

Standard Dargwa Corpus

Toldova Svetlana Juryevna, Sokur Elena Olegovna
National Research University "Higher School of Economics"

The research was supported by RSF (project No. 22-28-01648 "Variation in the discourse and lexicon: an investigation of closely related languages with digital methods")

Introduction

- The project on Dargwa languages (**Nakh-Dagestanian** languages)
- Muira, Kadar are in focus
- Standard Dargwa as well
- an ongoing project of creating Standard Dargwa Corpus
- the pilot version of the corpus is based on newspaper texts

Dargwa languages

- Dargwa is one of the written languages of the Republic of Dagestan (Russia)
- Standard Dargwa is the standardized language used in writing: in the Soviet period it was created based on the dialect of Aqusha.
- It has official status and is used in the media (television, radio, press)

Data

- newspaper texts from archives of the Dargwa newspaper “Zamana” (2010-2020)
- politics, sports, economy, society, culture, anti-terror, finance, etc.
- approximately 50 issues per year
- the total number of issues is 505.
- Each issue contains 40-50 texts of different length.
<https://zamana.info/>
- The corpus size is ~9.8 million wordforms.

Corpus platform

- Tsakorpus [Arkhangelskiy 2012].
- It allows to present data with different layers of annotation.
- The standard set of layers includes
 - wordforms,
 - lemmas,
 - grammatical annotation
 - glosses.
- <https://github.com/timarkh/tsakorpus>

Word #1

Word: замаһа

Lemma:

Grammar:

Gloss:

Language/tier: Dargwa

Full-text search: Precise match

Search sentences Search words / lemmata Select subcorpus

Welcome! Here is how you can find something:

- Type a word or a lemma (dictionary form) in the text box above. Wildcards (*) and even [regular expressions](#) are allowed.
- Or choose some tags, such as part of speech, in the *Grammar* box.
- Hit *Search sentences* to find randomly sorted examples of what you are looking for.
- Or hit *Search words* or */ lemmata* to get a table with words that conform to your query.

There are lots of other options! Click at the top to find out.

- Udmurt corpus
- Hill Mari
- Bashkir

Preprocessing

- Texts are annotated with the *Uniparser-morph* technology and published through the *Tsakorpus* corpus platform.
- The corpus size is ~9.8 million wordforms.
- Parsed coverage is 6.8 million wordforms (~69.7%).

Uniparser

- *Uniparser-morph* is a rule-based morphological analysis tool, developed by T.Arkhangel'skiy (Arkhangelskiy 2012).
- It requires two main files to be used: *lexemes.txt* which contains a list of all lexemes and *paradigms.txt* which contains a list of affixes linked to each other in a specific order.
- Each entry in *lexemes.txt* starts with a *-lexeme* line which opens a dictionary
- *lex* stands for the dictionary form of the word.
- The file *lexemes.txt* contains all lexemes from the Dargwa-Russian dictionary [Yusupov 2017], including geographical names, human names and Russian borrowings.

Uniparser

- *stem* describes the lexeme as a morpheme object: a dot indicates a place where affixes from an inflectional paradigm can be inserted.
 - In (1), a stem contains two dots: the first is for the class agreement paradigm, and the second is for inflectional morphology.
 - *gramm* includes grammatical tags describing a lexeme (e.g. part of speech, gender, transitivity, etc.).
 - *paradigm* is a link to an inflectional paradigm from the *paradigm.txt* file which contains a list of affixes that are used to create inflected wordforms of this lexeme.
 - *gloss* contains a translation.
- (1) -lexeme
 - lex: абалк
 - stem: а.алк.
 - gramm: V,tr
 - paradigm: vclass
 - gloss: зажечь

Dargwa morphological parser

- The file *lexemes.txt* contains all lexemes from the Dargwa-Russian dictionary [Yusupov 2017], including geographical names, human names and Russian borrowings.
- Most frequent Russian stems are taken from [Lyashevskaya, Sharoff 2009] (<http://dict.ruslang.ru/>). Nouns are assigned the most frequent nominal paradigm. Verb borrowings are formed as *Russian infinitive + bares/bires* (Dargwa verb ‘to do’), e.g. *защитить-бирес* ‘protect-do’ “to protect”.

Dargwa morphological parser

- автоматизироватьбарибси
- автоматизироватьбарнила
- автоматизироватьдарахъес
- автоматизироватьдарибти
- агропромышленностьла

Dargwa morphological parser

We are able to provide analysis for 20 741 lexemes (~16% of them are Russian borrowings with Dargwa morphology) and 179 335 individual wordforms (~6% are Russian borrowings), see Table 2.

	lexemes	wordforms
Russian borrowings	3390	11024
Dargwa	17351	168311
Total	20741	179335

- 340 978 entities remained unanalyzed.
- 24 782 are irrelevant: these are numbers, words in other languages (English, Arabic, etc.), random characters, etc.
- 49 439 are words that were tagged as Russian using the python package *pymorphy2* (Korobov 2015).
- The rest 266 757 are Dargwa wordforms that are not recognized by the current version of the uniparser and need further investigation.

Dargwa morphological parser

- (5) a. амч|<б>иреска<б>ир-ес
go.bald<N>-INF
- b. амч|иреска<й>р-ес
go.bald<M>-INF
- In Standard Dargwa, the majority of verbs have gender/number agreement markers, which are usually controlled by the argument in the absolutive case (the single argument for intransitive verbs, and the P argument (patient, stimulus, theme) for transitive verbs).
- The markers distinguish between masculine, feminine and neuter in the singular, and human and non-human in the plural. There are some cases when an Ergative argument triggers gender agreement.
- there can be from 1 to 3 gender/number markers in one verb.
- 19 strategies include irregular forms

Corpus search

Илгъуна

1. ил

ил-гъуна

он-подобный

?

замана нушачи гIергъила бархи савли бакIибсири

замана нуша гIергъиласи 1. бар 1. савли 1. бакIи

замана нуша-чи гIергъила бар-хи савли б-акI-иб-с-и-ри

время мы-SUPER следующий бар-IN утро N-прийти-AOR-1SG-1SG

N

N

? савли

? бакIи

- Thank you for your attention!