

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

**Правиловый 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	4
1.1	Shughni language	4
1.2	Morphology parsing TODO: rethink this heading	4
2	Existing methods and solutions	4
2.1	Deep learning methods	4
2.2	Rule-based methods	4
2.3	Existing solutions for Shughni	4
3	Data	4
3.1	Grammar descriptions	4
3.2	Dictionaries	4
3.3	Text corpora	4
4	Methods	4
4.1	Finite-state transducers	4
4.2	Rule declaration TODO: rethink heading	4
4.2.1	Nouns	4
4.2.2	Verbs	4
4.2.3	Adjectives	4
4.2.4	Pronouns(?)	4
4.2.5	Numerals(?)	4
4.2.6	Anything else(?)	4
4.3	Metrics	4
5	Results	4
6	Conclusion	4

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 (**melchenko_2021_parser**), 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_twol**). This project covers **TODO: what**

TODO: Review abstract after finishing the work

1 Introduction

1.1 Shughni language

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

4.2 Rule declaration **TODO: rethink heading**

4.2.1 Nouns

4.2.2 Verbs

4.2.3 Adjectives

4.2.4 Pronouns(?)

4.2.5 Numerals(?)

4.2.6 Anything else(?)

4.3 Metrics

5 Results

6 Conclusion