Journal of Siberian Federal University. Humanities & Social Sciences
2024 17(5): 905–915

EDN: QCUVBE
УДК 81’32; 811.35

Building an Open Corpus and a Morphological Parser for Corpus Annotation for Standard Dargwa

Svetlana Iu. Toldova* and Elena O. Sokur
National Research University “Higher School of Economics”
Moscow, Russian Federation

Received 01.02.2023, received in revised form 12.12.2023, accepted 08.04.2024

Abstract. This paper is devoted to the ongoing project of creating a Standard Dargwa Corpus (Standard Dargwa is a Nakh-Dagestanian language). A pilot version was released in 2022. The paper describes building a fully-functional version of the corpus. First, we describe the pipeline used to develop the corpus. Second, we discuss the procedure of building and enhancing the morphological parser. The parser provides morphological annotation. The layers include the morphemic structure of a word, the grammatical labels of morphemes, the translations of lexemes from the dictionary. Third, we discuss the drawbacks of the parser and ways for overcoming them. Finally, we describe the corpus usage functionality.

Keywords: under-resourced language, corpus development, Standard Dargwa, Nakh-Daghestanian languages, morphological parser, interlinear glossing.

The research is supported by RSF grant No. 17–18–01184. We are thankful for all the specialists providing us Dargwa data.

Research area: social structure, social institutions and processes.

Citation: Toldova S. Iu., Sokur E. Building an open corpus and a morphological parser for corpus annotation for Standard Dargwa. In: *J. Sib. Fed. Univ. Humanit. soc. sci*., 2024, 17(5), 905–915. EDN: QCUVBE

* Corresponding author E-mail address: toldova@yandex.ru

Опыт создания открытого корпуса текстов на литературном даргинском языке и разработки морфологического парсера для его аннотации

С.Ю. Толдова, Е.О. Сокур
Национальный исследовательский университет «Высшая школа экономики»
Российская Федерация, Москва

Аннотация. Статья посвящена продолжающемуся проекту по созданию корпуса текстов на литературном даргинском языке (нахско-дагестанская группа языков). Пилотная версия корпуса была создана в 2022 г. В настоящей статье описывается опыт разработки полной версии корпуса. Во-первых, дан полный цикл разработки корпуса (необходимая последовательность этапов его разработки). Во-вторых, обсуждается процедура разработки и оптимизации морфологического парсера, обеспечивающего поморфемную аннотацию текстов в корпусе. Слои такой аннотации для каждой словоформы в корпусе включают: (а) слой поморфемной сегментации, (б) слой словарной формы, (в) слой грамматической поморфемной аннотации, (в) перевод лексемы на русский язык по двуязычному словарю. В-третьих, анализируются проблемы, возникающие при применении выбранного инструмента для разработки системы морфологической аннотации (системы UniParser Т. Архангельского) к материалу даргинского языка, а также возможные пути их решения. Помимо этого, в статье описывается корпусной функционал.

Ключевые слова: малоресурсные языки, корпуса текстов, литературный даргинский, нахско-дагестанские языки, морфологический анализатор, глоссирование.

Исследование выполнено при поддержке гранта РНФ № 17–18–01184. Мы благодарны всем специалистам, предоставившим нам данные по Даргве.

Научная специальность: 5.4.4 – социальная структура, социальные институты и процессы.

Цитирование: Толдова С.Ю., Сокур Е.О. Опыт создания открытого корпуса текстов на литературном даргинском языке и разработки морфологического парсера для его аннотации. *Журн. Сиб. федер. ун-та. Гуманитарные науки*, 2024, 17(5), 905–915. EDN: QCUVBE

I. Introduction

This paper deals with the development of the Standard Dargwa corpus on the basis of its pilot version, released in 2022 and described in (Toldova, Sokur, 2022). We describe the development of the corpus including the procedure of text preparation and discuss the procedure of building the first morphological parser for Standard Dragwa. The parser is rule-based. It is based on Uniparser-morph technology (Archandelskii, 2012) which provides an easy and quick start for creating parsers for under-resourced languages. It can be usable by theoretical linguists with no prior knowledge of Natural Language Processing (NLP). In this paper we describe our corpus in detail and discuss the problems we have come across.

Dargwa languages are a group of Nakh-Dagestanian languages spoken in Dagestan. Dargins are one of the most numerous indig-

enous ethnic groups in Dagestan. The Dargwa group includes several Dargwa languages and quite a few dialect varieties. Standard Dargwa was created on the basis of the Agusha dialect. It is the standardized language used in writing. The core text collection in our corpus is based on newspaper texts.

The morphological parser is used for providing interlinear glossing for texts. It suggests the morpheme segmentation of words, the morphological labelling of word segments, and the translation into Russian of corresponding lexemes based on the Dargwa-Russian dictionary (Iusupov, 2017), see Fig. 1. This type of annotation is standard for language documentation projects.

It can also provide training data for automatic morphological taggers, morphological segmentation etc. The texts are accompanied by morphological annotation and word-by-word translation into Russian.

The testing of the pilot corpus has shown that the texts need some cleaning, e.g. deleting letters-to-the-editor in Russian from newspaper readers, surnames lists etc. The morphological parser created in 2022 also needs further improvements. There are several issues that need special treatment. These are (a) overgeneration, especially for the most frequent words, (b) some morphemes are unseen by the parser that should be considered it, (c) impossible morpheme sequences provided by the parser.

We provide background in section II, describe the data in section III, describe the morphological parser in section IV, and discuss the parsing errors and drawbacks, ways for improving the annotation, and illustrate corpus functionality in section V.

II. Background

A. Dargwa languages

The Dargin languages constitute a separate branch of the Nakh-Dagestanian family. Dargin people live primarily in south-central Dagestan. According to the 2021 census, there are more than 600 000 speakers of Dargin languages. There are four basic subgroups of Dargwa and many dialectal varieties (for dialect variations see Gasanova 1971). In order to have a language for official communication, the standardized variety based on the Agusha dialect was created. As mentioned, it is the written language used in official and media communication.

The grammar of Standard Dragwa is described in (Abdullaev, 1954; Musaev 1999; Van den Berg, 2001). One of the largest dictionaries is (Iusupov 2017), which provides very detailed morphological information and an exhaustive list of possible translations into Russian for every lexeme.

The Dargwa alphabet is based on Cyrillic. There are many digraphs for ejective and other consonants absent in Russian, e.g. *кI, mI, гъ* etc. Some of the digraphs include Latin symbols. These digraphs serve as a source of error in parsing and it is a source of numerous OCR mistakes.

Nouns in Dargwa have following gender-number categories: masculine, feminine, neuter for the singular, and first/second person, human, non-human in the plural. Dargwa nouns mark case distinctions. There are five non-locative cases and many locative forms. There are many allomorphs for plural depending on the noun stem. Many nouns have irregular forms for the plural. This information is reflected in the dictionary. Nouns have two stems. For

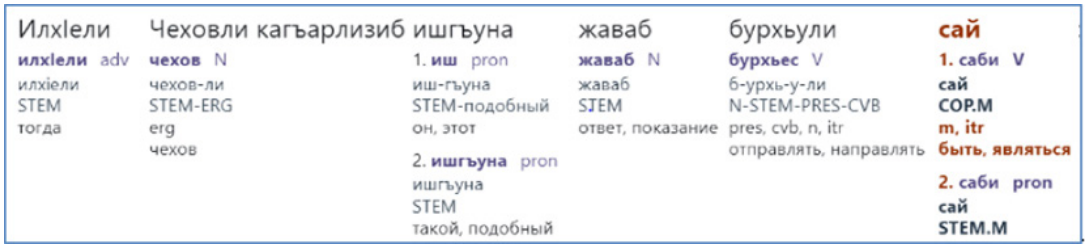


Fig. 1. An example of interlinear annotation with the following layers: (1) text, (2) segmentation into morphemes, (3) morphological tags, (4) Russian translation of a lexeme from the dictionary

example, the wordform *учительтани* (a borrowing from Russian "учитель" teacher) has the following interlinear morpheme glosses: *учитель-т-а-ни* STEM-PL-OBL.PL-ERG, where OBL.PL is a special form of the oblique plural stem.

Standard Dargwa, like other Daghestanian languages, has ergative alignment. That means that the Agent argument of transitive verbs does not trigger the class agreement. Verbs in Dargwa agree in class with a noun in the absolutive case: the single argument for intransitive verbs, and the P argument (patient, stimulus, theme) for transitive verbs.

Most verbal roots have a perfective and an imperfective form within a single verbal paradigm. Perfective and imperfective stems have different dictionary entries. Dargwa verbs, some postpositions, adjectives, adverbs, and particles agree in gender class with the noun in the Absolutive case, as in other Daghestanian languages. The gender marker is prefixed to a verb stem. Moreover, for many cases they are infixed into a wordform. Another noun/verb/adjective stem or prefix can precede the agreement marker in a wordform, as in *лас-б-ирхъ-ес* (around-N-do-INF) 'to look around'. There are a lot of compound verbs derived with a lexical stem and an auxiliary or light verb like 'to do', 'to become' and others. There can be up to three slot for class marker. This issue provides a challenge for Dargwa word-form parsing (see details below). Dargwa verbs also have a complex person agreement system.

B. Creating corpora for under-resourced languages and their interlinear annotation

At present, many languages of the world are endangered. There are many projects aimed at preserving language data (e.g. the Endangered Languages Documentation Programme (ELDP, https://www.eldp.net/)). Their aims are to collect texts in a particular language, to create digital collections, and to make them freely available online. In order for the morpheme structure of a language to be interpretable, the usual standard for presenting texts in a language documentation project is to provide interlinear glossing annotation, that is, with a split-into-morphemes layer, a morpheme labels layer, a part of speech layer (for the standards for language documentation corpora see (Himmelmann 1998; Comrie, Haspelmath, Bickel 2008; Goodman, Crowgey 2015; Arkhipov, 2020 etc.). The main glossing tool used in documentation projects is Fieldworks Language Explorer (FlEx, SIL). The example of layers is provided in Fig. 1 in section I.

The corpora collected in language documentation projects are usually small (10 000 up to 100 000 tokens). The texts are collected by linguists during fieldwork. They are annotated manually. The sentences usually have free translation into the target language. The translation is provided by native speakers. It is a highly time-consuming procedure.

Another direction of corpora development for low-resourced languages is creating web-corpora or so-called medium corpora, which include media texts and literary texts. The texts for these corpora are collected from the internet using crawlers with special tools for language recognition tuned for a particular language (Xingyuan, Ozaki, Anastasopoulos, Neubig, Levin 2020; Scannell, 2007). It is impossible to provide free-translation layers for web-corpora or corpora of written media texts. These texts are usually original texts without a parallel translation. However, usually the corpora are morphologically parsed by rule-based parsers based on bilingual dictionaries, cf. the procedure for Udmurt, described in (Arkhangelskiy, 2019). This is standard for medium corpora for under-resourced languages, e.g. Almaty Corpus of Kazakh language (http://web-corpora.net/KazakhCorpus/search/), Udmurt web-corpora (http://udmurt.web-corpora.net/index_en.html) and others.

Interest in parsing under-resourced languages has increased rapidly. The task of providing a text with morphological glosses with small resources for learning conditions is a challenge for modern NLP technologies. There are attempts to provide machine-learning based or neural network based morphological parsers. All of them require a certain amount of glossed data.

At present, there are no such data for Dargwa. Thus, our project provides a certain

amount of morphologically annotated Dragwa data for linguistic investigation and can serve as a first step for training more advanced systems.

C. A pipeline for building corpora for under-resourced languages

A standard pipeline for constructing corpora for written texts in minority languages is described by Arkhangelskiy, e.g. [11]. We take it as a basis for constructing our corpus.

We have chosen the Tsakorpus platform developed by Arkhangelskii (Archangelskii 2022, https://tsakorpus.readthedocs.io/en/latest/). This platform provides a convenient system for loading the data prepared by a user in the json format. It also provides a search engine and a convenient web-interface for search, accessing all the available layers.

To provide interlinear and translation layers, we use the Uniparser-morph tool, also developed by Archangelskii (Archangelskii 2012) (https://github.com/timarkh/uniparser-morph). This tool is designed for quick adaptation of the parser engine for under-resourced languages which do not have enough data for training statistical parsers.

It requires two main files: the file which contains a list of all lexemes, and the file, which contains a list of affixes. Thus, one only needs to prepare two files with language-specific data to start using the parser for a specific language.

The standard way of preparing data for the parser is to use a bilingual dictionary for extracting lemmas, stems, translations and the necessary grammatical word classes and to use grammatical description: a list of available morphemes and patterns of morpheme ordering. A more detailed description of the data for the parser is given in the following sections.

D. Dargwa language corpora

There exist several small corpora for different Dargwa dialects: Sanzhi Dargwa corpus, Muira Corpus, Kadar corpus etc. All these corpora present a small amount of texts collected during fieldwork. They are annotated manually with the FLEx tool. Dialect variation of Dargwa is great. For example, the number of noun cases vary from 6 to 85. Thus, these corpora cannot be used for morphological parser training for Standard Dargwa.

The Standard Dragwa corpus is a new corpus of one of the under-resourced languages and we have created a quite effective morphological parser for its annotation.

III. Data for corpus

A. The text data

At the first stage, we chose newspaper texts for our corpus data. We used the archives of Dargwa newspaper “Zamana” 2010–2020 (https://zamana.info/). There are approximately 50 issues per year. The total number of issues is 505. Each issue contains 20–50 texts of different lengths and on diverse topics such as politics, sports, economics, society, culture, anti-terror, finance, etc. There are also some literary texts in newspapers including poems.

For the pilot corpus we used a newspaper issue as a corpus item for the indexing procedure. At the next stage, we split the issues into separate short texts. The first reason is that the texts are very heterogeneous within an issue belonging to different genres. At present, we have no genre classifier for Dargwa, however we can build one in the future. The second reason is that newspaper issues contain specific textual data such as surname lists, sudoku etc. They also contain texts in Russian, e.g. the readers’ letters to the editor. We have excluded Russian texts and non-coherent texts from our corpus.

We are planning to add some literary texts within the month, namely tales and ballads. We have some problems with OCR texts, there are a great number of errors in recognition of Latin symbol I (I is used in digraphs for ejective consonants) and the Cyrillic symbol “и”.

While the first release of the corpus contained ~9.8 million wordforms, the cleaned version contains ~7.8 million wordforms. The parsed coverage is approximately 76 % (against 70 % for the pilot version).

B. The dictionary processing for morphological parser

As mentioned, we use the Dargwa Russian Dictionary by Iusupov (Iusupov 2017) as the basis for constructing the parser. It contains 40,000 lexemes. Lexical entries are well organized, contain all the necessary information for

АЛЛАГЬ/БУРЦ-ЕС [*мн.* **~дурц-ес], -у, -ули, -уси, -ен(–ена/я)**; III; *несов.* обожествлять; боготворять; **бацбархIи ~дурцес** обожествлять небесные светила (*солнце и луну*). ‖ *сов.* **аллагьбуцес**.

Fig. 2. An example of a dictionary entry for the verb аллагь/бурц-ес 'to divinify'

wordform formation. The format of lexical entries is regular and is amenable for parsing. An example of a dictionary entry is given in Fig. 2:

The verb *аллагьбурцес* 'to dignify' is a compound verb. It has two stems: the noun-stem *аллагь* for Allah (or the God) and the verb-stem of the light verb *бурцес ("to do")*. Symbol "/" denotes the place for the class-agreement morpheme. The wordforms inside the square brackets are the wordforms for other noun classes. There is only one more form available for class agreement in the example. This is the form for human plural objects. Some other inflectional forms are given in the dictionary entry of a verb (e.g. the aorist form), as well as the translation into Russian. Thus, the dictionary allows us to assign a verb its paradigm type and its translation.

The dictionary also contains some words borrowed from Russian (e.g. *промышленность* "industry"), that are frequent in newspapers.

We use the dictionary in order to get the citation form for a lexeme, its stem, and its formation class (paradigm type). We also extract all the translations excluding idioms and examples and some additional information.

IV. Uniparser for Standard Dargwa

A. Morphological parser for Dargwa

Uniparser-morph is a rule-based morphological analysis tool, developed by Arkhangelskii (Archangelskii 2012). It was created primarily for under-resourced languages. One has to provide two files for the parser: lexemes.txt containing a list of all lexemes and paradigms. txt containing 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. The example of an entry is given in Fig. 3.

'lex' is a citation form from the dictionary. 'stem' is a basic word stem for different wordform formation, a dot indicates a place where affixes from an inflectional paradigm can be inserted. In Fig 3, the stem contains two dots: the first is for a class agreement paradigm, and the second is for inflectional suffixes following the stem. 'gramm' includes grammatical tags containing information referring to a lexeme (e.g. part of speech, gender, transitivity, etc.). 'paradigm' is a link to an inflectional paradigm from the paradigm.txt file. 'trans_ru' contains the Russian translation of the lexeme from the dictionary.

The paradigms.txt file contains a collection of inflectional affixes grouped into paradigms. Each paradigm has a unique name (cf. vclass in Fig. 4):

'flex' introduces a morpheme. Within each morpheme a dot marks the position where it can be attached to a stem. In Fig. 4, each morpheme has three types of adpositioning: infix (.д.), suffix (.д) and prefix (д.). '.' means that a wordform continues with a subsequent paradigm that is linked at the end of the example (paradigm: vforms_1). 'gramm' includes

– lexeme
lex: сурбулхъес
stem: сур.улхъ.
gramm: V, itr
paradigm: vclass_bdel_middle
trans_ru: висеть, свешиваться, виснуть, свисать

Fig. 3. An example of a lexeme entry in the lexemes.txt file

```
– paradigm: vclass
– flex:.д.<.>//.д<.>//д.<.>
gramm: pl
gloss: PL
– flex:.б.<.>//.б<.>//б.<.>
gramm: n
gloss: N
…
paradigm: vforms_1
```

Fig. 4. An example of an item in the paradigm.txt file

grammatical tags that are associated with this morpheme, and 'gloss' is the grammatical class (POS) of a morpheme. A paradigm key introduces a link to the paradigm type.

The file lexemes.txt contains all lexemes from the Dargwa-Russian dictionary (Iusupov 2017) including geographical names, human names and Russian borrowings.

Building the Uniparser for a particular language is an iterative process. After each parsing procedure the unparsed words are analyzed. After the first iteration we removed the numbers, tokens containing non-alphabetic symbols, one-symbol tokens etc. We also found stems borrowed from Russian (in the Iusupov dictionary) with Dargwa affixes attached to them, e.g. турбаза-ла 'a tourist village – GEN'. In order to provide analysis for such borrowings, we took frequent Russian stems from (Lyashevskaya, Sharoff 2009 (http://dict.ruslang.ru/). Nouns were assigned the most frequent nominal paradigm. Verb borrowings are formed as Russian infinitive + a light verb (verbs *bares/bires* 'to do', биэс 'to become' etc.), e.g. адресовать-барес 'to.address-do' "to address".

Now, we are able to provide analysis for approximately 57,800 lexemes. There are 24,734 verbs among them.

B. Parsing Dargwa verbs for gender/number affixes

In Standard Dargwa, the majority of verbs have gender/number agreement markers. The markers distinguish between masculine, feminine, and neuter in the singular, and 1/2 person, human and non-human in the plural.

The lexeme.txt file includes verb stems from the Dargwa-Russian dictionary (Iusupov 2017). They are classified based on the type of their gender/number paradigm. The gender/number paradigms are distinguished based on how the markers are expressed, and what positions they occupy. Firstly, paradigms differ in how the singular masculine and the plural features are expressed. Secondly, there can be from 1 to 3 gender/number slots on one verb. There are some morphophonological processes when the masculine singular morpheme is attached to a verb stem. The allomorphs for *w*-affix are *j*- or null affix. The plural uses *d*- or *r*- affixes, *b*- is used for the neuter and *r*- for the feminine. For example, the wordforms in Fig. 5 contain two positions for gender/number markers. In (1a), the neuter singular is realized as the infix -b- in both positions. In (1b), the masculine singular is realized as a null infix in the first position and the infix -j- in the second position.

Taking into consideration all the morphonological variations in the paradigms and the number of gender/number positions there are 40 different paradigm types for verbs (morphological classes). 19 of the strategies include irregular forms in the masculine class: the neuter marker b+V (where V is the subsequent vowel) is replaced by *w, j*, or a null affix. For this particular case, we have to provide separate entries for the verbs in the masculine class.

The rarest paradigms 'vclass_double_v' and 'vclass_trio' were not found in texts. A verb belonging to the 'vclass_trio' paradigm type has three slots for inserting a gender/number marker (as in *бетбяхIбарес* 'to send

(1) a. амчI<б>ирескa<б>ир-ес
go.bald<N>-INF
b. амчIирескa<й>р-ес
go.bald<M>-INF

Fig. 5. An example of verb form with two class affixes

there' the stem field is 'stem:*.em.яxi.ap.*') and includes two patterns for marking masculine: 1) the masculine marker is *-w-* in all three positions; 2) the masculine marker is *-w-* in the first two positions and zero in the third. The vclass_double_v paradigm has two slots for a class marker and unites two patterns for masculine: 1) the first position is *-w-* and the second is *-ww-*; 2) both positions have *-w-*.

The most frequent paradigm is 'vclass_bdel' which has a zero marker for masculine (4,981 verbs in the dictionary). The rarest paradigm among those attested in the texts is 'vclass_double': it has geminate markers for neuter, feminine, and masculine, and *-rd-* for plural. For example, the verb *асарчеббарес* 'to strengthen' has the following stem field: stem: '*асарче.ap.*'

V. Improving the Dargwa Uniparser-morph

After the first iteration of the parser testing, we analyzed the results. The following types of errors were taken into consideration: (a) overgeneration, especially for most frequent words, (b) words containing segments absent in our paradigms file, (c) impossible morpheme sequences provided by the parser.

A. Overgeneration

For many wordforms, the parser suggests too many variants. Some of the generated annotations are impossible for a particular lexeme. Some of the annotations are possible, but too rare. On the one hand this is an expectable effect for rule-based parsing, usually, there are a lot of short grammatically ambiguous morphemes. Besides, basic verb-stems in Daghestanian language have simple one-syllable root, many of them consist of one or two letters.

The first type of mistake is due to the parsing procedure: the parser uses the regular paradigm file to parse a wordform. It is not able for the parser to use additional information from the dictionary. One of the most frequent mistakes is that the parser split a lexeme according to the hypothesis that a certain segment of a wordform could be a stem for the plural number formation with a regular number affix. For example, the wordform *декIар* is parsed as the plural form of the noun *декI* 'patch', while the dictionary suggests another plural form for it, namely, *декIни*. We have a case of erroneously assigned lemma for this wordform, this annotation should be deleted. To deal with this type of error a script was created for post-editing. It deletes an erroneous analysis if there is an irregular plural form in the dictionary.

The other case is the overgeneration for the most frequent wordforms. Most of the generated variants are very rare. This type of mistake needs to be corrected manually. We singled out the 100 most frequent wordforms. For the majority of them, it was possible to have the only one or two possible morphological analyses. However, for some lexemes from the list it is impossible to choose the only one most probable annotation. For instance, the wordform *sabi* has four quite probable annotations: it can be the pronoun 'self' or the verb 'to be'. The infixed *-b-* class marker is a class marker for neuter or for human plural gender.

B. Unparsed wordforms

The unparsed forms can be divided into the following classes: (a) a wordform contains a morpheme absent from our morpheme lists, (b) a wordform is a compound with a borrowed stem, (c) a wordform is a compound absent in the dictionary, while the first and the second parts are present in it as separate words, (d) other cases.

There are also some derivational morphemes in Dargwa that are regularly attached to a certain class of stems. For example, the

causative morpheme can be attached to almost any verb stem. Causative derivatives may be absent from the dictionary. In this case, the affix is added to the paradigm dictionary. A more difficult case is one of the negative morphemes. It can be inserted into different places in a wordform. This case remains unresolved.

Daghestanian languages have a rich system of particles/affixes. The latter are easily cliticized to any wordform. Usually, there is no exhaustive lists of such elements in grammars or dictionaries. This list can be built via several iterations of parsing. In Dargwa, there are many particles and morphemes forming temporal converbs (e.g., as.soon.as, when, by.time. of, etc.). They are freely cliticized to wordforms (ваибмадан – *в-а-иб-мадан* – m-stem-aor-as.soon.as, it is a converb form of the lexeme meaning 'to achieve, to arrive'). After testing the parser, we added a list of such elements to the basic list of morphemes.

An interesting class of cases is when a Russian word is used with Dargin affixes. The borrowings are formed using the full lexeme form (and not its stem, c.f. *станкостроение-ла*, where *-ла* is an attributivizer or a parker of Genitive). There are a lot of verbs derived according to this scheme in our corpus: c.f. *защитить-б-ар-ес* – protect-N-do, where *защитить* is an infinitive.

The (d) class of cases is when an existing in the dictionary stem is used with a light verb stem, as in ***белги-****б-ар-ес* 'definiteness-N-do', the first part is in the dictionary and the verb is not included there. This type of cases needs a special module for compound-word detection using different noun, adjective or adverb stems from the dictionary.

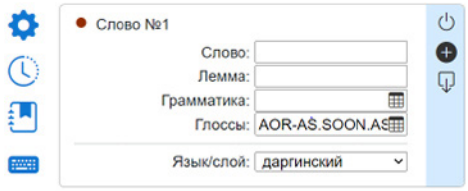
There are also cases of regular ambiguity of morphemes: *-la* is a suffix for Genitive, Locative and attributive form or *-b-* is a class marker for Neuter and for HumanPlural.

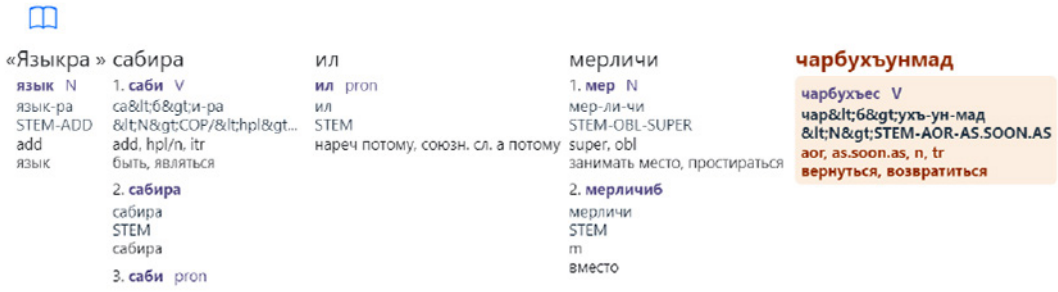
C. Impossible chains of morphemes

A separate case of overgeneration is when the system suggests an impossible sequence of morpheme tags. For instance, a tag for a noun case suggested after a verb stem, or a stem tag follows a tag for verb tense category which is an impossible sequence. This case can be overridden by enumerating different impossible tag sequences and deleting the corresponding analyses.

VI. Corpus functionality

The corpus is available for linguistic research through an online interface http://lingconlab.ru/standard_dargwa/search. One can search by wordforms, lexemes, translation, and glosses. Regular expressions are available. One can combine conditions for different layers in one search. The information on wordforms and lemmas frequencies is also available. See Fig. 6 for example of corpus query for concordance and for the result of word/lemmas mode.





Результат поиска: найдено 25 разных словоформ примерно в 40 документах, суммарная частотность: 49.

слово	лемма	грамм.	частотность					
каибмад	каэс	aor, as.soon.as, V, tr	2	> 50%	2	2		
ахъбуцибмад	ахъбуцес	aor, as.soon.as, n, V, tr	2	> 50%	2	2		
дурабухъунмад	дурабухъес	aor, as.soon.as, n, V, tr	1		1	1		
хіеризурмадан	хіербизес	aor, as.soon.as, m, V, tr	2	> 50%	2	2		
таманбарибмад	таманбарес	aor, as.soon.as, n, V, tr	2	> 50%	2	2		
багьурмадан	багьес	aor, as.soon.as, n, V, tr	1		1	1		
чеибмад	чеэс	aor, as.soon.as, V, tr	2	> 50%	2	2		
хіеръибмадан	хіерэс	aor, as.soon.as, V, tr	2	> 50%	2	2		
тамандиубмадан	таманбиэс	aor, as.soon.as, pl, V, tr	2	> 50%	2	2		
хъараахъурмад	хъараэс	caus, aor, as.soon.as, V, tr	2	> 50%	2	2		

Fig. 6. An example of a corpus search: (a) gloss query, (b) a sentence hit, (c) words and lemmas information

VII. Conclusions

We present a new corpus on an under-resourced language, namely, Standard Dargwa. We suggest a pipeline for corpus construction with no previous resources provided. The suggested pipeline provides a quick start. That is the usage of Tsakorpus as a corpus platform and of Uniparser-morph for building a parser for interlinear gloss annotation. We have discussed some peculiar cases of Dragwa morphology (interffix agreement morphemes, clitics etc.) and some ways of overcoming the parsing problems. Besides the corpus we also present the first rule-based morphological parser for Standard Dargwa.

References

Toldova S., Sokur E. Standard Dargwa Corpus. In Iarmakeev I. E., Kharasova F. Kh. (eds.) Sovremennaia lingvistika: ot teorii k praktike. Trudy I materialy III Kazanskogo mezhdunarodnogo lingvisticheskogo sammita. [The Proceedings of Kazan International Linguistic Summit 2022. Kazan, 16 November 2022]. Kazan', 2023, 1. 81–86.

Arkhangelskii T. Printsipy postroieniia morfologicheskogo parsera dl'a raznostrukturnykh jazykov [Principles of building a morphological parser for different-structure languages] Thesis cand. phil. sci. Moscow. Manuscript. 2012

Iusupov H., A. Darginsko-russkij slovar'. [Dargwa-Russian dictionary]. Moscow. Publishing House "Pero", 2017.

Gasanova S. M. Ocherki darginskoi dialectologii [Essays on Dargin dialectology]. Makhachkala: Dagestankii Filial Akad.Nauk SSSR. 1971

Abdullaev S. N. Grammatika darginskogo jazyka: Fonetika i morfologija [Grammar of the Dargin language: Phonetics and morphology]. Makhachkala: Dagestanskij Filial Akad. Nauk SSSR. 1954

Van den Berg H. Dargi Folktales. Oral Stories from the Caucasus and an Introduction to Dargi Grammar. Leiden: CNWS. 2001

Musaev M.-S., M. Darginskii iazyk. In Alekseev M. E. (red.). Iazyki mira. Kavkazskie iazyki. [The languages of the worls. The languages of the Caucasus] M. 1999

Himmelmann N. Documentary and descriptive linguistics. 1998.

Comrie B., Haspelmath M., Bickel B. "The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses." Max Planck Institute for Evolutionary Anthropology, Leipzig. Linguistics 2008, 36.161–195.

Goodman M. W., Crowgey J., Xia F., Bender E. M. Xigt: extensible interlinear glossed text for natural language processing. LREC, 2015. 49(2). 455–485.

Xingyuan Zh., Ozaki S., Anastasopoulos A., Neubig G., Levin L. S. Automatic Interlinear Glossing for Under-Resourced Languages Leveraging Translations. *In International Conference on Computational Linguistics*. 2020

Scannell K. P. The Crúbadán Project: Corpus building for under-resourced languages. *In Building and Exploring Web Corpora (WAC 3–2007): Proceedings of the 3rd Web as Corpus Workshop, Incorporating Cleaneval.* 2007, 4. 5.

Arkhangelskiy T. Corpora of social media in minority Uralic languages. *In In Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages.* 2019, 125–140.

Lyashevskaya O. N., Sharoff S. A. Frequency dictionary of modern Russian based on the Russian National Corpus [Chastotnyy slovar' sovremennogo russkogo jazyka (na materiale Nacional'nogo korpusa russkogo jazyka)], Azbukovnik, Moscow. 2009

Arkhipov A. INEL corpora general transcription and annotation principles. *In Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology* 5). Szeged & Hamburg: Department of Finno-Ugric Studies of the University of Szeged & Hamburger Zentrum für Sprachkorpora der Universität Hamburg. 2020. https://doi. org/10.14232/wpcl.