**NATIONAL RESEARCH UNIVERSITY**

**HIGHER SCHOOL OF ECONOMICS**

**SCHOOL OF LINGUISTICS**

**TERM PAPER**

**BUILDING MORPHOLOGICAL PARSER FOR AVAR IN LEXD AND TWOL**

Golovina Anna

group 213

Research Advisor

Associate Professor

George Moroz

**MOSCOW**

**2023**

**Contents**

[1. Introduction 3](#_heading=h.30j0zll)

[2. Main part 5](#_heading=h.1fob9te)

[2.1. Theoretical overview 5](#_heading=h.3znysh7)

[2.1.1. Language data 5](#_heading=h.2et92p0)

[2.1.2. Methodology 8](#_heading=h.tyjcwt)

[2.1.3. Literature review 9](#_heading=h.3dy6vkm)

[2.1.4. Drawbacks of the previous parser based on lexc 11](#_heading=h.1t3h5sf)

[2.2. Lexd 12](#_heading=h.2s8eyo1)

[2.2.1. Structure of the upgraded compiler 12](#_heading=h.17dp8vu)

[2.2.2. Evaluation of corpus coverage with lexd 13](#_heading=h.3rdcrjn)

[3. Conclusion 14](#_heading=h.26in1rg)

[References 15](#_heading=h.lnxbz9)

# Introduction

This work is devoted to building a morphological analyzer for the Avar language. Morphological analyzers are one of the most essential tools in the NLP, as they are capable of being a base for automatic translators, lemmatizers, spell-checkers and other NLP tools.

At the moment, there are few tools for processing texts in the Avar language, and the most significant is the parser from Apertium (ссылку). However, it was created 10 years ago and functioned on the basis of the lexc compiler. Moreover, the previous parser needed validation and completion of the data.

One of the main problems associated with checking the data and markup of the parser on lexc was that there are quite a few theoretical works on the Avar language, and, basically, these are teach-yourself books on the language. Most of the theoretical works are more focused on the verbal system of the Avar language, as, for example, (add references). This fact is explained by the richness of the Avar language in such phenomena as ergativity, reciprocity, and the presence of labile verbs.

The latest and most modern version of lexical compilers is lexd, so transferring data from lexc to lexd and supplementing the existing data seems to be a relevant task for upgrading NLP tools for the Avar language. There are two approaches to creating morphological parsers. They are based either on probabilistic models that underlie neural networks, or on the processing of a set of grammatical rules of the language. However, neural network parsers are only suitable if there is a large collection of texts in the language. According to the 2020 census data, 654,363 people speak Avar in the Russian Federation, and the amount of available data on the Avar language is small and often this data is not digitized. Accordingly, Avar can be attributed to low-resource languages, for which only a rule parser can be created.

To create this morphological parser, the Helsinki Finite-State Toolkit framework (Lindén et. al. 2011) is used. This set of tools allows you to work with the lexd lexical compiler and the twol morphonological rules engine. The programs listed are working on the base of transducers, bidirectional finite-state machines that are capable of not only parsing given material but also generating new tokens due to the rules written in it.

To test the quality of the resulting parser, we used the data from the Avar Wikipedia, namely, downloaded dumps of all articles, on the texts of which it was possible to check whether the parser recognizes certain forms.

The paper also discusses the shortcomings of the existing transducer for the Avar language and the result of its update and transfer from lexc to lexd. The relevance of the study lies in the fact that there are not enough tools for studying the Avar language outside the field, as a result of which Avar is studied mainly in Dagestan).

The creation of a morphological parser can contribute to an increase in interest in Avar and the emergence of new theoretical works about the language, in view of the fact that it provides more convenient automated text processing. Moreover, from the practical point of view, the parser can make a significant contribution to the UniMorph project (сюда и в библиографию ссылку на Sylak-Glassman 2016).

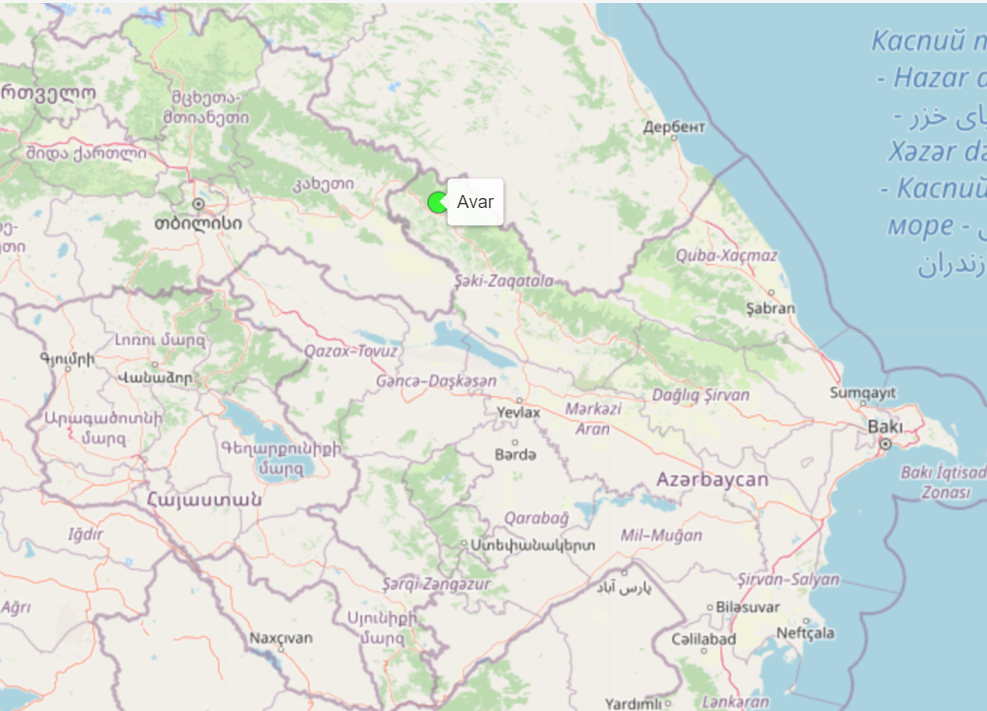
# Main part

## *Theoretical overview*

### Language data

The Avar language belongs to the Avar group of the Avar-Ando-Tsez branch of the Nakh-Dagestan family of languages. Its native speakers live on the territory of Dagestan in Russia and also in some regions of Azerbaijan.

Fig 1. Map of the area where native speakers of the Avar language live (Dryer, Haspelmath, 2013)



Avar is not only one of the official languages of Dagestan, which is taught in schools, but it also is the language of interethnic communication in Dagestan (ссылочку на языки Дагестана добавить). The dialects of the Avar language are divided into two large groups: the northern (which includes the eastern, Salatav and Khunzakh dialects) and the southern (Gid, Antsukh, Zakatal, Karakh, Andalal, Kakhib and Kusur dialects) dialect groups. The northern dialect formed the basis of the literary norm of the Avar language, called bolmats, which is considered in this work.

According to the data of the Russian population census for 2020, 654,363 people speak the Avar language. Officially, the Avar language has been given the status of vulnerable (ссылка на UNESCO + добавить ее же в библиографию).

Despite the relatively large number of people who speak the Avar language, it is still poorly studied and is studied mainly in Dagestan, which, of course, prevents a complete picture of the languages of the Caucasus at the international level.

From the point of view of the alignment, the Avar language is ergative. Morphologically, it is rather agglutinative, but allows some elements of fusion. The big difference between the number of lemmas and word forms in the language is due not only to the rich morphology (presence of nominal classes, hypertrophied case system, rich verbal system with labile verbs, a large number of aspect categories and reciprocal forms), but also to complex morphonological processes, for example, vowel syncope during stress shift, assimilation and consonant alternation. All these parameters must be taken into consideration during the parser development, as they directly affect the number of rules and size of lexicon in the parser.

Since the verbal system of the Avar language has been studied much better than the nominal one, it was necessary, first of all, to put information about nouns, adjectives, adverbs and numerals into the parser.

The Avar language has cardinal, ordinal, distributive, group and collective numerals. All of them, with the exception of the quantitative numeral *co* ‘one’, are formed by adding formative affixes or enclitics. Another distinctive feature of the language is the presence of a vigesimal numeral system.

In the paradigm of adjectives, the so-called short and long forms are distinguished. The derivational suffix involved in the formation of long forms of adjectives based on short ones is also considered as an adjectivizer.

It is also worth noting that it is often difficult to draw a clear distinction between adjectives and adverbs, long and short forms seem to be built into a continuum from adjective to adverb, since the short forms mentioned earlier can be used as an adverb. Long forms, on the contrary, seem to have more characteristics of conventional adjectives as, for example, gender agreement (the suffix agreement is controlled by the head of the noun phrase in which an adjective occurs). Adjective prefixes (which are more like class markers for nouns) also agree with the head of the noun phrase, but if there is an adjectival argument, they agree to that complement.

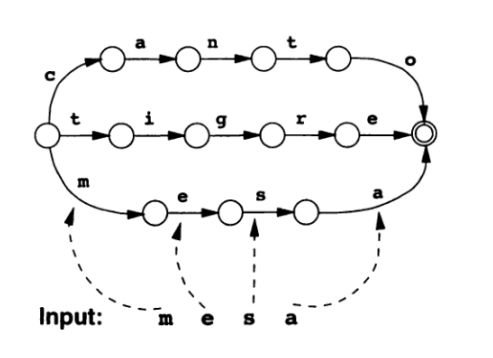
Nouns are marked by case, number and gender. At the same time, it is important to note that the grammatical masculine and feminine gender coincide with the biological sex of a person, and inanimate nouns are marked with indicators of the neuter gender. The cases of nouns can be divided into core and locative. The core ones are nominative, ergative, genitive and dative. There are 20 locative cases, and they are formed according to the principle of attaching the marker of localization and then the indicator of the case (lative, essive or ablative) to the stem. For nouns of the first and second nominal class, the ergative case form is also used as an oblique stem…

### Methodology

Lexd was used to describe the morphological rules directly, twol was used to describe the morphonological alternations, and the hfst toolkit was used to connect these two transducers and run them directly.

Lexd, receiving a string as input, processes it character by character. Thus, one or another substring is defined as a stem or as an affix in accordance with what is written in the transducer's lexicon…

Fig 2. Principle of how the finite state machine works (Beesley & Karttunen, 2003)



### Literature review

The earliest work on the study of the Avar language is (Uslar (write ref. in the normal form)). Already in the 20th century, Avar was researched (list of references from the bibliography). Most grammars describe the structure of literary Avar. (it is possible to add more about Charachidze, who investigated the Khuznakh dialect, but this dialect is still close to the literary norm, so we can take this paper into consideration)

However, there are not very many comprehensive descriptive grammars of the Avar language, and the names of their authors are repeated paper by paper, since this language is studied by a limited circle of scientists. Moreover, often grammars are not digitized, which complicates their processing and transfer of rules to lexd.

(Madieva 1981) and (Ebeling 1966) were chosen as the main source of inflectional and formative paradigms. Despite the fact that these are not the most modern editions, they contain the most complete and comprehensive information on the morphology of nominal categories.

The main dictionary is the Russian-Avar school dictionary, that is due to the fact that, to a large extent, the study of the Avar literary language takes place at school (тут ссылка на ФЗ) and often the first language of children in Dagestan is Russian or the national language of a particular region.

(according to my dialogue with a native speaker, children in Dagestan often acquire the Russian language as L1, and they learn the Avar literary language at school as the Republic language (basically, as a the cultural component of the motherland) and, after stopping classes, forget the rules and vocabulary)

The second dictionary chosen as the source of the lexicon is (Saidov 1967) despite the fact that this dictionary is older than (Madieva 1981), since the vocabulary of this dictionary is wider than that of the school one.

Most of the theoretical studies of the Avar language are devoted, as already mentioned, to the verb system. Nominal categories are studied rather from the point of view of syntactic characteristics (works about anaphors, reflexives regarding different grammatical relations), and nominal morphology remains outside the scope of individual studies and is more common in prefabricated grammars.

To test the quality of the resulting parser, it was necessary to find texts in the Avar language and try 2 corpora: (добавить нормальные ссылки на балтослав и веб-корпора)

Since it was difficult to get all the data from both corpora (self-written crawlers could not get all the texts), the Avar Wikipedia dump was used for testing.

For this, the texts of all articles of the Avar Wikipedia were downloaded. Then they were tokenized, cleared of spelling. To compile a test csv file, it was also necessary to remove word repetitions.

Thus, the parser was tested on preprocessed word forms from the Avar Wikipedia.

### Drawbacks of the previous parser based on lexc

A morphological parser for the Avar language has already been created as part of the Apertium open-source machine-translation project (ссылку добавить). However, the parser is based on the lexc transducer, which is an older version of lexd.

Lexc is an obsolete version of the transducer, it has cumbersome and non-intuitively understandable syntax, compared to lexd. However, the previous parser was in need of upgrade not only because of the relevance of the transducer version.

Even a cursory examination of the files shows that some code fragments are duplicated unnecessarily. The general form of writing the code was extensive, but not optimized.

When analyzing the parser, shortcomings were found from the point of view of not only computational, but also theoretical linguistics. The shortcomings concerned the structure of word forms registered in the transducer and the presence of gaps in the paradigms of different parts of speech and in the lists of translations.

In more detail, among the numerals these such as *fifty*, *seventy*, *ninety* are missing.

In numerals, the clitic -go is assigned to the indicators of numerals, as if it were a word-changing affix (but in (ссылка на Форкер и Мадиеву) this item is considered as an enclitic).

The list of adjectives included pronouns and postpositions (*буге-гьечӏеб, кинабго, кӏиябго*)

In the original lexc, there was no case differentiation of adjectives.

(+ add problems with nouns: incorrect tags of number; lack of splits (ergative base is not considered as an oblique stem in the 1st and the 2nd declension classes); pretty random groups of clitics)

## *Lexd*

### Structure of the upgraded compiler

Different groups of Avar numerals are mainly formed by manipulating the stem of cardinal numbers.

The numeral *co* ‘one’ is the only one that does not have morphemes in its composition, except for the root. To all cardinal numbers, except for *co* ‘one’, the clitic =go is added to the stem. Ordinal numbers are formed by adding the suffix -abile-GM, and distributive numbers by complete reduplication. The suffix -(j)a-GM=go is used to form collective numerals, and such forms are used to either select a subset or denote the number of elements of a set.

(добавить глоссы и примеры)

When writing lexd, the stems of cardinal numerals were added to the lexicon, and suffixes and clitics of other categories of numerals were added to the patterns.

As mentioned earlier, adjectives change in gender and number, but gender is not distinguished within the plural. In the declension paradigms of the neuter and feminine genders, only the forms of the absolutive differ, on the contrary, the forms of the ergative, genitive and dative, as well as the superessive, coincide.

(таблица из Алексеева и Атаева)

The main problem with the processing of adverbs was that some locative postpositions in the Avar language can act as adverbs, and in some contexts it is not possible to draw a clear distinction between them. In addition, adverbs or postpositions in the function of adverbs can be inflected by directional cases. The types of adverb stems that attach different affixes are quite heterogeneous, so they were recorded in the lexicons separately, and a separate pattern was prescribed for each type.

Nouns are divided into three declensional classes. In the first and second, the ergative form acts as an indirect stem. In the third class, the forms of indirect cases consist of the stem of the ergative form and case markers.

Another big problem associated with the processing of nouns is the presence of a large number of assimilations and mutations of consonants, which can be systematized only if separate groups of word forms are written into the lexd lexicon as different types of declensions.

The paradigms of plural and singular nouns differ in the number of cases that can be expressed. So, in the singular, the absolutive, ergative, genitive, dative, super-essive, apud-essive, apud-lative and apud-ablative are possible. Absolutive, ergative, dative and apud-essive are possible in the plural.

In addition, a location can be marked with a special series of spatial markers.

(сюда таблицу из Атаева и Алексеева)

### Evaluation of corpus coverage with lexd

# Conclusion

(about coverage)

In further development of the parser, it is necessary to add the verb declension and various small groups of words, such as abbreviations, interjections, conjunctions and postpositions.

переводчик: вопрос, успею или нет

# References

Beesley, Karttunen 2003 — K. Beesley, L. Karttunen (2003). *Finite-state morphology: Xerox tools and techniques*. CSLI, Stanford.

Federal State Statistics Service. (2020). *Results of the 2020 All-Russian Population Census*. Vol. 5. National composition and language use. <https://rosstat.gov.ru/vpn_popul>.

Lindén et al. 2011 — K. Lindén, E. Axelson, S. Hardwick, T. A. Pirinen, M. Silfverberg (2011). Hfst – framework for compiling and applying morphologies. *International Workshop on Systems and Frameworks for Computational Morphology* (pp. 67–85). Springer, Berlin, Heidelberg.

Swanson, Howell 2021 — D. Swanson, N. Howell (2021). Lexd: A Finite-State Lexicon Compiler for Non-Suffixational Morphologies. *Multilingual Facilitation* (pp. 133–146). University of Helsinki.

Federal law of 10.04.2014 N 48 “On education in the Republic of Dagestan”// LRS ConsultantPlus. <http://www.consultant.ru/regbase/cgi/online.cgi?req=doc&base=RLAW346&n=22373>

Forker, Polinsky 2021 — D. Forker, M. Polinsky (2021). Avar grammar sketch. *Handbook of the languages of the Caucasus*.

Alekseev, Ataev 1997 — M. Alekseev, B. Ataev (1997). *Avarskij jazyk*. Moscow: Academia.

Alekseev et al. 2014 — M. Alekseev, S. Alikhanov, B. Ataev, M. Magomedov, M. Magomedov, G. Madieva, P. Saidova, D. Samedov (2014). *Sovremennyj avarskij jazyk*. Makhachkala: Aleph.

Bokarev 1949 — E. Bokarev (1949). *Sintaksis avarskogo jazyka*. Moscow: Izd. AN SSSR.

Ebeling 1966 — C. Ebeling (1966). The grammar of literary Avar. *Studia Caucasica 2*. 58-100.

Khalilov, Khalilova 2016 — M. Khalilov, Z. Khalilova (2016). Avar. In P. Müller, I. Ohnheiser, S. Olsen, F. Rainer (eds.). *Word-Formation: An International Handbook of the Languages of Europe*. Volume 5. Berlin: De Gruyter Mouton.

Madieva 1967 — G. Madieva (1967). Avarskij jazyk. In E. Boakrev, K. Lomtatidze (eds.) *Jazyki narodov SSSR*, Vol. IV, Iberijsko-kavkazskie jazyki, 255–271. Moscow: Nauka.

Madieva 1981 — G. Madieva (1981). *Morfologija avarskogo literaturnogo jazyka*. Makhachkala.

Madieva 2000 — G. Madieva (2000). Avarskij jazyk. In G. Gamzatov (ed.) *Jazyki Dagestana*, 45–62. Maxačkala, Moscow: Rossijskaja Akademija Nauk, otdelenie literatury i jazyka.

Magomedov, Magomedov 2009 — M. Magomedov, M. Magomedov (2009). *Škol’nyj russko-avarskij slovar’*. Makhachkala: Lotos.

Saidov 1967 — M. Saidov (1967). *Avarsko-russkij slovar’*. Moscow: Sovetskaja Enciklopedija.

Yamada 2013 — H. Yamada (2013). *Reciprocal constructions in Avar*. Lingua 126. 150–171.

Letučij 2022 — A. Letučij (2022). *Tipologija labilʹnyh glagolov*. – Litres.

Polinsky et al. 2012 — M. Polinsky, C. Gallo, P. Graff, E. Kravtchenko (2012). *Subject preference and ergativity*. Lingua 122. 267–277

Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. WALS Online (v2020.3) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.7385533>

(Available online at https://wals.info, Accessed on 2023-05-10.)

(еще ссылку по поводу гипертрофированной падежной системы из грамсема Плунгяна)