ЯЗЫКОЗНАНИЕ

УДК 81.26

Особенности процессинга арабского языка: морфологическое моделирование*

О. А. Берникова, Н. А. Кижаева

Санкт-Петербургский государственный университет,
Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

Для цитирования: *Берникова О. А., Кижаева Н. А.* Особенности процессинга арабского языка: морфологическое моделирование // Вестник Санкт-Петербургского университета. Востоковедение и африканистика. 2023. Т. 15. Вып. 3. С. 459–484. https://doi.org/10.21638/spbu13.2023.302

В статье рассматриваются особенности морфологического моделирования арабского языка на основе определения специфики его формализации. Морфологическое моделирование является одним из ключевых этапов автоматического анализа текстов и включает в себя инструменты для возведения словоформы к основе, корню, определения части речи, автоматического построения (генерации) заданной словоформы и т. д. Задачи исследования носят междисциплинарный характер и включают в себя как теоретические аспекты изучения особенностей арабского языка, которые наиболее актуальны для его автоматической обработки, так и анализ имеющихся морфологических анализаторов и определение специфики их работы. Практическая часть основана на тестировании инструмента CAMeL Tools, одним из преимуществ которого является его комплексный характер, позволяющий осуществлять как препроцессинг текста, так и решение задач прикладного характера, включая сентимент-анализ. Критерии выборки примеров для тестирования учитывали особенности арабского языка, представляющие трудность для его формализации (сегментация служебных слов, имеющих слитное написание; морфологическая и лексическая омонимия и т. д.). Кроме того, принимается во внимание вариативность обобщенного понятия «арабский язык», которое объединяет классический арабский язык, современный арабский литературный язык и современные арабские диалекты. Тестирование инструментов для морфологического моделирования позволяет сделать выводы о необходимости совершенствования терминологического аппарата, вариативность которого отмечена при описании словоформ. Такого рода варьирование (расхождение с понятиями, принятыми в общем языкознании) потенциально ведет к искажению результатов лексико-семантического разбора.

* Исследование выполнено за счет гранта Российского научного фонда № 22-28-01046, https://rscf.ru/project/22-28-01046/.

В ходе анализа отмечены некоторые лакуны, связанные с определением частеречной принадлежности, описанием словоформ и т. д. Результаты исследования актуальны как для проведения лингвистических исследований, так и для совершенствования разработки программных приложений, направленных на процессинг арабского текста.

Ключевые слова: арабский язык, морфологическое моделирование, анализатор, процессинг.

Введение

Обработка естественного языка (Natural Language Processing, NLP) является сегодня одним из важнейших направлений искусственного интеллекта, необходимость развития которого в Российской Федерации определяется задачами «Национальной стратегии развития искусственного интеллекта на период до 2030 года» [1]. Несмотря на очевидную эффективность и актуальность применения автоматических методов обработки данных при исследовании арабского языка, они все еще недостаточно распространены в арабистике и исламоведении в силу преимущественного использования классических методов гуманитарных наук. Вместе с тем рост интернет-контента на арабском языке, все увеличивающееся количество программных приложений, осуществляющих процессинг арабского языка, предполагают необходимость тестирования корректности их функционирования с учетом особенностей формализации арабского языка.

Автоматическая обработка текстов позволяет изучать объем материала, анализ которого вручную невозможен, а также получать статистическую информацию по всему корпусу текстов. Одним из ключевых этапов всех алгоритмов автоматического анализа текстов является предобработка документа, состоящая из токенизации, фильтрации, лемматизации и/или стемминга. Токенизация — разбиение последовательности символов на части (фразы, слова, слоги), называемые токенами, — может включать в себя удаление определенных символов, например знаков пунктуации, гиперссылок, номеров страниц. Фильтрация заключается в удалении некоторых слов из текста, наиболее распространенным ее видом является удаление стоп-слов. Под стоп-словами понимаются такие слова, которые часто встречаются в тексте и не несут содержательной информации (предлоги, союзы и т. п.). Лемматизация включает в себя морфологический анализ слов, при котором различные формы группируются для того, чтобы их можно было обрабатывать как один объект. При лемматизации текстов для каждого слова необходимо определить часть речи, последнее представляет собой достаточно сложную задачу, в связи с чем на практике чаще используются методы стемминга — определения основы слова, которая зачастую не совпадает с его морфологическим корнем.

Таким образом, одним из ключевых этапов автоматического анализа текстов является предобработка документа, существенная часть которой относится к морфологическому моделированию — комплексному решению, включающему в себя инструменты для возведения словоформы к основе, корню, определение части речи (так называемый морфологический анализатор), автоматическое построение (генерацию) заданной словоформы и т. д.

Несмотря на то что разработка первых морфологических анализаторов для арабского языка началась несколько десятилетий назад, до сих пор отмечаются ла-

куны в функционировании такого рода решений. Интересен тот факт, что в основе многих инструментов для обработки арабского текста лежит морфологический анализатор Тима Бакволтера (Buckwalter Arabic morphological analyzer version 1.0), разработанный два десятилетия назад [2]. С тех пор было издано несколько версий анализатора, и этот инструмент широко используется как в корпусе Корана (The Quranic Arabic Corpus) [I], так и в корпусе арабского языка (ArabiCorpus) [II]. В настоящий момент для арабского языка существует несколько морфологических анализаторов в открытом доступе (пакет библиотек Natural Language Toolkit [III], Qutuf [IV] и др.).

Несмотря на распространение в последнее время технологий, позволяющих решать многие задачи без обращения к методам морфологического моделирования, очевидно, что последнее не теряет своей важности в силу эффективности его использования в качестве метода препроцессинга текста, а также важного ресурса для проведения лингвистических исследований. Важность морфологического моделирования особенно очевидна для языков, тяготеющих к синтетическому строю речи, предполагающих наличие разветвленной парадигмы словоизменительных и словообразовательных моделей. Арабский язык представляет собой один из ярких примеров такого рода языков, одной из отличительных особенностей которого является алгебраичность его морфологической структуры.

По последним оценкам, численность носителей арабского языка превышает 450 млн человек [3, p. 11]. За последние годы доля сегмента интернета на арабском языке значительно увеличилась, сегодня он занимает четвертое место в мире по степени распространенности в сети Интернет. Вместе с тем в докладе о статусе арабского языка, обнародованном Министерством культуры и молодежи ОАЭ в 2021 г. [3], отмечен широкий круг актуальных задач, касающихся функционирования арабского языка в цифровой сфере, которые необходимо решить в ближайшее время. Среди них отмечаются как расширение цифрового контента на арабском языке и оцифровка исторических документов, так и совершенствование инструментария для обработки арабского текста (классического и современного). Особое внимание уделяется арабскому языку как инструменту электронной коммуникации, во многом восходящему к арабским диалектам, процессингу которых обычно уделяется меньше внимания, чем обработке современного арабского языка.

Учитывая сказанное выше, значимость процессинга арабского текста в целом и морфологического моделирования для арабского языка в частности становится очевидной. Целью данного исследования является анализ особенностей морфологического моделирования арабского языка на основе изучения специфики его формализации и имеющихся программных решений для его процессинга. Задачи исследования носят междисциплинарный характер и опираются как на методы лингвистического анализа и корпусных исследований, так и на автоматизированный подсчет численных характеристик текста. Практическая часть работы основана на тестировании инструмента CAMeL Tools [V; 4], который в силу комплекса объективных факторов и своего функционала представляет собой удобное решение для процессинга арабского текста. Одним из преимуществ CAMeL Tools является его комплексный характер, позволяющий осуществлять как препроцессинг текста, так и решение задач прикладного характера, включая сентимент-анализ.

Особенности арабского языка в контексте задач морфологического моделирования

При разработке решений для морфологического моделирования арабского текста необходимо учитывать особенности арабского языка, который на системном и структурном уровнях значительно отличается от западных языков, на которые обычно первоначально настроены инструменты для обработки естественных языков. Рассмотрим примеры основных особенностей арабской морфологии, которые наиболее актуальны при разработке и адаптации морфологических анализаторов и иных инструментов процессинга арабского текста[1].

Как отмечалось выше, арабский язык характеризуется разветвленной системой словообразования и словоизменения. При этом значительную трудность составляет тот факт, что в процессе возведения словоформы к корню нужно учитывать не только префиксы и постфиксы, несущие грамматическое значение, но и внутренние модификации основы слова при изменении породы, залога, времени и иных грамматических категорий. Приведенные в представленной ниже табл. 1 примеры наглядно демонстрируют процесс словоизменения в арабском языке, затрагивающий и основу слова.

Сопоставительный анализ приведенных лексем показывает, что образование различных форм глагола от корня *q-w-l* происходит не только за счет присоединения префиксов и постфиксов, но и посредством изменения второго согласного корня *-w-*, которому в различных словоформах могут соответствовать гласные */ā/*, */ī/*, */ū/* или падение второго корневого. Аналогичные преобразования свойственны и именам, причастиям и масдарам (именам действия). Соответственно, разработка решений для морфологического моделирования должна учитывать не только стандартный перечень словоизменительных морфем, но и изменение основы. Последнее предполагает написание соответствующих правил, сгруппированных по разным типам корней (с первым слабым, со вторым слабым, с *хамзой* и т. д.).

Другой существенной особенностью арабского языка является влияние частеречной принадлежности слова на его словоизменительную парадигму. Для примера образуем форму множественного числа от слова كَاتِبٌ *kātib*. Одна и та же форма в значении существительного означает «писатель», а в значении причастия — «пишущий». При этом образование форм множественного числа зависит от частеречной принадлежности (рис. 1).

Таблица 1. **Особенности словоизменения глагола: корень قول** (*q-w-l*)

3 л., ед. ч., м. р., прошедшее время		**3 л., ед. ч., м. р., настоящее-будущее время**		**2 л., мн. ч., настоящее-будущее время**	
действ. залог	*страд. залог*	*действ. залог*	*страд. залог*	*м. р.*	*ж. р.*
قَالَ	قِيلَ	يَقُولُ	يُقَالُ	يَقُولُونَ	يَقُلْنَ
qāla	qīla	yaqūlu	yuqālu	yaqūlūna	Yaqulna
он сказал	ему сказали	он говорит	ему говорят	они (м. р.) говорят	они (ж. р.) говорят

[1] Задачи данного исследования охватывают современный и классический арабский язык и лишь косвенно касаются арабских диалектов, которые требуют отдельного рассмотрения.

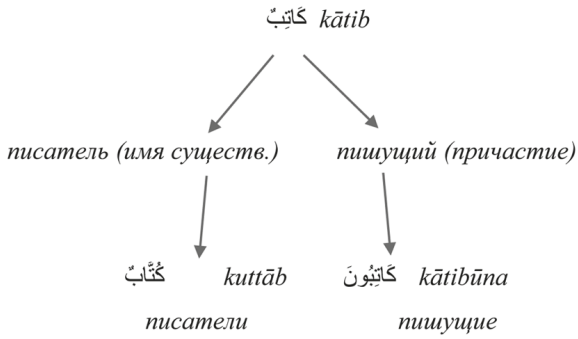

Рис. 1. Зависимость форм множественного числа от частеречной принадлежности

Таблица 2. **Спряжение глагола مَرّ «идти, проходить» в настоящем-будущем времени в единственном числе**

Грамматическая форма	**Написание с *ташдидом***	**Написание без *ташдида***	**Транскрипция**
3 л., ед. ч., м. р.	يمرّ	يمر	yamurru
3 л., ед. ч., ж. р.	تمرّ	تمر	tamurru
2 л., ед. ч., м. р.	تمرّ	تمر	tamurru
2 л., ед. ч., ж. р.	تمرّين	تمرين	tamurrīna
1 л., ед. ч., м. р. / ж. р.	أمرّ	أمر	’amurru

Таким образом, частеречная принадлежность является необходимым маркером, который важно учитывать при создании инструментов морфологического моделирования.

Возведение к корневой основе или образование в арабском языке производных может быть осложнено вариативностью написания знака, обозначающего удвоение согласного (т. е. *ташдида*, ◌ّ). Современные нормы арабского языка допускают наличие или отсутствие соответствующего знака на письме (включая печатные тексты). Однако в ходе автоматической обработки текста наличие или отсутствие *ташдида* может существенно влиять на корректность лемматизации или стемминга. Для решения данной проблемы необходим комплекс алгоритмов, учитывающий потенциальную возможность вариативности при написании одного и того же слова. В большей степени это свойственно удвоенным глаголам и их производным. Для примера рассмотрим спряжение удвоенного глагола в настоящем-будущем времени в единственном числе (табл. 2).

Спряжение глагола مَرَّ *marra*, наличие или отсутствие *ташдида* в слове демонстрируют необходимость учета всех потенциально возможных вариантов написания знаков, букв, слов при обработке арабского текста. Аналогичный подход необходимо применять и при написании *хамзы* с *алифом*, которая зачастую может опускаться или реализовываться с *васлой*. В табл. 3 представлены примеры графической реализации начальной *хамзы* в одном и том же слове. Характер написания может зависеть от контекстуального употребления либо носить индивидуальный характер (т. е. зависеть от подхода автора к набору текста на арабском языке).

Таблица 3. **Вариативность написания *хамзы* в начале слова на примере слова «сын»**

Написание *хамзы* с *алифом*	**Отсутствие знака *хамзы***	**Реализация *васлы* вместо *алифа* с *хамзой***	**Транскрипция**
إبن	ابن	ٱبن	’ibn
إـــ	اـــ	ٱـــ	

Таблица 4. **Варианты произношения омографа** جار

Написание без огласовок	**Написание с огласовками**	**Корень**	**Грамматическая характеристика**	**Перевод**
جار ğ-ā-r	جَارٌ ğārrun	ج ر ر	Причастие, I порода, ед. ч., м. р., И. п.	тянущий, волочащий; *грам.* управляющий родит. падежом
	جَارٍ ğārin	ج ر ي	Причастие, I порода, ед. ч., м. р., действ. залог	бегущий
	جَار ğāri	ج ر ي	Глагол, III порода, повел. накл., 2 л., ед. ч., м. р.	подражай; соревнуйся
	جَارَ ğāra	ج و ر	Глагол, I порода, прош. вр., 3 л., ед. ч., м. р.	отходить, отклоняться
	جَارٌ ğārun	ج و ر	Имя, ед. ч., м. р., И. п.	сосед

Учет всевозможных вариантов графической реализации *хамзы* обеспечивает корректную работу морфологического анализатора. В противном случае значительное количество словоформ могут оказаться нераспознанными. Это же касается и вариативности написания ي и ى. Несмотря на то что языковые нормы предполагают написание буквы ي с точками, в отличие от *алифа-максура* (ى), современные издания допускают написание ي без точек и наоборот. Соответственно, корректный поиск слов и возведение их к основе будут реализованы исключительно в случае учета реалий печатного арабского текста. Обычно при создании морфологических анализаторов создаются функции, направленные на нивелирование проблем с вариативностью написания различных букв.

Особенности современного арабского письма заключаются в том, что в большинстве случаев на письме не отображаются огласовки (знаки, используемые для передачи кратких гласных, их отсутствия и т. д.). Этот факт усложняет задачу возведения к корневой основе или образования производных от искомого слова. В зависимости от распределения гласных в слове меняется его значение и, как следствие, его начальная форма и корень. В табл. 4 рассматривается слово, одна графическая репрезентация которого может соответствовать разным корням и словоформам и, как следствие, иметь разные варианты произношения.

В таблице перечислены лишь некоторые примеры потенциально возможных вариантов прочтения и перевода слова جار, написанного без огласовок. Тем не менее

Таблица 5. **Варианты произношения омографа لكم**

Написание без огласовок	Написание с огласовками	Корень	Грамматическая характеристика	Перевод
لكم l-k-m	لَكُمْ lakum	ل كم	предлог la + слитное местоимение kum	для вас (мн. ч.)
	لَكَمَ lakama	ل ك م	глагол, I порода, прош. время, 3 л., ед. ч., м. р.	колотить

содержание приведенных примеров демонстрирует, что рассматриваемое слово может восходить к нескольким типам корней (удвоенным, пустым, недостаточным), различным частям речи (имени существительному, причастию, глаголу), нескольким породам (I, III). Очевидна и словоизменительная палитра приведенных форм: от действительного залога у причастий до глаголов в прошедшем времени и повелительном наклонении.

Многозначность слов, реализованных без диакритики, обусловлена и наличием в арабском языке слитных предлогов и частиц, а также слитных местоимений. Пример представлен в табл. 5.

Таким образом, неогласованный арабский текст характеризуется преобладанием в нем омографов, т. е. слов, имеющих одинаковое написание, но разное прочтение. Произношение данных слов предполагает выбор из множества вариантов с учетом контекста, морфологических характеристик и иных лингвистических факторов. При процессинге текста данная особенность арабского языка представляет наибольшую сложность.

Трудности сегментации арабского текста проявляются в большом количестве слитных частиц, предлогов, слитных местоимений. Приведем пример разбора фразы, содержащей союз, частицу, глагол и слитное местоимение.

فَلْيَشْتَرُوهُ

ف–ل–يشترو–هُ

и — пусть — они покупают — его

союз — частица — глагол — слитное местоимение

В сегменте данной фразы يشترو присутствует маркер VIII породы (инфикс ت), а также дификс و---ي, характеризующий форму условного наклонения, 3 л. мн. ч. м. р. Таким образом, лишь ش и ر относятся к корню. Учитывая, что большинство корней в арабском языке имеют трехсогласную основу, указанные буквы в нужной последовательности могут соответствовать согласным, входящим в состав корней с первым (وشر) или третьим (شري) слабым, а также характеризовать удвоенные основы (شرّ). Дальнейший анализ позволяет методом исключения определить, что в данном случае использован корень с третьим слабым (شري), так как словарный состав арабского языка не имеет глаголов в VIII породе, образованных от двух других приведенных выше корней. Безусловно, столь сложный процесс возведения к корневой основе значительно упрощается благодаря работе морфологических анализаторов, но их создание должно учитывать все нюансы словоизменительной и словообразовательной парадигмы арабского языка.

В этой связи скрупулезный анализ морфологических особенностей арабского языка, существенно влияющих на выбор методики для его автоматической обработки, требует гармоничного сочетания лингвистических и информационных компетенций разработчиков. Кроме того, необходимо учитывать и то, что знание языка не всегда означает умение осуществлять морфологический разбор того или иного слова и/или формализовать его «глазами компьютера». Корректность работы решений для морфологического моделирования арабского языка во многом зависит от соблюдения «строгости» при проведении междисциплинарных исследований, предполагающих равноценный учет специфики различных научных областей [5]. Приведенные примеры (перечень которых может быть существенно расширен) наглядно демонстрируют невозможность прямого использования для арабского языка решений, разработанных для западных языков, без существенной их адаптации к требованиям арабского языка.

Морфологическое моделирование арабского языка: обзор имеющихся решений

В данном разделе приведен обзор первых морфологических анализаторов для арабского языка. Алгоритмы, лежащие в их основе, оказали существенное влияние на развитие подходов к морфологическому моделированию. Современные системы автоматической обработки арабского языка в своем морфологическом компоненте в той или иной мере реализуют идеи, заложенные в этих работах.

В 1989 г. К. Бизли, Т. Бакволтер и С. Ньютон представили компьютерную программу, которая осуществляла морфологический разбор и поиск в словаре слов на арабском языке [6]. Работы велись в рамках проекта ALPNET, а затем позже были приобретены компанией Xerox, выпустившей коммерческий продукт. В 1998 г. К. Бизли переписал правила для конечного автомата в системе Xerox [7]. Это был первый морфологический анализатор, построенный с помощью конечных автоматов (англ. finite-state automata), реализующих так называемую двухуровневую морфологию. Конечный автомат — это абстрактная математическая модель дискретного устройства, которая имеет конечное множество возможных состояний. В каждый момент времени устройство может находиться только в одном из возможных состояний. Среди состояний выделяют одно входное состояние (вход) и одно или несколько выходных состояний (выходы). Теория конечных автоматов находит широкое применение в синтаксических и лексических анализаторах при компиляции и интерпретации языков программирования.

Идея двухуровневой морфологии была впервые представлена К. Коскенниеми в 1983 г. в работе по созданию компьютерной лингвистической модели для финского языка [8]. Однако алгоритм обрабатывал лишь последовательное соединение морфем в слове. Специфика арабского языка потребовала доработки этой модели. В частности, формирование основы из корня и паттерна было реализовано с помощью пересечения двух лексиконов — операции, поддерживаемой в конечных преобразователях (англ. finite-state transducers). Система Xerox состоит из четырех преобразователей, которые компилируются параллельно и генерируют все возможные слова в языке. На рис. 2 представлена схема анализатора на примере разбора слова *wakafy*. Верхний блок — фильтрующий автомат, который удаляет не-

правильно сформированные строки. Ниже расположен основной лексикон, состоящий из 4930 корней и 400 паттернов, трех кратких гласных, суффиксов и префиксов. Правила пересечения обеспечивают перебор комбинаций корней и паттернов, формирующих основы. Основа в системе Xerox — это комбинация трех компонентов основного лексикона: корня, гласных и паттерна. Рис. 2 демонстрирует схему работы анализатора [9, p. 44].

Для каждого из 4930 корней вручную отобраны паттерны, с которыми он совместим. Система генерирует более 90 000 основ, но если добавить к ним все возможные комбинации префиксов и суффиксов, то на выходе получается 72 000 000 «слов». Такая избыточная генерация слов демонстрирует вполне ожидаемое поведение, так как перебираются все комбинации морфем (в данном случае под морфемами понимаются компоненты основного лексикона, включая служебные части речи, которые пишутся слитно). Следовательно, необходим механизм фильтрации, чтобы удалять недопустимые в языке сочетания морфем.

Именно фильтрующий автомат обрабатывает зависимости между несколькими морфемами в одном слове. Для его работы необходимо правило, по которому можно определить, какие префиксы[2] разрешено объединять с суффиксами. Например:

لِمُعَلِّمِينَ
li-muʿallim-īna
у (для) преподавателей

После предлога имя всегда стоит в родительном падеже, что у форм множественного числа мужского рода может маркироваться окончанием *-īna*, что определяет зависимость в одной словоформе окончания от предлога. Знаки + и обозначают границы морфем и слова соответственно. Они используются во внутреннем представлении слова и реализуются как пустой символ на выходе. За отображение внутреннего представления в выходную строку отвечают правила вариации. В системе Xerox насчитывается 66 правил вариации, позволяющих обрабатывать особые случаи арабской фонологии и орфографии.

Алгоритм Buckwalter Arabic Morphological Analyser (BAMA) [10] имеет много общего с версией Xerox: как, например, сегментирование самого слова на префикс, основу и суффикс, так и разделение лексикона на префиксы, основы и суффиксы с целью установления зависимости между разными грамматическими категориями.

BAMA — один из самых распространенных морфологических анализаторов для научно-исследовательских задач. Так, университет Пенсильвании использовал его для проекта Arabic TreeBank [11; 12] по аннотированию большого набора новостных статей и созданию размеченного корпуса TREC-11 Arabic corpus, используемого в задачах информационного поиска.

В системе BAMA три лексикона: 421 префикса, 126 471 основа и 1170 суффиксов, а также три таблицы совместимости: префикс-основа (таблица AB), основа-суффикс (таблица BC), префикс-суффикс (таблица AC). В таблицах AB, BC и AC

[2] Значение понятий «суффикс» и «префикс» в лингвистике отличается от их применения при описании работы морфологических анализаторов. В последнем случае «префикс» означает не только морфему, но и любое слово (как правило, служебное), которое пишется слитно с последующим словом. «Суффикс» может охватывать не только морфему, но и слитное местоимение.

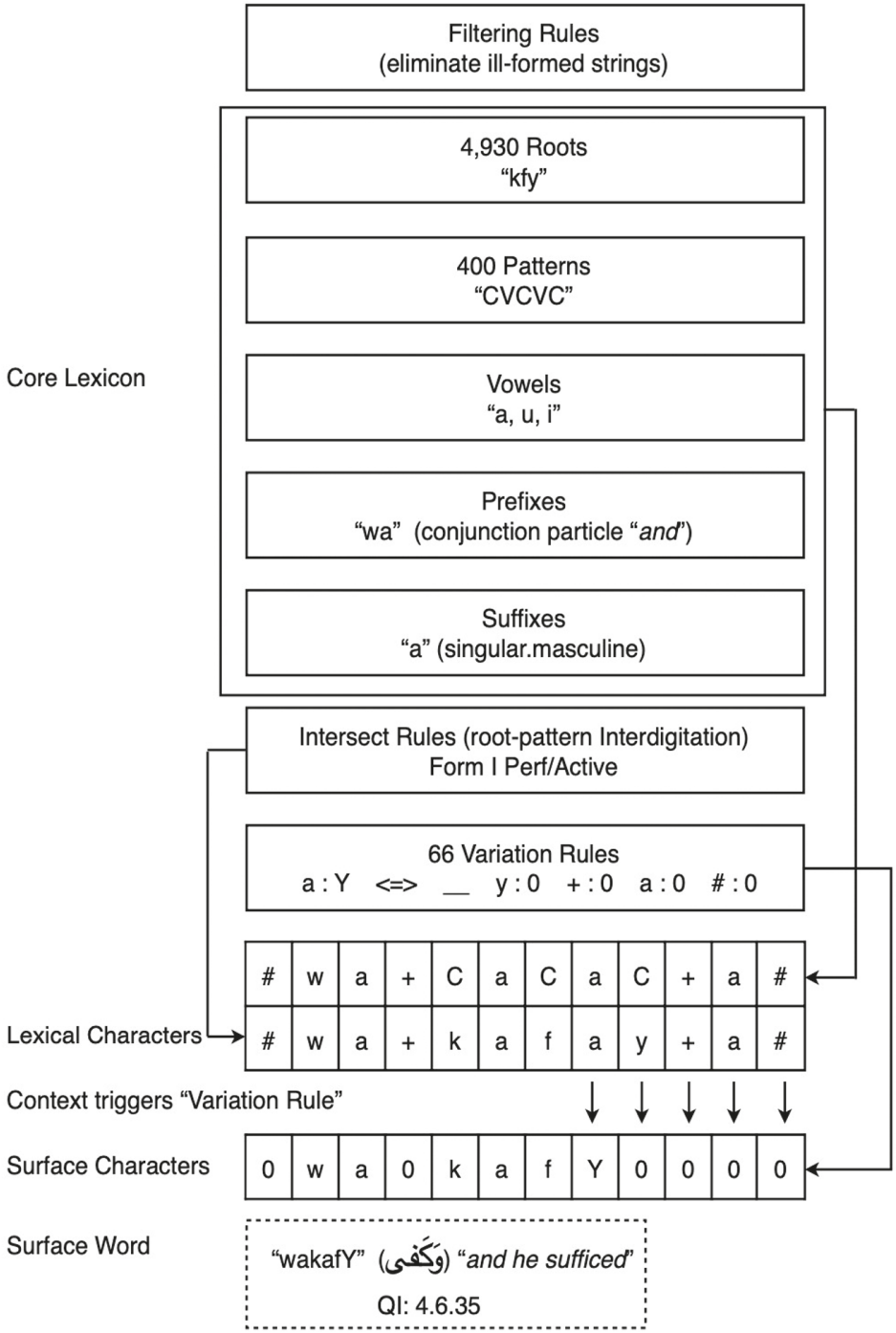

Рис. 2. Схема работы анализатора Xerox [9, p. 44]

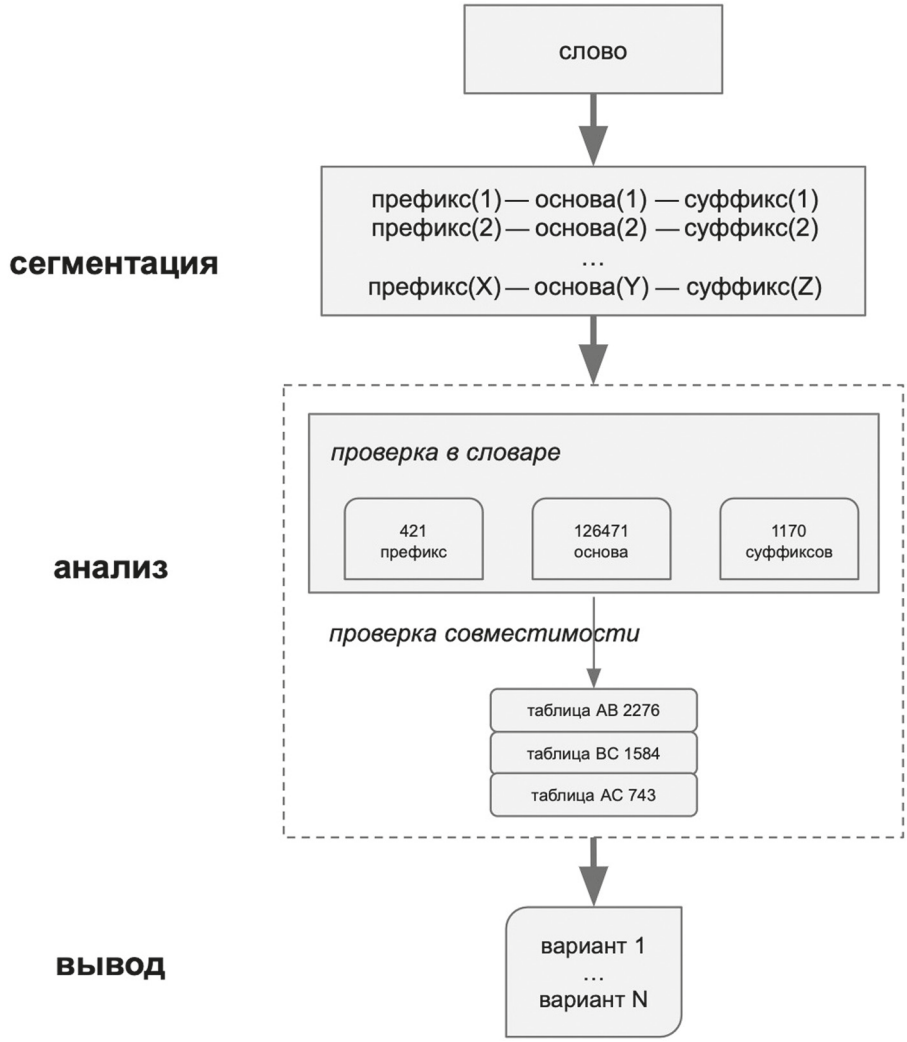

Рис. 3. Основные этапы Buckwalter Arabic Morphological Analyzer

содержится 2276, 1584 и 743 элемента соответственно. Рис. 3 отражает последовательность действия алгоритма.

Система считывает поданное на вход слово и проходит три этапа: разделение слова на сегменты, проверку полученных сегментов в словаре, проверку совместимости сегментов. На выходе — возможные варианты разбора.

Сегментация. Слово в лексиконе BAMA состоит из префикса длиной от 0 до 4 символов, основы длиной от 1 до максимального количества символов во входном слове и суффикса длиной от 0 до 6 символов. Входное слово разделяется на все возможные комбинации «префикс — основа — суффикс» при ограничении длины префикса ≤ 4, длины основы ≥ 1, длины суффикса ≤ 6. BAMA начинает с префикса длиной 0, затем пытается найти все возможные комбинации основ и суффиксов. Для каждого префикса комбинации основ и суффиксов включают в себя все осно-

вы, чья длина ≥ 1, и всего 7 возможных суффиксов. Таким образом, максимальное количество сегментов равно 35 (произведение 5 (допустимые префиксы) и 7 (допустимые суффиксы)). Алгоритм сегментации никак не учитывает допустимость полученных сегментов в языке, такая проверка происходит внутри функции «анализ» на следующем этапе.

Анализ. Для каждого сегмента, полученного на этапе сегментации, проводится сверка с таблицей совместимости комбинаций «префикс — основа — суффикс». Если сегменты существуют в лексиконе, то BAMA переходит к следующему этапу анализа. Когда определен шаблон «префикс — основа — суффикс», система проходит по таблицам совместимости и проверяет сначала совместимость префикса с основой и, если такая комбинация возможна, совместимость основы с суффиксом и префикса с суффиксом. Если вся комбинация допустима, она попадает в список вариантов разбора.

Алгоритмы Xerox и BAMA сформировали лингвистический лексический подход к морфологическому моделированию арабского языка. Множество работ посвящены развитию похода: адаптации для диалектов, ускорению вычислений, расширению баз данных и их эффективному хранению [13–16].

В то же время существуют подходы к морфологическому моделированию, использующие методы машинного обучения. Для алгоритмов обучения с учителем требуется корпус размеченных данных, так как вычисление параметров модели происходит статистически на большом объеме данных без дополнительных лингвистических знаний. Основательный обзор и классификация подходов к морфологическому моделированию представлены в работе [17]. Методы глубокого обучения в основном применяются в задаче устранения неоднозначности морфологического моделирования и выбора наиболее вероятного варианта разбора [18; 19].

Обработка естественного языка включает в себя множество взаимосвязанных задач. Задачи морфологического моделирования, парсинга, токенизации, стемминга и лемматизации, сегментации, выделения сущностей являются базовыми элементами высокоуровневых приложений, обеспечивающих машинный перевод, автоматическое исправление опечаток и т. д. Задачи сентимент-анализа или определения диалекта на первом этапе всегда будут включать предобработку текста и вычисление числовых характеристик. Таким образом, для решения аналитических задач требуется инструмент, позволяющий проводить все этапы обработки и анализа комплексно.

Программные пакеты для обработки арабского языка

Развитие ресурсов для обработки арабского языка идет стремительными темпами. Некоторые из библиотек фокусируются на отдельных задачах (токенизация, диакритизация), в то время как другие пытаются одновременно решать несколько задач. Большинство ресурсов посвящены определенному варианту арабского языка, например современному арабскому литературному языку или египетскому диалекту. Эти ресурсы также различаются по используемым языкам программирования, типам интерфейсов, которые они предоставляют, используемым представлениям данных и стандартам, а также степени общедоступности (например, с открытым исходным кодом, коммерческие, недоступные).

Ниже рассмотрены программные пакеты, объединяющие инструменты для решения спектра задач по автоматической обработке арабского языка.

MADAMIRA [20] предоставляет настраиваемый морфологический анализатор. Доступны частеречная разметка, токенизация, сегментация, лемматизация и выделение именованных сущностей. MADAMIRA разработан специально для арабского языка и поддерживает современный арабский литературный язык и египетский диалект. Предоставляет интерфейс командной строки и Java API.

Farasa [21; 22] — это коллекция библиотек на языке Java и интерфейсов командной строки для MSA. Включает в себя инструменты для диакритизации, сегментации, частеречной разметки, парсинга и выделения именованных сущностей.

Stanford CoreNLP [23] — библиотека на языке Java, интерфейс командной строки и сервер, предоставляющие различные компоненты NLP с варьирующейся поддержкой разных языков. Арабский язык поддерживается в парсинге, токенизации, частеречной разметке, разделении предложений и выделении именованных сущностей. Широко используемая библиотека Natural Language Toolkit (NLTK) [24] обеспечивает поддержку арабского языка через привязку к Stanford CoreNLP.

CAMeL Tools [25] — библиотека с открытым исходным кодом, состоящая из набора API (Application Programming Interface, программный интерфейс приложения) на языке Python и соответствующих интерфейсов командной строки. CAMeL Tools предоставляет утилиты для предобработки текста, морфологического моделирования, определения диалекта, определения именованных сущностей и сентимент-анализа.

Библиотека CAMeL Tools постоянно совершенствуется, имеет хорошую документацию и актуальный репозиторий на сайте GitHub[3]. В рамках исследования было решено протестировать морфологический модуль библиотеки с целью определения корректности его функционирования. Как было отмечено выше, выбор данного инструмента для проведения тестирования обусловлен его комплексным характером, позволяющим осуществлять как препроцессинг текста, так и решение задач прикладного характера. Кроме того, библиотека основана на последних достижениях в области обработки естественных языков и находится в свободном доступе.

Морфологическое моделирование с использованием CAMeL Tools

Морфологический модуль библиотеки CAMeL Tools является реализацией морфологического анализатора CALIMA-Star, представленного в статье [26]. Помимо морфологического анализа, модуль способен генерировать и изменять словоформы. Компоненты модуля оперируют базами данных в формате ALMORGEANA [27], которые содержат три таблицы для префиксов, суффиксов и основ и три таблицы совместимостей: префикс — суффикс, префикс — основа и основа — суффикс.

Под морфологическим анализом в рамках библиотеки понимается нахождение всех возможных вариантов реализации слова вне контекста. Каждый из вариантов определен списком лексических признаков, таких как лемма, толкование (англ. gloss), и морфологических признаков, таких как часть речи, род, число, падеж и т. д. Анализатор ожидает получить на вход слово и пытается сопоставить префикс, основу и суффикс из таблицы базы данных с полученной формой слова

[3] GitHub — CAMeL-Lab/camel_tools. URL: https://github.com/CAMeL-Lab/camel_tools (дата обращения: 04.08.2022).

на входе, при этом учитывая ограничения совместимости префиксов, основ и суффиксов между собой. Алгоритм в целом соответствует схеме BAMA, но также позволяет обрабатывать иностранные слова и знаки пунктуации.

Под генерацией понимается изменение леммы по набору морфологических категорий. Алгоритм генерации реализован согласно статье [27]. Генератор на входе принимает лемму с учетом частеречной принадлежности, на выходе производит все возможные словоформы с их полным анализом. Если при этом на входе указаны другие категории, например род, лицо, падеж, то на выходе будут только словоформы, обладающие этими признаками.

Изменение словоформы (реинфлекция, англ. reinflection) отличается от генерации тем, что на вход подается уже слово в некоторой словоформе вместе со списком морфологических признаков словоформы, которую хочется получить на выходе. Реинфлектор (англ. reinflector) не ограничен определенной леммой и частью речи, потому генерирует все возможные варианты разбора словоформы на входе, в том числе все возможные леммы и части речи. Затем, согласно списку морфологических признаков на входе, а также списку признаков, полученных в результате разбора, составляется новый список признаков, который поступает на вход генератору. Генератор производит новые словоформы для каждого из вариантов разбора.

Если рассмотреть программный интерфейс морфологического модуля, то он состоит из пяти подмодулей.

1. camel_tools.morphology.database.

Класс MorphologyDB парсит файл базы данных и возвращает индексы, которые затем используются анализатором, генератором и реинфлектором. Не нужно напрямую обращаться к экземплярам MorphologyDB, а только передавать их в качестве аргументов при создании новых экземпляров компонентов анализатора, генератора и реинфлектора. Вместе с CAMeL Tools поставляются следующие базы данных (БД):

— calima-msa-r13: БД для анализа современного стандартного арабского языка;
— calima-egy-r13: БД для анализа египетского арабского языка;
— calima-glf-01: БД для анализа диалектов Персидского залива.

2. camel_tools.morphology.analyzer.

Класс морфологического анализатора. Метод AnalyzedWord(word) возвращает все возможные варианты морфологического разбора слова (word).

3. camel_tools.morphology.generator.

Класс генератора. Метод generate (lemma, feats) генерирует словоформы и связанные с ними варианты разбора для леммы (lemma) и заданного набора признаков (feats).

4. camel_tools.morphology.reinflector.

Класс реинфлектора. Метод reinflect (word, feats) создает словоформы и связанные с ними разборы для слова (word) и набора признаков (feats).

5. camel_tools.morphology.errors.

Базовый класс для ошибок морфологического модуля и отдельные классы для каждого из подмодулей.

Полная документация библиотеки доступна на сайте CAMeL Tools Documentation[4].

[4] CAMeL Tools Documentation — camel_tools 1.5.2 documentation. URL: https://camel-tools.readthedocs.io (дата обращения: 04.08.2022).

Тестирование системы морфологического моделирования CAMeL Tools

Для определения корректности работы морфологического компонента CAMeL Tools был составлен первичный перечень примеров, содержащих как типовые и достаточно распространенные, так и редкие словоформы. Критерием выборки являлись не количественные, а качественные показатели, опирающиеся на особенности формализации арабского языка и отражающие описанные выше трудности его обработки. В зависимости от результатов тестирования первичный перечень расширялся. Кроме того, учитывалась вариативность обобщенного понятия «арабский язык», которое объединяет классический арабский язык, современный арабский литературный язык и современные арабские диалекты. В данной статье приведем примеры, которые преимущественно взяты из текста Корана и содержат лексику, как свойственную современному языку, так и характеризующую особенности языка Корана. Известно, что приложения для процессинга арабского текста демонстрируют бóльшую точность при работе с современным языком, поэтому анализ классического языка особенно актуален.

Критерии выборки примеров были также направлены на тестирование корректности работы анализатора с точки зрения трудностей формализации арабского языка, которые были описаны выше. Речь идет о слитном написании ряда частиц, предлогов, слитных местоимений, что вызывает проблемы с сегментацией данных компонентов. Другая особенность касается задачи снятия морфологической и лексической омонимии, что распространяется и на неоднозначность служебных слов. Так, фраза لِيَبْلُوَكُمْ *liyabluwakum* включает в себя частицу *li*, которая в зависимости от лексико-грамматического контекста имеет разные значения и влияет на наклонение последующего глагола. Сопоставительный анализ словоформ, содержащих данную частицу, позволит сделать выводы относительно корректности работы приложения с точки зрения семантики служебных слов и решения задачи снятия лексической и морфологической омонимии. В данном случае частица *li-* несет значение цели:

الَّذِي خَلَقَ الْمَوْتَ وَالْحَيَاةَ لِيَبْلُوَكُمْ أَيُّكُمْ أَحْسَنُ عَمَلًا وَهُوَ الْعَزِيزُ الْغَفُورُ

…который создал смерть и жизнь, **чтобы испытать вас**, кто из вас лучше по деяниям, — Он велик, прощающ! (67:2)

لِيَبْلُوَكُمْ ‘чтобы испытать вас’

li-yabluwa-kum[5]

чтобы-IPFV.3SG.M-испытывать<SUBJ>=2PL.M[6]

В данном случае разбор словоформы в целом был выполнен корректно, при этом было отражено значение, передаваемое частицей *li-*:

‘gloss’: ‘for_+_them_(people)_to+afflict;test+you’

يَوْمَئِذٍ يَصْدُرُ النَّاسُ أَشْتَاتًا لِّيُرَوْا أَعْمَالَهُمْ

[5] В данной строке здесь и далее представлено сегментирование рассматриваемой фразы.

[6] Здесь и далее представлено поморфемное глоссирование, оформленное в соответствии с Лейпцигскими правилами: The Leipzig Glossing Rules. URL: http://www.eva.mpg.de/lingua/pdf/LGR08.02.05.pdf (дата обращения: 10.07.2022), а также с учетом комментариев относительно применения глоссирования к арабскому языку, опубликованных в монографии под ред. В. Б. Касевича «Грамматика и семантика восточного текста» [28].

В тот день выйдут люди толпами, **чтобы им показаны были** их деяния; (99:6)
لِّيُرَوْا
li-yuraw
чтобы=IPFV.PASS.3PL.M-показывать<SUBJ>

Корректно определяя в данном примере значение частицы *li-*, наклонение глагола, число и род, анализатор не предлагает актуального для данной словоформы описания. Изучение разбора словоформы لِّيُرَوْا подтверждает, что значение страдательного залога не учтено ни при расстановке огласовок (см. 'diac': 'لِيَرَوا'), ни в передаче значения глоссы 'gloss': 'for_+_them_(people)_to+see;think;believe+[masc.pl.]'. Кроме того, значение данного глагола в данном контексте не «видеть», а «показывать», что соответствует значению четвертой породы глагола. Порядок корневых также отражен не вполне корректно: 'root': 'ر.#.#', вместо 'root': 'ر.#.#' (т. е. رأى). Что касается определения основной леммы, то система допускает ошибку, добавляя к основе отсутствующие в ней буквы ('lex': 'راوَنْد'). Ошибка такого рода отмечена и при анализе иных производных от рассматриваемого корня.

Следующий пример содержит усеченную форму частицы *li-*, которая при употреблении с союзом *fa* теряет гласный. В данном случае частица несет значение побуждения к действию для третьих и первых лиц (т. е. значение «пусть, давайте»):

فَلْيَعْبُدُوا رَبَّ هَذَا الْبَيْتِ

Пусть же они поклоняются Господу этого дома (106:3)

فَلْيَعْبُدُوا 'пусть они поклоняются'
fa-l-ya'budū
и=пусть=IPFV. 3SG. M-поклоняться<JUSS>

В разборе данного примера корректно представлено описание словоформы и отражено значение частицы *-l-*, несущей в данном случае значение побуждения к действию для первых и третьих лиц. Графическая омонимия широко распространена в арабском языке, что во многом обусловлено отсутствием в современных текстах огласовок, передающих краткие гласные или их отсутствие и т. д. Поэтому частица ل может соответствовать как упомянутой выше частице *li-*, так и *la-*, которая придает значение усиления последующему слову, что демонстрирует приводимый ниже пример.

وَإِنَّهُ لِحُبِّ الْخَيْرِ لَشَدِيدٌ

И, поистине, он **тверд** в любви к благам! (100:8)
لَشَدِيدٌ 'действительно, силен (зд. тверд)'
la šadīdun
действительно=сильный.SG. M-NOM-INDF
'bw': 'لَ/EMPHATIC_PART+/شَدِيدNOUN_PROP'

Разбор данной словоформы ('لَ/EMPHATIC_PART+شَدِيد/NOUN_PROP') подтверждает корректную передачу частицы *la-*. Однако при этом присутствует описание NOUN_PROP, что подразумевает имя собственное, которое отсутствует в приведенном выше примере.

Значение запрета действия в арабском языке не допускает использования формы императива (т. е. повелительного наклонения) как такового, а передается конструкцией, содержащей частицу *lā-* и глагол в условном наклонении. В частности, приведенный ниже пример был разобран корректно.

كَلَّا لَا تُطِعْهُ وَاسْجُدْ وَاقْتَرِبْ

Так нет! **Не подчиняйся ему**, и поклонись, и приблизься! (96:19)

لَا تُطِعْهُ ‘не подчиняйся’
lā tuṭi‘=hu
не IPFV.2SG.M-подчиняться<IMP>=3SG.M

Следующие примеры направлены на тестирование функции корректности определения части речи, а также корня слова. Так, глагол إِطْمَأَنَّ «успокаиваться» достаточно хорошо распространен как в классическом, так и в современном арабском языке. При этом его морфологический анализ зачастую содержит неточности. Например, в корпусе Корана отмечено, что глагол إِطْمَأَنَّ и его производные относятся к XII породе и образованы от трехсогласного корня ط م ن, хотя здесь речь идет о четырехсогласном корне ط م ء ن.

فَإِذَا اطْمَأْنَنتُمْ فَأَقِيمُواْ الصَّلاَةَ

А когда вы **успокоитесь**, то выстаивайте молитву. (4:103)
اطْمَأْنَنتُمْ букв. ‘вы (мн. ч.) успокоились’
’iṭma’anantum
PRF-успокаиваться-2PL.M

В данном случае словоформа, глосса и корень слова были определены правильно:

‘bw’: ‘اِطْمَأْنَن/PV+تُم/PVSUFF_SUBJ:2MP’
‘gloss’: ‘be_calm;be_reassured+you_[masc.pl.]_<verb>’
‘root’: ‘ط.م.#.ن’

Однако при разборе однокоренного причастия корень передается с отсутствием *хамзы* (‘root’: ‘ط.م.ن’), что не вполне корректно:

يَا أَيَّتُهَا النَّفْسُ الْمُطْمَئِنَّةُ

О ты, душа **успокоившаяся**! (89:27)

الْمُطْمَئِنَّةُ ‘успокоившаяся’
’al-muṭma’innata
DEF-успокоившийся<PTCT.PASS>-F-ACC

Разбор данной словоформы позволяет сделать вывод о неточности передачи частей речи. Так, *’al-muṭma’innata* — причастие действительного залога, однако в разборе оно маркируется как прилагательное (‘pos’: ‘adj’).

В качестве другого примера определения части речи используем аят, содержащий причастие страдательного залога.

فَجَعَلَهُمْ كَعَصْفٍ مَأْكُولٍ

И сделал Он их точно нива со **съеденными** зернами. (105:5)

مَأْكُولٍ ‘съеденный’
ma’kūlin
съеденный<PTCT.PASS>.SG.M-NOM-INDF

В разборе данное причастие страдательного залога отмечено как «имя» ‘pos’: ‘noun’, при этом отсутствует маркер страдательного залога.

Для анализа корректности разбора словоформы, которая специфична для языка Корана, но не свойственна современному арабскому языку, рассмотрим пример глагола *tanazzalu*. Его особенность состоит в том, что в слове утрачен префикс имперфекта, необходимый для передачи соответствующего грамматического значения.

تَنَزَّلُ عَلَى كُلِّ أَفَّاكٍ أَثِيمٍ

Нисходят они на всякого лжеца, грешника. (26:222)
تَنَزَّلُ ‘нисходят’
tanazzalu
нисходить-IMPF.SG

Варианты разбора данной словоформы содержат указание форм перфекта, I, II и IV пород имперфекта и ни одного описания потенциальной возможности реализации V породы имперфекта. Данный факт говорит о том, что особенности классического языка требуют дополнительного рассмотрения при совершенствовании работы библиотеки CAMeL Tools.

Приводимые ниже примеры содержат постфикс так называемого усиленного наклонения, т. е. содержат نَّ, несущий значение усиления действия, передаваемого глаголом. Данное явление распространено в тексте Корана, встречается и в современном языке, но его описание варьируется в имеющихся грамматиках, поэтому важно изучить его разбор, полученный в результате использования морфологического анализатора. Для тестирования данного явления мы рассмотрели несколько примеров.

فَلَنَسْأَلَنَّ الَّذِينَ أُرْسِلَ إِلَيْهِمْ وَلَنَسْأَلَنَّ الْمُرْسَلِينَ

Мы **спросим** тех, к которым были посланы, и спросим посланников. (7:6)
فَلَنَسْأَلَنَّ букв. ‘и мы обязательно спросим’
fa-la-nas’alanna
и=обязательно=IMPF. 1SG-спрашивать<ENERG>[7]
‘bw’: ‘فَ/CONJ+لِ/PREP+نَ/IV1P+سْأَل/IV’, ‘gloss’:
‘and;so_+_for_+_us_(to)+ask;inquire;request’,

يَقُولُونَ لَئِن رَّجَعْنَا إِلَى الْمَدِينَةِ لَيُخْرِجَنَّ الْأَعَزُّ مِنْهَا الْأَذَلَّ

Они говорят: «Если мы вернемся в город, то сильнейший там **изгонит** слабейшего». (63:8)

لَيُخْرِجَنَّ ‘обязательно изгонит’
la-yuḫriğanna
обязательно=IPFV. 3SG. M-изгонять<ENERG>

[7] ENERG — значение усиленного наклонения, образованного с помощью так называемого нуна усиления نَّ.

Постфикс усиленного наклонения ن в неогласованных текстах омонимичен морфеме женского рода множественного числа. Все разобранные нами примеры продемонстрировали, что потенциальная возможность реализации словоформ с нуном усиления не учтена, разборы подтверждают, что рассматриваемая морфема воспринимается как маркер женского рода, что демонстрирует фрагмент разбора словоформы لَيُخْرِجَنَّ : for_+_them_(women)_to+go_out;exit;leave+[fem.pl.].

وَلَقَدْ فَتَنَّا الَّذِينَ مِن قَبْلِهِمْ فَلَيَعْلَمَنَّ اللَّهُ الَّذِينَ صَدَقُوا وَلَيَعْلَمَنَّ الْكَاذِبِينَ

Мы испытали тех, кто был до них; **ведь знает** Аллах тех, которые правдивы, и знает лживых! (29:3)

فَلَيَعْلَمَنَّ букв. ‘Ведь, действительно, знает’

fa-lya‘lamanna

ведь=действительно=IPFV.3SG.M-знать<ENERG>

Третий пример подтверждает факт отсутствия в системе значения усиленного наклонения:

‘diac’: ‘فَلِيُعْلِمْنَ’, ‘lex’: ‘أَعْلَم’, ‘bw’: ‘فَ/CONJ+لِ/PREP+/يُIV3FP+عْلِمِ/IV+نَ/IVSUFF_SUBJ:FP’, ‘gloss’: ‘and;so_+_for_+_them_(women)_to+notify;inform+[fem.pl.]’,

Библиотека CAMeL Tools содержит приложения для анализа египетского диалекта и диалекта стран Персидского залива. Мы протестировали несколько частотных фраз на египетском диалекте и получили результаты, подтверждающие высокую степень корректности разбора рассматриваемых сочетаний. Примеры включали в себя именные и глагольные отрицания, маркирующие египетский диалект, например:

لغة وحدة بس ماتكفيش ابدا

Только одного языка никогда **недостаточно**.

ماتكفيش ‘недостаточно’

mā-tkafī-š

NEG=IMPRF.3SG.F-быть достаточным=NEG

Фрагмент разбора представлен следующим образом:

{‘diac’: ‘مَا_اِتْكَفَش’, ‘lex’: ‘كَفَى’, ‘bw’: ‘ما/NEG_PART+اِتْكَفَى/PV_PASS+(null)/PVSUFF_SUBJ:3MS+ش/NEG_PART’, ‘gloss’:

‘not_[CALIMA]+be_enough;suffice_[CALIMA]+he;it_+_not_[CALIMA]’, ‘pos’: ‘verb’…}

ماحَدِّشْ

никто

mā-ḥaddi-š

NEG=один.IND.SG.M=NEG

{‘diac’: ‘مَاحَدِّش’, ‘lex’: ‘حَدّ’, ‘bw’: ‘ما/VERB+حَدّ/NOUN+ش/NEG_PART’, ‘gloss’: ‘not_[CALIMA]+stop;halt;end;extent;limit;level;border;frontier_[CALIMA]+not_[CALIMA]’…}

Интересно отметить, что разбор приведенного выше слова без подключения базы данных для египетского диалекта рассматривает его как заимствование, что в анализаторе маркируется как «Foreign»:

‘pos’: ‘foreign’, ‘diac’: ‘ما حدش’, ‘lex’: ‘ما حدش’, ‘bw’: ‘ما حدش/FOREIGN’

При подключении базы данных египетского диалекта разбор корректен.

Другой пример для тестирования содержит глагол с префиксом, передающим действие, осуществляемое в настоящий момент времени. Аналогичный префикс отсутствует в арабском литературном языке.

أنَا بَحِبَّكْ

Я **люблю тебя**.
بَحِبَّكْ ‘люблю тебя’
baḥibb=ak
IMPF.1SG. любить=2SG.M

Приведенная в разборе глосса показывает корректность семантики, передаваемой префиксом ب (настоящее-будущее время). В арабском литературном языке отсутствует аналогичная морфема.

‘gloss’: ‘[present-tense]_+_I_[CALIMA]+love;like;want_[CALIMA]+you_[CALIMA]’,

Следующие примеры направлены на тестирование корректности разбора слитных местоимений и слов, отсутствующих в арабском литературном языке:

تْحِبِّي نِرْقِصِي مَعَايَا ؟

Ты хочешь потанцевать **со мной**?

معايا ‘со мной’
ma‘ā=yā
с=я.ACC
‘gloss’: ‘with_[CALIMA]+my_[CALIMA]’

إزَّيَّك

Как твои дела?
إزَّيَّكْ ‘как ты’
’*izzayya-k*
как=ты.ACC.SG.M

В данном примере система учла графическую омонимию слитных местоимений мужского и женского родов и предложила два соответствующих разбора.

Выборочное тестирование словоформ, содержащих типичные для египетского диалекта особенности, продемонстрировало высокую степень корректности работы пакета египетского диалекта библиотеки CAMeL Tools, что особенно важно, учитывая недостаточное внимание к процессингу диалектов по сравнению с арабским литературным языком, а также преимущественно устный характер функционирования диалектов, отсутствие стандартизированной системы письма для них. В этом отношении возможности, предлагаемые CAMeL Tools, безусловно, важны.

Выводы

Тестирование инструментов для морфологического моделирования CAMeL Tools позволяет сделать следующие выводы, которые могут быть актуальны как при использовании данного решения для проведения лингвистических исследований, так и при разработке других лингвистических программных приложений, направленных на процессинг арабского текста.

При проведении междисциплинарных исследований важную роль играет терминологический аппарат, который должен соответствовать каждому из используемых научных направлений. Зачастую в процессе междисциплинарных исследований создается новая терминология. Описание морфологических особенностей CAMeL[8] демонстрирует некоторое расхождение используемых понятий и концептов с терминами, принятыми в языкознании в целом и в арабистических штудиях в частности. Так, значения терминов «префикс» (*prefix*) и «суффикс» (*suffix*) распространяются не только на словоизменительные морфемы, но и на служебные слова и местоименные клитики. Например, к именным «префиксам» разработчики относят вопросительную частицу أ, союзы ف ، و, слитные предлоги, частицы, определенный артикль.

Описание служебных частей речи также имеет свои особенности. Так, сокращением PREP, означающим Preposition, т. е. «предлог», маркируются частицы, реализуемые перед глаголами, что демонстрируют приведенные выше примеры, что противоречит назначению предлогов, которые не могут стоять в препозиции к глаголам.

Обращает на себя внимание и факт отсутствия в перечне частей речи причастий, которые при описании, как правило, передаются как «noun», т. е. «имя». Причастия, действительно, зачастую реализуются в значении имени. Однако в наших примерах встретилось значительное количество причастий, передающих отсутствующие у имени категории действительного или страдательного залогов и несущих синтаксические функции причастия. Поэтому не вполне корректно обозначать такого рода слова как «имя». Если исходить из того, что авторы опирались на укрупненную классификацию частей речи, свойственную арабскому языкознанию (имя, глагол, служебные части речи) и не учитывающую промежуточные части речи, то тогда не должны выделяться и прилагательные, которые широко представлены в данном приложении. Присутствует даже обозначение степеней сравнения прилагательных. Все, что сказано про причастия, может быть отнесено и к масдарам, которые также не получили отдельного описания в рассматриваемом решении.

Тестирование работы морфологического анализатора выявило некоторые неточности в описании редких словоформ. В первую очередь следует отметить тот факт, что система не учитывает суффикс, служащий для передачи так называемого усиленного наклонения. Речь идет о морфеме ن, используемой для передачи усиления действия, выраженного глаголом. Ни один из рассмотренных примеров, содержащий *нун* усиления, не был корректно описан. При этом система воспринимала его в качестве маркера глагола женского рода множественного числа.

В качестве примеров для тестирования корректности анализа словоформ, отражающих особенности языка Корана, мы использовали аяты Корана, в которых

[8] CAMeL Morphology Features // camel_tools 1.5.2 documentation. URL: https://camel-tools.readthedocs.io/en/latest/reference/camel_morphology_features.html#other-features (дата обращения: 04.08.2022).

использованы глаголы в пятой породе имперфекта, реализованные без префикса *ta*. Однако система воспринимает данный глагол как перфект, т. е. не учитывается соответствующая особенность классического языка.

Приведенные выше неточности носят частный характер и не влияют на корректность работы системы. Некоторые отмеченные наблюдения следует учитывать при работе с инструментами, направленными на процессинг арабского текста. В целом тестирование решений для морфологического моделирования CAMeL Tools подтверждает высокую степень корректности его работы, что особенно важно в контексте процессинга арабского языка, характеризующегося сложной морфологической структурой и иными особенностями, затрудняющими вопросы его формализации. В данном случае следует особенно отметить высокую точность сегментации словоформ и снятие проблемы лексической омонимии даже на уровне служебных слов.

Несмотря на то что основной задачей исследования являлось тестирование работы морфологического анализатора, было осуществлено и выборочное тестирование генератора и реинфлектора. В первом случае речь шла об образовании производных словоформ от начальной формы слова. В результате мы обнаружили корректность всех форм глагола в перфекте, имперфекте, сослагательном и условном наклонениях. Но при этом важно отметить, что исходное слово должно быть полностью огласовано (что не требуется при работе анализатора).

Что касается реинфлектора, то система корректно возводила формы ломаного множественного числа имени к запрашиваемой форме единственного, но при образовании ряда глагольных форм происходили неточности, связанные с изменением времени глагола. Результаты тестирования данного типа морфологического моделирования требуют более детального рассмотрения.

Анализ особенностей арабского языка в контексте тестирования приложений для его процессинга продемонстрировал существенный прогресс, произошедший за последние десятилетия в области обработки естественного языка в целом и арабского языка в частности. Вместе с тем выявлены некоторые лакуны в разборе словоформ, связанные во многом с уникальными морфологическими особенностями арабского языка, часть которых не имеет прямых аналогов в других языковых системах. Выявленные неточности могут быть учтены при модернизировании имеющихся программных решений и разработке новых приложений. Рассмотренные темы вновь ставят вопрос о необходимости повышения качества проведения междисциплинарных исследований и расширения профессиональных компетенций их участников.

Электронные ресурсы

I. *The Quranic Arabic Corpus*. URL: https://corpus.quran.com (дата обращения: 04.08.2022).
II. *ArabiCorpus*. URL: https://arabicorpus.byu.edu (дата обращения: 04.08.2022).
III. *Natural Language Toolkit*. URL: https://www.nltk.org (дата обращения: 04.08.2022).
IV. *Qutuf*. URL: https://github.com/Qutuf/qutuf (дата обращения: 04.08.2022).
V. *CAMeL Tools*. URL: https://camel-tools.readthedocs.io/en/latest/overview.html (дата обращения: 04.08.2022).

Литература

1. Национальная стратегия развития искусственного интеллекта на период до 2030 года. 2019. 10 окт. № 490. URL: http://static.kremlin.ru/media/events/files/ru/AH4x6HgKWANwVtMOfPDhcbRpvd1HCCsv.pdf (дата обращения: 04.08.2022).

2. *Buckwalter T.* Buckwalter Arabic morphological analyzer version 1.0 // Linguistic Data Consortium, University of Pennsylvania. 2002. URL: https://doi.org/10.35111/7vzm-mb15 (дата обращения: 04.08.2022).

3. تقرير حالة اللغة العربية ومستقبلها إعداد وإشراف وزارة الثقافة والشباب في دولة الإمارات العربية المتحدة الرقم الدولي [Отчет о состоянии арабского языка и его будущем // Министерство культуры и молодежи Объединенных Арабских Эмиратов. 2021. 683 с.] (На араб. яз.)

4. *Obeid O., Zalmout N., Khalifa S., Taji D., Oudah M., Alhafni B., Inoue G., Eryani F., Erdmann A., Habash N.* CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing // Proceedings of the 12th language resources and evaluation conference. Marseille, France. European Language Resources Association. 2020. P. 7022–7032.

5. *Callaos N.* Rigor and Inter-disciplinary Communication: Intellectual Perspectives from Different Disciplinary and Inter-Disciplinary Fields. Independently published, 2020. 100 p.

6. *Beesley K.* Timothy Buckwalter, and Stewart Newton. Two-Level Finite-State Analysis of Arabic Morphology // The Seminar on Bilingual Computing in Arabic and English, Cambridge: University of Cambridge, 1989. P. 63–72.

7. *Beesley K.* Arabic Morphology Using Only Finite-State Operations // COLING-ACL'98 Proceedings of the Workshop on Computational Approaches to Semitic languages. Montreal, 1998. P. 50–57.

8. *Koskenniemi K.* Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. PhD thesis. Helsinki: University of Helsinki, 1983. 164 p.

9. *Algarni M.* Light Morphology and Arabic Information Retrieval. A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy. Canterbury: University of Canterbury, 2016. 157 p.

10. *Buckwalter T.* Buckwalter Arabic morphological analyzer version 1.0 LDC2002L49 Web Download. Philadelphia: Linguistic Data Consortium, 2002. https://doi.org/10.35111/7vzm-mb15

11. *Maamouri M., Bies A.* Developing an Arabic treebank: Methods, guidelines, procedures, and tools // Proceedings of the Workshop on Computational Approaches to Arabic Script-based languages. Geneva, Switzerland. COLING. 2004. P. 2–9.

12. *Maamouri M., Bies A., Buckwalter T., Mekki W.* The penn Arabic treebank: Building a large-scale annotated Arabic corpus // NEMLAR conference on Arabic language resources and tools. 2004. Vol. 27. P. 466–467.

13. *Habash N., Rambow O.* MAGEAD: A morphological analyzer and generator for the Arabic dialects // Proceedings of 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Sydney: Association for Computational Linguistics, 2006. P. 681–688.

14. *Soudi A., Cavalli-Sforza V., Jamari A.* A computational lexeme-based treatment of Arabic morphology // Proceedings of the Arabic Natural Language Processing Workshop, Conference of the Association for Computational Linguistics (ACL 2001). 2001. P. 50–57.

15. *Cavalli-Sforza V., Soudi A., Mitamura T.* Arabic morphology generation using a concatenative strategy // Proceedings of the 1st North American chapter of the Association for Computational Linguistics Conference. Stroudsburg: Association for Computational Linguistics, 2000. P. 86–93.

16. *Habash N.* Arabic morphological representations for machine translation // Arabic Computational Morphology: Knowledge-Based and Empirical Methods / A. Soudi, A. van den Bosch, G. Neumann (eds). Dordrecht: Springer Netherlands, 2007. P. 263–285.

17. *Alothman A., Alsalman A. M.* Arabic Morphological Analysis Techniques // International Journal of Advanced Computer Science and Applications. 2020. Vol. 11, no. 2. P. 214–222.

18. *Zalmout N., Habash N.* Don't throw those morphological analyzers away just yet: Neural morphological disambiguation for Arabic // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Copenhagen, 2017. P. 704–713.

19. *Zalmout N., Habash N.* Adversarial multitask learning for joint multi-feature and multi-dialect morphological modeling // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. P. 1775–1786.

20. *Pasha A., Al-Badrashiny M., Diab M., Habash N., Pooleery M., Rambow O.* MADAMIRA v2.0 User Manual. Center for Computational Learning Systems. Columbia University, 2015. 40 p.

21. *Abdelali A., Darwish K., Durrani N., Mubarak H.* Farasa: A fast and furious segmenter for Arabic

// Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations. 2016. P. 11–16.

22. *Darwish K., Mubarak H.* Farasa: A new fast and accurate Arabic word segmenter // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. P. 1070–1074.

23. *Manning C. D., Surdeanu M., Bauer J.* The Stanford CoreNLP natural language processing toolkit // Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations. 2014. P. 55–60.

24. *Loper E., Bird S.* Nltk: The natural language toolkit // Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. 2002. P. 63–70.

25. *Taji D., Khalifa S., Obeid O., Eryani F., Habash N.* An Arabic Morphological Analyzer and Generator with Copious Features // Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology. Brussels: Association for Computational Linguistics, 2018. P. 140–150.

26. *Taji D., Khalifa S., Obeid O., Eryani F., Habash N.* An Arabic Morphological Analyzer and Generator with Copious Features // Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology. Brussels: Association for Computational Linguistics, 2018. P. 140–150.

27. *Habash N.* Arabic Morphological Representations for Machine Translation // Arabic Computational Morphology: Knowledge-based and Empirical Methods / Antal van den Bosch et al. (eds). Text, Speech and Language Technology, vol. 38. Springer, Dordrecht. 2007. https://doi.org/10.1007/978-1-4020-6046-5_14

28. *Берникова О. А.* Арабский язык // Грамматика и семантика восточного текста. Квантитативные характеристики / отв. ред. В. Б. Касевич. СПб., 2011. С. 35–48.

Статья поступила в редакцию 3 октября 2022 г.,
рекомендована к печати 30 июня 2023 г.

Контактная информация:

Берникова Ольга Александровна — канд. филол. наук; o.bernikova@spbu.ru

Кижаева Наталья Александровна — канд. физ.-мат. наук; natalia.kizhaeva@gmail.com

Peculiarities of the Arabic Language Processing: Morphological Modeling*

O. A. Bernikova, N. A. Kizhaeva

St. Petersburg State University,
7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

For citation: Bernikova O. A., Kizhaeva N. A. Peculiarities of the Arabic Language Processing: Morphological Modeling. *Vestnik of Saint Petersburg University. Asian and African Studies*, 2023, vol. 15, issue 3, pp. 459–484. https://doi.org/10.21638/spbu13.2023.302 (In Russian)

The paper deals with the features of morphological modeling of the Arabic language based on the definition of the specifics of its formalization. Morphological modeling is one of the key stages of automatic text analysis and includes tools for building a word form to a stem, root, definition of a part of speech, automatic construction (generation) of a given word form, etc. The objectives of the study are interdisciplinary in nature and include both the theoretical aspects of studying the features of the Arabic language, which are most relevant for its automatic processing, and the study of existing morphological analyzers and determining the specifics of their work. The practical part is based on testing the CAMeL Tools, one of the advantages of which is its comprehensive nature, which allows both preprocessing of text and solving applied problems, including sentiment analysis. The criteria for selecting examples for testing took into account the features of the Arabic language, which are difficult for its formalization (segmentation of functional words with continuous spelling, morphological and lexical homonymy, etc.). The variability of the generalized concept of "the Arabic language" is taken into account,

* The research was carried out at the expense of the grant of the Russian Science Foundation no. 22-28-01046, https://rscf.ru/project/22-28-01046/.

which combines classical Arabic, Modern Standard Arabic and modern Arabic dialects. Testing tools for morphological modeling allows us to draw conclusions about the need to improve the terminological apparatus, the variability of which is noted in the description of word forms. Such kind of variation (divergence from the concepts accepted in general linguistics) potentially leads to a distortion of the results of lexico-semantic analysis. During the analysis, some gaps were noted related to the definition of part-of-speech belonging, the description of word forms, etc. The results of the study are relevant both for linguistic research and for improving the development of software applications aimed at processing the Arabic text.

Keywords: Arabic language, morphological modeling, analyzer, processing.

References

1. *National strategy for the development of artificial intelligence for the period up to 2030.* October 10, 2019. No. 490. Available at: http://static.kremlin.ru/media/events/files/ru/AH4x6HgKWANwVtMOfPDhcbRpvd1HCCsv.pdf (accessed: 04.08.2022). (In Russian)

2. Buckwalter T. Buckwalter Arabic morphological analyzer version 1.0. *Linguistic Data Consortium, University of Pennsylvania*. 2002. Available at: https://doi.org/10.35111/7vzm-mb15 (accessed: 04.08.2022).

3. Report on the status of the Arabic language and its future. *Ministry of Culture and Youth of the United Arab Emirates.* 2021. 683 p. (In Arabic.)

4. Obeid O., Zalmout N., Khalifa S., Taji D., Oudah M., Alhafni B., Inoue G., Eryani F., Erdmann A., Habash N. CAMeL Tools: An open source Python toolkit for Arabic natural language processing. *Proceedings of the 12th language resources and evaluation conference*. Marseille, France. European Language Resources Association. 2020, pp. 7022–7032.

5. Callaos N. *Rigor and Inter-disciplinary Communication: Intellectual Perspectives from Different Disciplinary and Inter-Disciplinary Fields*. Independently published, 2020. 100 p.

6. Beesley K. Timothy Buckwalter, and Stewart Newton. Two-Level Finite-State Analysis of Arabic Morphology. *The Seminar on Bilingual Computing in Arabic and English*. Cambridge, University of Cambridge, 1989, pp. 63–72.

7. Beesley K. Arabic Morphology Using Only Finite-State Operations. *COLING-ACL'98 Proceedings of the Workshop on Computational Approaches to Semitic languages*, Montreal, 1998, pp. 50–57.

8. Koskenniemi K. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. PhD thesis. Helsinki, University of Helsinki, 1983. 164 p.

9. Algarni M. *Light Morphology and Arabic Information Retrieval*. A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy. Canterbury: University of Canterbury, 2016. 157 p.

10. Buckwalter T. *Buckwalter Arabic morphological analyzer version 1.0 LDC2002L49 Web Download.* Philadelphia: Linguistic Data Consortium, 2002. https://doi.org/10.35111/7vzm-mb15

11. Maamouri M., Bies A. Developing an Arabic treebank: Methods, guidelines, procedures, and tools. *Proceedings of the Workshop on Computational Approaches to Arabic Script-based languages*, Geneva, Switzerland. COLING, 2004, pp. 2–9.

12. Maamouri M., Bies A., Buckwalter T., Mekki W. The penn Arabic treebank: Building a large-scale annotated Arabic corpus. *NEMLAR conference on Arabic language resources and tools*, 2004, vol. 27, pp. 466–467.

13. Habash N., Rambow O. MAGEAD: A morphological analyzer and generator for the Arabic dialects. *Proceedings of 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Association for Computational Linguistics, 2006, pp. 681–688.

14. Soudi A., Cavalli-Sforza V., Jamari A. A computational lexeme-based treatment of Arabic morphology. *Proceedings of the Arabic Natural Language Processing Workshop, Conference of the Association for Computational Linguistics (ACL 2001)*, 2001, pp. 50–57.

15. Cavalli-Sforza V., Soudi A., Mitamura T. Arabic morphology generation using a concatenative strategy. *Proceedings of the 1st North American chapter of the Association for Computational Linguistics Conference*. Stroudsburg, Association for Computational Linguistics, 2000, pp. 86–93.

16. Habash N. Arabic morphological representations for machine translation. *Arabic Computational Morphology: Knowledge-Based and Empirical Methods*, A. Soudi, A. van den Bosch, G. Neumann (eds). Dordrecht, Springer Netherlands, 2007, pp. 263–285.

17. Alothman A., Alsalman A. M. Arabic Morphological Analysis Techniques. *International Journal of Advanced Computer Science and Applications*, 2020, vol. 11, no. 2, pp. 214–222.

18. Zalmout N., Habash N. Don't throw those morphological analyzers away just yet: Neural morphological disambiguation for Arabic. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Copenhagen, 2017, pp. 704–713.

19. Zalmout N., Habash N. Adversarial multitask learning for joint multi-feature and multi-dialect morphological modeling. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Association for Computational Linguistics, 2019, pp. 1775–1786.

20. Pasha A., Al-Badrashiny M., Diab M., Habash N., Pooleery M., Rambow O. *MADAMIRA v2.0 User Manual*. Center for Computational Learning Systems. Columbia University, 2015. 40 p.

21. Abdelali A., Darwish K., Durrani N., Mubarak H. Farasa: A fast and furious segmenter for Arabic. *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*. 2016, pp. 11–16.

22. Darwish K., Mubarak H. Farasa: A new fast and accurate Arabic word segmenter. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pp. 1070–1074.

23. Manning C. D., Surdeanu M., Bauer J. The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 2014, pp. 55–60.

24. Loper E., Bird S. Nltk: The natural language toolkit. *Proceedings of the Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, 2002, pp. 63–70.

25. Taji D., Khalifa S., Obeid O., Eryani F., Habash N. An Arabic Morphological Analyzer and Generator with Copious Features. *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Brussels, Association for Computational Linguistics, 2018, pp. 140–150.

26. Taji D., Khalifa S., Obeid O., Eryani F., Habash N. An Arabic Morphological Analyzer and Generator with Copious Features. *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology.* Brussels, Association for Computational Linguistics, 2018, pp. 140–150.

27. Habash N. Arabic Morphological Representations for Machine Translation. *Arabic Computational Morphology: Knowledge-based and Empirical Methods.* Antal van den Bosch et al. (eds). Kluwer/Springer, 2007.

28. Bernikova O. A. The Arabic language. *Grammar and semantics of the Eastern text. Quantitative characteristics*. Ed. by V. B. Kasevich. St. Petersburg, 2011, pp. 35–48. (In Russian)

Received: October 3, 2022
Accepted: June 30, 2023

Author's information:

Olga A. Bernikova — PhD in Philology; o.bernikova@spbu.ru

Natalia A. Kizhaeva — PhD in Physics and Mathematics; natalia.kizhaeva@gmail.com