Capacity building workshop

Donostia, 28th of May 2026

CREATING LINGUISTIC CORPORA FOR LANGUAGES WITH NO UNIFIED WRITTEN STANDARD

Assistant Professor at the Faculty of "Artes Liberales"

Center for Research and Practice in Cultural Continuity

University of Warsaw, j.dolinska@al.uw.edu.pl

UNIVERSITY OF WARSAW

WYDZIAŁ ARTES LIBERALES

Center for Research and Practice in Cultural Continuity

Funded by the European Union

FOSTERLANG

Short introduction

PhD in linguistics in 2019

Studies: University of Warsaw (Poland), University of Leipzig (Germany), University of Duisburg-Essen (Germany), National University of Mongolia (Mongolia)

Post-doctoral fellowship: Max Planck Institute for the Science of Human History (Germany)

Visiting research stays: Smithsonian Institution (USA), Mahidol University (Thailand), University of Cambridge (UK), University of Groningen (The Netherlands), University of Strasbourg (France)

Research interests: sociolinguistic situation of minority language communities, Asian linguistics, computational linguistics, contact linguistics

UNIVERSITY OF WARSAW

WYDZIAŁ ARTES LIBERALES

Center for Research and Practice in Cultural Continuity

Funded by the European Union

FOSTERLANG

Previous experience with natural language processing and automatic speech recognition for under-resourced languages

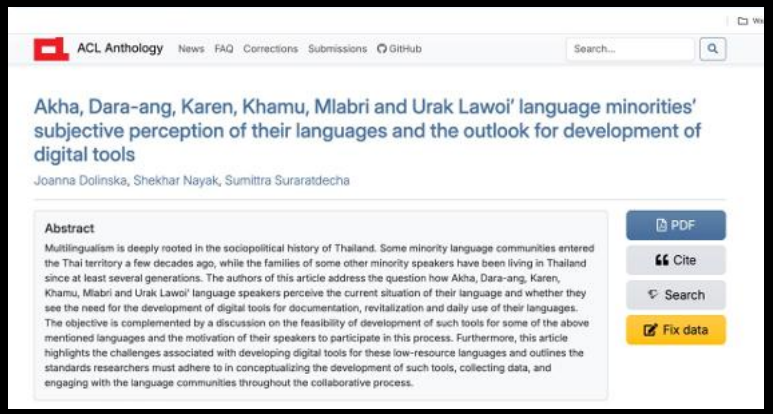
-6 years of experience in the roles of a computational linguist and product owner for automatic speech recognition (Samsung Electronics)

-Creating an experimental corpus of the Dagur language and conducting experiments with Prof. Delphine Bernhard from the University of Strasbourg

-Co-supervising a MA thesis devoted to text to speech synthesis for Manchu by Mr. Shenguan Ding, University of Groningen

-Surveying the opportunities for development of voice technologies for endangered and low-resource languages of Thailand

-Reviewing the available NLP tools for Mongolic languages



FOSTERLANG SURVEY – THE PART FOCUSED ON LANGUAGE TECHNOLOGIES FOR UNDER-RESOURCED LANGUAGES

UNIVERSITY OF WARSAW

Center for Research and Practice in Cultural Continuity

Funded by the European Union

FOSTERLANG

HORIZON EUROPE PROJECT FOSTERLANG

Survey dedicated to the research on:

- the presence of language technologies for minoritized languages in Europe,
- the collaboration with these communities on further development of such tools,
- potential opportunities and challenges in ditigal equality development.

Survey developed by:

- University of Warsaw (PL)
- University of the Highlands and Islands (UK)
- Dublin City University (IR)
- Max Planck Institute for Psycholinguistics (NL)

Language versions:

- **Basque**, **Spanish** and **French** language versions dedicated to Basque speaking communities in the Basque Country in Spain and in several regions in France
- **German** language version dedicated to Carinthia in Austria
- **Slovene** language version dedicated to Carinthia in Austria
- **English** language version dedicated to Scotland
- **Gaelic** language version dedicated to Scotland
- **English** language version dedicated to Ireland
- **Irish** language version dedicated to Ireland
- **Polish** language version dedicated to Poland

UNIVERSITY OF WARSAW

WYDZIAŁ ARTES LIBERALES

Center for Research and Practice in Cultural Continuity

Funded by the European Union

FOSTERLANG

FOSTERLANG PROJECT: SURVEY LAUNCHED IN POLAND, SCOTLAND, IRELAND AND AUSTRIA

SURVEY DEDICATED TO THE BASQUE SPEAKING COMMUNITY DELIVERED IN THE BASQUE, SPANISH AND FRENCH LANGUAGES - **COMING SOON!**☺

ODK-software coding:

- Joanna Dolińska
- Arvind Kumar

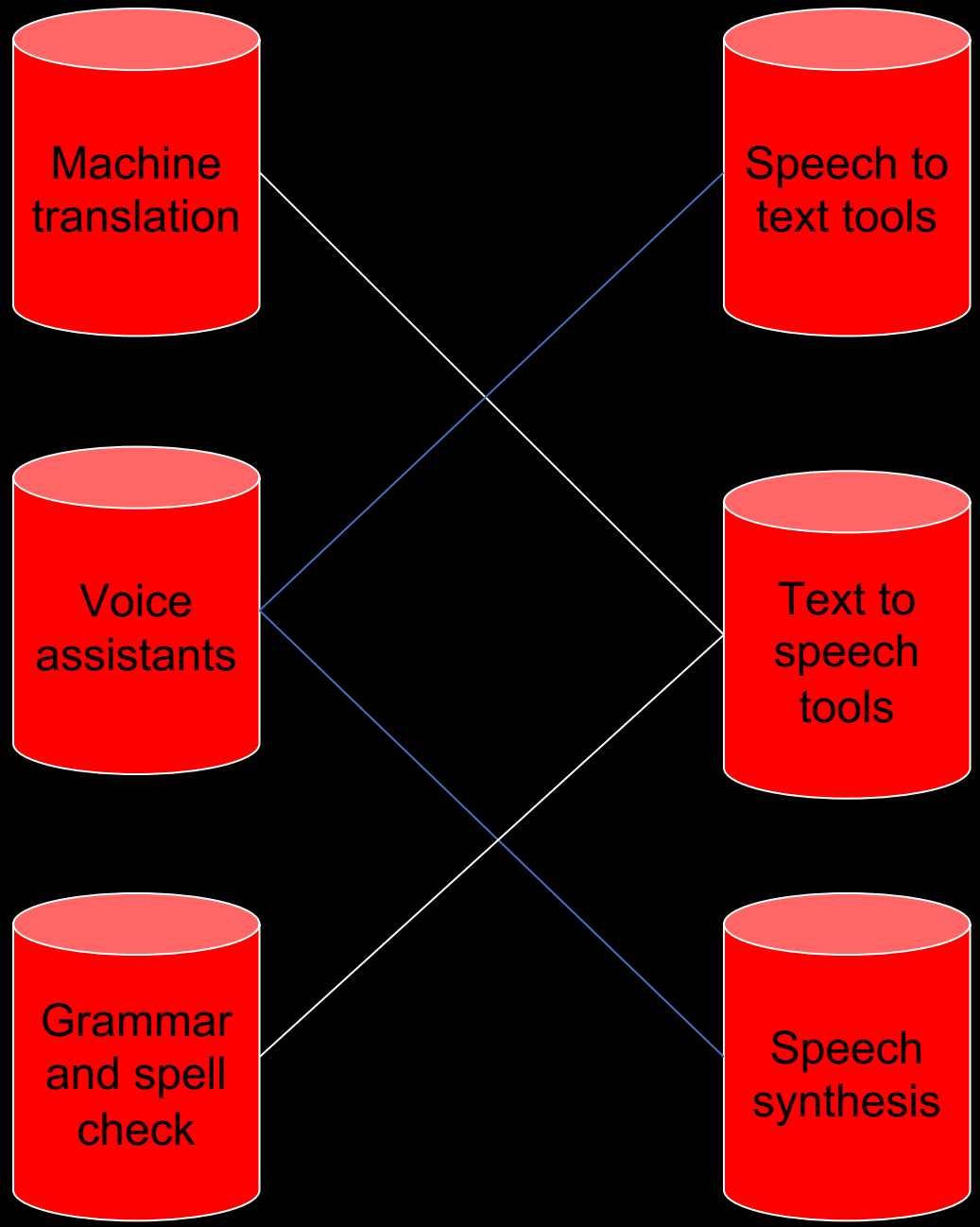
- **Students from the University of Warsaw:**
- Grzegorz Kaliński
- Krzysztof Dąbrowski
- Wojciech Soczyński

Joint collaboration with:

1. University of Teacher Education Carinthia
2. University of the Basque Country
3. Linguapax International
4. European Language Equality Network

LANGUAGE TECHNOLOGIES IN GENERAL

LANGUAGE TECHNOLOGIES EXAMPLES



UNIVERSITY OF WARSAW

WYDZIAŁ ARTES LIBERALES

Center for Research and Practice in Cultural Continuity

Funded by the European Union

FOSTERLANG

Before we think of developing digital tools

- Discussion concerning the nature of such tools and benefits to the community
- Worldview and data
- Competing orthographies
- Legacy/heritage data

UNIVERSITY OF WARSAW

WYDZIAŁ ARTES LIBERALES

Center for Research and Practice in Cultural Continuity

Funded by the European Union

FOSTERLANG

LANGUAGE DOCUMENTATION AND REVITALIZATION

Language documentation

- The method of researching (minority) languages and their use.
- Also called *documentary linguistics.*
- The beginning of this scientific approach to language lies in the 1990s.

Language documentation

- The aim of language documentation:
  - To create audio-visual samples of how language is used and performed.
  - To document everyday language use, narratives and ritualized activities.
  - To document speakers' own attitutde towards their language.
- The result: our knowledge of how a language is employed in various situations and contexts by various speakers.

Language documentation

- An organized collection of linguistic data is called *a corpus*.
- Why do we need corpora?
  - Mother-tongue education
  - The increase the social status of a language
  - Learning or re-learning the language, and thereby revitalizing it
- A copy of a corpus needs to be placed in an **archive**.
- The corpus needs to be accompanied by **metadata.**

Language revitalization

- The goal of revitalization is to **take actions** to revive the use of a particular language, broaden the circle of its users or protect it from extinction.

When data for langauge documentation is not suitable for language revitalization

„1) The records in the corpus may focus on **interesting or unusual linguistic features** rather than how conversations are organized in the particular community
(2) Conversations, narratives, and interviews may focus on the **past, looking back nostalgically to the 'good old days'** before social, cultural, and linguistic shifts began to take plac.
(3) The linguistic analyses created by language documenters, including transcriptions and grammatical annotations, may be **produced in orthographies or languages unknown** to the community and using specialized terminology which is not easily understandable to non-linguists;"

UNIVERSITY OF WARSAW
WYDZIAŁ ARTES LIBERALES
Center for Research and Practice in Cultural Continuity
Funded by the European Union
FOSTERLANG

Language corpora

- are needed for empirically sound descriptions of endangered languages.
- can be utilized in various ways in future computational linguistic studies on these languages.

Metadata in corpora

- Preceded by a hashtag “#”
- Inform about the genre of a given text
- Inform about the title of a given text
- Inform about the source of a given text
- Point to the author and the original date of the given text
- The more metadata the better!

UNIVERSITY OF WARSAW
Center for Research and Practice in Cultural Continuity
Funded by the European Union
FOSTERLANG

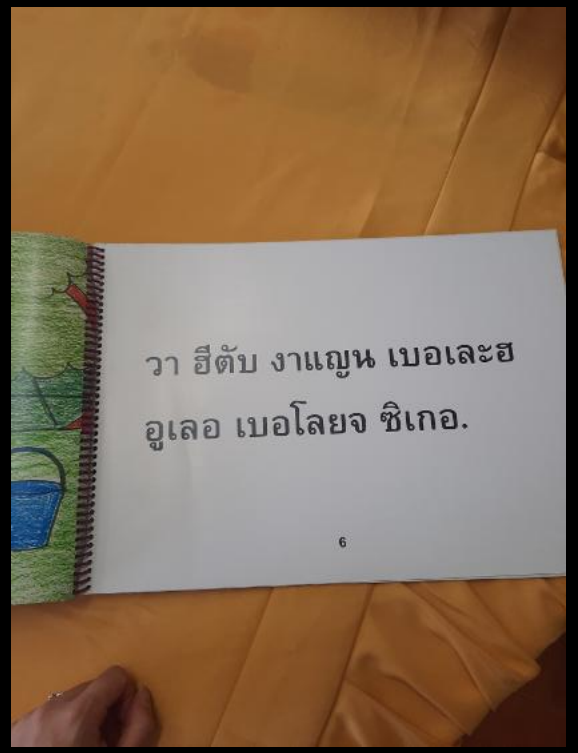
Data statements by Bender and Friedman (2018)

- It is recommended that the description of data follows the so-called “data statements” practice developed by Bender and Friedman (2018). The goal of this practice is to deliver digital tools that avoid the **risk of oversimplifying the situation of given speech communities, as well as their exclusion, underrepresentation or misrepresentation.**
  - Curator rationale
  - Language variety
  - Speaker Demographic
  - Annotator Demographic
  - Speech Situation
  - Text Characteristics
  - Recording Quality
  - Others
  - Provenance Appendix

Why languages have
versatile orthographies and scripts?

Could you share with us some examples?

URAK LAWOI’ COMMUNITY IN KRABI PROVINCE, THAILAND



UNIVERSITY OF WARSAW

WYDZIAŁ ARTES LIBERALES

Center for Research and Practice in Cultural Continuity

Funded by the European Union

FOSTERLANG

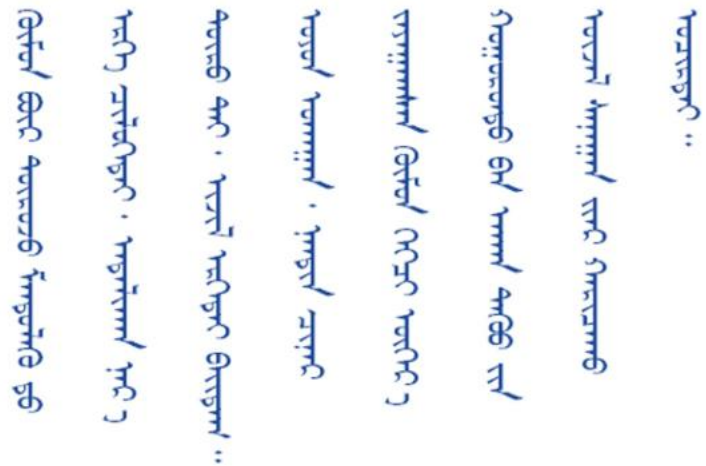
CONTEMPORARY MONGOLIAN LANGUAGE

Монгол Википедиа

Чөлөөт нэвтэрхий толь — Википедиагаас

2004 оны 2 сарын 28-наас **монгол хэлний** кирилл бичгээрх **Википедиа** ажиллаж эхэлсэн.

Өнөөдрийн байдлаар үүнд 17,986 бичлэг байгаа нь нийт 287 хэлний Википедиагийн дунд 116-р их тоот болж явна. 2011 оны 1 сарын 8-ны байдлаар 11,183 бүртгэгдсэн гишүүнтэй байсан бол 2013 оны 11 сарын 11-ны байдлаар 25,873 бүртгэгдсэн гишүүнтэй болсон байна. Эдгээрээс 9 нь зохицуулагч. Сүүлийн 1 сард хувь нэмэр оруулсан (идэвхтэй) гишүүний тоо 90 байна.



Can we have different orthographies of one language in a corpus?

- **Yes**, even though this variability is sometimes called *noise* in computational linguistics.
- EUD **Low Saxon** LSDC treebank (Siewert et al., 2021) is presented both in the original ad-hoc pronunciation spelling and in a recently proposed orthography for Low Saxon, Nysassiske Skryvwyse.
- UD Alemannic DIVITAL
- „There is no widely used written standard for Alsatian. Various spelling systems have been developed and proposed, to make it possible for all speakers to write in their own variety of Alsatian with shared grapheme to phoneme rules. However, **most speakers are not familiar with these spelling systems**, and there is thus a lot of variation in how speakers write Alsatian, depending both on the specific variety they speak **and on the degree of influence of French and Standard German spelling**" (Bernhard et al. 2025 after Beiner, 2022).

UNIVERSITY OF WARSAW

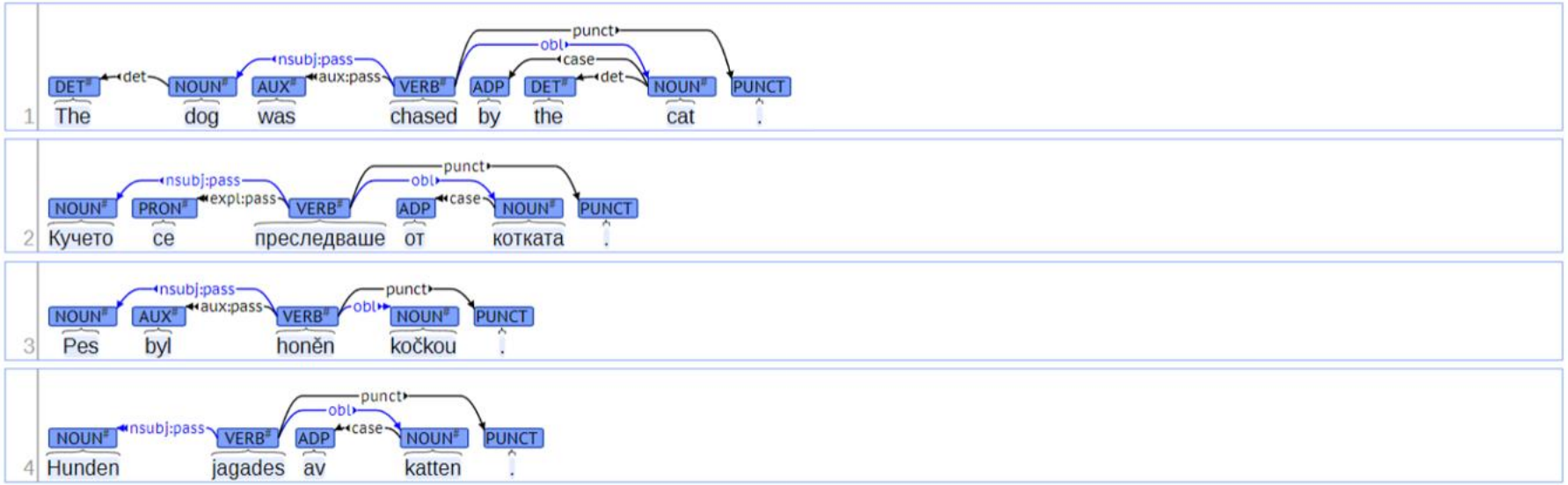
Center for Research and Practice in Cultural Continuity

Funded by the European Union

FOSTERLANG

Example from Universal Dependencies

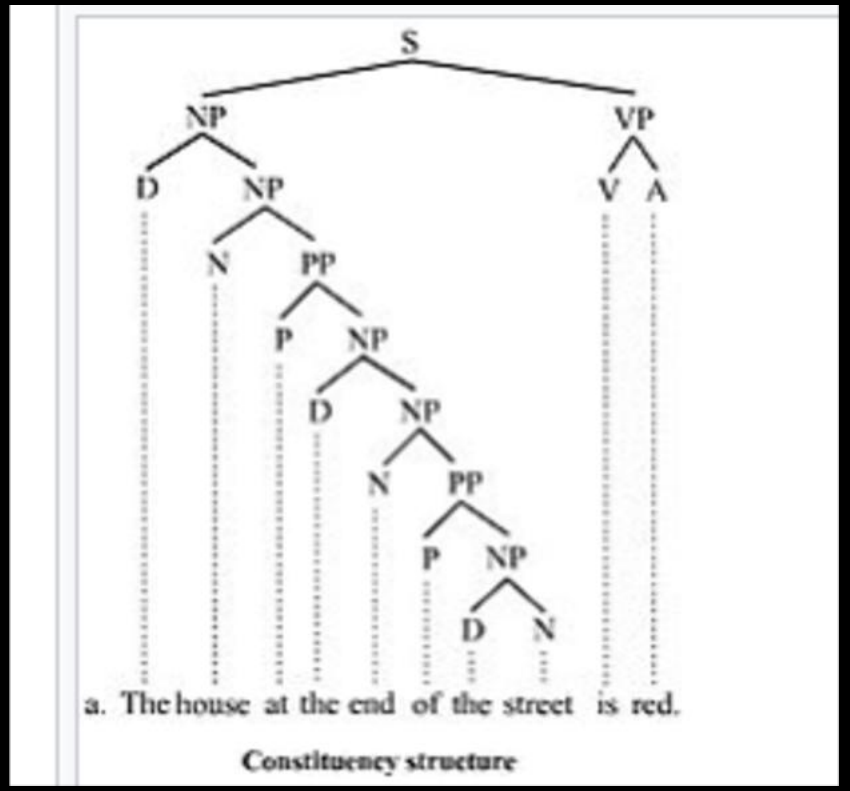
This is illustrated in the following parallel examples from English, Bulgarian, Czech and Swedish, where the main grammatical relations involving a passive verb, a nominal subject and an oblique agent are the same, but where the concrete grammatical realization varies.



UNIVERSITY OF WARSAW
WYDZIAŁ ARTES LIBERALES
Center for Research and Practice in Cultural Continuity
Funded by the European Union
FOSTERLANG

Parts of speech categories in universal dependencies

- Treebanks
- Various languages
- Please do not forget to quote the authors! ☺️



Source: An example from Wikipedia. https://en.wikipedia.org/wiki/Treebank

Linders, G.M. & Louwerse, M.M. (2023)

UNIVERSITY OF WARSAW

WYDZIAŁ ARTES LIBERALES

Center for Research and Practice in Cultural Continuity

Funded by the European Union

FOSTERLANG

ANALYZING TEXT

What are the basic examples of language processing tools?

- Word tokenization and segmentation
- PoS (parts of speech tagging)
- Named entity recognition

On which dimensions can a text be analyzed?

- Sentence level
- Paragraph level
- Document level

Linders, G.M. & Louwerse, M.M. (2023)

Data-driven and Dictionary-based tools

- Dictionary-based tools are those that include a dictionary or database to categorize words and consequently categorize a text.
- Data-driven methods instead rely on patterns in the text and quantify those using computational linguistic and statistical models.
- Hybrid approaches use a mixture of both dictionary-based and data-driven approaches.
- While it is difficult to create a dictionary-based tool in multiple languages because individual **dictionaries or databases need to be constructed for each language**, it is at the same time difficult to create data-driven tools in multiple languages because it would require natural language data that is annotated in a **unified manner** across different languages.

Linders, G.M. & Louwerse, M.M. (2023)

UNIVERSITY OF WARSAW

WYDZIAŁ ARTES LIBERALES

Center for Research and Practice in Cultural Continuity

Funded by the European Union

FOSTERLANG

HERITAGE/LEGACY DATA

UNIVERSITY OF WARSAW
WYDZIAŁ ARTES LIBERALES
Center for Research and Practice in Cultural Continuity
Funded by the European Union
FOSTERLANG

What is heritage data and why is it important?

- What do we need to take into account when working with heritage data?
  - Authors rights
- How do we access this data for processing?
  - Through manual retyping
  - Retrieving text from original digital files
  - Building our own OCR (Optical Character Recognition) models for digitization.

UNIVERSITY OF WARSAW

Center for Research and Practice in Cultural Continuity

Funded by the European Union

FOSTERLANG

MONGOLIC LANGUAGES IN THE WORLD

Fig. 1 Current distribution of Mongolic languages. Map designed on the basis of the PPT Depot https://pptdepot.com/

UNIVERSITY OF WARSAW
WYDZIAŁ ARTES LIBERALES
Center for Research and Practice in Cultural Continuity
Funded by the European Union
FOSTERLANG

Dagur language 1/2

Endangered, easternmost Mongolic language spoken mainly in northeast China.

It does not have one common, official written standard.

As late as in **1930** it was considered to be “almost completely unexplored” (Poppe, 1930).

In Nicholas Poppe’s grammar from 1930 it is still called a dialect (ДАГУРСКОЕ НАРЕЧИЕ)

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).

According to N. Poppe the variety containing various Manchu words was “less pure” than the variety recorded by him from the speakers originating from Hailar -> very normative approach

UNIVERSITY OF WARSAW
WYDZIAŁ ARTES LIBERALES
Center for Research and Practice in Cultural Continuity
Funded by the European Union
FOSTERLANG

Dagur language 2/2

Today Dagur is spoken mainly in the **Heihe region of the Middle Amur basin**, in the locations within the Nonni river basin, in the Ewenki Autonomous Banner **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 and also in Cyrillic script in the 1950's (Tsumagari, 2005). Nowadays, Dagur speakers use either **Manchu script or Chinese for writing.**

UNIVERSITY OF WARSAW
WYDZIAŁ ARTES LIBERALES
Center for Research and Practice in Cultural Continuity
Funded by the European Union
FOSTERLANG

SCRIPTS

MANCHU SCRIPT (E.G. SCHOOL HANDBOOK)

LATIN SCRIPT (E.G. SAMUEL MARTIN'S DAGUR GRAMMAR, 1961)

CYRILLIC SCRIPT (E.G. NICHOLAS POPPE, B. TODAEVA, 1986)

UNIVERSITY OF WARSAW
WYDZIAŁ ARTES LIBERALES
Center for Research and Practice in Cultural Continuity
Funded by the European Union
FOSTERLANG

1. woa̯irdā mu̯rgu̯lĕl″ch″ĭwē,
   holdā mō̆r′ĭl″ch″ĭwē.
2. nĕ̊ʿk kʿŭ̄ kʿŭ̄ u̯l bolŏ̆n,
   nĕ̊ʿk mōd ɣal′ĭ (var. sigī) u̯l bolŏ̆n.
3. aĭjĕ̊̄sh — u̯gu̯wē,
   al″ch″ĭjĕ̊̄sh — woanăwē.
4. hoarăm orʿt āɣāsā, kʿu̯l′ī or′ēbē̆;
   u̯su̯ɣ woalăŋ āɣāsā, bĕ̊jī orēbē̆.
5. chʿagĭ chʿag′ērā u̯l ān,
   chʿaʿkʿĭldu̯ɣ kʿu̯ʿkʿĕ̊̄rĕ̊̄ u̯l ān.

УЧЭКЭН КЭКУ

энэ нэк учэ̄кэн кэку ā̆сан. кумни дэ̄р гарā̆р толколсон,
рго магал олсо̆н; хурбисэн, курбус дэ̄л олсо̆н; гиркэсэн,
чо̄р олсон; алкусан, алар мор олсо̆н.
тэгэ̄р харгудэ̄ йн нэк мангэ̄jй вачирсан. тэрэ мангэ̄
сбдж хэлбэi: — шй хэр эjмэ̄р дуртй бajин болсонш̄э̄?
энэ учэ̄кэн кэку джā̆дж хэлбэi: — бй тэрэ кумнид гарā̆р
нколсон, торго магал олсон; алкусан, алар мор олсон.

POPPE (1930)

TODAEVA (1986)

11. A: Ušiken nas[]aase ~ni # gaade + iau.legaa~bele +
little age ABL as-for outside go-let if
sain.
good
Children can be let outdoors from the time they are little
tots.

(MARTIN 1961)

UNIVERSITY OF WARSAW

WYDZIAŁ ARTES LIBERALES

Center for Research and Practice in Cultural Continuity

Funded by the European Union

FOSTERLANG

OUR RESEARCH GOALS

TO CREATE A NEW EXPERIMENTAL DAGUR CORPUS (DOLIŃSKA & BERNHARD, 2024) BASED ON LEGACY DATA.

TO CARRY OUT EXPERIMENTS ON THIS CORPUS

TO ASSESS THE POTENTIAL FOR TRANSFER LEARNING FROM OTHER RELATED AND UNRELATED LANGUAGES AND TO IDENTIFY THE SETTINGS AND DATA TRANSFORMATIONS THAT PERFORM BEST.

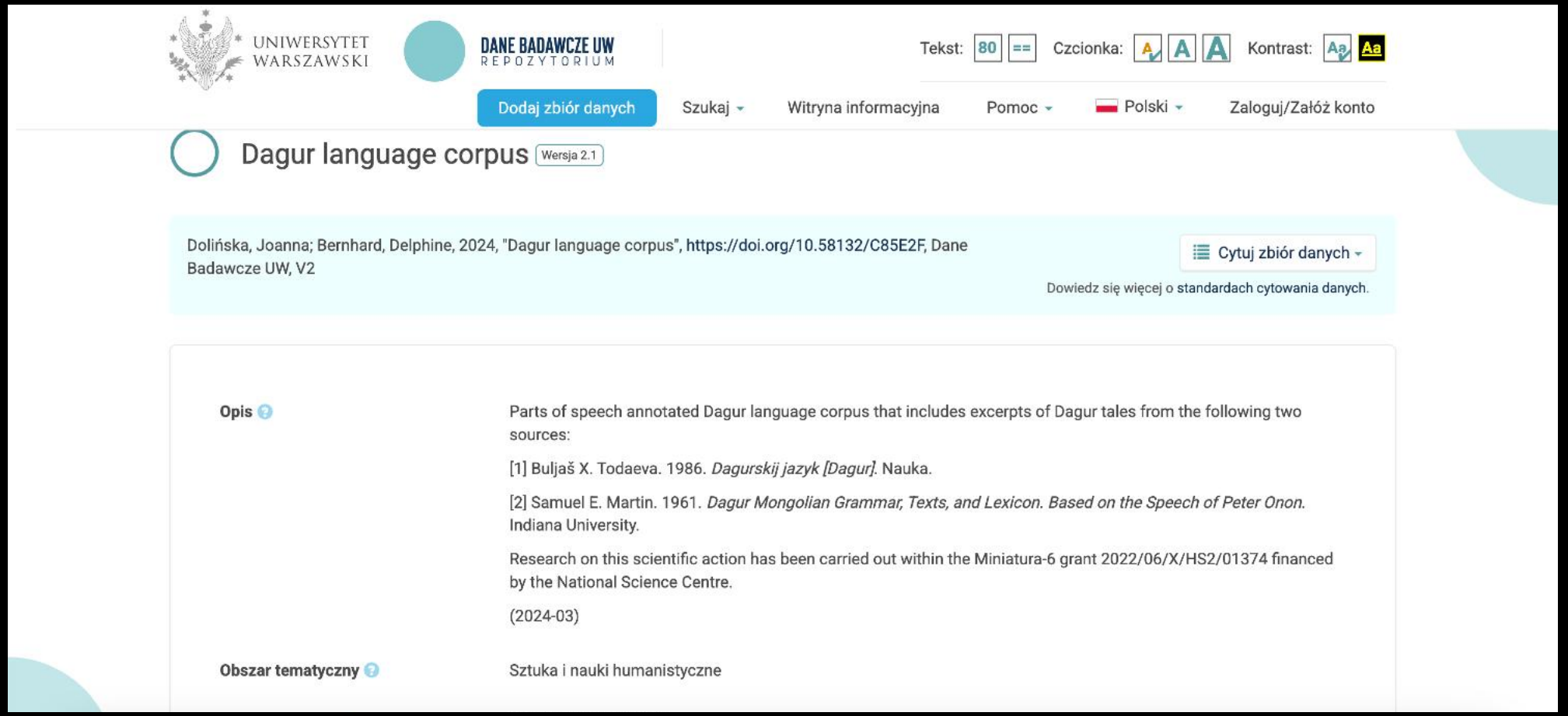
UNIVERSITY OF WARSAW

Center for Research and Practice in Cultural Continuity

Funded by the European Union

FOSTERLANG

CORPUS



UNIVERSITY OF WARSAW
WYDZIAŁ ARTES LIBERALES
Center for Research and Practice in Cultural Continuity
Funded by the European Union
FOSTERLANG

THE EXPERIMENTAL DAGUR CORPUS

The Dagur corpus contains 4,502 tokens and 550 sentences. It includes excerpts from Todaeva (1986) written in Cyrillic script and texts from Martin (1961) in Latin script. https://doi.org/10.58132/ C85E2F.

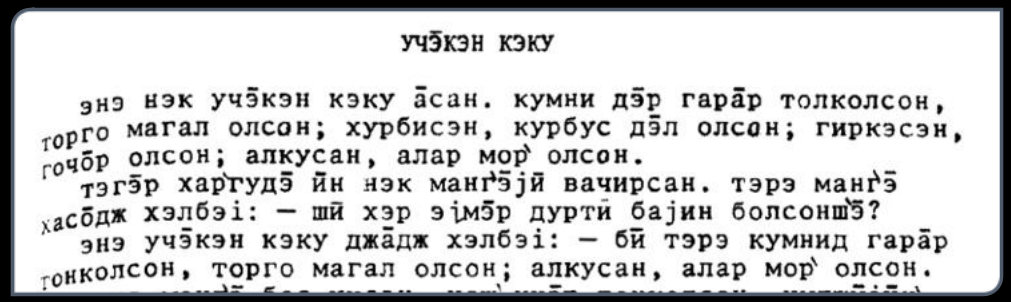
UNIVERSITY OF WARSAW
WYDZIAŁ ARTES LIBERALES
Center for Research and Practice in Cultural Continuity
Funded by the European Union
FOSTERLANG

Challenges when creating an experimental corpus for an endangered language

Challenges

- The choice of the character of the corpus
- Author's rights issues
- How to prepare a corpus
- Optical character recognition (OCR) issue when dealing with the photos of a given text
- Universal Dependences – helpful and sometimes misleading

Explanation



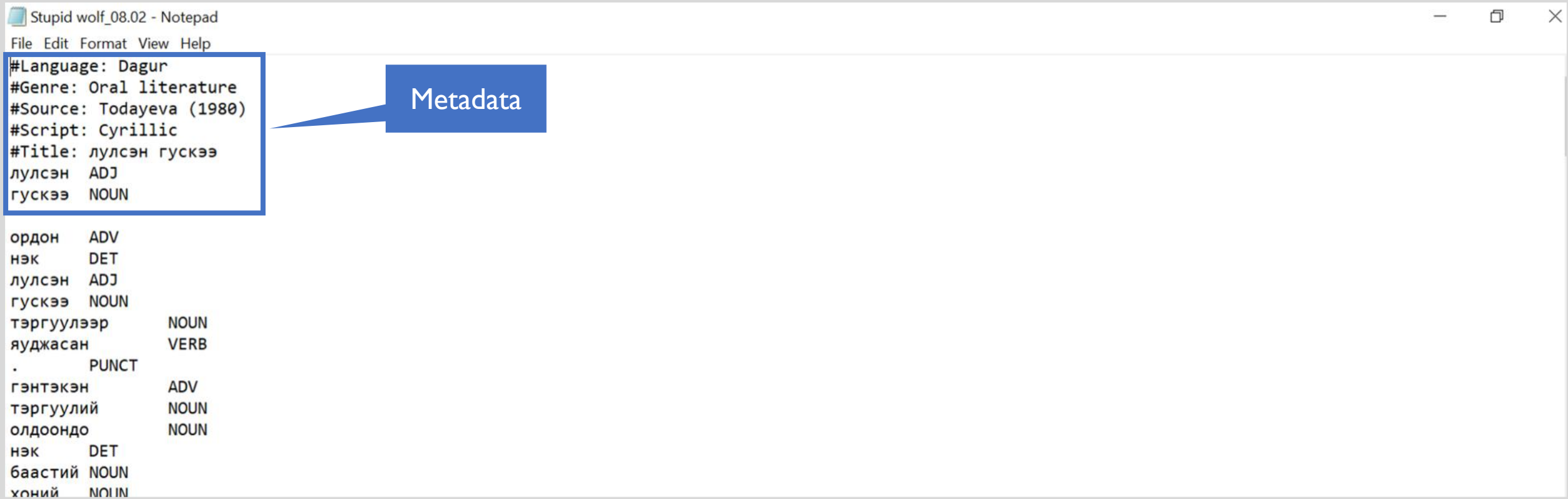
учӭкэн кэку

энэ нэк учӭкэн кэку ӑсан. кумни дӭр гарӑр толколсон, торго магал олсон; хурбисэн, курбус дӭл олсон; гиркэсэн, гочбр олсон; алкусан, алар мор олсон.
тэгӭр харгудӭ йн нэк мангӭйй вачирсан. тэрэ мангӭ хасбдж хэлбэі: — шй хэр эімбр дуртй бајин болсоншӭ?
энэ учӭкэн кэку джӑдж хэлбэі: — бй тэрэ кумнид гарӑр гонколсон, торго магал олсон; алкусан, алар мор олсон.

When a digitized text is a photograph, we need an OCR software to decipher the signs. Otherwise, we need to type down the text ourselves.

Source: Todaeva, B. Kh.(1986). Dagurskij jazyk. Nauka.

Example of a basic annotated corpus



Source: Excerpt from a corpus prepared by J. Dolinska and D. Bernhard 2024.

Bibliography

Austin, P. K. (2021). Language documentation and revitalization. In J. Olko & J. Sallabank (Eds.) *Revitalizing Endangered Languages*. Cambridge University Press, pp. 199-212. DOI: https://doi.org/10.1017/9781108641142

Beiner, N. 2022. Quelle(s) norme(s) pourl'.criture de l'alsacien en 2022 ? Master's thesis, Universit. de Strasbourg.

Bender, E. and Friedman., B. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association fo Computational Linguistics,* 6: 587-604.

Blokland, R., Partanen, N., Rießler, M. and Wilbur, J. (2019): "Using Computational Approaches to Integrate Endangered Language Legacy Data into Documentation Corpora: Past Experiences and Challenges Ahead." *Proceedings of the Workshop on Computational Methods for Endangered Languages*: Vol. 2 , Article 5.

Dolińska, J. (2022): The application of natural language processing (NLP) tools in relation to selected Mongolic languages: review of the current literature, available NLP tools and outlooks for the future. In: L. Becerra, B. Favre, C. Gardent & Y. Parmentier (Eds.): *LIFT-TAL. Actes des journées jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL), 14 au 15 novembre 2022 Marseille, France*, p. 188-197. HAL Id: hal-03846837

Dolińska, J. & Bernhard, D. (2023). POS Tagging for the Endangered Dagur Language. *The 2024 joint international conference on computational linguistics, language, resources and evaluation* (under review).

Elliott, R. (2021). Technology in Language Revitalization. In Olko, J. & Sallabank, J. (Eds.) Revitalizing Endangered Languages. Cambridge University Press, pp. 297-316.

Hoff, B., Beiner, N., & Bernhard, D. (2025). Universal Dependencies for the Alemannic Alsatian dialects. In S. Jablotschkin, S. Kübler, & H. Zinsmeister (Eds.), Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025) (pp. 10–22). Association for Computational Linguistics. https://aclanthology.org/2025.tlt-1.2/

Martin, S. E. (1961). "Dagur Mongolian. Grammar, Texts, and Lexicon", *Uralic and Altaic Series 4*. Indiana University

Poppe, N. (1930). Dagurskoe narechie [Dagur]. Izdatel'stvo Akademii Nauk SSSR.

Siewert, J., Scherrer, Y. and Tiedemann, J. (2021). Towards a balanced annotated Low Saxon dataset for diachronic investigation of dialectal variation. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, p. 242–246, Düsseldorf, Germany. KONVENS 2021 Organizers.

Todaeva, B. Kh. (1986). Dagurskij jazyk [Dagur]. Nauka.

Language resource references:

Badmaeva, E. and Tyers, F. (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.

Universal Dependencies: https://universaldependencies.org/

Dolińska, J. & Bernhard, D. (2024). Dagur language Corpus. https://doi.org/10.58132/C85E2F, Dane Badawcze UW, V2

JOANNA DOLIŃSKA

J.DOLINSKA@AL.UW.EDU.PL

THANK YOU FOR YOUR ATTENTION!