Building a Universal Dependencies Treebank for a Polysynthetic Language: the Case of Abaza

Alexey Koshevoy[1, 2, *], Anastasia Panova[3], and Ilya Makarchuk[4]

[1]Laboratoire de Psychologie Cognitive, Aix-Marseille University, CNRS, Marseille, France
[2]Institut Jean Nicod, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, Paris, France
[3]Department of Linguistics, Stockholm University, Stockholm, Sweden
[4]Linguistic Convergence Laboratory, HSE University, Russian Federation
[*]Corresponding author: alexey.koshevoy@univ-amu.fr

Abstract

In this paper, we discuss the challenges that we faced during the construction of a Universal Dependencies treebank for Abaza, a polysynthetic Northwest Caucasian language. We propose an alternative to the morpheme-level annotation of polysynthetic languages introduced in Park et al. (2021). Our approach aims at reducing the number of morphological features, yet providing all the necessary information for the comprehensive representation of all the syntactic relations. Besides, we suggest to add one language-specific relation needed for annotating repetitions in spoken texts and present several solutions that aim at increasing cross-linguistic comparability of our data.

1 Introduction

The Universal Dependencies (UD) framework (de Marneffe et al., 2021) provides a cross-linguistically universal annotation scheme, which, when necessary, allows language-specific extensions. The need for language-specific extensions becomes particularly evident when building a treebank for languages with polysynthetic morphology (on the features commonly considered polysynthetic, see, e.g., Mattissen (2017, p. 71-73)). For example, to account for nominal incorporation in Chukchi, Tyers and Mishchenkova (2020) introduce additional nodes for incorporated arguments of predicates.[1] In the treebank of St. Lawrence Island Yupik, Park et al. (2021) treat each morpheme as a token which requires seven additional dependency relations under the unspecified `dep` relation (`dep:infl`, `dep:aux`, etc.). In both cases the authors strive to create a more comprehensive representation of a polysynthetic language in UD, but the resulting annotation is not consistent with some of the UD principles. The UD approach assumes that the basic units are words (de Marneffe et al., 2021, p. 259), therefore, to achieve greater consistency, even in polysynthetic languages the word-based analysis should be favored over the morpheme-based analysis. In addition, we believe that such radical adjustments to UD as proposed for Yupik (Park et al., 2021) may be justified for one specific language (or one specific family) but generalization of this approach to all polysynthetic languages can be avoided. In this paper, we introduce a new way of dealing with polysynthetic morphology in UD and show that our approach aligns well with the UD framework. Specifically, we discuss the annotation of the treebank of Abaza, a polysynthetic Northwest Caucasian language.

2 Abaza

Abaza (ábaza bəzŝá, ISO 639-3: `abq`) is a polysynthetic language belonging to the Northwest-Caucasian family. According to the 2010 Russian census, it is spoken by approximately 38.000 speakers in the Karachay-Cherkes Republic, Russia. Additionally, Chirikba (2012) estimates that there are approximately 10.000 speakers of Abaza in Turkey. Abaza has two distinct variants — Tapanta and Ashkharywa. The data introduced in this paper comes from the Tapanta variant. Abaza has a rich written tradition, which has developped during the Soviet period. Abaza has a writing system that consists of a modified Cyrillic alphabet with the addition of one grapheme (I, "palochka"), which is used to indicate ejective consonants.

The polysynthetic nature of Abaza manifests itself mostly in verbal morphology. Abaza has a rich system of prefixes that are used for cross-referencing up to four verbal arguments. These prefixes indicate the number, person, gender, and grammatical role of each argument. As shown in Table 1, the personal prefixes form two distinct series – absolutive and oblique (including both ergative and indirect object markers). If the hearer

[1]See the same solution for Nahuatl in Pugh et al. (2022, p. 5018).

Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023), pages 1 - 6

can recover the subject, object, and indirect object from context, they don not need to be overtly expressed by independent nominals.

	absolutive	oblique		absolutive	oblique
1sg	s(ə)-/z-		1pl	h(ə)-/ʕ-	
2sg.M	w(ə)-		2pl	ŝ(ə)-/ẑ-	
2sg.F	b(ə)-/p-		3pl	j(ə)-/∅	r(ə)-/d(ə)-
3sg.M	d(ə)-	j(ə)-	Rel	j(ə)-	z(ə)-
3sg.F		l(ə)-			
3sg.N	j(ə)-/∅-	a-/na-			

Table 1: The system of verbal cross-referencing prefixes (adapted from Arkadiev (to appear)).

In addition to cross-referencing prefixes, there are more than a dozen types of affixes that can be attached to the verbal form. These include temporal markers, voice markers, markers of negation, locative affixes, etc. The ordering of those affixes is shown in Table 2.

Although the basic word order is SOV, some variation is allowed. For instance, there are cases of arguments appearing in the postverbal position. In example (1), the absolutive subject *a-waʕá* (def-people) occurs in the rightmost position in the clause.[2]

(1) abar-awəj a-pš-ta
EMP-DIST 3SG.N.IO-similar-ADV
j-bzaza-kʷa-d a-waʕa
3PL.ABS-live-PL-DCL DEF-people
'Thus lived the people.'

3 Spoken corpus of Abaza

The treebank presented in this paper is based on data from the Spoken Corpus of Abaza.[3] This corpus was built using the `tsacorpus` platform (see Arkhangelskiy (2020) for a brief description of the platform). It contains 25 spoken texts recorded from 8 different speakers. The recordings were made in the village of Inžič-Čukun in the Karachay-Cherkess Republic, Russia, between 2017 and 2019. The total duration of recorded data is approximately one hour.

The texts contained in the Spoken Corpus of Abaza were initially transcribed using the Abaza orthography. Further, these transcriptions were converted into an IPA-based transcription. The participants of the Abaza research group provided interlinear glosses for each text based on the translations obtained from the speakers of Abaza. The texts were annotated using the ELAN software (Wittenburg et al., 2006), therefore each sentences is aligned with a corresponding audio segment.

4 Preprocessing

We devised a specific pipeline to convert ELAN files into ten-column CoNLL-U format. We started by extracting the glossing abbreviations from the interlinear annotations. As the corpus uses an idiosyncratic notation for glosses, we have created a mapping between the interlinear glosses from the corpus and the corresponding morphological features compatible with the CoNLL-U format. We then used the script[4] written by Francis Tyers to convert the cleaned sets of Abaza morphological features into UD morphological features using our mapping. The untransformed glosses were also preserved to be included in the MISC section of the CoNLL-U format. As the final step, we manually added lemmas to each wordform in the CoNLL-U annotations.

We choose to use the original Cyrillic-based orthography instead of the phonological transcription in our treebank for the following reason. The Abaza orthography is used in non-annotated texts available online: newspapers and works of fiction. Since we plan to train an automatic parsing model to annotate more data for this treebank, the model needs to be trained on the data which has the same orthography as in the texts that it will be used to annotate.

5 Morphology

Many of the categories expressed by affixes in Abaza cannot be easily converted to UD annotation. First, verbal forms often have locative prefixes, which specify the meaning of the root. For example, in (2) the verbal root on its own means 'to fly', but with the addition of the locative prefix *tə-* 'out', its meaning changes to 'fly away.'

(2) a-warba a-ʕʷara j-tə-pssʕa-ṭ
DEF-eagle DEF-nest 3SG.N.ABS-loc:out-fly-DCL
'The eagle flew out of its nest.' (Klyčev, 1994, p. 140)

Second, Abaza verbs can be modified by so-called event operators, which express aspectual meanings or modify the Aktionsart (lexical aspect) of the predicate, cf. two repetitive markers in (3).

[2] The list of abbreviations for glosses is provided in the Appendix.

[3] http://lingconlab.ru/spoken_abaza/

[4] https://github.com/ftyers/ud-scripts/blob/master/conllu-feats.py

-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7
absolutive	subordination, negation	repetitive	potential, involuntative	applicatives	directional preverbs	locative preverbs	indirect object	ergative	negation	causative	sociative	root	directional suffixes	event operators	plural	aspect, tense	negation	past tense, modality	subordinators, force, emphasis

Table 2: The Abaza verbal template (adapted from Arkadiev (to appear)).

(3) h-ata-də-r-ca-χ-wa-n
1PL.ABS-rep-3PL.ERG-CAUS-go-re-IPF-PST
'they would again turn us [back]'

There is more than a hundred of different locative preverbs and more than fifteen event operators in Abaza (Klyčev, 1994; Tabulova, 1976, p. 204-215), so it is impossible to encode all these morphemes in an exhaustive set of UD morphological features. In addition, more than one locative preverb or event operator can occur in a wordform, therefore each of those affixes would require a separate feature. Finally, many locative prefixes and event operators are fully productive, so they cannot be easily attributed to derivation and thus ignored.

One way of dealing with this kind of morphology consists in treating all affixes as independent lexical items, as was suggested for St. Lawrence Island Yupik (Park et al., 2021), but this goes against the lexicalist approach adopted in the UD framework. Here we want to propose an alternative solution for the Abaza treebank. We suggest to keep the information about the meaning of each morpheme in the MISC section of CoNLL-U format. That way, it can still be available to the researchers that want to examine our data. At the same time, in the FEAT section we retain only those morphological features that are relevant to syntactic structure. Specifically, we decided to limit the grammatical features of Abaza encoded in UD to argument cross-referencing, valency-changing operations (reflexive, causative), finiteness, tense, mood, interrogativity and polarity. For example, compare the sentence (4) and its annotation in CoNLL-U format in Figure 1.

(4) s-pa a-q̇arč'a-k^{w}a
1SG.PR-son DEF-Karachai-PL
də-r-š'ə-n
3SG.H.ABS-3PL.ERG-kill-PST
d-sə-z-ʕa-r-g-χ-ṭ
3SG.H.ABS-1SG.IO-BEN-CSL-3PL.ERG-carry-RE-DCL
'the Karachays killed my son and brought [him to me]'

Example (4) contains two verbal forms — *də-r-š'ə-n* 'they killed him' and *d-sə-z-ʕa-r-g-χ-ṭ* 'they brought him to me'. Both verbs have cross-referencing markers which allow to identify syntactic relations between the predicates and their arguments. Final suffixes on verbal forms cumulatively express tense, mood and finiteness, and they are crucial for understanding the syntactic status of the predicate in the clause. The rest of the information present in the glosses of verbal forms — the cislocative prefix and the repetitive suffix — is not included in the FEAT section of the annotation because it is not relevant to syntax.

6 Syntax

In this paper, we propose to constrain the morphosyntactic information to the level of the morphological annotation so that the syntactic annotation does not differ from other languages present in UD. Overall, we were only required to add one language-specific relation (`dep:repeat`) and resolve minor complications with several expressions.

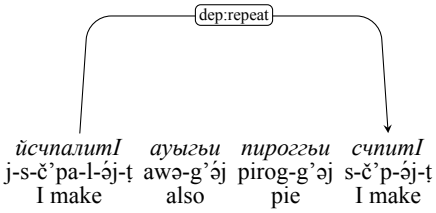
6.1 Repetitions

The data used in the Abaza treebank comes from spoken language and hence displays features that are not present in the written texts. In particular, our data contain multiple cases of word repetition, cf. the verb 'make' in (5).

(5) j-s-č'pa-l-əj-ṭ
3SG.N.ABS-1SG.ERG-make-HAB-PRS-DCL
awə-g'əj pirog-g'əj s-č'p-əj-ṭ <...>
DIST-ADD pie-ADD 1SG.ERG-make-PRS-DCL
'I make pies, I make...'

```
# sent_id = 57
# text_name = O_muzhe_SanashokovaCKh_13072017_checked.eaf
# text = спа акъарчаква дырщын дсызгIаргхтI
# text_orth = с-па а-къарча-ква ды-р-щы-н д-сы-з-гIа-р-г-х-тI
# text_transcription = s-pa a-q̇arč’a-kʷa də-r-š’ə-n d-sə-z-ʕa-r-g-χ-ṭ
# text_rus = Сына убили карачаевцы и привезли.
1 спа па NOUN _ Number[psor]=Sing|Person[psor]=1 3 obj _ Gloss=1sg.pr-сын
2 акъарчаква къарча NOUN _ Definite=Def|Number=Plur 3 nsubj _
Gloss=def-карачаевец-pl
3 дырщын щра VERB _ Gender[abs]=Com|Number[abs]=Sing|Number[erg]=Plur|Person[abs]=3|
Person[erg]=3|Tense=Past|VerbForm=Fin 0 root _ Gloss=3sg.h.abs-3pl.erg-убить-pst
4 дсызгIаргхтI гара VERB _ Gender[abs]=Com|Number[abs]=Sing|Number[erg]=Plur|
Number[io]=Sing|Person[abs]=3|Person[erg]=3|Person[io]=1|Tense=Aor|VerbForm=Fin
3 conj _ Gloss=3sg.h.abs-1sg.io-ben-csl-3pl.erg-нести-re-dcl
```

Figure 1: An annotation fragment.

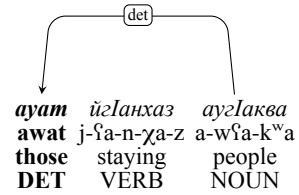


This phenomenon does not result from applying a grammatical rule (cf. reduplication expressing plurality or emphasis), and it does not fit into the existing set of UD syntactic relations. There is no speech repair, since the correct version of the word has already been uttered, so the `reparandum` relation cannot be employed in such cases. One might propose to use the `dislocated` relation but this relation is usually used for noun phrases that are fronted or postposed for reasons related to information structure (e.g., topicalization). However, what we are dealing with here is a result of hesitation about the next word, and the reason for repetition seems to be the intention of the speaker to fill the pause. Thus, we decided to introduce a new dependency relation `dep:repeat`, which encodes non-grammatical, non-repair repetitions.

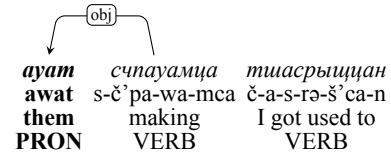
6.2 Demonstratives

Demonstratives in Abaza can be used both as determiners (6) and as third person pronouns (7). A similar pattern is observed in many other languages of the world (Bhat, 2013), such as Buryat, Hindi or Aleut. According to the current UD guidelines, the POS tag constrains the set of possible dependency relations. For instance, the dependent of an `advmod`-relation can only be an adverb (`ADV`). Likewise, in (6) we had to tag the demonstrative as a determiner (`DET`) so it can be a dependent of a `det`-relation. By contrast, in (7) we had to tag the same demonstative as a pronoun (`PRON`) so it can be an object of the verb. A similar solution was proposed for Punjabi in Arora (2022, p. 5706).

(6) awat j-ʕa-n-χa-z
dist.pl REL.ABS-CSL-LOC-stay-PST.NFIN
a-wʕa-kʷa <...>
DEF-people-PL
‘Those people who stayed <...>’



(7) awat s-č’pa-wa-mca
dist.pl 1SG.ERG-make-IPF-CVB
č-a-s-rə-š’ca-n <...>
RFL.ABS-3SG.N.IO-1SG.ERG-CAUS-get.used.to-PST
‘I got used to making **them** <...>’



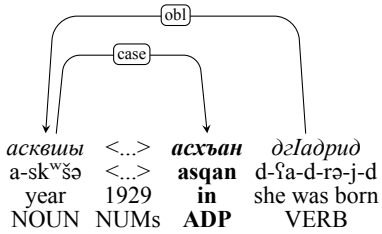
6.3 The word *asqan* ‘time’

The morphosyntactic status of the word *asqan* ‘time’ is not always clear. Usually, it heads nominal phrases denoting time periods (8) (‘in the time of the year 1929’) and subordinate temporal clauses (9)(‘in the time when she came back’).[5] Apparently, from the diachronic syntax perspective, in (8) *asqan* is a head of a noun phrase (‘the time of the year’), and in (9) it is a head of a relative clause (‘the time during which she came back’). However, we decided to simplify the annotation

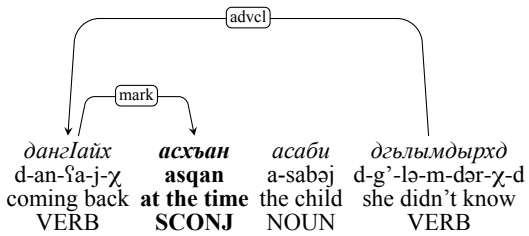
[5] Temporal clauses in Abaza represent a subtype of relative clauses. The predicate of the temporal clause is marked with the special prefix *an-* rel.tmp.

of *asqan* in our treebank for two reasons. First, *asqan* in these contexts may be seen as an already grammaticalized element and thus requiring a different analysis. Second, if we adhere to the diachronic analysis, a user who would like to find examples with adverbials and adverbial clauses in Abaza would miss those with *asqan*. Nominal phrases with *asqan* are functionally equivalent to English adpositional phrases with 'in' or 'during', and relative clauses with *asqan* are functionally equivalent to English *when*-clauses. That is why we decided to annotate Abaza *asqan*-constructions similarly to their English counterparts. Thus, in the current version of the Abaza treebank *asqan* introducing temporal nominals is analyzed as a postposition, and *asqan* introducing temporal clauses is analyzed as a subordinating conjunction.

(8) awəj a-skʷšə zk'-əj
DIST.SG DEF-year thousand-COORD
ĝ-ŝ-əj ʕʷaẑə ẑ-ba asqan
nine-hundred-COORD twenty nine-CL.N time
d-ʕa-d-rə-j-d
3SG.H.ABS-CSL-3PL.ERG-CAUS-be_born-DCL
'She was born **in** 1929.'



(9) <...> d-an-ʕa-j-χ asqan
3SG.H.ABS-REL.TMP-CSL-come-RE time
a-sabəj
DEF-child
d-g'-lə-m-dər-χ-d <...>
3SG.H.ABS-NEG.EMP-3SG.F.ERG-NEG-know-RE-DCL
'**When** she came back, she didn't know about the child <...>'



7 Conclusions

The Abaza treebank presented in this paper is the first case of a Northwest Caucasian language being added to UD. Abaza is a polysynthetic language, and thus it could be annotated on the level of individual morphemes, as suggested in (Park et al., 2021) for St. Lawrence Island Yupik. In this paper, we proposed a different approach which aims to minimize the morphological encoding, yet providing all necessary information for the analysis of syntactic relations. We showed that with the reduction of the number of the morphological features and some minimal adjustments to the set of dependency relations Abaza can be successfully annotated in the UD framework. Finally, we presented several solutions that aim at increasing the cross-linguistic comparability of our data.

Acknowledgements

This work is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University). We thank Peter Arkadiev for his help, as well as other members of the "Aspects of Abaza Grammar" research group. We also thank Thierry Poibeau and Niko Partanen for their extensive help during the initial stage of this project. And, finally, we want to thank the anonymous reviewers for their very helpful comments.

References

Peter M. Arkadiev. to appear. Abaza. In Yury Koryakov, Yury Lander, and Timur Maisak, editors, *The Caucasian Languages. An International Handbook*. De Gruyter Mouton, Berlin.

Timofey Arkhangelskiy. 2020. Web Corpora of Volga-Kama Uralic Languages. *Finno-Ugric Languages and Linguistics*, 9(1-2).

Aryaman Arora. 2022. Universal Dependencies for Punjabi. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5705–5711, Marseille, France. European Language Resources Association.

D. N. Shankara Bhat. 2013. Third person pronouns and demonstratives. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Vyacheslav A. Chirikba. 2012. Rasselenie abxazov i abazin v Turcii [Survey of the Abkhazians and Abazas in Turkey]. *Džigetskij sbornik. vyp. 1. Voprosy ètno-kul'turnoj istorii Zapadnoj Abxazii ili Džigetii [The Jiget collection. I. Studies in the ethnic and cultural history of Western Abkhazia or Jigetia]*, (1):21–95.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Uni-

versal Dependencies. *Computational Linguistics*, 47(2):255–308.

Rauf N. Klyčev. 1994. *Lokal□no-preverbnoe obrazovanie glagolov abazinskogo jazyka [The locative preverbial derivation of verbs in Abaza]*. Adž□pa, Cherkessk.

Johanna Mattissen. 2017. Sub-types of polysynthesis. In Michael Fortescue, Marianne Mithun, and Nicholas Evans, editors, *The Oxford Handbook of Polysynthesis*, pages 70–98. Oxford University Press.

Hyunji Hayley Park, Lane Schwartz, and Francis Tyers. 2021. Expanding Universal Dependencies for polysynthetic languages: A case of St. Lawrence Island Yupik. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 131–142, Online. Association for Computational Linguistics.

Robert Pugh, Marivel Huerta Mendez, Mitsuya Sasaki, and Francis Tyers. 2022. Universal Dependencies for western sierra Puebla Nahuatl. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5011–5020, Marseille, France. European Language Resources Association.

Nurja T. Tabulova. 1976. *Grammatika abazinskogo yazyka. Fonetika i morfologia [A grammar of Abaza. Phonetics and morphology]*. Karachaevo-Cherkesskoe otdelenie Stavropol'skogo knizhnogo izdatel'stva, Cherkessk.

Francis Tyers and Karina Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195–204, Barcelona, Spain (Online). Association for Computational Linguistics.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

A Appendix

A.1 Data availability

The current version of the treebank is available here: `https://github.com/UniversalDependencies/UD_Abaza-ATB/tree/dev`.

A.2 List of abbreviations

1 — 1st person; 3 — 3rd person; ABS — absolutive; ADD — additive; ADV — adverbial; BEN — benefactive; CAUS — causative; CL.N — classifier of non-humans; COORD — coordination; CSL — cislocative; CVB — converb; DCL — declarative; DEF — definite; DIST — distal demonstrative; EMP — emphasis; ERG — ergative; F — feminine; H — human; HAB — habitual; IO — indirect object; IPF — imperfective; LOC — locative preverb; N — non-human; NEG — negation; NFIN — non-finite; NPST — non-past; PL — plural; PRS — present; PST — past; RE, REP — repetitive; REL — relativization; RFL — reflexive; SG — singular; TMP — temporal subordination.