ANALYSING ACCURACY OF SLOVAK LANGUAGE LEMMATIZATION AND MSD TAGGING

Radovan Garabík – Denis Mitana

Jazykovedný ústav Ľ. Štúra SAV, v. v. i.
Panská 26, Bratislava
E- mail: radovan.garabik@kassiopeia.juls.savba.sk, denis.mitana@korpus.juls.savba.sk

GARABÍK, R. – MITANA, D. (2023): Analysing Accuracy of Slovak Language Lemmatization and MSD Tagging. In: *Slovenská reč*, 88/2, 129 – 140.

Abstract: Lemmatization and morphological tagging is an indispensable step in Slovak corpus linguistics. In this article, we evaluate two state-of-the-art Slovak language lemmatizers and MSD taggers. One is based on MorphoDiTa and the other is based on spaCy. We measured accuracy on the test subset of manually lemmatized and MSD annotated corpus and found that the combination of lemma and tag achieved 93.5% accuracy with MorphoDiTa, and 95.6% accuracy with spaCy. Most of the errors occurred in disambiguating MSD tags for homonymous uninflected parts of speech such as particles, conjunctions, and adverbs, and in disambiguating singular masculine inanimate nominative and accusative. In these cases, spaCy shows a noticeable improvement over MorphoDiTa, likely due to a better exploitation of the context of the words.

Keywords: lemmatization, MSD tagging, POS tagging, Slovak.

Article in brief:
- We analyze the accuracy of automatic lemmatization and morphosyntactic description of Slovak using two taggers, MorphoDiTa and spaCy.
- The accuracy of lemma + morphosyntactic tag is 93.5% for MorphoDiTa, 95.6% for spaCy.
- The accuracy of lemmatization is 99% for MorphoDiTa, 98.2% for spaCy.

Článok v skratke:
- Príspevok analyzuje presnosť automatickej lematizácie a morfologickej anotácie slovenčiny použitím dvoch nástrojov, MorphoDiTa a spaCy.
- Presnosť kombinovanej lematizácie a morfologickej anotácie je 93,5 % pre MorphoDiTa, 95,6 % pre spaCy.
- Presnosť lematizácie je 99 % pre MorphoDiTa, 98,2 % pre spaCy.

1. INTRODUCTION

Slovak, as a "typical" Slavic language, belongs to the group of moderately inflected languages. It has three or four genders and two grammatical numbers, which interact with the inflections in somewhat complex and unpredictable ways. Inflections are primarily realized by suffixes, but they exhibit numerous irregularities. One suffix encodes several grammatical categories, and the same suffix often reflects unrelated features. In other words, Slovak is a typical inflectional language that is not amenable to heuristic analysis.

Due to the nature of Slovak inflections, lemmatization is often an essential step in various text processing tasks, such as full-text search. Moreover, full morphosyntactic analysis or description (MSD) serves as the core of corpus linguistic research. Although simpler part-of-speech tagging (POS) is sometimes used, it is less useful for Slovak compared to morphologically simpler languages. In this article, we will use the term *tagging* to describe the process of automatically assigning morphosyntactic tags to individual word forms, including disambiguating possible multiple interpretations.[1]

Representative Slovak language corpora (Slovenský národný korpus 2020; Benko 2014) typically rely on lemmatization and MSD tagging using a morphological database developed at the Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, along with the MorphoDiTa tagger (Straková et al. 2014), which has been trained on a manually annotated Slovak language corpus[1]. Due to the vital role of lemmatization and MSD tagging in the field of Slovak corpus linguistics, it is important to acknowledge its limitations and recognize achievable levels of accuracy. However, to the best of our knowledge, no systematic evaluation has been published to date, though some preliminary and partial results were presented in Garabík – Mitana (2022). There were also internal evaluations conducted on previous versions of the morphological database, manually annotated corpus and MorphoDiTa. We want to address the situation by describing the current state of the lemmatization and MSD tagging, and by evaluating the accuracy of various common usage scenarios.

In corpus linguistics, the most frequently used outputs for moderately inflected languages are either lemma or lemma+MSD. In certain situations, users are also interested in case-insensitive lemmas.

1.1. MORPHOLOGICAL DATABASE

The morphological database forms the basis of subsequent linguistic analysis and comprises triplets of word form, lemma, and a morphological tag. At the time of writing, the database used for training MorphoDiTa consists of 3 816 295 entries (i.e. distinct word-lemma-MSD combinations), 114 634 unique lemmas, and 1 330 039 unique wordforms.

As anticipated, there are numerous words (potentially unlimited) absent from the database, and any sufficiently accurate tagger should take into account these words. If feasible, the tagger should make an effort to determine a plausible lemma and a MSD for them. This process is sometimes referred to as *guessing*. In the article,

[1] Some authors divide the process into two steps: "tagging" which assigns all possible or plausible combinations of lemma+MSD tag to the word, and the disambiguation, which selects the correct combination. For the sake of simplicity, we use the term *tagging* to refer to both steps.

we will refer to the word forms absent from the morphological database as out-of--vocabulary words, or OOV (words).

1.2. MANUALLY ANNOTATED CORPUS

Manually annotated (i.e. lemmatized and morphologically tagged) corpus *r-mak-6.0* comprises 1 199 793 tokens (137 505 unique, case sensitive), 55 090 unique lemmas, and 1354 unique morphosyntactic tags[1]. The corpus is composed of 30.5% journalistic, 50.5% fiction, 19.0% professional (scientific, technical) texts, in 720 documents, 77 671 sentences. For the purpose of training the taggers, the corpus was randomly divided with stratification over documents into *train*, *dev,* and *test* segments. The *train* segment contains 62 136 sentences, 959 737 tokens; the *dev* segment 7 767 sentences, 120 984 tokens; the *test* segment 7 768 sentences, 119 072 tokens.

1.3. MORPHODITA

MorphoDiTa: Morphological Dictionary and Tagger is an open-source tool for morphological analysis of natural language texts. It performs morphological analysis, morphological generation, tagging, and tokenization. The tool is available as both a standalone application and a library and comes with pre-trained linguistic models (Straková – Straka – Hajič 2016). While initially developed for Czech language lemmatization and tagging, MorphoDiTa is language independent (at least for languages of similar inflectional complexity).

MorphoDiTa has been widely utilized in Slovak language corpora, and a web interface has been developed to provide users with access to lemmatization and tagging (Garabík – Bobeková 2021). Evaluating various experiments with the tagger features (such as combinations of capitalization, word forms, lemmas, and parts of tags) lead to re-using the feature file used for the Czech model. This result is not surprising given the similarities of Czech and Slovak.

MorphoDiTa contains a statistical guesser for OOV words, trained on suffixes. We do not use prefixes for training the guesser (their function in Slovak morphology is mostly limited to superlatives, verb negation, and is connected to the verbal aspect). The training process of the guesser considers all the words in the training data, with configurable maximum suffix length (the suffix alone is used for training the guesser) and number of rules per suffix. We selected maximum suffixes of length at most 3, and 8 rules per suffix – these parameter give the best accuracy for Slovak, although the differences compared to other close values are only marginal.

To improve the accuracy of the guesser on real-world texts, we postprocess the guesser output and filter the list of possible lemmas to prefer tokens that appeared (as raw word forms) at least once in the corpora *prim-9.0-juls-sane* and *Araneum*

Slovacum IV Maximum (Benko 2014). We use a bloom filter to store the information about the word forms from the corpora in a space-efficient and performance favourable way. We also include several heuristic rules to filter out implausible combinations of lemmas and tags (e.g. if the word does not start with the prefix *naj-* we remove the tags indicating a superlative from the list of possible tags, or if the word does not start with the prefix *ne-* we remove the tags indicating a negated verb) and directly assign tags for numbers (written in Arabic digits), punctuation, and symbols.

1.4. SPACY

SpaCy[2] is an open-source tool for Natural Language Processing (NLP). It encompasses various components for Named Entity Recognition, Part-of-Speech tagging, dependency parsing, sentence segmentation, text classification, lemmatization, MSD tagging, entity linking, and more. Additionally, spaCy is language--independent and available as a Python package, offering support (in various levels of completeness) for over 72 languages.

SpaCy is a production-ready NLP library that comes with state-of-the-art NLP architectures such as the Transformer. It is easily extensible with custom components, and allows users to utilize custom models in the most frequently used deep learning frameworks like PyTorch and TensorFlow. Furthermore, it provides an easy to use training system, as well as model packaging and deployment.

As for the Slovak language, spaCy official support is limited to stop words and lexical attributes for general numerals. However, one online tool (Wencel 2021) is available that performs tokenization, MSD tagging, and dependency parsing. Regrettably, this tool is not described in detail and is unavailable as a free-to-use model in spaCy.

We deploy only MSD tagging and lemmatization components. For the morphological analysis component, we used the Transformer architecture based on the pre-trained multilingual BERT language model (Devlin 2019). Specifically, we used a *bert-base-multilingual-uncased* model that has 110 million parameters and 12 layers where each layer has 12 attention heads. We finetuned the model for the MSD tagging task on the *train* segment of the manually annotated corpus. We also also optimized the model hyperparameters on the *dev* segment of the corpus (see section 1.2.).

The lemmatization component is not trainable and thus is rule-based only. The rules applied are as follows, in this order:

1. If a given pair of word form and MSD tag is in the morphological database (see Section 1.1.), then use the assigned lemma.

2. Try to lemmatize a given pair of word form and MSD tag using the morphological suffix database.

To improve accuracy, we use postprocessing to assign tags directly for numbers, punctuation, and other symbols, following the same methodology as used in the Slovak MorphoDiTa tagging.

2. TRAINING AND EVALUATION

2.1. MORPHODITA

We present a summary of the accuracy of MorphoDiTa output in Table 1. For the subsequent tables, we take our evaluation on the lemma+MSD accuracy[2] as the baseline, with error rate defined as *errorrate = 1 − accuracy*. To indicate the change in error rate, we express it as a percentage reduction of the original error rate. A positive number thus means an improvement, while a negative number indicates a decrease in accuracy.

This provides the reader with an immediate overview of the changes in tagging errors depending on the output we measure. In the subsequent tables, the part of speech (POS) refers to the word's part of speech without any other grammatical categories (this is indicated by the first character of the MSD in the Slovak tagset we use).

Table 1: Accuracy of various (lemma, case insensitive lemma, MSD, POS, lemma+MSD, lemma+POS, case insensitive lemma+MSD) token annotations by Slovak MorphoDiTa. The entry in boldface is, in the authors' opinion, the most relevant one for linguistic purposes and serves as a baseline throughout the article. The accuracy was measured on the *test* segment of the manually annotated corpus (7 768 sentences).

	accuracy	*error rate*	*error rate decrease [%]*
Lemma only	0.9824	0.0176	73.0
Lemma (case insensitive) only	0.9895	0.0104	83.8
MSD only	0.9419	0.0581	10.6
POS only	0.9806	0.0194	70.2
Lemma+MSD	0.9350	0.0640	0.0
Lemma (case insensitive) + MSD	0.9403	0.0597	8.2
Lemma+POS	0.9687	0.0313	51.9

To estimate the role of the size of the morphological database, we select only OOV (as defined in section 1.1.) words and see the accuracy of the statistical guesser, described in Table 2. We observe that there remains a significant scope for accuracy improvement through the implementation of postprocessing or normalisation of lemma casing, which is indeed one of the planned features of the Slovak tagging process.

[2] Here, MSD is the complete morphosyntactic tag, as assigned by the tagging process and as used in several major Slovak corpora.

Table 2: Accuracy of various token annotations by Slovak MorphoDiTa, only word forms not present in the morphological database (OOV), as tested on the same *test* segment of the manually annotated corpus.

	accuracy
Lemma only	0.7760
Lemma (case insensitive) only	0.8824
Lemma+MSD	0.6503
Lemma (case insensitive)+MSD	0.7318

Since the assignment of disambiguated MSD tags and lemmas depends on the (tagged) context of the token, we cannot simply calculate the accuracy for the words present in the database. The presence of neighbouring OOV words might negatively affect the tagging. Therefore, we have created a filtered version of the *test* data, selecting only sentences where no OOV word appeared. The size of this data is 71 424 tokens, which represents 60% of the original *test* set. The accuracy of tagging is presented in Table 3, the error rate decrease is calculated with regard to the baseline (from Table 1). In a way, this represents an "ideal" achievable accuracy if we continue improving the coverage of the morphological database coverage indefinitely. However, extending the database also introduces additional homonymy, particularly for rare words and proper names.

Table 3: Accuracy of annotations by Slovak MorphoDiTa. No OOV words.

	accuracy	*error rate*	*error rate decrease [%]*
Lemma only	0.9909	0.0091	86.0
Lemma (case insensitive) only	0.9935	0.0065	90.1
MSD	0.9505	0.0495	23.9
POS	0.9848	0.0152	76.6
Lemma+MSD	0.9476	0.0524	19.5
Lemma (case insensitive)+MSD	0.9498	0.0502	22.8
Lemma+POS	0.9783	0.0217	66.6

2.1.1. Notable sources of errors (MorphoDiTa)

For the sake of brevity, we will focus on discussing only some of the significant sources of errors in the tagging process (in all the following text, we are working with the complete *test* set, i.e. including OOV words.). When analyzing the differences in POS tagging presented in Table 4, one noticeable source of errors is the (relative) inability to correctly distinguish between certain conjunctions (MSD tag O) and particles (MSD tag T). These differences account for 22.4% of the errors in POS assignment. Incorrect tagging of particles (tag T) as adverbs (tag D) and vice versa accounts for an additional 12.5% of errors. Erroneous tagging of nouns (tag S)

as foreign language elements (tag %) and vice versa accounts for additional 11.9% of the errors. The most notable examples of confusion between conjunctions and particles include words *a*, *ale*, *aj.* Confusion between particles and adverbs occurs with words *už* and *však*, where the distinction is purely syntactic or even semantic, often posing challenges even for trained linguists. To illustrate the impact, unifying the conjunctions and particles leads to an improvement in *lemma+MSD* accuracy from 0.9350 to 0.9393.

Examining the accuracy of MSD tags in Table 5, the most frequent error observed is once again the confusion between conjunctions (MSD tag O) and particles (MSD tag T), which accounts for 7.3% of the errors. This is followed by masculine inanimate nouns in the singular, where the errors are in incorrectly identifying the nominative and accusative cases (MSD tags SSis1 and SSis4), constituting 4.14% of the errors. In these cases, the nominative and accusative forms are morphologically identical. In total, errors in distinguishing the nominative and accusative cases, since they have an identical form, make up 22.7% of the errors in the tags.

Table 4: Notable sources of errors in POS tagging by Slovak MorphoDiTa. There are 2 309 errors in POS tagging.

source of error	*number of errors*	*ratio [%]*
O ↔ T	517	22.4
D ↔ T	288	12.5
% ↔ S	274	11.9
O ↔ P	154	6.7

Table 5: Notable sources of errors in MSD tagging by Slovak MorphoDiTa. There are 6 923 errors in MSD tagging.

source of error	*number of errors*	*ratio [%]*
O ↔ T	502	7.3
SSis1 ↔ SSis4	287	4.2
Dx ↔ T	285	4.1
SSns1 ↔ SSns4	136	2.0

2.2. SPACY

Similar to the previous evaluation, we summarise the accuracy of spaCy output in Table 6. As we can see, spaCy achieves higher accuracy in tagging. We suppose that more complex Transformer architecture handled a large number of output tags better. On the other hand, spaCy is worse in a lemmatization apparently due to the rule-based approach. Overall, for the combination of lemma+MSD, spaCy overcame MorphoDiTa.

Table 6: Accuracy of various token annotations by Slovak spaCy. The entry in boldface is a baseline for spaCy comparison.

	accuracy	*error rate*	*error rate decrease [%]*	*error rate decrease w.r.t. MorphoDiTa baseline [%]*
Lemma only	0.9823	0.0177	59.6	72.7
Lemma only (case insensitive)	0.9879	0.0121	72.4	81.4
MSD	0.9654	0.0346	21.3	46.8
POS	0.9847	0.0153	65.2	76.5
Lemma+MSD	0.9561	0.0439	0.0	32.5
Lemma (case insensitive)+MSD	0.9605	0.0395	10.2	39.3
Lemma + POS	0.9706	0.0294	33.1	54.8

2.2.1. Notable sources of errors (spaCy)

Upon examining Tables 7 and 8, it becomes evident that the primary sources of errors in POS and MSD tagging share similarities with those present in MorphoDiTa. Of these errors, the most frequent is the confusion between conjunctions (tag O) and particles (tag T). It is worth noting that in Slavic linguistics, particles are considered a separate part of speech for uninflected words that do not align with other POS categories and modify the meanings of the context, hence why particle identification does not quite fall under morphological distinctions, but rather a syntactic one.

One of the most relevant improvements that spaCy has over MorphoDiTa is its ability to disambiguate between singular masculine inanimate nominative (tag SSis1) and accusative (tag SSis4). Here, spaCy reduces the number of errors by two-thirds compared to MorphoDiTa. In this case, the accusative is indistinguishable from the nominative, and apparently the rather flexible word order in Slovak necessitates a better utilization of the context (or a larger context) to determine the correct case.

Table 7: Notable sources of errors in POS tagging by Slovak spaCy. There are 1 821 errors in POS tagging.

source of error	*number of errors*	*ratio [%]*
O ↔ T	469	25.8
D ↔ T	255	14.0
O ↔ P	164	9.0
% ↔ S	136	7.5

Table 8: Notable sources of errors in MSD tagging by Slovak spaCy. There are 4 119 errors in MSD tagging.

source of error	*number of errors*	*ratio [%]*
O ↔ T	453	11.1
Dx ↔ T	254	6.2
O ↔ PD	141	3.5
SSis1 ↔ SSis4	95	2.3

3. COMPARISON WITH OTHER MODELS AND LANGUAGES

In Table 9, we succinctly compare the accuracies between our Slovak models (*sk MorphoDiTa* and *sk spaCy*), Czech MorphoDiTa and state-of-the-art Czech lemmatization and MSD tagging models for different combinations of lemmas, MSD and POS tags (where available). Accuracy numbers for the state-of-the-art Czech models were taken from the MorphoDiTa manual and for the BERT from the study by Straka et al. (2019). In table 9, *sk* and *cs* denote Slovak and Czech languages, respectively. *cs BERT+WE+Flair* stands for Czech BERT model with word2vec and Flair embeddings (the best results). *cs BERT* (including morphological dictionary) is therefore probably the most directly comparable to our spaCy model (*sk spaCy*).

sk SlovakBERT result is from the study Pikuliak et al. (2022); for comparable purposes, the authors evaluated only the part of speech tagging accuracy. We emphasize that the accuracies of other models were obtained using different testing data (obviously in the case of the Czech models) and probably using slightly different methodologies. This should be taken into consideration when drawing conclusions from the results.

Table 9: Comparison of our models, SlovakBERT and the Czech models.

	sk MorphoDiTa	cs MorphoDiTa	sk spaCy	cs BERT	cs ERT+WE+Flair	sk SlovakBERT
Lemma	0.9824	0.9786	0.9823	0.9894	0.9898	
Lemma+MSD	0.9350	0.9506	0.9561	0.9751	0.9898	
Lemma+POS	0.9687	0.9766	0.9706			
POS	0.9806	0.9901	0.9847	0.9925	0.9934	0.9837
MSD	0.9419	0.9555	0.9654	0.9791	0.9805	

4. CONCLUSION

We calculated the accuracy of two state-of-the-art Slovak language lemmatizers and MSD taggers, one based on MorphoDiTa and the other one on spaCy. MorphoDiTa reaches 93.5% accuracy on the lemma+MSD combination; 96.9% on the lemma+POS, and if we are interested only in lemmas, the accuracy is 98.2%. Neglecting differences in case, the accuracy rises to 94.0% for the lemma+MSD, and 99.0% for the lemmas only. The previous numbers include words not present in the morphological database (OOV), these are lemmatized by a statistical guesser; if we limit ourselves to known words, the lemma+MSD accuracy will be 94.8%,

lemma+POS 97.8%, and lemmas 99.1%. SpaCy reaches 95.6% accuracy on the lemma+MSD combination, 97.1% on the lemma+POS, and 98.2% on lemmas only. Similar to MorphoDiTa, case insensitive accuracy on the lemma+MSD is 96.0% and 98.8% on the lemmas only. One of the most significant improvements of spaCy over MorphoDiTa is its ability to disambiguate between singular masculine inanimate nominative and accusative cases. This requires a better utilization of context to identify the correct case. It follows that spaCy can utilize the context better than MorphoDiTa, thanks to using the BERT model. In our future work, we plan to add postprocessing and normalization of lemma casing (so that the assigned lemma capitalization follows the "most likely" variant) and test new Slovak large language models.

The Slovak MorphoDiTa and the spaCy models are the de facto standards in lemmatization and MSD tagging of Slovak corpora at the Ľ. Štúr Institute of Linguistics (corpora of the Slovak National Corpus and other publicly available corpora). These corpora are extensively utilized in linguistic research throughout Slovakia. In this light, it is essential for corpus users to be aware of annotation accuracy and the prevalent sources of errors, in order to enhance the precision of their research or to take inaccuracies into account. This article aims to address this long-standing gap in Slovak corpus linguistics and provide a source of baseline accuracies for further analysis of NLP tools for Slovak.

SLOVENSKÉ RESUMÉ

V článku opisujeme a bližšie analyzujeme presnosť dvoch lematizátorov a morfologických taggerov pre slovenčinu, jedného používajúceho software MorphoDiTa a druhého spaCy.

MorphoDiTa dosahuje presnosť 93,5 % pre kombináciu lema + morfosyntaktická značka (tag); 96,9 % pre kombináciu lemy a slovného druhu, a ak určujeme iba lemmy, presnosť je 98,2 %. Pri zanedbaní rozdielov v malých a veľkých písmenách sa presnosť zvýši na 94,0 % pri určovaní kombinácie lema + tag a 99,0 % pri určovaní samostatnej lemy. Predchádzajúce čísla zahŕňajú slová, ktoré sa nenachádzajú v morfologickej databáze (OOV) a ktoré sú lematizované štatistickým guesserom; ak sa obmedzíme na známe slová, presnosť (so zachovaním veľkostí písmen) lema + tag bude 94,8 %, lema + slovný druh 97,8 % a samostatná lema 99,1 %.

Najčastejšie chyby pri určovaní slovných druhov prostredníctvom MorphoDita sú zámeny medzi homonymnými spojkami a časticami (22,4 % chýb); a medzi príslovkami a časticami (12,5 % chýb). Najčastejšie chyby pri určovaní morfosyntaktickej značky sú zámeny medzi spojkami a časticami (7,3 % chýb); medzi nominatívom a akuzatívom mužských neživotných substantív (4,2 % chýb) a medzi príslovkami v prvom stupni a časticami (4,1 % chýb).

SpaCy dosahuje presnosť 95,6 % pre kombináciu lema + morfosyntaktická značka, 97,1 % pre kombináciu lemy a slovného druhu, a 98,2 % pre lemy. Podobne ako MorphoDiTa pri zanedbaní rozdielov v malých a veľkých písmenách sa presnosť zvýši na 96,0 % pre kombináciu lema + morfosyntaktická značka a 98,8 % pre lemy.

Najčastejšie chyby pri určovaní slovných druhov prostredníctvom spaCy sú zámeny medzi homonymnými spojkami a časticami (25,8 % chýb); a medzi príslovkami a časticami (14,0 % chýb). Najčastejšie chyby pri určovaní morfosyntaktickej značky sú zámeny medzi spojkami a časticami (11,1 % chýb); medzi príslovkami v prvom stupni a časticami (6,2 % chýb); a medzi spojkami a príslovkovými zámenami (3,5 % chýb).

Najviditeľnejšie zlepšenie presnosti pri použití spaCy oproti MorphoDiTa je v dezambiguácii neživotných maskulín v singulári nominatívu a akuzatívu a v odlíšení častíc od homonymných spojok a príslovkiek. Tieto zlepšenia sú zrejme dôsledkom lepšieho využitia kontextu v BERT modeloch a poukazujú na perspektívy využitia veľkých jazykových modelov v lingvistike.

Literature

BENKO, V. (2014): Aranea: Yet Another Family of (Comparable) Web Corpora. In: Sojka, P. – Horák, A. – Kopeček, I. – Pala, K. (eds.): *Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic. Proceedings. LNCS 8655.* Switzerland: Springer, 257 – 264.

BENKO, V. – GARABÍK, R. (2018): Ensemble Tagging Slovak Web Data. In: *SlaviCorp 2018*. *Book of Abstracts.* Prague: Charles University, 26 – 28.

DEVLIN, J. – CHANG, M.-W. – LEE, K. – TOUTANOVA, K. (2019): BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Volume 1 (Long and Short Papers). Minneapolis (Minnesota): Association for Computational Linguistics, 4171 – 4186.

GARABÍK, R. (2006): Slovak morphology analyzer based on Levenshtein edit operations. In: *Proceedings of the WIKT'06 Conference.* Bratislava: UI SAV, 2 – 5.

GARABÍK, R. – BOBEKOVÁ, K. (2021): Lematizácia, morfologická anotácia a dezambiguácia slovenského textu – webové rozhranie. In: *Slovenská reč*, 86/1, 104 – 109.

GARABÍK, R. – MITANA, D. (2022): Accuracy of Slovak Language Lemmatization and MSD Tagging – MorphoDiTa and SpaCy. In: *LLOD Approaches for Language Data Research and Management.* Abstract Book. Vilnius: Mykolo Romerio universitetas, 93 – 95.

PIKULIAK, M. – GRIVALSKÝ, Š. – KONÔPKA, M. – BLŠTÁK, M. – TAMAJKA, M. – BACHRATÝ, V. – ŠIMKO, M. – BALÁŽIK, P. – TRNKA, M. – UHLÁRIK, F. (2022): SlovakBERT: Slovak Masked Language Model. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*, Association for Computational Linguistics, 7156 – 7168.

Slovenský národný korpus – prim-9.0-juls-sane. (2020): Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2020. Available at: https://korpus.juls.savba.sk

STRAKA, M. – STRAKOVÁ, J. – HAJIČ, J. (2019): Czech Text Processing with Contextual Embeddings: POS Tagging, Lemmatization, Parsing and NER. In: *Proceedings of the 22nd International Conference on Text, Speech and Dialogue – TSD 2019*, *Lecture Notes in Computer Science,* Springer International Publishing, 137 – 150.

STRAKOVÁ, J. – STRAKA, M. – HAJIČ, J. (2016): Open-source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* Baltimore (Maryland): Association for Computational Linguistics, 2014, 13 – 18.
WENCEL, M. (2021): Spracovanie prirodzeného jazyka - SpaCy Web App, 2021. Available at: https://spacy.tukekemt.xyz/analyze_sk

Internet resources

[1] https://korpus.sk/r-mak/ (accessed: 2023-06-01)
[2] https://spacy.io (accessed: 2023-06-01)