Comparing domain-specific and domain-general BERT variants for inferred real-world knowledge through rare grammatical features in Serbian

Sofia Lee
Vrije Universiteit Amsterdam
s.m.lee2@student.vu.nl

Jelke Bloem
University of Amsterdam
j.bloem@uva.nl

Abstract

Transfer learning is one of the prevailing approaches towards training language-specific BERT models. However, some languages have uncommon features that may prove to be challenging to more domain-general models but not domain-specific models. Comparing the performance of BERTić, a Bosnian-Croatian-Montenegrin-Serbian model, and Multilingual BERT on a Named-Entity Recognition (NER) task and Masked Language Modelling (MLM) task based around a rare phenomenon of indeclinable female foreign names in Serbian reveals how the different training approaches impacts their performance. Multilingual BERT is shown to perform better than BERTić in the NER task, but BERTić greatly exceeds in the MLM task. Thus, there are applications both for domain-general training and domain-specific training depending on the tasks at hand.

1 Introduction

The recent introduction of Transformer models (Vaswani et al., 2017) has precipitated a dramatic shift in the landscape of Natural Language Processing, bringing unprecedented gains in performance and accuracy. Of particular note is Bidirectional Encoder Representations for Transformers (BERT), a model which has become a baseline for NLP tasks (Devlin et al., 2018). As with other deep neural network models, however, it is largely unknown *how* BERT is able to achieve its performance. The bulk of NLP research focused on BERT, a sub-field that has come to be known as BERTology, has centred around probing underlying embeddings through various aptitude tests, comparing the performance of different BERT variants to each other as well as to human performance metrics. These tasks may consist of more traditional tasks such as the Cloze task, or more NLP-specific tasks such as named entity recognition or sentiment analysis. As English is the language of the original BERT models, these efforts have usually focused on investigating linguistic phenomena that exist in English. This leaves us with a knowledge gap of BERT's representation of linguistic typological features that are not shared with English but occur in other language families, such as the Slavic languages.

Due to the success of transformer models, many domain-specific derivatives of BERT have been produced. This includes topic-based domains such as scientific text model SciBERT (Beltagy et al., 2019), as well as domains of related languages such as the Finnish-Estonian model FinEstBERT (Ulčar and Robnik-Šikonja, 2020) and BERTić (Ljubešić and Lauc, 2021), a model for Bosnian, Croatian, Montenegrin and Serbian (BCMS). The introduction of domain-specific derivatives has sparked a debate within BERTology on how specific the dataset of the fine-tuning task should be. Should derivatives be trained on a more general dataset within a domain or should they be fine-tuned on a much more domain-restricted dataset?

We contribute to this debate by focusing on the case of closely related and under-resourced languages. We compare two variants of BERT: the domain-general Multilingual BERT (mBERT), and the language-specific model BERTić, trained from scratch on the BCMS languages. In particular, we explore how the two training approaches affect performance on rare grammatical phenomena in Serbian. Our case study is indeclinable nouns, a phenomenon typical of fusional languages where the same morphological form is used for all grammatical functions of a noun. This is a challenging phenomenon to model as it is typically infrequent and the usual morphological cues of the language aren't expressed. We create adapted versions of the common probing tasks of Masked Language Modeling and Named Entity Recognition specifically targeting this phenomenon.

We contend that such phenomena that do not occur in English pose unique challenges to language modelling, particularly in under-resourced

Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023), pages 47–60

languages, and can reveal some of the overlooked underlying representations learned by BERT derivatives. We aim to show areas in which transformer-based language model training can improve, as well as emphasize the importance of analysing the linguistic capabilities of non-English BERT variants.

2 Background

Typically, BERT makes use of mixed-domain transfer learning. The first stage of training uses general-domain data, such as base BERT's training on Wikipedia and BookCorpus, followed by a fine-tuning domain-specific stage. Domain-specific pre-training has been proposed to be more effective. Beltagy et al. (2019) compare the results of a more general scientific domain BERT variant SciBERT with that of biomedical-specific BioBERT (Lee et al., 2019). SciBERT outperformed BioBERT in biomedical text tasks. Gu et al. (2021) contend that SciBERT's higher performance stems from its from-scratch training on scientific domain text.

Non-English language modelling provides distinct challenges compared to domain-specific training. Human languages differ in ways that exceed that of domains of the same language. While related languages may share vocabulary and grammatical features, they often differ vastly in information structure and syntax. Languages may also have varying amounts of quality data available. High resource languages such as English or German can be trained on monolingual text, while under-resourced languages may have no options but transfer learning. Transfer learning is the predominant approach to building language-specific variants of BERT. On top of base BERT, Multilingual BERT (mBERT) is additionally trained on the text of 104 language-specific Wikipedias without any cross-lingual alignment. mBERT achieved impressive cross-lingual performance and itself is used as a base for countless language-specific BERT derivatives, taking a mixed-domain training approach (Wu and Dredze, 2019).

However, several studies have shown that language-specific BERT models trained on a dataset consisting of only one language still perform better than mBERT-based models, especially in the case of under-resourced languages (Wu and Dredze, 2020). Bhattacharjee et al. (2021) show that a Bangla-specific variant, BanglaBERT, outperforms both mBERT and a Bangla-English bilingual variant. Tanvir et al. (2021) similarly show that an Estonian-specific BERT outperforms multilingual variants in five out of seven tasks. Likewise, Martin et al. (2022) find that a BERT variant trained ground-up on a Swahili dataset outperforms multilingual models. BERTić, a variant trained on Bosnian, Croatian, Montenegrin and Serbian, also outperforms both mBERT and a trilingual Croatian, Slovene and English BERT in nearly every task (Ljubešić and Lauc, 2021).

2.1 Grammatical embedding and indeclinable nouns

BERT shows a surprising ability to perform grammatical generalisation. Madabushi et al. (2022) find that BERT even outperforms human subjects in a task predicting article use (e.g. *a/an*, *the*) in English and tends to agree with annotators when annotators agree with each other. Multilingual models have also demonstrated that synthetic transfer can occur between languages (Guarasci et al., 2022). Meanwhile, Haley (2020) show that BERT can perform the Wug test, a standard grammatical generalisation test (Berko Gleason, 1958), significantly better than chance in English, French, Spanish and Dutch.

However, there are still many gaps in this research. Firstly, high-resource languages are used for these studies, where a model will have more evidence to generalize over grammatical patterns. Although some patterns may be transferred to under-resourced languages, these languages may present a diverse range of unique or rare typological features. Secondly, few if any of the languages studied are fusional languages, meaning its inflectional endings encode several pieces of information at the same time (Bender, 2019). The nature of word paradigms in these languages provides significant challenges for generalisation and statistical modelling due to the multitude of forms for each word.

One phenomenon common to many fusional languages is that of the indeclinable noun. Indeclinable nouns are nouns which exhibit an extreme form of case syncretism in which the same form is used for all grammatical functions. In many cases, such nouns form some sort of semantic class, such as being loanwords or abbreviations. As an example, although English is not a highly inflected language, it does have indeclinable nouns which violate the usual *-s* suffix in forming the plural, such as 'moose'. This word retains the same form in the singular and plural, and this is said to be the case due to its being a loanword from Eastern Algo-

nquian. Fusional languages that have indeclinable nouns include Russian (Nedomová, 2013), Czech (Naughton, 2006), Upper Sorbian (Corbett, 1987), Lithuanian (Mathiassen, 1996), Latvian (Kalnača and Lokmane, 2021), Latin (Schmitz, 2004), and both modern and ancient Greek. Indeclinable nouns serve as a fitting rare phenomenon to probe because they are present in a variety of under-resourced languages, appear relatively infrequently in corpora, and often require some kind of intuition from a speaker in order to correctly identify and use. To date, no studies that focus specifically on indeclinable nouns and language modelling exist, although indeclinable nouns are shown to cause low performance in NER tasks in a Greek edition of BERT (Singh et al., 2021).

2.2 Serbian as a subject for BERTology

Serbian is one of four mutually intelligible varieties of a pluricentric language referred to collectively as Bosnian-Croatian-Montenegrin-Serbian (BCMS). It is a highly inflected language, inflecting for case, number and gender in nouns, adjectives and some verb participles. Serbian is also a fusional language, as the same endings may encode different features. Serbian also has indeclinable nouns.

As with many other highly inflected languages, nouns in Serbian fall under a variety of paradigms with different numbers of unique forms. Masculine and neuter nouns exhibit one less form than feminine nouns, while some nouns, such as *ljubavi* 'love' only distinguish between three forms (four in some dialects). Indeclinable nouns in Serbian are a particularly restricted class. Whereas other languages may not place semantic restrictions on indeclinable loanwords, Serbian reserves indeclinability to two types of words: certain numbers, and loanwords or foreign names with a female referent that do not end in *-a* (Fidler et al., 2005). The latter are particularly infrequent in Serbian corpora as a whole but also grow in frequency daily due to an ever-increasing amount of global news and celebrity gossip written in the language.

Although Željko Bošković (2006) and Fidler et al. (2005) observe that indeclinable nouns are not allowed in sentences without an adjective that clarifies the case assignment, recent Serbian tabloids have simply used indeclinable names in case assigning roles as with any other name. Example 1, a lyric from 'In corpore sano' by Konstrakta, Serbia's entry in 2022 Eurovision, demonstrates how the indeclinability of female proper names may still be assigned cases even without a preposition.

(1) Koj-a li je
which-.F.SG.NOM Q be.3.SG.PRS
tajn-a zdrav-e
secret-.F.SG.NOM healthy-.F.SG.GEN
kos-e Megan
hair-.F.SG.GEN Meghan.F.SG.GEN
Markl?
Markle.F.SG.GEN
'Just what is the secret to the healthy hair of Meghan Markle?'

While indeclinable nouns function the same way in Bosnian and Croatian, Serbian requires all names to be written phonetically. Names are thus obfuscated from their native spelling, making them less likely to benefit from transfer learning. Indeclinable nouns in Serbian are thus especially suited as indicators of named entity recognition ability, semantic awareness, and real world knowledge.

2.3 Serbian as an under-resourced language

In comparison to high-resource languages such as English, research on Serbian NLP is sparse. Miletic (2018) provides a treebank for Serbian consisting of 81K tokens. A Python package by Ostrogonac et al. (2020), *nlpheart*, provides text processing tools for Serbian, although at the time of writing it remains unavailable. As a whole, NLP studies on Serbian are few, and tools tend to be defunct. The situation is not significantly improved even when factoring in the related Croatian, Bosnian or Montenegrin languages. Many tools are also grouped in with the related but not mutually intelligible Slovene. Ljubešić and Dobrovoljc (2019) provide a NLP pipeline for Slovene, Croatian and Serbian consisting of a part-of-speech tagger, a lemmatiser, a tokeniser, a dependency parser, and a named-entity recogniser.

Ulčar and Robnik-Šikonja (2020) provide a multilingual BERT model, CroSloEngual BERT or cseBERT, which although trained on Croatian and Slovene, has been shown to perform well on Serbian NLP tasks. Moving closer to Serbian, BERTić (Ljubešić and Lauc, 2021) is trained on the CLASSLA web corpus, based on Bosnian, Croatian, Montenegrin, and Serbian websites, the Riznica corpus of Croatian literature and newspapers (Ćavar and Brozović Rončević, 2012), and the cc100 corpus (Conneau et al., 2020). The corpora on which BERTić is trained are currently the largest for the BCMS languages. Ljubešić and Lauc

Meaning	'Jelena' (name)	'Marko' (name)	'hill'	'joy'	'Jean' (name)
Nominative	Jelen-a	Mark-o	brd-o	radost-Ø	Džin-Ø
Genitive	Jelen-e	Mark-a	brd-a	radost-i	Džin-Ø
Dative/Locative	Jelen-i	Mark-u	brd-u	radost-i	Džin-Ø
Accusative	Jelen-u	Mark-a	brd-o	radost-Ø	Džin-Ø
Vocative	Jelen-o	Mark-o	brd-o	radost-i	Džin-Ø
Instrumental	Jelen-om	Mark-om	brd-om	radost-i	Džin-Ø

Table 1: Common noun declension paradigms, including indeclinable names.

(2021) find that BERTić outperforms both mBERT and the Slovene-Croatian-English model CroSloEngual BERT (Ulčar and Robnik-Šikonja, 2020) in morphosyntactic tagging, named entity recognition, social media geolocation prediction, and commonsense casual reasoning. They also find that despite the lack of exposure to Serbian in cseBERT, there are no significant improvements in Serbian performance between cseBERT and BERTić, aside from one Serbian NER task. For this last reason, we use BERTić for this study.

3 Methodology

We compare BERTić and mBERT on two tasks: a feminine Named Entity Recognition (NER) task, targeting the name domain in which the indeclinable noun phenomenon occurs, and Masked Langauge Modelling (MLM), a more intrinsic evaluation task. BERTić is pretrained with the ELECTRA training objective, where instead of masking tokens, tokens are corrupted and a detection task is performed (Clark et al., 2020). MBERT uses the standard BERT MLM training objective. Other multilingual BERT-based models are available such as XLM-R (Conneau et al., 2020) and mT5 (Xue et al., 2021), but these all use the standard MLM objective rather than ELECTRA. Out of these, we chose to compare to mBERT as this comparison was also made by Ljubešić and Lauc (2021). Both models use WordPiece subword tokenization (Schuster and Nakajima, 2012).

For the feminine NER task, NER-fine-tuned variants of both BERTić and mBERT are used. The BERTić variant we use is the *bcms-bertic-ner* variant, which has been fine-tuned on the Croatian hr500k dataset, Serbian SETimes.SR dataset, and the ReLDI-hr and ReLDI-sr Internet (Twitter) datasets in Croatian and Serbian respectively. In total, the dataset consists of 768k tokens. Since a NER variant is not readily available for mBERT at the time of testing, we use *bert-base-multilingual-cased-ner-hrl* instead. This model is fine-tuned on Arabic, German, English, Spanish, French, Italian, Latvian, Dutch, Portuguese and Chinese NER. The training process is not well documented, but appears to consist mainly of newspapers from the early- to mid-2000s. This datedness ensures that mBERT does not have an extra advantage from being exposed to a wider selection of modern names. SpaCy[1] is used as a baseline for comparison. For the MLM task we only use the base BERTić and mBERT models. All tasks are performed using a Dell Optiplex 7010 with an Intel i7 processor and 12GB of RAM.

3.1 Named Entity Recognition

We sourced a list of names of popular female celebrities from nationality category lists in the Serbian Wikipedia. All names are converted from Serbian Wikipedia's default Serbian Cyrillic script to Serbian Latin script and edited for capitalisation errors. Details of which names we included and spelling variation and exceptional cases can be found in Appendix A. In total, 1323 names are included, of which 812 names are completely indeclinable, meaning the name does not include any declinable element. 511 contain at least one declinable element, of which 13 appear to be of Southern Slavic origin. 30 names are fully declinable.

We take the log_{10} frequency of each name across all three Serbian BERTić training corpora as a weighting score for that name to use in the evaluation. For example, *Madona* ('Madonna', a singer), appearing 5,060 times in the corpus, scores approximately 9.36. Unattested names, such as *Zelda Rubenstejn* ('Zelda Rubinstein', actress in 'the *Poltergeist* film series), are given a score of 0. There are 166 unattested names and 92 names with one attestation. The greatest number of attestations is 11602. Scores follow a Zipfian distribution.

[1] https://spacy.io/

We generated a feminine NER test corpus by filtering the three Serbian-specific corpora, on which BERTić was trained, for lines containing names from the list. This process generates 97,981 sentences, which is reduced by 6,619 or 6.75% when pruned for duplicates. Names are annotated with their name type, which could be either *indeclinable*, *Slavic*, *fully declinable*, or *declinable*. Names of any declinable type are also labelled by one of five cases: Nominative, Genitive, Dative/Locative, Accusative or Instrumental. Vocative, which is virtually unseen in the dataset, is ignored, and Dative and Locative are combined due to their identical forms. All input, including names, is tokenized by the model's tokenizer. In evaluation, models are awarded a point only for complete, unbroken names identified, with the B-PER token in the beginning of the name and the I-PER token at the end. Other categories are discarded and names not included in the name list are ignored.

3.2 Masked Language Modelling

In the Masked Language Modelling task, a set of 216 sentences for each name in the name list is automatically generated using templates, totalling 285,758 sentences. A mask was inserted at a predetermined spot for the models to fill in. Each sentence could be of one or two types: a low-context type, in which there is one sentence containing the name and mask with minimal context, and a high-context type, in which the declinability of the name is demonstrated by one of eleven sentences that involve case assignment. This distinction is made to differentiate between the use of information from the embedding itself (low-context condition) and from the grammatical inflections in the contextual sentence (high-context condition). By only providing the nominative form in the low-context sentences, no information about the gender of the name is available if it does not have a feminine form, i.e ends in *-a*. High-context sentences provide inflectional information that can indicate gender through feminine inflections, either by having no inflections or through the native inflections. All sentences are written to be as gender neutral as possible otherwise.

Low-context sentences consist of one completely open-ended sentence (e.g 'Laura Dern is [MASK]') and sentence types that elicit particular parts-of-speech that may encode information about gender in Serbian, such as an adjective (e.g. 'Laura Dern is very [MASK]'). The high-context condition involves a context sentence containing a name paired with a high-frequency other name — three male names and three female names. Each of the cases are represented. An example is 'Vladimir is afraid of Laura Dern (genitive)'. This is then followed by a sentence with a mask as in the low-context condition. The full set of sentence types with glossing can be seen in Appendix B.

All input, including names, is tokenized by the model's tokenizer. All sentences include a single mask, in which any element from a model's vocabularly can be predicted, including subtokens. The top five suggestions for each sentence by each model are counted, regardless of model confidence. Responses are manually scored and only deemed correct if the suggested word is a word in Bosnian, Croatian, Montenegrin or Serbian and fall into one of the following word types: 1) a noun referring to a woman, such as *političarka* 'politician (f.)'; 2) an adjective with a feminine ending, e.g. *srećna* 'happy (f.)'; 3) the possessive feminine adjective, *njen* or *njezin*; 4) a feminine past participle, e.g. *pročitala* 'read (f.)'; or 5) the feminine plural past participle of *biti* 'to be', *bile*. Nouns of feminine gender that do not refer to humans, such as *ulica* 'street' or *reka* 'river' are not counted as correct. Nouns that are grammatically feminine but not semantically, such as *osoba* 'person' were not counted. All proper names, even if feminine, are ignored. A single animal word, *mačka* 'cat' which also double as slang term for a woman, is included, while others, such as *zmija* 'snake' or *riba* 'fish' are excluded. Words that are feminine but not in the BCMS lexicon are not considered correct. Finally, subtokens (word segments), even if ungrammatical, are scored as correct as long as it indicates a feminine gender.

4 Results

4.1 Named Entity Recognition

mBERT scores the highest in the feminine Named Entity Recognition task (87.49%), outperforming both BERTić (57.79%) and the spaCy baseline (35.98%).[2] Figure 1 visualizes these results by the log frequency of each name in the corpus as operationalized in Section 3.1. Furthermore, mBERT and BERTić both performed slightly worse with

[2] A χ^2 test of independence shows that there is a statistically significant association between correctness and model type, $\chi^2 = 45238.63$ (2, N = 300348), $p < 0.00001$.

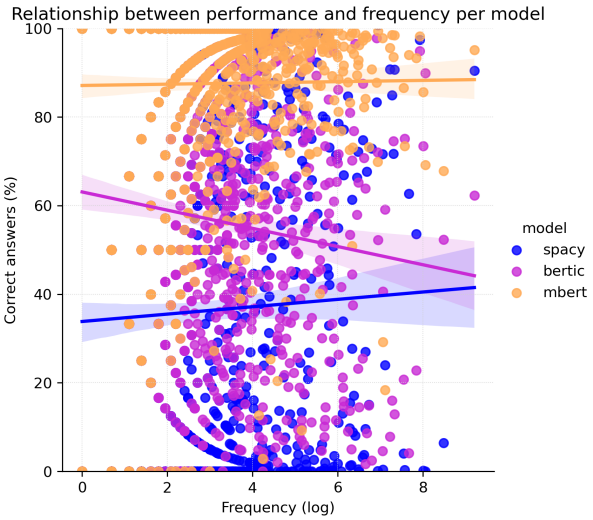


Figure 1: NER results

indeclinable names than the average (86.37% and 55% respectively) whereas spaCy saw a significant improvement with them (47.57%). Results with regression lines for each name type are shown in Figure 2 or in Appendix C for the spaCy baseline. BERTić shows a weak negative but significant correlation between performance and name frequency, r(1321) = -0.11, $p < 0.0005$. No such correlation is found for mBERT or spaCy.

4.2 Masked Language Modelling

BERTić provides feminine forms 49.16% of the time whereas mBERT only provides feminine forms 15.75% of the time.[3] Figure 3 visualizes these results by name frequency. Forms that are feminine but do not appear to be Serbian words were excluded.

BERTić shows higher performance (49.64%) in low-context sentences than high-context ones (44.15%) whereas mBERT performed worse in low-context sentences (15.08%) compared to high-context sentences (22.71%). For both BERTić and mBERT, declinable names of all types resulted in a feminine form more often than an indeclinable form. BERTić selects a feminine form 33.17% of the time with indeclinable names, 82.68% of the time with Slavic names, 75.67% of the time with fully declinable names, and 74.26% of the time with other declinable names (Figure 4a). mBERT only selects a feminine form 8.65% of the time with indeclinable names, 27.13% of the time with Slavic names, 27.45% of the time with fully declinable names, and 27.02% of the time with other declinable names (Figure 4b).

Since low-context sentences only use nominative case, we evaluate case performance only for high-context sentences. There is little variation between the performances per case of both BERTić (M = 50.26, SD = 2.38) and mBERT (M = 14.67, SD = 1.31). Cases rank from highest to lowest performance for BERTić are nominative (52.32%), accusative (51.56%), instrumental (51.46%), dative (50.27%) and genitive (45.66%), whereas for mBERT the order is dative (16.69%), genitive (15.25%), accusative (14.84%), instrumental (13.74%) and nominative (12.85%). We also compare the distribution of the feminine forms per name to the frequency of each name in the corpus. BERTić showed a very weak correlation between performance and frequency, r(1321) = .08, $p < 0.005$. Thus, BERTić is somewhat more likely to select a feminine form to complete a sentence when the sentence is focused on a more common the name in the corpus. This is especially the case when the sentence concerns an indeclinable name. No such correlation is found for mBERT.

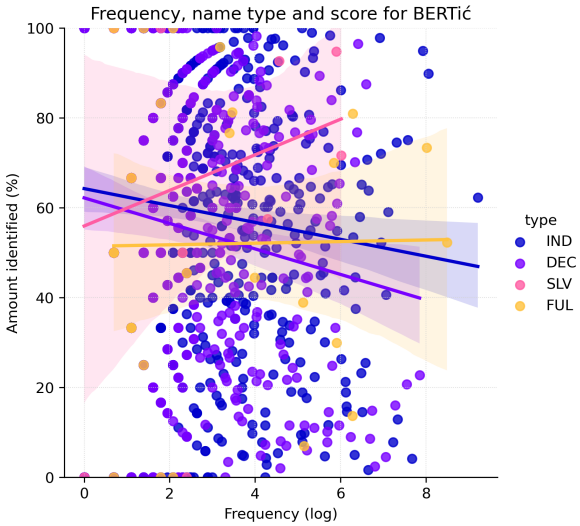
5 Discussion

5.1 Named Entity Recognition

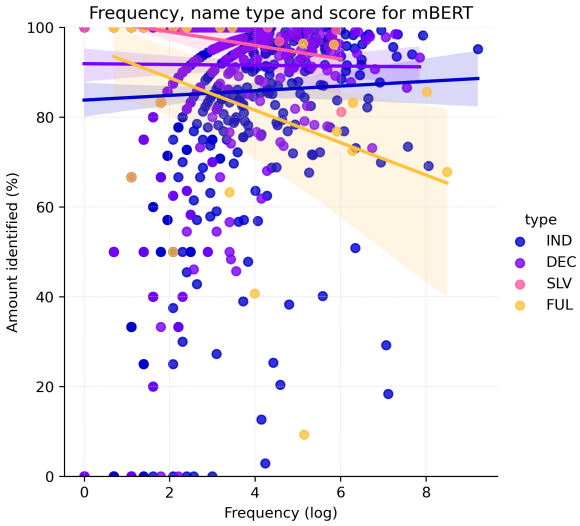
In contrast to the results of Ljubešić and Lauc's (2021) general NER task, BERTić trails significantly behind mBERT in our feminine NER task. In the general task both models reach near 90% accuracy, while in our task only mBERT did. Only when the name is fully declinable and in the accusative case both models perform similarly, but our dataset has only 30 of 1323 fully declinable names and indeclinable is the most common type (exact numbers are in Section 3.1).

An error analysis reveals that BERTić produces excessive span errors, exhibiting a tendency to oversegment all names. From the first 10000 lines of the srWAC celebrity sub-corpus, when looking at both male and female names, BERTić over-segments on 6348 lines a total of 12123 times, sometimes even twice in the same name. Understandably, the names in question, being uncommon, lack embeddings in BERTić and are thus tokenized into subtokens, but this does not explain why BERTić performs significantly worse than mBERT, which is even less likely to have full token embeddings for such a name. In many cases, BERTić and mBERT are tokenising

[3] A χ^2 test of independence shows that there is a statistically significant association between correctness and model type, $\chi^2 = 363597.04$ (2, N = 2857670), $p < 0.00001$.



(a) BERTić NER results per name type



(b) mBERT NER results per name type

Figure 2: NER results for both models with regression lines for each name type, including indeclinable (IND), declinable (DEC), Slavic (SLV) or fully declinable (FUL).

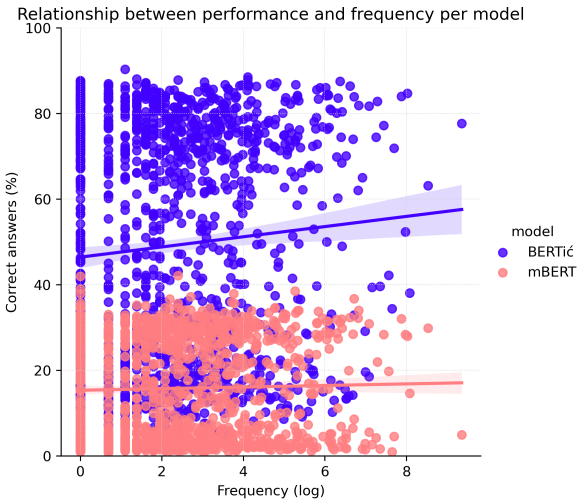


Figure 3: Overall MLM scores for BERTić and mBERT

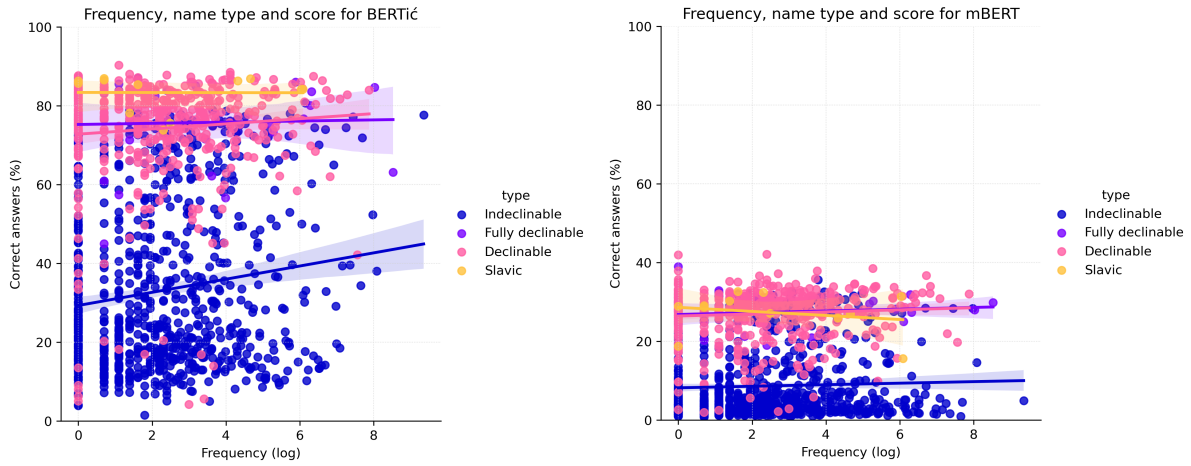
names into similar subtokens, but only mBERT consistently labels the beginning of a name with the correct B-PER tag instead of I-PER (indicating a separate name). For mBERT, over-segmentation occurs 495 times in the same sample. Of this, 207 occur with a name containing the characters *ž*, *š* or *đ*. As mBERT is trained with all diacritics stripped out, this hints at an encoding error.

However, not all of BERTić's low scores can be attributed purely to low performance. In some cases, BERTić provides answers that demonstrate more advanced comprehension of context. Phrases such as *vlada Margaret Tačer* 'the government of Margaret Thatcher' are labelled as organisations by BERTić whereas only the name *Maragret Tačer* is tagged by mBERT. BERTić performance here has higher practical significance. Although Singh et al. (2021) suggest that indeclinable nouns pose particular challenges in the NER task, we only see minor differences. This could be attributed to the fact that most female foreign names are indeclinable, potentially causing models, particularly the language-specific BERTić, to struggle with the whole semantic class of female foreign names (i.e. our entire dataset), including declinable ones.

5.2 Masked Language Modelling

The MLM task shows that indeclinable names are particularly challenging to both mBERT and BERTić. Unlike in the NER task, both models clearly fare worse when facing sentences with indeclinable names. BERTić performs better when a name is more common, suggesting that higher representation in a dataset helps. Interestingly, BERTić scores lower in the high-context sentences compared to the low-context sentences, whereas mBERT scores higher in low-context sentences. While mBERT may need more context in order to identify the language being used, it is unclear why BERTić sees a performance loss when working with high-context sentences. The effect of the divergent vocabularies of the tokenizers should be limited on this task as we also scored subtokens.

mBERT and BERTić, to varying degrees, both show evidence that names of famous people are being discussed. *poznata* 'famous' (f.) is among the



(a) MLM scores for BERTić

(b) MLM scores for mBERT

Figure 4: MLM scores for both models with regression lines for each name type.

top results for both mBERT and BERTić. However, BERTić shows a larger variety of words such as *zanimljiva* 'interesting' (f.) and *pametna* 'smart' (f.). In general, BERTić is able to produce 244 feminine words compared to mBERT's considerably smaller 62, a large amount of which are actually sub-words. BERTić, through its specialised training, appears able to produce more relevant descriptors.

5.2.1 Language identification

mBERT occasionally confuses the text with that of other Slavic languages, which is understandable given that it does not specialise in BCMS. The incredibly high occurrence of *v* ('in' in a considerable number of Slavic languages) suggests that mBERT is able to identify the text as being in some Slavic language, but not specifically Slovene, Czech, or Slovak. *v* however is not grammatical in any of the sentences given and has a low confidence score.

Slovene and Croatian in particular share a considerably large amount of vocabulary. Many of its top results (*za*, *dobra*, *velika*, *na*, *brzo* to name a few) are shared vocabulary with Slovene if not other Slavic languages, and some frequent responses with high scores, such as *objavil*, are most likely Slovene. Although such forms also exist in Kajkavian Croatian, this language variant is most likely unrepresented in mBERT's training set. This language confusion is probably a result of mBERT's domain-general training.

The issues that mBERT faces show one of the situations in which domain-general training may be ill-suited. These issues are exacerbated in low-context sentences. One of the ways that this may be rectified is through fine-tuning. A future study could explore how mBERT's performance could improve if fine-tuned for Serbian texts.

5.2.2 Language standards

Considering that the training set contains corpora in all variants of BCMS, BERTić mixes both Serbian standard spellings and spellings not considered standard Serbian in its responses. However, this occurs much less often than one would expect. BERTić shows a strong preference for Serbian forms for some words but uses non-standard or Bosnian, Croatian or Montenegrin forms for others. In some cases, the Serbian form of a word is not used at all. Table 2 shows some examples.

We also observe frequent output of Ijekavian spelling forms which are the standard in other BCMS regions, as opposed to Standard Serbian Ekavian spelling. This suggests that training a language model on a combined dataset of all language variants may induce negative transfer of a feature that is more common in other variants.

Twelve words in mBERT's result set are in Cyrillic, whereas BERTić has none. By not supporting Cyrillic, BERTić is effectively restricted to only Latin-using domains, ignoring the bi-alphabetism of Serbian. As the choice between the two alphabets is not arbitrary and can be tied to register, ideally a model would be trained on both Cyrillic and Latin text in their original scripts.

5.3 Implications for under-resourced languages

A known limitation of most large language models is that they reproduce social biases which are reflected in the training data (Mehrabi et al., 2019).

Lemma	Meaning	Serbian standard	Non-standard
lepa	'beautiful'	48512	20945
devojka	'girl'	10564	217
srećna	'happy, lucky'	0	3015
vredna	'valuable, worthwhile'	0	1528
volela	'love' (past participle)	0	68
pevačica	'singer' (f.)	0	50
poslednja	'last, final'	9	9
devojčica	'girl' (diminutive)	0	3

Table 2: Frequencies of Serbian standard and non-standard duplets in BERTić responses to the MLM task.

The effect of ethnic tensions in the Balkan region is well-known, and studied by sociolinguists, but less so in NLP. Considering that training on less data may amplify any biases within that data, BERTić or any other language model trained on corpora emerging from current or recent conflict will have a greater tendency to reproduce conflict discourse since the proportion of conflict-neutral training data is smaller. We observed evidence of this.

During the masked language modelling task, BERTić produces *Srbin* 'Serb' 8883 times and *Hrvat* 'Croat' 1115 times. In fact, *Srbin* is the 38th most common word in BERTić's answer set, while *Hrvat* is the 151th most common word. Additionally, BERTić also produces *musliman* ('a male follower of Islam', sometimes used to refer to Bosnians) 101 times. These forms largely surface in the most open-ended sentence in the MLM task. In contrast, mBERT does not produce any of these words once.

The fact that ethnic discourse is reproduced in BERTić has implications for other languages from conflict zones. Languages are not under-resourced simply because of neglect, but because of social, political and historical factors that create their present status. In the case of Serbian and its close relatives, political factors such as national language policies complicate the development of tools for each language standard. Both practical and political reasons impact the appropriateness of a BCMS-general model. Attempts to develop NLP tools for BCMS or any of the national standards must contend with the forces that continue to shape the identity of BCMS and its speakers.

6 Conclusion

We evaluated the performance of two BERT variants, multilingual BERT (mBERT) and BERTić, on Serbian indeclinable nouns, using a NER task and a MLM task. While in a general NER task, BERTić and mBERT show similar performance on Serbian (Ljubešić and Lauc, 2021), mBERT outperforms BERTić in our feminine NER task. In the MLM task, BERTić vastly outperforms mBERT and both models performed significantly worse on indeclinable names. BERTić produces a larger diversity of pragmatically correct responses overall. These results indicate that BERTić may encode information about gender and names, but whether the encoding can be considered a morphological feature of nouns or is specific to a semantic domain of names remains unclear. We only see that BERTić's performance is sensitive to name frequency. mBERT on the other hand produces feminine forms significantly less often, and produces responses from related languages such as Slovene and Czech.

The results from the NER task suggest that multilingual models perform better when the names are not native to the text language. On the other hand, language-specific tasks such as sentence completion will produce significantly more relevant results from models trained specifically for the language, as the embeddings contain a significantly larger amount of vocabulary for the target language.

Potential future directions include research on other typologically rare grammatical features, the behaviour of BERT with other kinds of fusional languages and probing how contextual real-world knowledge inferred from them may be encoded. The representation of bi-alphabetical languages in language modelling could be explored further, as well as the ways language-specific training compares to more general training when dealing with closely related variants. More broadly, we claim that research on closely related languages contributes to our knowledge of the conditions and factors that affect the choice between using a transfer learning or in-domain learning approach.

Acknowledgements

We would like to thank Lisa Bylinina for insightful comments on an earlier version of this work, as well as the workshop's anonymous reviewers for their helpful comments.

Limitations

Due to our starting point of studying existing resources, our study was limited to already existing models. It might have been possible to train or tune better-performing models for the Serbian language specifically by making our own model. The choice to use existing resources also comes with some methodological issues for the NER task - in particular, that there were most likely differences between the fine-tuning procedures on the NER task of both models. A controlled experiment in which both base models are tuned on the same NER data would exclude some possible sources of variation between the two approaches, but would have also cost significantly more training resources. Our choice also means we had no control over hyperparameters - perhaps a Serbian-specific tuning could improve performance.

Due to the limited resources available for Serbian, we had to use sentences from a corpus that BERTić was trained on for the NER evaluation. However, as this overlap is only with the pre-training dataset and the NER-specific tuned BERTić used different datasets we expect that this choice had limited consequences for NER performance on the evaluation set. We also did not have a proper NER gold standard available in which all names in text were annotated, so we were only able to report accuracy, not recall, on our own silver standard.

Our study is a case study of a specific phenomenon in a specific language, thus there is no way to ascertain that other rare grammatical phenomena in other under-resourced languages would also benefit from language-specific training on the basis of only our study.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Emily Bender. 2019. The #benderrule: On naming the languages we study and why it matters. *The Gradient*.

Jean Berko Gleason. 1958. The child's learning of English morphology. *Word*, 14.

Abhik Bhattacharjee, Tahmid Hasan, Kazi Samin, M. Sohel Rahman, Anindya Iqbal, and Rifat Shahriyar. 2021. Banglabert: Combating embedding barrier for low-resource language understanding. *CoRR*, abs/2101.00204.

Damir Ćavar and Dunja Brozović Rončević. 2012. Riznica: The Croatian Language Corpus. *Prace filologiczne*, 63:51–65.

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

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

Greville G. Corbett. 1987. The morphology/syntax interface: Evidence from possessive adjectives in Slavonic. *Language*, 63(2):299–345.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Masako Fidler, Stephen Wechsler, and Larisa Zlatić. 2005. The many faces of agreement. *The Slavic and East European Journal*, 49:170.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Raffaele Guarasci, Stefano Silvestri, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2022. Bert syntactic transfer: A computational experiment on italian, french and english languages. *Computer Speech & Language*, 71:101261.

Coleman Haley. 2020. This is a BERT. now there are several of them. can they generalize to novel words? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 333–341, Online. Association for Computational Linguistics.

Andra Kalnača and Ilze Lokmane. 2021. *Latvian Grammar*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

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.

Harish Tayyar Madabushi, Dagmar Divjak, and Petar Milin. 2022. Abstraction not memory: Bert and the english article system.

Gati Martin, Medard Edmund Mswahili, Young-Seob Jeong, and Jeong Young-Seob. 2022. SwahBERT: Language model of Swahili. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 303–313, Seattle, United States. Association for Computational Linguistics.

T. Mathiassen. 1996. *A Short Grammar of Lithuanian*. Slavica Publishers.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635.

Aleksandra Miletic. 2018. *Un treebank pour le Serbe: Constitution et exploitations*. Ph.D. thesis, Université de Toulouse - Jean Jaurès.

J. Naughton. 2006. *Czech: An Essential Grammar*. Routledge Essential Grammars. Taylor & Francis.

Zdeňka Nedomová. 2013. Paparazzi, matcho, guru, yeti - nesklonná životná apelativa v ruštině a češtině. *Studia Slavica*, 17(1):91–102.

Stevan Ostrogonac, Borko Rastovic, and Elizaveta Liliom. 2020. A python package for text processing for serbian: nlpheart. *Scientific Technical Review*, 70:41–45.

L. Schmitz. 2004. *Grammar of the Latin Language*. New Language Guides. Hippocrene Books.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE.

Pranaydeep Singh, Gorik Rutten, and Els Lefever. 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*, pages 128–137, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Hasan Tanvir, Claudia Kittask, Sandra Eiche, and Kairit Sirts. 2021. Estbert: A pretrained language-specific bert for estonian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 11–19.

Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pages 104–111. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Željko Bošković. 2006. Case checking versus case assignment and the case of adverbial nps. *Linguistic Inquiry*, 37(3):522–533.

A Building the names list

This appendix contains the details of manually filtering the list of names of popular authors, singers, actresses, and other female celebrities that we sourced from nationality category lists in the Serbian edition of Wikipedia.

With the exception of *Rijana* ('Rihanna', a Barbadian singer), most names belong to American,

Canadian, British or Australian figures. Additionally, five names belonging to politicans and other personalities are added, these being *Hilari Klinton* 'Hilary Clinton', *Margaret Tačer* 'Margaret Thatcher', *Sara Pejlin* 'Sarah Palin', *Kondoliza Rajs* 'Condoleezza Rice' and *Monika Levinski* 'Monica Lewinsky'. All names are converted from Serbian Wikipedia's default Serbian Cyrillic script to Serbian Latin script using an online converter and then edited for capitalisation errors. In some rare cases, we added names that we found in the corpus scraping phase into the name list alongside the names found on Wikipedia. This includes some doublets such as *Šeril Sandberg* ('Sheryl Sandberg', former chief operating officer of Meta Platforms), whose name is also spelt *Šeril Sendberg*, and *Anđelina Džoli* ('Angelina Jolie', spelt in Wikipedia as *Anđželina Džoli* but the former spelling is more commonly attested). These doublets are caused by ambiguities that arise when converting names to the Serbian phonetic system. Given that it is not possible to ensure that the models treat these doublets as the same name, they are treated as names of different people.

A few names are altered entirely from the Wikipedia titles. These names included the names of two rappers Saweetie and Megan Thee Stallion, whose names are replaced with their phonetic equivalents, *Saviti* and *Megan Di Stalion* respectively, as reflected by their spelling in Serbian tabloids. Conversely, two phonetic spellings of names, *Uma Terman* and *Vira Farmiga* are replaced with their corpus-attested spellings, *Uma Turman* and *Vera Farmiga* respectively, despite not reflecting the actual pronunciation of the names. One mononym, *Šeril* ('Cheryl', an English singer) is changed to *Šeril Kol* to avoid conflicts with other people named *Šeril*.

Finally, a number of names are pruned from the database. In cases where there are multiple people of the same name, duplicate entries are removed and treated as the same person. Some mononyms, are also removed for causing conflicts with common words. These names include *Niko*, the Serbian transliteration of American singer Nico, which is removed for being too similar to the common Serbian word *niko* 'no one'. *Keša* ('Ke$ha') is removed for being too similar to the genitive form of the slang word *kеš*. Additionally, three mononyms are removed for being too similar to Balkan names: Selena, the mononym of singer Selena Pérez, is removed for being a very common Serbian name, *Monika*, a common Croatian name, and *Alija*, a Bosnian name. Three more mononyms, *Benks*, *Eš*, and *Pink*, are also removed for being too common. *Lenka* is removed for causing conflicts, as is *Sijera*. *Vivika A. Foks* is removed due to the middle initial consisting of just 'a', causing a conflict in some of the evaluation procedures. In total, 1323 names are included, of which 812 names are completely indeclinable, meaning the name does not include any declinable element. 511 contain at least one declinable element, of which 13 appear to be of Southern Slavic origin. 30 names are fully declinable.

B Sentence type templates

This appendix provides an overview of the templates that were used to generate the sentences for the Masked Language Modeling task.

B.1 Low-context sentences

B.1.1 Open-ended sentence

(1) [NAME] je [MASK] .
[NAME] be.3.SG.PRS [MASK]
'[NAME] is [MASK] .'

B.1.2 Adjective sentences

These sentences use adverbs to encourage an adjective to be produced.

(2) [NAME] je veoma [MASK] .
[NAME] be.3.SG.PRS very [MASK]
'[NAME] is very [MASK] .'

(3) [NAME] je takođe [MASK] .
[NAME] be.3.SG.PRS also [MASK]
'[NAME] is also [MASK] .'

(4) [NAME] je vrlo [MASK] .
[NAME] be.3.SG.PRS very [MASK]
'[NAME] is very [MASK] .'

(5) [NAME] je sada [MASK] .
[NAME] be.3.SG.PRS now [MASK]
'[NAME] is now [MASK] .'

(6) [NAME] je trenutno [MASK] .
[NAME] be.3.SG.PRS currently [MASK]
'[NAME] is currently [MASK] .'

(7) [NAME] je [MASK] širom
[NAME] be.3.SG.PRS [MASK] throughout
svet-a.
world-SG.GEN
'[NAME] is [MASK] throughout the world.'

B.1.3 Past participle sentences

These sentences are constructed to be filled with a past participle.

(8) [NAME] je [MASK] u
[NAME] AUX.3.SG.PRS [MASK] in
grad-Ø juče.
city-SG.ACC yesterday
'[NAME] was [MASK] in the city yesterday.'

(9) [NAME] je [MASK]
[NAME] AUX.3.SG.PRS [MASK]
knjig-u juče.
book-SG.ACC yesterday
'[NAME] was [MASK] a book yesterday.'

(10) [NAME] je [MASK] da
[NAME] AUX.3.SG.PRS [MASK] REL
ode.
leave-3.SG.PRS
'[NAME] [MASK] to leave.'

B.1.4 Possessive sentences

(11) [NAME] i [MASK] otac
[NAME] and [MASK] father-M.SG.NOM
razgovar-aju.
converse-3.PL.PRS
'[NAME] and [MASK] father are conversing.'

(12) [NAME] i [MASK] drugaric-a
[NAME] and [MASK] friend.F-F.SG.NOM
razgovar-aju.
converse-3.PL.PRS
'[NAME] and [MASK] friend-M.SG.NOM are conversing.'

(13) [NAME] i [MASK] pas
[NAME] and [MASK] father-M.SG.NOM
šet-aju se.
walk-3.PL.PRS REFL
'[NAME] and [MASK] dog are walking.'

B.1.5 Plural past participles sentences

These sentences explore how the models handle feminine plural past participles.

(14) [NAME] i jedn-a
[NAME] and one-F.SG.NOM
žen-a [MASK] su ovde
woman-F.SG.NOM AUX.3.PL.PRS here bit
malo ranije.
earlier
'[NAME] and some woman were here a bit earlier.'

(15) [NAME] i njen-a
[NAME] and her-F.SG.NOM
sestr-a [MASK] su ovde
woman-F.SG.NOM AUX.3.PL.PRS here bit
malo ranije.
earlier
'[NAME] and her sister were here a bit earlier.'

(16) [NAME] i [MASK] sestr-a
[NAME] and [MASK] sister-F.SG.NOM
bil-e su ovde malo
be-PTCP.F.PL AUX.3.PL.PRS here bit
ranije.
earlier
'[NAME] and [MASK] sister were here a bit earlier.'

B.1.6 Adjective embedded clauses

These sentences are constructed to be completed with an adjective inside an embedded clause.

(17) Veruj-e se da je
believe-3.SG.PRS REFL REL be.3.PL.PRS
[NAME] trenutno [MASK] .
[NAME] currently [MASK]
'It is believed that [NAME] is currently [MASK] .'

(18) Izjavil-o se da je
announce-PTCP.N.SG REFL REL be.3.PL.PRS
[NAME] trenutno [MASK] .
[NAME] currently [MASK]
'It was announced that [NAME] is currently [MASK] .'

B.2 High-context sentences

High-context sentences consists of two parts: a contextual sentence followed by one of three masked sentences.

B.2.1 Serbian names

These are the names used in the high-context sentences, taken from lists of most common Serbian names. Three are feminine, and three are masculine.

Feminine	Dragana, Jelena, Milica
Masculine	Marko, Ivan, Vladimir

B.2.2 Contextual sentence

Contextual sentences contain a common Serbian name [SN] as the subject or agent of a sentence, followed by one of the target names at the end in one of the cases.

Nominative

(19) [SN] je viši/viša nego [NAME]
[SN] be.3.PL.PRS taller than [NAME]
'[SN] is taller than [NAME]'

Genitive

(20) [SN] je velik-i
[SN] be.3.PL.PRS big-M.SG.NOM
fan-Ø [NAME]
fan-M.SG.NOM [NAME]
'[SN] is a big fan of [NAME]'

(21) [SN] se plaši [NAME]
[SN] REFL fear-3.SG.PRS [NAME]
'[SN] is afraid of [NAME]'

(22) [SN] stiže kod [NAME]
[SN] arrive-3.SG.PRS by [NAME]
'[SN] is arriving at [NAME]'s house'

Dative/Locative

(23) [SN] se divi [NAME]
[SN] REFL admire-3.SG.PRS [NAME]
'[SN] admires [NAME]'

(24) [SN] daj-e poklon [NAME]
[SN] give-3.SG.PRS gift-M.SG.NOM [NAME]
'[SN] gives a gift to [NAME]'

(25) [SN] čita članak o
[SN] read-3.SG.PRS article-M.SG.NOM about
[NAME]
[NAME]
'[SN] reads an article about [NAME]'

Accusative

(26) [SN] voli [NAME]
[SN] love-3.SG.PRS [NAME]
'[SN] loves [NAME]'

(27) [SN] ne zn-a za [NAME]
[SN] NEG know-3.SG.PRS for [NAME]
'[SN] do not know of [NAME]'

Instrumental

(28) [SN] se druž-i sa [NAME]
[SN] REFL socialise-3.SG.PRS with [NAME]
'[SN] is hanging out with [NAME]'

(29) [SN] id-e u centar-Ø
[SN] go-3.SG.PRS in centre-M.SG.NOM
grad-a sa [NAME]
city-M.SG.GEN with [NAME]
'[SN] is going downtown with [NAME]'

B.2.3 Masked sentences

Each contextual sentence is paired with one of three mask sentences.

(1) [NAME] je [MASK] .
[NAME] be.3.SG.PRS [MASK]
'[NAME] is [MASK] .'

(2) [NAME] je vrlo [MASK] .
[NAME] be.3.SG.PRS very [MASK]
'[NAME] is very [MASK] .'

(3) [NAME] je [MASK]
[NAME] AUX.3.SG.PRS [MASK]
knjig-u .
book-SG.ACC
'[NAME] was [MASK] a book yesterday.'

C NER results for spaCy baseline

This appendix shows the NER result visualizations for the spaCy baseline separated by name type, including indeclinable (IND), declinable (DEC), Slavic (SLV) or fully declinable (FUL).

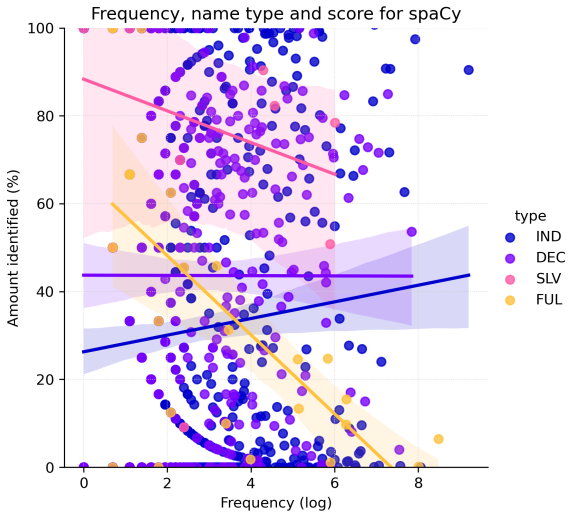


Figure 5: spaCy NER results per name type