ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа

«Фундаментальная и компьютерная лингвистика»

Землянская София Александровна

АВТОМАТИЧЕСКАЯ ОБРАБОТКА СОВРЕМЕННОГО ЗАПАДНОГО АРАМЕЙСКОГО ЯЗЫКА

Выпускная квалификационная работа студента 4 курса бакалавриата группы БКЛ202

Академический руководитель образовательной программы
канд. филологических наук, доц.
Ю.А. Ландер

_____

« » _____ 2024 г.

Научный руководитель

Доцент, Школа лингвистики
Толдова Светлана Юрьевна

_____

Научный консультант
Приглашенный преподаватель, Школа лингвистики
Сериков Олег Алексеевич

_____

Москва 2024

ОГЛАВЛЕНИЕ

1. Введение

Современный западный арамейский язык (далее — СЗА) принадлежит к северозападной ветви семитской языковой семьи и является одним из представителей новоарамейских языков (Hammarström et al. 2024). На нём говорят в нескольких деревнях в Сирии в 60 км к северу от Дамаска в районе горного хребта Каламун (Arnold 2006: xv). Сами же новоарамейские языки делятся на восточную и западную ветви. В то время как современных восточных арамейских языков сейчас много и на них говорят несколько тысяч носителей на достаточно обширной территории, СЗА является единственным ныне живущим представителем западной ветви и находится под угрозой исчезновения (Duntsov et al. 2022: 359). СЗА принято делить на три диалекта в зависимости от деревень, в которых на этом идиоме говорят или говорили: это диалекты Маалулы, Бахи и Джуббадина. Также носители СЗА есть и за пределами этих трёх деревней, в частности, в Дамаске и Бейруте. Общее количество носителей оценивается в не более чем 15 000 человек (Arnold 2011: 685). На диалекте Джуббадина говорят около 10 000 человек. Баха была полностью разрушена в ходе гражданской войны в Сирии; все её жители были вынуждены бежать в другие населённые пункты, поэтому оценить количество носителей этого диалекта не представляется возможным (Duntsov et al. 2022: 359). На диалекте Маалулы по разным оценкам сейчас говорят от 350 до 500 человек, и исчезнуть он может в ближайшие десятилетия (Duntsov et al. 2022: 360; Bromirskaya et al. 2023: 1). По этим причинам СЗА представляет интерес, во-первых, для семитологов, так как является ценным источником информации для сравнительно-исторических исследований семитских языков в целом и для изучения таких близких мёртвых языков, как христианско-палестинский, иудейско-палестинский и пр., а во-вторых, для библеистов. Также СЗА очень консервативен в фонетике и морфологии, что делает его очень близким к среднеарамейскому, существовашему в 500–1000 гг., и это ещё одна из причин, по которой он интересен исследователям (Jastrow 1997: 334).

Сейчас существует несколько научных центров, занимающихся СЗА. Один из них находится в НИУ «Высшая школа экономики» — это научно-учебная группа «Грамматика современных арамейских языков» (https://iocs.hse.ru/grammararamaic/), берущая своё начало от научно-учебной группы «Исследование новоарамейских языков» (https://iocs.hse.ru/aramaic/). Её участники проводят исследования, семинары и регулярные экспедиции в Маалулу, в которых было собрано и записано много полевого

материала. Также с 2010 г. существует Московский арамеистический кружок (МАС), участники которого занимаются как западными, так и восточными новоарамейскими языками. За рубежом лингвисты занимаются СЗА в Университете Гейдельберга, Университете Дюссельдорфа, Университете Лейдена, Ратгерском университете и др. Также сейчас регулярно проводятся конференции и воркшопы, посвящённые новоарамейским языкам, например, NALCon (Neo-Aramaic Languages Conference), а на конференциях, посвящённых любым языкам, часто выступают с докладами про СЗА и другие новоарамейские языки (Grishin, Bromirskaya 2023, Kashintseva 2023).

Несмотря на популярность изучения, для СЗА отсутствуют какие бы то ни было NLP-инструменты, поэтому сейчас нет возможности, например, проводить сложные корпусные исследования и быстро анализировать и издавать полевые тексты. В связи с этим мы решили разработать NLP-инструменты, которые помогут облегчить исследования всем лингвистам, занимающимся СЗА. Так как три диалекта значительно различаются между собой, а в центре внимания учёных находится именно диалект Маалулы, было решено создавать инструменты в первую очередь для него. В эти инструменты входят: 1) морфологический парсер, который для каждого слова выдаёт его лемму, часть речи, грамматические признаки; 2) инструмент для глоссирования в виде функции на языке Python, которая принимает на вход полевой текст и выдаёт его глоссированный вариант, который потом можно использовать в статьях; 3) параллельный арамейско-немецкий корпус с морфологической аннотацией для арамейского языка (полученной с помощью созданного нами парсера), в котором возможен поиск не только по словоформам, но и по леммам, частям речи, грамматическим признакам, а также по метаданным. Помимо этого, нами были созданы инструменты для оценки качества работы морфологического парсера и инструмента для глоссирования, а также проведены исследования омонимии и глаголов в форме субъюнктива на материале полученного корпуса. Все созданные инструменты, парсер и информация о корпусе доступны на GitHub по ссылке: https://github.com/Sophia179/aramaic_NLP.

Эта работа имеет следующую структуру: в Разделе 2 представлен литературный обзор материалов и грамматики СЗА, обзор морфологических парсеров и платформ для создания лингвистических корпусов; в Разделе 3 приводится краткое описание грамматики СЗА; в Разделе 4 описаны методы, которыми мы пользовались при создании NLP-инструментов для диалекта Маалулы, и процесс создания этих

инструментов; в Разделе 5 представлены результаты; в Разделе 6 приводятся примеры мини-исследований на материале полученного корпуса; в Разделе 7 обсуждаются дальнейшие направления исследований.

2. Литературный обзор

2.1 Материалы по современному западному арамейскому

Несмотря на активность изучения, на сегодняшний день СЗА достаточно плохо описан. Из материалов есть несколько грамматик, например, (Arnold 1990, Spitaler 1938), учебник с элементами грамматики и текстами (Arnold 2006), единственный словарь (Arnold 2019). Ни одна из этих книг, за исключением словаря, не является распознанной. Многие материалы появились лишь недавно. Значительный вклад в изучение СЗА внёс немецкий лингвист В. Арнольд. Многие монографии, посвящённые этому языку, были написаны им по результатам почти двухлетней экспедиции в деревни Маалула, Баха и Джуббадин в 1985–1987 гг. На тот момент СЗА описывается В. Арнольдом как «необыкновенно живой», несмотря на то что уже тогда было распространено мнение, что это вымирающий язык (Arnold 1990a: xix). В грамматике (Arnold 1990a) описаны фонетика и морфология всех трёх диалектов, но практически не описан синтаксис.

Что касается текстов, то, во-первых, есть сборник из 28 сказок, записанных в Маалуле в 1869 году А. Примом и Е. Социном, который был опубликован Г. Бергштрессером в 1915 году (Bergsträsser 1915). Во-вторых, есть четыре сборника рассказов на СЗА: по одному сборнику из Бахи и Джуббадина (Arnold 1989, 1990b) и ещё два из Маалулы (Arnold 1991a, 1991b). В-третьих, регулярно публикуются новые полевые тексты в отдельных статьях, посвящённых СЗА, например (Bromirskaya et al. 2023; Duntsov et al. 2022).

Из электронных ресурсов на данный момент есть только параллельный корпус текстов из всех трёх деревень с переводом на немецкий язык. В корпус входят тексты из (Arnold 1989, 1900b, 1991a, 1991b, 2002, 2006). Объём корпуса составляет приблизительно 110 000 токенов, из них около 70 000 приходятся на тексты из Маалулы (Barsky). В настоящее время этот корпус не аннотирован: поиск в нём возможен только по точным словоформам или с помощью регулярных выражений, что существенно сужает круг исследований, которые можно на нём провести. В частности, сейчас на нём затруднительно проводить исследования в области синтаксиса, а это —

одно самых приоритетных направлений исследований, поскольку синтаксис СЗА практически не описан: единственный материал по этой области был издан более сорока лет назад (Correll 1978).

Помимо письменных материалов существует некоторое количество аудиозаписей речи на диалекте Маалулы. Так, во время экспедиции В. Арнольд записал на магнитофонную плёнку множество полевых текстов на темы, охватывающие почти все сферы деревенской жизни: среди них рассказы о традициях, сказки, анекдоты и пр. Затранскрибированные тексты к этим аудиозаписям и вышли потом в виде двух сборников (Arnold 1991a, 1991b), которые потом вошли в имеющийся корпус. Аудиозаписи к этим текстам находятся в свободном доступе. Также доступны отдельные аудиозаписи полевых текстов, сделанные лингвистами в ходе более недавних экспедиций (например, https://iocs.hse.ru/en/lullaby).

Несмотря на то, что носителей диалекта Маалулы в разы меньше, чем диалектов Джуббадина и Бахи, именно он стал центром внимания лингвистов и на данный момент является наиболее изученным (Коган, Лёзов 2009: 705). Отчасти это объясняется тем, что Маалула является христианской деревней, в то время как Баха и Джуббадин — мусульманские (Fassberg 2019: 633; Коган, Лёзов 2009: 705, Jastrow 1997: 334). По этой причине организовывать экспедиции было проще именно в неё. Сейчас регулярно выходят новые статьи и монографии, посвящённые диалекту Маалулы, например, (Eid 2024; Bromirskaya et al. 2023, Duntsov et al. 2022). В 2022 году был опубликован датасет *The Maalula Aramaic Speech Corpus* (далее — MASC), составленный лингвистами из Университета Дюссельдорфа (Ghattas et al. 2022). Он содержит тексты из (Arnold 1991a, 1991b), лемматизированные варианты этих текстов, список и количество их лемм и встретившихся уникальных словоформ для каждой леммы. Лемматизация текстов происходила вручную и сверялась с информантом. Также в датасете есть аудиозаписи всех текстов и метаданные к каждой из них. В метаданных содержится информация об имени, поле, возрасте (на 1986 год) и профессии говорящего, а также о теме рассказа.

2.2 Морфологические парсеры

Морфологический парсер — инструмент, который анализирует слова и присваивает им леммы, части речи и грамматические характеристики. Существует несколько способов их создания. Глобально их можно разделить на

правило-ориентированные (rule-based parsers) и подходы с использованием машинного обучения. Парсеры, созданные в рамках первого способа, часто представляют собой в сущности сложную систему регулярных выражений. Работа таких парсеров обычно базируется на двух или трёх файлах: в одном из них содержится информация о лексемах, в другом — о парадигмах, по которым должно происходить словоизменение, в третьем (который присутствует не во всех фреймворках) — информация о морфофонологических чередованиях. Часто, но не всегда, на последнем этапе происходит конвертация данных в конечный автомат, что ускоряет работу анализатора.

Правило-ориентированные парсеры особенно подходят для малоресурсных и малоописанных языков. Примерами таких фреймворков и парсеров могут служить UniParser (Arkhangelskiy et al. 2012), lexd & twol (Swanson and Hollow 2021), HFST — Helsinki Finite State Technology (Lindén et al. 2009), SFST — Stuttgart Finite State Transducer (Schmid 2005), pymorphy2 (Korobov 2015) и др. Однако некоторые из таких инструментов подходят для создания парсеров только для некоторых языков, а не для любых. В частности, многие из фреймворков не поддерживают несегментную морфологию, такую как инфиксы, трансфиксы и редупликация, поскольку в основе их работы у лексем есть неизменяемая основа, к которой могут присоединяться префиксы и суффиксы, а изменения самой основы (кроме морфофонологических чередований) не допускаются.

Другой подход заключается в использовании методов машинного обучения. Обычно в этом случае требуется большое количество аннотированных или неаннотированных данных для обучения, что невозможно для малоресурсных языков. Более того, при обучении на малом объёме данных существует риск переобучения (Sorokin 2019: 2). Поэтому в случае малоресурсных языков прибегают к таким методам, как transfer-learning, annotation projection и др. (Hwa et al. 2002). Однако несмотря на трудности в реализации иногда всё же удаётся получить хорошие результаты для таких языков. Например, в соревновании LowResourceEval-2019 три команды использовали RNN, CNN и модели Маркова для создания морфологических анализаторов для эвенкийского, селькупского и других малых языков, и результаты были весьма успешны (Klyachko et al. 2020).

2.3 Корпусные платформы

После применения морфологических парсеров к большому количеству текстов удобно организовывать их как аннотированный корпус, который можно сделать доступным онлайн для всех. Существует множество платформ для создания таких лингвистических корпусов, среди них SketchEngine и NoSketchEngine (Rychlý 2007, Kilgarriff et al. 2014), tsakorpus (Arkhangelskiy 2017), ANNIS (Krause and Zeldes 2016), Corpus Workbench (CW) и др. Некоторые из них поддерживают не только лингвистическую аннотацию и поиск по всему корпусу, но и возможность прикрепления мультимедийных файлов и поиск по подкорпусам.

3. Некоторые сведения о грамматике СЗА

СЗА является бесписьменным языком. В начале XXI в. предпринималась попытка разработать для него алфавит на основе древнеарамейского, однако он не прижился (Fassberg 2019: 633). В научных материалах лингвисты используют специальную орфографию на основе латиницы с диакритиками, разработанную В. Арнольдом, или символы международного фонетического алфавита.

Все носители СЗА владеют дамасским вариантом арабского языка и литературным арабским, который имеет наибольший престиж (Arnold 1990a: i, Arnold 2011: 685, Bromirskaya et al. 2023: 1). На СЗА общаются в основном в кругу семьи. Большинство жителей Маалулы большую часть времени проводят в Дамаске, приезжая домой только на лето. В связи с этим в СЗА проникло множество завимствований из арабского, часть из которых арамеизировалась, а часть — нет.

Диалекты Маалулы, Бахи и Джуббадина значительно различаются между собой на всех уровнях: фонетическом, морфологическом, синтаксическом и лексическом. Например:

Таблица 1. Сравнение диалектов Маалулы, Бахи и Джуббадина.

черта	**Маалула**	**Баха**	**Джуббадин**
Префикс 2 л. ед. ч. ж. р. у глаголов настоящего времени	*č-*	*š-*	*š-*
Вокализм в начальных формах глаголов на примере глагола ‘спать’	*id̲mex, yid̲mux*	*id̲mex, yud̲mux*	*id̲max, yud̲mux*

Императив ед. ч. м. р. на примере глагола ‘жить’	*ḥā*	*iḥḥa*	*iḥḥō*
Различение рода в формах мн. ч. глаголов	есть	нет	есть
Степень палатализации звука /k/	сильная (k’’)	слабая (k’)	очень сильная (č)
‘кто’	*mōn*	*man*	*mūn*
‘я спасался бегством’	*šamṭiṯ*	*šimṭiṯ*	*nhazmiṯ*
‘(он) говорит им’	*amerlun*	*amellun*	*ameləl*
‘червяк’	*ṯawlaʕča*	*dūda*	*ṯawlaʕča*
‘шакал’	*naččawīṯa*	*wawīṯa*	*bawwōyṯa*
Различия в значении объектных аффиксов на примере глагола ‘дать’	*mayṯīl* ‘(он) даёт мне’	?	*mayṯīl* ‘(они) дают им’
Различия в значении некоторых лексем	*afaš* ‘ну/так/теперь’	*afaš* ‘(он) остался’	*afaš* ‘(он) плыл’
Выражение притяжательности	*tīḏ(i)* ‘мой’	*ci līl* ‘мой’	*tīḏay* ‘мой’

Фонемный инвертарь диалекта Маалулы представлен в Таблице 2 (Duntsov et al. 2022: 363–364). Согласные бывают глухими и звонкими, эмфатическими и неэмфатическими, гласные — долгими и краткими (на письме долгота обозначается чертой над гласным: e ~ ē). Также на письме отображается звук шва, который на самом деле не является фонемой: он появляется только для удобства произношения, когда возникает кластер из трёх и более согласных. В таблице в квадратных скобках даны фонемы, которые встречаются только в неассимилированных заимствованиях. В круглых скобках показано, как конкретный звук записывается в латинской транскрипции В. Арнольда, если буква алфавита не совпадает с символом МФА.

Таблица 2. Согласные диалекта Маалулы.

Взрывные и аффрикаты

	губно-губные	губно-зубные	межзубные	альвеолярные	постальвеолярные	палатальные	велярные	увулярные	фарингальные	глоттальные
глухие	*p*			*t*	*ʧ* (č)	*c* (k)	*k* (ḳ)			*ʔ*
звонкие	*b*			[*d*]			[*g*]			
эмфатические				*tˤ* (ṭ)						

Фрикативные

	губно-губные	губно-зубные	межзубные	альвеолярные	постальвеолярные	палатальные	велярные	увулярные	фарингальные	глоттальные
глухие		*f*	*θ* (t̲)	*s*	*ʃ* (š)			*χ* (x)	*ħ* (ḥ)	*h*
звонкие			*ð* (d̲)	*z*	*ʒ* (ž)			*ʁ* (ġ)	*ʕ*	
эмфатические			*ðˤ* (ḏ)	*sˤ* (ṣ), [*zˤ*] (ẓ)						

Сонорные

	губно-губные	губно-зубные	межзубные	альвеолярные	постальвеолярные	палатальные	велярные	увулярные	фарингальные	глоттальные
носовые	*m*			*n*						
латеральные				*l*						
апикальные				*r*						
аппроксиманты	*w*					*j* (y)				

Таблица 3. Гласные диалекта Маалулы.

	передний ряд	**средний ряд**	**задний ряд**
верхний подъём	/i/, /iː/	/u/, /uː/	
средний подъём	/ɛ/, /ɛː/	(ə)	/o/, /oː/
нижний подъём	/a/, /aː/		

дифтонги: /au/, /aɪ/

В СЗА происходит очень много ассимиляций. Так, *b* оглушается до *p* в позициях перед глухими согласными и на конце слова (например, *ʕinbō* ‘виноград.Pl’ — *ʕenəpt̲a* ‘виноград.Sg’). Согласный *n* может ассимилироваться к *č* или *l*. Например, есть вариация в произношении *inḥeč/iḥḥeč* ‘спускаться’, а *ešna* ‘год’ при присоединении суффикса *-l* превращается в *ešl*. Также суффикс *-l* почти всегда ассимилируется к так называемым «солнечным согласным», а местоимение *hōd̲* ‘эта’ — практически ко всему, что стоит рядом. Например, *ḥessil reḳḳa* → *ḥessi**r**‿reḳḳa* ‘голос рябка’, *ʕemmil samkōt̲a* → *ʕemmi**s**‿samkōt̲a* ‘с рыбами’, *ṭaḥniččil ṭaḥənt̲a* → *ṭaḥničči**ṭ**‿ṭaḥənt̲a* ‘молол зерно’, *ḥmiččil hōd̲ ktīšča* → *ḥmiččil‿**l**ō**k**‿ktīšča* ‘(я) увидел эту лошадь’.

Морфология СЗА достаточно сложна. В языке активно используются трансфиксы при склонении существительных, прилагательных, при спряжении глаголов, в словообразовательных моделях и пр. Например, глагол *id̲mex* ‘спать’ при спряжении может иметь следующие формы: *d̲imx-at̲, yi-d̲mux, č-d̲umx-un, d̲mōx, d̲mūx, d̲ōmex, d̲ōmx-in, d̲mīx-a.* А прилагательное *izʕur* ‘маленький’ следующие: *zʕōr, zʕūrin, zuʕrōt̲a.*

Существительные бывают мужского и женского рода. Чаще всего первые оканчиваются на *-a*, вторые — на *-t̲a* или *-ča*. В обоих случаях *-a* будет показателем свободной формы, т. е. формы, когда существительное не является вершиной в посессивных конструкциях. В противном случае форма называется связанной и маркируется либо суффиксом *-l*, либо посессивным аффиксом. Например:

(1) *t̲arʕ-il* *forn-a*
дверь-HD духовка-FREE
‘дверь духовки’ (‘дверь’ — *t̲arʕa*)

(2) *waʕ-yō-t̲-iš*
платье-P-F-POSS.2fs
‘твои платья’ (‘платья’ — *waʕyōt̲a*)

Помимо существительных, посессивные аффиксы могут присоединяться к предлогам. Например, *ġapp-e* ‘с ним’, *ġapp-ax* ‘с тобой (m.)’ и т. д.

Глаголы делятся на одиннадцать пород: I, II, III, IV, I_2, II_2, III_2, IV_2, I_7, I_8, I_{10}. Также есть слабые глаголы, которые делятся на типы I?, Iw, Iy, IIw, IIy, IIIy и др. Иногда к слабым глаголам также причисляют некоторые глаголы типа In и глаголы с одинаковым вторым и третьим согласным. Породы III, III_2, I_7, I_8, I_{10} заимствованы из арабского, I_2 и

IV_2 унаследованы из арамейского, остальные представляют собой смешение заимствованных и исконных глаголов. Корни глаголов чаще всего трёхсогласные. Глаголы с четырёхсогласными корнями относятся к породам I и I_2. Также есть глаголы со слабыми согласными *w* и *y*, которые проявляются только в некоторых формах, например: *aḥək* ‘он говорил (PST)’ — *maḥəkyin* ‘они (m.) говорят’ (корень *ḥky*), *intar* ‘он ходил покругу (PST)’ — *tayyer* ‘он ходил покругу (PRF)’ (корень *tyr*).

Глаголы изменяются по таким грамматическим категориям Tense-Aspect-Mood, как презенс, претерит, перфект, субъюнктив и императив. Формы претерита относятся к префиксальному спряжению, субъюнктива, презенса и перфекта — к суффиксальному, а императива — ни к тому, ни к другому. Презенс и перфект развились из старых причастий, поэтому в этих формах, в отличие от претерита и субъюнктива, будет различие рода в 1 лице, а объектные суффиксы, выражающие прямые и непрямые объекты, будут совпадать (Arnold 1990a: 54). Будущее время выражается аналитически с помощью сочетания вспомогательного глагола *batte* ‘хотеть’ и смыслового глагола в форме субъюнктива. Например, *batte yiḏmux* — ‘он будет спать’ (Arnold 1990a: 193).

Помимо лично-числовых показателей к глаголам могут присоединяться объектные суффиксы. Они бывают трёх видов: суффиксы прямых объектов, непрямых объектов и двойные суффиксы, которые несут в себе одновременно значение прямого и непрямого объекта. В некоторых формах часть суффикса двойного объекта оказывается в неожиданной позиции между основой и лично-числовым показателем глагола, как, например, в примере (6):

(3) *faṯḥ-aṯ*
открыть.PST-3fs
‘она открыла’

(4) *faṯḥ-ač-č-e*
открыть.PST-3fs-PLEO-DO.3ms
‘она открыла его’

(5) *faṯḥ-al-le*
открыть.PST-3fs-IO.3ms
‘она открыла ему’

(6) *faṯəḥ-l-al-le*
открыть.PST-DO-3fs-IO.3ms
‘она открыла его ему’

В презенсе и перфекте суффиксы прямых и непрямых объектов совпадают, вычислить их точное значение можно только из контекста.

Синтаксис СЗА практически не описан. Порядок слов в предложении считается относительно свободным (Коган, Лёзов 2009: 745).

В языке обширно представлена омонимия. Чаще всего она проявляется при спряжении глагола. Так, например, всегда будут омонимичны формы 3ms и 3cp в претерите: *iftaḥ* — ‘он открыл’/‘они открыли’. В субъюнктиве совпадает сразу несколько форм: *čišmaʕ* — это одновременно и ‘(чтобы) она услышала’, и ‘(чтобы) ты услышала’, и ‘(чтобы) ты услышал’, а *nišmaʕ* — ‘(чтобы) я услышал(а)’ и ‘(чтобы) мы услышал(и)’.

4. Методы

4.1 Создание морфологического парсера

Для создания морфологического парсера мы решили выбрать UniParser. Этот инструмент был спроектирован так, чтобы с его помощью можно было создать парсер для языка с любой морфологией. В частности, в нём возможна реализация трансфиксов, которые являются ключевой чертой морфологии любого семитского языка. Это была одна из причин, по которой мы решили выбрать именно этот фреймворк. Другая причина — с помощью UniParser уже были созданы парсеры для языков мира с самой разной морфологией, например, для албанского (индоевропейская семья) (Arkhangelskiy and Morozova 2019), адыгейского (абхазо-адыгская семья) (Lander, Arkhangelskiy 2018), бурятского (монгольская семья) (Arkhangelskiy 2021), мокшанского (уральская семья) (Arkhangelskiy 2018), языка яварана (карибская семья) (Matter 2022) и др. Что для нас более важно — с помощью UniParser также были сделаны морфологические анализаторы для некоторых современных восточных арамейских языков, а именно для туройо (Arkhangelskiy et al. 2018) и христианского урмийского (Arkhangelskiy 2019), которые в своей морфологии схожи с СЗА. Третий аргумент в пользу UniParser — это удобный пользовательский интерфейс и понятная документация.

Но несмотря на большое количество положительных сторон, у UniParser есть и минусы. Во-первых, на больших объёмах данных он может работать медленно, так как на финальном этапе создания парсера не происходит конвертации в трансдьюсер. Однако для нашего случая это неважно, так как СЗА — малоресурсный язык и у нас небольшой объём данных. Во-вторых, в UniParser, как и во многих других правило-ориентированных морфоанализаторах, не анализируются out-of-vocabulary words, то есть слова, которые не содержатся в файле с лексемами. Это представляет собой некоторую проблему, так как в полевых текстах из Маалулы часто попадаются слова, которых нет в словаре (Arnold 2019). Для решения этой задачи можно дополнительно реализовать гессер, однако это уже не входит в рамки этой работы. В-третьих, UniParser чувствителен к орфографии. Если записать слово не так, как оно представлено в словаре, оно не будет распознано — это частный случай пункта выше. Поэтому в нашем случае важно придерживаться традиционной орфографии, предложенной Арнольдом.

В качестве данных мы использовали словарь (Arnold 2019), грамматику (Arnold 1990a), поскольку в ней представлены наиболее полно словоизменительные парадигмы, и датасет MASC (Ghattas et al. 2022).

UniParser требует наличия двух файлов — lexemes.txt и paradigm.txt. В первом перечислены лексемы: лемма, основа (или варианты основ, если их несколько), часть речи (и неизменяемые грамматические признаки, если есть), название парадигмы из файла с парадигмами, по которой эта лексема словоизменяется, а также перевод лексемы и другие языки. В файле с парадигмами содержатся парадигмы, в каждой из которых есть флексия, её граммемы и то, как они должны выглядеть при глоссировании. Также возможны дополнительные файлы, например, clitics.txt, в которых содержится информация о клитиках, но эти файлы не являются обязательными и мы их не использовали.

Файлы с лексемами и парадигмами заполнялись следующим образом: сначала в грамматике (Arnold 1990a) ищется парадигма и читается описание к ней, если оно есть. Обычно одна парадигма соответствует словам, подходящим под какой-нибудь паттерн, например, KiKKa (где K — согласный), или слова, оканчивающиеся на определённый суффикс (например, на *-ōna*). На основании этой информации пишется регулярное выражение, и затем из таблицы лемм из датасета MASC извлекаются все слова,

подходящие под паттерн. Далее при необходимости может происходить проверка слов по словарю, поскольку иногда под паттерны могли попасть слова, принадлежащие к другим частям речи или изменяющиеся по другим парадигмам.

Например, одни существительные, подходящие под паттерн KōKa образуют множественное число по модели KaKō, другие — по KuKō. На синхронном уровне определить, какой из двух вариантов будет, нельзя — эти различия обусловлены диахронически происхождением гласного. В таких случаях проходила итерация по распознанному словарю, где выводились строки, в которых есть одновременно исходная словоформа и один из двух вариантов множественно числа. Далее эти строки отсматривались и принималось решение, к какой из парадигм каждую лексему отнести. Также после получения списка служебных слов по нему проводились итерации, чтобы убрать случайно попавшиеся служебные слова из парадигм для других частей речи.

Как уже было сказано, одна из самых ярких и сложных в реализации черт семитских языков — трансфиксы. СЗА в этом плане не исключение. В UniParser можно задавать это явление несколькими способами. Первый — перечислить все возможные варианты основ и затем указать в парадигме, какая при какой флексии должна использоваться. Второй — выделить основу без трансфикса и затем в парадигме расставить гласные нужным образом. Например, слово *xefa* ‘камень’, множественное число у которого выглядит как *xifō*, можно задать такими двумя способами:

Листинг 1. Пример лексемы в UniParser.

```
-lexeme: xefa
 -stem: xef.|xif.
 -gramm: NOUN,m
 -paradigm: NOUN_KeKa1
```

```
-lexeme: xefa
 -stem: x.f.
 -gramm: NOUN,m
 -paradigm: NOUN_KeKa2
```

В первом случае через символ | перечисляются алломорфы основы, во втором — через точки указываются места «склейки» основы с флексией. Парадигмы для каждого из случаев будут выглядеть так:

Листинг 2. Примеры парадигм в UniParser.

```
-paradigm: NOUN_KeKa1
 -flex: <0>.a
  gramm: Sg
 -flex: <1>.ō
```

```
-paradigm: NOUN_KeKa2
 -flex: .e.a
  gramm: Sg
 -flex: .i.ō
```

```
  gramm: Pl                        gramm: Pl
```

Первый способ удобен, если в определённой группе слов есть несколько разных типов вокализма. Например, среди двухсогласных существительных помимо модели KeKa ~ KiKō единственное и множественное число могут выглядеть как KoKa ~ KaKō, KoKa ~ KuKō, KēKa ~ KiKō и пр. В таких случаях удобно перечислить варианты основ для единственного и множественного числа через | и составить для всех них одну парадигму, так как окончания во всех случаях будут одинаковы (*-a* и *-ō*). Это экономнее, чем для каждого варианта составлять новую парадигму с новым типом вокализма.

Второй способ удобен тогда, когда приходится перечислять слишком много основ. В случае выше у нас было всего две основы, однако, например, при спряжении глаголов количество таких основ у одного глагола может доходить до 12. Чтобы не перечислять их все, удобным оказывается как раз использование второго способа с точками.

Иногда оба способа могут комбинироваться. Например, когда у глагола помимо обычной основы есть основа с удвоенным согласным. Тогда лексема и её парадигма могут выглядеть так:

Листинг 3. Пример глагола IV породы и фрагмент его парадигмы.

```
-lexeme
 lex: aḥref yaḥref
 stem: .ḥ.r.f.|.ḥ.rr.f.
 gramm: VERB,IV

-paradigm: VERB_IV_ae
 -flex: <0>[a]..[e].
  gramm: Praet,3,Sg,m
 -flex: <0>[a]..[e].
  gramm: Praet,3,Pl,c
 -flex: <0>[a]...|aṯ//<0>[a].[ə]..|aṯ
  gramm: Praet,3,Sg,f
  gloss: 3fs
 -flex: <0>[a]...|ič//<0>[a].[ə]..|ič
  gramm: Praet,2,Sg,m
  gloss: 2ms
 -flex: <0>[a]...|iš//<0>[a].[ə]..|iš
```

```
 gramm: Praet,2,Sg,f
 gloss: 2fs
-flex: <0>[a]...|ičxun//<0>[a].[ə]..|ičxun
 gramm: Praet,2,Pl,m
 gloss: 2mp
-flex: <0>[a]...|ičxen//<0>[a].[ə]..|ičxen
 gramm: Praet,2,Pl,f
 gloss: 2fp
-flex: <0>[a]...|iṭ//<0>[a].[ə]..|iṭ
 gramm: Praet,1,Sg,c
 gloss: 1S
-flex: <0>[a]..[i].innaḥ
 gramm: Praet,1,Pl,c
 gloss: 1P

-flex: <1>n|.[a].[e].
 gramm: Perf,1,Sg,m
 gloss: 1
-flex: <1>č|.[a].[e].
 gramm: Perf,2,Sg,m
 gloss: 2
-flex: <1>.[a].[e].
 gramm: Perf,3,Sg,m
-flex: <1>n|.[a].[ī].|a
 gramm: Perf,1,Sg,f
 gloss: 1|F
-flex: <1>č|.[a].[ī].|a
 gramm: Perf,2,Sg,f
 gloss: 2|F
-flex: <1>.[a].[ī].|a
 gramm: Perf,3,Sg,f
 gloss: F
```

К глаголам и существительным могут присоединяться объектные и посессивные суффиксы соответственно. Для таких случаев были созданы отдельно парадигмы с этими аффиксами, к которым могли отсылать отдельные формы из парадигм глаголов и сущесвительных. В случае с глаголами аффиксы могли присоединяться не напрямую, а через наращение на основе, поэтому итоговая парадигма получалась вдвойне вложенной. Например:

Листинг 4. Фрагмент парадигмы глагола с прямообъектными суффиксами.

```
-paradigm: VERB_I_au
 -flex: <0>.[a]..<.>//<0>.[a].[ə].<.>
  gramm: Praet
  paradigm: VERB_O1

-paradigm: VERB_O1
 -flex: .<.>
  gramm: 3,Sg,m
  paradigm: VERB_O2
 -flex: .n<.>
  gramm: 3,Sg,m
  gloss: PLEO
  paradigm: VERB_O2
 -flex: .an<.>
  gramm: 3,Sg,m
  gloss: 3ms
  paradigm: VERB_O22
 -flex: .an|n<.>
  gramm: 3,Sg,m
  gloss: 3ms|PLEO
  paradigm: VERB_O22
 -flex: .ač<.>
  gramm: 3,Sg,f
  gloss: 3fs
  paradigm: VERB_O2
 -flex: .ač|č<.>
  gramm: 3,Sg,f
  gloss: 3fs|PLEO
  paradigm: VERB_O2
 -flex: .un<.>
  gramm: 3,Pl,m
  gloss: 3mp
  paradigm: VERB_O2
 -flex: .un|n<.>
  gramm: 3,Pl,m
  gloss: 3mp|PLEO
  paradigm: VERB_O2
 -flex: .an<.>
  gramm: 3,Pl,f
```

```
  gloss: 3fp
  paradigm: VERB_O2
 -flex: .an|n<.>
  gramm: 3,Pl,f
  gloss: 3fp|PLEO
  paradigm: VERB_O2
 -flex: .īč|n<.>//.īčə|n<.>
  gramm: 2,Sg,m
  gloss: 2ms|PLEO
  paradigm: VERB_O2
 -flex: .īš|n<.>//.īšə|n<.>
  gramm: 2,Sg,f
  gloss: 2fs|PLEO
  paradigm: VERB_O2
 -flex: .čun|n<.>
  gramm: 2,Pl,m
  gloss: 2mp|PLEO
  paradigm: VERB_O2
 -flex: .čan|n<.>
  gramm: 2,Pl,f
  gloss: 2fp|PLEO
  paradigm: VERB_O2
 -flex: .ič<.>
  gramm: 1,Sg,c
  gloss: 1S
  paradigm: VERB_O2
 -flex: .ič|n<.>//.ič|č<.>
  gramm: 1,Sg,c
  gloss: 1S|PLEO
  paradigm: VERB_O2
 -flex: .laḥl<.>//.naḥl<.>//.laḥəl<.>//.naḥəl<.>
  gramm: 1,Pl,c
  gloss: 1P
  paradigm: VERB_O2

-paradigm: VERB_O2
 -flex: .l//.il//.əl//.lə
  gramm: DOM
  std: .l
  gloss: =DOM
 -flex: .e//.u
```

```
  gramm: O.3ms
  gloss: =3ms
 -flex: .a
  gramm: O.3fs
  gloss: =3fs
 -flex: .un//.
  gramm: O.3mp
  gloss: =3mp
 -flex: .en
  gramm: O.3fp
  gloss: =3fp
 -flex: .ax
  gramm: O.2ms
  gloss: =2ms
 -flex: .iš
  gramm: O.2fs
  gloss: =2fs
 -flex: .xun
  gramm: O.2mp
  gloss: =2mp
 -flex: .xen
  gramm: O.2fp
  gloss: =2fp
 -flex: .i//.//.in
  gramm: O.1cs
  gloss: =1S
 -flex: .aḥ
  gramm: O.1cp
  gloss: =1P

-paradigm: VERB_O22
 -flex: .un//.
  gramm: O.3mp
  gloss: =3mp
 -flex: .en
  gramm: O.3fp
  gloss: =3fp
 -flex: .xun
  gramm: O.2mp
  gloss: =2mp
 -flex: .xen
```

```
  gramm: O.2fp
  gloss: =2fp
-flex: .aḥ
  gramm: O.1cp
  gloss: =1P
```

Такая парадигма позволяет получать такие формы, как *faṯḥ-ač-č-e* (см. пример (3)) и др.

Также в процессе создания парсера некоторые парадигмы могли сливаться, если это не приводило к неверным разборам. Например, у некоторых собирательных существительных дефектная парадигма — у них нет множественного числа. Чтобы не делать для них отдельную парадигму, в целях экономии им приписывалась обычная парадигма, которая уже есть у каких-либо других существительных. Так как собирательное существительное никогда не встретится во множественном числе, можно не опасаться таких разборов, поскольку их не будет. Другой случай, когда могло происходить слияние парадигм — это расхождение лишь в одной из форм. Например, существительные женского рода в единственном числе оканчиваются на суффикс *-t̲a* или *-ča*. Причём такого, чтобы одна и та же основа могла присоединять оба из этих суффиксов и чтобы при этом это были разные лексемы, не бывает. Поэтому можно объединить парадигмы для таких двух типов существительных, написав `-flex: .t̲a//.ča`, поскольку во всех остальных формах они имеют одинаковые окончания.

Парадигмы получали название вида «часть речи + паттерн (+ род)» в случае существительных и прилагательных (например, NOUN_KaKKa_m) и «часть речи + порода (+ гласные из паттерна)» в случае глаголов (например, VERB_I_aa). Иногда после части речи приводилось какое-нибудь отдельное существительное или глагол или пр. (например, NOUN_bhimca). Это случаи, когда для слова нужна была отдельная парадигма, которая не вписывалась в другие. При этом не все слова, для которых указана парадигма с паттерном в названии, относятся к этому паттерну: названия парадигм следует понимать не как «слова, относящиеся к этому паттерну», а как «слова, которые изменяются как слова, подходящие под этот паттерн».

Для обозначения частей речи использовались стандартные тэги Universal Dependencies. Примеры слов на каждый из тэгов:

- NOUN: *ṣafrōna* ‘птица’, *ḳeṣṣ̲a* ‘история’, *ffō* ‘лицо’.

- VERB: *išḳal yišḳul* ‘брать’, *zappen yzappen* ‘продавать’, *inəfṯaḥ yinəfṯaḥ* ‘быть открытым’, *amar yīmar* ‘говорить’.
- AUX: *ōb* ‘быть, существовать’, *īle*/*ūle* ‘иметь возможность’, *batte* ‘хотеть’ (вспомогательный глагол при образовании форм будущего времени)
- ADJ: *izʕur* ‘маленький’, *malleḥ* ‘солёный’, *šōṭar* ‘умелый’
- ADV: *elʕel* ‘вверху’, *imōḏ* ‘сегодня’, *baḥar* ‘много’
- PRON: *ana* ‘я’, *mōn* ‘кто’, *anu* ‘чей’, *ḥrīṯa* ‘другая’, *flanō* ‘такой-то’
- NUM: *iṯər* ‘два’, *ṯarč* ‘две’, *ṯēn* ‘второй/вторая’
- DET: *hanna* ‘этот’, *hōṯe* ‘тот’
- PREP: *b* ‘в’, *ʕemmil* ‘с’
- CCONJ: *w* ‘и’, *aw* ‘или’, *amma* ‘но’
- SCONJ: *ḏukkil* ‘когда’, *bōṯar* ‘после’, *ḥetta* ‘чтобы’
- PART: *ya* (вокативная частица), *la* ‘не’, *lorkaʕ* ‘больше не’
- INTJ: *uf* (звукоподражание), *ahūha* (возглас радости), *mpala* ‘напротив!/определённо!’, *čfaḏḏāl* ‘пожалуйста’
- PROPN: *Ḥanān* (имя), *Žaržūra* (имя), *Alṭun* (имя), *Berkṯa* (имя), *Ḏemseḳ* ‘Дамаск’.

К тэгу AUX отсносятся вспомогательные глаголы и так называемые «псевдоглаголы» — слова, которые имеют некоторые глагольные признаки, но которые из-за своей дефектной парадигмы не могут быть причисленны к полноценным глаголам (Arnold 1990a: 184).

Грамматические признаки, которые приписывал парсер, для каждой из частей речи были следующими:

- существительные: род (m — мужской, f — женский), число (Sg — единственное, Pl — множественное, ZPl — счётная форма), посессивные суффиксы (POSS.3ms, POSS.2fp и пр.), связанная форма (HD — head dependent);
- глаголы: TAM (Praet — претерит, Subj — субъюнктив, Imp — императив, Prs — презенс, Perf — перфект), число (Sg, Pl), род (m, f, c — общий), лицо (1, 2, 3), объектные суффиксы (O — прямой объект, D — непрямой (дативный) объект, Dopp — двойной объект);
- прилагательные: род (m, f), число (Sg, Pl), определённость (Def — определённая форма, Indef — неопределённая), лицо (1, 2, 3), элатив (Elat);
- числительные: тип (Card — количественные, Ord — порядковые), род (m, f);

- местоимения: род (m, f), число (Sg, Pl), тип (Pers — личные, Inter —вопросительные , Rel — относительные, Poss — посессивные, Tot — общие, Ind — неопределённые, Refl — рефлексивные);
- предлоги: посессивные суффиксы (POSS.3ms, POSS.2fp и др.) и свободная форма (Free);
- наречия: времени (Temp), места (Loc), модальные (Mod) и другие (other);
- вспомогательные глаголы и псевдоглаголы: число (Sg, Pl), род (m, f), лицо (1, 2, 3);
- имена собственные: имена людей (PER), место (LOC), число (Sg), род (m, f).

4.1.1 Некоторые фонетические замечания

Орфография В. Арнольда фонетическая, то есть отображает произношение. Как было сказано в Разделе 3, звук шва в СЗА не имеет статуса фонемы, а возникает лишь для удобства произношения в ситуациях, когда появляется кластер из трёх и более согласных. Поэтому на письме он часто пишется непоследовательно. Одно и то же слово может встретиться в корпусе как с «ə», так и без. Например, *maḥkyin* и *maḥəkyin*. Поэтому в случаях, когда возможно появление шва, для соответствующих основ или морфем указывались оба варианта через // (см. пример в Листинге 3). То же касается и случаев с ассимиляциями, описанными в Разделе 3.

4.1.2 Некоторые морфологические замечания

Тэги PER и LOC, а также род и число в настоящее время указаны не для всех имён собственных.

Так как в словаре не содержится информации о роде существительных, то женский род присваивался всем словам, у которых было окончание женского рода *-ṯa/-ča*, а мужской — у которых было просто окончание *-a*. Однако не всегда такое разделение верно, например, *ʕayna* ‘глаз’ и *ḥaḳla* ‘поле’ относятся к женскому роду, а *payṯa* ‘дом’ — к мужскому. Но так как в словаре в большинстве случаев не содержится информации о роде, проверить всё было невозможно, и поэтому род присваивался исключительно на основании окончания и в дальнейшем при проверке точности работы парсера не учитывался.

Как было сказано в Разделе 3, посессивность может выражаться специальными посессивными аффиксами, которые присоединяются прямо к существительным.

Например, *xalpa* ‘собака’ — *xalp-iš* ‘твоя (f) собака’. При этом грамматический показатель 3 л. ед. ч. ж. р. совпадает с таковым для свободной формы. Например:

(7) *xalp-a*
собака-FREE
‘собака’

(8) *xalp-a*
собака-POSS.3fs
‘её собака’

Поэтому этот аффикс не указывался в парадигмах, так как иначе все свободные формы существительных ед. ч. получили бы омоним в виде формы с посессивным суффиксом 3fs, а установить его точное наличие возможно лишь в редких контекстах.

Что касается глагола, то в словарях он традиционно задаётся двумя формами — 3 л. ед. ч. м. р. претерита и субъюнктива. По одной из этих форм невозможно однозначно определить вторую. Например, глагол, у которого форма 3 л. ед. ч. м. р. претерита выглядит как *iḳlab,* может в субъюнктиве выглядеть как *yiḳlub* (и тогда это ‘поворачивать’) или *yiḳlab* (и тогда это ‘опрокидываться’). Поэтому в качестве леммы для глаголов указывались обе эти формы через пробел.

В настоящее время не все породы поддерживаются. На данный момент есть только I, II, III, IV, I_7, I_8, I_{10}, II_2, III_2 породы и подтипы Iy, Iw, In, Iʔ, IIy, IIIy, IVy. Остальных пород нет. В частности, не поддерживаются редкие породы, например, порода IV_2 встречается очень редко и уже не известна молодому поколению носителей (Arnold 1990a: 91).

Также на данный момент не для всех пород поддерживается распознавание всех объектных суффиксов. Таблица, в которой показано, что поддерживается для первых четырёх, представлена ниже. Помимо этих пород, объектные суффиксы могут поддерживаться у некоторых глаголов, для которых были созданы отдельные парадигмы.

Таблица 3. Какие из объектных суффиксов для каких пород глаголов поддерживаются в настоящее время (A — суффикс прямого объекта, D — суффикс непрямого объекта, 2 — двойной суффикс).

порода	**претерит**			**субъюнктив**			**императив**			**презенс**			**перфект**		
	A	D	2	A	D	2	A	D	2	A	D	2	A	D	2

I	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
II	+	+	+	+	+	+				+	+	+	+	+	+
III	+	+	+	+	+	+				+	+	+	+	+	+
IV	+	+	+	+	+	+				+	+	+	+	+	+

UniParser поддерживает омонимию. Её можно снять различными способами, например, с помощью модуля ConstraintGrammar, но это не входит в рамки этой работы.

4.2 Перенос на корпусную платформу

Чтобы перенести размеченные тексты на платформу для корпусов, мы решили использовать *tsakorpus*, разработанный также Т. Архангельским. Корпус представляет собой удобную платформу, на которой можно осуществлять поиск как по словоформам и леммам, так и частям речи и грамматическим признакам и их сочетаниям. *tsakorpus* пользуется популярностью среди исследователей НИУ ВШЭ.

Tsakorpus использовался для создания корпусов многих малых языков, таких как чукотский (чукотско-камчатская семья) (Stenin, Garanina 2018), адыгейский (абхазо-адыгская семья) (Arkhangelskiy et al.), эвенкийский (тунгусская семья) (Казакевич и др.), башкирский (тюркская семья) (Ovsyannikova et al. 2017).

В качестве корпусной платформы мы выбрали *tsakorpus*, во-первых, потому что у него богатый функционал: он позволяет осуществлять сложные запросы (искать по морфологическим характеристикам, леммам, частям речи), а также делать отрицательные запросы, задавать расстояние между словами, выбирать подкорпус на основе метаданных, создавать параллельные корпуса, слушать аудио- и видеофрагменты предложений и пр. Во-вторых, на нём также были выложены корпуса для туройо (Lyavdanskiy et al. a) и христианского урмийского (Lyavdanskiy et al. b), поэтому мы продолжим традицию и в этом.

Наш корпус состоит из 191 текста. 99 из них взяты из сборника (Arnold 1991a), 74 — из сборника (Arnold 1991b), 6 — из учебника (Arnold 2006) и 12 — из сборника стихотворений и песен (Arnold 2002). Все тексты снабжены немецким переводом, кроме двух текстов из учебника, для которых перевод отсутствует.

Чтобы загрузить морфологически размеченные тексты на *tsakorpus*, нужно привести тексты в нужный формат. *Tsakorpus* предлагает готовые конверторы, но для

того чтобы ими воспользоваться, нужно всё равно привести данные в определённый формат. Наши данные были из разных источников и очень разнородные по формату, поэтому мы решили написать функцию, которая сразу приведёт их в нужный для корпуса формат — файл .json определённой структуры. Для каждого текста указываются метаданные: говорящий, пол говорящего, возраст говорящего, его профессия, деревня, в которой записан разговор, и тема разговора. Метаданные были извлечены из датасета MASC.

Предложения в файле .json разбираются на слова, для арамейских слов приводятся леммы, части речи и морфологический разбор, для немецких — только леммы и части речи, так как при проведении арамейских исследований изредка бывает нужно выяснить, как переводится то или иное слово на арамейский. Леммы и части речи для немецких слов были получены с помощью библиотеки *stanza* (Qi et al. 2020).

Так как корпус параллельный, для каждой пары арамейского и немецкого предложения указывается уникальный id, который позволяет их сопоставлять. Наш корпус содержал аудиофайлы, поэтому для каждого предложения ещё приводились тайминг и ссылка на аудио. Тайминг доставался из TextGrid файлов из датасета MASC с помощью модуля *textgrid*.

Наличие аудиофайлов в корпусе является большим плюсом, поскольку позволяет:

- проверять правильность затранскрибированного текста. Хотя тексты и были выверены много раз, ошибки всё равно могут быть допущены.
- проверять эмфатический/неэмфатический согласный, потому что носители могут по-разному эмфатизировать звуки в разных словах (Arnold 1990a: 16).

Также важны и метаданные, так как они позволяют проводить на корпусе социолингвистические исследования.

4.3 Создание инструмента для глоссирования

Поскольку НУГ из НИУ ВШЭ и лингвисты из других университетов используют разные системы глоссирования, инструмент был заточен под нужны НУГ НИУ ВШЭ.

Инструмент представляет собой функцию на Python, на вход которой подаётся текст и которая возвращает его в отглоссированном формате. Выдача состоит из двух рядов предложений, каждой слово отделено табуляцией. Токенизация происходила с помощью бибилотеки *NLTK*. Далее каждое слово получает морфологический анализ с

помощью разработанного парсера. Поскольку омонимия на данный момент не снята, инструмент на данный момент возвращает первый из возможных анализов. В дальнейшем планируется сделать так, чтобы он возвращал все возможные варианты через дробь или в другом удобном для пользователя виде.

Первый ряд выдачи представляет собой поморфемную сегментацию согласно правилам глоссирования НУГ. Во втором ряду приводится отсглоссированный вариант, причём поскольку на данный момент ещё нет арамейско-немецких (и арамейско-английских) пар «слово — перевод», все переводы заменяются словом «STEM». Исключение представляют собой некоторые служебные слова и наречия, которые имеют всегда один и тот же перевод вне зависимости от контекста (например, местоимения, некоторые предлоги, частицы, наречия времени и пр.). В этом есть и свой плюс: во-первых, многие существительные, прилагательные и глаголы имеют по несколько значений, выбор которых зависит от контекста; во-вторых, многие слова принадлежат к лексике, связанной с местной культурой, сельским хозяйством или другими местными реалиями, перевод которых на английский язык будет затруднён. Поэтому на данном этапе предполагается, что после получения отглоссированного текста лингвист самостоятельно определит, какие переводы подходят лучше, и пропишет их вместо слов «STEM».

Пример выдачи инструмента для глоссирования выглядит следующим образом:

Листинг 5. Пример входного текста для инструмента для глоссирования.

isleḳ ʕa ʕarḳūba, ḥmull ʕazīz, laḳṭunne.

ʕazīz m-zawʕe miskīna šayšar bə-brōḳe.

amellun: «w ḥayyil alō, čūb ana, hanna fēris ti taḥəklil xēfa».

iṯḳen mtawwrin aʕəl, ana niṭmer, la sčahət aʕəl.

ʕirpaṯ šimša w ana nḳayyam elʕel.

Листинг 6. Выдача инструмента для глоссирования для текста из Листинга 5.

isleḳ	ʕa=ʕarḳūb-a	UNK	ʕazīz	laḳṭ-un-n=e
STEM.PST	PP=STEM-FREE		UNK PN	STEM.PST-3mp-PLEO=3ms

ʕazīz	b=zawʕ=e	miskīn-a	šayšar b=brōḳ=e

PN	in=STEM=3ms	STEM-FREE		STEM.PST		in=STEM=3ms			
amel-lunw	UNK	alō	čūb	ana	hanna	fēris	ti	UNK	xēf-a
STEM.PST-3mp.IO		and	UNK	God	STEM	I	this.M	PN	REL
UNK	STEM-FREE								
iṯ̲ken	mtawwr-in	aʕəl	ana	ni-ṭmer	la	UNK	aʕəl		
STEM.PRF	STEM.PRS-MP	to	I	1-STEM.PRF		NEG	UNK	to	
UNK	šimš-a	w	ana	UNK	elʕel				
UNK	STEM-FREE	and	I	UNK	above				

5. Результаты

5.1 Парсер для диалекта Маалулы

Был создан морфологический парсер для диалекта Маалулы. Ниже представлены таблицы, показывающие точность парсера на разных частях речи по трём параметрам: точность определения лемм, частей речи и грамматических признаков. Так как в корпусе не снята омонимия, определение считалось верным, если хотя бы один из разборов совпадал с эталоном. Определение грамматических признаков считалось верным, если совпадали все признаки. Если не совпадал хоть один грамматический показатель, разбор считался неверным.

В качестве эталона были размечены 100 предложений. При выборе мы руководствовались правилом, что в эталоне не должно быть двух предложений из одного текста (поскольку тогда повышается вероятность повторения одних и тех же слов, о которых идёт разговор в тексте, а для оценки качества лучше иметь как можно более разнообразные слова). Из датасета MASC стало известно, что тексты в корпусе из томов III и IV принадлежат к 11 разным тематикам (см. Таблицу 4). Поэтому было решено выбрать предложения автоматически случайным образом с учётом баланса этих тем, а именно: 2 предложения из текстов про басни, 6 — из суеверий, 6 — из рассказов о мусульманских традициях, 8 — из песен, и по 11предложений из текстов всех остальных тем. Так как в сумме это дало 99, а в корпусе, помимо рассказов из томов III и IV, есть ещё тексты из учебника и песни из сборника песен, было решено

добавить ещё одно предложение из сборника песен (так как не для всех текстов из учебника есть немецкий перевод, что затруднило бы проверку правильности разметки).

Таблица 4. Распределение текстов из томов III и IV по тематикам.

Тема	**Количество текстов**
Личный опыт и события (Personal experiences and events)	34
Шутки и анекдоты (Jokes and anecdotes)	30
Сказки (Fairy tales)	23
Занятия дома и в деревне (Activities at home and in the village)	20
Христианские традиции (Religious traditions and beliefs (the Christian community)	20
Сельское хозяйство (Occupational and agricultural activites)	13
Профессиональные знания (Lore)	11
Песни и поэмы (Songs and poems)	8
Мусульманские традиции (Religious traditions and beliefs (the Muslim community)	6
Суеверия (Superstitions)	6
Басни (Fables)	2

Результаты представлены в четырёх таблицах. В Таблице 5 приведены показатели только для словоупотреблений, которые распознались парсером, в Таблице 6 — для всех словоупотреблений (в том числе не распознанных), в Таблице 7 представлены показатели для уникальных распознанных словоформ, в Таблице 8 — для всех уникальных словоформ, а в Таблице 9 приведены общие показатели на основе средневзвешенного значения.

Таблица 5. Точность определения грамматических характеристик в зависимости от части речи (только на распознанных словоупотреблениях).

Часть речи	**Всего**	**Точность лемм**	**Точность частей речи**	**Точность грамматических признаков**
Существительные	229	226 (98,69%)	227 (99,13%)	225 (98,25%)
Глаголы	228	227 (99,56%)	228 (100%)	222 (97,37%)

Вспомогательные глаголы и псевдоглаголы	38	38 (100%)	38 (100%)	38 (100%)
Прилагательные	19	19 (100%)	19 (100%)	19 (100%)
Наречия	49	47 (95,92%)	47 (95,92%)	47 (95,92%)
Местоимения	48	48 (100%)	48 (100%)	48 (100%)
Числительные	23	23 (100%)	23 (100%)	23 (100%)
Сочинительные союзы	19	18 (94,74%)	18 (94,74%)	18 (94,74%)
Подчинительные союзы	83	83 (100%)	83 (100%)	83 (100%)
Детерминативы	29	29 (100%)	29 (100%)	29 (100%)
Предлоги	134	134 (100%)	134 (100%)	134 (100%)
Частицы	32	32 (100%)	32 (100%)	32 (100%)
Междометия	8	8 (100%)	8 (100%)	8 (100%)
Имена собственные	13	12 (92,31%)	12 (92,31%)	12 (92,31%)
Всего	**952**	**944 (99,16%)**	**946 (99,37%)**	**938 (98,53%)**

Полностью совпавших разборов (т. е. разборов, где были одновременно верно определены и лемма, и часть речи, и все грамматические признаки): 938. Как можно видеть, это все разборы из последней колонки, т. е. если получилось, что если правильно разбираются грамматические признаки, то правильно разбирается и всё остальное.

Можно заметить, что лучше всего анализируются служебные части речи, поскольку они представляют собой закрытые классы и не обладают богатым словоизменением. Самый низкий показатель точности наблюдается у глагола и имён собственных. В случае с глаголами это связано с их богатой морфологией и обилием грамматических признаков, в случаях с именами собственными — отсутствие их полного списка и невозможность его получения простым путём ввиду того, что они не подчиняются какому-то набору паттернов.

Таблица 6. Точность определения грамматических характеристик в зависимости от части речи на всех словоупотреблениях (в том числе на нераспознанных):

Часть речи	**Всего**	**Точность лемм**	**Точность частей речи**	**Точность грамматических признаков**
Существительные	251	226 (90,04%)	227 (90,44%)	225 (89,64%)

Глаголы	296	227 (76,69%)	228 (77,03%)	222 (75%)
Вспомогательные глаголы и псевдоглаголы	38	38 (100%)	38 (100%)	38 (100%)
Прилагательные	21	19 (90,48%)	19 (90,48%)	19 (90,48%)
Наречия	53	47 (88,68%)	47 (88,68%)	47 (88,68%)
Местоимения	48	48 (100%)	48 (100%)	48 (100%)
Числительные	23	23 (100%)	23 (100%)	23 (100%)
Сочинительные союзы	20	18 (90%)	18 (90%)	18 (90%)
Подчинительные союзы	83	83 (100%)	83 (100%)	83 (100%)
Детерминативы	29	29 (100%)	29 (100%)	29 (100%)
Предлоги	135	134 (100%)	134 (100%)	134 (100%)
Частицы	33	32 (96,97%)	32 (96,97%)	32 (96,97%)
Междометия	9	8 (88,89%)	8 (88,89%)	8 (88,89%)
Имена собственные	16	12 (75%)	12 (75%)	12 (75%)
Всего	**1055**	**944 (89,48%)**	**946 (89,67%)**	**938 (88,91%)**

Среди нераспознанных глаголов были: глаголы, относящиеся к тем породам, которые сейчас не поддерживаются; глаголы с объектными суффиксами, относящиеся к тем породам, которые сейчас поддерживаются, но для которых не все объектные суффиксы сейчас доступны; глаголы с орфографией, отличной от стандартной (иногда такое бывает в корпусе, когда автор транскрипции захотел отразить частный случай нестандартного произношения, например, показать эмфатизацию).

Среди нераспознанных существительных были: заимствования из арабского (которые не подчиняются арамейским паттернам, а потому не были извлечены из датасета MASC); существительные с орфографией, отличной от стандартной; существительные подходящих паттернов, но которые по разным причинам не извлеклись из датасета MASC.

Таблица 7. Уникальные распознанные словоформы.

Часть речи	**Всего**	**Точность лемм**	**Точность частей речи**	**Точность грамматических признаков**
Существительные	188	185 (98,40%)	186 (98,94%)	184 (97,87%)
Глаголы	192	191 (99,48%)	192 (100%)	186 (96,88%)

Вспомогательные глаголы и псевдоглаголы	18	18 (100%)	18 (100%)	18 (100%)
Прилагательные	18	18 (100%)	18 (100%)	18 (100%)
Наречия	25	23 (92%)	23 (92%)	23 (92%)
Местоимения	21	21 (100%)	21 (100%)	21 (100%)
Числительные	14	14 (100%)	14 (100%)	14 (100%)
Сочинительные союзы	9	8 (88,89%)	8 (88,89%)	8 (88,89%)
Подчинительные союзы	2	2 (100%)	2 (100%)	2 (100%)
Детерминативы	16	16 (100%)	16 (100%)	16 (100%)
Предлоги	38	38 (100%)	38 (100%)	38 (100%)
Частицы	11	11 (100%)	11 (100%)	11 (100%)
Междометия	6	6 (100%)	6 (100%)	6 (100%)
Имена собственные	13	12 (92,31%)	12 (92,31%)	12 (92,31%)
Всего	**571**	**563 (98,60%)**	**565 (98,95%)**	**557 (97,55%)**

Таблица 8. Все уникальные словоформы.

Часть речи	**Всего**	**Точность лемм**	**Точность частей речи**	**Точность грамматических признаков**
Существительные	208	185 (88,94%)	186 (89,42%)	184 (88,46%)
Глаголы	258	191 (73,03%)	192 (74,42%)	186 (72,09%)
Вспомогательные глаголы и псевдоглаголы	18	18 (100%)	18 (100%)	18 (100%)
Прилагательные	20	18 (90%)	18 (90%)	18 (90%)
Наречия	29	23 (79,31%)	23 (79,31%)	23 (79,31%)
Местоимения	21	21 (100%)	21 (100%)	21 (100%)
Числительные	14	14 (100%)	14 (100%)	14 (100%)
Сочинительные союзы	10	8 (80%)	8 (80%)	8 (80%)
Подчинительные союзы	2	2 (100%)	2 (100%)	2 (100%)
Детерминативы	16	16 (100%)	16 (100%)	16 (100%)
Предлоги	39	38 (97,44%)	38 (97,44%)	38 (97,44%)

Частицы	12	11 (91,67%)	11 (91,67%)	11 (91,67%)
Междометия	7	6 (85,71%)	6 (85,71%)	6 (85,71%)
Имена собственные	16	12 (75%)	12 (75%)	12 (75%)
Всего	**670**	**563 (84,03%)**	**565 (84,33%)**	**557 (83,13%)**

Таблица 9. Общая точность парсера (средневзвешенное).

	Точность лемм	**Точность частей речи**	**Точность грамматических признаков**
На распознанных словоупотреблениях	99,18%	99,39%	98,55%
На всех словоупотреблениях	89,49%	89,68%	88,92%
На распознанных уникальных словоформах	98,60%	98,95%	97,55%
На всех уникальных словоформах	84,03%	84,33%	83,13%

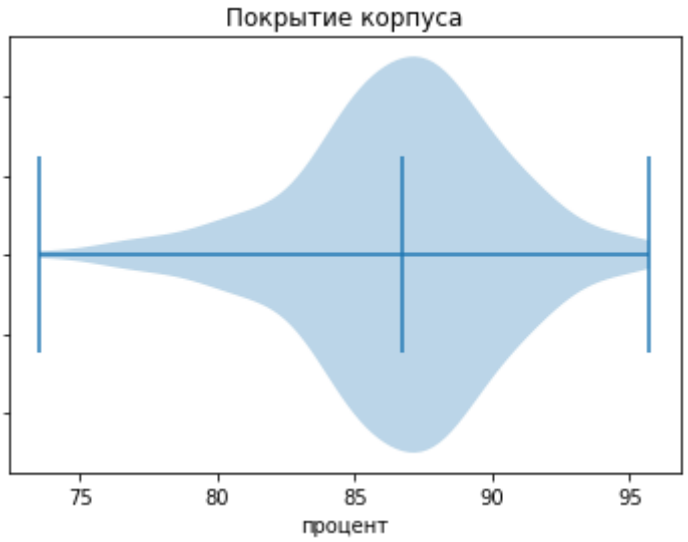
Таким образом, точность парсера в определении лемм, частей речи и грамматических характеристик на всех словах составляет около 83–89%. Если же слова относятся к категории распознанных, то хотя бы один из разборов будет верен в 97–98% случаев.

5.2 Аннотированный мультимедийный корпус для диалекта Маалулы

С помощью парсера были размечены все имеющиеся в нашем распоряжении тексты из Маалулы. Полученный объём арамейского корпуса — 191 текст, 5802 предложения. Всего токенов: 68710, из них 12584 — уникальных, 8927 — не распознанных, 4909 — уникальных нераспознанных. Таким образом, покрытие по всему корпусу, то есть процент разобранных токенов, составляет 87.01%. Если же смотреть по отдельным текстам, то минимальное покрытие составляло 73.53% (текст принадлежал к теме «шутки и анекдоты»), а максимальное — 95.71% (текст принадлежал к теме «личный опыт и события»).

На Диаграмме 1 показано распределение текстов по покрытию. Видно, что основной массив сосредоточен в районе 85–87%.

Диаграмма 1. Покрытие корпуса в процентах.



Всего было 11 тематик текстов, а именно: занятия дома и в деревне, шутки и анекдоты, христианские религиозные традиции и поверья, мусульманские религиозные традиции и поверья, профессиональные и сельскохозяйственные занятия, профессиональные знания, басни, сказки, суеверия, песни и поэмы, личный опыт и события. На Диаграмме 2 показано среднее покрытие текста в зависимости от его тематики. На Диаграмме 3 показан разброс среднего покрытия в зависимости от тематики текста.

Диаграмма 2. Покрытие корпуса в процентах в зависимости от тематики текста.

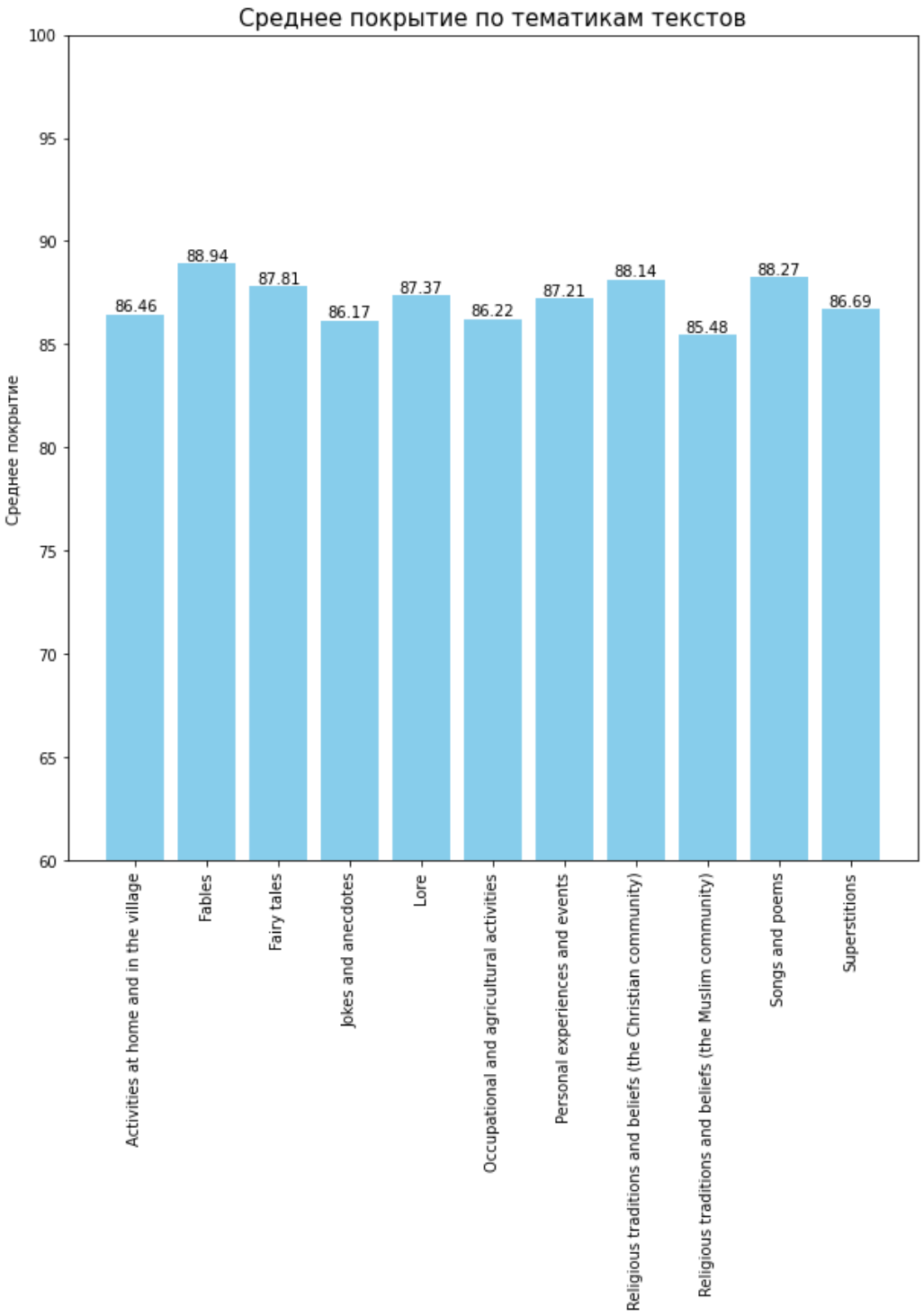
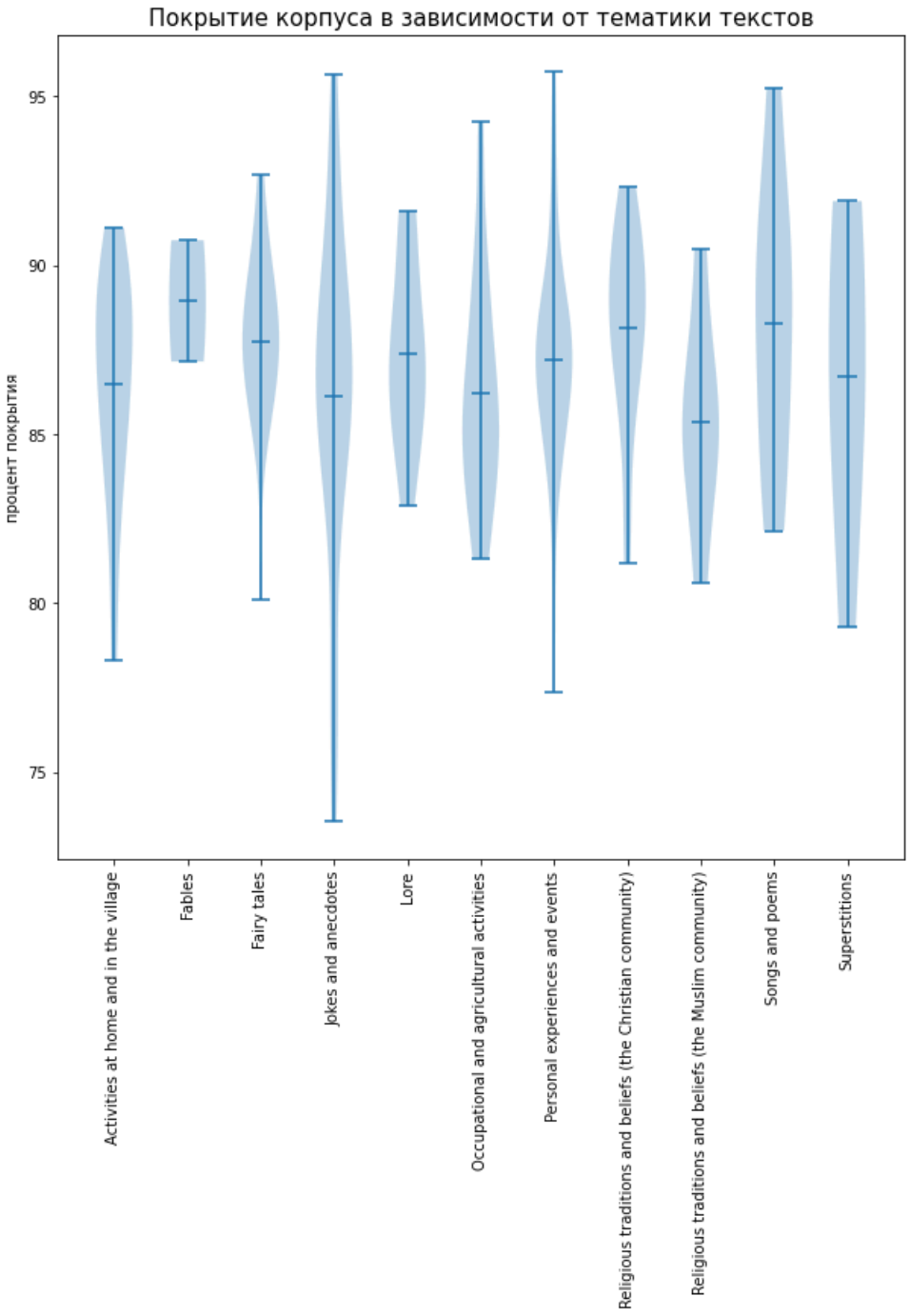

Среднее покрытие по тематикам текстов
Среднее покрытие
100
95
90
85
80
75
70
65
60
86.46
88.94
87.81
86.17
87.37
86.22
87.21
88.14
85.48
88.27
86.69
Activities at home and in the village
Fables
Fairy tales
Jokes and anecdotes
Lore
Occupational and agricultural activities
Personal experiences and events
Religious traditions and beliefs (the Christian community)
Religious traditions and beliefs (the Muslim community)
Songs and poems
Superstitions

Диаграмма 3. Покрытие текста в зависимости от тематики.



Из Диаграммы 2 видно, что среднее покрытие текста не сильно зависит от его темы. На Диаграмме 3 показано, что самый большой разброс наблюдался у текстов, относящихся к темам «шутки и анекдоты» и «личный опыт и события». Наименьший

разбор наблюдается у басен, поскольку к басням относились всего два текста. Разброс по текстам остальных тем находится в пределах 10–12%.

Что касается среднего покрытия, то наименьшее среднее наблюдается у текстов, относящихся к тематике «мусульманские традиции». Это может быть связано с тем, что в таких текстах присутствует особо много арабизмов, которые не соответствуют арамейским паттерном, а потому не извлеклись из датасета MASC и не попали в файл с лексемами.

У всех нераспознанных словоформ в корпусе частота составляет 28 словоупотреблений и менее. Туда вошли:

- заимствования из арабского: как слова, связанные с арабской культурой, например, *imōma* ‘имам’, *ḳurʔān* ‘Коран’, так и обычные заимствования, такие как *čiḳrīban* ‘приблизительно’.
- заимствования из других языков: *matte* ‘мате (напиток)’, *kīlo* ‘килограмм’, *káhraba* ‘электричество’, *šufēr* ‘водитель’.
- имена собственные: например, имена *Brōm*, *Krunbe*.
- глаголы с объектными суффиксами, для которых ещё не поддерживается разбор с этими суффиксами: например, *applēle* ‘он дал ему его’, *maytyill* ‘он приносит (с аффиксом DOM)’.
- глаголы, отсносящиеся к породам или подтипам основных пород, которые ещё не поддерживаются: например, *mʕannyin* ‘они (m.) поют’, *čbaḳḳ* ‘он остался’.
- некоторые наречия: *bnawb* ‘вообще, совсем’.
- слова с неожиданными чередованиями: например, *ʕimmiš* ‘с тобой (f.)’ вместо ожидаемого *ʕemmiš*, *yinfuḳ* ‘(чтобы) он вышел (SBJV)’ вместо *yiffuḳ*.
- слова, не являющиеся заимствованиями, но не попавшие ни под один паттерн по другим причинам: например, *šoppṯa* ‘суббота’, *sʕarō* ‘ячмень’.

5.3 Инструмент для глоссирования

Была написана функция на Python, на вход которой подаётся текст и которая возвращает его в отглоссированном виде. Также был создан инструмент для оценки работы этой функции, который считает точность морфологической сегментации и морфологической аннотации. Точность работа инструмента для глоссирования

проверялась на эталонном тексте размером 47 предложений. Результаты представлены в Таблице 8.

Таблица 8. Точность работы инструмента для глоссирования.

	Точность морфологической сегментации	**Точность морфологической аннотации**
На распознанных словоупотреблениях	89.963%	76.208%
На всех словоупотреблениях	73.556%	62.31%
На распознанных уникальных словоформах	91.169%	78.043%
На всех уникальных словоформах	74.006%	62.691%

Среди неверных случаев сегментации и аннотации оказались случаи, когда:

- вместо дефиса использовался знак равенства или наоборот. Например, STEM-3ms вместо STEM=3ms.
- для предлога выдавался перевод по умолчанию (который используется в большинстве контекстов), но в данном конкретном случае он не подходил по контексту. Например, in=STEM=HD вместо on=STEM=HD для *b=ḥaṣṣ=l* ‘на спине’.
- выделялся аффикс, который выделяется в обычном случае, но не для данной конкретной лексемы. Например, *ō* в слове *ffō* ‘лицо’ на самом деле не показатель плюралиса, а часть основы, поэтому когда к этому слову присоединяются посессивные аффиксы, его следует глоссировать не как *ff-ōy=e* STEM-P=3ms (что верно для обычных существительных), а как *ffōy=e* STEM=3ms.
- неверно распозналось слово. Например, *ṭálama* ‘вот’ отглоссировалось по правилам для существительных, хотя это на самом деле частица.

6. Исследования на материале полученного корпуса

6.1 Исследование омонимии

При разметке предложений возникла ожидаемая омонимия — это формы 3 л. претерита, некоторые формы субъюнктива, перфект I и II пород, рефлексивы и существительные, от которых они образованы, некоторые формы прилагательных и

глаголов, некоторые предлоги и частицы. Омонимию можно будет в дальнейшем снять с помощью ConstraintGrammar или другими способами. Примеры на каждый из случаев перечислены ниже.

Омонимичными будут формы глаголов, стоящих в 3 л. ед. ч. и мн. ч. претерита:

(7) *aḳam aḳīme walla **iščaḥ** erraʕ menne ġurnōyṯiḏ ḏahba.*

aḳam	*aḳīm-e*	*walla*
встать.PST.3ms	отодвигать.PST.3ms-DO.3ms	и

iščaḥ	*erraʕ*	*menn-e*	*ġurnōy-ṯ-il*	*ḏahb-a.*
найти.PST.3ms	внизу	из-POSS.3ms	тарелка-F-HD	золото-FREE

‘Он взял и отодвинул его и **нашёл** внизу тарелку, полную золота.’

(8) *iʕber l-elġul willa **iščaḥ** ṣaḥna uppe biʕō šlīḳan.*

iʕber	*l-elġul*	*willa*	*iščaḥ*
войти.PST.3cp	к-внутрь	и	найти.PST.3cp

ṣaḥn-a	*upp-e*	*biʕ-ō*	*šlīḳ-an.*
тарелка-FREE	AUX-3ms	яйцо-P	вареный-FP.INDF

‘Они вошли и **нашли** тарелку, в которой были варенные яйца.’

В субъюнктиве наблюдается омонимия у форм 3fs, 2ms и 2fs: *čizbun* это одновременно и ‘(чтобы) она купила’, и ‘(чтобы) ты купил’, и ‘(чтобы) ты купила’ (Arnold 1990a: 72).

Также совпадают по форме перфект I и II породы: например, *šammīʕa* может быть перфектом 3 л. ед. ч. ж. р. как глагола I породы *išmeʕ, yišmaʕ* ‘слушать’, так и глагола II породы *šammaʕ, yšammaʕ* ‘позволять слушать’

В СЗА многие рефлексивы произошли от существительных, поэтому все они омонимичны соответствующим существительным, например: *ōmar b-leppe* ‘он сказал себе (букв. он сказал в своём сердце)’, *ōmar b-ʕaḳle* ‘он сказал себе (букв. он сказал в своём разуме)’, *ōmar b-nefše* ‘он сказал себе (букв. он сказал в своей душе)’ (Arnold 1990a: 48).

Многие прилагательные исторически произошли от глаголов через стадию причастия, поэтому многие из них до сих пор омонимичны глаголам в презенсе и перфекте, поскольку именно эти времена также произошли от причастий. Например, *ixfen* ‘он проголодался/он голодный’ *nixfen* ‘я проголодался/я голодный’.

Среди служебных частей речи также есть омонимия. Например, *ya* может быть как вокативной частицей, так и сочинительным союзом. Например, *ya eppay* ‘о мой отец’, *ya ʕīla ya ʕīlča* ‘или молодой осёл, или молодая ослица’.

Также изредка возникала омонимия между разными частями речи, например, когда существительное совпало по форме с глаголом или имя собственное совпало с существительным, поскольку было от него образовано. В первом случае примером может служить слово *ḥalba* ‘молоко’, которое теоретически может совпасть с формой 3 л. ед. ч. м. р. претерита с объектным суффиксом 3 л. ед. ч. ж. р. ‘(он) подоил её’, однако на практике второго значения не встретилось. Во втором случае, например, есть фамилия *Šōʕra*, которая буквально значит ‘поэт’.

Есть омонимия, которая снимается фонетическими условиями, например, предлог *b* ‘в’ перед словами, начинающимися на *m-*, имеет аллофон *m* В то же время есть предлог *m* ‘из’. Соответственно, если перед словом, начинающимся не на *m-*, стоит предлог *m*, то это точно будет ‘из’, а не ‘в’. Например, *m-maʕlūla* — ‘в Маалуле / из Маалулы’, *m-payṯa* — ‘из дома’ (но не ‘в доме’).

Другой тип омонимии снимается синтаксическими условиями. Например, у некоторых существительных счётная форма и форма с POSS.1cs могут совпадать, например, *ḥūn* ‘мой брат’ и *iṯər ḥūn* ‘два брата’. Однако если перед словом стоит числительное, то это с большей вероятностью будет именно счётная форма. Другим примером омонимии этого типа могут быть предлог и частица прогрессива *ʕa*: если после *ʕa* следует существительное, то это предлог, а если глагол — то частица. Например, *ʕa ḏayra* — ‘к монастырю’, *ʕa nḏōḥkin* — ‘(мы) смеёмся (прямо сейчас)’.

Третий тип омонимии снимается контекстом. Сюда относятся такие случаи, как совпадение 3 л. ед. ч. и мн. ч. в претерите, совпадение форм субъюнктива, форм с аккузативными и дативными объектными суффиксами у глаголов в презенсе и перфекте. Пример на последний из случаев: *ftiḥōle* — ‘она открыла его/она открыла ему’.

6.2 Исследование субъюнктива

В грамматике (Arnold 1990a) и учебнике (Arnold 2006) все примеры субъюнктива даются только в составе аналитической конструкции со вспомогательным глаголом *batte*, в связи с чем поднимается вопрос, могут ли глаголы в форме субъюнктива употребляться без этого вспомогательного глагола. Для этого в корпусе был задан запрос “(-batte) & (VERB,Subj)”. В выдаче оказалось 368 предложений, из них были отобраны первые 30 подходящих. В результате выяснилось, что субъюнктив может употребляться при наличии определённых слов, задающих необходимую модальность, таких как некоторые псевдоглаголы, подчинительные союзы “когда”, “если”, “после того как” и пр. Также встретилось употребление субъюнктива без каких бы то ни было подобных слов в устойчивых выражениях, где субъюнктив выражает пожелание. Примеры, разделённые на категории, которые удалось выделить, представлены ниже (полужирным выделены формы субъюнктива).

Употребление со вспомогательными глаголами или псевдоглаголами:

(9) *sḳillinnaḥ ṯlōṯa arpʕa yūm b-ḏemseḳ, la aktrinnaḥ* ***nislaḳ*** *m-ṯelka.*

sḳill-innaḥ	ṯlōṯa	arpʕa	yūm	b=ḏemseḳ	la
остаться.PST-1P	три.M	четыре.M	день.NUM	в=Дамаск	NEG

aḳtr-innaḥ	ni-slaḳ	m=ṯelk-a
мочь.PST-1P	1-подниматься.SBJV	из=снег-FREE

‘Мы оставались три, четыре дня в Дамаске, (так как) не могли **подняться** из-за снега.’

(10) *la affne* ***yisḳaṯ****, lā ʕal_ommṯa w lā ʕal_arʕa* [...]

la	aff-n-e	yi-sḳaṯ	lā
NEG	позволить.PST-PLEO-3ms	3M-падать.SBJV	NEG

ʕal=ommṯ-a	w	lā	ʕal=arʕ-a
к=люди-FREE	и	NEG	к=земля-FREE

‘Это не позволило ему **упасть**, ни на людей, ни на пол, [...]’

(11) *amrilla: «uppiš* ***čallix*** *w* ***əṯṯammṣill*** *ʕaynōš?»*

amril-la	upp=iš	č-allix	w	č-ṭammṣil-l

говорить.PRS-3fs.IO AUX=2fs 2-идти.SBJV и 2-закрыть-DOM
ʕayn-ō=š
глаз-P=2fs
‘Я сказал ей: «Можешь ли ты **пойти** и при этом **закрыть** глаза?»’

(12) *w t̲ēle ōblə ḥd̲ūt̲a mnaḳḳeṭla, w mižčamʕin marōylə ḥd̲ūt̲a xullun sawa, w ti bōʕ* ***ynaḳḳeṭ*** *m-xull lanna žamʕa, t̲ēle mnaḳḳeṭlun.*

w	t̲ē-le	ōb=l	ḥd̲ūt̲-a	mnaḳḳeṭ-la	w
и	идти.PRS-3ms	отец=HD	жених-FREE	дарить.PRS-3fs	и

mi-žčamʕ-in	mar-ōy=l	ḥd̲ūt̲-a	xull=un
3-собираться.PRS-MP	господин-P=HD	жених-FREE	весь=3mp

sawa	w	ti	bōʕ	y-naḳḳeṭ	m=xull	l-hanna
вместе	и	REL	AUX	3M-дарить.SBJV	из=весь	к-этот.M

žamʕ-a	t̲ē-le	mnaḳḳeṭ-lun
собрание-FREE	идти.PRS-3ms	дарить.PRS=3mp

‘И приходит отец жениха и одаряет её, и родственники жениха собираются все вместе, и кто хочет, **делает подарки** от этого всего собрания, он приходит одаряет их.’

(13) *amerle hanna ōġa: «lōzim* ***črōžaʕ*** *leʕle hōš,* ***čaḥəmtenne*** *w* ***čmalle*** *“ysallmell d̲wōt̲ax”, w illa d̲arrarlax.»*

amer-le	hanna ōġ-a	lōzim	č-rōžaʕ	
говорить.PST-3ms.IO	этот.M ага-FREE	AUX	2-возвращаться.SBJV	
leʕl=e hōš	č-aḥəmt-enn=e		w	
к=3ms сейчас	2-благодарить.SBJV-PLEO=3ms		и	
č-mal-le	y-sallmell	d̲w-ō-t̲-ax		w
2-говорить.PST-3ms.IO	3M-сохранить.SBJV	рука-P-F=2ms		и
illa	d̲arrar-lax			
чтобы.не	вредить.PST-2ms.IO			

‘Ага сказал ему: «Ты должна **вернуться** к нему немедленно, чтобы **поблагодарить** его и **сказать** “спасибо” (букв. “да сохранит (Господь) твои (m.) руки”), чтобы он не причинил тебе вреда.’ (ага — титул землевладельца)

(14) *xaṭərṯa nob* ***nallex*** *ana w lawandyus ebər ḥōl, willa ḥminnaḥ ḏapparīṯa b-arʕa.*

xaṭər-ṯ-a	n-ob	n-allex	ana	w	lawandyus
случай-F-FREE	1-AUX	1-идти.PRF	я	и	Лавандиус

ebər	ḥōl	willa	ḥm-innaḥ	ḏapparī-ṯ-a
сын	дядя.1S	и	видеть.PST-1P	пчела-F-FREE

b=arʕ-a
в=земля-FREE

‘Однажды я **шёл**, я и Лавандиус, сын моего дяди по материнской линии, и мы увидели пчелу на земле.’

(15) *l-ōš* ***nfakkar*** *ḏukkin nōb* [...]

l=hōš	n-fakkar	*ḏukkil*	n-ōb
к=сейчас	1-помнить.SBJV	когда	1-AUX

‘До сих пор я **помню**, как я была (там) [...]’

Употребление в придаточных цели и в сочетании с другими невспомогательными глаголами, которые указывают на цель или каузацию действия:

(16) *salma bōṯar mil ayṯaṯ ṯlōṯa arpʕa ibər amrōl beʕla: «yā ġabrōna, ana īl p-xōṭra* ***nzill*** [...].*»*

salma	bōṯar.mil	ayṯ-aṯ	ṯlōṯa	arpʕa	ibər
Сальма	после.того	родить.PST-3fs	три.M	четыре.M	сын.NUM

amrō-l	beʕl-a	yā	ġabrōn-a	ana
говорить.PRS-3mp.IO	муж-FREE	VOC	супруг-FREE	я

īl	b=xōṭr-a	n-zi-ll
AUX	в=смысл-FREE	1-идти.PRS-S

‘Сальма, после того как она родила трёх, четырёх детей, сказала своему мужу: «О муж, я думаю (букв. «имею в голове») **уйти**, чтобы [...]’

(17) *xaṭərṯa iččžaʕ ḥūn w emmay wayba ču ḳattīra **čaḥḥčenne** ʕa ḏemseḳ xett, niḥčiṯ ʕemmil eppay.*

xaṭər-ṯ-a	iččža ʕ	ḥūn	w	emmay
случай-F-FREE	болеть.PST	брат.1S	и	мать.1S

w-ayb-a	ču	ḳattīr-a	č-aḥḥč-enn=e	ʕa=ḏemseḳ
PST-быть-F	NEG	мочь.PRF-3fs	2-везти.SBJV-PLEO=3ms	в=Дамаск

xett	niḥč-iṯ	ʕemmil	eppay
поэтому	ехать.PST-1S	с	отец.1S

‘Однажды мой брат заболел, и моя мать также была не в состоянии **отвезти** его в Дамаск (в больницу), поэтому я поехал со своим отцом.’

(18) *«ayṯāy **naḥšem**!» ṯōle aḥšem.*

ayṯ-āy	n-aḥšem	ṯō-le	aḥšem
нести.IPV-F	1-ужинать.SBJV	идти.PST-3ms	ужинать.PST

‘«Принеси их, (чтобы) мы **поужинали**!» Он пришёл и ужинал.’

(19) *amrōla: «alō yaffinniš, ana ġarīpča w čūl ḏokkṯa **niḏmux**.*

amrō-la	alō	y-aff-inn-iš	ana	ġarīp-ča
говорить.PRS-3fs.IO	Бог	3-воздать.SBJV-PLEO-2fs	я	чужой-FS.DEF

w	čūl	ḏokk-ṯ-a	ni-ḏmux
и	NEG	место-F-FREE	1-спать.SBJV

‘Она сказала ей: «Да воздаст тебе Бог, я чужая здесь и мне негде **переночевать** (букв. “у меня нет места, (чтобы) переночевать”).’

(20) *eppay šattar ommṯa **ytawwrun** aʕəl, azaʕ aʕəl, liʔannu ana nwaḥtōnay nōb ġappe p-payṯa.*

eppay	šattar	ommṯ-a	y-tawwr-un	aʕəl
отец.1S	отправлять.PST	люди-FREE	3-искать.SBJV-MP	к.1S

azaʕ	aʕəl	liʔannu	ana	n-waḥtōn-ay
бояться.PST	к.1S	так.как	я	1-единственный-MS.INDF

n-ōb	ġapp=e	b=payṯ-a
1-AUX	c=3ms	в=дом-FREE

‘Мой отец отправил людей **искать** меня, потому что он боялся за меня, так как я был у него единственным ребёнком в семье.’

(21) *bōṯar meṯṯa mḥáttitin yōma, **ynuḥčun yxuṯpull** xṯōba, hanna b-nesəpṯa s-sarḳōy.*

bōṯar	meṯṯ-a	mḥáttit-in	yōm-a	y-nuḥč-un
после	время-FREE	определить.PRS-MP	день-FREE	3-ехать.SBJV-MP

y-xuṯp-ul=l		xṯōb-a	hanna
3-писать.SBJV-PLEO=DOM		договор-FREE	этот.M

b=nesəp-ṯ-a	l=sarḳ-ōy
в=отношение-F-FREE	к=мусульманин-MP.DEF

‘Через некоторое время они устанавливают день, чтобы **съездить** (в город), чтобы они **написали** брачный договор, это касается только мусульман.’

(22) *ṯēle **yiʕbar** — bahheč.*

ṯē-le	yi-ʕbar	bahheč
идти.PRS-3MS	3M-входить.SBJV	стыдиться.PRF

‘Он пришёл, (чтобы) **войти** — но ему было стыдно.’

(23) *nzayyeʕ minnayn, yaʕni **yḳuṭlunn**, bessi ʕanmičẓōhar, inne ču nzayyeʕ.*

n-zayyeʕ	minnay=n	yaʕni	y-ḳuṭl-unn-∅
1-бояться.PRF	от=3P	значит	3-избить-PLEO-1S

bess	ʕa=n-mičẓōhar	inne	ču	n-zayyeʕ
но	PP-1-набраться.смелости.PRS	будто	NEG	1-бояться.PRF

‘Я боялся их, то есть что они меня **побьют**, но я набрался смелости, как будто я не боюсь.’

(24) *ayṯunne l-ʕal-anna malka imeṯ, fakkar hanna malka inne: ya rēt, yīb ntaššīrle p-šōrʕa* ***yīmuṯ****, w lā yīb mawčṯe ʕayatt.*

‘После того как они привели его к королю мёртвым, думал король так: Если бы я его на улице оставил, чтобы он **умер**, тогда он бы умер не из-за меня.’

(25) *waḳčlə mṭinnaḥ b-žubaylō —, atar b-žubaylō ōṯ ḳaṣṭla, šawwiyille ḥetta* ***ymarrḳull*** *mōya bē.*

waḳčil	mṭ-innaḥ	b=žubaylō	atar	b=žubaylō
когда	прибыть.PST-1P	в=Žubaylō	значит	in=Žubaylō

ōṯ	ḳaṣṭl-a	šawwiy-il-le	ḥetta
AUX	водопроводная.труба-FREE	делать.PRF-PLEO-3ms	чтобы

y-marrḳ-ul=l	mōy-a	bē
3-направлять.SBJV-PLEO=1S.IO	вода-FREE	в.3ms

‘Когда мы прибыли в Žubaylō —, значит, в Žubaylō есть водопроводная труба, которую сделали, чтобы через неё воду **направить**.’

(26) *ē, marōylə ḥḏūṯa rōfḏin* ***yiščun*** *ḳahwe.*

ē	mar-ōy=lə	ḥḏūṯ-a	rōfḏ-in
да	человек-P=HD	жених-FREE	сопротивляться.PST-3MP

yi-šč-un	ḳahwe
3-пить.SBJV-MP	кофе

‘Да, родственники жениха отказываются **пить** кофе.’

Употребление со значением будущего времени, но без *batte*:

(27) *ṯōle ebr əwzīra, amelle: «hann ḥiməš ḏahəb, w zēx ḳʕāx p-ḳahwe haččˇ, w ana* ***nzill*** *b-ḏokkṯax bil-lēlya!»*

ṯō-le	ebr	wzīr-a	amel-le
идти.PST-3ms	сын	министр-FREE	говорить.PRS-3ms.IO

hann	ḥiməš	ḏahb	w	z-ēx	ḳʕā-x

этот.P	пятьдесят	золото.NUM	и	идти.IPV-MS	сидеть.IPV-MS
b=ḳahwe	haччˇ	w	ana	n-zi-ll	b=ḏokk-ṯ=ax
в=кофе	ты.M	и	я	1-идти.SBJV-S	в=место-F=2ms

b=lēly-a
в=ночь-FREE

‘Пришёл сын минситра и сказал ему: «Здесь пятьдесят золотых монет, и теперь иди и сиди в кофейне, а я **пойду** вместо тебя ночью!»’

Также встретился субъюнктив в значении завершённого действия в прошлом. Такое употребление может трактоваться как то, что формы субъюнктива здесь используются, поскольку их требует конкретный союз, который вводит придаточное:

(28) *iḳḏum ma **yḥasslun** mnə-ṣlōṯa p-ḳalles, mfarrḳin šamʕa ʕal_ommṯa, w manəhrišš šamʕa* [...]

iḳḏum.ma	y-ḥassl-un	mn=ṣlō-ṯ-a	b=ḳalles
после.того	3-быть.готовым-MP	из-молитва-F-FREE	в=немного

mfarrḳ-in	šamʕ-a	ʕal=ommṯ-a	w
раздавать.PRS-MP	свеча-FREE	к=люди-FREE	и

manəhril-l	šamʕ-a
зажигать.PRS-DOM	свеча-FREE

‘Перед тем, как они **закончили** с молитвой, они раздают свечи людям и они зажигают свечи [...]’

(29) *orḥa wōb marreḳ mṭānyus taʕsēn, hanna ḳašīša imōḏ, yōməl wōb, iḳḏum mil **yiṯḳan** ḳašīša.*
Однажды проходил мимо Mṭānyus Taʕsēn, который сейчас священник, это было в день, когда он ещё не **стал** священником.

(30) *bess **yḥasslun** mḳallyillun w ḳaʕyillun ōxlin.*
‘Когда вы будете готовы, испеките их и садитесь и ешьте.’

(31) [...] w hōḏ ġrōrča ʕamġōrsa, ḥetta **čiġrus** mutta, iṯər mutti, ṯlōṯa mutti, ti hinnun.

‘[...] и эта ручная мельница мелет, пока она не **намелет** один мутт, два мутта, три мутта, (сколько там) их.’ (мутт — мера объёма).

(32) *bess **čislaḳ** šimša mʕōwet ṭayra markeš* [...]

‘Как только солнце **взошло**, хищная птица проснулась Sobald die Sonne aufging, wachte der Raubvogel wieder auf [...]’

(33) *bess **yiṯḳan** xann čiʕčam tunya ḳalles, mabətya ḥafəlṯa, mabətya ḥafəlṯa bīma? — p-tabəkṯa.*

‘Как только **стало** чуть темнее, начинается праздник, и с чего он начинается? — с хоровода.’

(34) *l-muhimm ṯiḳniṯ nimsayyarle p-ḥakya ana, ḥetta **ymaḏḏell** waḳča ḥetta **yiṯyillun** ommṯa m-ġuppaʕōḏ.*

‘В любом случае я начал подбадривать его речами, чтобы **скоротать** время, пока не **придут** люди из Джуббадина.’

Версию про то, что субъюнктив задаётся конкретным союзом, подтверждает следующий пример, где после предлога *bess* также идёт форма субъюнктива:

(35) *ōmar: «lā sulḳōn, sulḳōn! — aḥsan! bess **yiṯḳan** oḳra aḥsan mallxa.»*

‘Он сказал: «Нет, садитесь, садитесь! — (Это) лучше! Когда (у автобуса) **есть** (большой) вес, он едет лучше.»’

Устойчивые выражения (пожелания) с субъюнктивом:

(36) *«alō **yašəḥnenxun** b-raḥəmṯe!» ṭaleblaḥ* [...]

‘«Пусть Господь **согреет** вас Своей милостью!» он пожелал нам [...]’

(37) *amerle hanna ōġa: «lōzim črōžaʕ leʕle hōš, čaḥəmtenne w čmalle “**ysallmell** ḏwōṯax”, w illa ḏarrarlax.»*

‘Ага сказал ему: «Ты должен вернуться к нему немедленно, чтобы поблагодарить его и сказать “спасибо” (букв. “да **сохранит** (Господь) твои (m.) руки”), чтобы он не причинил тебе вреда.’

(38) *amrōla: «alō **yaffinniš**, ana ġarīpča w čūl ḏokkṯa niḏmux.*

amrō-la	alō	y-aff-inn-iš		ana	ġarīp-ča
говорить.PRS-3fs.IO	Бог	3-воздать.SBJV-PLEO-2fs		я	чужой-FS.DEF

w	čūl	ḏokk-ṯ-a	ni-ḏmux
и	NEG	место-F-FREE	1-спать.SBJV

‘Она сказала ей: «Да **воздаст** тебе Бог, я чужая здесь и мне негде переночевать (букв. “у меня нет места, чтобы переночевать”).’

Также встретился один пример субъюнктива в выражении желания, просьбы:

(39) *ḳʕōle mallex roḥla, mōmar: «yā alō taxīlax **čayṯ** ḥmōrč ya ʕīla ya ʕīlča!»*

ḳʕō-le	mallex	roḥl-a	mōmar	
сидеть.PST-PAST	идти.PRS	сзади-3fs	говорить.PRS	

yā	alō	taxīl=ax	č-ayṯ	ḥmōr-č	ya
VOC	Бог	пожалуйста=2ms	2-родить.SBJV	осёл.1S-F	или

ʕīl-a	ya	ʕīl-č-a	
молодой.осёл-FREE	или	молодой.осёл-F-FREE	

‘Он подошёл к ней сзади сказал: «О Господи, ппрошу тебя, пусть моя ослица **родит** молодого ослика или молодую ослицу!»’

Итого, можно сделать вывод, что формы субъюнктива может требовать не только вспомогательный глагол *batte*, но и некоторые другие вспомогательные глаголы, а также некоторые подчинительные союзы, даже если они не вводят ирреальной ситуации (эта гипотеза требует дальнейшего подтверждения на остальном материале корпуса). Без каких бы то ни было подобных слов субъюнктив может употребляться при выражении пожеланий или желаний. Также встретился один пример с употреблением субъюнктива в значении будущего времени, но без вспомогательного глагола *batte*. Этот пример трубет дальнейшего анализа.

7. Дальнейшие направления исследований

Дальнейшую работу можно проводить по двум направлениям: внутри получившегося корпуса и за его пределами. В первом случае в работу входит

совершенствование самого корпуса — очистка от омонимии, пополнение новыми текстами, разработка гессера для лексем, которых нет в словаре, исправление ошибок. Во втором случае — это создание парсеров и корпусов для диалектов Бахи и Джуббадина, которые смогли бы позволить проводить сравнительные исследования внутри СЗА.

8. Заключение

В рамках этой работы мы создали морфологический парсер с помощью UniParser для диалекта Маалулы современного западного арамейского языка, разметили с помощью него тексты из сборников (Anrold 1991a, Arnold 1991b) и учебника (Arnold 2006). На основе созданного парсера был сделан инструмент для глоссирования, которому на вход подаётся текст на диалекте Маалулы и который возвращает его в отглоссированном виде. Затем на платформе *tsakorpus* нами был создан параллельный арамейско-немецкий корпус диалекта Маалулы с этими текстами, каждое предложение сопровождалось соответствующим аудиофайлом. Также были созданы инструменты для оценки точности работы морфологического парсера (замерялась точность определения лемм, частей речи и грамматических признаков слова) и инструмента для глоссирования (замерялась точность морфологической аннотации и сегментации). Точность парсера по всем трём параметрам оказалась около 98% на всех распознанных токенах и на распознанных уникальных словоформах и около 89% на всех токенах и около 84% на всех уникальных словоформах. Точность инструмента для глоссирования составила около 90% на распознанных токенах и около 73,7% на всех токенах в случае морфологической сегментации, и около 77% и 62,5% в случае морфологической аннотации соответственно. Процент распознанных токенов в корпусе оказался около 87%.

Помимо этого, были проведены исследования омонимии в СЗА и использования форм субъюнктива. Первое исследование показало, что в языке обширно представлена омонимия среди глагольных форм, как с объектными суффиксами (совпадение суффиксов прямых и непрямых объектах в презенсе и перфекте), так и без них (в случаях 3 лица претерита и форм перфекта I и II породы). Также была замечена омонимия между разными частями речи, такими как глагол и существительное, частица и союз, прилагательное и глагол. Во втором исследовании были рассмотрены контексты употребления субъюнктива без вспомогательного глагола *batte*. Среди них

оказались контексты с другими вспомогательными глаголами и псевдоглаголами, которые вводят нужную модальность, а также с некоторыми союзами, которые, судя по всему, требуют после себя только формы субъюнктива, поскольку ситуации, описываемые придаточными с ними, вполне реальны. Также субъюнктив может употребляться отдельно при выражении желаний и пожеланий, особенно в устойчивых выражениях.

Современный западный арамейский — единственный ныне живущий представитель западной ветви новоарамейских языков. Он находится он под угрозой исчезновения и представляет собой интерес для многих лингвистов. Несмотря на активность изучения, для него раньше не существовало никаких NLP-инструментов, что существенно затрудняло его изучение.

Благодарности

Автор выражает благодарность своему научному консультанту Олегу Алексеевичу Серикову, а также Тимофею Александровичу Архангельскому, Антону Бузанову и участникам НУГ «Грамматика современных арамейских языков», в особенности Анне Бромирской, Николаю Гришину, Ксении Кашинцевой и Юлии Маккавеевой за консультации по многочисленным вопросам, помощь в реализации корпуса и разметке эталонных текстов и за поддержку.

Список глосс

1 — первое лицо

2 — второе лицо

3 — третье лицо

c — общий род

AUX — вспомогательный глагол или псевдоглагол

DO — прямой объект

f (F) — женский род

FREE — связанная форма

HD — аффикс сопряжённой формы

INDF — неопределённая форма

IO — непрямой объект

m — мужской род

NEG — отрицание

p (P) — множественное число

POSS — посессивный суффикс

PST — претерит

PRF — перфект

PRS — презенс

s — единственное число

SBJV — субъюнктив

VOC — вокативная частица

Литература

Arkhangelskiy et al. 2012 — Arkhangeskiy, T., Belyaev, O., and Vydrin, A. (2012). *The creation of large-scale annotated corpora of minority languages using uniparser and the eanc platform.* pp. 83–92.

Arnold 1989 — Arnold, Werner. 1989. *Das Neuwestaramäische. I. Texte aus Baxʿa.* (Semitica Viva; Bd. 4/I), Wiesbaden 1989.

Arnold 1990a — Arnold, Werner. 1990. *Das Neuwestaramäische V: Grammatik.* (Semitica Viva, 4.) Wiesbaden: Harrassowitz. xxi+410pp.

Arnold 1990b — Arnold, Werner. 1990. *Das Neuwestaramäische. II. Texte aus Jubbʿadīn.* (Semitica Viva; Bd. 4/II), Wiesbaden 1990.

Arnold 1991a — Arnold, Werner. 1991. *Das Neuwestaramäische. III. Volkskundliche Texte aus Maʿlūla.* (Semitica Viva; Bd. 4/III), Wiesbaden 1991.

Arnold 1991b — Arnold, Werner. 1991. *Das Neuwestaramäische. IV. Orale Literatur aus Maʿlūla.* (Semitica Viva; Bd. 4/IV), Wiesbaden 1991.

Arnold 2002 — Arnold, Werner. 2002. *Neue Lieder aus Maʿlūla*. In: ”Sprich doch mit deinen Knechten aramäisch, wir verstehen es!” 60 Beiträge zur Semitistik. Festschrift für Otto Jastrow zum 60. Geburtstag. Wiesbaden 2002

Arnold 2006 — Arnold, Werner. 2006. *Lehrbuch des Neuwestaramäischen.* (Semitica viva: Series didactica, 1.) 2nd edn. Wiesbaden: Harrassowitz. 152pp.

Arnold 2011 — Arnold, Werner. 2012. *Western Neo-Aramaic.* in Weninger, S. (2012). *The Semitic Languages: An International Handbook.* Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110251586

Arnold 2019 — Arnold, Werner. 2019. *Das Neuwestaramäische. VI. Wörterbuch. Neuwestaramäisch – Deutsch* (Semitica Viva; Bd. 4/VI). Wiesbaden: Harrassowitz. xix+1018pp. (German and Aramaic).

Bergsträsser — Bergsträsser, Gotthelf (Hg.): *Neuaramäische Märchen und andere Texte aus Malula.* Leipzig: F.A. Brockhaus, 1915, S. 106-107.

Correll 1978 — Correll, Christoph. 1978. *Untersuchungen zur Syntax der neuwestaramäischen Dialekte des Antilibanon: Ma'lūla, Baḫ'a, Ǧubb'Adīn*. Mainz: Deutsche Morgenländische Gesellschaft. xx+220pp.

Duntsov, Alexey & Haberl, Charles & Loesov, Sergey. (2022). *A Modern Western Aramaic Account of the Syrian Civil War.* Word. 68. 359-394. 10.1080/00437956.2022.2084663.

Eid 2024 — Eid, Ghattas. (2024). *The Phonology of Maaloula Aramaic.* Düsseldorf University Press.

Fassberg 2019 — Fassberg, Steven Ellis. 2019. *Modern Western Aramaic.* In John Huehnergard and Na‘ama Pat-El (eds.), The Semitic Languages, 632–652. 2nd edn. London & New York: Routledge.

Ghattas et al. 2022 — Ghattas Eid, Esther Seyffarth, and Ingo Plag. 2022. *The Maaloula Aramaic Speech Corpus (MASC): From Printed Material to a Lemmatized and*

Time-Aligned Corpus. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 6513–6520, Marseille, France. European Language Resources Association.

Grishin, Bromirskaya 2023 — Grishin Nikolay, Bromirskaya Anna. *Verbless and existential clauses in Modern Western Aramaic.* 5th Neo-Aramaic languages conference, 26-27 October 2023.

Hwa et al. 2002 — Hwa, R., Resnik, P., Weinberg, A., and Kolak, O. (2002). *Evaluating translational correspondence using annotation projection.*

Jastrow 1997 — Jastrow, Otto. 1997. *The Neo-Aramaic Languages.* In: Robert Hetzron (ed.), The Semitic Languages, 334-377. London and New York: London & New York: Routledge.

Kashintseva 2023 — Kashintseva Kseniya. *Adnominal possession marking in Ṭuroyo language.* 5th Neo-Aramaic languages conference, 26-27 October 2023.

Kilgarriff et al. 2014 — Kilgarriff, Adam, et al. *The Sketch Engine: Ten Years on Lexicography*. 2014, 1.1: 7-36.

Klyachko et al. 2020 — Klyachko, E., Sorokin, A., Krizhanovsky, N., Krizhanovsky, A., and Ryazanskaya,G. (2020). *LowResourceEval-2019: a shared task on morphological analysis for low-resource languages.*

Korobov 2015 — Korobov, M. (2015). *Morphological Analyzer and Generator for Russian and Ukrainian Languages.* In: Khachay, M., Konstantinova, N., Panchenko, A., Ignatov, D., Labunets, V. (eds) Analysis of Images, Social Networks and Texts. AIST 2015. Communications in Computer and Information Science, vol 542. Springer, Cham. https://doi.org/10.1007/978-3-319-26123-2_31

Krause and Zeldes 2016 — Krause, Thomas & Zeldes, Amir (2016). *ANNIS3: A new architecture for generic corpus query and visualization.* In: Digital Scholarship in the Humanities 2016 (31). http://dsh.oxfordjournals.org/content/31/1/118

Lindén et al. 2009 — Lindén, K., Silfverberg, M., and Pirinen, T. (2009). *HFST tools for morphology — an efficient open-source package for construction of morphological analyzers.* volume 41, pp. 28–47.

Qi et al. 2020 — Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages.* In Association for Computational Linguistics (ACL) System Demonstrations. 2020.

Rychlý 2007 — Rychlý, Pavel. *Manatee/Bonito-A Modular Corpus Manager.* In: RASLAN. 2007. p. 65–70.

Schmid 2005 — Schmid, H. (2005). *A programming language for finite state transducers.* pp. 308–309.

Sorokin 2019 — Sorokin, Alexey. 2019. *Morphological parsing of low-resource languages.* РГГУ М. in Компьютерная лингвистика и интеллектуальные Технологии. По материалам ежегодной международной конференции Диалог-2019. pp. 636–647.

Spitaler 1938 — Spitaler, Anton. 1938. *Grammatik des neuaramäischen Dialekts von Ma'lula (Antilibanon).* (Abhandlungen für die Kunde des Morgenlandes, 23.1.) Leipzig. xxvi+225pp.

Swanson and Howell 2021 — Swanson, D. and Howell, N. (2021). *Lexd: A Finite-State Lexicon Compiler for Non-Suffixational Morphologies.*

Коган, Лёзов 2009 — Коган Л. Е., Лёзов С. В., *Маалулы язык.* В кн.: Языки мира: Семитские языки. Аккадский язык. Северозападносемитские языки.

Электронные ресурсы:

Arkhangelskiy 2017 — Archangelskiy, Timofey. 2017.

Tsakorpus 2.0.

(Available online at: https://github.com/timarkh/tsakorpus.)

Arkhangelskiy 2018 — Arkhangelskiy, Timofey. (2018).

Moksha morphological analyzer.

(Available online at: https://github.com/timarkh/uniparser-grammar-moksha.)

Arkhangelskiy et al. 2018 — Arkhangelskiy, T., Barsky, E., Furman, Y., and Kashintseva, K. (2018).

Turoyo morphological analyzer.

(Available at: https://github.com/margisk/uniparser-grammar-turoyo.)

Arkhangelskiy 2019 — Arkhangelskiy, Timofey. (2019).

Urmi morphological analyzer.

(Available at: https://github.com/timarkh/uniparser-grammar-urmi.)

Arkhangelskiy and Morozova 2019 — Arkhangelskiy, T. and Morozova, M. (2019).

Albanian morphological analyzer.

(Available at: https://github.com/timarkh/uniparser-grammar-albanian.)

Arkhangelskiy 2021 — Arkhangelskiy, T. (2021). Buryat morphological analyzer.

(Available at: https://github.com/timarkh/uniparser-grammar-buryat.)

CW — The IMS Open Corpus Workbench. URL: https://cwb.sourceforge.io/

Hammarström et al. 2024 — Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian. 2024.

Glottolog 5.0.

Leipzig: Max Planck Institute for Evolutionary Anthropology.

https://doi.org/10.5281/zenodo.10804357

(Available online at http://glottolog.org, Accessed on 2024-05-10.)

MAC — Moscow Aramaic Circle. Available at: https://moscow-aramaic-circle.netlify.app/

Matter 2022 — Matter, F. (2022). *A morphological parser for Yawarana.*

(Available at: https://github.com/fmatter/uniparser-yawarana.)

Электронные корпуса:

Arkhangelskiy et al. — Timofey Arkhangelskiy, Irina Bagirokova, Yury Lander, Anna Sorokina.

West Circassian (Adyghe) Corpus.

(Available online at: adyghe.web-corpora.net).

Barsky — Barskiy, E. *Corpora of Maalula, Syriac, Turoyo and Christian Urmi.*

Available at: https://evb0110.github.io/aramaicsite/hello

Lander, Arkhangelskiy 2018 — Lander, Yury, Arkhangelskiy, Timofey. (2018).

Adyghe morphological analyzer.

(Available online at: https://github.com/timarkh/uniparser-grammar-adyghe.)

Lyavdanskiy et al. a — Alexey Lyavdansky et al. *Turoyo Corpus.* (Available online at: neo-aramaic.web-corpora.net/turoyo_corpus/search).

Lyavdanskiy et al. b — Alexey Lyavdansky et al. *Christian Urmi Corpus.* (Available online at: neo-aramaic.web-corpora.net/urmi_corpus/search).

Ovsyannikova et al. 2017 — Ovsynnikova, M., Say, S., Aplonova, E., Smetina, A., and Sokur, E. (2017). *Ustnyy corpus bashkirskogo yazyka der. Rakhmetova i s. Baimovo.* (Available at: http://lingconlab.ru/spoken_bashkir/.)

Stenin, Garanina 2018 — Stenin, I. and Garanina, E. (2018).

Chukchi multimedia corpus.

(Available at: https://chuklang.ru/corpus.)

Казакевич и др. — О. А. Казакевич, Е. Л. Клячко, Н. К. Митрофанова. *Корпус эвенкийского языка* (https://minlang.iling-ran.ru/corpora/evenki). Лицензия: https://creativecommons.org/licenses/by-sa/4.0/