NLP for preserving Torlak, a vulnerable low-resource Slavic language

Li Tang, Teodora Vuković

University of Zurich, Linguistic Research Infrastructure, Zurich, Switzerland
li.tang@uzh.ch, teodora.vukovic2@uzh.ch

Abstract

Torlak is an endangered, low-resource Slavic language with a high degree of areal and inter-speaker variation. In previous work, interviews were performed with Torlak speakers in Serbia, near the Bulgarian border, and the transcripts annotated with lemma and morphosyntactic descriptions at token level. As such token-level annotations facilitate cross-language comparison in the context of the Balkan Sprachbund, where multiple languages influenced Torlak over time, including Serbian and Bulgarian. Here, we aim to improve the prediction of morphosyntactic annotations for this low-resource language using the fine-tuning of large language models, comparing several predictive models. We also further fine-tuned the large language models for scoring the degree of 'Torlakness' of a sentence by labeling likely Torlak tokens, to facilitate the documentation of additional Torlak transcribed speech with a high degree of Torlak-style non-standard features compared to standard Serbian. Taken together, we hope that these contributions will help to document this endangered language, and improve digital access for its speakers.

1 Introduction

While the contemporary standard languages of Serbian, Bulgarian and Macedonian all belong to the South Slavic branch, their noun phrase and verbal systems diverge in crucial aspects from each other, as Macedonian and Bulgarian (but not Serbian) exhibit Balkan morphosyntactic features, traits characteristic of the Balkan linguistic area (Vuković et al., 2022; Vuković, 2021). From an areal point of view, they can also share a number of morphosyntactic innovations with neighboring non-Slavic languages. The resulting variation becomes most distinct in the Torlak dialects spoken in Southern Serbia (Vuković et al., 2022; Vuković, 2021). Stigma associated with the use of those Torlak dialects in Serbian society, together with a perception of lower socio-economic status, is a key factor that can complicate research and language conservation efforts (Vuković, 2021; Vuković, 2022; Vuković, 2024), as it applies pressure against the use of this dialect by Torlak speakers. Due to such pressures, Torlak is listed as an vulnerable language by the UNESCO (Salminen, 2010).

Our work builds on transcriptions of audio materials collected in narrative interviews with Torlak speakers in Southern Serbia, who live in villages near the border with Bulgaria, together with lemma and morphosyntactic annotations at token level, to help preserve knowledge about Torlak (Vuković, 2021; Vuković, 2020). Here, we aim to improve the predictive modeling of morphosyntactic annotations using large language models (LLM), comparing several models in this dataset.

Lemmatization is a related linguistic annotation task that labels words with their dictionary forms. This can be especially important for the documentation of morphologically complex, highly inflectional

Proceedings of the 31st International Conference on Computational Linguistics, pages 6338–6347
January 19–24, 2025. ©2025 Association for Computational Linguistics

languages with less standardized spelling, such as Torlak, where the relationship between surface forms and their dictionary forms are often opaque (Manning & Schütze, 1999; Vuković, 2021).

Part-of-speech (POS) tagging is another important concept in the documentation of morphologically complex languages. Besides its obvious use, that is, searching for words that belong to a certain POS class, POS tags can also be used, to some extent, to disambiguate lemmatization. It was shown that for Slavic languages with rich inflectional and derivational morphology, such as Bulgarian (one of the languages that influenced Torlak), working with a large number of more informative morphosyntactic descriptions, compared to a smaller number of POS tags, can capture important aspects of morphological complexity in such languages (Georgiev et al., 2012; Erjavec, 2010; Tomic, 2006). Here, we therefore try to improve the ability to predict such detailed morphosyntactic descriptions for each token in a sentence, in total 712 labels.

With LLM trained with both Torlak and Serbian data, we then aim to score the degree of 'Torlakness' of texts from this region, compared to standard Serbian, to facilitate the automated discovery of additional Torlak materials with a high degree of non-standard features. In previous work, a more fine-grained but also labor-intensive human expert-driven approach based on five distinguishing features to cluster Torlak speakers, by their degree of non-standardness compared to standard Serbian, was developed (Vuković et al., 2022). With this transformer-based modeling of the degree of 'Torlakness', linguistic expert knowledge of Serbian and Torlak is not required when searching for the most non-standard Torlak texts in a collection, therefore enabling a higher degree of automation in the annotation of Torlak texts.

We hope that these NLP-based contributions will aid in preserving this endangered language, and improve digital access for its speakers.

2 Related Work

Much research has been done on the historical and social processes of language contact in this region of Europe, specifically in the linguistic area which has been referred to as the 'Balkan Sprachbund' (Friedman, 2011; Friedman & Joseph, 2015; Lindstedt, 2000). In the history of this Sprachbund, the Torlak dialects have experienced influences from many neighboring languages over time, including Slavic (e.g. Serbian, Bulgarian, Macedonian) and non-Slavic languages (Vuković, 2021; Lindstedt, 2000). As the most non-standard varieties of Torlak can be unintelligible to standard Serbian speakers (Vuković, 2021).

In terms of the application of Artificial Intelligence (AI), specifically the use of Natural Language Processing (NLP), for Slavic languages, the SlavicNLP workshop in 2023 at EACL (the European chapter of the ACL) provides an example of recent progress made in NLP-enabled research in this field (Piskorski et al., 2023), on a variety of datasets, NLP tasks and research questions - while the field of NLP in recent years experienced a shift from statistical to neural modeling, with major performance gains across various prediction tasks often due to the use of pre-trained LLM (Goldberg, 2017; Zhou et al., 2020; Min et al., 2023; Ulcar & Robnik-Sikonja, 2020; Tang, 2020; Conneau et al., 2020)

However, despite the impressive recent progress made in NLP overall, most of today's NLP research is still focused on just 20 of the 7000 languages of the world, leaving the vast majority of languages understudied, as those remaining languages are then considered 'low-resource languages' in NLP-enabled research (Magueresse et al., 2020; Hedderich et al., 2021; Conneau et al., 2020). Typical attributes of such low-resource languages are that they are not only scarce in resources, but also less studied, less computerized, less privileged, and less commonly taught.

In this context, an adaptation of NLP approaches to the special challenges we face with low-resource languages could help to improve digital access for low-resource language speakers, help to prevent language extinction, increase access to knowledge expressed in that language, and facilitate academic research in this area, as NLP-enabled experiences with some low-resource languages may also help to deal with other low-resource languages in an analogous manner (Mangueresse et al., 2020, Wiemerslage et al., 2022; Hedderich et al., 2021; Conneau et al., 2020).

Detailed morphosyntactic annotation of low-resource languages with a large number of tags, as in Georgiev (2012) and Vuković (2021), can make a contribution to low-resource language documentation, as it provides rich information at token and sentence level that facilitates comparisons with neighboring languages, including part-of-speech (POS), gender, number, case, and mood aspects of morphosyntax. For such purposes, the MULTTEXT-East Version 4 specification for capturing morphosyntactic descriptions (MSD) has been developed by Erjavec (2010) and applied to different languages of the Balkan Sprachbund (Tomic, 2006), including Slavic languages that have influenced Torlak such as Serbian, Bulgarian and Macedonian.

With increasing automation of such fine-grained morphosyntactic annotation, low-resource language documentation can be facilitated and accelerated, as MSD also facilitate cross-lingual 'transfer learning' from well-resourced languages (with better coverage of relatively rare phenomena) to low-resource languages (Wiemerslage et al., 2022). Note that when assigning each token a MSD, the context of a word needs to be considered, by human annotators and in AI-based automation (Erjavec, 2010).

In this context, efforts have been made to develop predictive models that learn patterns from large sets of human expert MDS (and lemma) annotations, also to allow for a higher degree of automation in such low-resource language conservation efforts. While CRFs (Conditional Random Fields), a classic probabilistic modeling method (Sutton & McCallum, 2012), can provide good predictive modeling of MULTTEXT-East style MSD (Vuković, 2021; Ljubešić et al., 2016) for Slavic languages. However, in recent years, neural models have often been able to provide an improvement of predictive performance for NLP of low-resource languages, including an improved modeling of non-standard, highly variable features that can be a challenge in low-resource language projects (Wiemerslage et al., 2022). Such improvements have already been shown for the predictive modeling of MSD for Serbian and Croatian (Ljubešić, 2018).

In summary, in recent years, it was shown that neural LLMs have performed very well in many related NLP tasks, making them attractive language modeling options for academic research on low-resource languages (Koroteev, 2021; Chakkarwar et al., 2023; Conneau et al., 2020). In this context, transfer learning from languages with more resources and standardization to low-resource languages with higher variability is an attractive possibility, e.g. using patterns learned by LLM (Rybak, 2024; Conneau et al., 2020).

3 Data and Methods

3.1 Dataset

The dataset we used is based on a previously published corpus of the vulnerable Torlak dialect from Southeast Serbia, near the Bulgarian border (Vuković, 2021; Vuković, 2020). Expressing considerable variation in the use of non-standard features under the influence of standard Serbian (Vuković, 2022). Between 2015 and 2017, semi-structured interviews were conducted in the field, eliciting spontaneous speech in the form of long narratives about traditional culture and history (Vuković, 2021). The majority of speakers in this Torlak dataset are older people whose language represents the highly non-standard variety, as older people tend to use more non-standard, dialectal features (Vuković, 2021). The corpus comprises 500,697 tokens of semi-orthographic transcripts representing 80 hours of recording from locations evenly distributed across the Timok area of the Torlak dialect zone (Vuković, 2021).

For model training and evaluation, a Torlak dialect sample of 59,612 manually verified tokens with lemma and MSD annotations ('tor') was merged with an existing training set for standard Serbian ('sr') with such annotations, consisting of 70,971 tokens, creating the 'torsr' dataset (totally 130,583 tokens, annotated with 10,256 distinct lemma labels and 712 distinct MSD labels). Performance evaluation of predictive models was performed on a test set containing 7683 manually annotated Torlak tokens, a subset of the larger 'torsr' dataset.

MSD tags in this dataset follow the MULTEXT-East Version 4 specification (Erjavec, 2010), to facilitate comparison with neighboring languages of the Balkan Sprachbund, such as Serbian, Bulgarian and Macedonian.

To enable the modeling of 'Torlakness' (i.e. the occurrence of Torlak-style non-standard features) of a text, a dedicated 'labeled version' of the 'torsr' dataset was prepared, in which each token was labeled with either '1' for Torlak or '0' for standard Serbian ('language labels'), based on their origin in either the 'tor' or 'sr' dataset, resulting in 50,687 tokens with label 1 and 51,908 tokens with label 0. No annotations of MSD or lemma for those tokens were used in LLM fine-tuning in this task, to make the fine-tuned LLM independent of such annotations in its input. Here, a split of training / development / testing data of 61,121 / 10,812 / 30,662 tokens was generated, each containing a balanced representation of both language labels.

3.2 Modeling

For predictive modeling the following models were used: LLM on Huggingface were by default fine-tuned for 15 epochs on the 'torsr' dataset, if not stated otherwise. Fine-tuning on only the 'tor' or 'sr' data consistently resulted in much lower predictive performance, as this did not allow for the inteded 'sr'-to-'tor' transfer learning. Test data were always 'tor' only, to see if the LLM's MSD prediction can deal well with the high variability of Torlak. Note that, while all LLM we used here are multi-language, they use either a 'many languages' (e.g. about a hundred) or 'a few languages' (e.g. 3-4) pre-training strategy. Which means that their 'knowledge' is either focused on the patterns that occur in a few selected languages, or in a wide variety of diverse languages.

Baseline model: a custom model of the ReLDI tagger based on the probabilistic CRF implementation we call 'Baseline-CRF' here (Vuković, 2021; Ljubešić et al., 2016), which used the same dataset for predictive modeling. https://github.com/bravethea/Torlak-ReLDI-Tagger-2019

BERTic: a BERT-style transformer focused on Slavic languages, as it was pre-trained on 8 billion tokens of crawled text from Serbian, Bosnian, Croatian and Montenegrin web domains (Ljubešić & Lauc, 2021). https://huggingface.co/classla/bcms-bertic

mBERT is the multi-language version of the seminal BERT transformer, pre-trained on more than 100 languages found on Wikipedia, including several Slavic languages, e.g. Serbian, Bulgarian and Croatian (Devlin et al., 2019). https://huggingface.co/google-bert/bert-base-multilingual-cased

cseBERT is a BERT-style transformer focused on Slavic languages, that was pre-trained on Croatian, Slovenian and English for cross-lingual knowledge transfer (Ulcar & Robnik-Sikonja, 2020). https://huggingface.co/InfoCoV/Cro-CoV-cseBERT

XLM-ROBERTa: the multi-lingual version of the ROBERTa transformer, was

Model	Accuracy	F1 score
Baseline-CRF (torsr)	84.61%	-
BERTic	92.27%	0.9227
mBERT	92.65%	0.9265
cseBERT	92.34%	0.9235
canine-c	93.16%	0.9316
XLM-ROBERTa	93.20%	0.9320
Flair-ROBERTa	93.22%	0.9313

Table 1: MSD predictions.

Model	Accuracy	F1 score
Baseline-CRF	92.62%	-
BERTic	91.41%	0.9141
mBERT	92.47%	0.9247
cseBERT	91.71%	0.9171
canine-c	92.40%	0.9240
XLM-ROBERTa	93.19%	0.9319
Flair-ROBERTa	93.09%	0.9234

Table 2: Lemma predictions.

trained on data from 100 languages, and compares well with mBERT including low-resource languages (Conneau et al., 2020) https://huggingface.co/FacebookAI/xlm-roberta-base

Flair-ROBERTa: combines XLM-ROBERTa and Flair embeddings, which model words as sequences of characters (Akbik et al., 2018).

canine-c: is a tokenization-free language model pre-trained on more than 100 languages (the same data as mBERT). Being tokenization-free, it does not require WordPiece-style subword tokenization as BERT-style transformers typically use. Instead, it works at character level (Clark et al., 2022). It was run for 30 instead of 15 epochs as it required longer fine-tuning than the other transformers to get to its best performance.

https://huggingface.co/google/canine-c

Note that all LLM used here aim at enabling cross-language knowledge transfer including low-resource language settings, in which patterns modeled from the data-rich language would ideally inform the modeling of the low-resource language.

Hyper-parameter optimization was performed to increase the predictive performance of the LLM, comparing performance parameters after fine-tuning, varying the following parameters: epochs, batch size, learning rate, and dropout rate.

To predict the degree of ‘Torlakness’ (Torlak-style non-standard features in Serbian texts), several BERT-style LLMs were fine-tuned using a special version of the ‘torsr’ dataset in which each token was labeled with either 1 for Torlak or 0 for standard Serbian, see section 3.1.

3.3 Evaluation and Statistics

For the comparison of all predictive models, % accuracy of prediction and the weighted F1 score were calculated on the relevant test set, see section 3.1. To compare LLM predictive performance with the published data on the Baseline-CRF (Vuković, 2021).

4 Results

4.1 MSD and lemma prediction: An overview of results obtained with the baseline (CRF) and different LLM in the Torlak test set is provided in Table 1 (MSD predictions) and Table 2 (lemma predictions). Accuracy for the Baseline-CRF model was taken from Vuković (2021), for both MSD and lemma predictions. Hyper-parameter optimization showed best performance for LLM with a relatively small batch size of 2, as this parameter had a considerable effect on predictive performance in both tasks. A learning rate of 5e-5 was selected, 15 epochs of fine-tuning (unless otherwise specified) and a dropout rate of 0.1 for fully connected and attention layers.

In Table 1, we can see that all LLM tested here clearly outperform the CRF baseline model, predicting 712 different MSD labels.

While the observed performance gain using LLM, compared to the Baseline-CRF, is in line with observations in the recent literature (see section 2), the comparison among the different LLM and how they were pre-trained can be informative, as discussed by Ulcar & Robnik-Sikonja (2020). In that sense, depending on the task, it can be advantageous to use a transformer that is focused in its pre-training on only a few (e.g. 3-4) relevant languages, compared to others, like mBERT, canine-c and XLM-ROBERTa, that were pre-trained on more than 100 languages, to focus the model on patterns that are more likely to be relevent in the focus language(s). However, our results indicate that pre-training on many languages rather than a few can actually provide an advantage, at least for this particular dataset and task. As with other NLP tasks in which XLM-ROBERTa reportedly outperformed mBERT (Conneau et al., 2020), including low-resource languages, we also see a slightly better performance of XLM-ROBERTa (and its Flair variant) over mBERT, in terms of both accuracy and F1 score for MSD predictions. Flair-ROBERTa achieves slightly higher accuracy than XLM-ROBERTa, due to the character-level Flair embeddings, but also runs a bit slower.

Table 2 shows the results for the related lemma prediction task. Here, somewhat surprisingly, only XLM-ROBERTa (and its Flair variant) was able to slightly outperform the Baseline-CRF model. Other LLM performed near the baseline. Therefore, in this task, LLM we tested did not provide a substantial advantage over the classic probabilistic predictive modeling method.

4.2 Prediction of 'Torlakness' at token level: Here, several BERT-style LLM pre-trained with relevant languages were used, see section 3.1. After training for 18 epochs, the achieved accuracies and F1 scores are shown in Table 3. mBERT and cseBERT performed best, looking at both performance measures, with a clear advantage over BERTic. Here, we don't see a clear difference between the 'many languages' mBERT model and the 'few languages' models (BERTic, cseBERT), as before. Such an LLM-based approach for enriching Torlak could therefore be useful for detecting Torlak materials that exhibit Torlak-style non-standard features compared to standard Serbian.

Model	**Accuracy**	**F1 score**
BERTic	94.28%	0.9330
mBERT	98.66%	0.9799
cseBERT	98.59%	0.9823

Table 3: Torlakness predictions. The models were all trained with the language-labeled data.

Lemma	Freq	Unique Tokens
biti	6014	nesmo, si, biti, nee, neje, este, nisu, smoo, ste, budu, so, ne, bilo, se, je, beše, bio, jeste, smo, za, nismo, bili, bih, bi, bila, nisəm, nije, səm, e, nesam, sə, su, bude, bile, jeee, niste, nesəm, nesu, sam, j, čə, bil
u	3226	vu, u, v, uu, uuu
i	3175	i, iii, l, iə, ii, u
da	2806	da, daa, ta, d
sebe	1807	si, e, s, še, sə, se
na	1302	naaa, naa, na
za	1224	za
taj	861	toga, tom, teja, to, toe, tuj, tu, toj, taj, tog, te, toa, ti, ta, tija
koji	818	koji, koja, koje, koj, kuj
sa	778	s, səs, sə, sa, sas, səg
on	740	mu, je, ona, gu, nje, joj, nju, gi, ono, go, njom, m, njemu, ju, njega, njeg, ga, g, on, njoj, njim
hteti	629	ćeš, nečeš, neče, nećeš, nećemo, neču, neći, neḱe, oć, ću, će, oče, oćeš, ćeli, ćemo, če, oč, ču, neće, ḱe, ćete, čə
od	605	ot, odi, ood, od
pa	565	pa, pə, ba, paa, p
ne	557	nə, nee, ne, nećemo, neee, n
a	537	aə, aaa, ə, a, əəə, aa, daa, əə
imati	425	imali, nemaš, imalo, nema, imam, nemamo, imal, imaš, nemam, imaju, imao, imaše, imala, ima, imale, imamo, imate, im
godina	416	godine, godinu, godina, godin, godiina
ja	407	j, meni, ni, mene, me, moj, mii, men, mi, m, ja
iz	350	iz

Table 4: Most frequent Torlak lemmas in the language-labeled version of ‘torsr’, with frequency of occurrence (freq) and a list of unique tokens.

The most frequent lemmas in the Torlak part of the labeled ‘torsr’ dataset are listed in Table 4, along with a list of the unique tokens that were mapped to that particular lemma, in this dataset. It provides an impression of the morphological complexity and variability of Torlak variants captured in this dataset, at least regarding the most frequently occurring lemmas. Note how the relationship between surface forms (unique tokens) and their dictionary forms (lemma) are often opaque, even to human experts. As this may help explain why it’s difficult to further improve predictive modeling (Tables 1,2).

5 Conclusions

Despite the high degree of variation found in this Torlak dataset, based on the variable occurrence of non-standard linguistic features (Vuković, 2022), different LLM trained on Slavic (and other) languages were able to achieve very good performance in the prediction of MSD, compared to classic CRF-based probabilistic models. However, the performance gains achieved were less striking in the lemma prediction. The performance gains achieved by the LLMs in MSD prediction may be due to their excellent language modeling abilities.

In addition, we were able to further finetune our models to perform in another new prediction task that we defined, the degree of ‘Torlakness’, which predicts which tokens in a sentence are more similar to Torlak-style non-standard features than to standard Serbian. With this approach it may then be easier to enrich the most non-standard Torlak dialect found in a dataset of mixed Serbian/Torlak texts, e.g. from transcribed audio samples, to further document this vulnerable language. Using such NLP approaches could then help to increase throughput and automation in the face of problematic access to expert-level manual annotation (a common problem for understudied low-resource languages).

With the much improved MSD prediction and the prediction of ‘Torlakness’, we hope to contribute to the conservation of Torlak and improved digital access for Torlak speakers, by improving the ability to automate.

6 Limitations

In terms of known limitations of this work, considering how the LLM models were trained on a high-quality dataset with expert-annotated tokens and careful capture of spoken language variability, the predictive performance of the models may be different when using other kinds of (less controlled or otherwise different) input data, e.g. crawled from the Web. While we did not use the MSD and lemma token-level annotation in the ’Torlakness’ prediction task, to make the model independent of at least these annotations, the above limitations may still be relevant, in terms of the quality of predictive modeling and automation we can expect in such settings.

7 Ethical Considerations

By being aimed at the conservation of a low-resource, vulnerable language, namely Torlak, this work aims at making a contribution to a field that could benefit greatly from similar applications of NLP. As it may help to improve digital access for Torlak speakers, social connectivity and public services, thereby potentially helping to prevent language extinction, or, at least, reduce the everyday pressures on the survival of this language. But if such improved digital access would be sufficient to minimize the stigma experienced by Torlak speakers, as a key force applying pressure on the use of the dialect, is another question that we cannot comment on here.

Ethical considerations can also apply beyond the original intended use of an AI tool, as the same tool could in some cases be used in other situations that present a different ethical scenario from that original setting (Ghotbi, 2024). In this particular case we can not anticipate any major ethical issues, considering the rich literature on the many different multi-language BERT-style transformers available on Huggingface, and their highly task-specific nature (after fine-tuning) e.g. in cross-language knowledge transfer.

With any automation comes the possibility of a risk of a ‘loss of human control’, in this particular case related to the knowledge of a Serbian/Torlak expert annotating tokens manually instead of a more automated approach using AI, such as the one outlined here. But when the dataset of such expert annotations is large enough to let AI learn the patterns that guide such annotation, even for more rare cases, and access to such expertise is limited, the use of AI tools such as those investigated here seems worth considering. Although we may expect quality differences in token annotations in some cases, between expert-annotated and transformer-annotated texts, if we assume that human annotation is perfect. In the case of many low-resource languages, the problems with access to such expert knowledge can be considerable, further making the case for careful AI use and automation, to preserve knowledge about those languages and increase digital access for its speakers.

References

Alan Akbik, Duncan Blythe, Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. https://aclanthology.org/C18-1139.pdf

Vrishali Chakkarwar, Sharvari Tamane, and Ankita Thombre. 2023. A review on BERT and its implementation in various NLP tasks. ICAMIDA 2022, ACSR 105, pp. 112–121. https://doi.org/10.2991/978-94-6463-136-4_12

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805

Tomaz Erjavec. 2010. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA). https://aclanthology.org/L10-1086/

Victor A. Friedman and Brian D. Joseph. 2017. Reassessing Sprachbunds: A view from the Balkans. In *The Cambridge handbook of areal linguistics.* 55:87

Victor A. Friedman. 2011. The Balkan languages and Balkan linguistics. *Annual Review of Anthropology* 40(1):275-291.

Georgi Georgiev, Valentin Zhikov, Petya Osenova et al. 2012. Feature-rich part-of-speech tagging for morphologically complex languages: Application to Bulgarian. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 492–502, Avignon, France. Association for Computational Linguistics. https://aclanthology.org/E12-1050.pdf

Nader Ghotbi. 2024. Ethics of Artificial Intelligence in academic research and education. In *Second Handbook of Academic Integrity*. Springer.

Yoav Goldberg. 2017. Neural network methods in natural language processing. Morgan & Claypool Publishers. ISBN 978-1627052986

Michael A. Hedderich, Lukas Lange, Heike Adel et al. 2021. A survey on recent approaches for Natural Language Processing in low-resource scenarios. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* pages 2545-2568. Association for Computational Linguistics. https://aclanthology.org/2021.naacl-main.201

Mikhail V Koroteev. 2021. BERT: A review of applications in Natural Language Processing and Understanding. arXiv:2103.11943

Jouko Lindstedt. 2000. Linguistic Balkanization: Contact-induced change by mutual reinforcement. In: Languages in Contact, pages 231-246. Brill

Nikola Ljubešić and Davor Lauc. 2021. BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing, pages 37–42, Kiyv, Ukraine. Association for Computational Linguistics. https://aclanthology.org/2021.bsnlp-1.5

Nikola Ljubešić. 2018. Comparing CRF and LSTM performance on the task of morphosyntactic tagging of non-standard varieties of South Slavic languages. Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 156–163 Santa Fe, New Mexico, USA. https://aclanthology.org/W18-3917

Nikola Ljubešić, Filip Klubicka, Željko Agić, Ivo-Pavao Jazbec (2016). New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portoroz: European Language Resources Association (ELRA). https://aclanthology.org/L16-1676/

Alexandre Magueresse, Vincent Charles, Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. arXiv:2006.07264v1

Christopher Manning, Hinrich Schütze. 1999. Foundations of statistical natural language processing. MIT Press, Cambridge MA, USA. ISBN 0-262-13360-1.

Bonan Min, Hayley Ross, Elior Sulem et al. 2023. Recent advances in Natural Language Processing via large pre-trained language models: a survey. ACM Computing Surveys 56(2):1-40. https://dl.acm.org/doi/abs/10.1145/3605943

Jakub Piskorski, Michał Marcińczuk, Preslav Nako et al. 2023. Proceedings of the 9th workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023). Association for Computational Linguistics. https://aclanthology.org/2023.bsnlp-1.0

Piotr Rybak. 2024. Transferring BERT capabilities from high-resource to low-resource languages using vocabulary matching. arXiv:2402.14408

Tapani Salminen. 2010. Europe and Caucasus. In *Atlas of the World's languages in danger*, pages 32-42. Paris, UNESCO Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000187026

Charles Sutton and Andrew McCallum. 2012. An introduction to Conditional Random Fields. *Foundations and Trends in Machine Learning*, Vol. 4, No. 4. Pages 267-373.

Li Tang. 2020. UZH at SemEval-2020 Task 3: Combining BERT with WordNet Sense Embeddings to Predict Graded Word Similarity Changes. *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval)*, pages 166–170, Barcelona (online). International Committee for Computational Linguistics. https://aclanthology.org/2020.semeval-1.19/

Olga Miseska Tomic. 2006. Balkan Sprachbund Morpho-syntactic features. *Studies in Natural Language and Linguistic Theory.* Springer. ISBN-10 1-4020-4487-9.

Matej Ulcar and Marko Robnik-Sikonja. 2020. FinEst BERT and CroSloEngual BERT: Less is more in multilingual models. Text, Speech, and Dialogue: *23rd International Conference, TSD 2020*, Brno, Czech Republic. Springer International Publishing. arXiv:2006.07890.

Teodora Vuković. 2024. Empirical approaches to variation. The case of Timok variety of Torlak. PhD thesis, University of Zurich. https://www.zora.uzh.ch/id/eprint/260570/1/Vukovic_Teodora_Dissertation.pdf

Teodora Vuković, Anastasia Escher, and Barbara Sonnenhauser. 2022. Degrees of non-standardness: Feature-based analysis of variation in a Torlak dialect corpus. *International Journal of Corpus Linguistics*, 27(2):220-247. https://doi.org/10.1075/ijcl.20014.vuk.

Teodora Vuković. 2021. Representing variation in a spoken corpus of an endangered dialect: The case of Torlak. *Language Resources & Evaluation,* 55:731-756. https://www.zora.uzh.ch/id/eprint/195800/

Teodora Vuković. 2020. Spoken Torlak dialect corpus 1.0 (transcription). *Slovenian language resource repository CLARIN.SI.* ISSN 2820-4042. http://hdl.handle.net/11356/1281.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, et al. 2022. Morphological processing of low-resource languages: where we are and what's next. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 988–1007, Dublin, Ireland. Association for Computational Linguistics. https://aclanthology.org/2022.findings-acl.80

Ming Zhou, Nan Duan, Shujie Liu and Heung-Yeung Shum. 2020. Progress in neural NLP: Modeling, Learning, and Reasoning. *Engineering* 6(3):275-290. https://doi.org/10.1016/j.eng.2019.12.014 s