UNIVERSITY OF WARSAW

Université de Strasbourg

THE 2024 JOINT INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, LANGUAGE RESOURCES AND EVALUATION, 20-25 MAY 2024, TORINO, ITALY

POS Tagging for the Endangered Dagur Language

Joanna Dolińska[1] & Delphine Bernhard[2]

1 University of Warsaw, Faculty of "Artes Liberales", Poland

2 Université de Strasbourg, LiLPa UR 1339, France

Contents

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 minority languages.
- NLP tools can be applied both in the documentation and revitalization of low-resource languages.

Objectives

- 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 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).

Our contributions

- 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 (https://doi.org/10.58132/ C85E2F.).
- 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:** 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).

Dagur language

- Endangered easternmost **Mongolic** language spoken mainly in 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.
- As late as in **1930** it was considered to be "almost completely unexplored" (Poppe, 1930).
- 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).
- 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.
- **Butha Dagur** 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.

Data statement 1/2

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	• We digitized and processed heritage Dagur data for the purpose of creating an experimental corpus. The choice of the source (B. Kh. Todaeva's *Dagurskij jazyk* from 1986) was dictated by its digital availability, richness of vocabulary, enclosed Dagur-Russian glossary and the script in which the Dagur tales were represented (Cyrillic).
Language variety	• Butha (Buteha) Dagur language from the Inner Mongolia Autonomous Region, China, used in oral literature spoken in the 1980's in Northeast China.
Speaker demo-graphic	• Age: most likely elderly speakers. Gender: both male and female. Ethnicity: Mongolic. Native language: Dagur. Other languages: Chinese. Socioeconomic status: Agricultural society from the northeast borderlands of China.

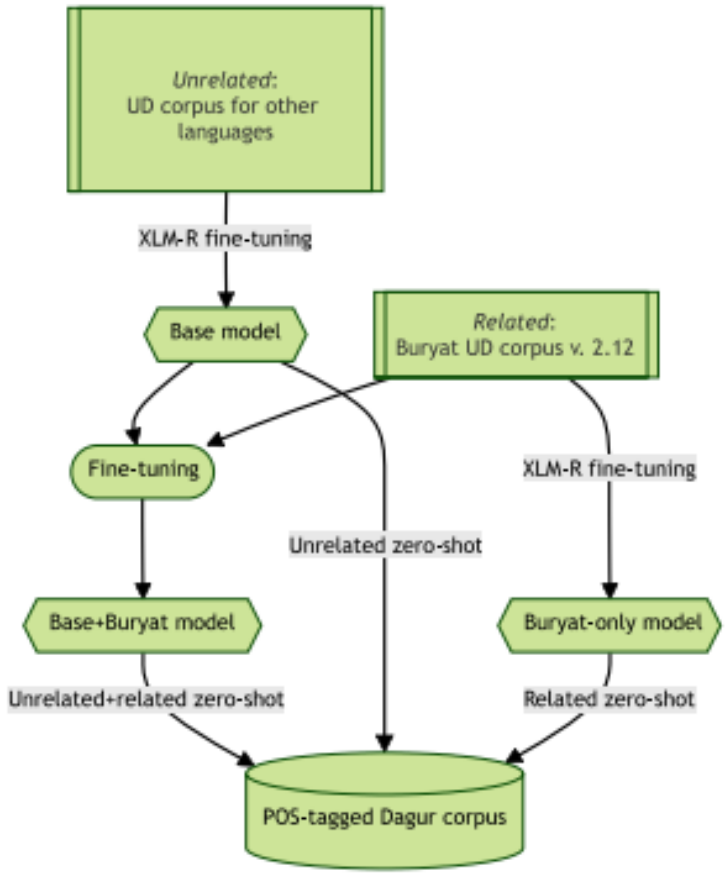
Data statement 2/2

Annotator demo-graphic	• One female researcher with expertise in Mongolic languages and experience in computational linguistics.
Speech situation	• Time and place: 1980's in Northeast China. Modality (spoken/signed, written): Most likely oral modality. Scripted/edited vs. spontaneous: The texts have been edited. Intended audience: Non-native Dagur speakers.
Text charac-teristics	• 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.

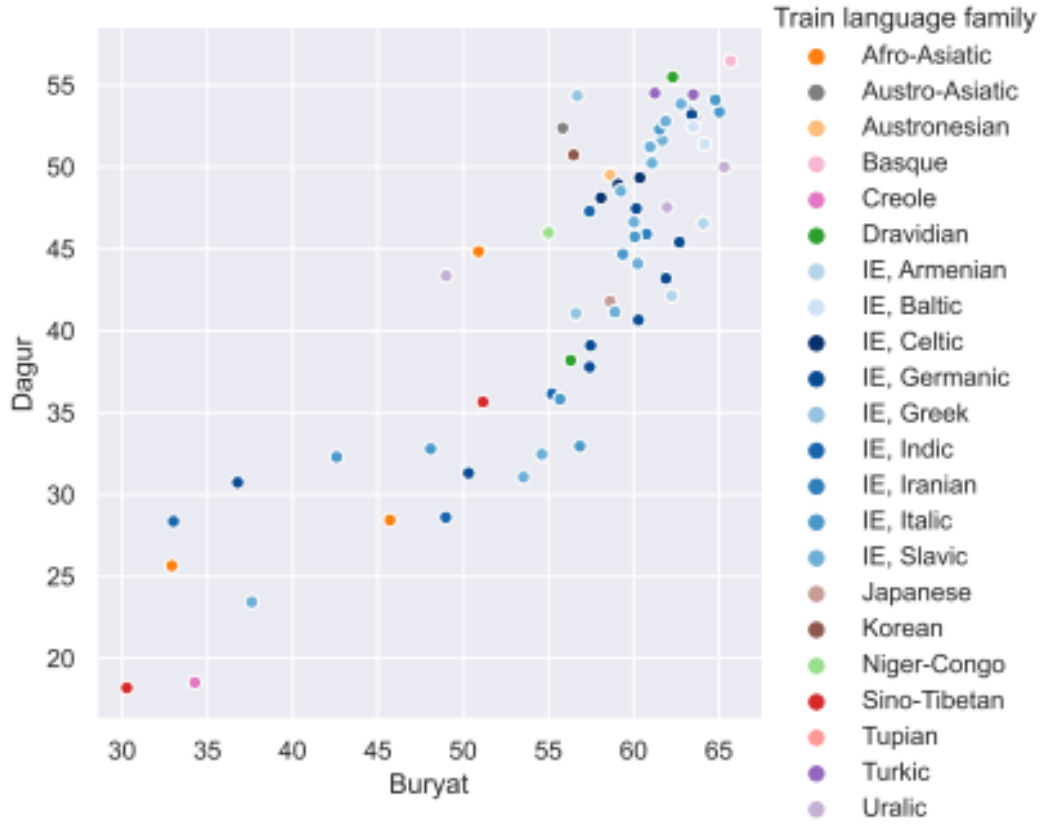
Methodology

- 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.
    - 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.**
    - ***Related zero-shot:*** we fine-tune the **XLMRoBERTa** 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 finetuning 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.
    - ***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.

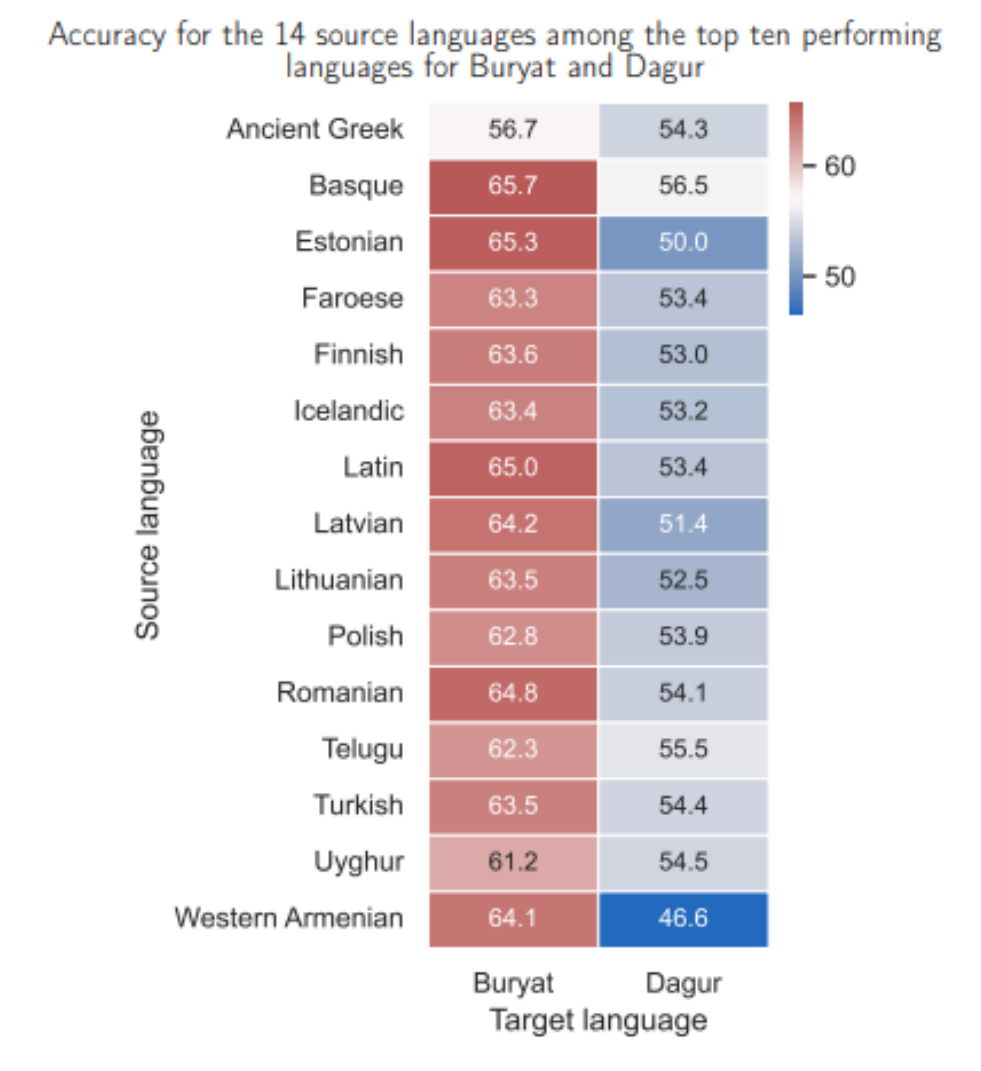
POS Tagging Experiments



Accuracy scores for Buryat and Dagur highlighting the language family of the source language used for training.



Results



Accuracy before (base.) and after (+Buryat) fine-tuning on the Buryat UD corpus for the 10 models which perform best on Dagur.

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

Conclusions

- Among the best performing source languages in the unrelated zero-shot setting are Uyghur (54.5) and Turkish (54.4), which are agglutinative languages just like Dagur. They belong to the Transeurasian (or Altaic) group of languages that bear resemblance in terms of morphology, syntax, phonology and semantics.
- The script of the source corpora does not necessarily lead to better results.
- The related zero-shot approach shows that training on Buryat (same family and script) contributes to a better performance.
- In the unrelated + related zero-shot setting, Uyghur benefited less from further training on Buryat than some other languages, including Latin. Training on two close languages possibly leads to less diversity on the training data and hence fewer generalisation capabilities to a new language.
- The presented corpus and its expanded version with 4,502 tokens can be found at https://doi.org/10.58132/C85E2F

Bibliography

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.

Frederic Blum. 2022. Evaluating zero-shot transfers and multilingual models for dependency parsing and POS tagging within the lowresource 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 WalkerPainemilla. 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.

Bibliography

Chatchawarn Hansakunbuntheung, Ausdang Thangthai, Nattanun Thatphithakkul, and Altangerel Chagnaa. 2011. Mongolian speech corpus for text-to-speech development. In *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.

Diane Nelson, Nhenety Kariri-Xocó, Idiane KaririXocó, 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. pandasdev/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.

Oleg Rinchinov. 2019. Structural markup of the mongolian-script buryat chronicles for the diachronic corpus of buryat language.

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.

Bibliography

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

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.

Thank you very much

for your attention!