The Abkhaz National Corpus

Paul Meurer
University of Bergen
Bergen, Norway
paul.meurer@uib.no

Abstract

In this paper, we present the Abkhaz National Corpus, a comprehensive and open, grammatically annotated text corpus which makes the Abkhaz language accessible to scientific investigations from various perspectives (linguistics, literary studies, history, political and social sciences etc.). The corpus also serves as a means for the long-term preservation of Abkhaz language documents in digital form, and as a pedagogical tool for language learning. It now comprises more than 10 million words and is continuously being extended. Abkhaz is a lesser-resourced language; prior to this work virtually no computational resources for the language were available. As a member of the West-Caucasian language family, which is characterized by an extremely rich, polysynthetic morphological structure, Abkhaz poses serious challenges to morphosyntactic analysis, the main problem being the high degree of morphological ambiguity. We show how these challenges can be met, and what we plan to further enhance the performance of the analyser.

Keywords: Abkhaz National Corpus, lesser-resourced languages, polysynthetic languages, parsing of rich morphology

1. Motivation, sociolinguistic situation

Abkhaz, a West-Caucasian language, is spoken in Abkhazia by approx. 100,000 people, and by a considerable number of people in the Abkhazian diaspora in Turkey, Syria, Jordan, Adzharia (Georgia), Russia and elsewhere, with exact numbers unknown.

Abkhaz has a rather young literary tradition, reaching back no further than to the end of the 19th century, when Russian linguists devised a script for the language. In the 20th century, a vivid literary production developed, and the language was taught in schools and used as a medium of mainstream communication.

However, already in the 19th century, when Abkhazia became part of the Russian Empire and the muslim Abkhazians were deported to Turkey, the Abkhazians were in the minority. Because of the Soviet and Georgian rule, and the continuous influx of Georgians and other non-Abkhaz population, Abkhaz came, in the course of the 20th century, even more in the position of a minority language, with many ethnic Abkhazians giving up their language in favour of Russian. In 1989, before the Abkhaz-Georgian war, only 17.8% of the population of the Autonomous Republic Abkhazia were ethnic Abkhazians.[1]

Because of the almost total exodus of ethnic Georgians from Abkhazia during the Abkhaz-Georgian war, the Abkhazians are now again the majority population in the region (50,7% according to the 2011 census). However, Russian, which is the second official language in the self-declared Abkhaz state, has remained the dominant language of mainstream communication.

Since the Abkhaz de-facto independence, there is a growing awareness in the Abkhaz public of the need to improve the viability of the language. There are also political aspirations to make Abkhaz more widely used in public, and even a law (“Law on the official language of the Republic Abkhazia” from 2006) that tries to enforce the use of Abkhaz in the state administration and in public life, with the long-term aim to have Abkhaz replace Russian as the dominant language.

But the implementation of these measures, overzealous as they appear from the outset, is severely hampered by the lack of qualified teachers and adequate or even basic printed or electronic teaching material.

On this background, the Abkhaz National Corpus (AbNC)[2] project has the following aims:

- To establish a linguistically annotated Abkhaz language corpus of reasonable size
- To develop freely available computational resources (electronic dictionaries and lexicons, parsers and more) for the language
- To give learners of the language a tool that provides electronic support for reading texts online
- To serve as a long-time storage repository for Abkhaz language texts, dictionaries and other resources in electronic form (as a CLARIN[3] resource)

2. Previous work

The project I am reporting on is the first computational work being done on Abkhaz, and indeed on any West-Caucasian language. The only corpus resource besides the AbNC is a small corpus of 2 million words without grammatical annotation.[4] There are plans to create an annotated Circassian/Adyghe corpus (Arkhangelskij and Lander, 2016), but this work seems still to be in a preparatory phase.

3. The project

The AbNC project is in the fortunate situation that there exists a large online collection of freely available Abkhaz

[1] See https://en.wikipedia.org/wiki/Demographics_of_Abkhazia for these numbers.

[2] http://clarino.uib.no/abnc

[3] https://www.clarin.eu

[4] http://baltoslav.eu/apsua

texts.[5] The maintainers of the site are enthusiasts who put a lot of effort into collecting and scanning Abkhaz books and other documents, with the aim to build a comprehensive archive of Abkhaz text resources. In addition to scanned texts, they also make available digitally-born PDF files from Abkhaz publishing houses, and it is mainly these digital PDF texts (originating from 2010 or later, but often constituting reprints of older editions) the AbNC project draws upon. In addition, internet news sites were harvested, some important texts were scanned and OCRed in the project; and some texts came directly from the authors.

The texts belong to a variety of genres, with novels, traditional texts and news texts in the majority. Table (1) shows approximate sizes for the genres represented in the corpus. The sizes are in 1000 tokens, where tokens comprise both words and punctuation.

(1)

Genre	*Size*
fiction	
traditional (fairy tales, Nart saga etc.)	1,000
religious (New Testament)	200
novels and other prose	6,700
poetry	400
drama	300
non-fiction	
journalism and news	1,250
political	20
textbook	140

The texts were thoroughly preprocessed and annotated with rough structural markup adhering to TEI P5 standards that captures section headings, paragraphs, page breaks, footnotes, lines of poetry and the structure of drama texts. In addition, necessary metadata was included in the TEI header. The result is a very clean corpus with texts that can be read nicely in a web browser.

The project was able to agree with the Abkhaz partners on an open license that poses very few restrictions on the use of the texts in the confines of the project. In technical terms, the license is a CLARIN PUB license (CLARIN_PUB-BY-NC-ND).

4. Typological features of Abkhaz

Phonetically, Abkhaz is characterized by a large number of consonant phonemes (between 56 and 65, depending on the dialect), whereas there are few vowel phonemes. Only *a* and *ə* are phonemic, whereas *e, i, o* and *u* can appear in loan words and as phonetic realizations of *a* and *ə* in certain contexts. The nominal morphology is rather simple; there is no case marking, but the language exhibits noun-noun and noun-adjective compounding. Postpositions can be suffixed to the noun or adjective. Remarkably, cardinal numbers can be prefixed to the noun. The verbal morphology can be characterized as agglutinative and polysynthetic, with a huge number of verbal prefixes and suffixes. The position of the affixes in the verb is rigid, they fit into given slots of a general template. Morphemes are generally monosyllabic (*Ca, Cə*), with a high degree of homonymy. There is however little suppletion, few irregularities and few phonological processes. Abkhaz has free and dynamic word stress which is not coded in the orthography.

5. The analyser

Since there is no disambiguated training data available, and because of the extraordinarily rich morphology of the language, we did not consider a statistical approach to morphological tagging at the outset. Instead, we developed a finite-state morphological analyser using the FST platform (Beesley and Karttunen, 2003), and Constraint Grammar rules are used for disambiguation. We are however planning to train a statistical disambiguator that operates on the partially disambiguated output of the CG parser once a sufficiently large gold standard corpus has been created.

5.1. The lexicon

The lexicon of the analyser is based on a large bilingual (Abkhaz-Russian) dictionary (Kaslandzia, 2005), which we had to convert into a suitable machine-readable format in a semi-automatic process. All lexicon entries of the dictionary are marked for word stress, which is distinctive in Abkhaz. For the nouns, plural forms are given. The verb entries, with the verbal noun as lemma form, contain information on transitivity; static and dynamic verbs are distinguished, and for each entry, a sample inflected form is given, in addition to translations and example phrases. Unfortunately, there are many typographic and other errors in the dictionary entries, and the important transitivity information is not very reliable. We have corrected these errors manually, and often we had to consult other dictionaries (Yanigasawa, 2010; Šʼaḳrəl and Kondžʼaria, 1986) and Abkhaz native speakers.

To be useful for the analyser, the verb lemma entries had to be segmented correctly into preverb, stem and affix morphemes, and to those segments the correct morphological features had to be assigned. The segmentation candidates were generated by an auxiliary finite state transducer that is basically the subset of the full finite-state analyser restricted to verbal nouns. In general, there are many segmentation candidates, and we did choose the correct analyses manually.

As an example, the annotation of the correct segmentation of the verb ‘аитцанаршәшәара’ *áic̣anarš*^{w}*š*^{w}*ara* ‘to be shaken (in a car etc.)’ is *á-ai/R-c̣a/A-na/S-r:š*^{w}*š*^{w}*a-ra*, with markers for reciprocal (*/R*), local preverb taking an object (*/A*), inanimate agent (*/S*), and causative (*r:*) before the root. The prefix *á-* is the generic article and the suffix *-ra* is the verbal noun marker, both of which are always present.

A specific problem is to properly segment the stem of intransitive verbs. A verbal stem is either simple (consisting of the verbal root only), or it is composed of a (simple or complex) preverb and the root. For transitive verbs, the boundary between preverb and root is obvious from the inflected form, because the agent affix (normally *-i-, -l-* or *-r-* in the sample forms) is placed immediately before the root. Intransitive verbs however have no agent affix; here, the boundary is only visible in comparatively rare negated

[5] `http://apsnyteka.org`

aorist forms, where negation is marked by the negation infix *-m-* placed between preverb and root. When such a form was not contained in the example phrases, we had to find evidence for the correct segmentation by a corpus search or from informants.

The analyser lexicon comprises 6,750 verb stems, 16,000 nouns, 3,000 adjectives and 1,350 adverbs.

5.2. The feature set

The morphological analyser (Meurer, 2011) takes as input an orthographic surface form in the Abkhaz Cyrillic script and returns, for each found reading, a lemma form (containing stress information), and a bag of morpho-syntactic features.

The feature set is quite rich, it comprises more than 350 features. There are 160 features for inflected verb forms alone. This high number is due to the polysynthetic nature of the Abkhaz verb; there are affixes not only for the subject, but also for direct, indirect and locative objects in the verb, there are applicative markers with associated person affixes, as well as adverbial, conjunctional, question and clitic elements. Person affixes include relative and reciproque markers. The verb form in Example (2) will be analysed as the lemma form ‘аҽамадара’ *a-č·á·ma·da-ra* with feature set (*V Dyn Tr NonFin Pres Neg Refl:Rel Reln:Pot RO:Rel LO:3Pl QWho Excess*).

(2) зҽызызрымамдацәода
z.čə-zə.z-rə-ma-m-da-c^w*a.wa-da*
Rel.Refl-Rel.Pot-3Pl-PV-Neg-ROOT-Exc-Pres-Who?
‘who cannot trust them too much?’, lit.: ‘who is it who cannot betrust whose head to them in excess?’

5.3. Ambiguity

Since nominal and verbal stems and affixes often consist of only one simple (CV) syllable, there is considerable homonomy among those morphemes. Although morphemes of equal syllable structure may differ in their accent status, this information is not (or in some cases only indirectly) available in the orthographic word form, since stress is not marked in present-day orthography.

As a consequence, most word forms are highly ambiguous; often there are many ways to segment a word form into preverb(s), stem, and affixes, and those stems and affixes may by themselves be ambiguous. In addition, lemma forms may differ solely in their stress pattern.

For example, the word ‘илаба’ *ilaba* (with its most frequent reading ‘his stick’) gets 42 analyses, consisting of inflected and predicative forms of *a-labá* N ‘stick’, *a-lába* N ‘male dog’, *abá* N ‘textile’, *abá* A ‘dry’, *á-la·ba-ra* V ‘suck up’, *a-ba-rá* V ‘see’.

5.4. Disambiguation

Abkhaz poses serious challenges to disambiguation due to the virtual absence of dependent marking (e.g., no case marking), the weak significance of word order, and the extreme prevalence of homonymy.

To keep ambiguity not higher than necessary, we tried as much as possible to avoid overgeneration in the morphological analyser. Obviously, the prefixes and suffixes that can fill the slots of the verb template cannot combine freely. The negation marker *-m-*, for example, can only occur in one position at a time, although there is a prefix and a suffix slot position for negation. The position of the marker depends on a combination of tense, finiteness, and whether the form is dynamic or static. Relational prefixes are restricted to non-finite tense suffixes. Other restrictions depend on syntactic and semantic properties of a verb; e.g., agent prefixes are obviously restricted to transitive verbs only. The relevant grammatical literature (e.g., (Spruit, 1986; Šaduri, 2006; Hewitt, 2010), to mention a few important works) is however in many cases vague about those combinatorial restrictions, and we had to work out specific rules to be used in the implementation of the morphology.

One such case is the possibility of relative markers in a verb form. A general rule restricts relative markers to non-absolutive non-finite forms. However, a corpus search reveals that the situation regarding absolutives is more complicated:

In Abkhaz, absolutives are used in serial verb constructions, where the absolutive and the verb following it are part of the same event. This can be seen in (3), where the verb ‘амазаара’ *á-ma-zaa-ra* ‘have’ initiates a serial construction.

(3)	Ача	иманы	иааит.
	a-č’á$_i$	ϕ_i-i_j*-ma-nə́*	i_j*-áai-ṭ*
	the-bread	(it)-it-have.Abs	it-come.Aor.Fin

‘It came with (lit. having) the bread.’

However, if the main verb of the serial construction is a non-finite relativized verb like in (4), its subject relative marker (*-i-*) is taken up by the absolutive (in the form of the indirect object relative marker *-z-*). Thus, we get an absolutive form with a relative marker.

(4)	ача	зманы	иааиз
	a-č’á$_i$	ϕ_i-z_j*-ma-nə́*	i_j*-áai-z*
	the-bread	(it)-Rel-have.Abs	Rel-come.Aor.NonFin
	амашьына		
	a-maš’ə́na		
	the-car		

‘the car which came with the bread’

In devising the Constraint Grammar rules, several strategies were used to disambiguate homonymous forms:

- Highly unlikely (albeit possible) analyses of frequent words were ruled out.
- Syntactic rules were used where possible.
- To disambiguate further, semantic information in the form of selectional restrictions was used.
- As a heuristics, analyses that match a dictionary entry were preferred.

5.5. Treatment of predicate phrases

In Abkhaz, stative verbs can be formed from nouns and adjectives, but also from two-element noun phrases, with predicative meaning: ‘X is NP’. This is shown in (5), where the brackets indicate the NP that is surrounded by verbal morphological material.

(5) Дысҩыза бзиоуп.
də-[*s-y*^{w}*ə́ẓa* *bẓío*]*-up̣.*
he-[my-friend good]-be.Stat.Pres.Fin
'He is my good friend.'

Аҟәа шқалақь бзиоу
Áq^{w}*a* *ø-š-*[*kalak'* *bẓío*]*-u*
Sukhum (it)-how-[city good]-be.Stat.Pres.NonFin
'what a good city Sukhum is'

The close analogy to true static verbs is apparent from (6).

(6) Дгылоуп.
d-[*gə́lo*]*-up̣.*
he-[stand]-Stat.Pres.Fin
'He is standing.'

ишгылоу
i-š-[*gə́lo*]*-u*
it-how-[stand]-Stat.Pres.NonFin
'how it is standing'

The question arises whether to analyse such constructions as verbs, or to keep the part of speech of the underlying noun or noun phrase components. Treating a verbal form derived from a complex NP as a verb would entail that the construction would have to be represented as a single token containing whitespace. This would be unfortunate in the context of a searchable corpus, for several reasons. First, it would not be clear how the lemma form of such a complex static verb should look like – verbal nouns cannot be derived from two-element NPs. And even if a solution to this problem were found, it would be complicated and unintuitive to search for the components of the NP stem in such constructions.

Therefore, the two components of the predicate NP are treated as two separate tokens, each being annotated with the part of speech of the underlying component. In addition, the components are annotated with a *Pred* feature, indicating that they are parts of a predicative static verb construction, and with a *LHP* (left half predicate) and *RHP* (right half predicate) feature, respectively, showing that they are the left and right parts of a full predicate construction. They also get the features that other affixes of both components might contribute. Such an analysis might not be ideal, but it is satisfactory in the context of a searchable corpus. The analysis of the phrase 'шқалақь бзиоу' *ø-š-*[*kalak' bẓío*]*-u* would thus look like this:

(7) *á-kalak'* Noun NH Sg Pred LHP How S:3 S:Ad
a-bẓía Adj Sg Pred RHP NonFin Pres

This analysis is also extended to static verbs derived from single nouns and adjectives.

5.6. Performance

At the present stage of development, it is premature to give exact precision and recall figures for the analyser. For the corpus as a whole, 93.6% of the tokens and 74.0% of the types are recognized by the analyser. A non-formal evaluation of the parser on a section of a novel chapter (medium complexity) indicates that more than 85% of the tokens are assigned the correct part of speech. Precision for all features will obviously be somewhat lower.

6. The corpus

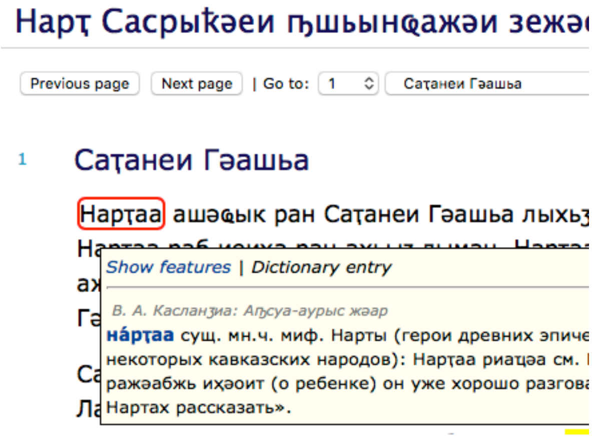
The Abkhaz National Corpus is available in the corpus management platform Corpuscle (Meurer, 2012), which offers advanced search capabilities and a user-friendly web interface. To make the corpus accessible to Abkhazians, who are often uncomfortable with English, the tool was localized to Russian and to Abkhaz. All Abkhaz text, which by default is shown in the modern Abkhaz Cyrillic alphabet, can also be viewed in scientific transliteration, which makes it easier for non-specialist linguists to use the corpus.

6.1. Features for language learners

In addition to being a corpus-linguistic resource and tool, the corpus also serves as a digital library and a pedagogical tool for language learning.

From the catalogue page, where all texts are listed by author, title, creation date and genre, the user can select a text for reading. On the reading page, one page of the text is shown, which in most cases corresponds to a page of the printed edition. The user can maneuvre to the next or previous page, or select a page or a book section from a drop-down menu. The current reading position can be stored as a bookmark.

Most importantly, by clicking on a word in the text, the user is shown grammatical information about that word, including lemma form and morphosyntactic features, and the lemma can be looked up in the integrated Abkhaz-Russian Dictionary (Kaslandzia, 2005) (see Fig. 1).



Нарҭ Сасрыҟәеи ҧшьынҩажәи зежә

1 Саҭанеи Гәашьа

Нарҭаа ашәҩык ран Саҭанеи Гәашьа лыхьӡ

Show features | Dictionary entry

В. А. Касланӡиа: Аҧсуа-аурыс жәар

на́рҭаа сущ. мн.ч. миф. Нарты (герои древних эпиче
некоторых кавказских народов): Нарҭаа риаҵәа см.
ражәабжь иҳәоит (о ребенке) он уже хорошо разгова
Нартах рассказать».

Figure 1: Dictionary lookup.

7. Conclusions and future work

Abkhaz, a polysynthetic language, exhibits a high degree of morphological complexity and word form homonymy. We have outlined how a morphosyntactic analyser can be implemented that adresses these complexities. Although much remains to be done to improve precision and recall of the analyser, it is already in its present state a useful tool for annotating a text corpus of medium size, the Abkhaz National Corpus. When the feature set has stabilized, a gold corpus will be manually annotated, so that exact performance figures can be calculated. The gold corpus will also be used to train a statistical disambiguator that will be used on top of the Constraint Grammar-based disambiguator.

The corpus itself is continually growing, and the long-term aim is to include all important texts that have been written in Abkhaz.

8. Acknowledgments

The Abkhaz National Corpus project started in 2015. Project partners are the Goethe University Frankfurt, Institute of Empirical Linguistics, the Centre for Civil Integration and Inter-Ethnic Relations (CCIIR), Tbilisi, and the organization Business Women of Abkhazia, Sukhum.

The project is funded by the United States Agency for International Development (USAID).

We would like to thank the Abkhaz project partners for their valuable help in resolving questions about Abkhaz morphology and the lexicon.

9. Bibliographical References

Arkhangelskij, T. A. and Lander, Y. A. (2016). Developing a polysynthetic language corpus: problems and solutions. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016"*, Moscow.

Beesley, K. R. and Karttunen, L. (2003). *Finite State Morphology*. CSLI Publications, Stanford.

Hewitt, G. (2010). *Abkhaz. A Comprehensive Self-Tutor*. LINCOM, München.

Kaslandzia, V. (2005). *Apsua-aurəs ž^{w}ar (Abkhaz-Russian Dictionary)*. Abkhazian State Academy, Sukhum.

Meurer, P. (2011). A finite state approach to Abkhaz morphology and stress. In *Lecture Notes in Computer Science 2011, Volume 6618*, pages 271–282.

Meurer, P. (2012). Corpuscle: a new corpus management platform for annotated corpora. In *Exploring newspaper language: using the web to create and investigate a large corpus of modern Norwegian*, pages 31–49. John Benjamins Publishing Company.

Spruit, A. (1986). *Abkhaz studies. Doctoral dissertation*. Leiden.

Yanigasawa, T. (2010). *Analytic Dictionary of Abkhaz*. Hituzi Syobo Publishing, Tokyo.

Šaduri, I. (2006). *Morfologija abxazskogo jazyka*. Tbilisi.

Š'aḳrəl, K. S. and Kondž'aria, V. H. (1986). *Apsua bəzš^{w} a až^{w}ar (Dictionary of the Abkhaz language)*. Alašara, Sukhum.