# Universal Dependencies for Dargwa Mehweb

**Alexandra Kozhukhar**
National Research University Higher School of Economics
Faculty of Humanities
School of Linguistics
Russia
sasha.kozhukhar@gmail.com

## Abstract

The Universal Dependencies (UD) project aims to create the unified annotation schemes across languages. With its own annotation principles and abstract inventory for parts of speech, morphosyntactic features and dependency relations, UD aims to facilitate multilingual parser development, cross- lingual learning, and parsing research from a language typology perspective. This paper provides the description for the way Dargwa Mehweb (East Caucasian language family) meets UD scheme.

## 1 Introduction

The Universal Dependencies (UD) (Nivre et al., LREC 2016) is a project dealing with consistent cross-linguistic morphological and syntactic mark-up. The UD is currently in version 2 and covers 52 languages with 10 more languages yet to be included. While UD covers 11 language families, it does not include languages of the Caucasus and, in particular, East Caucasian languages.

The guidelines of UD are based on the Google Universal Part-of-Speech Tagset (Petrov et al. 2012) for parts of speech, the Interset framework (Zeman 2008) for morphological features, and Stanford Dependencies (De Marneffe et al. 2006, De Marneffe et al. 2014) for syntactic relations. The approach aims at making typological features of the languages of the world scalable and at simplifying the cross-linguistic comparison.

The paper is structured as follows: Section 2 provides general information about the Dargwa Mehweb language – area of distribution, number of speakers, some sociolinguistic data and a short grammar overview; Section 3 describes part of speech mapping; Section 4 discusses relevant features of Dargwa Mehweb; Section 5 explains the syntactic dependencies of Dargwa Mehweb in terms of the UD approach; Section 6 discusses the cases of the language change and grammaticalization; and Section 7 presents the conclusions.

## 2 Dargwa Mehweb: General Information

Mehweb belongs to Dargwa group of East Caucasian language family. It is often considered a dialect of Dargwa (Magometov 1982), although according to (Koryakov & Sumbatova 2004, Khaidakov 1985) Mehweb is a separate language rather than a dialect. For the following research we consider Mehweb a separate language. According to lexicostatistic analysis, Mehweb is a member of the north-central group of Dargwa languages (Koryakov 2013).

Dargwa Mehweb is spoken in the village Megeb (Republic of Daghestan, Russian Federation). The language is spoken by approximately 800 people but is not quite at risk of becoming endangered (Dobrushina et al. 2017). Megeb is the only Dargwa village in the area surrounded by Lak and Avar villages. The official language of the village is Avar, children are taught Avar in school. Most of the speakers are bi- or trilingual (Kozhukhar & Barylnikova 2013).

Dargwa Mehweb was first mentioned in (Uslar 1892). There are two reference grammars of Mehweb (Magometov 1982, Khaidakov 1985), both published in 1980s, and selected essays on different aspects of the Mehweb grammar (Dobrushina et al. 2017) published in 2017. Dictionaries and texts were obtained during field trips to Megeb in 1990s and 2010s organized by Lomonosov Moscow State University and Higher School of Economics.



*Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 92-99,
Pisa, Italy, September 18-20 2017

Mehweb is notliterate. Native speakers use Avar orthography that does not match completely the phonemic inventory of Mehweb. During the field trips in 2010s a new orthography based on the IPA was introduced. All previous texts were converted into new orthography. In the following paper we use the orthography invented in 2010s.

In terms of its features Dargwa Mehweb is a typical East Caucasian language. Mehweb demonstrates agglutinative morphology. Mehweb is ergative in terms of agreement and case marking. There are five non-spatial cases in Mehweb. Spatial forms are bimorphemic: the first morpheme defines the spatial domain (‘on’, ‘near’, ‘in’ etc.) and the second one defines the orientation (Goal, Source, Path). There are no adpositions in Mehweb – all spatial relations are expressed with spatial cases and directional prefixes on verbs. Most of the verbs have perfective and imperfective stems from which all the verbal forms are derived. The formal relation between the stems is irregular and involves alternations, infixation and loss of class agreement slots. Most of the verbs bear a class agreement marker referential to the absolutive argument of the clause. Class agreement distinguishes feminine, masculine and neuter in singular, human and non-human in plural. Mehweb is a typical East Caucasian language with basic SOV order.

Most of the East Caucasian languages, and Mehweb as well, allow using non-finite forms as heads of simple clauses. Clausal coordination is encoded by joining clauses headed by non-finite verb forms and the matrix clause. Apart from citation and reported speech contexts, where finite verb form is obligatory, all subordinate clauses bear non-finite verb forms such as action nominals, infinitives, participles and converbs. There is no clausal coordination in Mehweb. Mehweb has reflexive pronouns which can also be used in logophoric function.

In the following paper we use Leipzig Glossing Rules (Comrie et al. 2008) to indicate grammatical features in the examples. The list of abbreviations used in the paper is given in Appendix A.

## 3 POS Mapping

POS mapping is simple since there are exclusively verbal (participle morpheme, TAM markers etc.) and exclusively nominal morphemes (number, case). Cross-categorial morphemes are considered clitics (coded as PART), for example, additive clitic *=ra*, which functions as CCONJ ‘and’, class markers or emphatic clitic *=al*, which can be combined with pronouns and numerals.

Most of the adjectives in Mehweb are marked with attributivizing morpheme *-(i)l*. There is a closed set of adjectives that also bear a class marker of the head.

(1) *ħunt’a-l qul-le-šu*
red-ATR house-PL-AD(LAT)

(2) *har-il* *urši-li-s*
each-ATR boy-OBL-DAT

(3) *r=igu-l*
F1-engaged-ATR

Adverbs derived from adjectives are marked with adverbalizing morpheme *-le*. All other adverbial meanings, especially spatial ones, are expressed with verbal prefixes or spatial cases. Mehweb has auxiliaries, they are used with adjectives in predicative position (*sa=b=i*) and analytical progressive verb forms (*le=w*). Mehweb uses negative copula *agwara* with affirmative auxiliaries.

There are four deictic pronouns that are mapped as DET. DETs can also be used as personal pronouns (cf. Table 1).

| **Meaning** | **Pronoun** | **Meaning** | **Pronoun** |
|---|---|---|---|
| ‘near the hearer’ | il | ‘higher than hearer’ | ič’ |
| ‘far from hearer’ | it | ‘lower than hearer’ | iχ |

Table 1: Mehweb deictic pronouns

There are special pronouns that function as wh-words, for example, *sik’al* meaning ‘what’. Wh-words can be used in affirmative sentences as well as in interrogative ones.

There are no CCONJ and SCONJ in Mehweb since all the subordinates are encoded by non-finite verb forms. Thus ‘Ali-[ERG] hit-[finite] Fatima run away-[converb]’ would mean that ‘Ali hit Fatima and ran away’ but not that ‘Running away Ali hit Fatima’.

In Mehweb CCONJ and SCONJ are distinguished by the type of the converb used in the subordinate clause – for CCONJ general converbs are used, for SCONJ, specialized ones. We can assume that Mehweb had SCONJs since the texts from 40 years ago contain a special conjunction, but contemporary texts lack this conjunction. For further information see Section 6.

Table 2 gives the overview on the POS mapping for Mehweb.

| POS | Mehweb | POS | Mehweb |
|---|---|---|---|
| ADJ | + | CCONJ | – |
| ADV | + | DET | + |
| NOUN | + | NUM | + |
| VERB | + | PART | + |
| ADP | – | SCONJ | – |
| AUX | + | PRON | + |
| INTJ | + | PROPN | + |

Table 2. Overview on the parts of speech in Mehweb

## 4 Features Mapping

The following section deals with feature mapping for Dargwa Mehweb. Subsection 4.1. deals with nominal features; subsection 4.2. covers verbal features.

### 4.1 Nominal Features

Animacy feature with values `Hum` and `Nhum` is used for class agreement in plural since Mehweb does not distinguish female, masculine and neuter in plural. Table 2 presents clitics marking class (gender) in Mehweb. Gender feature with values `Fem`, `Masc` and `Neut` are used in singular.

| | Sg | Pl | |
|---|---|---|---|
| **M** | *w* | *b* | **HPL** |
| **F** | *d* | | |
| **F1** | *d-r* | | |
| **N** | *b* | *d-r* | **NPL** |

Table 3: Class markers in Mehweb

Case feature is used to distinguish between cases. There are five non-spatial cases, namely, absolutive, ergative, comitative, dative, genitive. Vocative case is formed by stress shift (Dobrushina et al. 2017). Spatial cases consist of two morphemes. We propose our own mapping for spatial cases, where each spatial form has marker of Localization and Orientation:

Localizations: Super ('On'), In ('In'), Inter ('Inside'), Apud ('Near');

Orientations: Lat ('Move towards'), Ess ('Staying in place'), El ('Move away from'), Trans ('Moving through').

All orientation markers are expressed by special suffix except the lative, which is zero in Mehweb. It worths noting that zero lative is very rare in East Caucasian languages. Usually it is the essive which is expressed by zero.

There are only singular (coded as `Sing`) and plural (coded as `Plur`) number. There are also *pluralia tantum* nouns in Mehweb.

The Person feature is polarity dependent. There is a distinction between the first person singular and second person singular: in affirmative clauses suffix -ra marks the first person singular, whereas in negative and interrogative clauses the same suffix marks second person singular. Polarity feature thus has to be coded using three values instead of two: `Pos`, `Neg` and `Interrogative`.

Mehweb pronouns can be coded using the following values: `Dem`, `Rcp`, `Int`, `Neg`, `Ind`. There are no separate personal pronouns in Mehweb — demonstratives (in matrix clauses) and reflexives (in subordinate clauses) are used instead.

There is a set of ordinal (coded as `Ord`) and cardinal numerals (coded as `Card`). Cardinal numerals are derived from ordinals using emphatic clitic *=al*.

### 4.2 Verbal Features

Aspect feature is used to distinguish between imperfective (coded as `Imp`) and perfective (coded as `Perf`) verb stems.

Mehweb demonstrates a wide range of moods, thus the Mood feature with the values `Ind`, `Imp`, `Cnd`, `Pot`, `Jus`, `Opt`, `Prp` is required. Some of the values of the Mood feature are encoded by special converbs, for example, apprehensive meaning is expressed by a separate morpheme, and it is still under discussion whether the values of the feature Mood should be used or it requires introducing some separate values.

All verb forms bear a tense marker: `Past`, `Fut`, `Pres`. The following verb forms are relevant for Mehweb: `Fin`, `Inf`, `Part`, `Conv`, `Vnoun`. There are also analytical verb forms which consist of an auxiliary and a general converb, for example, the for of progressive.

Table 4 gives an overview on the feature mapping in Mehweb.

| **Feature** | **Relevant Values** |
|---|---|
| Animacy | `Hum`, `Nhum` |
| Aspect | `Imp`, `Perf` |
| Case | `Abs`, `Erg`, `Com`, `Dat`, `Gen`, `Voc` |
| Foreign | `Yes` |
| Gender | `Fem`, `Masc`, `Neut` |
| Mood | `Ind`, `Imp`, `Cnd`, `Pot`, `Jus`, `Opt`, `Prp` |
| NumType | `Ord`, `Card` |
| Number | `Sing`, `Plur`, `Ptan` |
| Person | `1`, `2` |
| Polarity | `Pos`, `Neg` |
| Reflexive | `Yes` |
| Tense | `Fut`, `Past`, `Pres` |
| VerbForm | `Fin`, `Inf`, `Part`, `Conv`, `Vnoun` |
| Voice | `Cau` |
| PronType | `Dem`, `Rcp`, `Int`, `Neg`, `Ind` |

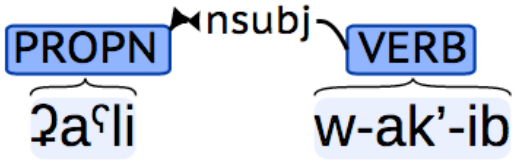
Table 4: Feature mapping in Mehweb

## 5 Syntactic dependencies mapping

Syntactic structure of East Caucasian languages, and Darwa Mehweb as well, differs significantly from languages described in terms of UD approach earlier. In this section we discuss some cases of dependency relations mapping for Mehweb.

In Mehweb as in a language with ergative alignment `nsubj` is marked with ergative case in transitive clauses and with absolutive in intransitive clauses. The causer is also marked with ergative.

```
text: ʔaʕli w-ak'-ib
gloss: Ali(nom) m-come.pfv-aor
text[eng]: 'Ali came'.
ʔaʕli NSUBJ PROPN
wak'ib ROOT VERB
```




In the example above the `root` of the clause is the verb *wak'ib* 'came' which is intransitive. Ali is `nsubj` marked with absolutive case. Verb also bears a class agreement marker w- referring to Ali, i.e. masculine and singular (see Table 3).

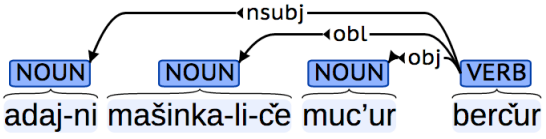
However, there are cases where Mehweb does not fit the ergative logic. In the sentence with transitive and intransitive verb forms conjoined, since there is no coordination in Mehweb, transitive and intransitive verb forms have ergative argument as `nsubj`. Thus 'Ali-[ERG] hit-[transitive] Fatima-[ABS] and ran away-[intransitive]' would mean that 'Ali hit and Ali ran away' and not that 'Ali hit Fatima and Fatima ran away'.

There is also a list of experiential verbs that have subject marked with oblique cases such as `Dat` and `Inter:Lat`. Such cases will be marked as `obl:exprnc`.

The `obj` is always marked with `Abs` including the cases with transitive experiential verbs, for example, verb *gwes* 'to see' marks

its nsubj with Inter:Lat and is obj with Abs. Transitive verbs get class agreement from OBJ. If the subordinate clause turns out to be a direct object of the verb, i.e. ccomp, the verb form gets Neut class agreement. iobj can be marked with oblique non-spatial cases and all spatial cases.

```
text: adaj-ni mašinka-li-če
muc'ur b-erč-ur
gloss: father-erg
hair.cutter-obl-super(lat)
beard(nom) n-cut.hair.pfv-aor
text[eng]: 'The father cut his
beard with a hair cutter'.
adajni NSUBJ NOUN
mašinkaliče OBL NOUN
muc'ur OBJ NOUN
berčur ROOT VERB
```

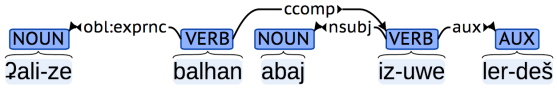



In the example above the root of the clause is transitive verb *berčur* 'shove'. Father ('adaj') is nsubj marked with ergative case. The verb bears an agreement marker *b-* (i.e. neuter, singular) referring to the beard ('muc'ur') which is in absolutive case, i.e. obj. The cutter ('mašinka-li-če') is marked with Super:Ess marker *-če* and a zero marker of lative, therefore it is considered an obl.

The relevance of csubj in Mehweb is under discussion since it was not directly tested yet. advcl and acl are the main strategies of the clause composition. In case subordinate clause is headed by the participle it is assigned acl label. In case subordinate clause is headed by the specialized or general converb it is assigned advcl. If the subordinate clause is headed by the infinitive or nomen actionis it is assigned xcomp. If the matrix verb bears a neuter singular class marker (*b-*) and there are no neuter singular obj in the matrix clause then b- is considered a subordinate clause agreement marker and the subordinate is assigned ccomp label.

In case of citation and indirect speech the cited phrase is headed by the finite verb. These types of clauses were labeled parataxis. However, cited phrases can be connected with matrix clauses by reflexive pronoun used in subject position in the cited phrase and co-referential to the subject of the matrix clause. Thus the parataxis label is under discussion.

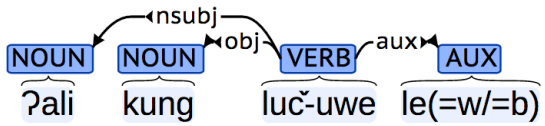
```
text: ʔali-ze b-alh-an abaj iz-
uwe le-r-deš
gloss: Ali-inter n-know:ipf-hab
mother(abs) be.sick:ipf-cvb
cop-f-nmlz
text[eng]: 'Ali knows that mother
is sick.'
ʔalize NSUBJ PROPN
balhan ROOT VERB
abaj NSUBJ NOUN
izuwe CCOMP VERB
lerdeš AUX AUX
```



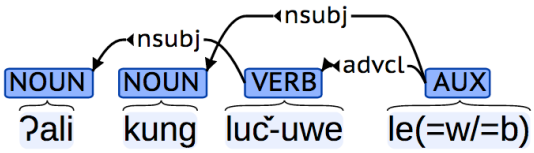

Example above consists of two clauses: *ʔalize balhan* 'Ali knows' and *abaj izuwe lerdeš* 'mother is sick'. The root of the whole sentence is an experiential transitive verb balhan 'knows'. Since balhan is an experiential verb its obl:exprn is marked with interlative, i.e. Inter:Lat, case. The direct object of the root verb is a second clause abaj izuwe lerdeš, thus it is labeled ccomp. Root verb bears the class agreement marker *b-* (neuter, singular) which refers to the subordinate clause. The head of the second clause is an auxiliary verb lerdeš (being). Auxiliary verb bears the class agreement marker *-r-* (feminine, singular) referring to the nsubj of the ccomp mother ('abaj') and a nominalization suffix *-deš* since the second clause is subordinate and has to be non-finite. Subordinate clause contains also a converb *izuwe* 'being sick' which is a part of an analytic progressive form izuwe *lerdeš* standing for 'is being sick'. Converb *izuwe* is labeled VERB.

Mehweb does not allow to use more than one auxiliary in a single analytical verb form. However, Mehweb demonstrates so-called biabsolutive constructions with analytical verb forms:

(4) ʔali kung
Ali(ABS) book(ABS)
luč-uwe le(=b/=w)
read.IPF-CVB aux(=N/=M)
'Ali is reading a book'.




The preceding tree is impossible since Ali must be in ergative case in order to be considered `nsubj` of the transitive verb 'read'.




We propose treating biabsolutive constructions as if they were `advcls` since they turn out to be two clauses: matrix clause headed by copula and subordinate clause headed by converb. Therefore (4) will be literally translated as 'The book is so that Ali reads it'.

Table 5 gives the overview on the syntactic dependencies mapping.

| UD | Mehweb |
|---|---|
| nsubj | + |
| obj | + |
| iobj | + |
| csubj | not attested |
| ccomp | + |
| xcomp | + |
| obl | + |
| vocative | + |
| expl | – |
| acl | + |
| advcl | + |
| advmod | + |
| aux | + |
| cop | + |
| mark | – |
| nmod | + |
| det | + |
| clf | – |

Table 5: Dependency relations (UD 2.0) as attested in Mehweb

Mehweb does not have mark tags since the clauses headed by converbs are encoded instead, marked as `advcl`. Mehweb also lacks expletive subject and classifiers. There is no `nmod` since there are no prepositional groups (the genitive case is used instead).

## 6 Language change and grammaticalization

Language material we are basing on is heterogeneous since half of the texts are contemporary and half of them is 40 years old. These two groups of texts demonstrate some differences in how conjunction is expressed. This was considered a case of a language change.

Texts from (Magometov 1982) have a special autonomous word form *wa* which functions as 'and' and is used as a conjunction between nouns and clauses. Contemporary texts lack *wa*. Instead of *wa* `acls` and `advcls`, i.e. subordinate clauses headed by non-finite verb forms, are used to conjoin clauses and *=ra* clitic is used to conjoin nouns. For cases with *wa* SCONJ POS label and `conj` dependency relation label are used, although conjoining clauses and nouns with *wa* is considered unnatural for Mehweb.

The following case is considered a grammaticalization of an non-finite verb form. In Mehweb when matrix clause is headed by the speech verb, e.g. 'say' or 'know', converb *ile* 'being said' occurs in the end of the subordinate clause. For majority of native speakers *ile* is optional. Some native speakers recognize *ile* as an autonomous verb form, whereas others do not provide a subtle translation for it.

We provide two ways of treating *ile*. The first one is treating it as an optional citation particle, i.e. PART. The second one is consid-

ering ile a separate `advcl` since it is a converb. Cited phrase that precedes *ile* then is considered its `ccomp` in that case.

## 7 Conclusion

In this paper we described how well the UD approach covers the features of one of the typical East Caucasian language, Dargwa Mehweb. Three types of guidelines were applied: POS mapping, feature mapping and dependency relations.

Some features were raised due to mapping. First is the way grammaticalization cases should be treated since there are more than one possible way of representing them on a dependency tree. Second is a clausal conjunction since Dargwa Mehweb use the sequence of non-finite clauses instead of expected SCONJs. Third is a lack of difference between moods and special converbs since there are special mood markers that can be combined with converbs only.

## References


Bernard Comrie, Martin Haspelmath and Balthasar Bickel. 2008. *The Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses.* URL: https://www.eva.mpg.de/lingua/resources/glossing-rules.php

Marie-Catherine de Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. *Generating typed dependency parses from phrase structure parses.* In Proceedings of 5th International Conference on Language Resources and Evaluation (LREC), volume 6, pages 449– 454.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. *Universal Stanford Dependencies: a cross-linguistic typology*. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Nina Dobrushina, Michael Daniel, Dmitry Ganenkov, George Moroz , Daria Barylnikova, Marina Kustova, Alexandra Kozhukhar, Yuri Lander, Maria Sheyanova and Ilia Chechuro. 2017. *Mehweb. Selected essays on phonology, morphology and syntax.* Berlin: Language Science Press.

Said Khajdakov. 1985. *Darginskij i megebskij yazyki (printsipy slovoizmeneniya) [Dargwa and Mehweb languages]*. Makhachkala.

Yuri Korjakov and Nina Sumbatova. 2007. *Darginskie jazyki [The Dargwa Languages]* In BRJe, tom 8. Moskva: Bol'shaja rossijskaja enciklopedija, pages 328–329.

Yuri Koryakov. 2013. *Convergence and divergence in the classification of Dargwa languages*. In 46th Annual Meeting of the Societas Linguistica Europaea (SLE 2013). 18–21 September 2013. Book of abstracts. Part 1. Split: University of Split.

Alexandra Kozhukhar and Daria Barylnikova. 2013. *Multilingualism in Dagestan.* Higher School of Economics Research Paper No. WP BRP, 4.

Alexander Magometov. 1982. *Megebskij dialekt darginskogo jazyka (Issledovanie i teksty) [The Mehweb Dialect of Darwga Language]*. Tbilisi: Mecniereba.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. *Universal dependencies v1: A multilingual treebank collection.* In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016).

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. *A universal part-of- speech tagset.* In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), pages 2089–2096.

Petr Uslar. 1892. *Etnografija Kavkaza. Jazykoznanie [Ethnography of the Caucasus]*. V. Hjurkilinskij jazyk, Tiflis.

Daniek Zeman. 2008. *Reusable tagset conversion using tagset drivers.* In Proceedings of

the 6th International Conference on Language Resources and Evaluation (LREC), pages 213–218.

## Appendix A. List of abbreviations

| | |
|---|---|
| ABS | absolutive |
| AD | localization ad |
| AOR | aorist |
| ATR | attributivized |
| AUX | auxiliary |
| CVB | converb |
| DAT | dative |
| F | feminine |
| F1 | feminine |
| HPL | human plural |
| IPF | imperfective |
| LAT | lative |
| M | masculine |
| N | neuter |
| NMLZ | nominalization |
| NPL | non-human plural |
| OBL | oblique |
| PFV | perfective |
| PL | plural |