Processing and Information Extraction, among others.

V. Mireles (✉) · A. Revenko
Semantic Web Company GmbH, Vienna, Austria
e-mail: victor.mireles@semantic-web.com

A. Revenko
e-mail: artem.revenko@semantic-web.com

S. Billib
Bergen-Belsen Memorial, Lohheide, Germany
e-mail: stephanie.billib@stiftung-ng.de

S. Jänicke
Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
e-mail: stjaenicke@imada.sdu.dk

F. Uiterwaal
NIOD, Institute for War, Holocaust and Genocide Studies, Amsterdam, The Netherlands
e-mail: f.uiterwaal@niod.knaw.nl

P. Pecina
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
e-mail: pecina@ufal.mff.cuni.cz

P. Haber et al. (eds.), *Data Science—Analytics and Applications*,
https://doi.org/10.1007/978-3-031-42171-6_10

The creation of these formalised knowledge systems, in the form of ontologies, taxonomies, and knowledge graphs, is increasingly recognised as necessary for the organisation of large masses of sources. In the context of the modern period, knowledge graphs of different types have been built for describing, e.g., the Finnish winter and continuation wars [5], or the history of cities [8], as well as cataloguing important concepts in the history of WW2 and the Holocaust [12].

This structured information is typically made accessible to domain scholars through user interfaces, and the utility and impact of such frameworks [17] and their underlying methodologies are well documented in a recent survey on visualisation strategies for cultural heritage collections [18]. Casual users or domain experts are enabled to analyse large text corpora from different angles, follow down patterns they find interesting, and inspect details on data items suitably contextualised. In addition to such visual analytical environments that support information seeking, digital storytelling [16] has been proven beneficial for exploring and narrating cultural heritage data [9].

A large diversity of exploration means exists particularly for text-based data sources [6], including ego documents like letters [2] or diaries [7]. The latter project on exploring the diaries of the Bergen-Belsen concentration camp revealed the potential but also limitations of current approaches to quantitatively analyse such sensitive data resources. For example, state-of-the-art approaches for sentiment analysis, usually trained on modern texts [11], can hardly extract the emotion in these diaries. For the same reason, the accuracy of named entities depicting persons or geospatial references is insufficient and calls for new methods for automated named entity extraction.

There have been successful applications of Named Entity Recognition techniques which are specifically tuned for historical documents, such as hmBert [15] for German, English, French, Swedish, and Finnish, the work of Rovera et al. [14] for Italian, or that of Hubkova and Král [3] for Czech. Furthermore, systems that work beyond text and incorporate phonetic information for audio recordings have also been produced (e.g., Psutka et al. [13]). For a further survey focused on early-modern materials, the reader is referred to [4]. These methods and their successes can very well complement the formalised knowledge systems such as ontologies and taxonomies, to power the rich information organisation and exploration described above.

This work is an exploration of the commonalities and differences between human-centred accounts of daily life and the historiographic organisation of knowledge about the same historical context. We do this by analysing diaries written by people interned in a Nazi concentration camp using expert-generated vocabularies about WW2 and Named Entity Recognition tools.

2 Datasets

In this work, we process a corpus consisting of six diaries of people interned in the Bergen-Belsen concentration camp. Since life inside Bergen-Belsen is not well documented, prisoner-written texts like diaries are important sources. The different backgrounds (in terms of, for example, age, gender or nationality) of prisoners result in situations or events being described with different languages and interpreted in different forms. The wider the distribution of author backgrounds, the better equipped are historians to highlight the personal perspectives, and to estimate the details of the actual event. The six diaries chosen here cover a broad range of perspectives regarding nationality, gender, age, language, and duration. Among the authors are two female and four male authors, the youngest being 16 years and the oldest 46 years old at the time of liberation. Four authors were imprisoned in the same sub-camp (Star camp), while one author was a prisoner of the Hungarians' camp and one in the Special camp for Polish prisoners.

Three of the six diaries are kept until the time of the liberation in April 1945, the earliest of them starting in October 1943 and thus covering the longest time span of $2\frac{1}{2}$ years. The shortest diary covers four months only. All diaries are handwritten and original languages range from Dutch to Greek, German, Hungarian, Polish and Serbian reflecting the background of their author. Here, we use a digital version of said diaries, specifically of their German-language editions which comprise, in total, 228245 tokens. While only one of them was originally written in German, this is the language of researchers and visitors at the Bergen-Belsen memorial, and therefore most sources are translated by experts into German.

The diaries deal with daily life in the Bergen-Belsen concentration camp. Different aspects of camp life are described and interpreted from the perspective of the three different sub-camps showing different conditions and rules in each of them. A common topic in all the diaries is hunger and the lack of food as well as the social behaviour in the authors' surrounding. In all six diaries the vocabulary is informal but rich with general terms related to Nazi persecution and concentration camps. In addition, there are words used in Bergen-Belsen specifically or even in one of the sub-camps only.

We compare these human-centred documents with a formalisation of the domain done for historiographic purposes, the WW2 Thesaurus.[1] This thesaurus has, over the past decade, become the glue that holds together digital collections with a

[1] NIOD website on the WW2 thesaurus. URL: https://www.niod.nl/en/collections/ww2-thesaurus (last accessed on 3 January 2023).

thematic relation to the Second World War. It originated in 2016 when the Dutch NIOD Institute for War, Holocaust, and Genocide Studies transformed the collection of subject headings that it had built, into a semantic graph of meaningful relations. The thesaurus includes over two thousand entities including events, persons, organisations, camps, and juridical concepts [12]. It has been used, for example, to catalogue over a million historical sources in Oorlogsbronnen.[2]

The thesaurus currently contains 6,135 entities with 17,646 SKOS-XL[3] labels in Dutch, German, and English. More entities and labels are constantly added through a community process coupled with expert curation, and linking to Wikidata and Geocodes allows for interoperability with the wider Semantic Web. Because no two entities have the same label, this thesaurus can be used in entity extraction without the need for disambiguation.

The structure of the thesaurus is multi-hierarchical, with entities related to one another through broader/narrower relations indicating some notion of hypernymy. For example, events can be grouped by the year in which they happened, by the type of event (e.g., Attack, Evacuation, Mass executions), or by some other categorisation. This hierarchical structure has a depth of up to 8 levels deep in some cases. The top categories, known as Concept Schemes, are: Organisations, Events, Interment Camps, Places, and General Concepts. They contain, respectively 1431, 1489, 1629, 443, and 5676 different entities.

3 Methods

The exploratory analysis presented here is the result of applying two methods: Entity Extraction and Named Entity Recognition (NER). Entity Extraction (also known as Concept Extraction, or Entity Linking) is the task of identifying mentions of entities from a controlled vocabulary (the WW2 Thesaurus) in a text. In this work, entity extraction was done using a commercial tool[4] that is tolerant to basic inflections of the German language. Since in the thesaurus used in this work, it is often the case that entities have multiple (alternative) labels, entity extraction has the effect of normalising entity mentions. This process results in a set of *concept matches*, which indicate the id of the entity identified, the character offset where it appeared, and the text that the tool deemed as matching text.

For NER we used a German BERT model fine-tuned on the GermEval14 dataset [1] to recognise mentions of named entities. The model is publicly available[5] and the authors report strict F_1 scores of 86.89 and 85.52 on the manually annotated GermEval 2014 and German CoNLL-2003 test data-sets, respectively.[6] Although specialised models [15] have shown better results (in terms of the F_1 score) on historic data-sets, the chosen model is only a few points behind state of the art and outperforms specialised models in non-historic corpora. As our corpus contains documents written in modern German we opt for the German BERT model. The model recognises the “big four” NER classes: LOC (Locations), ORG (Organizations), PER (People), OTH (Other). The model also recognises nested entities, however, at this stage of analysis we considered only the larger of the nested entities. The *Named Entity Matches* which result from this process are not normalised, in the sense that two different matches might refer to the same real-world entity being mentioned using different names or, simply, different linguistic variants (e.g., declensions of case).

4 Results

The data-set consisting of the diaries of six inmates of the Bergen-Belsen concentration camp was analysed using two different approaches: entity extraction and named entity recognition. While the former is based on expert-curated knowledge in the form of a thesaurus, so that entities of interest to historians are identified, the latter aims at identifying entities based on the lexical patterns in text. We present the results from each of the methods, and compare them to find commonalities.

[2] Oorlogsbronnen website: https://www.oorlogsbronnen.nl/ (last accessed on 3 January 2023).

[3] Simple Knowledge Organization System–Extension for Labels. An ontology for denoting taxonomic relations between entities, with support for rich statements about their labels. https://www.w3.org/TR/skos-reference/skos-xl.html (published 18th of August 2009, last Accessed 3rd of April 2023).

[4] PoolParty Semantic Suite https://www.poolparty.biz/ (last accessed 3rd of April, 2023).

[5] https://huggingface.co/fhswf/bert_de_ner.

[6] https://github.com/stefan-it/fine-tuned-berts-seq.

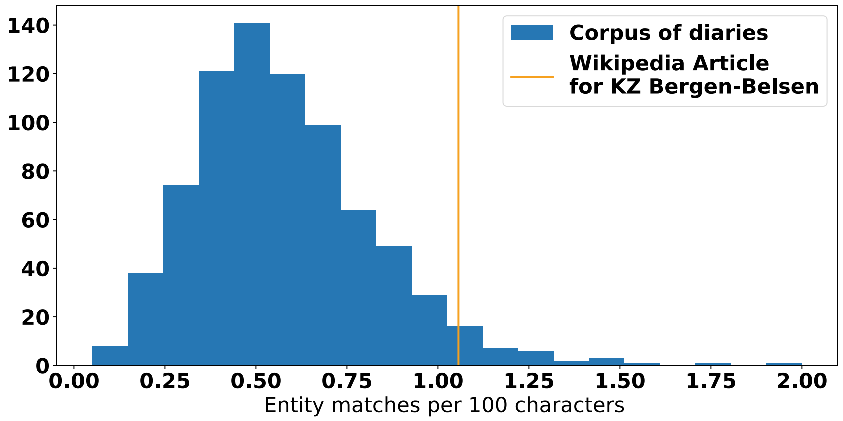

Fig. 1 **Histogram of entity matches per 100 characters, using the WW2 thesaurus.** The corpus was partitioned into overlapping windows of 1000 characters, and the number of entity matches in each was noted

Table 1 **Prevalence of entities from different concept schemes.** Shown are the number of concept matches found for every concept scheme, and in parenthesis, the number of distinct entities in said concept scheme found in each document. Note that the NIOD WW2 Thesaurus is a multi-hierarchy, so a given concept can belong to more than one concept scheme

Document	Internment camps	Organizations	Events	Concepts	Total
Diary 1	134 (13)	18 (7)	37 (10)	1429 (211)	**1618**
Diary 2	12 (2)	3 (2)	7 (2)	566 (116)	**588**
Diary 3	107 (11)	77 (6)	2 (2)	1807 (163)	**1993**
Diary 4	33 (7)	8 (3)	10 (3)	482 (115)	**533**
Diary 5	47 (7)	30 (3)	22 (7)	983 (146)	**1082**
Diary 6	32 (6)	0 (0)	43 (10)	3106 (221)	**3181**
Whole corpus	365 (20)	136 (12)	121 (19)	8373 (405)	

4.1 Entity Extraction

Entity extraction of the six diaries resulted in a total of 8428 concept matches. Given the corpus size, this corresponds to approximately 0.57 concept matches per 100 characters. We contrast this to the German-language Wikipedia article on Bergen-Belsen concentration camp,[7] which exhibits around 1.056 concept matches per 100 characters. See Fig. 1 for a comparison.

The concept matches were distributed among different concept schemes as seen in Table 1, with the most frequently mentioned entities of each shown in Table 2. In general, we see great overlap in the entities mentioned by the different authors.

4.2 Named Entity Recognition

Executing the Named Entity Recognition method on the corpus resulted in a total of 7217 occurrences of 2486 different named entities. These were distributed across entity types as shown in Table 3. Some of these were recognised as entities of different types, so that there were a total of 2355 different strings associated to a named entity.

[7] https://de.wikipedia.org/wiki/KZ_Bergen-Belsen last accessed 10th of February 2023.

Table 2 Most common concepts per concept scheme. Places are not shown since in the thesaurus these have only Dutch labels

Organisations	Concepts	Events	Internment camps
Durchgangslager Westerbork	Ostern	Warschauer Aufstand Beschusse	Lieben
Trawniki	Weihnachten	Ostfront	Vittel
Jewish Agency for Palestine	Generalgouvernement	Schlacht	Ku
Sonderkommando	Tauschhandel	Abmarsch	Grünberg
SS	Desinfektion	Offens	Fossoli
American Jewish Joint Distribution Committee	Entlausung	Luftkrieg	Theresienstadt
World Jewish Congress	Hygiene	Luftangriff	Drancy
Waffen-SS	Täter	Landungen	Durchgangslager Westerbork
Rex	Nachschub	Invasion	Dachau
Reichssicherheitshauptamt	Hunde	Machtübernahme	Oranienburg
Luftwaffe	Helfer	Luftschutzkeller	Treblinka
Germanen	Ghetto	Attentat	Sachsenhausen
	Arbeitslager	Abwerfen	Poniatowa
	Kapo	Abzug	Neuengamme

Table 3 Distribution of different types of entities in corpus documents

Document	LOC	ORG	OTH	PER	Total
Diary 1	280	105	141	429	**955**
Diary 2	45	3	4	22	**74**
Diary 3	152	28	32	169	**381**
Diary 4	93	11	14	217	**335**
Diary 5	192	21	17	63	**293**
Diary 6	245	16	23	598	**882**
Whole corpus	688	166	213	1419	2486

Table 4 Intersection between concept schemes and entity types

	LOC	ORG	OTH	PER
Concepts	19	3	6	6
Internment camps	18	0	1	3
Organisations	2	3	0	1
Events	0	0	0	0

4.3 Contrasts Between the Results of Both Methods

Of the 2355 different strings that correspond to named entities in the corpus, 45 were also recognised, by the entity extraction tool, as surface forms of an entity in the thesaurus. These 45 entities make up a total of 390 occurrences, and their classification according to named entity type and concept scheme is shown in Table 4. Given that the thesaurus has no concept scheme corresponding to people, any named entities recognised as class PER are errors of the NER tool. A close inspection reveals that, for example, the strings *Brande, Poniatowa* and *Drancy*, which are names of internment camps, are wrongly labelled as people. While such false positives from NER methods are known to happen, previous evaluations of the method used [10] suggest that at least 70% of recognised named entities are correct matches. Identifying the issues that arise in a particular corpus can help fine tune the method or post-process its results.

5 Conclusion and Next Steps

In this exploratory analysis, we used a thesaurus developed to describe a given historical period (WW2 and the Holocaust), to examine first-hand accounts of events from that time. We show that this thesaurus, which has been successfully used for cataloguing historical materials and is well suited for analysing historiographic texts, is less effective for describing texts narrating personal experiences. As evidence supporting this statement, consider the case of the 166 organisations that the NER method has detected in the corpus. Although the thesaurus has a concept scheme containing 1431 organisations, it only contains 3 out of those 166.

We believe this discrepancy exposes the need for augmenting formalised knowledge sources in order to treat documents of a more personal nature. The results also highlight the role that Named Entity Recognition can play in this augmentation process. However, it must be noted that NER results must always be curated by experts when used to enlarge a knowledge organization system (e.g., a thesaurus). This becomes clear in this work, as inspecting the results reveals that several demonyms are detected as locations.

In the next steps, we will create a knowledge graph that brings together the results of NER, the expert-curated vocabularies such as the one used here, and the large collections of entities from bibliographic authority controls and the linked open data cloud.[8]

Acknowledgements The work presented here is part of project MEMORISE. This project is funded by the European Union's Horizon Europe research and innovation programme under grant agreement No. 101061016.

References

1. Benikova, D., Biemann, C., Reznicek, M.: Nosta-d named entity annotation for german: Guidelines and dataset. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 2524–2531. European Language Resources Association (ELRA), Reykjavik, Iceland (2014)
2. Edelstein, D., Findlen, P., Ceserani, G., Winterer, C., Coleman, N.: Historical research in a digital age: Reflections from the mapping the republic of letters project. Am. Hist. Rev. **122**(2), 400–424 (2017)
3. Hubková, H., Kral, P.: Transfer learning for Czech historical named entity recognition. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pp. 576–582. INCOMA Ltd., Held Online (2021)
4. Humbel, M., Nyhan, J., Vlachidis, A., Kim, S., Ortolja, A.: Named entity recognition for early-modern textual sources: a review of capabilities and challenges with strategies for the future. J. Document. **6** (2021)
5. Hyvönen, E.: "sampo" model and semantic portals for digital humanities on the semantic web. In: Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020). CEUR-WS.org (2020)
6. Jänicke, S., Franzini, G., Cheema, M.F., Scheuermann, G.: Visual text analysis in digital humanities. In: Computer Graphics Forum, vol. 36, pp. 226–250. Wiley Online Library (2017)
7. Khulusi, R., Billib, S., Jänicke, S.: Exploring life in concentration camps through a visual analysis of prisoners' diaries. Information **13**(2) (2022)
8. Krabina, B.: Building a knowledge graph for the history of vienna with semantic mediawiki. J. Web Semant. **76**, 100771 (2023)
9. Kusnick, J., Andersen, N.S., Beck, S., Doppler, C., Koch, S., Liem, J., Mayr, E., Seirafi, K., Windhager, F., Jänicke, S.: A survey on visualization-based storytelling in digital humanities. In: Computer Graphics Forum. Wiley Online Library (2023) (in review)
10. Labusch, K., Kulturbesitz, P., Neudecker, C., Zellhöfer, D.: Bert for named entity recognition in contemporary and historical german. In: Proceedings of the 15th Conference on Natural Language Processing, pp. 8–11. Erlangen, Germany (2019)
11. Nandwani, P., Verma, R.: A review on sentiment analysis and emotion detection from text. Soc. Network Anal. Min. **11**(1), 81 (2021)
12. van Nispen, A., Jongma, L.: Holocaust and world war two linked open data developments in the netherlands. Umanistica Digitale **4** (2019)
13. Psutka, J., Švec, J., Psutka, J.V., Vaněk, J., Pražák, A., Ircing, P.: System for fast lexical and phonetic spoken term detection in a czech cultural heritage archive. EURASIP J. Audio Speech Music Process. 1–11 (2011)
14. Rovera, M., Nanni, F., Ponzetto, S., Goy, A.: Domain-specific Named Entity Disambiguation in Historical Memoirs, pp. 287–291 (2017)
15. Schweter, S., März, L., Schmid, K., Çano, E.: hmbert: Historical multilingual language models for named entity recognition. In: Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (eds.) Proceedings of the Working Notes of CLEF 2022–Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th–to–8th, 2022. CEUR Workshop Proceedings, vol. 3180, pp. 1109–1129. CEUR-WS.org (2022)
16. Segel, E., Heer, J.: Narrative visualization: Telling stories with data. IEEE Trans. Vis. Comput. Graph. **16**(6), 1139–1148 (2010)
17. Vassilakis, C., Kotis, K., Spiliotopoulos, D., Margaris, D., Kasapakis, V., Anagnostopoulos, C.N., Santipantakis, G., Vouros, G.A., Kotsilieris, T., Petukhova, V., Malchanau, A., Lykourentzou, I., Helin, K.M., Revenko, A., Gligoric, N., Pokric, B.: A semantic mixed reality framework for shared cultural experiences ecosystems. Big Data Cogn. Comput. **4**(2) (2020)
18. Windhager, F., Federico, P., Schreder, G., Glinka, K., Dörk, M., Miksch, S., Mayr, E.: Visualization of cultural heritage collection data: state of the art and future challenges. IEEE Trans. Vis. Comput. Graph. **25**(6), 2311–2330 (2018)

[8] https://memorise.sdu.dk/.