Proceedings of the EURALI Workshop @LREC2020, pages 61–64
Marseille, 20 June 2022

Digital Resources for the Shughni Language

Yury Makarov, Maksim Melenchenko, Dmitry Novokshanov
HSE University
105066, Moscow, Staraya Basmannaya Ulitsa, 21/4
yurmkrv@gmail.com, mgmelenchenko@edu.hse.ru, danovokshanov@edu.hse.ru

Abstract

This paper describes the Shughni Documentation Project consisting of the Online Shughni Dictionary, morphological analyzer, orthography converter, and Shughni corpus. The online dictionary has not only basic functions such as finding words but also facilitates more complex tasks. Representing a lexeme as a network of database sections makes it possible to search in particular domains (e.g., in meanings only), and the system of labels facilitates conditional search queries. Apart from this, users can make search queries and view entries in different orthographies of the Shughni language and send feedback in case they spot mistakes. Editors can add, modify, or delete entries without programming skills via an intuitive interface. In future, such website architecture can be applied to creating a lexical database of Iranian languages. The morphological analyzer performs automatic analysis of Shughni texts, which is useful for linguistic research and documentation. Once the analysis is complete, homonymy resolution must be conducted so that the annotated texts are ready to be uploaded to the corpus. The analyzer makes use of the orthographic converter, which helps to tackle the problem of spelling variability in Shughni, a language with no standard literary tradition.

Keywords: Shughni, online dictionary, morphological analyzer

1. The Shughni Language

1.1 General Information

Shughni is one of the Pamir languages, which belong to the Iranian branch of the Indo-European family. As estimated by (Edelman and Dodykhudoeva, 2009), it is spoken by circa 100,000 people in the Mountainous Badakhshan Autonomous Region of Tajikistan and in the neighbouring Badakhshan Province of Afghanistan. Our project revolves around the Shughni-Rushani subgroup of the Pamir languages. This subgroup consists of the following closely related idioms: Shughni (with Bajuwi), Rushani (with Khufi), Bartangi (with Roshorvi), and Sarikoli. The resources described below are focused on the Shughni language, especially on its variety spoken in Tajikistan.

Tajik is the official language of Tajikistan, taught in schools and exerting strong influence on Shughni, resulting in numerous loanwords and metatypic changes in the latter. A considerable part of Shughni speakers in Tajikistan also know Russian.

1.2 Writing System

Shughni has no official status in contemporary Tajikistan and there is no common writing tradition for it. Before the 20th century Arabic script was used sporadically. In the 1930s the Soviet authorities introduced a Latin-based Shughni alphabet but a decade later switched to a Cyrillic-based script. However, due to political reasons writing in Shughni was not welcomed until the 1980s, and no stable orthography was established. In the 1990s there appeared a few other writing systems. Scholars of various fields of study use different orthographies or even continue to create their own. Despite the absence of established orthography, there is a non-negligible written literature in Shughni (poetry, prose fiction, journalistic articles).

2. Online Shughni Dictionary

2.1 Benefits of Having an Online Dictionary

2.1.1 Online Dictionary as Research Tool

There is a number of reasons why online dictionary is instrumental in linguistic work. First, it facilitates creating a corpus by providing a fast interface for finding word meanings and hence making glossing an easier task. Moreover, building a lexical database is required for automatic analyzers (see 3 below), which also contribute to annotating texts for corpora, and other NLP tools. Second, if a dictionary contains a lot of examples of word usage, for some purposes it can be used as a corpus itself. Third, using existing dictionaries as a basis, researchers can update and expand them, and even turn them into a detailed lexical database by establishing an elaborate label system. Forth, while usually dictionaries of lesser-resourced languages are unidirectional (i.e., English-French and not vice versa), after digitalizing they become available for searching in both languages. Lastly, using website means that groups of researchers from different parts of the world can work on the same project cooperatively.

2.1.2 Online Dictionary for Local Communities

Online dictionary serves as an essential resource for language learning and revitalization, especially in the context of multilingualism. Speaking of teaching at schools, it is hard to imagine the codification of a language in the absence of an accessible dictionary. Available on the Internet through an easy-to-use interface, it is likely to become popular among a wider audience. This is also supposed to help those interested in developing literature in that language, and in reading existing texts.

2.2 The Case of the Shughni Language

2.2.1 Existing Dictionaries

During the 20th century two main dictionaries of the Shughni languages were published: one by Ivan Zarubin (1960), another by Dodkhudo Karamshoev (1988–1999). Both were written in Russian. While the former cannot be called comprehensive, the latter has three volumes totaling about 16,000 lexemes and contains massive illustrative material (based on high-quality field records made in the 1960–1970s in different Shughni-speaking areas). This played a decisive part in using that dictionary as a basis for

the project and also in using Russian as the main language for the interface of the website[1].

2.2.2 Digitalization of the Karamshoev's dictionary

The digitalization of the Karamshoev's Shughni-Russian dictionary came in several steps. First, we scanned the volumes and used automatic recognition software to get text files. The dictionary uses two Cyrillic writing systems (for Russian and for Shughni) with a few diacritics and supplementary symbols from Greek script (see 4.1 below). This made proofreading an essential part of the process. At the same time entries were annotated, so that it would be possible to transfer every part of an entry to the corresponding section in the database.

2.3 Website Architecture

2.3.1 Searching and Database Structure

Online Shughni Dictionary is supposed to be useful both for linguists and general public (be it native speakers, language activists or language learners). This means that apart from basic functions such as finding a particular lexeme in the dictionary and giving a link to it in the output, there have to be more search options.

In our database every lexeme is represented as a network: it has different forms (spelling and phonetic variants, different grammar forms), meanings and idiomatic units (for Shughni, compound verbs and idioms are discriminated). Further, every idiomatic unit has its own set of meanings, and every meaning (both of lexemes and idiomatic units) can be illustrated with examples.

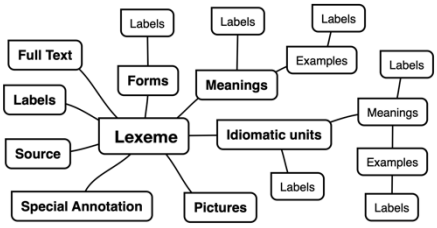

Figure 1: Representation of a lexeme in the database

Labels have different scope and hence can be applied to the lexeme in general (e.g., when it belongs to a certain dialect) or to a particular form, meaning or example. A lexeme may have multiple labels, but they will be arranged in a way that is not random but rather logical. There is also room for special annotation, be it comments about word usage or etymology. Every lexeme can be illustrated with some pictures. The online dictionary can aggregate different sources, so for each lexeme there is information about the dictionary from which it was taken.

Such database architecture makes it possible to search in particular domains, e.g., in meanings only. It is useful when one looks up 'iron' to find out what this metal is called in Shughni but gets a word for 'shirt' instead because in this entry there is an example saying, 'They *iron* their shirts every day.' Nevertheless, full texts of entries are also available as a search domain. The system of labels allows one to make conditional search queries, e.g., to find all obsolete verbs of falling in a certain dialect (provided that there are labels 'obsolete', 'verb of falling' and 'dialect X').

2.3.2 Search Engine and User Interface

In Shughni, there is no standard orthography. Some of the writing systems exploit symbols not easily accessible through every keyboard. That is why we provide users with the virtual keyboard. Despite that, one can choose not to discriminate between similar special symbols, and query with an ambiguous symbol will be parsed as having a set of analogous symbols in one place, e.g., 'wин' will be considered having either 'и' or 'ӣ', and the output will include both entries for 'wин' and 'wӣн.' Another handy function is looking up parts of words. It is possible to enter sequence 'wин' and get entries 'wин' and 'парwӣн' in the output.

Apart from the Cyrillic script illustrated above there is a Latin orthography used by some Iranists. We support both systems, so it is possible to search and view the entries using the two of them.

Another feature of the Online Shughni Dictionary is adaptive design, so it is comfortable to access the website through practically every kind of device.

2.3.3 Editing the Online Dictionary

Online Shughni Dictionary allows editors to change the database via web interface. The editing interface is intuitive and can be used by those who do not have programming skills. Every part of a lexeme (see 1.3.1 above) can be modified, e.g., it is possible to add new grammar forms, edit existing meanings, add new idiomatic units, etc. Editors are allowed to add new and delete existing entries or hide the latter from users.

While using the online dictionary users often notice typos or even mistakes. From our perspective, collecting such feedback is of vital importance. That is why there is an option to report a mistake while viewing entries. Such messages are sent to the special section of the editor's interface on the website.

2.4 Future Development

Such website architecture is suitable not only for the dictionaries of the Shughni language and its dialects but at least for other languages of Pamir as well. Digitalizing other sources in the same manner as with the Karamshoev's dictionary makes it feasible to develop a lexical database for multiple Iranian languages. Such database will be instrumental in typological studies as well as in comparative linguistics. Some of the most obvious functions that this database must have is the ability to form a list of cognates and compare same concepts (or lexemes) across different languages of the region.

We plan to create a new version of the dictionary where, among other things, special labels will be used for annotating the Tajik and Russian borrowings. Other dictionaries of the languages of the Shughni-Rushani group will be added to the platform. In addition to that, we will

[1] It is worth mentioning that most researchers studying the Shughni language have some knowledge of Russian since the latter is one of the languages widely spoken in the region.

use our corpus (see 4.2) to describe more lexemes. It will help minimize the effects of the problem of the dictionary size (see 3.3.1).

3. Automatic Morphological Analyzer

3.1 Principles of Automatic Analysis

Another instrument available on the website of our project is the morphological analyzer for Shughni. The algorithm is written in Python and uses the same database with the online dictionary. Once put into the analyzer, text is sentenized and tokenized using *nltk* module (Bird et al., 2009). Then, for each token the program tries to find correspondences in the *Forms* section of the database (see 2.3.1 above) so that the token contains a root from the database. If the correspondence is found, the program attempts to identify the remaining parts of the token (before and after the root); there is also an algorithm that makes sure that the chain of affixes or clitics in these parts are possible in Shughni.

The main goal of the analyzer is to identify every morpheme or clitic in the analyzed sentence. The set of affixes and clitics was collected manually using the grammar descriptions and dictionaries of the Shughni language. Every morpheme and clitic have special restrictions on the grammatical and phonetic contexts where they can appear. For example, suffix *-um*, which is used for forming ordinal numerals, is attached only to the stems of cardinal numerals. Lative suffix has several phonetically conditioned variants, one of which, *-rd*, appears only after vowels. These restrictions are ascribed to every morpheme and clitic, and the algorithm considers them in the process of analysis.

Grammar rules implemented in the analyzer are hardcoded. For example, verbs in the present tense get a person suffix, whereas in the past and perfect tenses person markers are clitics which are not necessarily attached to the verb. This means that if the analyzer suggests that a morpheme chain contains a present verb stem with no person suffix attached, such analysis is rejected.

3.2 Export to Corpus

The output of the analyzer can be presented in two ways. One is via intuitive web interface, another is a json-file which can be uploaded to the Shughni corpus (see 4.2 below). The thing is that there often are competing analyses in the output from which one has to choose, i.e., *homonymy resolution* must be performed before exporting the annotated text to the corpus. The website interface allows one to do it easily.

3.3 Problems of Automatic Analysis

3.3.1 Size of the Dictionary

The analysis that the program conducts is helpful but far from perfect. There are several obstacles that the algorithm currently faces. Some lexemes found in Shughni texts are simply absent from the online dictionary. Among them are collocations and proper names, e.g., names of settlements, rivers, and other geographical objects in the Badakhshan region. Another large group of words not found in the dictionary is Tajik borrowings, which are numerous since the Shughni language spoken in Tajikistan is heavily influenced by the state language, namely Tajik. It is often difficult to distinguish between borrowing and code switching as the Shughni speakers in Tajikistan are at least bilingual. The same is true for the Russian borrowings though they appear to be less frequent.

3.3.2 Spelling Variation

In the absence of a codified spelling system, a lot of words vary in how they are written. This is particularly acute when it comes to loanwords, e.g., *miloim* (corresponding to *miloyim* 'soft' in the dictionary), *salomati* (*salůmati* 'health'), *tavallud* (*tawallud* 'birthday'), *avtomobīli* (*aftamubīl* 'car') are not recognized by the analyzer. Sometimes variability arises from dialectal or even interspeaker variation.

3.3.3 Ignoring Diacritics and Special Symbols

The problem of spelling variation (see 3.3.2 above) can be partly solved by ignoring certain differences between special symbols with and without diacritics. For example, Shughni has a set of short and long vowels which correspond to symbols with and without diacritics in writing, cf. *i* and *ī*. Such symbols are often confused in texts (even by natives). That is why during processing tokens with confusable symbols the same algorithm as described in 2.3.2 above is by default applied. Analyzing 'winč' will result in getting the correct output despite the fact that 'wīnč' is the correct spelling for this stem.

This feature, however, can lead to the increase of incorrect analyses in the output (cf. minimal pairs with short and long vowels). To avoid such scenario one can switch off ignoring diacritics and special symbols in the setting.

4. Other Resources

4.1 Orthography Converter

As noted in 1.2 above, the Shughni language spoken in Tajikistan uses different orthographies based on Cyrillic and Latin alphabets. They often include diacritics or digraphs. Orthography converter is designed to solve this problem and automatically convert Shughni texts in various orthographies to the unified Latin-based spelling system used by our project for research and documentation.

Some of the existing orthographies use the same symbol for different phonemes. For example, letter *j* can denote phonemes /j/, /d͡ʒ/ and /d͡z/. The converter must be able to identify for what phoneme each problematic symbol stands. For example, for *Xudowandard půnd jītet at wi roh yen rost kinet* 'Prepare ye the way of the Lord, make his paths straight' in the input, the converter understands that letter *y* is used for /j/; then, letter *j* cannot stand for the same phoneme and should be interpreted as /d͡ʒ/ (the next frequent meaning) instead. The output of the looks like *Xudowandard půnd ǰītet at wi roh yen rost kinet*, where *y* is for /j/ and *j* is for /d͡ʒ/.

It is presumed that, according to the law of large numbers, if the text is sufficiently big, there will be enough different symbols in it for the program to identify the phonetic meaning of them correctly. For special cases, users can manually choose the phonetic meanings of the problematic symbols in the settings.

The converter is accessible as a separate instrument. It also serves as a pre-processing tool in morphological analysis

as the analyzer is supposed to work only with the texts in the unified Latin orthography of the project.

4.2 Corpus

The Shughni corpus contains oral and written texts of different genres including fairy tales, prose fiction, poetry, etc. Some of these are stories which were recorded during fieldwork and can be played. There are also parallel Shughni-Tajik texts. Annotation consists of layers for token, morphemes, glosses, part of speech and meaning. Glosses are mostly in English, whereas the translations are in Russian. Metadata contains information about the author or speaker, title, source, place and date of the recording, its genre and whether the text is annotated manually. The corpus runs on the open-source Tsakorpus platform developed by Timofey Arkhangelskiy (see https://github.com/timarkh/tsakorpus). The current volume of the corpus is circa 40,000 tokens.

5. Bibliographical References

Bird, S., Klein, E., and Loper, E. (2009). Natural Language Processing with Python. Sebastopol: O'Reilly Media Inc. https://www.nltk.org/book/

Edelman, D. I. and Dodykhudoeva, L. R. (2009). Shughni. In G. Windfuhr (Ed.), *The Iranian Languages*. London and New York: Routledge, pp. 787–852.

Karamshoev, D. (1988–1999). Shughni-Russian dictionary in 3 volumes. Moscow: Izd-vo "Nauka".

Zarubin, I. I. (1960). Shughni dictionary and texts. Moscow-Leningrad: Izd-vo Akademii nauk SSSR.

6. Language Resource References

Online Dictionary of the Shughni-Rushani Language Group. (2022). Languages of Pamir: Online Dictionary / Памирские языки: онлайн-словарь, distributed via HSE University (Yury Makarov), 2.0, ISLRN 865-656-162-260-6. https://pamiri.online/

Shughni corpus. (2022). Shughni language corpus, distributed via HSE University, 1.0, ISLRN 728-407-900-078-8. https://linghub.ru/shughni_corpus/search

7. Acknowledgements

This work was supported by the Humanitarian Research Foundation of the Faculty of Humanities, HSE University in 2020, Project "Computational and Linguistic Resources for the Shughni Language" (Russian: Компьютерные и лингвистические ресурсы для поддержки шугнанского языка), and in 2021, Project "Computational and Corpus Instruments for Iranian Studies" (Russian: Компьютерные и корпусные инструменты для иранистических исследований).

We thank the anonymous reviewers for their useful comments and suggestions.