

Правительство Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего
образования
Национальный исследовательский университет
«Высшая школа экономики»

Факультет гуманитарных наук
Образовательная программа
«Фундаментальная и компьютерная лингвистика»

Картина Элен Геннадьевич

**Правильный морфологический парсер для шугнанского языка:
существительные, глаголы и прилагательные**

Выпускная квалификационная работа студента 4 курса бакалавриата группы БКЛ211

Академический руководитель образовательной программы
канд. филологических наук, доц.
Ю.А. Ландер

«_____» _____ 2025 г.

Научный руководитель
канд. филологических наук, доц.
Г.А. Мороз

Научный консультант
Стажёр-исследователь
М.Г. Мельченко

Москва 2025

Contents

1	Introduction	3
1.1	Shughni	3
1.2	Morphology parsing TODO: rethink this heading	5
2	Existing methods and solutions	5
2.1	Deep learning methods	5
2.2	Rule-based methods	5
2.3	Existing solutions for Shughni	5
3	Data	5
3.1	Grammar descriptions	5
3.2	Dictionaries	5
3.3	Text corpora	5
4	Methods	5
4.1	Finite-state transducers TODO: Move to introduction?	5
4.2	Transliteration	5
4.3	Russian lemmas TODO: rethink heading	5
4.4	Testing	5
4.5	Metrics	5
4.6	Rule declaration TODO: rethink heading	5
4.6.1	Nouns	5
4.6.2	Verbs	5
4.6.3	Adjectives	5
4.6.4	Pronouns(?)	5
4.6.5	Numerals(?)	5
4.6.6	Anything else(?)	5
4.7	Metrics	5
5	Results	5
6	Conclusion	5
	References	6

Abstract

In this work I present a rule-based morphological analysis tool based on Helsinki Finite-State Technology (HFST) for the Shughni language (ISO: sgh; glottocode: shug1248), a language of the Iranian branch of the Indo-European family, a member of ‘Pamiri’ areal language group. While one existing rule-based parser exists for Shughni (Melenchenko, 2021), it does not utilize finite-state transducer technology. This work proposes the first HFST-based morphological parser implementation for Shughni, offering the advantages of this well-established framework for morphological analysis. The parser is presented in two variations: a morphological parser that breaks each word-form into stem and morphemes and assigns morphological tags to each one of them; a morphological generator that outputs word-forms taking a stem and morphological tags as an input. **TODO: prev sentence is questionable** This is a continuation my previous work, where nouns, pronouns, prepositions and numerals were implemented (Osorgin, 2024). This project covers **TODO: what**

TODO: Review abstract after finishing the work

1 Introduction

1.1 Shughni

The Shughni language (ISO: sgh; Glottolog: shug1248) is a language of the Iranian branch of the Indo-European family (Plungian, 2022, p. 12). As of June 1997, it was estimated to be spoken by approximately 100,000 people (Edelman & Yusufbekov, 1999, p. 225) **TODO: найти http://old.iea.ras.ru/publications_new/kalandarov.html ? там тоже есть оценка** in the territories of Tajikistan and Afghanistan. Both countries have a subregion where Shughni is the most widely spoken native language. The Shughni-speaking subregion of Tajikistan is called ‘Shughnon’ and it belongs to the ‘Gorno-Badakhshan Autonomous’ province. In Afghanistan, the Shughni-speaking region is called ‘Shughnan’ and it lies within the territory of ‘Badakhshan’ province (Parker, 2023, p. 2).

TODO: Pamiri

1.2 Morphology parsing **TODO: rethink this heading**

2 Existing methods and solutions

2.1 Deep learning methods

2.2 Rule-based methods

2.3 Existing solutions for Shughni

3 Data

3.1 Grammar descriptions

3.2 Dictionaries

3.3 Text corpora

4 Methods

4.1 Finite-state transducers **TODO: Move to introduction?**

4.2 Transliteration

4.3 Russian lemmas **TODO: rethink heading**

4.4 Testing

4.5 Metrics

4.6 Rule declaration **TODO: rethink heading**

4.6.1 Nouns

4.6.2 Verbs

4.6.3 Adjectives

4.6.4 Pronouns(?)

4.6.5 Numerals(?)

4.6.6 Anything else(?)

4.7 Metrics

5 Results

References

- Edelman, D. I., & Yusufbekov, S. (1999). Шугнанский язык [Shughni language]. In *Языки мира: Иранские языки. III: Восточноиранские языки [Languages of the world: Iranian languages. III. Eastern Iranian languages]*. <https://iling-ran.ru/web/ru/publications/langworld/volumes/7>
- Melenchenko, M. G. (2021). *Автоматический морфологический анализ шугнанского языка [Automatic full morphology analysis for Shughni]*, NRU HSE.
- Osorgin, I. G. (2024). *Создание морфологического парсера для шугнанского языка в системе lexd u twol [Creating a morphological parser for the Shughni language using lexd and twol systems]*, NRU HSE.
- Parker, C. (2023). *A grammar of the Shughni language* [Doctoral dissertation, Department of Linguistics, McGill University].
- Plungian, V. (2022). The study of shughni: The past and the future. *RSUH/RGGU Bulletin: "Literary Teory. Linguistics. Cultural Studies", Series*. 2022;(5):11-22.