On the Role of Morphological Information for Contextual Lemmatization

Olia Toporkov
HiTZ Center - Ixa, University of the
Basque Country UPV/EHU
olia.toporkov@ehu.eus

Rodrigo Agerri
HiTZ Center - Ixa, University of the
Basque Country UPV/EHU
rodrigo.agerri@ehu.eus

Lemmatization is a natural language processing (NLP) task that consists of producing, from a given inflected word, its canonical form or lemma. Lemmatization is one of the basic tasks that facilitate downstream NLP applications, and is of particular importance for high-inflected languages. Given that the process to obtain a lemma from an inflected word can be explained by looking at its morphosyntactic category, including fine-grained morphosyntactic information to train contextual lemmatizers has become common practice, without considering whether that is the optimum in terms of downstream performance. In order to address this issue, in this article we empirically investigate the role of morphological information to develop contextual lemmatizers in six languages within a varied spectrum of morphological complexity: Basque, Turkish, Russian, Czech, Spanish, and English. Furthermore, and unlike the vast majority of previous work, we also evaluate lemmatizers in out-of-domain settings, which constitutes, after all, their most common application use. The results of our study are rather surprising. It turns out that providing lemmatizers with fine-grained morphological features during training is not that beneficial, not even for agglutinative languages. In fact, modern contextual word representations seem to implicitly encode enough morphological information to obtain competitive contextual lemmatizers without seeing any explicit morphological signal. Moreover, our experiments suggest that the best lemmatizers out-of-domain are those using simple UPOS tags or those trained without morphology and, lastly, that current evaluation practices for lemmatization are not adequate to clearly discriminate between models.

1. Introduction

Lemmatization is one of the basic NLP tasks and consists of converting an inflected word form (e.g., *eating*, *ate*, *eaten*) into its canonical form (e.g., *eat*), usually known as the lemma. Thus, we follow the formulation of lemmatization as defined by the

Action Editor: Rico Sennrich. Submission received: 23 February 2023; revised version received: 5 July 2023; accepted for publication: 26 July 2023.

https://doi.org/10.1162/coli_a_00497

Table 1
Examples of inflected forms of the word 'cat' in Basque, English, Spanish, and Russian.

English	Spanish	Russian	Basque
cat	gato	кот	katu
cats	gata	коты	katuak
	gatos	кота	katua
	gatas	коту	katuari
		котом	katuarekin
		коте	katuek
		котов	katuekin
		котам	katuei
		котами	katuen
		котах	katurik
			katuarentzat
			katuentzat

SIGMORPHON 2019 shared task (Aiken et al. 2019). Lemmatization is commonly used when performing many NLP tasks such as information retrieval, named entity recognition, sentiment analysis, word sense disambiguation, and so forth. For example, for morphologically rich languages named entities are often inflected, which means that lemmatization is required as an additional process. Thus, lemmatization is more challenging for languages with rich inflection, as the number of variations for every different word form in such languages is very high. Table 1 illustrates this point by showing the differences in inflections of the word "cat" for four languages with different morphological structure. This language sample offers a spectrum of varied complexity, ranging from the more complex ones, Basque and Russian, to the less inflected ones, such as Spanish and English, in that order.

As we can see in Table 1, the word "cat" can vary in English by changing from singular to plural. In Spanish gender (masculine/feminine) is also marked. Things get more complicated with languages that mark case. For example, in Russian there are six cases and for Basque there are 16, some of which can be doubly inflected.

Both the context in which it occurs and the morphosyntactic form of a word play a crucial role to approach automatic lemmatization (McCarthy et al. 2019). Thus, in Figure 1 we can see a fragment of a Russian sentence in which each inflected word form has a corresponding lemma (in red). Furthermore, each inflected form has an associated number of morphosyntactic features (expressed as tags) depending on its case, number, gender, animacy, and so on. Morphological analysis is crucial for lemmatization as it explains the process required to produce the lemma from the word form, which is why it has traditionally been used as a stepping stone to design systems to perform lemmatization.

As in many other tasks in NLP, the first approaches to lemmatization were rule-based, but nowadays the best performing models address lemmatization as a supervised task in which learning in context is crucial. Regardless of the learning method used, three main trends can be observed in current contextual lemmatization: (i) those that use gold standard or learned morphological tags to generate features to learn

Figure 1
Example of a morphologically tagged and lemmatized sentence in Russian using the UniMorph annotation scheme.

lemmatization in a pipeline approach (Chrupala, Dinu, and van Genabith 2008; Yildiz and Tantuğ 2019); (ii) those that aim to jointly learn morphological tagging and lemmatization as a single task (Müller et al. 2015; Malaviya, Wu, and Cotterell 2019; Straka, Straková, and Hajic 2019); and (iii) systems that do not use any explicit morphological signal to learn to lemmatize (Chakrabarty, Pandit, and Garain 2017; Bergmanis and Goldwater 2018).

Research on contextual (mostly neural) lemmatization was greatly accelerated by the first release of the Universal Dependencies (UD) data (de Marneffe et al. 2014; Nivre et al. 2017), but especially by the contextual lemmatization shared task organized at SIGMORPHON 2019, which included UniMorph datasets for more than 50 languages (McCarthy et al. 2019). It should be noted that the best models in the task used morphological information either as features (Yildiz and Tantuğ 2019) or as part of a joint or a multitask approach (Straka, Straková, and Hajic 2019). However, the large majority of previous approaches have used all the morphological tags from UniMorph/UD assuming that fine-grained morphological information must be always beneficial for lemmatization, especially for highly inflected languages, but without analyzing whether that is the optimum in terms of downstream performance.

In order to address this issue, in this article we empirically investigate the role of morphological information to develop contextual lemmatizers in six languages within a varied spectrum of morphological complexity: Basque, Turkish, Russian, Czech, Spanish, and English. Furthermore, previous work has shown that morphological taggers substantially degrade when evaluated out-of-domain, be that any type of text different from the data used for training in terms of topic, text genre, temporality, and so forth. (Manning 2011). This point led us to research whether lemmatizers based on fine-grained morphological information will degrade more when used out-of-domain than those requiring only coarse-grained UPOS tags. We believe that this is also an important point because lemmatizers are mostly used out-of-domain, namely, to lemmatize data from a different distribution with respect to the one that was used for training.

Taking these issues into consideration, in this article we set to investigate the following research questions with respect to the actual role of morphological information to perform contextual lemmatization. First, is fine-grained morphological information really necessary, even for high-inflected languages? Second, are modern context-based word representations enough to learn competitive contextual lemmatizers without including any explicit morphological signal for training? Third, do morphologically enriched lemmatizers perform worse out-of-domain as the complexity of the morphological features increases? Fourth, what is the optimal strategy to obtain robust

contextual lemmatizers for out-of-domain settings? Finally, are current evaluation practices adequate to meaningfully evaluate and compare contextual lemmatization techniques?

The conclusions from our experimental study are the following: (i) fine-grained morphological features do not always benefit, not even for agglutinative languages; (ii) modern contextual word representations seem to implicitly encode enough morphological information to obtain state-of-the-art contextual lemmatizers without seeing any explicit morphological signal; (iii) the best lemmatizers out-of-domain are those using simple UPOS tags or those trained without explicit morphology; (iv) current evaluation practices for lemmatization are not adequate to clearly discriminate between models, and other evaluation metrics are required to better understand and manifest the shortcomings of current lemmatization techniques. The generated code and datasets are publicly available to facilitate the reproducibility of the results and further research on this topic.[1]

The rest of the article is structured as follows. The next section discusses the most relevant work related to contextual lemmatization. The systems and datasets used in our experiments are presented in Sections 4 and 3, respectively. Section 5 presents the experimental setup applied to obtain the results, which are reported in Section 6. Section 7 provides a discussion and error analysis of the results. We finish with some concluding remarks in Section 8.

2. Background

First approaches to lemmatization consisted of systems based on dictionary lookup and/or rule-based finite state machines (Karttunen, Kaplan, and Zaenen 1992; Oflazer 1993; Alegria et al. 1996; van den Bosch and Daelemans 1999; Dhonnchadha 2002; Segalovich 2003; Carreras et al. 2004; Stroppa and Yvon 2005; Jongejan and Dalianis 2009). Grammatical rules in such systems, either hand-crafted or learned automatically by using machine learning, were leveraged to perform lemmatization together with the use of lexicons or morphological analyzers that returned the correct lemma. The problem of unseen and rare words was solved by generating a set of exceptions added to the general set of rules (Karttunen, Kaplan, and Zaenen 1992; Oflazer 1993) or by using a probabilistic approach (Segalovich 2003). Such systems resulted in very language-dependent approaches, and in most of the cases they required huge linguistic knowledge and effort, especially in the case of those languages with more complex, high-inflected morphology.

The appearance of large annotated corpora with morphological information and lemmas facilitated the development of machine learning methods for lemmatization in multiple languages. One of the core projects that gathered annotated corpora for more than 90 languages is the Universal Dependencies (UD) initiative (Nivre et al. 2017). This project offers a unified morphosyntactic annotation across languages with language-specific extensions when necessary. Based on the UD data, the Universal Morphology (UniMorph) project (McCarthy et al. 2020) converted the UD annotations into UniMorph, a universal tagset for morphological annotation (based on Sylak-Glassman [2016]), where each inflected word form is associated with a lemma and a set of morphological features. The current UniMorph dataset includes 118 languages, including extremely low-resourced languages such as Quechua, Navajo, and Haida.

1 https://github.com/oltoporkov/morphological-information-datasets.

The assumption that context could help with unseen and ambiguous words led to the creation of supervised contextual lemmatizers. The pioneer work on this topic is perhaps the statistical contextual lemmatization model provided by Morfette (Chrupala, Dinu, and van Genabith 2008). Morfette uses a Maximum Entropy classifier to predict morphological tags and lemmas in a pipeline approach. Interestingly, instead of learning the lemmas themselves, Chrupala, Dinu, and van Genabith (2008) propose to learn automatically induced lemma classes based on the shortest edit script (SES), which consists of the number of edits necessary to convert the inflected word form into its lemma. Morfette has influenced many other works on contextual lemmatization, such as the system of Gesmundo and Samardžić (2012), IXA pipes (Agerri, Bermudez, and Rigau 2014; Agerri and Rigau 2016), Lemming (Müller et al. 2015), and the system of Malaviya, Wu, and Cotterell (2019). The importance of using context to learn lemmatization is investigated in the work of Bergmanis and Goldwater (2018). They compare context-free and context-sensitive versions of their neural lemmatizer Lematus and evaluate them across 20 languages. Results show that including context substantially improves lemmatization accuracy and it helps to better deal with the out-of-vocabulary problem.

The next step in the development of contextual lemmatization systems came with the supervised approaches based on deep learning algorithms and vector-based word representations (Chakrabarty, Pandit, and Garain 2017; Dayanik, Akyürek, and Yuret 2018; Bergmanis and Goldwater 2018; Malaviya, Wu, and Cotterell 2019). The parallel development of the transformer architecture (Vaswani et al. 2017) and the appearance of BERT (Devlin et al. 2019) and other transformer-based masked language models (MLMs) offered the possibility to significantly improve lemmatization results. Thus, most of the participating systems in the SIGMORPHON 2019 shared task on contextual lemmatization for 66 languages were based on MLMs (McCarthy et al. 2019). The baseline provided by the task was based on the work of Malaviya, Wu, and Cotterell (2019), a system that performs joint morphological tagging and lemmatization.

To the best of our knowledge, current state-of-the-art results in contextual lemmatization are provided by those models that achieved best results in the SIGMORPHON 2019 shared task. The highest overall accuracy was achieved by UDPipe (Straka, Straková, and Hajic 2019). Using UDPipe 2.0 (Straka 2018) as a baseline, they added pre-trained contextualized BERT and Flair embeddings as an additional input to the network. The overall accuracy (average across all languages) was 95.78, the best among all the participants.

The second-best result (95 overall word accuracy) in the task was obtained by the CHARLES-SAARLAND system (Kondratyuk 2019). This system consists of a combination of a shared BERT encoder and joint lemma and morphology tag decoder. The model uses a two-stage training process, in which it first performs a multilingual training over all treebanks, and then executes the same process monolingually, maintaining the previously learned multilingual weights. Morphological tags in this case are calculated jointly and lemmas are also represented as SES. The experiments are performed using multilingual BERT in combination with the methods introduced by UDify (Kondratyuk and Straka 2019) for BERT fine-tuning and regularization.

The third best result (94.76) was reported by Morpheus (Yildiz and Tantuğ 2019). Morpheus uses a two-level LSTM network which gets as input the vector-based representations of words, morphological tags, and SES. Morpheus then aims to jointly output, for a given sequence, their corresponding morphological labels and the SES representing the lemma class, which is later decoded into its lemma form.

Thus, it can be seen that a common trend in current contextual lemmatization is to use the morphological information provided by the full UniMorph labels without

taking into consideration whether this is the optimal setting. Furthermore, lemmatization techniques are only evaluated in-domain, resulting in extremely, and perhaps deceptive, high results for the large majority of the 66 languages included in the SIGMORPHON 2019 data.

3. Languages and Datasets

In order to address the research questions formulated in the Introduction, we selected the following six languages: Basque, Turkish, Russian, Czech, Spanish, and English. Such a choice will allow us to compare the role of fine-grained morphological information to learn contextual lemmatization within a range of languages of varied morphological complexity. In this section we briefly describe general morphological characteristics of each language as well as the specific datasets used.

3.1 Languages

Basque and Turkish are agglutinative languages with morphology mostly of the suffixing type. Basque is a language isolate and does not belong to any language group, and Turkish is a member of the Oghuz group of the Turkic family. These two languages have no grammatical gender, with some particular exceptions for domestic animals, people, and foreign words (Turkish) or in some colloquial forms when the gender of the addressee is expressed for the second person singular pronoun (Basque). Turkish and Basque have two number types (singular and plural), and in Basque there is also the unmarked number (undefined or *mugagabea*). In both Turkish and Basque the cases are expressed by suffixation.

Basque is an ergative-absolutive language containing 16 cases, meaning that the grammatical case marks both the subject of an intransitive verb and the object of a transitive verb. The verb conjugation is also specific for this language: The majority of the verbs are formed by a combination of a gerund form and a conjugated auxiliary verb.

Turkish has six general cases; nouns and adjectives are not distinguished morphologically and adjectives can also be used as adverbs without modifications or by doubling of the word. For verbs there are 9 simple and 20 compound tenses. There is a relatively small set of core vocabulary and the majority of Turkish words originate from applying derivative suffixes to nouns and verbal stems.

The two Slavic languages, Russian and Czech, which have a fusional morphological system, exhibit a highly inflectional morphology and a wide number of morphological features. Russian belongs to the East Slavic language group, while Czech is a West Slavic language. These two languages have nominal declension that involves 6 main grammatical cases for Russian and 7 for Czech. Both languages distinguish between two number (singular and plural) and three gender types (masculine, feminine, and neuter). Furthermore, the masculine gender is subdivided into animate and inanimate. Verbs are conjugated for tense (past, present, or future) and mood.

Spanish is a Romance language that belongs to the Indo-European language family. It is a fusional language, which has a tendency to use a single inflectional morpheme to denote multiple grammatical, syntactic, or semantic features. Nouns and adjectives in Spanish have two gender (male, female) and two number types (singular and plural). Additionally, some articles, pronouns, and determiners also possess a neuter gender. There are 3 main verb tenses (past, present, and future) and each verb has around fifty

conjugated forms. Apart from that, Spanish has 3 verboid forms (infinitive, gerund, past participle), perfective and imperfective aspects for past, 4 moods, and 3 persons.

Finally, English is a Germanic language, also part of the Indo-European language family. It has lower inflection in comparison to previously mentioned languages. Only nouns, pronouns, and verbs are inflected, while the rest of the parts of speech are invariable. In English, animate nouns have two genders (masculine or feminine) and the third-person singular pronouns distinguish three gender types: masculine, feminine, and neuter, while for most of the nouns there is no grammatical gender. Nouns have only a genitive case and personal pronouns are mostly declined in subjective and objective cases. English has a variety of auxiliary verbs that help to express the categories of mood and aspect and participate in the formation of verb tenses.

3.2 Datasets

The datasets we used are distributed as part of the data used for the SIGMORPHON 2019 shared task (McCarthy et al. 2019). The source of the original datasets comes from the UD project (de Marneffe et al. 2014), but the morphological annotations are converted from UD annotations to the UniMorph schema (Kirov et al. 2018) with the aim of increasing agreement across languages. As our experiments will include both in-domain and out-of-domain evaluations, we selected some datasets for each of the settings.

With respect to in-domain, we chose one corpus per language using the standard train and development partitions. For Basque we used the Basque Dependency Treebank (BDT) (Aldezabal et al. 2008), which contains mainly literary and journalistic texts. The corpus was manually annotated and then automatically converted to UD format. For Czech we used the CAC treebank (Hladká et al. 2008) based on the Czech Academic Corpus 2.0. This corpus includes mostly unabridged articles from a wide range of media such as newspapers, magazines, and transcripts of spoken language from radio and TV programs. The corpus was annotated manually and then converted to UD format. With respect to English we chose English Web Treebank (EWT) (Silveira et al. 2014). This corpus includes different Web sources: blogs, various media, e-mails, reviews, and Yahoo! answers. In the EWT corpus the lemmas were assigned by UD-converter and manually corrected. UPOS tags were also converted to UD format from manual annotations. For Russian we used the GSD corpus, extracted from Wikipedia and manually annotated by native speakers. In the case of Spanish we selected the GSD corpus as well, consisting of texts from blogs, reviews, news, and Wikipedia. Finally, for Turkish we used ITU-METU-Sabanci Treebank (IMST) (Sulubacak et al. 2016). It consists of well-edited sentences from a wide range of domains, manually annotated and automatically converted to UD format.

For the out-of-domain evaluation setting we picked the test sets of other datasets included in UniMorph, different from the ones selected for in-domain experimentation. In the case of Basque, only one corpus was available in the UD project, so we used the Armiarma corpus, which consists of literary critics semi-automatically annotated using Eustagger (Alegria et al. 1996). For Czech and Turkish we used the PUD data—part of the Parallel Universal Dependencies treebanks created for the CoNLL 2017 shared task (Zeman et al. 2017). The corpora consist of 1,000 sentences from the news domain and Wikipedia annotated for 18 languages. The Czech language PUD data was manually annotated and then automatically converted to UD format. For Turkish the original data was automatically converted to UD format, but later manually reannotated (Türk et al. 2019). In the case of English we used the Georgetown University Multilayer

(GUM) corpus (Zeldes 2017). This corpus presents a collection of annotated Web texts from interviews, news, travel guides, academic writing, biographies, and fiction from such sources as Wikipedia, Wikinet, and Reddit. Its lemmas were manually annotated, while UPOS tags were converted to UD format from manual annotations. In the case of Russian we used SynTagRus (Lyashevkaya et al. 2016), which consists of texts from a variety of genres, such as contemporary fiction, popular science, as well as news and journal articles from the 1960–2016 period. Its lemmas, UPOS tags, and morphological features were manually annotated in non-UD style and then automatically converted to UD format. For Spanish we chose the AnCora corpus (Taulé, Martí, and Recasens 2008), which contains mainly texts from news. All the elements of this corpus were converted to UD format from manual annotations.

4. Systems

In this section we present the systems that we will be applying in our investigation. First, research on the role of fine-grained morphological information for contextual lemmatization will be performed in-domain using the statistical lemmatizer from the IXA pipes toolkit (Agerri and Rigau 2016) and Morpheus, the third best system in the SIGMORPHON 2019 shared task. These two systems were chosen for several reasons: (i) both use morphological information as features to learn lemmatization; (ii) both systems use SES to represent automatically induced lemma classes; and (iii) they both address contextual lemmatization as sequence tagging.

In order to investigate whether modern contextual word representations are enough to learn competitive lemmatizers both in- and out-of-domain, we train baseline models using Flair (Akbik, Blythe, and Vollgraf 2018), multilingual MLMs mBERT and XLM-RoBERTa (Devlin et al. 2019; Conneau et al. 2020), as well as language-specific MLMs for each of the languages: BERTeus for Basque (Agerri et al. 2020), slavicBERT for Czech (Arkhipov et al. 2019), RoBERTa for English (Liu et al. 2019), Russian ruBERT (Kuratov and Arkhipov 2019), Spanish BETO (Cañete et al. 2020), and BERTurk for Turkish.[2] As with Morpheus and IXA pipes, we treat contextual lemmatization as a sequence tagging task and fine-tune the language models by adding a single linear layer to the top of the model. The experiments were implemented using the HuggingFace Transformers API (Wolf et al. 2020).

4.1 Systems Using Morphology

IXA pipes is a set of multilingual tools which is based on a pipeline approach (Agerri, Bermudez, and Rigau 2014; Agerri and Rigau 2016). IXA pipes learn perceptron (Collins 2002) models based on shallow local features combined with pre-trained clustering features induced over large unannotated corpora. The lemmatizer implemented in IXA pipes is inspired by the work of Chrupala, Dinu, and van Genabith (2008), where the model learns the SES between the word form and its lemma. IXA pipes are able to learn lemmatization using gold-standard or learned morphological tags.

Morpheus is a neural contextual lemmatizer and morphological tagger that consists of two separate sequential decoders for generating morphological tags and lemmas. The input words and morphological features are encoded in context-aware vector representations using a two-level LSTM network and the decoders predict both the

2 https://github.com/stefan-it/turkish-bert.

morphological tags and the SES, which are later decoded into its lemma (Yildiz and Tantuğ 2019). Morpheus obtained the third best overall result in the SIGMORPHON 2019 shared task (McCarthy et al. 2019).

4.2 Systems Without Explicit Morphological Information

We train a number of models that use modern contextual word representations by addressing lemmatization as a sequence tagging task. Thus, the input consists of words encoded as contextual vector representations and the task is to assign the best sequence of SES to a given input sequence.

Flair is a NLP framework based on a BiLSTM-CRF architecture (Huang, Xu, and Yu 2015; Ma and Hovy 2016) and pre-trained language models that leverage character-based word representations that, according to the authors, capture implicit information about natural language syntax and semantics. Flair has obtained excellent results in sequence labeling tasks such as named entity recognition, POS tagging, and chunking (Akbik, Blythe, and Vollgraf 2018). The library includes pre-trained Flair language models for every language except Turkish.

With respect to the MLMs, we use two multilingual models and 6 language models trained specifically for each of the languages included in our study. Multilingual BERT (Devlin et al. 2019) is a transformer-based masked language model, pre-trained on the Wikipedias of 104 languages with both the masking and next sentence prediction objectives. Furthermore, we also use XLM-RoBERTa (Conneau et al. 2020), trained on 2.5TB (295K millions of tokens) of filtered CommonCrawl data for 100 languages. XLM-RoBERTa is based on the BERT architecture but (i) trained only on the MLM task, (ii) on larger batches, (iii) on longer sequences, and (iv) with dynamic mask generation. Thus, multilingual BERT was trained with a batch size of 256 and 512 sequence length for 1M steps, using both the MLM and NSP tasks. Regarding XLM-RoBERTa, both versions (base and large) were trained over 1.5M steps with batch 8,192 and sequences of 512 length.

Details about the six language-specific MLMs used are provided in Table 2. BERTeus (Agerri et al. 2020) is a BERT-base model trained on the BMC Basque corpus, which includes the Basque Wikipedia and news articles from online newspapers. Apart from the training data, the other difference from original BERT is the subword tokenization, which is closer to linguistically interpretable strings in Basque. BERTeus significantly outperforms multilingual BERT and XLM-RoBERTa in tasks such as POS tagging, named entity recognition, topic modeling, and sentiment analysis.

BERTurk[3] is a cased BERT-base model for Turkish. This model was trained on a filtered and sentence segmented version of the Turkish OSCAR corpus (Ortiz Suárez, Sagot, and Romary 2019), together with Wikipedia, various OPUS corpora (Tiedemann 2016), and data provided by Kemal Oflazer, which resulted in a total size of 35GB (4,404M tokens total).

For Czech we used slavicBERT (Arkhipov et al. 2019), developed by taking multilingual BERT as a basis and further pre-trained using Russian news and the Wikipedias of four Slavic languages: Russian, Bulgarian, Czech, and Polish. The authors also rebuilt the vocabulary of subword tokens, using the subword-nmt repository.[4]

3 https://github.com/stefan-it/turkish-bert.
4 https://github.com/rsennrich/subword-nmt/.

Table 2
List of language-specific models used in the experiments for each of the target languages.

Language	Model	Architecture	Training corpus and number of tokens
Basque	BERTeus	BERT	35M tokens (Wikipedia) + 191M tokens (online)
Czech	slavicBERT	BERT	Russian news and Wikipedia in Russian, Bulgarian, Czech, and Polish
English	RoBERTa	BERT	BookCorpus (800M tokens), CC-News (16,000M tokens), OpenWebText (8,706M tokens), CC-Stories (5,300M tokens)
Russian	ruBERT	BERT	Dataset for original BERT (BookCorpus(800M tokens)), English Wikipedia (2,500M tokens), Russian news and Wikipedia for subword vocabulary
Spanish	BETO	BERT	Wikipedia and OPUS project in Spanish (3,000M tokens)
Turkish	BERTurk	BERT	OSCAR corpus, Wikipedia, OPUS corpora, corpus of Kemal Oflaizer (4,404M tokens total)

RuBERT was developed in a similar fashion as slavicBERT but only with Russian as the target language using the Russian Wikipedia and news corpora (Kuratov and Arkhipov 2019). They generated a new subword vocabulary obtained from subword-nmt which contains longer Russian words and subwords.

For Spanish we used BETO (Cañete et al. 2020), a BERT-base language model, trained on a large Spanish corpus. The authors of this model upgraded the initial BERT model by using the Dynamic Masking technique, introduced in RoBERTa. BETO performed 2M steps in two different stages: 900K steps with a batch size of 2,048 and maximum sequence length of 128, and the rest of the training with a batch size of 256 and maximum sequence length of 512. We use the version trained with cased data, which included the Spanish Wikipedia and various sources from the OPUS project (Tiedemann 2012) in a final corpus size of around 3 billion words.

RoBERTa-base is the model chosen for English. RoBERTa (Liu et al. 2019) is an optimized version of BERT, as commented above. To train this model the authors, apart from the standard datasets used to train the BERT model, also used the CC-news dataset, including English news articles from all over the world published between January 2017 and December 2019. The total size of the training data exceeds 160GB of uncompressed text (more than 30 billion tokens).

4.3 Baselines

We use two models as baselines. First, the system used as a baseline for the SIGMORPHON 2019 shared task (McCarthy et al. 2019), a joint neural model for morphological tagging and lemmatization presented by Malaviya, Wu, and Cotterell (2019). This system performs morphological tagging by using a LSTM tagger described in Heigold, Neumann, and van Genabith (2017) and Cotterell and Heigold (2017). The lemmatizer is a neural sequence-to-sequence model (Wu and Cotterell 2019) which includes a hard

attention mechanism with a training scheme based on dynamic programming. The tagger and lemmatizer are connected together by jackknifing (Agić and Schluter 2017), which allows us to avoid exposure bias and improve lemmatization results.

The second baseline is the winner of the SIGMORPHON'19 shared task (Straka, Straková, and Hajic 2019). UDPipe is a multitask model which jointly learns morphological tagging and lemmatization. The system architecture consists of three bidirectional LSTMs that process the input and softmax classifiers that generate lemmas and morphosyntactic features. Lemmatization is performed as a multiclass classification task, where the system predicts the correct lemma rule or SES.

5. Experimental Setup

The systems described above were trained on the datasets listed in Section 3.2 using the following methodology. For the two IXA pipes models (using gold-standard and learned morphology) we used the default feature set, with and without clustering features, specified in Agerri and Rigau (2016). The default hyperparameters were also applied to train Morpheus (Yildiz and Tantuğ 2019). The input character embedding length d_a is set to 128, the length of the word vectors d_e to 1,024, and the length of the context-aware word vectors d_c to 2,048. Moreover, the length of character vectors in the minimum edit prediction component d_u and the length of the morphological tag vectors d_v are set to 256. The hidden unit sizes in the decoder LSTMs d_g and d_q are set to 1,024. The Adam optimization algorithm is used with learning rate 3e-4 to minimize loss (Kingma and Ba 2015).

Flair is used off-the-shelf with FastText CommonCrawl word embeddings (Grave et al. 2018) combined with Flair contextual embeddings for each of the languages. The hidden size of the LSTM is set to 256 with a batch of 16.

The MLMs were fine-tuned for lemmatization as a sequence tagging task by adding a single linear layer on top of the model being fine-tuned. A grid search of hyperparameters was performed to pick the best batch size (16, 32), epochs (5, 10, 15, 20, 25), and learning rate (1e-0, 2e-5, 3e-5, 5e-5). We pick the best model on the development set in terms of word accuracy and loss. A fixed seed is used to ensure reproducibility of the results.

For multilingual BERT we used a maximum sequence length of 128, batch size 32, and 5e-5 as learning rate while for XLM-RoBERTa we used the same configuration but with a batch of 16. For Russian we perform grid search on two language-specific models, namely, ruBERT and slavicBERT. RuBERT obtained the best results with a maximum sequence length of 128, batch size 16, and a 5e-5 value for learning rate over 15 epochs. For the rest of the models the best configuration was that of XLM-RoBERTa over 5 epochs for BETO and RoBERTa-base, 10 epochs for BERTeus, 15 epochs for BERTurk, and 20 epochs with slavicBERT for Czech.

6. Experimental Results

In this section we present the experiments to empirically address the following research questions with respect to the actual role of morphological information to perform contextual lemmatization, namely: (i) Is fine-grained morphological information really necessary, even for agglutinative languages? (ii) Are modern context-based word representations enough to learn competitive contextual lemmatizers without including any explicit morphological signal during training? (iii) Do morphologically enriched lemmatizers perform worse out-of-domain as the complexity of the morphological

features increases? (iv) What is the optimal strategy to obtain robust contextual lemmatizers for out-of-domain settings? and (v) Are current evaluation practices adequate to meaningfully evaluate and compare contextual lemmatization techniques?

Unlike the vast majority of previous work on contextual lemmatization, which has been mostly evaluated *in-domain* (McCarthy et al. 2019), we also report results in *out-of-domain* settings. It should be noted that by *out-of-domain* we mean to evaluate the model on a different data distribution from the data used for training (Manning 2011).

First, Section 6.1 studies the in-domain performance of contextual lemmatizers depending on the type of morphological features used to inform the models during training. The objective is two-fold: to determine whether complex (or any at all) morphological information is required to obtain competitive lemmatizers and to establish whether modern contextual word representations and MLMs allow us to perform lemmatization without any morphological information.

Second, in the *out-of-domain* evaluation presented in Section 6.2 we analyze the performance of morphologically informed lemmatizers. Furthermore, comparing them with contextual lemmatizers developed without an explicit morphological signal would allow us to obtain a full picture as to what is the best strategy for out-of-domain settings (the most common application scenario).

6.1 In-domain Evaluations

For the first experiment we train the two variants of the IXA pipes statistical system, ixa-pipe-gs and ixa-pipe-mm (Agerri and Rigau 2016), and one neural lemmatizer, Morpheus (Yildiz and Tantuğ 2019). As explained in Section 4, all three require explicit morphological information and they all apply the SES to automatically induced lemma classes from the training data.

Furthermore, we combined the UniMorph morphological tags to generate labels of different complexity. Thus, taking UPOS tags as a basis we obtain 5 different morphological tags, as shown in Table 3. The first 4 are combinations of UPOS, case, gender, and number. The last label includes UPOS and every feature present for a given word in UniMorph in the following order: {UPOS+Case+Gender+Number+Rest-of-the-features}. For some word types, such as prepositions or infinitives, UniMorph only includes the UPOS tag. In order to illustrate this, Table 4 provides an example originally in Russian including the information required to train contextual lemmatizers, namely, the word, some morphological tag, and the lemma.

Putting it all together, Table 5 characterizes the final datasets used for in- and out-of-domain evaluation. The number of tokens, unique labels per category, and unique SES

Table 3
List of UniMorph morphological tags used.

Morphological label
UPOS
UPOS+Case+Gender
UPOS+Case+Number
UPOS+Case+Gender+Number
UPOS+AllFeaturesOrdered

Table 4
An example of the data used to train contextual lemmatizers with morphological information.

Word form	Morphological label {UPOS+Case+Gender}	Lemma
Проект[*Project*]	NNOMMASC	проект[*project*]
сильно[*a lot*]	ADV	сильно[*a lot*]
отличался[*differed*]	VMASC	отличаться[*to differ*]
от[*from*]	ADP	от[*from*]
предыдущих[*previous*]	ADJGEN	предыдущий[*previous*]
подлодок[*submarines*]	NGENFEM	подлодка[*submarine*]
.	-	.

Table 5
Language complexity reflected in the number of labels according to the augmentation of morphological features, number of lemma classes and corpus tokens.

language	corpus	number of tokens	upos	upos+case +gender +number	upos +allfeat. ord.	upos +allfeat. not.ord.	SES (lemma) class
Basque	train (BDT)	97,336	15	205	1,143	1,683	1,306
	dev (BDT)	12,206	14	148	556	787	432
	test (BDT)	11,901	14	153	545	773	428
	test (Armiarma)	299,206	–	–	–	–	1,495
Czech	train (CAC)	395,043	16	332	1,266	1,784	946
	dev (CAC)	50,087	16	298	876	1,129	536
	test (CAC)	49,253	15	284	827	1,036	556
	test (PUD)	1,930	14	175	288	292	151
Russian	train (GSD)	79,989	14	241	851	1,384	553
	dev (GSD)	9,526	14	191	435	673	235
	test (GSD)	9,874	14	203	455	713	258
	test (SynTagRus)	109,855	15	247	757	1,243	896
Spanish	train (GSD)	345,545	25	116	287	510	310
	dev (GSD)	42,545	23	100	208	342	200
	test (GSD)	43,497	23	103	222	387	200
	test (AnCora)	54,449	15	75	178	309	298
English	train (EWT)	204,857	16	43	94	173	233
	dev (EWT)	24,470	16	41	88	160	120
	test (EWT)	25,527	16	41	85	156	115
	test (GUM)	8,189	17	42	72	124	80
Turkish	train (IMST)	46,417	15	124	1,541	1,897	211
	dev (IMST)	5,708	15	95	605	748	106
	test (IMST)	5,734	16	100	589	725	104
	test (PUD)	1,795	15	66	217	220	59

(calculated using the UDPipe method) illustrate the varied complexity of the languages involved.[5] Thus, those languages with more complex morphology have a higher number of unique labels that include additional morphological features. The same pattern can be seen in the amount of lemma classes (SES), significantly larger for the languages with more complex morphology. In the case of Turkish the low number of lemmas could be explained by the fact that most Turkish words are formed by applying derivative suffixes to nouns and verbal stems. Moreover, the core vocabulary in this particular corpus is rather small. Finally, we decided to order the subtags comprising the full UniMorph labels as the number of unique labels decreased significantly.

Table 6 reports the in-domain results of training the three systems for the 6 languages with the 5 different types of morphological labels. First, the results show that the neural lemmatizer Morpheus outperforms the statistical lemmatizers for every language except English. In fact, for languages with more complex morphology, such as Basque and Turkish, the differences are larger. Second, if we look at the impact of including fine-grained morphological features, it can be seen that no single morphological tag performs best across systems and languages. Thus, while adding case, number, and/or gender seems to be slightly beneficial, differences in performance are substantial when training the statistical lemmatizer using gold-standard morphological labels (ixa-pipe-gs) and especially for languages with more complex morphology (Basque, Russian, Turkish). Third, the results clearly show that adding every available morphological feature is not beneficial *per se*. Fourth, the statistical lemmatizer trained with learned morphological tags (ixa-pipe-mm) performs significantly worse in every case except for English and Spanish. Finally, adding a special label "no-tag" with no morphological information shows that performance decreases significantly for every system and language.

Summarizing, *in-domain* performance for high-inflected languages improves when some fine-grained morphological attributes (case and number or gender) are used to train the statistical lemmatizers. However, for English and Spanish using UPOS seems to be enough. Thus, in the case of neural lemmatization with Morpheus (the best of the models using morphological information), we can see that no substantial gains are obtained by adding fine-grained morphological features to UPOS tags, not even for agglutinative languages such as Basque or Turkish.

This point is reinforced by the results of computing the McNemar test of statistical significance to establish whether the differences in the results obtained by Morpheus (the best among the models trained with morphology) informed only with UPOS labels or with the best morphological label (as by Table 6 above) are statistically significant or not (null hypothesis). The result of the test showed that for every language the differences were not significant ($\alpha = .05$, with 0.936 p-value for Basque, 0.837 for Czech, 0.511 for Russian, and 0.942 for Spanish).

Taking this into consideration, the next natural step is to consider whether it is possible to learn good contextual lemmatizers without providing any explicit morphological signal during training. Previous work on probing contextual word representations

5 Even though it is not required for out-of-domain evaluation, the UniMorph information is not available for the Basque Armiarma corpus because it is not part of the UniMorph project.

Table 6
In-domain lemmatization results on the development sets for systems that use morphology to train contextual lemmatizers. ixa-mm: IXA pipes with learned morphological tags; ixa-gs: IXA pipes with gold standard morphology; morph = Morpheus; UCG: UPOS+Case+Gender, UCN: UPOS+Case+Number, UCGN:UPOS+Case+Gender+Number: UALLo: UPOSAllFeaturesOrdered.

	English					
	no-tag	UPOS	UCG	UCN	UCGN	UAllo
ixa-mm	–	98.97	98.97	99.03	98.97	98.86
ixa-gs	96.98	99.51	99.49	99.58	99.59	99.65
morph	97.60	98.20	98.12	98.13	98.19	98.14
	Spanish					
ixa-mm	–	98.75	98.74	98.71	98.78	98.74
ixa-gs	98.36	98.82	98.78	98.82	98.80	98.88
morph	98.17	98.09	98.93	98.96	98.92	98.91
	Russian					
ixa-mm	–	94.85	95.37	95.69	95.50	95.53
ixa-gs	91.85	95.05	96.95	96.45	96.99	97.04
morph	96.50	96.92	96.91	97.10	97.18	97.24
	Basque					
ixa-mm	–	93.19	93.22	93.14	93.30	93.49
ixa-gs	91.68	93.50	94.33	94.58	94.58	96.50
morph	95.48	96.30	96.43	96.54	96.37	96.42
	Czech					
ixa-mm	–	97.76	97.17	97.29	97.10	97.10
ixa-gs	95.64	97.68	98.10	97.93	98.09	98.20
morph	98.37	98.78	98.84	98.83	98.82	98.80
	Turkish					
ixa-mm	–	84.83	84.51	85.06	85.06	83.95
ixa-gs	85.97	88.81	88.89	89.14	89.14	90.52
morph	96.04	96.41	96.53	95.95	96.27	96.50

and transformer-based MLMs suggests that such models implicitly encode information about part-of-speech and morphological features (Manning et al. 2020; Akbik, Blythe, and Vollgraf 2018; Conneau et al. 2018; Belinkov et al. 2017). Following this, for this experiment we fine-tune various well-known multilingual and monolingual language models (detailed in Section 4) by using only the word forms and the automatically induced SES as implemented by UDPipe (Straka, Straková, and Hajic 2019).

Figure 2 reports the results. From left-to-right, the first three bars correspond to the best statistical and Morpheus models using explicit morphological information as previously reported in Table 6. The next four list the results from Flair, mBERT, XLM-RoBERTa-base, and a language-specific monolingual model (none of these four use any explicit morphological signal) whereas *base* (dark purple) refers to the system of Malaviya, Wu, and Cotterell (2019), used as a baseline for the SIGMORPHON 2019 shared task (McCarthy et al. 2019). For state-of-the-art comparison, the last column on

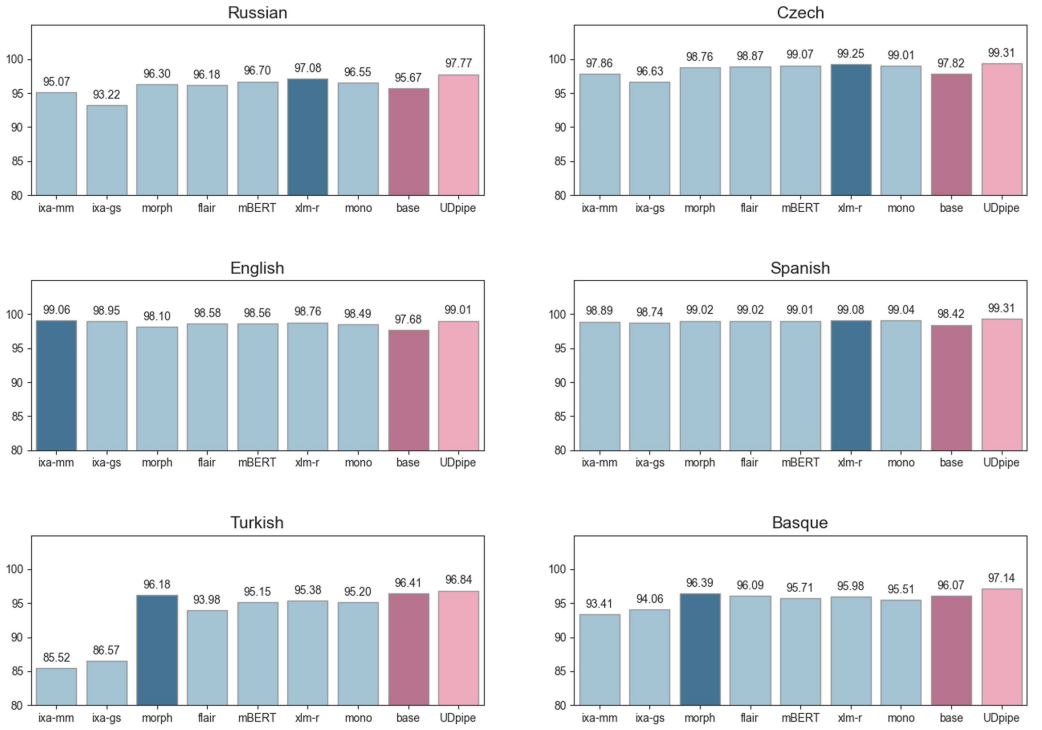

Figure 2
Overall *in-domain* lemmatization results on the test data for models trained with and without explicit morphological features; monolingual transformers: Russian - ruBERT, Czech - slavicBERT, Basque - BERTeus, Turkish - BERTurk, English - RoBERTa, Spanish - BETO.

the right provides the results from UDPipe (Straka, Straková, and Hajic 2019) (light purple color). Finally, the dark blue bars represent the best result for each language without considering either the baseline system or UDPipe.

The first noticeable trend is that every model beats the baseline except the IXA pipes-based statistical lemmatizers, which perform over the baseline and comparatively to the other models for English and Spanish only, the languages with the less complex morphology.

The second and, perhaps, most important fact is that the four models (Flair, mBERT, XLM-RoBERTa, and mono) that do not use any morphological signal for training obtain a remarkable performance across languages, XLM-RoBERTa-base being the best overall, even better than language-specific monolingual models. In fact, XLM-RoBERTa-base outperforms Morpheus for 4 out of the 6 languages, a neural model which was the third best system in the SIGMORPHON 2019 benchmark and which uses all the morphological information available in the UniMorph data. The McNemar test of significance shows that the differences in results obtained by Morpheus and XLM-RoBERTa are statistically significant ($\alpha = .05$) for Russian, Spanish, and English (in XLM-RoBERTa's favor), and for Basque and Turkish (Morpheus over XLM-RoBERTa).

An additional observation is that our XLM-RoBERTa-base lemmatization models perform competitively with respect to UDPipe, which obtains the best results for 5 out of the 6 languages included in our study. UDPipe's strong performance is somewhat

expected as it was the overall winner of the SIGMORPHON 2019 lemmatization task. It should be noted that UDPipe is a rather complex system consisting of a multitask model to predict POS tags, lemmas, and dependencies by applying three shared bidirectional LSTM layers which take as input a variety of word and character embeddings, the final model being an ensemble of 9 possible embedding combinations. However, the results obtained by the language models we trained without any explicit morphological signal, such as XLM-RoBERTa-base, are based on a simple baseline setting, where the transformer models are fine-tuned using the automatically induced SES as the target labels in a token classification task. These results seem to confirm that, as it was the case for POS tagging and other tasks (Manning et al. 2020), contextual word representations implicitly encode morphological information which made them perform strongly for lemmatization.

However, we can see that for agglutinative languages such as Basque and Turkish, the neural models using explicit morphological features (Morpheus, Malaviya et al. 2019, and UDPipe) still outperform those without it (although for Basque the differences are much smaller). Still, the overall results show that, apart from Basque and Turkish, differences between XLM-RoBERTa and the best model for each language are rather minimal. This demonstrates that it is possible to generate competitive contextual lemmatization without any explicit morphological information using a very simple technique, although a more sophisticated approach or larger language model may be required to be competitive with the state-of-the-art currently represented by UDPipe.

6.2 Out-of-domain Evaluation

Although lemmatizers are mostly used out-of-domain, the large majority of the experimental results published so far do not take this issue into account when evaluating approaches to contextual lemmatization. In this section we empirically investigate the out-of-domain performance of the lemmatizers from the previous section to establish whether: (i) using fine-grained morphological information causes cascading errors in the lemmatization performance; (ii) the lack of morphological information helps to obtain more robust lemmatizers across domains.

For a better comparison, Table 7 presents both the in-domain results presented in the previous section together with their corresponding out-of-domain performance on the datasets presented in Section 3.

Table 7 allows us to see the general trend in performance across domains and with respect to the type of morphological information used. First, and, as expected, out-of-domain performance is substantially worse for every evaluation setting and particularly significant for highly inflected languages. Second, in terms of the type of morphological label, there are no clear differences between the models using just UPOS tags or those using more fine-grained information, the exception being Russian and Turkish with the *ixa-pipe-mm* system, for which the highest result with {UPOS+Case+Number} is around 1 point in word accuracy better than UPOS. Furthermore, there is not a common type of morphological information that works best across languages. Third, while the statistical lemmatizers are competitive for Spanish and English, they are clearly inferior for Basque and Turkish. Finally, when looking at the results in terms of the models using gold-standard morphological annotations (*ixa-pipe-gs* and Morpheus), it is interesting that they degrade less out-of-domain than the model using learned morphological tags for most of the cases except for Russian. Summarizing, we can conclude that adding fine-grained morphological information to UPOS does not in general result in better out-of-domain performance.

Table 7
In-domain and out-of-domain test results for systems trained with explicit morphological information: ixa-mm: IXA pipes with learned morphological tags; ixa-gs: IXA pipes with gold standard morphology; morph: Morpheus; UCG: UPOS+Case+Gender, UCN: UPOS+Case+Number, UCGN: UPOS+Case+Gender+Number: UALLo: UPOSAllFeaturesOrdered. Underline: Best model per language and type of label; *: best overall per language.

	IN-DOMAIN			OUT-OF-DOMAIN		
	NO TAG					
	ixa-mm	ixa-gs	morph	ixa-mm	ixa-gs	morph
en	–	96.34	97.51	–	90.40	92.47
es	–	98.53	98.17	–	89.75	89.70
ru	–	92.81	95.31	–	83.95	86.84
eu	–	90.61	95.69	–	85.64	88.25
cs	–	96.37	98.31	–	91.50	91.61
tr	–	87.11	95.62	–	77.16	84.07
	UPOS					
	ixa-mm	ixa-gs	morph	ixa-mm	ixa-gs	morph
en	99.11*	98.91	98.10	95.38*	95.25	92.92
es	98.91	98.76	98.94	97.53	97.41	90.29
ru	94.36	93.74	96.20	90.00	89.40	87.59
eu	93.11	92.29	96.39	85.22	86.79	88.97
cs	97.86	97.28	98.75	92.33	93.68	91.66
tr	84.65	87.76	96.44*	79.22	81.67	84.96*
	UCG					
	ixa-mm	ixa gs	morph	ixa-mm	ixa-gs	morph
en	99.10	98.92	97.99	95.20	95.24	92.97
es	98.94	98.70	98.98	97.54	97.43	90.31
ru	94.85	93.30	96.21	90.97	89.33	87.67
eu	92.65	92.39	96.34	85.23	86.74	89.09
cs	97.29	96.64	98.76*	91.61	91.35	91.92
tr	85.09	87.09	96.18	80.06	81.23	84.74
	UCN					
	ixa-mm	ixa-gs	morph	ixa-mm	ixa-gs	morph
en	99.06	98.87	98.01	95.16	95.16	92.86
es	98.92	98.75	99.02*	97.56	97.44	90.35
ru	95.07	93.70	96.20	91.00*	89.60	87.58
eu	93.03	92.35	96.39	85.47	86.36	89.03
cs	97.44	96.87	98.71	91.04	92.07	92.23*
tr	85.52	87.18	96.11	80.33	81.00	84.40
	UCGN					
	ixa-mm	ixa-gs	morph	ixa-mm	ixa gs	morph
en	99.08	98.96	97.99	95.21	95.15	92.95
es	98.89	98.71	98.97	97.59*	97.44	90.38
ru	95.00	93.08	96.44*	90.80	89.13	87.66
eu	93.03	92.28	96.39	85.38	86.55	88.86
cs	97.17	96.68	98.70	91.71	91.50	91.97
tr	85.52	87.18	96.20	80.33	81.00	84.46
	UAllo					
	ixa-mm	ixa-gs	morph	ixa-mm	ixa-gs	morph
en	99.04	98.95	98.06	95.08	95.13	93.15
es	98.86	98.74	99.00	97.54	97.45	90.34
ru	94.75	93.22	96.30	90.88	88.66	87.57
eu	93.41	94.06	96.50*	85.33	86.31	89.11*
cs	97.03	96.63	98.70	91.19	91.81	92.02
tr	84.90	86.57	96.22	79.39	80.50	84.96

Following this, we would like to evaluate the out-of-domain performance when not even UPOS labels are used for training. From what we have seen in-domain, the systems that operate without morphology achieve competitive results with respect to the models using morphological information. Figure 3 provides an overview of both the in- and out-of-domain results obtained for both types of systems, confirming this trend. Thus, it is remarkable that the XLM-RoBERTa model scores best out-of-domain for Turkish and Czech, and a very close second in Russian. The results for Spanish and English deserve further analysis, as the IXA pipes statistical models clearly outperform every other system for these two languages, with the differences around 7 points in word accuracy.

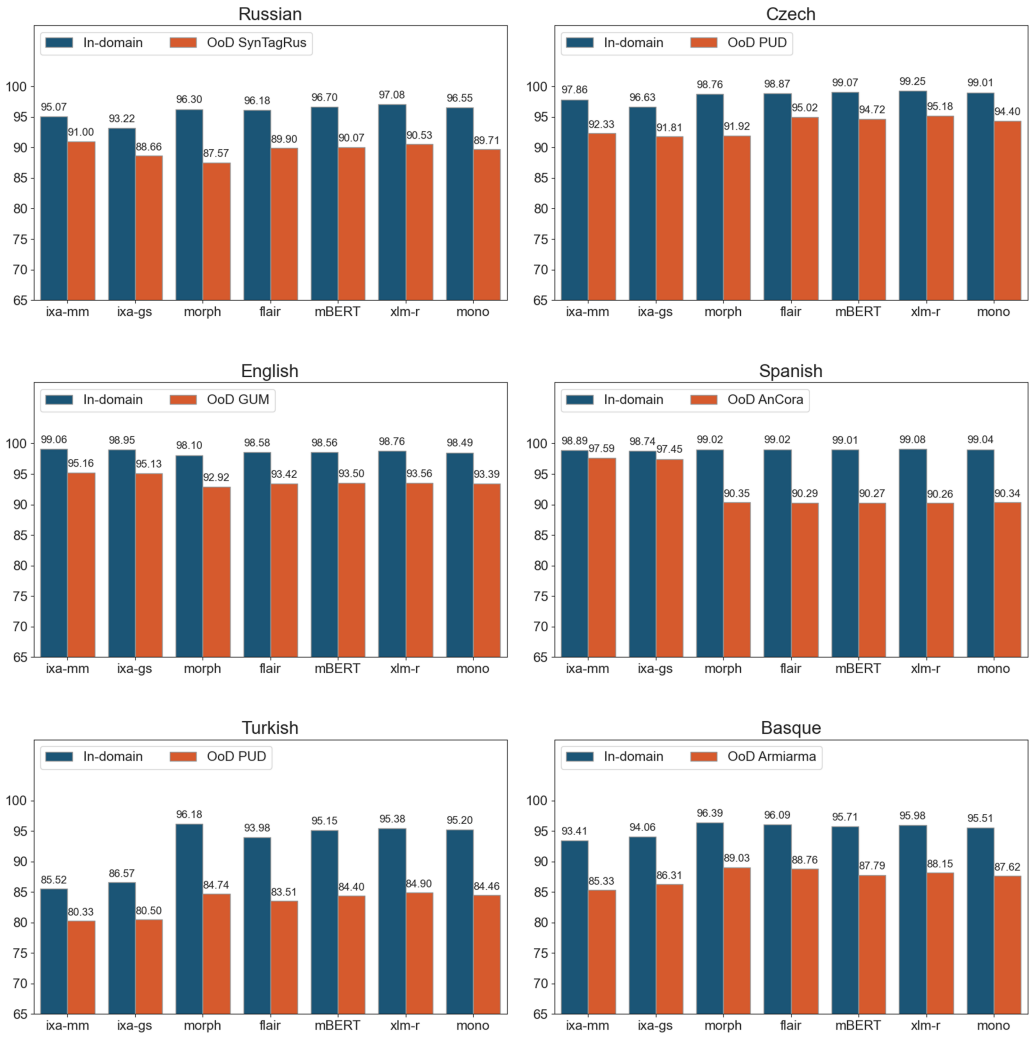

Figure 3
Overall in-domain and out-of-domain results.

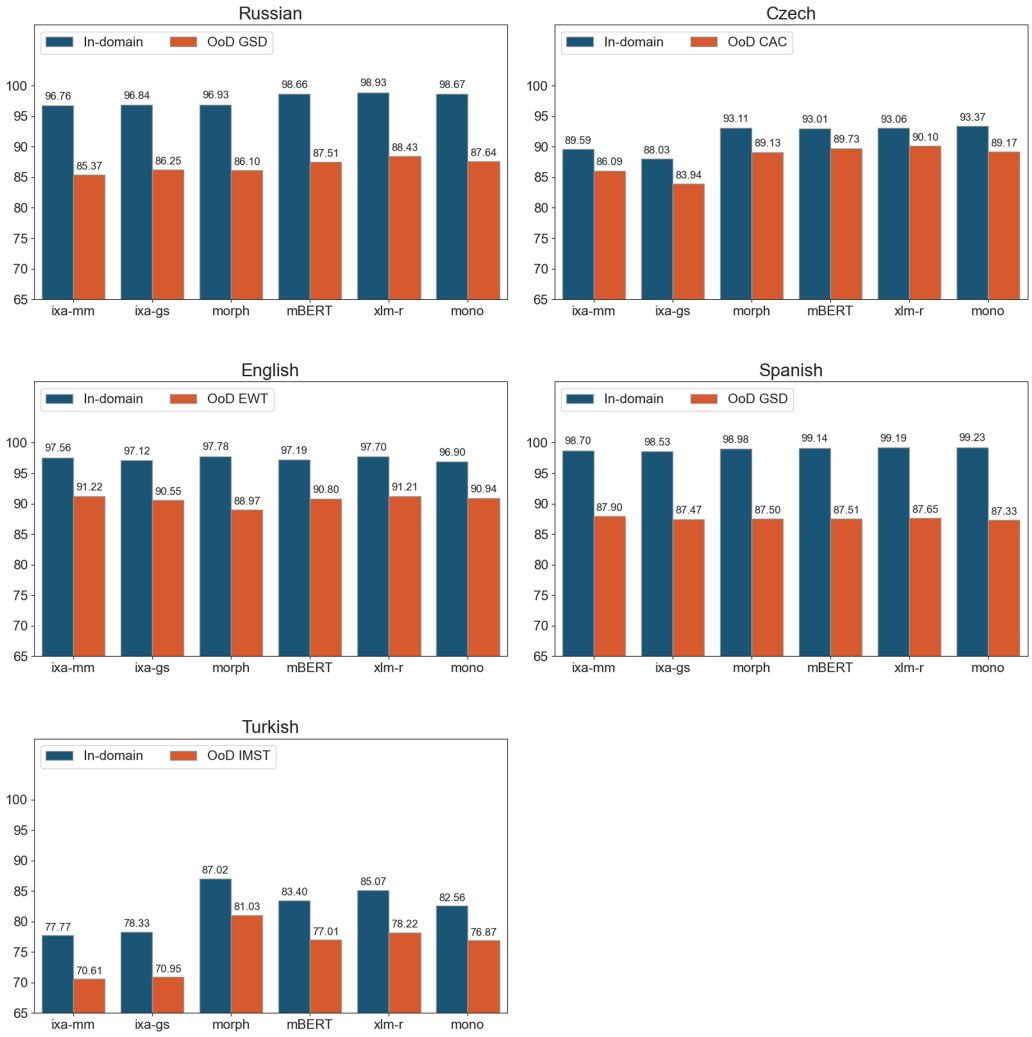

Figure 4
Overall in- and out-of-domain results in the reversed setting.

Figure 4[6] presents the reversed results of those presented in Figure 3, namely, the test set of the in-domain corpora becomes the out-of-domain test data while the models are fine-tuned on the training split of the out-of-domain data. Doing this experiment allows us to discard that the out-of-domain behavior exhibited in previous results could be due to differences in size between the training in-domain data and the testing out-of-domain test sets. Good examples of this are Russian and Spanish, for which SynTagRus and AnCora are used as in-domain data in the reversed setting. These two datasets are much larger than the GSD corpora for those languages (used as in-domain data in

6 Basque is not present in this evaluation due to the fact that the Armiarma corpus does not include UniMorph annotations.

the original setting). Thus, results in the reversed setting demonstrate that: (i) out-of-domain performance worsens substantially regardless of the language and model; (ii) language models fine-tuned without explicit morphological information outperform every other model for all languages except Turkish for the in-domain evaluation setting; and (iii) the out-of-domain results of XLM-RoBERTa-base are the best for Russian and Czech and similar to other models in English and Spanish.

In any case, Figures 3 and 4 show that the results of every model significantly degrade when evaluated out-of-domain, the most common application of lemmatizers. Thus, even for high-scoring languages such as English and Spanish, out-of-domain performance worsens between 3 and 5 points in word accuracy. For high-inflected languages the differences are around 8 for Basque and more than 10 for Turkish.

Given that pre-trained language models such as XLM-RoBERTa-base can be leveraged to learned competitive lemmatizers without using any explicit morphological signal, we propose a final experiment to address the following two additional research questions. First, will lemmatization results get closer to the state-of-the-art by using a larger transformer-based model such as XLM-RoBERTa-large? Second, can we improve the performance of a language model such as XLM-RoBERTa by adding morphological information during fine-tuning?

Table 8 shows the results of experimenting with XLM-RoBERTa-base and XLM-RoBERTa-large to learn lemmatization as a sequence labeling task with and without adding morphology as explicit handcrafted features. For each language we pick the best morphological configuration from Table 7 and encode the morphological labels as feature embeddings. Both feature and encoded text embeddings are then sent into a softmax layer for sequence labeling (Wang et al. 2022). The first observation is that the large version of XLM-RoBERTa obtains the best results both in- and out-of domain. It is particularly noteworthy that fine-tuning XLM-RoBERTa-large with only the SES classes helps to outperform any other model for every language and evaluation setting. Furthermore, adding morphology as a feature seems to be beneficial. In fact, the morphologically informed models are the best in 4 out of 6 in-domain evaluations and for all 6 out-of-domain cases.

We compute the McNemar test to establish whether the differences obtained with and without morphological features are actually statistically significant. It turns out

Table 8
In- and out-of-domain results for XLM-RoBERTa-base and XLM-RoBERTa-large models with and without morphological features during training.

	xlm-r base				xlm-r large			
	in-domain		out-of-domain		in-domain		out-of-domain	
	without morph.	with morph.	without morph.	with morph.	without morph.	with morph.	without morph.	with morph.
en	98.76	98.74	93.56	93.72	98.85	98.92	93.82	93.86
es	99.08	99.10	90.26	90.42	99.12	99.15	90.48	90.53
eu	95.98	96.45	88.15	88.60	96.66	96.70	88.75	88.81
ru	97.08	97.25	90.53	90.92	97.63	97.96	91.60	91.71
cz	99.25	99.32	95.18	94.72	99.40	99.23	95.42	96.06
tr	95.38	95.19	84.90	85.34	96.30	96.13	85.18	85.40

that for XLM-RoBERTa-large, results are rather mixed. Thus, only for Russian (p-value 0.003) and Czech (0.000) are the results significant at $\alpha = .05$. For Turkish and Basque the results are not conclusive (p-value 0.0495) while for the rest the null hypothesis cannot be rejected (0.423 for Spanish, 0.242 in English, and 0.547 in Basque). Regarding XLM-RoBERTa-base, in 4 out of 6 languages the results are statistically significant at $\alpha = .01$ (the McNemar test), failing to reject the null hypothesis for Russian and Turkish.

To sum up, our experiments empirically demonstrate that fine-grained morphological information to train contextual lemmatizers does not lead to substantially better in- or out-of-domain performance, not even for languages of varied complex morphology, such as Basque, Czech, Russian, and Turkish. Thus, only for Basque and Turkish did Morpheus (using UPOS tags) outperform XLM-RoBERTa models.

Taking this into account, and as previously hypothesized for other NLP tasks (Manning et al. 2020), modern contextual word representations seem to implicitly capture morphological information valuable to train lemmatizers without requiring any explicit morphological signal. We have proved this by training off-the-shelf language models to perform lemmatization as a token classification task obtaining state-of-the-art results for Russian and Czech, and very close performance to UDPipe in the rest. Finally, statistical models are only competitive to perform contextual lemmatization on languages with a morphology on the simple side of the complexity spectrum, such as English or Spanish.

Thus, the results indicate that XLM-RoBERTa-large is the optimal option to learn lemmatization without any explicit morphological signal for every language and evaluation setting.

7. Discussion

In this article we performed a number of experiments to better understand the role of morphological information to learn contextual lemmatization. Our findings can be summarized as follows: (i) fine-grained morphological information does not help to substantially improve contextual lemmatization, not even for high-inflected languages; using UPOS tags seems to be enough for comparable performance; (ii) contextual word representations such as those used in transformer and Flair models seem to encode enough implicit morphological information to allow us to train good performing lemmatizers without any explicit morphological signal; (iii) the best-performing lemmatizers out-of-domain are those using either simple UPOS tags or no morphology at all; (iv) evaluating lemmatization on word accuracy is not the best strategy—results are too high and too similar to each other to be able to discriminate between models. By using word accuracy we are assigning the same importance to cases in which the lemma is equivalent to the word form (e.g., "the") as to complex cases in which the word form includes case, number, and/or gender information (e.g, *medikuarenera*, which in Basque means 'to the doctor', with its corresponding lemma *mediku*). We believe that this may lead to a high overestimation in the evaluation of the lemmatizers.

In this section, we address some remaining open issues with the aim of understanding better the main errors and difficulties still facing lemmatization. First, we discuss the convenience of using an alternative metric to word accuracy. Second, we analyze the performance of XLM-RoBERTa-base by evaluating accuracy per SES. Third, we examine the generalization capabilities of XLM-RoBERTa-base by computing word accuracy for in-vocabulary and out-of-vocabulary words. We also discuss any issues regarding test data contamination. Finally, we perform some error analysis on the

out-of-domain performance of the XLM-RoBERTa-base model for Spanish, to see why it is different from the rest of the languages, as illustrated by Figure 3.

7.1 Sentence Accuracy

Looking at the in-domain results for lemmatization reported in the previous sections and in the majority of recent work (Malaviya, Wu, and Cotterell 2019; McCarthy et al. 2019; Yildiz and Tantuğ 2019; Straka, Straková, and Hajic 2019), with word accuracy in-domain scores around 96 or higher, it is not surprising to wonder whether contextual lemmatization is a solved task. However, if we look at the evaluation method a bit more closely, things are not as clear as they seem. As has been argued for POS tagging (Manning 2011), word accuracy as an evaluation measure is easy because you get many free points for punctuation marks and for the many tokens that are not ambiguous with respect to its lemma, namely, those cases in which the lemma and the word form are the same. Following this, a more realistic metric might consist of looking at the rate of getting the whole sentence correctly lemmatized, just as was proposed for POS tagging (Manning 2011).

Figure 5 reports the sentence accuracy of the six languages we used in our experiments both for in- and out-of-domain. In contrast to the word accuracies reported in Figure 3, we can see that the corresponding sentence accuracy results drop significantly. In addition to demonstrating that lemmatizers have a large margin of improvement, sentence accuracy allows us to better discriminate between different models. We can see this phenomenon in the English and Spanish results. Thus, while every model obtained very similar in-domain word accuracy in Spanish, using sentence accuracy helps to discriminate between the statistical and the neural lemmatizers. Furthermore, it also shows that among the neural models XLM-RoBERTa clearly outperforms the rest of the models by almost 1 percent.

The effect of sentence accuracy for the in-domain evaluation is vastly magnified when considering out-of-domain performance, with the extremely low scores across languages providing further evidence of how far lemmatization remains from being solved.

7.2 Analyzing Word Accuracy per SES

The next natural step in our analysis is identifying which specific cases are most difficult for lemmatizers. In order to do so, we look at the word accuracy for each of the SES labels automatically induced from the data. In order to illustrate this point, we take XLM-RoBERTa-base as an example use case and analyze its predictions for the languages which could be inspected in-house, namely, Basque, English, Spanish, and Russian. Thus, Table 9 presents examples and results for the 10 most frequent SES for each of these 4 languages' development sets.

As we can see in Table 9, the most common lemma transformation to be learned is based on the edit script "do nothing", namely, the lemmatizer needs to learn that the lemma and the word have the same form. It is also interesting to see how the ratio of such lemma type changes across languages, from English, where such cases are observed in almost 77% of the cases, to Basque, where only half of the lemmas correspond to this rule. However, in terms of word accuracy, the results are remarkably similar for all 4 languages, in the range of 99%–99.30%. This demonstrates that the traditional evaluation method greatly overestimates the lemmatizers' performance.

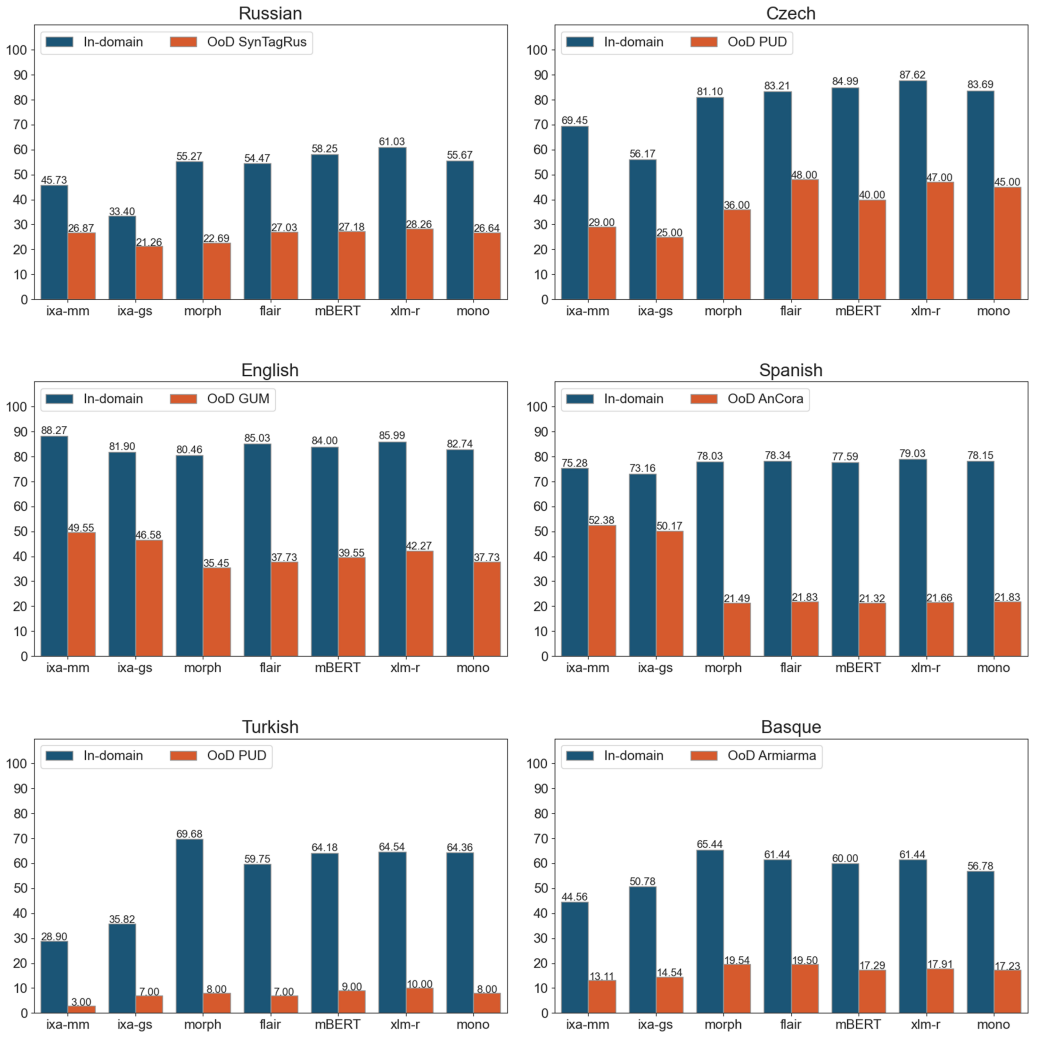

Figure 5
Sentence accuracy results for in- and out-of-domain settings.

By looking at other specific cases, we can see that in English problematic examples to learn are those related to the casing of some characters (e.g., Martin → *Martin*, NASA → *NASA*). Another noticeable issue refers to the verbs in gerund form (e.g., try**ing** → *try*, driv**ing** → *drive*).

With respect to Spanish, interesting difficult lemmas are observed with articles in feminine form (e.g., la → *el*, una → *uno*), where the masculine form is considered the canonical form or lemma, and feminine articles and adjectives should be lemmatized by changing the gender of the word from female to male.

In Russian the most challenging case corresponds to the lemmatization of the nouns that end with a soft sign ь with the word accuracy for this SES as low as 93.94%. The possible reason for such low accuracy could be the absence of a specific grammar rule that defines the gender of such nouns and, therefore, the termination these nouns have in different cases. The second lowest accuracy among the 10 most popular SES

Table 9
10 most frequent SES, brief description, corresponding word accuracy, weight (in %) in the corpus and examples of words and their lemmas for English, Spanish, Russian, and Basque; SES are computed following UDPipe's method (Straka, Straková, and Hajic 2019).

	SES	Casing	Edit script	W.acc	%	Examples
	↓0;d¦+	all low	do nothing	99.29	**76.87**%	positive→*positive*
	↑ 0¦ ↓1;d¦+	1st up	do nothing	96.29	6.97%	Martin→*Martin*
	↓0;d¦-+	all low	remove last ch	98.58	5.52%	things→*thing*
	↓0;abe	all low	ignore form, use *be*	99.81	2.02%	is→*be*
en	↓0;d¦–+	all low	remove 2 last ch	97.42	1.52%	does→*do*
	↓0;d¦—+	all low	remove 3 last ch	96.45	1.10%	trying→*try*
	↑ 0¦ ↓-1;d¦+	all up	do nothing	94.22	0.68%	NASA→*NASA*
	↓0;d–+b¦+	all low	first 2 char to *b*	99.33	0.59%	are→*be*
	↓0;d¦-+v+e+	all low	last ch to *ve*	100.00	0.51%	has→*have*
	↓0;d¦—+e+	all low	3 last ch to *e*	96.23	0.42%	driving→*drive*
	↓0;d¦+	all low	do nothing	99.36	**72.40**%	acuerdo→*acuerdo*
	↓0;d¦-+	all low	del last ch	97.22	5.29%	estrellass→*estrella*
	↓0;d+e¦-+	all low	add *e*, del last ch	96.73	3.36%	la→*el*
	↓0;d¦-+o+	all low	del last ch, add *o*	96.21	2.37%	una→*uno*
es	↓0;d+e¦–+	all low	add *e*, del 2 last ch	99.78	2.13%	los→*el*
	↓0;d¦–+	all low	del 2 last ch	97.36	1.40%	flores→*flor*
	↓0;aél	all low	ignore form, use *él*	99.83	1.32%	se→*él*
	↓0;d¦+r+	all low	add *r*	100.00	0.91%	hace→*hacer*
	↓0;d¦+o+	all low	add *o*	97.73	0.91%	primer→*primero*
	↓0;d¦–+a+r+	all low	del last ch, add *ar*	98.07	0.83%	desarrolló→*desarollar*
	↓0;d¦+	all low	do nothing	99.16	**57.80**%	Петербург→Петербург
	↓0;d¦–+	all low	del last ch	97.67	6.97%	церковью→церковь
	↓0;d¦–+a+	all low	del last ch, add a	96.65	3.32%	экономику→экономика
	↓0;d¦–+й+	all low	del last ch, add й	96.08	3.10%	городское→городской
ru	↓0;d¦–+	all low	del 2 last ch	99.03	2.10%	странами→страна
	↓0;d¦–+e+	all low	del last ch, add e	98.04	2.07%	моря→море
	↓0;d¦–+я+	all low	del last ch, add я	97.83	1.86%	историю→история
	↓0;d¦–+т+ь+	all low	del last ch, add ть	98.88	1.81%	получил→получить
	↓0;d¦–+ь+	all low	del last ch, add ь	93.94	1.67%	сентября→сентябрь
	↓0;d¦–+т+ь+	all low	del 2 last, add ть	98.10	1.60%	были→быть
	↓0;d¦+	all low	do nothing	99.05	**49.63**%	sartu→*sartu*
	↓0;d¦–+	all low	remove 2 last ch	97.72	9.93%	librean→*libre*
	↓0;d¦–+	all low	remove last ch	96.27	6.54%	korrikan→*korrika*
	↓0;d¦—+	all low	remove 3 last ch	93.24	3.60%	aldaketarik→*aldaketa*
eu	↑ 0¦↓1;d¦+	1st up	do nothing	98.54	3.46%	MAPEI→*Mapei*
	↓0;d¦—-+	all low	del 4 last ch	93.00	2.52%	lagunaren→*lagun*
	↑ 0¦↓1;d¦–+	1st up	del 2 last ch	95.54	1.88%	Egiptora→*Egipto*
	↓0;d-+i+z ¦+n+	all low	del 1st ch, add *iz*,*n*	100.00	1.38%	da→*izan*
	↑ 0¦↓1;d¦-+	1st up	del last ch	90.08	1.10%	Frantziak→*Frantzia*

in Russian is for adjectives, cases in which to obtain the lemma one should delete the last character of the word and add a letter й (pronounced as *iy kratkoe*, short y), that in Russian determines the suffix for some masculine nouns and adjectives in singular and nominative case. The words could be in different cases and genders, so it is necessary to know such information for correct lemmatization (e.g., городское → городской [neutral gender, nominative case], семейным → семейный [masculine gender, instrumental case]).

Table 10
Word accuracy for in-vocabulary and out-of-vocabulary words for XLM-RoBERTa-base model (original setting). Corpora: English - EWT (in-domain), GUM (out-of-domain); Spanish - GSD (in-domain), AnCora (out-of-domain); Basque - BDT (in-domain), Armiarma (out-of-domain); Russian - GSD (in-domain), SynTagRus (out-of-domain); Czech - CAC (in-domain), PUD (out-of-domain); Turkish - IMST (in-domain), PUD (out-of-domain).

	In-domain		Out-of-domain	
	In-vocabulary	Out-of-vocabulary	In-vocabulary	Out-of-vocabulary
en	97.25	90.55	92.56	81.11
es	98.06	93.54	82.53	60.07
eu	96.65	82.85	87.92	71.08
ru	98.62	90.23	89.19	77.50
cz	99.21	93.28	98.08	88.66
tr	97.84	84.54	92.34	68.39

Finally, for Basque the most problematic cases with a rather low word accuracy of only 90.08% can be found among the nouns in ergative (e.g., Frantzia**k** → *Frantzia*) or locative cases (e.g., Mosku**n** [in Moscow] → Mosku, Katalunia**n** [in Catalonia] → Katalunia). The other two most difficult SES occur when the word forms are in possessive case (e.g., lagun**aren** → lagun) and for nouns in indefinite form (e.g., aldaketa**rik** [change] → aldaketa).

It should be noted that an extra obstacle to improving some of these difficult cases is the low number of samples available. Nonetheless, this analysis shows that lemmatizers still do not properly learn to lemmatize relatively common word forms.

7.3 Generalization Capabilities of Language Models

In this subsection we aim to analyze the generalization capabilities of a MLM such as XLM-RoBERTa-base in the lemmatization task. More specifically, we will discuss two issues: (i) whether MLMs simply memorize the SES lemma classes during fine-tuning and (ii) whether the good performance of MLMs in this task might be due to some test data contamination.[7]

In order to address the first point, we evaluate the performance of XLM-RoBERTa-base, fine-tuned without morphological features, for those words seen during fine-tuning (in-vocabulary words) with respect to out-of-vocabulary occurrences.

Tables 10 and 11 report the results for both original and reversed settings and in- and out-of-domain evaluations. It is noticeable that the model performs very well on out-of-vocabulary words, also in the out-of-domain evaluation, which would seem to indicate that XLM-RoBERTa is generalizing beyond the words seen during training. This seems to be confirmed also by looking at the Spanish and Russian results. It should be remembered that, while in the reversed setting the training data for Spanish (AnCora, 500K tokens) and Russian (SynTagRus, 900K words) is much larger than in the original setting (both GSD), the obtained results reflect roughly the same trend.

7 https://hitz-zentroa.github.io/lm-contamination/.

Table 11
Word accuracy for in-vocabulary and out-of-vocabulary words for XLM-RoBERTa-base model (reversed setting). Corpora: English - GUM (in-domain), EWT (out-of-domain); Spanish - AnCora (in-domain), GSD (out-of-domain); Russian - SynTagRus (in-domain), GSD (out-of-domain); Czech - PUD (in-domain), CAC (out-of-domain); Turkish - PUD (in-domain), IMST (out-of-domain).

	In-domain		Out-of-domain	
	In-vocabulary	Out-of-vocabulary	In-vocabulary	Out-of-vocabulary
en	96.48	88.95	89.84	75.94
es	98.54	93.11	81.71	47.10
ru	98.70	92.28	89.22	59.13
cz	96.26	83.51	90.42	82.35
tr	90.11	70.15	88.63	58.79

Finally, we should consider whether a MLM such as XLM-RoBERTa has already seen the datasets we are experimenting with during pre-training, namely, whether XLM-RoBERTa has been contaminated.[8] First, it should be noted that CC-100, the corpus used to generate XLM-RoBERTa, was constructed by processing the CommonCrawl snapshots from between January and December 2018. Second, the SIGMORPHON data we are using was released in 2019[9] with the test data including gold standard lemma and UniMorph annotations being released in April 2019. Third and most importantly, XLM-RoBERTa does not see the lemmas themselves during training or inference, but the SES classes we automatically generate in an ad-hoc manner for the experimentation. The datasets containing both the words and the SES classes used have not been made publicly available.

Based on this, it is possible to say that XLM-RoBERTa seems to generalize over unseen words and that its performance is not justified by any form of language model contamination.

7.4 Analyzing Spanish Out-of-domain Results

In Section 6.2 we saw that out-of-domain performance of transformer-based models for Spanish was not following the pattern of the rest of the languages. Instead, they were 6%–7% worse than the results obtained by the IXA pipes statistical lemmatizers (ixa-pipe-mm and ixa-pipe-gs). By checking the most common error patterns of XLM-RoBERTa-base, we found that most of the performance loss was caused by inconsistencies in the manual annotation of lemmas between the data used for in-domain and out-of-domain evaluation. More specifically, the GSD Spanish corpus included in UniMorph wrongly annotates lemmas for proper names such as Madrid, London, or Paris entirely in lowercase, namely, madrid, london, and paris. However, the AnCora Spanish corpus used for out-of-domain evaluation correctly annotates these cases specifying their corresponding lemmas with the first character in uppercase. This inconsistency results in 3,781 examples of proper names in the AnCora test set which are all lemmatized following the pattern seen during training with the GSD training set. Con-

8 https://hitz-zentroa.github.io/lm-contamination/blog/.
9 First GitHub commit December 19, 2018.

sequently, the word accuracy obtained by the model for these types of examples in the AnCora test set is 0%. In order to confirm this issue, we corrected the wrongly annotated proper names in the GSD training data, fine-tuned again the model and saw the out-of-domain performance of XLM-RoBERTa-base go up from 90.26% to 96.75%, a more consistent result with respect to the out-of-domain scores for the other 5 languages.

This issue manifests the importance of consistent manual annotation across corpora from different domains in order to fairly evaluate out-of-domain performance of contextual lemmatizers.

8. Concluding Remarks

Lemmatization remains an important natural language processing task, especially for languages with high-inflected morphology. In this article we provide an in-depth study on the role of morphological information to learn contextual lemmatizers. By taking a language sample of varied morphological complexity, we have analyzed whether a fine-grained morphological signal is indeed beneficial for contextual lemmatization. Furthermore, and in contrast to previous work, we also evaluate lemmatizers in an out-of-domain setting, which constitutes, after all, their most common application use. Our results empirically demonstrate that informing lemmatizers with fine-grained morphological features during training is not that beneficial, not even for agglutinative languages. In fact, modern contextual word representations seem to implicitly encode enough morphological information to obtain good contextual lemmatizers without seeing any explicit morphological signal. Finally, good out-of-domain performance can be achieved using simple UPOS tags or without any explicit morphological signal.

Therefore, our results suggest that an optimal solution among all the options considered would be to develop lemmatizers by fine-tuning a large MLM such as XLM-RoBERTa-large without any explicit morphological signal. Addressing lemmatization as a token classification task results in highly competitive and robust lemmatizers with results over or close to the state-of-the-art obtained with more complex methods (Straka, Straková, and Hajic 2019).

Furthermore, we have discussed current evaluation practices for lemmatization, showing that using simple word accuracy is not adequate to clearly discriminate between models, as it provides a deceptive view regarding the performance of lemmatizers. An additional analysis looking at specific lemma classes (SES) has shown that many common word forms are still not properly predicted. The conclusion is that lemmatization remains a challenging task. Future work is therefore needed to improve out-of-domain results. Furthermore, it is perhaps a good time to propose an alternative word-level metric to evaluate lemmatization that, complemented with sentence accuracy, may provide a more realistic view of the performance of contextual lemmatizers.

Appendix A. Detailed Lemmatization Results

Table A1
Overall in-domain lemmatization results for models trained with and without explicit morphological features; monolingual transformers: Russian - ruBERT, Czech - slavicBERT, Basque -BERTeus, Turkish - BERTurk, English - RoBERTa, Spanish - BETO.

	ixa-mm	ixa-gs	morph	flair	mBERT	xlm-r	mono	base	UDPipe
en	99.06	98.95	98.10	98.58	98.56	98.76	98.49	97.68	99.01
es	98.89	98.74	99.02	99.02	99.01	99.08	99.04	98.42	99.31
ru	95.07	93.22	96.30	96.18	96.70	97.08	96.55	95.67	97.77
eu	93.41	94.06	96.39	96.09	95.71	95.98	95.51	96.07	97.14
cz	97.86	96.63	98.76	98.87	99.07	99.25	99.01	97.82	99.31
tr	85.52	86.57	96.18	93.98	95.15	95.38	95.20	96.41	96.84

Table A2
Overall out-of-domain lemmatization results for models with and without explicit morphological features; monolingual transformers: Russian - ruBERT, Czech - slavicBERT, Basque -BERTeus, Turkish - BERTurk, English - RoBERTa, Spanish - BETO.

	ixa-mm	ixa-gs	morph	flair	mBERT	xlm-r	mono
en	95.16	95.13	92.92	93.42	93.50	93.56	93.39
es	97.59	97.45	90.35	90.29	90.27	90.26	90.34
ru	91.00	88.66	87.57	89.90	90.07	90.53	89.71
eu	85.33	86.31	89.03	88.76	87.79	88.15	87.62
cz	92.33	91.81	91.92	95.02	94.72	95.18	94.40
tr	80.33	80.50	84.74	83.51	84.40	84.90	84.46

Table A3
In-domain sentence accuracy results; monolingual transformers: Russian - ruBERT, Czech - slavicBERT, Basque -BERTeus, Turkish - BERTurk, English - RoBERTa, Spanish - BETO.

	ixa-mm	ixa-gs	morph	flair	mBERT	xlm-r	mono
en	88.27	81.90	80.46	85.03	84.00	85.99	82.74
es	75.28	73.16	78.03	78.34	77.59	79.03	78.15
ru	45.73	33.40	55.27	54.47	58.25	61.03	55.67
eu	44.56	50.78	65.44	61.44	60.00	61.44	56.78
cz	69.45	56.17	81.10	83.21	84.99	87.62	83.69
tr	28.90	35.82	69.68	59.75	64.18	64.54	64.36

Table A4
Out-of-domain sentence accuracy results; monolingual transformers: Russian - ruBERT, Czech - slavicBERT, Basque -BERTeus, Turkish - BERTurk, English - RoBERTa, Spanish - BETO.

	ixa-mm	ixa-gs	morph	flair	mBERT	xlm-r	mono
en	49.55	46.58	35.45	37.73	39.55	42.27	37.73
es	52.38	50.17	21.49	21.83	21.32	21.66	21.83
ru	26.87	21.26	22.69	27.03	27.18	28.26	26.64
eu	13.11	14.54	19.54	19.50	17.29	17.91	17.23
cz	29.00	25.00	36.00	48.00	40.00	47.00	45.00
tr	3.00	7.00	8.00	7.00	9.00	10.00	8.00

Table A5
Overall in-domain lemmatization results (reversed setting) for models with and without explicit morphological features; monolingual transformers: Russian - ruBERT, Czech - slavicBERT, Basque -BERTeus, Turkish - BERTurk, English - RoBERTa, Spanish - BETO.

	ixa-mm	ixa-gs	morph	mBERT	xlm-r	mono	base	UDPipe
en	97.56	97.12	97.78	97.19	97.70	96.90	97.41	98.63
es	98.70	98.53	98.98	99.14	99.19	99.23	98.54	99.46
ru	96.76	96.84	96.93	98.66	98.93	98.67	95.92	98.92
cz	89.59	88.03	93.11	93.01	93.06	93.37	93.58	98.13
tr	77.77	78.33	87.02	83.40	85.07	82.56	86.02	89.03

Table A6
Overall out-of-domain lemmatization results (reversed setting) for models with and without explicit morphological features; monolingual transformers: Russian - ruBERT, Czech - slavicBERT, Basque -BERTeus, Turkish - BERTurk, English - RoBERTa, Spanish - BETO.

	ixa-mm	ixa-gs	morph	mBERT	xlm-r	mono
en	91.22	90.55	88.97	90.80	91.21	90.94
es	87.90	87.47	87.50	87.51	87.65	87.33
ru	85.37	86.25	86.10	87.51	88.43	87.64
cz	86.09	83.94	89.13	89.73	90.10	89.17
tr	70.61	70.95	81.03	77.01	78.22	76.87

Acknowledgments

This work has been supported by the HiTZ center and the Basque Government (research group funding IT-1805-22). Olia Toporkov is funded by a UPV/EHU grant "Formación de Personal Investigador". We also thank the funding from the following MCIN/AEI/10.13039/501100011033 projects: (i) DeepKnowledge (PID2021-127777OB-C21) and ERDF A way of making Europe; (ii) Disargue (TED2021-130810B-C21) and European Union NextGenerationEU/PRTR; (iii) Antidote (PCI2020-120717-2) and European Union NextGenerationEU/PRTR. Rodrigo Agerri currently holds the RYC-2017-23647 (MCIN/AEI/10.13039/501100011033 and ESF Investing in Your Future) fellowship.

References

Agerri, Rodrigo, Josu Bermudez, and German Rigau. 2014. IXA pipeline: Efficient and ready to use multilingual NLP tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3823–3828.

Agerri, Rodrigo and German Rigau. 2016. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238(2):63–82. Chakrabarty. https://doi.org/10.1016/j.artint.2016.05.003

Agerri, Rodrigo, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: The case for Basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4781–4788.

Agić, Željko and Natalie Schluter. 2017. How (not) to train a dependency parser: The curious case of jackknifing part-of-speech taggers. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 679–684. https://doi.org/10.18653/v1/P17-2107

Aiken, Brad, Jared Kelly, Alexis Palmer, Suleyman Olcay Polat, Taraka Rama, and Rodney Nielsen. 2019. Sigmorphon 2019 task 2 system description paper: Morphological analysis in context for many languages, with supervision from only a few. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 87–94. https://doi.org/10.18653/v1/W19-4211

Akbik, Alan, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Aldezabal, I., M. J. Aranzabe, A. Diaz de Ilarraza, and K. Fernández. 2008. From dependencies to constituents in the reference corpus for the processing of Basque. *Procesamiento del Lenguaje Natural*, (41):147–154.

Alegria, Iñaki, Xabier Artola, Kepa Sarasola, and Miriam Urkia. 1996. Automatic morphological analysis of Basque. *Literary and Linguistic Computing*, 11:193–203. https://doi.org/10.1093/llc/11.4.193

Arkhipov, Mikhail, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93. https://doi.org/10.18653/v1/W19-3712

Belinkov, Yonatan, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James R. Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872. https://doi.org/10.18653/v1/P17-1080

Bergmanis, Toms 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. https://doi.org/10.18653/v1/N18-1126

Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.

Carreras, Xavier, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004. FreeLing: An open-source suite of language analyzers. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 239–242.

Chakrabarty, Abhisek, Onkar Arun Pandit, and Utpal Garain. 2017. Context sensitive lemmatization using two successive bidirectional gated recurrent networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1481–1491. https://doi.org/10.18653/v1/P17-1136

Chrupala, Grzegorz, Georgiana Dinu, and Josef van Genabith. 2008. Learning morphology with Morfette. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Collins, Michael. 2002. Discriminative training methods for Hidden Markov Models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 1–8. https://doi.org/10.3115/1118693.1118694

Conneau, Alexis, 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. https://doi.org/10.18653/v1/2020.acl-main.747

Conneau, Alexis, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136. https://doi.org/10.18653/v1/P18-1198

Cotterell, Ryan and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759. https://doi.org/10.18653/v1/D17-1078

Dayanik, Erenay, Ekin Akyürek, and Deniz Yuret. 2018. MorphNet: A sequence-to-sequence model that combines morphological analysis and disambiguation. *CoRR*, abs/1805.07946.

de Marneffe, Marie Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186.

Dhonnchadha, Elaine Uí. 2002. A two-level morphological analyser and generator for Irish using finite-state transducers. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*.

Gesmundo, Andrea 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.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Heigold, Georg, Guenter Neumann, and Josef van Genabith. 2017. An extensive empirical evaluation of character-based morphological tagging for 14 languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 505–513. https://doi.org/10.18653/v1/E17-1048

Hladká, Barbora, Jan Hajic, Jirka Hana, Jaroslava Hlavácová, Jirí Mírovský, and Jan Raab. 2008. The Czech academic corpus 2.0 guide. *The Prague Bulletin of Mathematical Linguistics*, 89:41–96. https://doi.org/10.2478/v10108-009-0003-9

Huang, Zhiheng, Wei Xu, and Kailiang Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Jongejan, Bart and Hercules Dalianis. 2009. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 145–153. https://doi.org/10.3115/1687878.1687900

Karttunen, Lauri, Ronald M. Kaplan, and Annie Zaenen. 1992. Two-level morphology with composition. In *COLING 1992 Volume 1: The 14th International Conference on Computational Linguistics*. https://doi.org/10.3115/992066.992091

Kingma, Diederik P. and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Kirov, Christo, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA).

Kondratyuk, Dan. 2019. Cross-lingual lemmatization and morphology tagging with two-stage multilingual BERT fine-tuning. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*,

pages 12–18. https://doi.org/10.18653/v1/W19-4203

Kondratyuk, Dan and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. 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 2779–2795. https://doi.org/10.18653/v1/D19-1279

Kuratov, Yuri and Mikhail Y. Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. *CoRR*, abs/1905.07213.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Lyashevkaya, Olga, Kira Droganova, Daniel Zeman, Maria Alexeeva, Tatiana Gavrilova, Nina Mustafina, and Elena Shakurova. 2016. Universal Dependencies for Russian: A new syntactic dependencies tagset. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2859998

Ma, Xuezhe and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074. https://doi.org/10.18653/v1/P16-1101

Malaviya, Chaitanya, Shijie Wu, and Ryan Cotterell. 2019. A simple joint model for improved contextual neural lemmatization. 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)*, pages 1517–1528. https://doi.org/10.18653/v1/N19-1155

Manning, Christopher D. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing*, pages 171–189. https://doi.org/10.1007/978-3-642-19400-9_14

Manning, Christopher D., Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. In *Proceedings of the National Academy of Sciences*, 117:30046–30054. https://doi.org/10.1073/pnas.1907367117, PubMed: 32493748

McCarthy, Arya D., Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931.

McCarthy, Arya D., Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244. https://doi.org/10.18653/v1/W19-4226

Müller, Thomas, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274. https://doi.org/10.18653/v1/D15-1272

Nivre, Joakim, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*.

Oflazer, Kemal. 1993. Two-level description of Turkish morphology. In *Sixth Conference of the European Chapter of the Association for Computational Linguistics*. https://doi.org/10.3115/976744.976810

Ortiz Suárez, Pedro Javier, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019*, pages 9–16.

Segalovich, Ilya. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a Web search engine. In *MLMTA*, page 273.

Silveira, Natalia, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer,

and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904.

Straka, Milan. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207.

Straka, Milan, Jana Straková, and Jan Hajic. 2019. UDPipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morphological categories, corpora merging. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 95–103. https://doi.org/10.18653/v1/W19-4212

Stroppa, Nicolas and François Yvon. 2005. An analogical learner for morphological analysis. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 120–127. https://doi.org/10.3115/1706543.1706565

Sulubacak, Umut, Memduh Gokirmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal Dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454.

Sylak-Glassman, John. 2016. The composition and use of the universal morphological feature schema (UniMorph schema). Technical report.

Taulé, Mariona, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218.

Tiedemann, Jörg. 2016. OPUS – parallel corpora for everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, *Baltic Journal of Modern Computing*.

Türk, Utku, Furkan Atmaca, Şaziye Betül Özateş, Abdullatif Köksal, Balkiz Ozturk Basaran, Tunga Gungor, and Arzucan Özgür. 2019. Turkish treebanking: Unifying and constructing efforts. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 166–177. https://doi.org/10.18653/v1/W19-4019

van den Bosch, Antal and Walter Daelemans. 1999. Memory-based morphological analysis. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 285–292. https://doi.org/10.3115/1034678.1034726

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Wang, Zeqiang, Yile Wang, Jiageng Wu, Zhiyang Teng, and Jie Yang. 2022. YATO: Yet another deep learning based text analysis open toolkit. *arXiv preprint arXiv:2209.13877*. https://doi.org/10.18653/v1/2023.emnlp-demo.11

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Wu, Shijie and Ryan Cotterell. 2019. Exact hard monotonic attention for character-level transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537. https://doi.org/10.18653/v1/P19-1148

Yildiz, Eray and A. Cüneyd Tantuğ. 2019. Morpheus: A neural network for jointly learning contextual lemmatization and morphological tagging. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 25–34. https://doi.org/10.18653/v1/W19-4205

Zeldes, Amir. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612. https://doi.org/10.1007/s10579-016-9343-x

Zeman, Daniel, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan

Hajič Jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19. https://doi.org/10.18653/v1/K17-3001