LLMs in Ottoman Turkish: From MLM to NER

Enes Yılandiloğlu

University of Helsinki, Finland
enes.yilandiloglu@helsinki.fi

Abstract

This paper introduces three foundational contributions to Digital Ottoman Turkish Studies. It presents: (1) three masked language models (MLMs) trained on over 11 million words from 144 works spanning from the 15th to 20th century, (2) a state-of-the-art Named Entity Recognition (NER) model (F1 = 89.94%) trained on 9,960 manually annotated entities, and (3) a state-of-the-art Universal Dependency (UD) parsing model for Ottoman Turkish. This work differs from others by deploying IJMES-transliterated documents for training and evaluation in order to prevent loss of information due to the change of the script from Perso-Arabic to Latin. The paper further explores probabilistic manuscript reconstruction in preliminary experiments, showing that MLMs can recover unread sections in historical documents with 77.8% top-1 accuracy when a list of candidate words is provided. Followed by a discussion, the paper outlines the future directions as building century-aware MLMs and expanding the training data across genres to enhance model generalization.

Keywords: Ottoman Turkish, Masked Language Models, Named Entity Recognition, Universal Dependencies, Low Resource Languages

1. Introduction

Ottoman Turkish Studies have been dominated by qualitative readings which often overlook computational approaches. When computational methods are leveraged, the scope is often to build Optical Character Recognition models (Taşdemir et al., 2024; Dölek & Kurt, 2023) or to apply statistical analysis (Öncel, 2020) and geo-spatial analysis (Aladağ, 2023), and engagement with MLMs remains almost untouched. Although many manuscripts have been digitized and their printed editions publicly released, most textual resources remain trapped in poorly encoded PDF formats or require laborious preprocessing. Moreover, Ottoman Turkish, when rendered in Latin script via the IJMES transliteration system (International Journal of Middle East Studies, n.d.), employs characters (e.g., ḳ and ġ) that are unsupported by conventional Turkish Natural Language Processing (NLP) pipelines, particularly for the tokenizers. To address these limitations, this work presents three principal resources for Digital Ottoman Turkish Studies. First, we train three MLMs using a corpus of 11,531,618 words from 144 literary works from poetry to prose and the 15th to 20th century. The models were trained on IJMES-transliterated documents rather than modern Turkish alphabet-based documents, which improves their performance in capturing the nuances of Ottoman Turkish. Second, the MLMs were evaluated on downstream tasks, Named Entity Recognition (NER) and Universal Dependencies (UD) parsing showing the superior performance of ota-xlm-roberta-base. Third, manuscript reconstruction in a probabilistic manner is explored leveraging the MLMs. Although the current models are not yet ready for scholarly use in manuscript reconstruction, they hold promise for reconstructing unreadable segments in historical Ottoman Turkish manuscripts in future. To make the language resources publicly accessible, we release all models and code demonstrating their use cases on Zenodo under a CC-BY-NC license[1].

2. Background

Since Devlin et al. (2018) introduced BERT, which started a new phase in computational linguistics, MLMs provide language representations for target language(s). Based on this architecture, MLMs have been developed for various historical low-resource languages such as ancient Greek (Singh et al., 2021), and old English (Harju & van der Goot, 2025). On the other hand, while several MLMs have been introduced in Turkish (Schweter, 2020), Ottoman Turkish remained underrepresented, except for a few MLMs. Aside from the models introduced in this work, HuggingFace, a hub for open-source NLP models, contains only 3 Ottoman Turkish MLMs. The MLMs presented in this study cover Ottoman Turkish with full length from 14th to 20th century. Furthermore, while there are a few NER models for Ottoman Turkish, such as the models trained on entities from only one book (İlter et al., 2025; Ünlü, 2025), which hinders the generalization ability of the model, the current work leverages six works spanning from the 15th to the 18th century as training data and has the highest F1 score. Additionally, such models utilized the modern Turkish alphabet to transliterate the Perso-Arabic original script. However, this work employs the Ottoman Turkish transliteration alphabet suggested by the IJMES Transliteration System (International Journal of Middle East Studies, n.d.), a standardized method for converting the Perso-Arabic script into the Latin alphabet while

[1] Models and code can be found here: 10.5281/zenodo.19043420.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3517–3522
11–16 May 2026.

preventing information loss. To illustrate, the IJMES transliteration distinguishes *ḥāl* (حال, meaning state, condition) from *ẖāl* (خال, meaning mole), which are written identically as hâl (or hal) in modern Turkish orthography. Without the IJMES transliteration alphabet, models simplify Ottoman Turkish orthography and orthographically equalize حال *and* خال, whereas the IJMES preserves the original phonological contrast as much as possible. Transliteration through IJMES is indeed an academic standard for corpora such as QHOD and journals like *Turkish Journal of Islamic Studies*, *Journal of the Muhyiddin Ibn Arabi Society*, and *Keshif*. The vast majority of Ottoman manuscripts and documents still remain in their original Perso-Arabic script, held as scanned images in library collections worldwide without any Latinized transcription. Providing the exact volume of IJMES-transliterated material is therefore difficult, as no comprehensive catalogue exists. Nevertheless, IJMES is the predominant standard among scholars who do produce Latinized Ottoman Turkish text, and as digitization and transcription efforts continue, the volume of IJMES-compatible text is expected to grow, making the NLP methods developed here directly applicable to an expanding body of material. Moreover, UD is a treebank project for the languages of the world in the same scheme (Nivre et al., 2016). There is no model for Ottoman Turkish that can be used to parse an IJMES-transliterated sentence for all required fields for the UD-style annotation task. Yet, UDPIPE (Straka, 2018) provides a UD parser for Ottoman Turkish data that is Latinized but not in the IJMES scheme. Consequently, this study fills the gap in Digital Ottoman Turkish Studies by providing the most comprehensive and state-of-the-art MLM and NER model, in addition to a model to be used for UD parsing. Furthermore, while various studies, such as Levine et al. (2024), focused on manuscript reconstruction for historical texts with computational methods, no previous study investigated Ottoman Turkish manuscripts as this study does, reconstructing manuscripts using MLMs.

3. Data

A heterogeneous corpus of over 11 million words was assembled from 144 Ottoman Turkish literary works, comprising both poetry and prose, composed between the fifteenth and twentieth centuries. Domain experts manually transliterated the original Perso-Arabic script into the IJMES Latin alphabet. All source materials were supplied as PDF documents, sometimes ill-encoded, requiring a robust extraction and normalization process.

3.1 PDF extraction and rule-based cleaning

PDF pages were processed via PyMuPDF Python library (McKie, 2025) to obtain the body text. Two deterministic heuristics were utilized to remove non-body elements. First, spans with font sizes equal to or smaller than 90% of the median font size on a page were classified as editorial footnotes and removed. Second, header lines were excluded if text strings located within the top 10% of a page were repeated on at least 30% of all pages in the document. Subsequently, the body text was concatenated into UTF-8 encoded .txt files.

3.2 Character normalization and token filtering

Despite expert transliteration, the source PDFs contained glyphs mapped to Private Use Area (PUA) codepoints or anomalous Unicode characters. Hence, a two-stage postprocessing was applied. Firstly, the data contained the original Ottoman Turkish letters. Since the aim is to train models for Latin-transcribed Ottoman Turkish, such phrases were automatically removed leveraging the encoding values. The words whose Arabic and Persian character ratio is higher than 0.2 were removed as this was enough to detect all Perso-Arabic script in the documents except badly transliterated words. Secondly, it was observed that characters in the PDF files were systematically misencoded. For instance, the letter ṣ was represented by the letters “ś”, “ã”, “§”, and “š” in different works. Therefore, such badly encoded characters were converted to their counterparts in the IJMES transliteration alphabet. Regex-based rules for 150 characters were manually prepared, and the files were postprocessed with the rules. Additionally, the characters “ú”, “*”, “ù”, “Ù”, “ê”, “ÿ”, “ä”, “å”, “á”, “ë”, “ï”, “µ”, and two badly encoded characters, in total 14 characters, were identified as being used to represent different characters in different documents, although there was always a consistent use in a single document. Hence, such characters were manually converted in each document. Lastly, the remaining noise was manually removed, although such cases were rare. This postprocessing process ensured a high-quality, noise-reduced corpus for model training.

3.3 Tokenizer coverage evaluation

One of the questions this paper seeks to answer is which tokenizer yields better results for Ottoman Turkish. Thus, the data after postprocessing was tokenized via three different tokenizers. Table 1 presents the results.

Table 1. Tokenizer coverage results.

Tokenizer	Token count	Unknown token count	Unknown token ratio (%)
bert-turkish-cased	29,556,369	2,160,552	7.31

(Schweter, 2020)			
xlm-roberta-base (Conneau et al., 2019)	38,217,516	710,484	1.86
mdeberta-v3-base (He et al., 2021)	38,732,566	0	0

The zero unknown-token ratio exhibited by mdeberta-v3-base underscores its successful alignment with IJMES-transliterated Ottoman Turkish. Table 1 also emphasizes the importance of the tokenizer in addition to the power of representation of a given MLM. The unknown tokens for the other two models are particularly due to the characters in the IJMES such as Ḥ, *ḥ*, *Ż*, and *ż* that are not used in many modern alphabets including the current Turkish alphabet.

3.4 MLMs

We finetuned three different MLMs using BERT-turkish-cased (Schweter, 2020), XLM-RoBERTa-base (Conneau et al., 2019), and mDeBERTa (He et al., 2021). Selection criteria included the size of the models, considering the size of the corpus, and their suitability for Turkish language. Meanwhile, we aimed to test whether multilingual or Turkish-specific models can yield better performance in Ottoman Turkish. After splitting the data 95/5 (training and validation set), the same hyperparameters were utilized for all three models during the training phase including a batch size of 32, 5 epochs, a linear learning scheduler, a learning rate of 0.00002, a warmup ratio of 0.06, and a weight decay of 0.01. The hyperparameters were decided after several experiments.

Once the training was done, the models were evaluated on downstream tasks, NER, UD parsing.

4. Results

The MLMs were evaluated on downstream tasks, particularly NER and UD parsing in addition to an experiment on manuscript reconstruction.

4.1 NER

For the NER task, 9,960 entities in 1,600 128-token chunks were manually annotated. The data was chosen from six sources written between the 15th and 18th centuries. The labeling scheme consists of four labels, PER, LOC, ORG, and MISC. Subsequently, three MLMs were finetuned for the NER task. The data was split as 80/10/10 and the same hyperparameters, linear learning scheduler, 0.00003 learning rate, 0.06 warmup ratio, 16 batch size, 0.01 weight decay, and 10 epochs, were applied to all three models. The results in comparison with other NER models can be seen in **Error! Reference source not found.**. The first three models were evaluated on our test set while others report scores from their own test sets; hence, the comparison should be interpreted with caution.

Table 2. Comparison of NER model performance.

Models	Precision	Recall	Span-level F1
ota-bert-turkish-cased	83.24	84.34	83.79
ota-xlm-roberta-base	88.99	**90.91**	**89.94**
ota-mdeberta-v3-base	86.88	89.18	88.01
MemoirNER-BERTurk (İlter et al., 2025)	-	-	82.56
ottoman-ner-latin (Özateş et al., 2025)	**89.4**	87.1	88.2

Table 2 reveals the state-of-the-art performance of ota-xlm-roberta-base model in terms of span-level F1 while Özateş et al.'s (2025) model yields the best result in terms of precision which means it can fit better for historical analysis. A significant difference between the first three models and the other two in Table 2 is that while the former utilized the IJMES transliteration alphabet, misaligning the data with modern Turkish orthography, the latter did not.

4.2 UD parsing

For the UD parsing task, the largest Ottoman Turkish treebank with 2,000 sentences from UD was utilized as training data (Yılandiloğlu, 2025). The training was applied with MaChAmp architecture (van der Goot et al., 2021). The default hyperparameters but with 50 epochs were deployed. Table 3 demonstrates the performance of the existing models for UD parsing.

Table 3. Performance of the models for UD Parsing.

Models	Dependency (LAS)	Lemmatization	Morph. analysis	UPOS	XPOS
ota-bert-turkis	64.78	85.69	76.16	89.93	90.81

h-cased					
ota-xlm-roberta-base	**70.81**	**91.30**	**83.02**	**93.63**	**93.01**
ota-mdeberta-v3-base	69.67	90.61	81.86	93.37	92.88
Straka (2018)	53.06	82.36	80.96	86.18	90.81

The result for UD parsing aligns with the NER results. ota-xlm-roberta-base outperforms the other two models for all tasks. ota-mdeberta-v3-base appears as the second-best model while ota-bert-turkish-cased performs the poorest. The UD parsing model by Straka (2018) was trained and evaluated on modern Turkish alphabet-based Ottoman Turkish data, making direct comparison with our IJMES-transliterated models infeasible due to the mismatch with characters. Their reported scores are therefore included for reference only and should not be interpreted as a controlled comparison.

4.3 Manuscript reconstruction

This work also demonstrates the use of MLMs to reconstruct the unreadable or lost historical Ottoman Turkish documents. For this section, we create a case where historians cannot read a piece of the manuscript; however, they have some candidates on what the text can be based on visible letters. Subsequently, MLMs evaluate which candidate fits the best for the context. We evaluate in two levels: (1) token-level, where only one token was masked in a 128-token chunk and (2) word-level with pseudo-log likelihood (PLL) (Salazar et al., 2020), a realistic manuscript reconstruction scenario where an entire word with more than two characters is masked. For both scenarios, we used the validation set consisting of 5,000 randomly sampled and unseen chunks of 128 tokens by masking one token and providing a list of 5 candidates for the masked token or word while only one of them is correct. The other 4 candidates were chosen among the most similar tokens or words based on bigram similarity. The results indicate that at the token level, ota-mdeberta-v3-base achieves the highest accuracy at 94.52% (MRR 0.9691), followed by ota-xlm-roberta-base at 94.12% (MRR 0.9652), and ota-bert-turkish-cased at 85.60% (MRR 0.9121). At the word level, ota-xlm-roberta-base leads with 77.80% accuracy (MRR 0.8697), closely followed by ota-mdeberta-v3-base at 76.56% (MRR 0.8646), while ota-bert-turkish-cased achieves 67.78% (MRR 0.7935).

To illustrate, we provide one successful and one incorrect word-level prediction from ota-xlm-roberta-base. In the first example, the correct word *reyḥân* was masked and *Reyḥân*, *Keyḥân*, *Seyḥân*, and *Ceyḥân* were chosen as candidates within the context of …*Var ṣafḥa-i ḫaddinde nigâruñ ḫaṭ-ı [___]…* (On the page/surface of the beloved's cheek, there is the *reyhani* script). The model correctly predicts *reyḥân* as the most probable word for this context. In the second example where the model fails, among the candidate words, *pâküñe* (to your purity), *pâkümi* (my purity), *ḫâküñ* (your soil/world), *bâküñ* (your fear), and *pâküñ* (your purity), the last is the correct word, *pâküñe* (to your purity) gets the highest probability for the context …*Egerçi ʿırḳ-ı [___] medḫali vardur necābetde…* (Actually, your pure lineage has an influence on nobility) while *pâküñ* (your purity) is the 2nd probable word.

The preliminary findings demonstrate the potential of MLMs to reconstruct Ottoman Turkish documents when a fragment of text cannot be read but the reader has some possible candidates for the missing part. These three masked language models yielded promising but insufficient performance for accurate manuscript reconstruction. Thus, seq2seq architectures such as T5 may be suited better for this task, and future work will therefore focus on exploring and refining such architectures.

5. Discussion

Our results reveal that although ota-mdeberta-v3-base yields 0% unknown token ratio, it is not the best model in downstream tasks, NER, UD parsing, and manuscript reconstruction. On the other hand, ota-xlm-roberta-base appears as a state-of-the-art MLM to finetune for NER and UD parsing for five fields, and the manuscript reconstruction task on word-level. This demonstrates the strength of ota-xlm-roberta-base for Ottoman Turkish NLP tasks.

The superior performance of ota-xlm-roberta-base on NER, UD parsing, and manuscript reconstruction tasks presumably results from its multilingual pretraining. Since Ottoman Turkish has often borrowed words and phrases from Arabic and Persian that are absent in modern Turkish, its cross-lingual capacity plays a key role in Ottoman Turkish NLP tasks. Additionally, this study reveals better suitability of multilingual models such as XLM-RoBERTa-base for Ottoman Turkish compared to Turkish-specific MLM, BERT-turkish-cased.

Another significant topic is that although the models presented in this study are trained on IJMES-transliterated documents, which hinders the capability of the model to transfer knowledge

from modern Turkish knowledge it has, our best model (F1=89.94) outperformed others (88.2 and 82.56) that did not utilize IJMES-transliterated documents, in the NER task, in terms of F1 score. This demonstrates that preserving the orthographic knowledge in Ottoman Turkish documents, despite making the data distant from modern Turkish spelling, models can still perform well with sufficient training data.

The advantage of IJMES transliteration becomes more significant when applying the model on the manuscript reconstruction task. For instance, *ḳ* (ق) and *k* (ك) indeed refer to different letters in Perso-Arabic script. Thus, the words *ḥaḳḳ* (حق, truth or right) and *ḥakk* (حك, to scratch or to rub) differ by this consonant. Models trained on modernized transliterations that collapse these letters into a single k cannot distinguish such minimal pairs, often leading to semantic confusion in prediction. By contrast, the IJMES transliteration preserves these distinctions and helps the model to predict the correct word.

These resources can immediately catalyze Ottoman Turkish scholarship to find new research questions and answers. For instance, NER models can be leveraged to build a co-occurrence network based on appearing in the same text to uncover the connections among individuals in a vast amount of data. Scholars can apply diachronic analysis to Ottoman Turkish language using the UD parsing model. Lastly, the manuscript reconstruction can facilitate scholarly efforts to read damaged or ill-written manuscripts. In short, the models presented in this paper can serve as foundational tools in Digital Ottoman Turkish Studies.

6. Future work and conclusion

This paper presents the state-of-the-art MLM resources for Ottoman Turkish using IJMES transliteration. We train and evaluate three MLM architectures (BERT, XLM-RoBERTa, and DeBERTa-v3) on a diverse corpus of 11.5 million words from 144 literary works spanning five centuries. Our contributions include: (1) three publicly available MLMs, with ota-xlm-roberta-base achieving the best performance across tasks; (2) a state-of-the-art NER system (F1=89.94%) trained on 9,960 manually annotated entities; and (3) a state-of-the-art UD parsing model for Ottoman Turkish. Preliminary experiments demonstrate the potential of MLMs for manuscript reconstruction, although practical implementation requires further development. These resources address a critical gap in NLP tools for Ottoman Turkish, enabling computational analysis of historical texts that have remained largely inaccessible to digital methods. By training on IJMES-transliterated text, our models preserve linguistic information lost in the modern Turkish alphabet conversions, making them particularly suitable for scholarly applications. Future work will focus on three directions: (1) developing seq2seq models for manuscript reconstruction; (2) creating century-aware models to capture diachronic language change; and (3) expanding training data across additional genres and sources. Integration with digital humanities platforms will facilitate adoption by historians, linguists, and literary scholars, advancing interdisciplinary research on Ottoman Turkish Studies.

7. Bibliographical References

Aladağ, F. (2023). Spatial humanities and Ottoman studies: The potential of geographic information system (GIS) for Ottoman urban and administrative history. *Kadim*, 0(5): 47–68. https://doi.org/10.54462/kadim.1153648

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116. http://arxiv.org/abs/1911.02116

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding [Preprint]. arXiv. https://arxiv.org/abs/1810.04805

Dölek, İ., and Kurt, A. (2023). Ottoman Optical Character Recognition with deep neural networks. *Journal of the Faculty of Engineering and Architecture of Gazi University,* 38(4): 2579–2593. https://doi.org/10.17341/gazimmfd.1062596

Harju, A., and van der Goot, R. (2025). How to age BERT well: Continuous training for historical language adaptation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025)*, pages 258–267. Association for Computational Linguistics. https://aclanthology.org/2025.loreslm-1.21

He, P., Gao, J., and Chen, W. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing [Preprint]. arXiv. https://arxiv.org/abs/2111.09543

International Journal of Middle East Studies. (n.d.). *IJMES transliteration system for Arabic, Persian, and Turkish* [Transliteration chart]. Cambridge University Press. Retrieved March 15, 2026, from https://www.cambridge.org/core/services/aop-file-

manager/file/57d83390f6ea5a022234b400/TransChart.pdf

Levine, L., Li, C., Bremer-McCollum, L., Wagner, N., and Zeldes, A. (2024). Lacuna language learning: Leveraging RNNs for ranked text completion in digitized Coptic manuscripts. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 61–70. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.ml4al-1.8

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC '16)*, pages 1659–1666. European Language Resources Association (ELRA). https://aclanthology.org/L16-1262/

Öncel, F. (2020). Rural Manufacturing in Mid-Nineteenth-Century Ottoman Countryside: Textile Workers in Three Plovdiv Villages. In E. Kabadayı and L. Papastefanaki (Eds.), *Working in Greece and Turkey: A Comparative Labour History from Empires to Nation States 1840-1940*. Berghahn Books, pp. 174–203.

Özateş, Ş. B., Tıraş, T. E., Adak, E. E., Doğan, B., Karagöz, F. B., Genç, E. E., and Bilgin-Taşdemir E. F. (2025). *Building foundations for natural language processing of historical Turkish: Resources and models* [Preprint]. arXiv. https://arxiv.org/abs/2501.04828

Salazar, J., Liang, D., Nguyen, T. Q., and Kirchhoff, K. (2020). Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.240

Singh, P., Rutten, G., and Lefever, E. (2021). A pilot study for BERT language modelling and morphological analysis for ancient and medieval Greek. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2021)*, pages 128–137. Association for Computational Linguistics.

Straka, M. (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. Association for Computational Linguistics. https://doi.org/10.18653/v1/K18-2020

Taşdemir, E. F. B., Tandoğan, Z., Akansu, S., Kızılırmak, F., Sen, M., Akcan, A., Kuru, M., and Yanıkoğlu, B. (2024). Automatic transcription of Ottoman documents using deep learning. In G. Sfikas and G. Retsinas (Eds.), *Document Analysis Systems*, vol 14994, pp. 422–435. Springer, Cham. https://doi.org/10.1007/978-3-031-70442-0_26

van der Goot, R., Üstün, A., Ramponi, A., Sharaf, I., and Plank, B. (2021). Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197. Association for Computational Linguistics.

Yılandiloğlu, E. (2025). DUDU: A Treebank for Ottoman Turkish in UD Style. In *The Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025)*, pages 74–79. University of Tartu Library.

8. Language Resource References

İlter, M., Onuç, E., Evecen, D., Erşahin, B., Özcan Gönülal, Y., Karabulut, S., Berci, İ., and Tekir, S. (2025). *MemoirNER-BERTurk: Named Entity Recognition for Ottoman Turkish Memoirs* [Machine learning model]. Hugging Face. https://doi.org/10.57967/hf/6141

Karagöz, F. B. (2025). *Ottoman-NER Latin: A Named Entity Recognition Model for Transliterated Ottoman Turkish* [Machine learning model]. Hugging Face. https://huggingface.co/fatihburakkaragoz/ottoman-ner-latin

McKie, J. X. (2025). *PyMuPDF* (Version 1.26.3) [Computer software]. Artifex Software, Inc.

Schweter, S. (2020). *BERTurk - BERT models for Turkish* (Version 1.0.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3770924

Ünlü, C. (2025). *BERTurk_HisTR_NER_v2* [Machine learning model]. Hugging Face. https://huggingface.co/cihanunlu/BERTurk_HisTR_NER_v2