

АВТОМАТИЧЕСКИЙ МОРФОЛОГИЧЕСКИЙ АНАЛИЗ ДЛЯ ОСЕТИНСКОГО ЯЗЫКА НА МАТЕРИАЛЕ УСТНОГО КОРПУСА

А. С. Шатских¹ [0009-0006-9025-7783], А. А. Сорокин² [0000-0003-3877-4223]

¹Институт языкознания РАН, г. Москва, Россия

²Московский государственный университет имени М. В. Ломоносова,
г. Москва, Россия

²Яндекс, г. Москва, Россия

¹anapriselec@gmail.com, ²alexey.sorokin@list.ru

Аннотация

Работа посвящена созданию первого корпуса для осетинского языка с морфологической разметкой в формате Universal Dependencies. Корпус состоит из размеченных вручную текстов Устного корпуса осетинского языка, содержит 5454 предложения и имеет суммарный объем 73042 токена. Представлен морфологический анализатор на основе архитектуры BERT, достигающий пословной корректности 95.60%. Приведены результаты экспериментов по улучшению качества классификации. Сделан вывод, что фильтрация выдачи модели с помощью бесконтекстного анализатора и многозадачный подход не приводят к значительному улучшению. Представлены также результаты тестирования анализатора на внедоменных данных и продемонстрировано что на них модель достигает пословной корректности 91.45%.

Ключевые слова: осетинский язык, иронский диалект, стандарт морфологической аннотации, морфологический анализ.

ВВЕДЕНИЕ

Осетинский язык является представителем индоевропейской семьи (иранская группа) индоевропейской языковой семьи. Язык распространен преимущественно на Северном Кавказе. Согласно Всероссийской переписи населения 2020 г., число его носителей на территории России составляет 344907 человек. На сегодняшний день существуют два основных корпуса осетинского

языка: Осетинский национальный корпус (ОНК)¹ и Устный корпус осетинского (УКО)².

Осетинский национальный корпус содержит литературные тексты преимущественно современных авторов (1999–2004 гг.), а также произведения наиболее известных авторов XX в. Кроме того, он включает тексты литературных журналов «Мах дуг» («Наша эпоха») и «Ногдзау» («Пионер») и материалы интернет-газеты «Спутник», а также наиболее полное собрание Нартовского эпоса. Объем корпуса составляет приблизительно 12 млн словоупотреблений. В корпусе автоматически размечены части речи, леммы и морфологические признаки; аннотация выполнена с помощью анализатора Uniparser-Morph [1]. Поскольку анализатор применяется бесконтекстно, в корпусе не снята омонимия, и для каждого словоупотребления предлагается множество всех возможных разборов.

Устный корпус осетинского содержит устные тексты, собранные в ходе лингвистических экспедиций 2007–2013 гг. в различных районах Республики Северная Осетия – Алания. Корпус также содержит выпуски радиопередачи «Одноклассники». Объем корпуса составляет приблизительно 60 тыс. словоупотреблений. Он снабжен поморфемными разборами (глоссами), которые выполнены вручную в программе SIL FieldWorks Language Explorer.

Несмотря на то что из формата FieldWorks возможно извлечь частеречные метки и морфологические признаки, аннотация УКО является неполной и недостаточно подробной. С одной стороны, это затрудняет автоматическое приведение данных к международному стандарту морфологической аннотации Universal Dependencies [2], с другой – это препятствует использованию корпуса в образовательных и исследовательских целях. Таким образом, существует необходимость в создании корпуса осетинского языка с морфологической разметкой со снятой омонимией, который бы соответствовал стандартам Universal Dependencies версии 2 (UD v2). Настоящее исследование направлено на решение этой задачи. Настоящая статья является расширенной версией работы, представленной на конференции «Диалог-2026» [3].

¹<http://corpus.ossetic-studies.org> (дата обращения 14.08.2026)

²<https://www.ossetic-studies.org/ru/texts/iron> (дата обращения 14.08.2026).

ЛИНГВИСТИЧЕСКИЕ СВЕДЕНИЯ

Осетинский язык подразделяется на два диалекта: иронский и дигорский. В настоящей работе термины «осетинский язык» и «иронский диалект осетинского языка» используются как синонимы.

Осетинский язык обладает богатой морфологией. Именная морфология является агглютинативной, однако демонстрирует регулярные чередования гласных и согласных при образовании форм множественного числа [4]. Глагольная морфология имеет ряд синтетических черт, таких как наличие словоизменяемых классов, супплетивное или нерегулярное образование основы прошедшего времени, а также кумулятивное выражение лица и числа. Продуктивным способом образования новых глаголов является создание сложных глаголов. Сложный глагол состоит из именной части и легкого глагола, который выбирается из закрытого класса.

Порядок слов в клаузе свободный, немаркированным порядком является SOV. Как правило, вершина составляющей в осетинском языке расположена справа. Порядок слов в именной группе фиксирован и следует иерархии Poss > Det > Num > Adj, Gen > N [5], где Poss – генитив посессора, Gen – низкий нереперентный генитив.

ИСТОЧНИК ДАННЫХ

Цель настоящего исследования – создание корпуса с морфологической аннотацией для осетинского языка. В качестве источника данных выбран корпус УКО, поскольку он уже содержит ручную морфологическую аннотацию в нестандартном формате со снятой омонимией. Исходные морфологические метки приводятся к формату CoNLL-U согласно правилам разметки UD v2 [2] и их расширениям, которые описаны в следующем разделе.

Приведение данных УКО к формату CoNLL-U включает следующие этапы:

- 1) фильтрация данных не на целевом языке, в частности удаление всех полностью русскоязычных предложений;
- 2) ручная сегментация предложений по многоточию и символу верхнего регистра, а также сегментация многословных токенов;
- 3) ручное приведение меток частей речи и морфологических признаков к универсальному перечню;

4) ручное исправление аннотации корпуса в соответствии с универсальными правилами разметки и их расширениями. Этот этап требует знания осетинской грамматики и является наиболее трудоемким;

5) автоматическое приведение данных к формату CoNLL-U, включая заполнение по умолчанию тех значений признаков, которые отсутствуют в исходных по-морфемных разборах (например, именительный падеж существительных).

Итоговый датасет состоит из 5454 предложений и имеет суммарный объем 73 042 токена. Датасет разделен на обучающую, валидационную и тестовую выборки в отношении 8:1:1. Распределение частот поджанров в каждой из подвыборок приведено в Приложении А, распределение частот морфологических меток – в Приложении Б.

ПРАВИЛА АННОТАЦИИ

Ряд особенностей осетинской грамматики требует расширения стандартного набора признаков UD v2 и их значений. В этом разделе описаны расширения, принятые в исходной системе аннотации корпуса и введенные нами в настоящем исследовании.

Многие исследователи (например, В. И. Абаев [4]) отмечают, что в осетинском языке не существует четких границ между существительными, прилагательными, наречиями и послелогоми. Одна и та же последовательность символов может выполнять функции различных частей речи; так, например, слово «хорз» может быть переведено как «хороший» или как «хорошо». В УКО существительным, прилагательным, послелогам и некоторым наречиям присваивается одна и та же метка части речи. Система UD v2, в свою очередь, задает универсальный инвентарь частей речи, а часть речи словоформы в рамках этой системы определяется на основании морфосинтаксических критериев. Для различения послелогов, наречий, прилагательных и существительных приняты следующие эвристики:

1) имена без падежно-числовой морфологии, модифицирующие глагол и управляющие генитивом или посессивной проклитикой, считаются послелогам (ADP) (например, «бын» в примере 1);

2) имена без падежно-числовой морфологии, модифицирующие глагол, считаются наречиями (ADV) (например, «рæсугъд» в примере 2);

3) имена без падежно-числовой морфологии, модифицирующие другие имена (кроме закрытого класса интенсификаторов типа тынг 'очень'), считаются прилагательными (ADJ) (например, «ирон» в примере 3);

4) имена с падежно-числовой морфологией и имена без падежно-числовой морфологии, не покрытые в предыдущих пунктах, считаются существительными (NOUN).

Приведём примеры.

(1) *Æмæ зæй-ы бын фе-сты адæм...*
и лавина-gen под pv-быть.3pl человек.pl
'И под лавиной оказались люди...'

(2) *Ды тынг рæсугъд дзур-ыс.*
2sg очень красиво говорить-prs.2sg
'Ты очень красиво говоришь.'

(3) *Фæлæ ирон æгъдæу-тт-æ...*
но осетинский обычай-pl-nom
'Но осетинские обычаи...'

В именных предикациях и сложных глаголах синтаксические критерии, описанные выше, неприменимы. В таких случаях имя по умолчанию считается существительным; в случае, если имя имеет показатель сравнительной степени или зависимый интенсификатор, оно относится к прилагательным (ADJ). Так, прилагательным считается «кадджын» в примере 4.

(4) *Æмæ гъе уыдон ны-ххаст-а, æмæ уый тынг*
и ну DemDist.pl pv-брать-pst.3sg и DemDist очень
кадджын уыд-и.
почетный быть-pst.3sg
'И он взял их, и это было очень почетно.'

В осетинском языке существует закрытый класс глаголов, способных функционировать в качестве вспомогательных; однако все представители этого класса также могут выступать в роли лексических глаголов. Хотя в исходной разметке УКО глаголы в лексической и вспомогательной функциях аннотируются одинаково, мы утверждаем, что их различие не только методологически необходимо, но и осуществимо на основе синтаксических критериев. Глагол считается вспомогательным (AUX) в следующих конфигурациях:

- 1) легкий глагол в сложных глаголах (например, «кодтат» в примере 5);
- 2) глагол-связка в именных предикациях (например, «уыди» в примере 6);
- 3) собственно вспомогательный глагол, т. е. финитный глагол в аналитических глагольных конструкциях (например, «нæй» в примере 7).

Приведём примеры.

(5) *Куыд ахуыр код-т-ат скъола-йы?*

как учеба делать-tr-pst.2pl школа-in

‘Как вы учились в школе?’

(6) *Алагирский район-ы хицау уыд-и.*

Алагирский район-gen хозяин быть-pst.3sg

‘Он был главой Алагирского района.’

(7) *Уым дæр нæй фыст.*

DemDist.in add быть.neg.prs.3sg писать.ptcp.pst

‘Этого там тоже не написано.’

Для классификации местоименных единиц на местоимения-существительные (PRON) и местоимения-прилагательные (DET) применяются те же морфосинтаксические критерии, которые используются для различения существительных (NOUN) и прилагательных (ADJ). Решение обусловлено тем, что лексемы, обычно выступающие как именные модификаторы, могут подвергаться субстантивации, т. е. возглавлять именную группу, выступать в роли аргумента глагола или выражать грамматические категории существительных (например, «ахæмтæ» в примере 8). В таких случаях им присваиваются метка PRON и признаки падежа и числа. Те же принципы применяются к лексеме-плейсхолдеру «йед» (как в примере 9). Следует отдельно отметить, что у лексемы «йед» отсутствует признак местоименного разряда (PronType), так как в инвентаре значений этого признака отсутствует подходящее.

Рассмотрим примеры.

(8) *Уыд-и дзы ахæм-т-æ, æртын фондз*

быть-pst.3sg 3sg.in такой-pl-nom тридцать пять

аз-ы кæ-уыл цыд-и.

год-num кто-super идти-pst.3sg

‘Были там и те, кому было 35 лет.’

(9) *Æмæ дын йед-мæ куы ба-хæццæ дæн...*

и 2SG.DAT HES-ALL когда PV-смешанный быть.PRS.1SG

бын-мæ...

низ-ALL

‘И когда я пришла туда... вниз...’

Субстантивация местоименных наречий в первую очередь происходит путем присоединения падежной морфологии. При этом получающиеся словоформы продолжают функционировать как модификаторы имен или глаголов. Чтобы подчеркнуть отличие субстантивированных наречий от полноценных случаев субстантивации, описанных выше, мы не присваиваем признаку числа значение по умолчанию (Number = Sing) и аннотируем этот признак только тогда, когда число выражено ненулевым показателем (например, «афтæтæ» в примере 10).

(10) Ам дæр афтæ-т-æ нæ уыд-ысты.

DemProx.IN ADD так-PL-NOM NEG быть-3PL

‘И здесь так не было.’

В именной парадигме можно выделить от девяти до двенадцати падежей [6]. Директив и рецессив, которые были грамматикализованы позднее других, отсутствуют в классических грамматиках; однако мы включаем их в парадигму в соответствии с отмеченным исследованием [6]. Для директива и рецессива вводятся дополнительные значения признака Case с названиями Dir и Rcs.

Статус нумератива как отдельного падежа для существительных в конструкциях с числительными остается предметом дискуссий. В первую очередь это связано с тем, что показатель счетной формы (-ы) омонимичен показателю генитива. Однако, можно привести как минимум два убедительных аргумента в пользу его включения в падежную парадигму.

Во-первых, в дигорском диалекте осетинского языка показатель счетной формы может сочетаться с любым падежом. Это указывает на то, что нумератив представляет собой отдельную грамматическую функцию, а не одну из функций генитива. С другой стороны, в то время, как дигорский диалект предоставляет достаточные основания для отнесения нумератива к категории числа, в иронском диалекте таких оснований недостаточно, при этом необходимость различать нумератив и генитив существует.

Во-вторых, в иронском диалекте у существительных *pluralia tantum* нумератив омонимичен номинативу, а не генитиву. Согласно двум приведенным соображениям, Num вводится в качестве дополнительного значения признака падежа.

Когда конструкция с числительным выступает в роли прямого дополнения, для определения падежа существительного применяются правила дифференцированного маркирования объекта (DOM) [7]. Согласно этим правилам одушевленные прямые дополнения, обозначающие людей, маркируются генитивом, неодушевленные прямые дополнения – номинативом. Соответственно, в конструкциях с числительными генитив присваивается существительным, обозначающим людей (например, одушевленное *лæппуы* в примере 11), нумератив – неодушевленным существительным (например, неодушевленное *рæнхъы* в примере 12). Поскольку правила DOM для одушевленных объектов, которые не обозначают людей, недостаточно описаны, такие имена обрабатываются аналогично неодушевленным существительным согласно эвристике (например, *карчы* в примере 13).

Приведём примеры.

- (11) *Уыцы дыууæ лæппу-йы с-хъомыл код-т-а.*
 DemDist два мальчик-GEN PV-взрослый делать-TR-
 PST.3SG

‘Она вырастила этих двух мальчиков.’

- (12) *...аст рæнхъ-ы ны-ффыст-он.*
 восемь строка-NUM PV-писать-PST.1SG

‘...я написала восемь строк.’

- (13) *...æмæ уу-дæс карч-ы а-ргæвст-он,*
 и один-десять курица-NUM PV-забивать-PST.1SG
ууæндæс карч-ы.
 одиннадцать курица-NUM

‘...и забила около десяти куриц, одиннадцать куриц.’

Широкий перечень нефинитных глагольных форм в осетинском языке требует введения новых значений для признака VerbForm. Введены значения Inf2, Part2, Dest и Red. Соответствие нефинитных форм значениям признака VerbForm приведено в табл. 1 на примере глагола *дзурын* ‘говорить’.

Табл. 1. Нефинитные глагольные формы.

Форма	Значение VerbForm	Функция
дзур-ын	Inf	Инфинитив
дзур-æн	Inf2	II инфинитив
дзырд	Part	Причастие прошедшего времени (страдательное)
дзур-инаг	Part	Причастие будущего времени
дзур-æг	Part	Причастие настоящего времени (действительное)
дзур-аг	Part2	Причастие настоящего времени (действительное, хабитуальное)
дзур-гæ	Conv	Конверб
дзур-æггаг	Dest	Дестинативное причастие
ра-дзур-ба-дзур	Red	Номинализация в сложных глаголах

Использование для осетинских глагольных форм признака вида (Aspect) привело бы к формированию неоднородных категорий, не соответствующих традиции описания языка. Вместо этого предлагается использовать признак Preverb = Yes для словоформ, содержащих глагольные приставки (превербы), поскольку именно они выражают видовые значения (например, æрбахонон в примере 14).

(14) *Æрба-хон-он дын æй?*
 PV-звать-SBJV.1SG 2SG.DAT 3SG.GEN
 ‘Мне позвать ее?’

Кроме того, в языке существуют единицы æд и æнæ, которые могут быть отражены на письме как именные приставки или предлоги (см. æнæ в примерах 15 и 16). Нами вводится признак Prenoun = Yes и используется в первом случае для указания на наличие именной приставки в словоформе.

Представим примеры.

(15) *Адæм æнæ-хæлд гъе уæд уыд-ысты.*

Человек.PL NEG-портить.PTCP.PST ну тогда быть-PST.3PL

‘Люди тогда не были испорченными.’

(16) *Æз та æнæ очки-т-æй нæ уыд-т-он бинаг*
1SG PTCL без очки-PL-ABL NEG видеть-TR-PST.1SG нижний
фыст.

писать.PTCP.PST

‘Без очков я не увидела нижнюю надпись.’

СОЗДАНИЕ МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА

Помимо создания корпуса с морфологической аннотацией, целью нашего исследования было создание морфологического анализатора, обученного на данных этого корпуса. Учитывая успех подхода трансферного обучения и предобученных энкодерных языковых моделей в задаче морфологического анализа (см., в частности, [8, 9]), мы предложили анализатор на основе архитектуры BERT [10]. Поскольку языковые модели для осетинского языка отсутствуют, модель архитектуры BERT обучается стандартной задаче маскированного языкового моделирования, а затем дообучается задаче морфологического анализа. Предобученная модель архитектуры BERT также может быть использована для других задач автоматической обработки текста на осетинском языке.

Для предобучения используются неразмеченные данные ОНК. Данные проходят автоматическую предобработку, которая состоит из следующих этапов:

- 1) NFC-нормализация;
- 2) локальная дедупликация путем удаления повторяющихся строк в пределах одного документа;
- 3) исключение фрагментов HTML и JavaScript с помощью регулярных выражений;
- 4) глобальная дедупликация строк, встречающихся более 50 раз во всем корпусе;
- 5) фильтрация документов по порогу минимальной длины в 200 символов;
- 6) применение алгоритма LSH (Locality Sensitive Hashing) для обнаружения и удаления приближенных дубликатов.

Предбработанные данные разделяются на предложения с помощью регулярного выражения, и в них нормализуется регистр. Кроме того, удаляются предложения, соответствующие по крайней мере одному из следующих критериев:

- предложение содержит только издательскую информацию или метаданные;
- в предложении отсутствуют кириллические символы или символы нижнего регистра;
- в предложении содержатся кириллические и латинские символы в пределах одного токена (за исключением латинского символа æ).

В качестве кандидатов на оптимальную отправную точку используются следующие модели:

- 1) mBERT: мультиязычная модель архитектуры BERT от Google среднего размера, учитывающая регистр³;
- 2) осетинская модель BERT: версия mBERT с замененным слоем эмбеддингов, который адаптирован к осетинскому языку согласно методу [11].

Для совместимости мультиязычной модели с символьным составом обучающей выборки стандартная нормализация токенизатора mBERT дополняется следующими шагами:

- кириллический символ æ приводится к его латинскому аналогу, поскольку только последний присутствует в словаре модели mBERT;
- различные варианты написания кавычек и апострофа (U+2018, U+2019, U+02BC, U+0027, U+00B4, U+0060) приводятся к символу U+02BC. Единообразие представлений различных вариантов критически важно, поскольку символ апострофа или кавычки в осетинском языке выполняют функцию показателя сандхи на границе слов.

Инициализация весов осетинской модели BERT производится следующим образом. На данных ОНК обучается токенизатор системы WordPiece с размером словаря в 25000 токенов. Для сохранения знаний исходной модели каждое вхождение из словаря нового токенизатора разбивается с помощью токенизатора исходной модели и эмбеддинги полученных токенов усредняются. После

³<https://huggingface.co/google-bert/bert-base-multilingual-cased> (дата обращения 14.08.2026)

этого слой эмбедингов исходной модели заменяется на конкатенацию эмбедингов вхождений осетинского словаря.

Обе модели дообучаются задаче маскированного языкового моделирования в течение пяти эпох с размером батча 64 и темпом обучения $5e-5$. Значения функции потерь моделей после обучения представлены в табл. 2. Отметим, что эти значения нельзя сравнивать напрямую, поскольку словари моделей несопоставимы. Это отражено в среднем числе токенов на слово, которое измерено на тестовой выборке датасета Ossetic-COT и также приведено в табл. 2.

Табл. 2. Результаты предобучения моделей.

Модель	Значение функции потерь	Среднее число токенов на слово
Мультиязычная	0.6086	3.29
Осетинская	3.4311	1.21

В табл. 3 представлены результаты дообучения в течение 10 эпох для мультиязычной модели и в течение 13 эпох для осетинской модели с размером батча 8 и темпом обучения $5e-5$. Значения усреднены по результатам пяти независимых запусков обучения. Для каждой модели оптимальные гиперпараметры подобраны с помощью валидационной выборки, и результаты приведены только для наилучшего набора гиперпараметров.

Табл. 3. Результаты дообучения моделей.

Модель	Корректность	Корректность по предложениям
Мультиязычная	95.60	61.54
Осетинская	95.47	60.11

Поскольку различие между моделями в корректности классификации незначительно, в качестве итоговой модели выбран классификатор на основе мультиязычной модели BERT (mBERT). Значения точности, полноты и F-меры для этой модели приведены в табл. 4. Таблица содержит значения, усредненные по

всем классам с помощью макроусреднения и взвешенного усреднения, а также значения, усредненные по всем классам, соответствующим каждой части речи, с помощью макроусреднения и взвешенного усреднения. Все значения также усреднены по результатам пяти независимых запусков обучения.

Табл. 4. Расширенные метрики мультязычной модели.

Часть речи	Точность		Полнота		F-мера	
	Макро	Вес	Макро	Вес	Макро	Вес
ADJ	0.78	0.91	0.81	0.92	0.79	0.91
ADP	0.99	0.99	0.96	0.96	0.97	0.97
ADV	0.84	0.95	0.88	0.97	0.86	0.96
AUX	0.64	0.86	0.66	0.88	0.64	0.86
CCONJ	0.99	0.99	0.99	0.99	0.99	0.99
DET	0.89	0.96	0.87	0.93	0.88	0.94
INTJ	0.97	0.98	0.97	0.98	0.97	0.98
NOUN	0.78	0.93	0.78	0.94	0.78	0.93
NUM	0.85	0.99	0.85	0.98	0.85	0.98
PART	0.97	0.98	0.99	0.99	0.98	0.99
PRON	0.81	0.92	0.82	0.92	0.80	0.92
PROPN	0.44	0.95	0.50	0.96	0.46	0.96
PUNCT	1.00	1.00	1.00	1.00	1.00	1.00
SCONJ	0.93	0.86	0.91	0.82	0.92	0.84
VERB	0.77	0.92	0.79	0.92	0.77	0.92
X	0.87	0.87	0.66	0.66	0.75	0.75
Макро	0.77		0.78		0.77	
Вес	0.95		0.96		0.95	

Для фильтрации выдачи модели на этапе получения предсказаний создается бесконтекстный морфологический анализатор на основе конечного преобразователя в программе Foma [12], а также список основ из закрытых классов и словоформ из закрытых классов для единиц с нерегулярными парадигмами. Алгоритм фильтрации для одного токена состоит из следующих этапов.

1. Поиск словоформы в списке словоформ из закрытых классов.

2. Порождение множества разборов с помощью бесконтекстного анализатора.
3. Поиск основы из каждого из разборов в списке основ из закрытых классов.
4. Получение распределения вероятностей по классам из нейронной сети для классификации.
5. Приведение к нулю вероятностей всех классов, не входящих в множество бесконтекстных разборов для данного токена.
6. Выбор наиболее вероятного класса из распределения в качестве предсказания для данного токена.

Из токенов тестовой выборки лишь для меньше 1% наиболее вероятный класс в распределении нейросетевого классификатора исключается бесконтекстным анализатором. Таким образом, фильтрация представляется избыточной и не используется в итоговой модели.

Для качественной оценки работы модели использован набор псевдоминимальных пар предложений для основных случаев морфологической омонимии в осетинском языке. Каждая пара состоит из предложений, где целевое слово имеет различный верный разбор, при этом окружающий контекст отличается минимально. Задачей классификатора является предсказание правильного разбора в обоих случаях. Тестовый набор представлен в Приложении В, целевые слова выделены полужирным. В табл. 5 приведены предсказания модели для каждой пары в тестовом наборе.

Табл. 5. Разрешение омонимии мультязычной моделью.

Часть речи	Контекст 1	Контекст 2
Adj/Noun	Adj	Noun
Adj/Adv	Adj	Adv
Noun/Adp	Noun	Noun
1SG/INF	1SG	INF
GEN/IN	GEN	IN
GEN/NUM (неодуш.)	GEN	NUM
GEN/NUM (одуш.)	NUM	NUM

IN/PRS.3SG	GEN	GEN
DemDist.NOM/DemDist.GEN	DemDist.NOM	DemDist.GEN
1PL.NOM/1PL.GEN	1PL.NOM	1PL.GEN
1SG.POSS/1SG.GEN (ЭНкл.)	1SG.POSS	1SG.GEN (ЭНкл.)
1SG.POSS/союз 'и'	1SG.POSS	союз 'и'
2SG.POSS/2SG.GEN (ЭНкл.)	2SG.POSS	2SG.GEN (ЭНкл.)
2SG.GEN (ЭНкл.)/быть.2SG	2SG.GEN (ЭНкл.)	быть.2SG
1PL.POSS/1PL.GEN (ЭНкл.)	1PL.POSS	NEG
1PL.POSS/NEG	1PL.POSS	NEG
EMPH/NEG	EMPH	NEG
ADD/NEG	ADD	NEG
ITER/'один'	ITER	'один'

Анализ результатов показывает, что наиболее сложными являются следующие случаи омонимии: Noun/Adp (существительное и послелог), GEN/NUM (генитив и нумератив) для существительных, обозначающих людей, 1PL.GEN/NEG (генитив местоимения 1-го лица единственного числа и показатель отрицания) и IN/PRS.3SG (инессив и показатель 3-го лица единственного числа настоящего времени глагола).

Отметим, что для некоторых случаев омонимии предсказанная метка отличается для разных запусков обучения. Кроме того, незначительные лексические изменения в тестовых примерах могут приводить к предсказанию неверного разбора для целевых словоформ. Эти явления могут быть обусловлены количественными и качественными ограничениями датасета Ossetic-COT. Отмеченные ограничения можно преодолеть с помощью обогащения обучающей выборки более объемными и лексически разнообразными языковыми данными.

Детальный анализ ошибок модели на тестовой выборке позволяет выделить следующие основные классы ошибок:

- неразрешимая неоднозначность, или случаи, в которых контекст не позволяет определить верный разбор либо существует более одного верного разбора;
- ошибки из-за несбалансированности данных, или случаи, когда для словоформы существует более одного верного разбора, при этом один из них значительно более частотен в обучающей выборке, чем другие, и ошибочно предсказывается вместо других;
- орфографические ошибки, или случаи, когда моделью предсказан верный разбор для словоформы, похожей на целевую, но не для целевой;
- дистантные зависимости, или случаи, когда фактор, определяющий верный разбор, находится в предложении на значительном расстоянии от целевой словоформы.

Стоит отметить языковые категории, представляющие наибольшую сложность. Ошибки возникают при определении нефинитных глагольных форм, а также при разграничении вспомогательных и лексических употреблений глаголов.

ОГРАНИЧЕНИЯ

Примечательно, что как валидационная, так и тестовая выборки содержат комбинации признаков, отсутствующие в обучающей выборке. Более того, подмножество теоретически возможных комбинаций признаков не встречается ни в одной из подвыборок. Проблема распознавания таких незнакомых классов может быть решена с помощью подхода многозадачного обучения (см., например, [13, 9]). При таком подходе значение каждого морфологического признака предсказывается моделью отдельно. В табл. 6 представлены результаты обучения многозадачных морфологических анализаторов на основе мультязычной и осетинской моделей BERT. Использованы размер батча 8 и темп обучения $5e-5$, обучение длится 25 эпох для мультязычной модели и 20 эпох для осетинской модели. Можно видеть, что использование многозадачного подхода не позволяет существенно превзойти показатели классификатора, предсказывающего цельные морфологические метки. Таким образом, приоритет отдается последнему в силу его простоты и вычислительной эффективности.

Табл. 6. Результаты дообучения многозадачных моделей.

Модель	Корректность	Корректность по предложениям
Мультиязычная	95.54	62.39
Осетинская	95.45	60.92

Как было отмечено при анализе ошибок, корпус обладает ограниченным лексическим разнообразием. Более того, синтаксические структуры в большинстве предложений либо просты, либо характерны для устной речи. Эти особенности данных приводят к более низкому качеству при применении модели к данным из других источников и других жанров. Так, при измерении качества модели на выборке из 200 предложений из осетинской Википедии корректность по словам составляет 91.45, корректность по предложениям – 41.00. Наиболее часто ошибки совершаются при распознавании сокращений, собственных имен и числительных, записанных цифрами. Адаптация анализатора к данным различных жанров остается направлением для будущих исследований.

ЗАКЛЮЧЕНИЕ

Представлен первый корпус осетинского языка с морфологической аннотацией, выполненной в соответствии со стандартом Universal Dependencies версии 2. Корпус подходит как для обучения нейросетевых морфологических анализаторов, так и для исследовательских и образовательных целей, поскольку его разметка отражает большинство морфологических признаков и различий, присущих языку. Кроме того, представлена базовая модель, которая может служить для оценки сложности датасета и является первым нейросетевым морфологическим анализатором для осетинского языка. Корпус и модель находятся в свободном доступе⁴⁵.

Основным направлением будущих исследований является преодоление текущих количественных и качественных ограничений корпуса путем его обогащения разнообразными литературными текстами из ОНК.

⁴<https://github.com/ania3000/Ossetic-COT> (дата обращения 14.08.2026)

⁵<https://huggingface.co/ossetic-encoders/ossbert-morph> (дата обращения 14.08.2026)

Благодарности

Авторы выражают благодарность Олегу Игоревичу Беляеву за идею исследования и за помощь в разработке специфичных для языка правил аннотации, а также Олегу Игоревичу Беляеву и Арсению Павловичу Выдрину за предоставление неразмеченных данных ОНК.

Исследование поддержано проектом FMNE-2025-0002, выполняемом в Институте языкознания РАН.

Список сокращений

1 – 1 лицо, 2 – 2 лицо, 3 – 3 лицо, ABL – аблатив, ADD – аддитивная частица, ALL – аллатив, DAT – датив, DemDist – дальний дейктик, DemProx – ближний дейктик, GEN – генитив, HES – плейсхолдер, IN – инессив, INF – инфинитив, ITER – итеративная частица, NEG – показатель отрицания, NOM – номинатив, NUM – нумератив, PL – множественное число, POSS – притяжательное местоимение, PRS – настоящее время, PST – прошедшее время, PTCL – частица, PTCP – причастие, PV – преверб, SVJV – конъюнктив, SG – единственное число, SUPER – суперэссив, TR – показатель переходности.

СПИСОК ЛИТЕРАТУРЫ

1. Arkhangelskiy T., Belyaev O., Vydrin A. The Creation of Large-Scale Annotated Corpora of Minority Languages using UniParser and the EANC platform. // Proc. of COLING 2012: Posters, Mumbai, India. The COLING 2012 Organizing Committee: 2012. P. 83–92.
2. Nivre J., de Marneffe M., Ginter F., Hajic J., Manning C., Pyysalo S., Schuster S., Tyers F., Zeman D. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. // Proc. of the Twelfth Language Resources and Evaluation Conference, Marseille, France. European Language Resources Association: 2020. P. 4034–4043.
3. Shatskikh A., Sorokin A. Ossetic-COT: Designing a morphologically annotated corpus and morphological analyzer for Ossetic. // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue” (2026). 2026. P. 544-555.
4. Abaev V. A grammatical sketch of Ossetic. The Hague: Mouton, 1964.

5. *Беляев О.И.* Коррелятивная конструкция в осетинском языке (диссертация на соискание степени кандидата филологических наук). 2014.
6. *Belyaev O.* Evolution of Case in Ossetic // Iran and the Caucasus. 2010. No. 14 (2). P. 287–322.
7. *Serdobolskaya N., Tuzhik O.* Differential Object Marking in Modern Ossetic: Referential Properties and Animacy // Tomsk Journal of Linguistics and Anthropology. 2025. No. 48 (2). P. 91–107.
8. *Kondratyuk D., Straka M.* 75 Languages, 1 Model: Parsing Universal Dependencies Universally // Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China. Association for Computational Linguistics: 2019. P. 2779–2795.
9. *Straka M., Straková J., Hajic, J.* UDPipe at SIGMORPHON 2019: Contextualized Embeddings, Regularization with Morphological Categories, Corpora Merging. // Proc. of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, Florence, Italy. Association for Computational Linguistics: 2019. P. 95–103.
10. *Devlin J., Chang M., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, Minneapolis, Minnesota. Association for Computational Linguistics: 2020. P. 4171–4186.
11. *Kuratov Y., Arkhipov M.* Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. 2019.
12. *Hulden M.* Foma: a Finite-State Compiler and Library // Proc. of the Demonstrations Session at EACL 2009, Athens, Greece. Association for Computational Linguistics: 2009. P. 29–32.
13. *Inoue G., Shindo H., Matsumoto Y.* Joint Prediction of Morphosyntactic Categories for Fine-Grained Arabic Part-of-Speech Tagging Exploiting Tag Dictionary Information // Proc. of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada. Association for Computational Linguistics: 2017. P. 421–431.

ПРИЛОЖЕНИЕ А. РАСПРЕДЕЛЕНИЕ ПОДЖАНРОВ

	Обучающая	Валидационная	Тестовая
Радио	253 (6.15%)	37 (6.79%)	29 (5.32%)
Экспедиции	4111 (93.85%)	508 (93.21%)	516 (94.68%)

ПРИЛОЖЕНИЕ Б. РАСПРЕДЕЛЕНИЕ МОРФОЛОГИЧЕСКИХ МЕТОК

	Обучающая	Валидационная	Тестовая
PUNCT	14913 (25.58%)	1865 (24.92%)	1824 (25.16%)
NOUN, Case = Nom Number = Sing	3269 (5.61%)	450 (6.01%)	444 (6.12%)
PART	3236 (5.55%)	419 (5.60%)	394 (5.43%)
CCONJ	2794 (4.79%)	351 (4.69%)	355 (4.90%)
PROPN, Case = Nom	1888 (3.24%)	253 (3.38%)	277 (3.82%)
ADV	1027 (1.76%)	109 (1.46%)	127 (1.75%)
ADV, PronType = Int	1009 (1.73%)	112 (1.50%)	116 (1.60%)
NOUN, Case = Nom Number = Plur	1003 (1.72%)	126 (1.68%)	140 (1.93%)
ADV, PronType = Dem	999 (1.71%)	123 (1.64%)	139 (1.92%)
PART, Polarity = Neg	988 (1.69%)	111 (1.48%)	124 (1.71%)
ADJ	882 (1.51%)	120 (1.60%)	104 (1.43%)
ADV, Deixis = Remt PronType = Dem	851 (1.46%)	132 (1.76%)	111 (1.53%)
VERB, Mood = Ind Number = Sing Person = 3 Tense = Pres VerbForm = Fin	830 (1.42%)	100 (1.34%)	85 (1.17%)
VERB, Mood = Ind Number = Sing Person = 3 Preverb = Yes Tense = Past VerbForm = Fin	749 (1.28%)	95 (1.27%)	87 (1.20%)

INTJ	714 (1.22%)	113 (1.51%)	132 (1.82%)
ADV, Deixis = Prox PronType = Dem	701 (1.20%)	77 (1.03%)	92 (1.27%)
VERB, Mood = Ind Number = Sing Person = 3 Tense = Past VerbForm = Fin	673 (1.15%)	96 (1.28%)	98 (1.35%)
PRON, Case = Gen Number = Sing Person = 3 PronType = Prs	643 (1.10%)	82 (1.10%)	88 (1.21%)
NUM, Case = Nom NumType = Card	610 (1.05%)	63 (0.84%)	81 (1.12%)
NOUN, Case = Gen Number = Sing	590 (1.01%)	95 (1.27%)	66 (0.91%)

ПРИЛОЖЕНИЕ В. ТЕСТОВЫЙ НАБОР ДЛЯ РАЗРЕШЕНИЯ ОМОНИМИИ

	Контекст 1	Контекст 2
Adj/Noun	Зæронд ус æрбацыди.	Зæронд æрбацыди.
Adj/Adv	Уый рæсугъд чызгимæ кафы.	Уый рæсугъд кафы.
Noun/Adp	Лæппу хæдзары сæр уыны.	Лæппу хæдзары сæр бады.
1SG/INF	Æз кусын.	Æз кусын байдыдтон.
GEN/INE	Залинæ хæдзары мидæг уыди.	Залинæ хæдзары уыди.
GEN/NUM (неодуш.)	Залинæ хæдзары хицауы федта.	Залинæ дыууæ хæдзары федта.
GEN/NUM (одуш.)	Дыууæ лæджы федтон.	Дыууæ лæджы æрбацыдысты.
INE/PRS.3SG	Мах хъæуы барæгбон кæнæм.	Мах хъæуы барæгбон.
DemDist.NOM/DemDist.GEN	Уый æрбацыдис.	Уый фæстæ æрбацыдтæн.
1PL.NOM/1PL.GEN	Мах уым цæрæм.	Мах бинонтæ уым цæрынц.

1SG.POSS/1SG.GEN (энкл.)	Уый мæ мады федта.	Уый мæ федта.
1SG.POSS/союз 'и'	Мæ чызг уым ахуыр кодта.	Лæппу мæ чызг уым ахуыр кодтой.
2SG.POSS/2SG.GEN (энкл.)	Уый дæ фыды федта.	Уый дæ федта.
2SG.GEN (энкл.)/быть.2SG	Ирон лæг дæ федта.	Ирон лæг дæ .
1PL.POSS/1PL.GEN (энкл.)	Уый нæ чызджы федта.	Уый нæ федта.
1PL.POSS/NEG	Уый нæ хæдзар федта.	Уый хæдзар нæ федта.
EMPH/NEG	Радзур- ма исты.	Ма дзур.
ADD/NEG	Хъæумæ ма цыдтæн.	Хъæумæ ма цæу.
ITER/'один'	Æмæ- иу лæг æрбацыдис.	Æмæ иу лæг æрбацыдис.

AUTOMATIC MORPHOLOGICAL TAGGING FOR OSSETIC BASED ON DATA FROM THE CORPUS OF ORAL TEXTS

A. S. Shatskikh¹ [0009-0006-9025-7783], A. A. Sorokin² [0000-0003-3877-4223]

¹*Institute of Linguistics of the Russian Academy of Sciences, Moscow, Russia*

²*Lomonosov Moscow State University, Moscow, Russia*

²*Yandex, Moscow, Russia*

¹anapriselec@gmail.com, ²alexey.sorokin@list.ru

Abstract

In this work we present the first morphologically annotated corpus for Iron Ossetic that conforms to the Universal Dependencies schema. The corpus includes 5454 manually annotated sentences from the Iron Ossetic Corpus of Oral Texts, containing 74032 tokens. We use this corpus to train a BERT-based morphological analyzer. The analyzer achieves tag accuracy of 95.60%. Furthermore, the paper presents the results of experiments aimed at improving classification quality. It is shown that neither filtering the model's output using a context-free analyzer nor multi-task approach lead to a significant improvement. Finally, the paper presents the results of testing the analyzer on out-of-domain data and shows that the model achieves tag accuracy of 91.45% on them.

Keywords: *Iron Ossetic, Universal Dependencies, Morphological Tagging, Morphological Analysis*

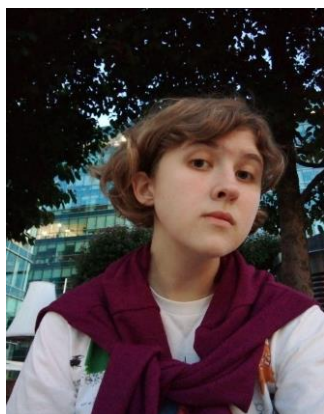
REFERENCES

1. Arkhangelskiy T., Belyaev O., Vydrin A. The Creation of Large-Scale Annotated Corpora of Minority Languages using UniParser and the EANC platform // Proc. of COLING 2012: Posters, Mumbai, India. 2012. P. 83–92.
2. Nivre J., de Marneffe M., Ginter F., Hajic J., Manning C., Pyysalo S., Schuster S., Tyers F., Zeman D. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection // Proc. of the Twelfth Language Resources and Evaluation Conference, Marseille, France. European Language Resources Association: 2020. P. 4034–4043.
3. Shatskikh A., Sorokin A. Ossetic-COT: Designing a morphologically annotated corpus and morphological analyzer for Ossetic. // Computational

Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2026). 2026. P. 544-555.

4. *Abaev V.* A grammatical sketch of Ossetic. The Hague: Mouton, 1964.
5. *Belyaev O.* Korreljativnaja konstrukcija v osetinskom jazyke v tipologicheskom osveschenii (PhD thesis). 2014.
6. *Belyaev O.* Evolution of Case in Ossetic. // Iran and the Caucasus. 2010. No. 14 (2). P. 287–322.
7. *Serdobolskaya N., Tuzhik O.* Differential Object Marking in Modern Ossetic: Referential Properties and Animacy // Tomsk Journal of Linguistics and Anthropology. 2025. No. 48 (2). P. 91–107.
8. *Kondratyuk D., Straka M.* 75 Languages, 1 Model: Parsing Universal Dependencies Universally // Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th International Joint Conf. on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China. 2019. P. 2779–2795.
9. *Straka M., Straková J., Hajic J.* UDPipe at SIGMORPHON 2019: Contextualized Embeddings, Regularization with Morphological Categories, Corpora Merging // Proc. of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, Florence, Italy. 2019. P. 95–103.
10. *Devlin J., Chang M., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. // Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, Minneapolis, Minnesota. Association for Computational Linguistics: 2020. P. 4171–4186.
11. *Kuraton Y., Arkhipov M.* Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. 2019.
12. *Hulden M.* Foma: a Finite-State Compiler and Library // Proceedings of the Demonstrations Session at EACL 2009, Athens, Greece. 2009. P. 29–32.
13. *Inoue G., Shindo H., Matsumoto Y.* Joint Prediction of Morphosyntactic Categories for Fine-Grained Arabic Part-of-Speech Tagging Exploiting Tag Dictionary Information // Proc. of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada. 2017. P. 421–431.

СВЕДЕНИЯ ОБ АВТОРАХ



ШАТСКИХ Анна Сергеевна – лаборант-исследователь в секторе прикладной лингвистики Института языкознания РАН. Область научных интересов: обработка естественного языка, языковые модели, компьютерная морфология, малоресурсные языки, иранские языки, осетинский язык.

Anna S. SHATSKIKH is a research assistant at the Department of Applied Linguistics, Institute of Linguistics of the Russian Academy of Sciences. Research interests: natural language processing, large language models, computational morphology, low-resource languages, Iranian languages, Ossetic language.

email: anapriselec@gmail.com

ORCID: 0009-0006-9025-7783



СОРОКИН Алексей Андреевич – старший научный сотрудник Центра ИИ МГУ, доцент кафедры математической логики и теории алгоритмов механико-математического факультета МГУ. Область научных интересов: обработка естественного языка, языковые модели, компьютерная морфология, обработка малоресурсных языков, автоматическое исправление грамматических ошибок.

Alexey A. SOROKIN is a senior researcher at the MSU AI Centre. Research interests: natural language processing, large language models, computational morphology, NLP for low-resource languages, grammatical error correction.

email: alexey.sorokin@list.ru

ORCID: 0000-0003-3877-4223

Материал поступил в редакцию 25 августа 2026 года