POS Tagging for the Endangered Dagur Language

Joanna Dolińska, Delphine Bernhard

▶ **To cite this version:**

Joanna Dolińska, Delphine Bernhard. POS Tagging for the Endangered Dagur Language. LREC-COLING 2024, May 2024, Torino, Italy. pp.12906-12916. hal-04597542

HAL Id: hal-04597542
https://hal.science/hal-04597542

Submitted on 7 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

POS Tagging for the Endangered Dagur Language

Joanna Dolińska[1], Delphine Bernhard[2]
[1] University of Warsaw, ul. Krakowskie Przedmieście 26/28, 00-927 Warsaw, Poland
[2] Université de Strasbourg, LiLPa UR 1339, F-67000 Strasbourg, France
j.dolinska@al.uw.edu.pl, dbernhard@unistra.fr

Abstract

The application of natural language processing tools opens new ways for the documentation and revitalization of under-resourced languages. In this article we aim to investigate the feasibility of automatic part-of-speech (POS) tagging for Dagur, which is an endangered Mongolic language spoken mainly in northeast China, with no official written standard for all Dagur dialects. We present a new manually annotated corpus for Dagur, which includes about 1,200 tokens, and detail the decisions made during the annotation process. This corpus is used to test transfer of models from other languages, especially from Buryat, which is currently the only Mongolic language included in the Universal Dependencies corpora. We applied the models trained by de Vries et al. (2022) to the Dagur corpus and continued training these models on Buryat. We analyse the results with respect to language families, script and POS distribution, in three different zero-shot settings: (1) unrelated, (2) related and (3) unrelated+related language.

Keywords: POS tagging, multilingual models, Dagur, Mongolic, zero-shot

1. Introduction

Until recently, the main focus of Natural Language Processing (NLP) tools has been mostly laid on the so-called "dominant" languages (English, French, Russian, Arabic, Chinese, Spanish, German). This rapid development of language technology has not been inclusive in terms of language equality and it mostly ignored the so-called extremely-under-resourced languages. Nevertheless, over the last decades there has been a growing interest in the perception of language loss as a part of the global extinction crisis (Gorenflo et al., 2012; Nelson et al., 2023). It has been namely noticed in many regions of the world that the shrinking multilingualism more often than not is accompanied by the decreased biodiversity. This notion represents language endangerment from a new perspective that puts, in fact, an obligation on the worldwide community to take care of our disappearing languages just like we try to care about the disappearing species of flora and fauna. NLP tools can be applied not only in the documentation, but also in revitalization of the lesser-resourced languages, for example in documenting the state-of-the-art condition of endangered languages, helping multilingual students from minorities communities catch-up with their peers at school representing the official language of a given country, help persons with disabilities communicate with their communities, create educational opportunities for young generations willing to return to the linguistic roots of their communities or facilitate the work of local doctors, social sector employees and legal authorities when working with language minority speakers.

In this paper, we focus on Dagur, an endangered Mongolic language spoken in the Northeast China. It does not have one common, official written standard and, to our knowledge, there is no part-of-speech (POS) annotated corpus for Dagur yet. In this article, we present a new manually annotated corpus for Dagur, which includes about 1,200 tokens, and describe the decisions made during the annotation process. We also investigate the feasibility of automatic POS tagging for the Dagur language, given the small size of the annotation dataset and the lack of one common, official written standard for all dialects of this language. In particular, we evaluate transfer learning from other languages, including Buryat (Badmaeva and Tyers, 2017) (Elena Badmaeva and Francis Tyers, 2023), which, for the time being, is the only Mongolic language included in the Universal Dependencies (UD) corpora (De Marneffe et al., 2021).[1]

Our contributions are as follows:

- We present the first small-size experimental corpus in Dagur manually annotated with Universal POS tags following the POS annotation for the Buryat language presented on the UD Homepage (Elena Badmaeva and Francis Tyers, 2023). The corpus has been released in the University of Warsaw Research Data Repository (Joanna Dolińska and Delphine Bernhard, 2024).
- We contribute to the language documentation of the Dagur language through the digitization of excerpts from Dagur language tales.
- We describe automatic POS tagging for the Dagur language using *zero-shot classification*:

[1] Nevertheless, it can be noticed that there have been steps taken to add Khalkha Mongolian (official language spoken in Mongolia) treebanks to the UD corpora.

LREC-COLING 2024, pages 12906–12916
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

a multilingual language model is fine-tuned for the POS tagging task with annotated corpora for languages other than Dagur and then it is used to tag the Dagur corpus.

- We investigate the linguistic factors which may account for the results obtained in our experiments, while taking into account essentially such parameters as script, the agglutinative morphology system of Dagur and its historical affiliation to the group of languages presently called "Transeurasian" (Robbeets and Savelyev, 2020) and historically linked to the term "Altaic" (Poppe, 1965).

2. Description of the Dagur Language

The Dagur language is the easternmost member of the Mongolic language family. As late as in 1930 it was considered to be "almost completely unexplored" (Poppe, 1930). The first accounts about the Dagur communities come from the 17th century and mention a sedentary, farming community inhabiting the upper Amur river region alongside Tungusic and other Mongolic communities (Todaeva, 1986). Due to a high number of Tungusic words in the Dagur language, the academic debate in the first half of the 20th century was focused on the question whether the Dagur language belongs to the Mongolic or Tungusic language family (Nugteren, 2020; Poppe, 1930; Todaeva, 1986). Dagur communities originally inhabited the region of upper Amur (Todaeva, 1986) and middle Amur river (Tsumagari, 2005), from where they moved at least in the 17th century to the region in the present day Northeast China (Tsumagari, 2005). Today, the Dagur language is spoken primarily in the Heihe region of the Middle Amur basin, in the locations within the Nonni river basin, in the Ewenki Autonomous Banner of Hulun Buir League and in the Xinjiang province in China with a total number of speakers of approximately 130,000 (Yamada, 2020). There are four main dialects of the Dagur language: Butha, Qiqihar, Hailar and Xinjiang, while the Butha Dagur is usually considered to be the standard dialect of the Dagur language and it served as the basis for the development of a standard writing system for Dagur in the Latin script in the 1960's (Yamada, 2020). However, there have been other attempts to standardize the Dagur literary language in the past as well - in the late Qing dynasty with the help of the Manchu script, in the Latin script in the 1930's and also in Cyrillic script in the 1950's (Tsumagari, 2005). Nowadays, Dagur speakers use either Manchu script or Chinese for writing. There are textbooks written in Dagur for the elementary schools in the Morii Dawaa district in Inner Mongolia, but written works in Dagur language are not common for the whole Dagur speaking community in China (Yamada, 2020).

3. Related Work

Mongolic languages, which encompass Dagur and several other languages, have been the subject of research in the NLP field since the late first decade of the 21st century. One of the main issues has been the lack of sufficient amounts of digitized Mongolic texts (Gao et al., 2008).The research on the Mongolic languages has been focused mainly on the dominant Mongolic languages: Khalkha Mongolian spoken in Mongolia and the Mongolian language spoken in the region of Inner Mongolia, China (Wei and Gao, 2014; Hansakunbuntheung et al., 2011). An exception here is a publication by Rinchinov (2019) on a corpus of Buryat language. Nevertheless, given the lack of research literature on this topic, it can be asserted that no attempts have been taken so far to apply NLP tools to the Dagur language.

Concerning POS tagging for extremely-low-resource languages, Lauscher et al. (2020) showed that transfer performance for POS tagging is primarily affected by the similarity in syntactic properties between source and target language. This analysis was confirmed by de Vries et al. (2022) who investigated zero-shot cross-lingual transfer learning with multilingual pre-trained models for the task of POS tagging. XLM-RoBERTa (Conneau et al., 2020) is used as the multilingual pre-trained model. They used 65 source languages for training and 105 target languages for testing. Among the 65 source languages, none belongs to the Mongolic language family. The only Mongolic language is Buryat, which is part of the test languages. For this language, the accuracy scores for the 10 best source languages ranged from 63.09 for Icelandic to 66.62 for Basque. This is higher than the overall mean of 57.4, but still rather low for the task of POS tagging. de Vries et al. (2022) show that the inclusion of the target language –and, to a lesser degree, the source language– in the training dataset for the multilingual pre-trained model is of particular importance. Being part of the same language family also has an effect on the accuracy, as well as sharing the writing systems.

Wu and Dredze (2020) have analysed another pre-trained multilingual language model, multilingual BERT (mBERT), on several tasks, including POS tagging. While their general conclusion was that mBERT performs worse on low resource languages, this particularly holds for more complex tasks such as NER and the differences are not as large for POS tagging.

Blum (2022) presents zero-shot experiments for languages from the low resource Tupían family.

The results show that the proximity of languages is a strong predictor of performance and that combining several related languages can also be useful.

4. Description of the Dagur Corpus

4.1. Data Statement

In this section, we describe the Dagur corpus we collected, following the professional practice called "data statements" developed by Bender and Friedman (2018) aiming at delivering more ethically responsive NLP tools which help the authors avoid primarily exclusion, overgeneralization, and underexposure of given language communities.

Curator rationale: The goal of selecting excerpts of tales collected by B. Kh. Todaeva in the Dagur language was based on the availability of these texts in the work *Dagursjik jazyk* from 1986, which includes a description of Dagur phonology, grammar and syntax, alongside 19 examples of Dagur literature (folktales, riddles and proverbs) and a Dagur-Russian glossary. The experimental corpus of 1,200 tokens has been based on the excerpts of six tales. They constitute heritage (Blokland et al., 2019) Dagur data that the authors digitized and processed for the purpose of creating an experimental corpus. The tales have a relatively concise structure and are accompanied by a word to word translation in Russian. Even though the availability of Russian translation facilitated the process of annotation, the knowledge of Dagur and other Mongolic languages was indispensable to carry out this task.

Language variety: The represented variety is Butha (Buteha) Dagur language from the Inner Mongolia Autonomous Region, China, used in oral literature spoken in 1980's in Northeast China. This language variety was surely represented by bilingual speakers of Dagur and Mandarin, as there has been a strong tendency in Northeast China for the Mandarin language to grow as a dominant language.

Speaker demographic: The speakers sharing orally the Dagur texts with B. Kh. Todaeva were well-versed in the Butha Dagur language, representing most likely an older generation. Unfortunately, it is not known from Todaeva's work exactly how many Dagur speakers shared their oral literature with her. Several represented excerpts mention the stories of elderly women and men, which might also point to the fact that they shared tales concerning the topics that are close to their everyday lives. The tales always represent some moral and hence, serve as a tool for knowledge exchange among the Dagur language speakers. Therefore, the speaker demographic is most likely as follows. *Age:* elderly. *Gender:* both male and female. *Race/ethnicity:* Mongolic. *Native language:* Dagur. *Socioeconomic status:* agricultural society from the borderlands of China. *Number of different speakers represented:* it is not clear how many Dagur speakers contributed to the collection of tales presented by (Todaeva, 1986).

Annotator demographic: Due to its experimental value, the corpus has been annotated by one researcher with the expertise in Mongolic languages and computational linguistics. The knowledge of Dagur dictionaries supported the annotator in the annotation process, while the Russian translation of the annotated texts played an auxiliary role in the work process. *Age:* middle-aged. *Gender:* female. *Race/ethnicity:* European. *Native language:* Polish. *Socioeconomic status:* urban, obtained tertiary education in the European Union. *Training in linguistics/other relevant discipline:* PhD in linguistics, trained in computational linguistics.

Speech situation: *Time and place:* B. Kh. Todaeva collected the tales from native Dagur speakers in 1980's in Northeast China. *Modality (spoken/signed, written):* It is most probable that the texts represented oral modality. *Scripted/edited vs. spontaneous:* The texts have been edited. *Synchronous vs. asynchronous interaction:* We do not possess relevant data to assess it. *Intended audience:* Non-Dagur speakers versed in Russian language who are primarily interested in Dagur folklore and language.

Text characteristics: The excerpts represent oral traditional tales. The vocabulary encompasses generic terms referring to gender and age, nature, distances and physical conditions of the main characters. The language is vivid and abounds in exclamations and rhetorical means that keep the readers in suspense.

Recording quality: N/A

Other: N/A

Provenance appendix: N/A

4.2. Digitisation and Manual Annotation

The annotation encompassed six excerpts of previously unpublished Butha Dagur tales collected and compiled by Buljaš X. Todaeva (Todaeva, 1986). Each excerpt contained 50 to 70 percent of the whole content of a tale. The annotated excerpts referred to the following tales: "Old woman"

(Dagur: Этээгу, Russian: Старуха, 126 tokens), “Bad friend” (Dagur: Моо гучу, Russian: Плохой друг, 72 tokens), “Stupid wolf” (Dagur: Лулсэн гускээ, Russian: Глупый волк, 288 tokens), “Little boy” (Dagur: Учээкэн Кэку, Russian: Маленький мальчик, 63 tokens), “Foolish people” (Dagur: Мэдэл увэі хуу, Russian: Неразумные люди, 212 tokens), “Old man and a lion” (Dagur: Сардии утаачи болоор арсалан, Russian: Старик и лев, 459 tokens), which summed up altogether to 1,220 tokens. The choice of this particular source was motivated by several factors. First of all, it is written down in Cyrillic script, which coincided with the availability of the Buryat language in Universal Dependencies. Buryat is so far the only Mongolic language represented in Universal Dependencies. It is written in Cyrillic script and spoken mainly in the region of the Buryat Republic, the Aga National District of Chita Province, the Ust'-Orda National District of Irkutsk Province (Russian Federation), in northern and eastern Mongolia and in Inner Mongolia (China) (Skribnik, 2005). Since the authors' aim was to check whether the choice of the same script plays any role in the comparison of typologically similar languages, the annotated corpus needed to represent the same script as the one compared to in Universal Dependencies. The copy of (Todaeva, 1986) was available to the annotator as a scanned version and the excerpts were OCR-ed with the help of the open-source gImageReader software[2] where Cyrillic and Mongolic scripts were chosen. The results of the OCR process were not entirely satisfactory and the OCR-ed version required manual adjustments in ca. 25% of the tales. B. Kh. Todaeva's choice of transliteration differed from the transliteration proposed, for example, by Tsybenov and Tumurdei (2014) in one of the most recent Dagur dictionaries. Characters that had to be permanently changed by the annotator included primarily the long vowels spelled by B. Kh. Todaeva as “ō” (transliterated by the annotator as “oo”), “ā” (transliterated by the annotator as “aa”) etc. Furthermore, the original spelling proposed by Todaeva has been preserved in reference to the following letters: “i” and “и”. The letter “i” is present in the diphthongs, whereas “и” occurs in the following syllable pattern: consonant-vowel-consonant (CVC), at the beginning of a word when followed by a consonant and at the end of a word when preceded by a consonant. Letter “ӣ”, representing a long vowel, has been written down by the annotator as “ии”. Excerpts of some tales compiled by B. Kh. Todaeva included inconsistencies. For example, there seems to be an error in the story “Мэдэл увэі хуу” (Russian: Неразумные люди, English: Foolish people), where the word ‘brothers’ (“ака дэу”) was misspelled as “ага дэу”. However, in the syllabus at the end of the book Todayeva provided the correct spelling “ака дэу”. Furthermore, the word “маучан” (‘rifle’) in the same tale is not represented in the syllabus accompanying the book. However, this expression could be found in a dictionary from 2014 (Tsybenov and Tumurdei, 2014) under a slightly different spelling “мяучаан” (’rifle’), which resolved the difficulty concerning the translation and annotation of this term. Furthermore, Todaeva used two different spelling versions for the question word “what”: “jō” and “jū”, whereby the variant “jō” was dominant. These inconsistencies slowed down the process of annotation as they influenced the search for the meaning of particular words, yet they are understandable given that one official common written standard for all Dagur varieties is not present neither in Cyrillic, nor in Latin or Manchu scripts.

The Dagur experimental corpus has been annotated following the POS annotation rules for the Buryat language represented in the Universal Dependencies. The Buryat treebank has been annotated by native speakers and it includes sentences from grammar books, fiction books and news reports. Parts of speech that were unambiguous in the annotation process were: ADP, DET, INTJ, NOUN, NUM, PART, PUNCT and SYM. Since Dagur, similarly to other Mongolic languages, abounds in converbial forms that play various roles in the sentence and might have a different meaning depending on their location in the syntax, a decision had to be made whether they should be classified as VERB or ADJ. In order to keep the consistency, they were classified as VERB. In addition, verbs which mean ”be” or ”become” and that form compound forms with other parts of speech (and do not denote the physical existence in a particular place at a particular time) have been tagged as AUX, following the Buryat annotation pattern. Furthermore, CCONJ and SCONJ were not numerous in the presented corpus, which might result from the represented text genre and the fact that the tales are based on oral literature, where long, convoluted subordinated sentences are not that frequent. The numerously repeated phrase “элджи хэлсэн” (non-literary translation ‘thus saying’), has been annotated as PART (элджи ‘thus’) and VERB (хэлсэн ‘said’), even though “элджи” (‘thus’) is actually an imperfective converbum form. It was treated by the annotator as a particle that has originated from the imperfective converbial form and that has lost its converbial character by now.

5. POS Tagging Experiments

In this section, we describe the automatic POS tagging experiments performed on the Dagur corpus.

[2] https://github.com/manisandro/gImageReader

5.1. Methodology

We evaluate three different *zero-shot* settings for POS tagging. In all settings, the performance is evaluated on our manually annotated Dagur corpus, plus Buryat (Elena Badmaeva and Francis Tyers, 2023) for the unrelated zero-shot setting.

1. *Unrelated zero-shot*: we use each of the 65 models provided by (de Vries et al., 2022) and apply them directly to the Dagur corpus and the UD v. 2.12 Buryat corpus, for comparison and evaluation. None of these models has been trained on a Mongolic language.

2. *Related zero-shot*: we fine-tune the XLM-RoBERTa base model (Conneau et al., 2020) on the Buryat UD corpus v. 2.12 (Elena Badmaeva and Francis Tyers, 2023). Buryat is the only other Mongolic language in UD. It has not been used as training data for fine-tuning by de Vries et al. (2022) due to the small size of the dataset: the train set contains 19 sentences and 153 tokens, while the test set contains 908 sentences and 10,032 tokens. In the experiments, we have reversed the data and used the larger test dataset for training and the train dataset for validation. de Vries et al. (2022) have fine-tuned their models using 10K training samples (sentences) and oversampled languages with fewer than 10K training samples using multiple epochs. Their experiments show that accuracies start reaching a plateau with 2.5K training samples, and start decreasing with 10K samples, which they chose as a threshold. We trained the model for Buryat for 10 epochs, which is close to 10K training samples (9,080 samples). We report the average results on the Dagur corpus for 5 training runs.

3. *Unrelated+related zero-shot*: We continue fine-tuning on the Buryat UD corpus for the 10 models contributed by de Vries et al. (2022) which perform best on Dagur. As a consequence, these models are trained on a combination of two languages: a non-Mongolic and a Mongolic language. We have compared the validation results on Buryat using 1 to 3 epochs (908 to to 2,724 training samples), and chose to use 3 epochs, based on the analysis of the evolution of both the accuracy and the F1 score on the Buryat validation data. We report the average results on the Dagur corpus for 5 training runs.

All the experiments and analyses are performed using the following main Python libraries and tools: Hugging Face[3] for models,[4] the *Transformers v. 4.30.2* and *Datasets v. 2.13.0* libraries (Tunstall et al., 2022), *PyTorch v. 2.0.1*,[5] *pandas v. 2.0.1* (pandas development team, 2023), *scikit-learn v. 1.3.0* (Pedregosa et al., 2011), *matplotlib v. 3.7.2* (Hunter, 2007) and *seaborn v. 0.12.2* (Waskom, 2021).

5.2. Experimental Results

5.2.1. Unrelated Zero-shot

Figure 1 displays the accuracy which could be reached for the 15 (out of 65) train languages among the top ten performing languages for Buryat and Dagur.[6] The results for Dagur are consistently lower than those obtained for Buryat: for Buryat, an accuracy of 65.7 is reached for Basque as the training language, and for Dagur, the best accuracy is only 56.5, also with Basque as the training language. This low accuracy is not exceptional and is almost on par with the average for cross-lingual accuracy observed by de Vries et al. (2022) for all target languages (57.4).

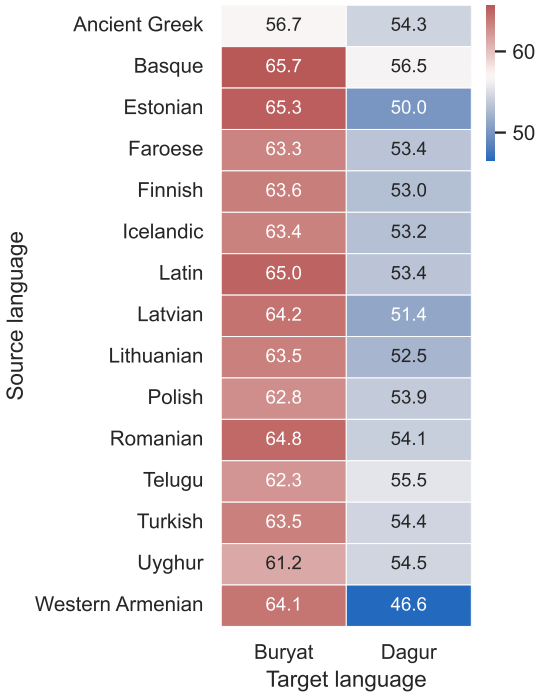

Figure 1: Accuracy for the 14 source languages among the top ten performing languages for Buryat and Dagur.

[3] https://huggingface.co

[4] https://huggingface.co/wietsedv/xlm-roberta-base-ft-udpos28-[languagecode] and https://huggingface.co/xlm-roberta-base

[5] https://pytorch.org/

[6] The accuracy values are slightly different from those reported for Buryat by (de Vries et al., 2022) because there were some changes in the UD corpus for Buryat between versions 2.8 and 2.12.

Figure 2 displays the joint values of the accuracy scores obtained for Buryat and Dagur while highlighting the language family. There is a clear relationship between both languages, which tend to have the highest and lowest accuracies with the same source training languages (correlation=0.85). However, language families tend to be scattered all over the graph, so that there is no clear tendency as to which language family yields the best results, except perhaps for both Turkic languages (Turkish and Uyghur) which perform well. Considering the language script (see Figure 3), the tendency is somewhat clearer, with languages written in cyrillic mostly found in the upper right hand quarter (Belarusian, Bulgarian, Old East Slavic, Russian and Ukrainian), except for Old Church Slavonic which yields much lower accuracies. But even though the Dagur corpus uses the Cyrillic script, training on languages with the same script does not necessarily lead to better results.

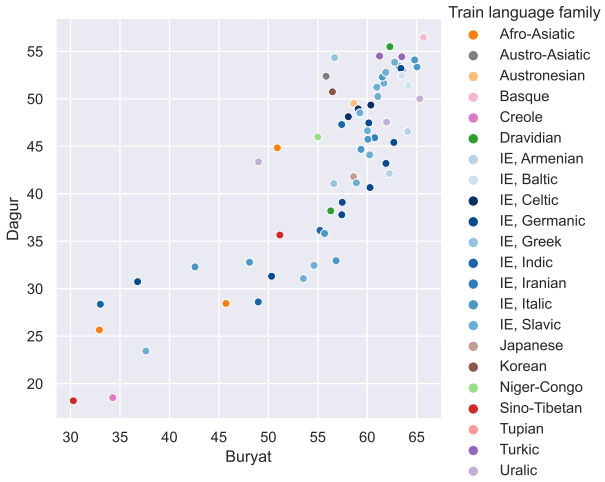

Figure 2: Scatter plot of the accuracy scores for Buryat and Dagur highlighting the language family of the source language used for training.

In order to better understand the linguistic characteristics which could account for better performance with some of the source languages, we have studied several statistics attached to individual POS tags in the Dagur corpus (see Table 1): their relative frequency, the average F1 score obtained over all source training languages, and the correlation between the accuracy and the frequency of the POS tag in the source training corpus. PUNCT (punctuation) obtains the highest average F1 score, well over the NOUN and VERB categories. This could be expected given that punctuation marks are not specific to a given language and can be learned rather efficiently from other, even distant, languages. It is also the third most frequent POS, after NOUN and VERB: accuracy is boosted by

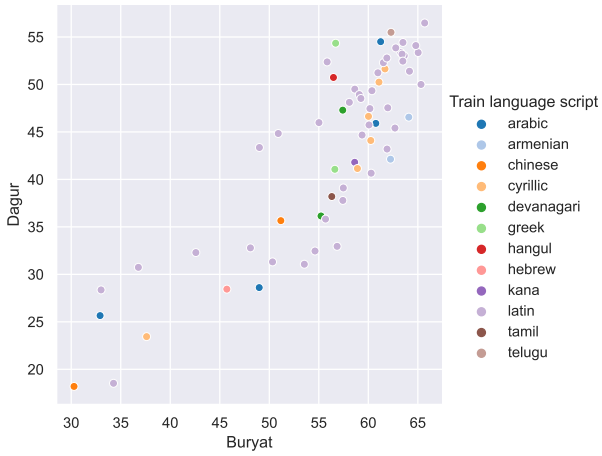

Figure 3: Scatter plot of the accuracy scores for Buryat and Dagur highlighting the script of the source language used for training.

the frequency of punctuation marks, which are also easy to label in a cross-lingual zero-shot setting. We also observe a positive correlation of .463 between the relative frequency of the PUNCT tag in the source training corpus and the accuracy: corpora with larger proportions of punctuation tend to produce more accurate models for Dagur. Overall, this shows that the models' capabilities are actually very limited for Dagur in a zero-shot cross-lingual setting.

POS	F1 score	Rel. freq.	Correlation
ADJ	0.056	0.045	0.013
ADP	0.216	0.009	-0.230
ADV	0.175	0.050	-0.081
AUX	0.051	0.024	-0.057
CCONJ	0.102	0.007	-0.027
DET	0.386	0.026	-0.158
INTJ	0.121	0.002	-0.314
NOUN	0.473	0.266	0.134
NUM	0.004	0.041	-0.015
PART	0.087	0.063	-0.151
PRON	0.104	0.055	-0.084
PUNCT	**0.901**	0.197	**0.463**
SCONJ	0.003	0.005	-0.158
VERB	0.438	0.210	-0.054

Table 1: Average F1 score for each POS and relative frequency in the Dagur corpus. The correlation is calculated between the accuracy for Dagur and the frequency of the POS tag in the source training corpus.

Finally, we use the KL_{cpos^3} measure to assess the similarity between languages, using the Kullback-Leibler divergence of the distributions of POS trigrams (Rosa and Žabokrtský, 2015). It has been proposed in the context of source treebank

selection for delexicalized parsing and used to measure the annotation consistency of different treebanks for the same language (Aggarwal and Zeman, 2020). Table 2 shows the KL_{cpos^3} values for the same languages as in Figure 1: the values which are closest to zero correspond to closer POS trigrams distributions. For Buryat, the closest language is Turkish, followed by Dagur. For Dagur, Buryat is the closest language, followed by Uyghur. Uyghur is also the third best source language for Dagur (see Figure 1). There is also a moderate negative correlation of -0.55 between KL_{cpos^3} and accuracy for Dagur, showing that KL_{cpos^3} might be used as an indicator to select the best source language to train on.

Language	Buryat	Dagur
Ancient Greek	1.89	2.34
Basque	0.96	1.88
Estonian	1.17	2.53
Faroese	2.64	3.12
Finnish	1.08	1.82
Icelandic	2.23	2.97
Latin	1.46	2.08
Latvian	1.09	1.74
Lithuanian	1.18	1.79
Polish	1.49	2.09
Romanian	2.17	2.79
Telugu	1.04	1.31
Turkish	**0.51**	1.71
Uyghur	0.82	**1.14**
Western Armenian	0.95	1.82
Buryat		0.98
Dagur	0.64	

Table 2: KL_{cpos^3} values.

5.2.2. Related Zero-shot

In this second setting, the accuracy for Dagur goes up to 60.11 (+4.6 from the best previous result obtained from fine-tuning on Basque), showing that training on Buryat, which is a language from the same family and the same script, yields improvements over training on unrelated languages. It translates in a improvement of the macro F1 score from 0.25 to 0.36 (averaged over 5 runs), showing that rarer POS tags are better taken into account by the model trained on Buryat: this latter model is therefore more balanced and versatile. These observations are also consistent with the proximity measured between Buryat and Dagur using KL_{cpos^3}.

5.2.3. Unrelated+related Zero-shot

In this final setting, the accuracy improves for all source languages (see Table 3). The overall performance benefits from further fine-tuning on Buryat, reaching an accuracy of 61.13 by combining Latin with Buryat. The results achieved are better than those obtained by training on Buryat only: 7 out of 10 combined models are slightly superior. This shows that a combination of languages, even if one of them is distant from the target language, may be somewhat beneficial in a zero-shot cross-lingual setting. Interestingly, one of the models which benefits less from further training on Buryat is Basque (which obtained the best results in the unrelated zero-shot setting).

Source lang.	base.	+Buryat	Δ	std
Ancient Greek	54.34	60.80	6.46	0.52
Basque	**56.48**	59.77	3.30	0.57
Faroese	53.36	59.57	6.21	1.02
Icelandic	53.20	60.16	6.97	0.78
Latin	53.36	**61.13**	**7.77**	0.79
Polish	53.85	60.67	6.82	0.42
Romanian	54.10	60.39	6.30	0.78
Telugu	55.49	58.05	2.56	0.76
Turkish	54.43	60.69	6.26	0.65
Uyghur	54.51	60.34	5.84	0.65
Buryat only	60.11			0.74

Table 3: Accuracy before (base.) and after (+Buryat) fine-tuning on the Buryat UD corpus for the 10 models which perform best on Dagur. Δ corresponds to the increase over the base. column. Results for the fine-tuned models are averaged over 5 training runs and the standard deviation is reported (std).

Figure 4 compares the results for each POS tag and shows that most of them benefit from fine-tuning on Buryat. This is particularly evident for the AUX, CCONJ, DET, PART and PRON tags. These tags are also categorised with higher performance when training on Buryat than on the other source languages (compare the red dashed line with the blue boxes). Moreover, the interquartile ranges are usually smaller, meaning there is less dispersion in the F1 scores, and thus less variability across the models: the differences between all models are evened out after continued training on Buryat.

6. Discussion and Conclusions

The highest accuracy in the unrelated zero-shot setting for the Dagur language was obtained with Basque (56.5%). This low accuracy is close to the average close-lingual accuracy reported by de Vries et al. (2022) for the entirety of target

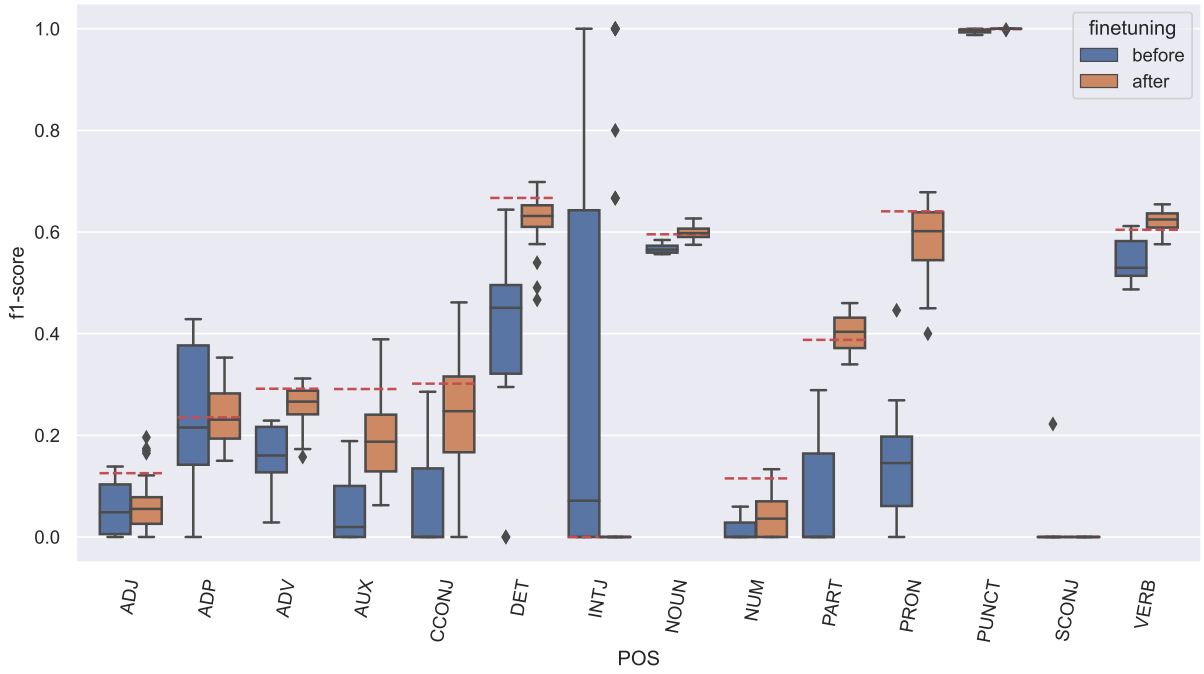

Figure 4: Boxplot of the F1 score for each POS before and after further fine-tuning on Buryat. The dashed red line marks the F1 obtained by training on Buryat only.

languages. Among some of the best performing source languages are Uyghur (54.5%) and Turkish (54.4%). This can be explained by at least two factors: a) They are agglutinative languages just like Dagur. Furthermore, they belong to the so-called Transeurasian (alternatively, Altaic) group of languages that bear resemblance in terms of morphology, syntax, phonology and semantics. Even though Uyghur and Turkish are Turkic languages, Mongolic and Turkic families of languages have been historically perceived as the core of the Altaic language group due to their numerous similarities. b) Dagur has retained some archaic Mongolic features (Todaeva, 1986; Sansheev, 1953). Therefore, it is similar to the oldest variety of the Mongolic language that has been attested so far (while putting the Para-Mongolic languages aside). This most archaic (known) version of the Mongolian language was spoken around the time of dispersal of Mongolic tribes under the command of Chinggis Khan in the first half of the 13th century and it is called Middle Mongolian. If Dagur contains many archaic elements of the Middle Mongolian language and Mongolic languages were once related to the Turkic languages, then it is likely that Uyghur turns out to be one of the closest languages in the unrelated zero-shot setting. The discussion on the potential genetic affinity between the Turkic and Mongolic languages can be potentially supported by such experimental results.

In order to continue this line of argumentation, corpora with a higher number of tokens would need to be studied. Furthermore, we noticed that the script of the analysed corpora is not of high importance for the accuracy in the unrelated zero-shot setting. The relatively high scores in Figure 2 for the Turkic and Uralic language families with respect to Dagur and Buryat results from the fact that Uralic and Turkic languages have an agglutinative structure and in the past were even believed to form one big Uralo-Altaic language family (Klaproth, 1831) In addition, the KL_{cpos^3} measure only confirms what we know from historical comparative linguistics: that Buryat and Dagur languages are close to each other in terms of POS as they belong to one Mongolic language family and that Uyghur and Turkish, being Turkic languages, bear the closest resemblance among the source languages to Buryat and Dagur. Furthermore, in contrast to the unrelated zero-shot setting, the related zero-shot approach showed that using the same family and script contributes to a better performance. Finally, the unrelated + related zero-shot setting shows a surprising result: the Uyghur language model benefited less from further training on Buryat than some other languages, including Latin. This is possibly because training on two close languages leads to less diversity on the training data and hence fewer generalisation capabilities to a new language.

While the first, experimental corpus described in this article includes around 1,200 tokens, representing only the Butha (Buteha) Dagur language

from Inner Mongolia (China), its expanded version includes 4,502 tokens and it encompasses examples of different Dagur heritage data (Martin, 1961). The expanded corpus illustrates also the Dagur language variety represented by the Dagur speaker Peter (Uregungee) Onon from the region of the Nonni river in Inner Mongolia (China). Expanding the size of the Dagur experimental corpus will allow the authors in the future to carry out further experiments in relation to multilingual models. The expanded Dagur experimental corpus with 4,502 tokens has been stored in the University of Warsaw Research Data Repository (Joanna Dolińska and Delphine Bernhard, 2024)

7. Acknowledgements

The authors would like to acknowledge the High Performance Computing Center of the University of Strasbourg for supporting this work by providing scientific support and access to computing resources. Part of the computing resources were funded by the Equipex Equip@Meso project (Programme Investissements d'Avenir) and the CPER Alsacalcul/Big Data. In addition, the Authors acknowledge the National Science Centre in Poland for awarding the Miniatura-6 grant 2022/06/X/HS2/01374 to Joanna Dolińska which made it possible for the Authors to commence the cooperation on the research presented in this article.

8. Limitations

The annotation has been carried out by only one person. Therefore, there is no inter-annotator agreement study. Furthermore, the size of the manually annotated corpus is rather small. In addition, due to the lack of one common written standard for all varieties of Dagur, B. Kh. Todaeva's choice of transliteration was arbitrary and therefore differed from the contemporary Latin script transcriptions. Last but not least, the corpus is limited to Butha Dagur oral literature, which is not representative of all the Dagur language varieties.

9. Ethics Statement

This article does not infringe upon the intellectual and property rights of the "Nauka" Publishing House, where the compiled and edited by B. Kh. Todaeva Dagur tales appeared in print in 1986. Only excerpts of the above-mentioned tales have been digitized, annotated and processed.

The annotated corpus does not contain sensitive, confidential, inappropriate or offensive information.

10. Bibliographical References

Akshay Aggarwal and Daniel Zeman. 2020. Estimating POS annotation consistency of different treebanks in a language. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 93–110.

Elena Badmaeva and Francis M. Tyers. 2017. Dependency Treebank for Buryat. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 1–12.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Rogier Blokland, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2019. Using computational approaches to integrate endangered language legacy data into documentation corpora: Past experiences and challenges ahead. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 2, pages 24–30.

Frederic Blum. 2022. Evaluating Zero-Shot Transfers and Multilingual Models for Dependency Parsing and POS Tagging within the Low-Resource Language Family Tupían. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 1–9.

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

Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the Best of Cross-lingual Transfer: Evidence from POS Tagging with over 100 Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Guanglai Gao, Wei Jin, Fei Long, and Hongxu Hou. 2008. A first investigation on mongolian information retrieval. In *EVIA@NTCIR*.

Larry James Gorenflo, Suzanne Romaine, Russell A. Mittermeier, and Kristen Walker-Painemilla. 2012. Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. *Proceedings of the National Academy of Sciences*, 109(21):8032–8037.

Chatchawarn Hansakunbuntheung, Ausdang Thangthai, Nattanun Thatphithakkul, and Altangerel Chagnaa. 2011. Mongolian speech corpus for text-to-speech development. In *Proceedings of the 2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, pages 130–135.

John D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Julius Klaproth. 1831. *Asia Polyglotta*. Verlag von Heideloff & Campe.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers. *arXiv preprint arXiv:2005.00633*.

Samuel E. Martin. 1961. *Dagur Mongolian Grammar, Texts, and Lexicon. Based on the Speech of Peter Onon*. Indiana University.

Diane Nelson, Nhenety Kariri-Xocó, Idiane Kariri-Xocó, and Thea Pitman. 2023. "We Most Certainly Do Have a Language": Decolonizing Discourses of Language Extinction. *Environmental Humanities*, 15(1):187–207.

Hans Nugteren. 2020. The classification of the mongolic languages. In Martine I. Robbeets and Alexander Savelyev, editors, *The Oxford Guide to the Transeurasian Languages*, pages 92–104. Oxford University Press.

pandas development team. 2023. pandas-dev/pandas: Pandas v2.0.1. 10.5281/zenodo.7857418.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nicholas Poppe. 1930. *Dagurskoe narechie [Dagur]*. Izdatel'stvo Akademii Nauk SSSR.

Nicholas Poppe. 1965. *Introduction to Altaic Linguistics*. Otto Harrassowitz.

Oleg Rinchinov. 2019. Structural markup of the mongolian-script buryat chronicles for the diachronic corpus of buryat language. *Culture of Central Asia: written sources [Культура Центральной Азии: письменные источники]*, 12:106–117.

Martine I. Robbeets and Alexander Savelyev. 2020. *The Oxford Guide to the Transeurasian Languages*. Oxford University Press.

Rudolf Rosa and Zdeněk Žabokrtskỳ. 2015. Klcpos3-a Language Similarity Measure for Delexicalized Parser Transfer. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 243–249.

Garma D. Sansheev. 1953. *Sravnitelnaja grammatika mongolskih jazykov. Tom 1*. Izdatel'stvo Akademii Nauk SSSR.

Elena Skribnik. 2005. Buryat. In Juha Janhunen, editor, *The Mongolic Languages*, pages 102–128. Routledge.

Buljaš X. Todaeva. 1986. *Dagurskij jazyk [Dagur]*. Nauka.

Toshiro Tsumagari. 2005. Dagur. In Juha Janhunen, editor, *The Mongolic Languages*, pages 129–153. Routledge.

Bazar D. Tsybenov and G. Tumurdei. 2014. *Kratkiij Dagursko-Russkij Slovar*. Russian Academy of Sciences.

Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. 2022. *Natural Language Processing with Transformers*. O'Reilly Media, Inc.

Michael L. Waskom. 2021. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

Hongxi Wei and Guanglai Gao. 2014. A keyword retrieval system for historical mongolian document images. *International Journal on Document Analysis and Recognition (IJDAR)*, 17(1):33–45.

Shijie Wu and Mark Dredze. 2020. Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Yonei Yamada. 2020. Dagur. In Martine I. Robbeets and Alexander Savelyev, editors, *The Oxford Guide to the Transeurasian Languages*, pages 321–333. Oxford University Press.

11. Language Resource References

Elena Badmaeva and Francis Tyers. 2023. *UD Buryat-BDT Treebank. Universal Dependencies v2.12*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, https://github.com/UniversalDependencies/UD_Buryat-BDT/tree/master.

Joanna Dolińska and Delphine Bernhard. 2024. *Dagur language corpus*. University of Warsaw Research Data Repository, University of Warsaw, https://doi.org/10.58132/C85E2F.