Rule-based morphological parser for Shughni language: nouns, verbs and adjectives

Elen Kartina
National Research University Higher School of Economics

Project Proposal

Research Advisor: George Moroz

March 2025

Contents

Testing page

Русский текст ёмаёюж

Consonant = б в w г □ □ □ д □ ж з □ й к □ л м н п р с т □ ф х х̱ □ ц ч ш □ ;

Vowel = а □ е □ и □ о у □ у̱ ;

Abstract

Automatic morphological analysis is a crucial task of computational linguistics. In this work I propose a rule-based morphological analysis tool for Shughni language based on Helsinki Finite-State Technology (HFST). The tool set is planned to contain two types of tools: a morphological parser that breaks word-forms into stems and morphemes and assigns morphological tags to each one of them; a morphological generator that outputs word-forms taking stems and morphological tags as an input. This project aims to cover at least three main parts of speech: nouns, verbs and adjectives. **TODO: Extend the abstract and review it later**

1 Introduction

Morphological parser is a fundamental tool, a wide range of computational linguistics' tasks rely on some form of morphological model. For morphologically rich languages it is close to impossible to list and manually define all the possible word-forms. The only reasonable way to approach such problem is to model a language's morphology.

For high-resource languages morphology modeling is usually approached using deep learning (DL) models, which are trained on large amounts of data. This method is not available for low-resource languages that lack digital textual data, for such languages linguists usually apply rule-based approach. Shughni is a low-resource language with very few data available, which leaves us the rule-based option.

Shughni language (ISO: sgh; glottolog: shug1248) is a morphologically rich low-resource language. It belongs to the Iranian branch of the Indo-European family (Plungian 2022: 12), and it is spoken by circa 80 000 - 100 000 people (Edelman and Dodykhudoeva, 2009) in two regions: Mountainous Badakhshan Autonomous Region (Tajikistan) and Badakhshan Province (Afghanistan). Both regions have a subregion, where

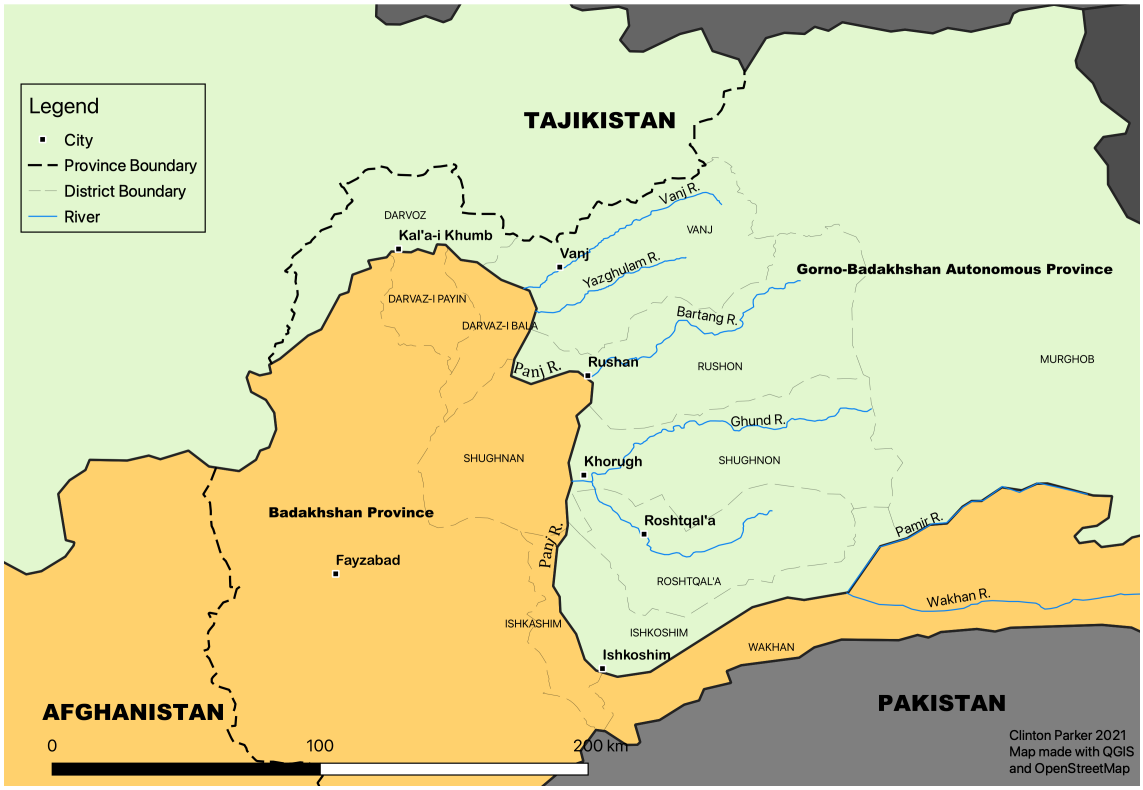

Figure 1: Mountainous Badakhshan Autonomous Province of Tajikistan and Badakhshan Province of Afghanistan

Shughni language is the most spoken native language, the subregions are called 'Shughnon' in Tajikistan and 'Shughnan' in Afghanistan (Parker 2023: 2), see Figure 1 for details. Shughni has a mixed morphological typology type (Parker 2023: 94), which means that grammatical meanings can be carried by morphemes, words or clitics. There are three scripts for Shughni language: Latin, Cyrillic and Arabic. The Arabic script is used on the territory of Badakhshan Province of Afghanistan, and Cyrillic and Latin scripts are used in the Mountainous Badakhshan Autonomous Region of Tajikistan. The Cyrillic script was created and gained popularity in 1930s, after it was set as the primary script for teaching in schools on the Shughni-speaking territory of Tajikistan.

Morphology analysis tools in question will focus on the variation of Shughni that is spoken in Tajikistan. Cyrillic and Latin script will be supported: the core analysis tool will be implemented in Cyrillic script, and Latin script support will be implemented via transliteration.

2 Literature review

TODO: refine references; заменить дословные переводы названий работ чем-то получше

2.1 Shughni

Shughni belongs to the Iranian family (Plungian 2022: 12). It is considered to belong to 'Pamiri' languages. According to Edelman and Dodykhudoeva (2009) estimates, approximately 80 000 - 100 000 people speak Shughni. There are two regions that speak Shughni: one on the territory of Afghanistan's Badakhshan Province - 'Shughnan', and the other on the territory of Tajikistan's Mountainous Badakhshan Autonomous Province - 'Shughnon' (Parker 2023: 2) (see Figure 1).

There are two main dictionaries of the Shughni language: 'Shughni texts and dictionary' (Zarubin 1960) and 'Shughni-Russian dictionary' (Karamshoev 1988-1999), both are written using Cyrillic script and include Russian translations. Some early dictionaries are 'Brief grammar and dictionary of Shughni' (Tumanovich 1906), that is also using Cyrillic and translates to Russian, and 'Shughni dictionary by D. L. Ivanov' (Salemann 1895), that translates to Russian but uses Arabic script alongside Cyrillic transcriptions for Shughhi word-forms.

Several Shughni grammar descriptions were written throughout the years, starting from basic grammar description done by D. L. Ivanov (Salemann 1895: 274-281). Lat-

est significant works were 'Shughni language' (Эдельман Дж.И., Юсуфбеков Ш.П. 1999: 225-242), 'Сравнительная грамматика восточноиранских языков' (Edelman, Dodykhudoeva 2009) and 'A grammar of the Shughhi language' by (Parker 2023). The latter grammar by Parker will be the main theoretical source for the development of the morphology analysis tool since it is the biggest existing grammar, the most detailed and the most recent one.

A significant contribution to the Shughni NLP field is 'Digital Resources for the Shughni language' (Makarov et al., EURALI 2022). The authors, among other tools and resources, developed a rule-based morphological analysis tool for the Shughni language. The parser proposed in this work, while also being rule-based, differs in its implementation through the use of Helsinki Finite-State Technology (HFST).

2.2 Morphology modeling

2.2.1 Neural approach

One of the most recent and widely adopted approaches to morphology modeling involves the use of the Transformer-based deep learning models. This approach usually requires large amounts of training data in form of manually tagged word-forms, which is not available for Shughni. There are texts available to me, that were manually tagged at the Linguistic Convergence Laboratory of HSE university, which consist of 3453 tokens in total. While this amount of training data is small, it is worth mentioning that it is possible to train a Transformer-based model with such small datasets, as shown by Kondratyuk and Straka (2019). Their approach includes fine-tuning a pre-trained multilingual BERT model, authors conclude, that multilingual learning is most beneficial for low-resource languages, even ones that do not possess a training set.

2.2.2 Rule-based approach

Finite-state technology (FST) is a finite-state machine with two tapes, one for input strings and one for output strings. The machine maps the alphabet of the first string to the alphabet of the second string, this concept was first proposed by Mealy (1955) and Moore et al. (1956). Eventually linguists noticed this technology and started applying it to model natural languages' grammar. Woods (1970) suggested Recursive Transition Networks (RTN) for sentence structure parsing, RTN essentially is a finite-state machine applied to syntax.

TODO: efficiency of fst

FST is widely applied when it comes to creating rule-based tools. Some of the examples of FST-based morphological tools are: morphological parser for the Tamil language by Sarveswaran et al. (2021), a morphological transducer for Kyrgyz by Washington et al. (2012), a morphological analyzer and generator for Sakha by Ivanova et al. (2022) and a morphological analyser for the Laz language by Onal and Tyers (2019).

Linden et al. (2007)

3 Methods

4 Expected outcomes

5 Conclusions

6 Discussion and future

7 References