From web to dialects: how to enhance non-standard Russian lects lemmatisation?

Ilia Afanasev
HSE University
Moscow, Russia
ilia.afanasev.1997@gmail.com

Olga Lyashevskaya
HSE University
Moscow, Russia
Vinogradov Russian Language Institute RAS
Moscow, Russia
olesar@yandex.ru

Abstract

The growing need for using small data distinguished by a set of distributional properties becomes all the more apparent in the era of large language models (LLM). In this paper, we show that for the lemmatisation of the web as corpora texts, heterogeneous social media texts, and dialect texts, the morphological tagging by a model trained on a small dataset with specific properties generally works better than the morphological tagging by a model trained on a large dataset. The material we use is Russian non-standard texts and interviews with dialect speakers. The sequence-to-sequence lemmatisation with the help of taggers trained on smaller linguistically aware datasets achieves the average results of 85 to 90 per cent. These results are consistently (but not always), by 1-2 per cent. higher than the results of lemmatisation with the help of the large-dataset-trained taggers. We analyse these results and outline the possible further research directions.

1 Introduction

Lemmatisation is a natural language processing (NLP) task that is a part of the basic language resource toolkit (BLARK) (Krauwer, 1998, 2003; Piotrowski, 2012). Lemmatisation may be defined as a transformation of a given token into the dictionary form, the latter being called a lemma. There may be different ways of lemmatisation, such as classifying a token by its particular supposed lemmatisation rule and the subsequent transformation by this rule (for instance, such model may classify *shown* into the group of tokens that are lemmatised with «delete last *n* and then add to before the token» rule, and then transformed by this rule into *to show*) (Anastasyev, 2020). In this paper, we focus on the sequence-to-sequence approach, which takes input sequence and transforms it into output sequence directly (Sutskever et al., 2014; Cho et al., 2014).

Sequence-to-sequence approach generally requires additional information for the token, because it is difficult for the model to lemmatise bare tokens (Kanerva et al., 2021). Many smaller lects[1] do not possess gold morphological tagging. However, they are located nearby a closely-related high-resource lect, for which there are a lot of gold morphological datasets.

We hypothesise that there is a reliable way to find a dataset with the specific distributional properties, train a tagger on it, use this tagger on a new, rather different dataset, and then lemmatise the tokens of this dataset with a preliminary fine-tuned large language model. We also presume that this approach is preferable to gathering the biggest data amount possible.

We believe that the thoroughness in morphological training data selection becomes gradually more important with increasing variation within the lemmatisation evaluation data. So, if overall the better tactic is to get the largest and the most heterogeneous dataset possible, for some types of data one needs a more nuanced approach.

We are going to demonstrate this on the material of the non-standard Russian lects. This includes the web as corpora material, social media texts, and dialect texts, presenting the continuum of lects getting further away from the standard Russian in terms of distributional properties. We hypothesise the following:

H1: Morphological tagging efficiency directly influences the lemmatisation accuracy.

H2: If the model trains on the larger dataset, the morphological tagging it performs will present stable satisfactory results.

H3: For the non-standard data, the distributional properties of the training dataset are generally more important than the sheer size.

We impose a set of restrictions. Both the model we use for morphological tagging and the lemmat-

[1] In this paper, we use lect as a neutral term for any given language variety, whether it is a standard, a dialect, or a sociolect.

Proceedings of the 2023 CLASP Conference on Learning with Small Data, pages 167–175
September 11–12, 2023.

iser (at least in prediction mode) should be able to run on an individual device with no more than 6GB V-RAM (the specs of NVIDIA GeForce GTX1060, currently the most widespread GPU). The training data may vary in size, however, the datasets that we select on distributional properties basis should not exceed 500 000 tokens.

Section 2 contains previous research on the topics of lemmatisation in general and Russian lemmatisation in particular. In section 3, we describe the data. Section 4 includes a description of the method. Section 5 describes the experiments and the analysis of their results. In section 6, we wrap up the research, stating either confirmation or refutation for each of the hypotheses, as well as the possible future directions of the research.

2 Related Work

Currently, there are two predominant approaches to lemmatisation. The first is the classification approach: the model determines the rule of lemmatisation for a given token and then applies the rule (Mills, 1998; Chrupała, 2006; Plisson et al., 2008; Gesmundo and Samardžić, 2012; Radziszewski, 2013). This approach tends to be monolingual (Anastasyev, 2020; Torre Alonso, 2022). The second is the generally multilingual sequence-to-sequence approach when the input (token and its features) is transformed directly to the output sequence, lemma (Straka and Straková, 2017; Bergmanis and Goldwater, 2018; Kanerva et al., 2021).

Russian lemmatisation currently dominates the East Slavic lemmatisation landscape (Anastasyev, 2020), including historical varieties, with both rule-based and automatic methods (Berdičevskis et al., 2016; Pedrazzini and Eckhoff, 2021). However, territorial lects have not yet gained the same kind of attention, while the lemmatisers designed for specific corpora are not open-source (Kryuchkova and Goldin, 2011, 2015). Russian web as corpora and social media texts are included in the evaluation pipelines but generally are not the centre of attention (Sorokin et al., 2017).

There are different ways to enhance the performance of a lemmatisation model, morphological tagging being the most common (Anastasyev, 2020). The ensemble models that enhance lemmatisation efficiency with external resources (Milintsevich and Sirts, 2021) are gaining popularity, especially for historical low-resource territorial lects (de Graaf et al., 2022). And given that social media texts are similar to them (Piotrowski, 2012), the contemporary vocabulary dictionaries are going to be of use in further research.

3 Data

We employ two groups of datasets: the training datasets and the evaluation datasets. Training datasets are generally well-established through Russian NLP and mostly contain standard Russian texts. The evaluation datasets group contains both the well-established ones and the ones that are not yet heavily adopted in the Russian NLP.

The largest training dataset is a collection of different Russian National Corpus[2] texts that vary diachronically (from the 1700s to 2010s), orthographically (containing texts in modern orthography, as well as premodern, used mostly before 1917), and genre-wise (including news, poetry, fiction, and social media). We later refer to this dataset as RNC-sampled. This dataset contains nearly 2 million tokens. It is also the dataset the lemmatisation model trained on. We also employ two subsets of RNC-sampled. The first one is Taiga (Shavrina and Shapovalova, 2017), which aims to represent texts from social media that demonstrate a higher level of variation and colloquiality. Taiga contains 197 000 tokens. The second is SynTagRus (Droganova et al., 2018), the biggest Universal Dependencies Russian corpus, containing fiction, non-fiction and news texts. The original SynTagRus contains 1.5 million tokens, we downsampled it to 195 000 tokens for effective comparison with Taiga.

We use three sets of data for evaluation. The first is the tagged part of the Russian General Internet Corpus (Belikov et al., 2018), designed for MorphRuEval-2017 (Sorokin et al., 2017). It is 270264 tokens in size. Later we refer to this dataset as GIKRYA. GIKRYA consists of different texts from the Internet, which possess a high degree of variation and lack orthographical normalisation. The tagging of the GIKRYA part that we employ is human-checked.

The second evaluation dataset is the scraped collection of tweets from 2022 to 2023, selected based on them containing words мокша ‘Moksha’, эрзя ‘Erzya’, and Саратов ‘Saratov’. The tweets contain texts from the regional mass media, as well as everyday communication, concerning current politics, by speakers of different origins and

[2] ruscorpora.ru

backgrounds. We slightly manually normalised the texts, correcting the most obvious errors, such as *проектрование > проектирование 'design'. The variation degree in this dataset, despite the minor edits, remains high, mostly due to the non-standard compounds, such as иворовал (<ива + воровал 'willow + steal.PAST.3.SG.M'), and non-standard orthography, for instance, расейськая 'Russian'. We provide human-checked lemmata (without PoS/morphological tagging) for this dataset. We later refer to this dataset as MES-Tweets. MES-Tweets contains 6100 tokens.

The third group of evaluation datasets is the transcribed recordings of interviews with speakers of Russian continuum dialects (small territorial lects) Belogornoje (Saratov Region, Russia, southern type, the territory where Russian speakers arrived after the Russian dialect system had formed), and Megra (Vologda Region, Russia, northern type, the territory where Russian speakers had arrived before the split of the Old East Slavic dialect continuum). We take the material for both Belogornoje and Megra (as we refer to them later) from Saratov dialectological corpus (Kryuchkova and Goldin, 2011, 2015). These datasets are in themselves homogeneous, yet they differ from the training datasets, representing small territorial lects, rather than variation within the standard. Belogornoje and Megra together contain 4372 tokens, with Megra being slightly larger (2856 versus 1516 tokens). Both datasets possess gold lemmatisation and morphological tagging, though annotation schema differences make the use of the latter hardly applicable to this study.

We present the short summary for each dataset in Table 1.

4 Method

To determine the degree, to which the morphological properties of a training dataset may influence the lemmatisation efficiency of an evaluation dataset, we present the following experiment pipeline.

Beforehand, we fine-tune the lemmatisation model with the largest morphologically tagged dataset available, RNC-sampled. The lemmatisation model is a sequence-to-sequence one, employing the BART architecture with the largest number of parameters (430M) (Lewis et al., 2020). This model, BART-large, is used for all the lemmatisation experiments.

For part-of-speech tagging, we use the Stanza tagger (Qi et al., 2018, 2020), modified for the low-resource lects (Scherrer, 2021). We train Stanza on three different datasets, RNC-sampled, Taiga, and SynTagRus-downsampled. RNC-sampled has the largest variation degree and the largest size. Taiga, being relatively smaller, consists of social media texts that inherently possess a high degree of variety. SynTagRus-downsampled is comparable in size to Taiga, but it is much more homogeneous genre-wise.

Training yields three taggers for each of the datasets (RNC-sampled, Taiga, and SynTagRus). We then test these models. For this, we use GIKRYA as a dataset both completely independent from the Russian National Corpus and possessing a significant variation degree. This provides us with the preliminary idea of whether the knowledge acquired through RNC-sampled, Taiga, and SynTagRus-downsampled data, may aid the model in tagging a completely different dataset.

Then we perform the three stages of the lemmatisation experiments. As a baseline for each stage, we use two different tactics. The first is a simple token-to-lemma method when each token is taken as its own lemma. The second is using BART-large on bare tokens (with input in the form of [token] [part-of-speech information] [morphological tagging information] and lemma as a desired output). For each stage, we tag the datasets of GIKRYA (stage 1), MES-Tweets (stage 2) and Belogornoje and Megra (stage 3) with each of the morphological tagging models available, providing silver (non-human-checked, yet performed by a model that generally produces satisfactory results) morphological tagging. The stages represent the growing degree of distance between standard Russian and the variations that form the datasets. After that, we lemmatise each of the acquired datasets with BART-large. We compare the results of the lemmatisation against the baseline. As GIKRYA provides the gold morphological tagging, for stage 1 we also lemmatise tokens with gold tagging to set the highest possible bar.

For evaluation, we use accuracy score, combined with different string similarity measures: Levenshtein distance (Levenshtein, 1966), Damerau-Levenshtein distance (Damerau, 1964), and Jaro-Winkler distance (Jaro, 1989; Winkler, 1990). Levenshtein distance that scores additions, deletions, and substitutions of characters gives a more precise picture of sequence-to-sequence model per-

Dataset name	Dataset group	Previous morphological tagging presence	Token number
RNC-sampled	Training	Present	2000000
Taiga	Training	Present	197000
SynTagRus	Training	Present	1500000
GIKRYA	Evaluation	Present	270264
MES-Tweets	Evaluation	Non-present	6100
Belogornoje	Evaluation	Present (different annotation schema)	1516
Megra	Evaluation	Present (different annotation schema)	2856

Table 1: Datasets used in the study

formance in comparison to the accuracy score, reducing the cost of small mistakes and putting the models that generalise over the models that only memorise. Damerau-Levenshtein distance adds substitutions, providing an even more fine-grained picture. Jaro-Winkler distance shows exactly how well models capture the concept of lemmatisation in Slavic languages, favouring the sequences that match from the beginning. We also use normalised versions of these metrics (Grubbs, 1969). Normalisation generally highlights the ability of a model to generalise: if the normalised score is less than its raw counterpart, the model possibly learned to remember particular token-lemma pairs rather than to lemmatise.

5 Experiments and Analysis

We split the experiments into the morphological tagging section and the lemmatisation section. The lemmatisation section consists of three stages. For the first, we use GIKRYA, the web corpus that contains texts of different genres and variations, some further from the standard Russian than others. The second includes the lemmatisation of the MES-Tweets dataset, which possesses a higher variation degree. For the third, we take dialect data, pushing the ability of the models to generalise to the limit.

5.1 Morphological tagging

The morphological tagging results for GIKRYA are in Table 2.

The model trained on RNC-sampled was overfitting. It has the least out-of-vocabulary rate while performing worse than the models trained on Taiga and SynTagRus-downsampled. The model trained on SynTagRus-downsampled performed the best, especially in the exact match category (UFeats). Probably, the homogeneity and the small size of SynTagRus-downsampled allow the model to concentrate on the morphological tagging concept rather than attempting to grasp variation within it. However, all the models achieved relatively high scores, which may make their tagging relevant for the lemmatisation.

5.2 Lemmatisation (GIKRYA)

The results of measuring the efficiency of GIKRYA, morphologically tagged with these models' lemmatisation (later referred to by the name of the dataset we trained them on), are in Table 3.

The results show that gold tagging predictably is the most desired option for the lemmatiser. Models, however, are still able to easily outperform both baselines. The synTagRus-downsampled-trained model demonstrates the highest accuracy score, while the RNC-sampled-trained one shows the highest Jaro-Winkler distance score. Levenshtein and Damerau-Levenshtein distances, including the normalised ones, are the same. Each model helps the lemmatiser to achieve a consistently high score and to understand that Russian lemmata generally start with the same characters as tokens. Importantly, mistakes that the lemmatiser makes are often caused by differences in the lemmatisation policy and not incorrect morphological tagging, cf. регулировать 'control' instead of регулирующий 'the controlling one': in RNC-sampled, Taiga and SynTagRus-downsampled the participles are treated as verbs and lemmatised to an infinitive, while in GIKRYA the participle is a full-fledged part-of-speech category, and the participles are lemmatised to their nominative singular masculine form.

However, the results do not correlate directly with the morphological tagging results, as the RNC-sampled-trained model performs the worst in morphological tagging, yet here it helps the lemmatiser the most to grasp the concept of lemmatisation, and overall to score pretty well. SynTagRus-downsampled-trained model, the best for morpho-

Training dataset	PoS	PoS+Feats	UFeats	OOV
RNC-sampled	83.17	77.65	54.90	**15.59**
Taiga	**85.57**	80.89	54.75	35.03
SynTagRus-downsampled	**85.57**	**82.51**	**60.69**	34.55

Table 2: The efficiency of GIKRYA dataset morphological tagging with Stanza (Qi et al., 2018, 2020; Scherrer, 2021), evaluated by Micro-F1 score, %.The best results here and after are highlighted in **bold**.

Model	A	L	L(N)	D-L	D-L(N)	J-W	J-W(N)
Token-to-lemma	51.79	0.86	0.84	0.86	0.84	93.99	97.21
Bare token	49.07	0.87	0.85	0.87	0.85	86.34	96.5
RNC-sampled	90.41	0.19	0.19	0.19	0.19	98.94	98.94
Taiga	89.93	0.19	0.19	0.19	0.19	98.92	98.92
SynTagRus-downsampled	90.51	0.19	0.19	0.19	0.19	98.93	98.93
Gold	**94.79**	**0.07**	**0.07**	**0.07**	**0.07**	**99.54**	**99.54**

Table 3: The results of GIKRYA lemmatisation evaluation by accuracy score (A, %), raw (L) and normalised(L(N)) Levenshtein, raw (D-L) and normalised (D-L(N)) Damerau-Levenshtein, raw (J-W) and normalised (J-W(N), %) Jaro-Winkler distances.

logical tagging, enables the lemmatiser to do the best in terms of accuracy, but also the latter gets worse Jaro-Winkler results. Only the Taiga-trained model still lags behind.

Morphological tagging mistakes may play some role in the downfalls of the models, for instance, in cases such as ar instead of ага 'yeah', which the tagger treats as a noun in the genitive singular form, misleading lemmatiser that afterwards applies the wrong tactic.

5.3 Lemmatisation (MES-Tweets)

GIKRYA is still a human-checked, heavily normalised dataset. To get the picture of the model's performance in what is functionally terra incognita, we attempt to lemmatise MES-Tweets. The results are in Table 4.

The results differ from the previous experiments. The tagging by the Taiga-trained model aids lemmatiser the most, even if by a slight margin in each given metric. It seems that here the Taiga lemmatisation approach coincides with the target dataset, as it correctly predicts размышляющий 'the thinking one' as the participle lemma in contrast to the infinitive размышлять 'to think', that, for example, SynTagRus lemmatisation rules propose. It also detects some complex nouns, such as финно-угр 'Finno-Ugric', which, for instance, SynTagRus-downsampled-trained model perceives as an adjective, yielding lemma финно-угрый. Morphological tagging yet again heavily defines the dives in the performance of the lemmatiser, but now the Taiga-trained model is seemingly the best with the given dataset. It may be explained by the closeness of the dataset domains: both Taiga and MES-Tweets are social media texts.

To check this, we turn to the dialect datasets, which are close to social media in terms of variation within themselves and when compared to the standard Russian. Results are presented in tables 5 and 6.

5.4 Lemmatisation (Dialect datasets)

Dialect datasets yet again show different results. In Megra, none of the models beat the token-to-lemma baseline by the normalised Jaro-Winkler distance metric, which signals the morphological tagging issues. Incorrect morphological tag detection leads to incorrect sequence-to-sequence transformation, as the confused model applies different rules. For instance, it may predict брести 'to wander' instead of бремя 'burden'. Despite that, Taiga achieves the best score by every other metric.

In Belogornoje, the lemmatiser benefits the most from RNC-sampled-trained model tagging, with Taiga getting close. There may be different factors at play here: the Belogornoje dataset is only a thousand tokens and is closer to the standard Russian, probably, 20th-century fiction, than Megra.

Morphological tagging still seems unable to solve some critical issues. The question of how to treat compound lemmata in the dataset remains, cf. дак и 'so' that is lemmatised only as и 'and' by the model. <ë> is necessary for dialects, though, in the

Model	A	L	L(N)	D-L	D-L(N)	J-W	J-W(N)
Token-to-lemma	58.42	0.71	0.53	0.71	0.53	96.19	98.12
Bare token	53.48	0.77	0.6	0.77	0.6	88.55	97.52
RNC-sampled	86	0.25	0.25	0.25	0.25	98.49	98.49
Taiga	**86.38**	**0.24**	**0.24**	**0.24**	**0.24**	**98.59**	**98.59**
SynTagRus-downsampled	86.1	0.25	0.25	0.25	0.25	98.47	98.47

Table 4: The results of MES-Tweets lemmatisation evaluation.

Model	A	L	L(N)	D-L	D-L(N)	J-W	J-W(N)
Token-to-lemma	60.54	0.86	0.8	0.86	0.8	90.76	**96.8**
Bare token	58.65	0.9	0.84	0.9	0.83	90.3	95.67
RNC-sampled	82.67	0.37	0.37	0.37	0.37	95.65	95.65
Taiga	**83.89**	**0.35**	**0.35**	**0.35**	**0.35**	**95.66**	95.66
SynTagRus-downsampled	81.97	0.4	0.4	0.4	0.4	95.4	95.4

Table 5: The results of Megra dialect lemmatisation evaluation.

standard Russian dataset, it is normalised to <е>. Rare word changing models for verbs like помирать ‘to be dying’, the forms of which lemmatiser treats as the forms of помереть ‘to die’ under the influence of more productive models, present the problem as well.

Significant dialect features, for instance, *jakanje*, if shown in lemma, also lead to errors (cf. выдояти ‘to milk’ that is lemmatised as standard Russian выдоить). Non-standard forms, such as мни ‘I-DAT’ (cf. standard мне) confuse both the tagger and the lemmatiser, leading to incorrect tagging and subsequent assignment of the token as its lemma, instead of я. But the most significant issue is still the lemmatisation policy, the differences between understanding what should be a lemma for a token in a dataset.

6 Conclusion

The experiments prove that the silver morphological tagging allows a lemmatiser to perform much more efficiently than without any information on morphological tagging (over 40% improvement). We show that silver morphological tagging aids almost as efficiently as gold morphological tagging, lagging only by 4% for web as corpora datasets, such as GIKRYA. This is achieved with the BART-large model, fine-tuned for the standard language. Both the prediction run of BART-large and any training run of modified Stanza (Scherrer, 2021) did not take more than 4 GB GPU on RTX 3060 (mobile). Thus, even if fine-tuning large lemmatiser models themselves on personal computer hardware is still going to remain a focus of further study, morphological tagging and lemmatisation itself may be performed on the relatively small data. The lemmatiser enhanced with data provided by both the Taiga-trained and the SynTagRus-downsampled-trained taggers often performs better than the lemmatiser enhanced with data provided by the RNC-sampled-trained tagger. Even when the situation is opposite, the distance between the results rarely exceeds five per cent.

The hypotheses that we stated at the beginning of the research were the following:

H1: Morphological tagging efficiency directly influences the lemmatisation accuracy.

H2: If the model trains on the larger dataset, the morphological tagging it performs will present stable satisfactory results.

H3: For non-standard data lemmatisation, the distributional properties of the training dataset are generally more important than the sheer size.

The first hypothesis, as GIKRYA experiments show, holds only partially. SynTagRus-downsampled-trained tagger performs the best in terms of morphological tagging, but RNC-sampled-trained tagger performs the best as an aide for the lemmatiser.

The second hypothesis holds: there are no sudden falls in lemmatisation accuracy when the RNC-sampled-trained tagger provides additional data, even if the results achieved are not the best.

The third hypothesis generally holds. The less the dataset resembles the standard Russian, the more efficient becomes the enhancement with data acquired from the Taiga-trained tagger, and the less

Model	A	L	L(N)	D-L	D-L(N)	J-W	J-W(N)
Token-to-lemma	59.37	0.83	0.78	0.83	0.78	92.34	97.1
Bare token	58.05	0.85	0.8	0.85	0.79	92.22	97.16
RNC-sampled	**84.89**	**0.29**	**0.29**	**0.29**	**0.29**	**97.85**	**97.85**
Taiga	83.71	0.31	0.31	0.31	0.31	97.73	97.73
SynTagRus-downsampled	83.25	0.33	0.33	0.33	0.33	97.61	97.61

Table 6: The results of Belogornoje dialect lemmatisation evaluation.

efficient becomes the enhancement with data acquired from the SynTagRus-downsampled-trained tagger. This is because Taiga, social media texts, is much more heterogeneous than SynTagRus-downsampled. Additional morphological information from RNC-sampled-trained tagger run beats the one that Taiga provides, but only for Belogornoje. It is important to remember that parts of Taiga are included in RNC-sampled, and interaction between these parts and other parts of the RNC-sampled enabled the lemmatiser to process Belogornoje especially well. However, this case is an outlier.

The future direction of the research becomes clear: further search for a dataset that provides the best silver morphological tagging for dialect data as well as attempts at efficiently using small transformers (such as TinyBART (Shleifer and Rush, 2020)) that one can fine-tune with personal computer hardware.

References

Dan Anastasyev. 2020. Exploring pretrained models for joint morpho-syntactic parsing of Russian. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2020"*, pages 1–12. Moscow.

Vladimir Belikov, Nikolay Kopylov, Alexander Piperski, Vladimir Selegey, and Serge Sharoff. 2018. Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In *Web as Corpus Workshop (WAC-8)*, pages 24–29, Lancaster, UK. WAC-8 Organising Committee.

Aleksandrs Berdičevskis, Hanna Eckhoff, and Tatiana Gavrilova. 2016. The beginning of a beautiful friendship: rule-based and statistical analysis of Middle Russian. In *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Trudy mezhdunarodnoj konferencii «Dialog»*, pages 99–111, Moscow, Russia. RSSU.

Toms Bergmanis and Sharon Goldwater. 2018. Context sensitive neural lemmatization with Lematus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1391–1400, New Orleans, Louisiana. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Grzegorz Chrupała. 2006. Simple data-driven context-sensitive lemmatization. *Proces. del Leng. Natural*, 37:121–137.

Fred J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176.

Evelien de Graaf, Silvia Stopponi, Jasper K. Bos, Saskia Peels-Matthey, and Malvina Nissim. 2022. AGILe: The first lemmatizer for Ancient Greek inscriptions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5334–5344, Marseille, France. European Language Resources Association.

Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data conversion and consistency of monolingual corpora: Russian UD treebanks. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, pages 52–65, Oslo University, Norway. Linköping University Electronic Press.

Andrea Gesmundo and Tanja Samardžić. 2012. Lemmatisation as a tagging task. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 368–372, Jeju Island, Korea. Association for Computational Linguistics.

Frank E. Grubbs. 1969. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.

Matthew A. Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406):414–420.

Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2021. Universal Lemmatizer: A sequence-to-sequence model for lemmatizing Universal Dependencies treebanks. *Natural Language Engineering*, 27(5):545–574.

Steven Krauwer. 1998. ELSNET and ELRA: Common past, common future. *ELRA Newsletter*, 3(2).

Steven Krauwer. 2003. The Basic Language Resource Kit (BLARK) as the first milestone for the language resources roadmap. In *Proceedings of the International Workshop "Speech and Computer", SPECOM 2003*, pages 8–15, Moscow, Russia. Moscow State Linguistic University.

Olga Kryuchkova and Valentin Goldin. 2011. Corpus of Russian dialect speech: concept and parameters of evaluation. In *Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog–2011"*, pages 359–367, Moscow, Russia. RSSU.

Olga Kryuchkova and Valentin Goldin. 2015. The parameters of text processing for the Russian dialect corpus. In *Proceedings of the international conference "Corpus linguistics — 2015"*, pages 307–314, Saint Petersburg, Russia. SPbU.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Kirill Milintsevich and Kairit Sirts. 2021. Enhancing sequence-to-sequence neural lemmatization with external resources. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3112–3122, Online. Association for Computational Linguistics.

Jon Mills. 1998. Lemmatisation of the corpus of Cornish. In *Workshop on Language Resources for European Minority Languages, LREC First International Conference on Language Resources and Evaluation*, pages 1–6, Granada, Spain.

Nilo Pedrazzini and Hanne Martine Eckhoff. 2021. OldSlavNet: A scalable Early Slavic dependency parser trained on modern language data. *Software Impacts*, 8:100063.

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Springer.

Joël Plisson, Nada Lavrac, Dunja Mladenić, and Tomaž Erjavec. 2008. Ripple down rule learning for automated word lemmatisation. *AI Commun.*, 21:15–26.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal Dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Adam Radziszewski. 2013. Learning to lemmatise Polish noun phrases. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 701–709, Sofia, Bulgaria. Association for Computational Linguistics.

Yves Scherrer. 2021. Adaptation of morphosyntactic taggers. In *Similar Languages, Varieties, and Dialects: A Computational Perspective*, Studies in Natural Language Processing, page 138–166. Cambridge University Press.

Tatiana Shavrina and Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning: «Taiga» syntax tree corpus and parser. In *Proceedings of the International Conference "CORPORA 2017"*, Saint-Petersbourg, Russia.

Sam Shleifer and Alexander M. Rush. 2020. Pre-trained summarization distillation.

Alexey Sorokin, Tatiana Shavrina, Olga Lyashevskaya, Victor Bocharov, Svetlana Alexeeva, Kira Droganova, Alena Fenogenova, and Dmitry Granovsky. 2017. MorphoRuEval-2017: an evaluation track for the automatic morphological analysis methods for Russian. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017"*, pages 1–17, Moscow, Russia. RSSU.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *ArXiv*, abs/1409.3215.

Roberto Torre Alonso. 2022. Automatic lemmatization of Old English class III strong verbs (L-Y) with ALOEV3. *Journal of English Studies*, 20:237–266.

William E. Winkler. 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research*, pages 354–359.