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

1 Introduction 3

2 Literature review 4

3 Methods 4

4 Expected outcomes 4

5 Conclusions 4

6 Discussion and future 4

7 References 4

Testing page

Русский текст ёмаёюж

Consonant = б в w г □ $\overset{□}{□}$ $\overset{□}{□}$ д □ ж з □ й к □ л м н п р с т □ ф х $\overset{□}{\text{х}}$ □ ц ч ш □ ;

Vowel = а □ е □ и □ о у □ $\overset{□}{\text{у}}$;

Abstract

Automatic morphological analysis is a crucial task of computational linguistics. The Shughni language, being a minority language, is one of the languages that lacks tools for such an essential task. 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.

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

TODO: Указать какую версию шугнанского я покрываю, почему, и сослаться на https://aclanthology.org/2022.eurali-1.9.pdf

For high-resource languages this problem is usually being solved using deep learning (DL) models, which require large amounts of training data. This method is not available for low-resource languages that lack digital text data. For such languages linguists usually apply rule-based approach.

2 Literature review

3 Methods

4 Expected outcomes

5 Conclusions

6 Discussion and future

7 References