DOI: 10.24833/2410-2423-2019-2-18-113-118

ЭЛЕКТРОННЫЙ СЛОВАРЬ ПУШТУ: СОЗДАНИЕ БАЗЫ ДАННЫХ МОРФОЛОГИИ[1]

Ю.П. Лалетин, В.О. Сорвёнков, М.А. Тимофеев

Московский государственный институт международных отношений (университет) МИД России, 119454, Россия, Москва, пр. Вернадского, 76.

В статье рассматривается создание электронного словаря на базе «Пушту-русского словаря» М.Г. Асланова, который является в настоящее время наиболее полным словарём пушту. Бумажные словари неизбежно устаревают, в то время как электронные словари обладают целым рядом неоспоримых преимуществ по сравнению с традиционными. Работа над электронным словарём включает три этапа: 1) составление словника (или использование уже готового), 2) создание базы морфологии, 3) работу с синтаксисом, заключающуюся в создании корпуса текстов, что позволит выявить несвободную сочетаемость слов, начиная от словосочетания. Статья посвящена главным образом созданию базы данных морфологии пушту для электронного словаря, что было сделано впервые в мире. Основное внимание в работе уделено именам существительным и прилагательным. На основе грамматических переменных выделены парадигматические классы указанных частей речи, а в их рамках – все возможные формы слова (морфемы). Для имён существительных описано двадцать шесть парадигматических классов, а для имён прилагательных – восемь. Некоторые классы разделяются на подклассы, в каждый из которых входит одна лексема. Это относится к так называемым исключениям из правил. Для каждого класса приведено в качестве образца наиболее характерное слово. Каждая морфема (равно как и каждое значение слова) выступает отдельной словарной единицей, что даёт возможность пользователю легко находить нужное слово, а также осуществлять обратный перевод. Данная статья рассчитана на афганцев, владеющих языком пушту, а также русскоговорящих, работающих с пушту. Особый интерес результаты исследования представляют для тех, кто составляет или собирается составлять электронные словари, особенно для редких языков.

Ключевые слова: *язык пушту, компьютерная лексикография, электронный словарь, база данных морфологии, имя существительное, имя прилагательное, парадигматический класс, грамматические переменные*

В современной России в условиях модернизации системы профессионального образования выдвигается требование обеспечить подготовку по иностранному языку, включающую в себя формирование у обучающихся компетенций межкультурной коммуникации, оптимизирующей эффективность контактов в конкретной сфере и ситуации общения, а также достижение определённых целей коммуникации. Без полноценного словаря достичь подобной цели вряд ли возможно. В настоящее время наиболее полным словарём пушту является «Афганско-русский словарь» Мартироса Григорьевича Асланова, во втором издании названный «Пушту-русским словарём». Словарь содержит около 50 тысяч слов.

Известно, что лексика неизбежно устаревает, каждый день появляется новая. В настоящее время появление новой лексики приняло лавинообразный характер. Между тем упомянутый

[1] Статья подготовлена по гранту на выполнение научных работ молодыми исследователями под руководством докторов и кандидатов наук МГИМО.

словарь основан на материалах картотеки, которая собиралась автором многие годы, начиная с начала 1930-х годов, путём росписи лексики из разнообразных источников. Для составления словаря были использованы также личные записи автора, сделанные в Афганистане в 1933-1939 гг. Работа по сбору материала была в основном закончена к 1952 году, вышел словарь в 1966 году, а второе издание увидело свет в 1985 году с небольшим приложением списка новых слов.

Современные технологии позволяют воспользоваться неоспоримыми преимуществами электронной версии словаря, к которым относятся расширенные возможности представления содержания словарной статьи, возможность использования средств мультимедиа, удобный и быстрый поиск, характеристики всех особенностей включаемых в словарь слов, расширенные возможности описания словарной единицы, большой, практически неограниченный объём словарной базы, постоянное обновление последней, актуальность и динамичность, возможность менять направление перевода, вариативность использования, доступность [3].

Ограниченность имеющейся бумажной версии словаря пушту на фоне открывающихся в настоящее время перспектив побудили авторов настоящей статьи приступить к созданию компьютерного словаря пушту на базе российского сайта multitran.com, автором и директором которого является Андрей Валентинович Поминов, разработавший программное обеспечение для создания электронных словарей. В настоящее время на сайте выложены словари почти всех европейских языков.

Методологической основой исследования послужили современные подходы к обучению иностранным языкам – коммуникативный, компетентностный, личностно-ориентированный, концепция информатизации образования, концепция дистанционного обучения, концепция Веб 2.0 и теория использования социальных сервисов Веб 2.0 в обучении.

Теоретической основой исследования служат комплексный и междисциплинарный подходы, методы исследования и основные положения компьютерной лексикологии и лексикографии. Определённую помощь при проведении исследования оказала база данных по русской и английской морфологии [2].

Использование данных методов и подходов позволило выработать представления о **принципах и этапах подготовки электронного словаря**. Прежде всего на компьютере набирается сам словник. Это можно делать как самостоятельно, так и используя уже имеющиеся бумажные словари. В данном случае за основу электронного пушту-русского словаря взят упоминавшийся выше словарь М.Г. Асланова. Он аккуратно отсканирован, в настоящее время осуществляется проверка текста и вносятся необходимые исправления.

Так формируется полноценная основа, с которой можно будет работать в дальнейшем.

Следующим этапом работы выступает преобразование бумажной версии словаря в электронную таким образом, чтобы были представлены все формы того или иного слова и чтобы каждая такая форма имела статус отдельной словарной единицы. Тем самым преобразуется словарная статья. В электронном словаре единицей выступает не словарная статья, а либо одно значение лексемы, либо одна морфема. Все устойчивые выражения и фразеологизмы с этим словом, включая идиомы, пословицы и поговорки, выступают отдельными единицами. Особое внимание будет уделено передаче сочетаемости слов.

Каждый язык имеет свои способы кодирования грамматического смысла, в частности, словообразования и словоизменения. И пушту здесь не исключение. Систематическое описание этих способов и, как следствие, наличие всех возможных форм слова позволит пользователю при переводе быстро находить нужный ему вариант. Всё это наряду с превращением словарной единицы в значение слова позволит с лёгкостью осуществлять обратный перевод.

Поэтому на втором этапе авторы решали задачу формирования морфологической базы языка пушту.

На последующих этапах создания словаря происходит работа над синтаксисом, когда рассматривается связь слов в словосочетании и в предложении. Для этого в словарь добавляются словосочетания, особенно устойчивые глагольные сочетания глаголов.

Проведённая работа даст возможность учесть почти все особенности включаемых в словарь слов, что, в свою очередь, создаст условия для точного определения компьютером при поисковом запросе каждого слова. Подобный электронный словарь даст возможность не ограничиваться такими операциями, как перевод и толкование (хотя дают возможность изменить направление перевода и воспользоваться ус-

лугой компьютерного перевода с нужного нам языка), но и явится информационной базой для комбинаций на уровне предложений и текстов [1]. В итоге появится возможность создать полноценный электронный словарь пушту.

Далее рассмотрим, как конкретно авторами осуществлялось создание **базы данных морфологии пушту**. В статье основное внимание уделено именам существительным и прилагательным.

Поскольку в предлагаемой нами морфологической модели принят словарь основ, то база данных помимо основ учитываемых лексем содержит словарь списков флексий, соответствующих каждому парадигматическому классу. С каждой флексией связан набор значений грамматических переменных, приписываемый основе с данной флексией. Если в морфологической модели учитываются какие-либо типичные особенности словоизменения (например, чередование букв в основе), то информация о них также должна храниться в базе данных [4].

Работа начинается с разделения имён существительных на парадигматические классы (склонения). Выделение классов и морфем в рамках класса осуществляется по особенностям словоизменения на базе грамматических переменных. Применительно к именам существительным к таким переменным относятся: одушевлённость /неодушевлённость, род (мужской, женский), число (единственное, множественное), падеж (прямой, косвенный).

Далее выбирается слово в качестве образца того или иного класса и даются все его формы, чаще в виде окончаний, но в случае необходимости и всей лексемы. Результаты разработки морфологии представлены ниже в виде таблицы (в машинном варианте таблиц нет, формы слов образуют строки и столбцы). Каждая строка соответствует одному классу, причём вначале приводятся наиболее многочисленные группы. В первом столбце указан номер класса (в самой базе данных нумерация сквозная для всех частей речи), во втором приводится существительное в прямом падеже, в третьем существительное в косвенном падеже, в четвертом – множественное число в прямом падеже, затем множественное число в косвенном падеже и в заключение – перевод.

Приведённые в статье парадигматические классы **имен существительных** включают наибольшее число лексем. Остальные классы содержат по одному слову. Первые двенадцать классов представлены существительными мужского рода. В первый парадигматический класс входят имена существительные одушевлённые, во второй – неодушевлённые.

В 1-4-м классах словоизменение осуществляется путём прибавления окончания, а в пятом классе – его (окончания) изменения, а субстантивированные существительные шестого класса объединяют оба варианта. Во всех остальных классах (6-12) словоизменение осуществляется путём внутренней флексии (изменения основы). При этом существительные 6-го класса являются исключениями, поэтому в базе данных они приводятся все, и каждое из них образует свой подкласс.

№	Ед. ч. прямой падеж	Ед. ч. косвенный падеж	Мн. ч. прямой падеж	Мн. ч. косвенный падеж	Перевод
1	mez	mez	mezuna	mezuno	стол
2	talib	talib	talibān	talibāno	студент
3	chaku	chaku	chakugān	chakugāno	нож
4	sāda	sāda	sādagān	sādagāno	простак
5	saṛay	sari	sari	sario	мужчина
6	zalmay	zalmi	zalmiyān	zalmiyāno	юноша
7	xar	xrə	xrə	xro	осёл
8	špun	špānə	špānə	špano	пастух
9	ṭopak	ṭopak	ṭopək	ṭopəko	ружьё
10	plār	plār	plaruna	plaruno	отец
11	zoy	zoy	zāmən	zāməno	сын
12	vror	vror	vruṇa	vruṇo	брат

У имён существительных женского рода 13-15-го парадигматических классов словоизменение происходит изменением окончания, 16-17-го классов – добавлением окончания, 18-23-го – присоединением суффикса, а 24-26-го – внутренней флексией. Существительные 17-го класса выписываются все с образованием каждым из них своего подкласса.

№	Ед. ч. прямой падеж	Ед. ч. косвенный падеж	Мн. ч. прямой падеж	Мн. ч. косвенный падеж	Перевод
13	taxta	taxte	taxte	taxto	доска
14	malgəre	malgəre	malgəre	malgəro	подруга
15	bazi	bazəy	bazəy	bazəyo	игра
16	kəṛkəy	kəṛkəy	kəṛkəy	kəṛkəyo	окно
17	vradz	vradze	vradze	vradzo	день
18	bizo	bizo	bizogāne	bizogāno	обезьяна
19	mlā	mlā	mlāgāne	mlāgāno	поясница
20	xvā	xvā	xvāve	xvāvo	сторона
21	nave	nāve	nāveyāne	nāveyāno	невеста
22	tafrih	tafrih	tafrigāne	tafrigāno	отдых
23	tafrih	tafrih	tafrihāt	tafrihāto	отдых
24	mor	mor	mende	mendo	мать
25	xor	xor	xvende	xvenvo	сестра
26	lur	lur	luṇe	luṇo	дочь

Что касается **имен прилагательных**, то наибольшее их количество представлено в приведённых ниже восьми классах. Для прилагательных число морфем в одном классе увеличивается при уменьшении числа самих классов, поскольку прилагательные изменяются не только по числам и падежам, но и по родам. Вместе с тем все классы, за исключением первого, шестого и седьмого, содержат ограниченное число прилагательных, формы которых нужно будет расписать.

Словоизменение прилагательных первого и второго парадигматических классов характеризуется аффиксацией, третьей, шестой и седьмой – флексией аффиксов, а остальные, кроме восьмого, – внутренней флексией. К восьмому классу относятся неизменяемые прилагательные. Прилагательные 2-5-го и 8-го классов выписываются все с образованием каждым из них своего подкласса. Второй и третий классы включают по три подкласса, пятый класс – пять и восьмой – шесть подклассов.

№	Ед.ч. м.р. прям. падеж	Ед.ч. ж.р. прям. падеж	Мн.ч. м.р. прям. падеж	Мн.ч. ж.р. прям. падеж	Ед.ч. м.р. косв. падеж	Ед.ч. ж.р. косв. падеж	Мн.ч. м.р. косв. падеж	Мн.ч. ж.р. косв. падеж	Перевод
1	tor	tora	tor	tore	tor	tore	toro	toro	чёрный
2	um	uma	umə	ume	umə	ume	umo	umo	незрелый
3	bidə	bida	bidə	bide	bidə	bide	bido	bido	спящий
4	spor	spara	spārə	spare	spārə	spare	sparo	sparo	верховой
5	sur	sra	srə	sre	srə	sre	sro	sro	красный
6	nəvay	nəve	nəvi	nəve	nəvi	nəve	nəvo	nəvo	новый
7	praday	pradəy	pradi	pradəy	pradi	pradəy	pradio	pradiyo	чужой
8	abi	abi	abi	abi	abi	abi	abi	abi	голубой

Наибольшее количество форм имеют **глаголы**. Выделяются семь групп глаголов, которые образуют множество парадигматических классов, подклассов и форм в зависимости от таких грамматических переменных, как время (настоящее, прошедшее, будущее, перфект, плюсквамперфект), вид (несовершенный, совершенный), наклонение (изъявительное, повелительное, условно-желательное), лицо (первое, второе, третье), число (единственное, множественное), форма причастия (действительное причастие настоящего времени, действительное причастие прошедшего времени), потенциальная форма.

В качестве примера приведём два первых парадигматических класса, включающих формы настоящего времени одноосновных глаголов, то есть имеющих одну основу в инфинитиве и в настоящем времени (kavəl – делать) и двуосновных глаголов, то есть полностью меняющих основу в настоящем времени (kedəl – становиться, делаться). Первая буква k обоих глаголов поставлена в скобки, поскольку с их помощью образуется не одна тысяча глаголов действительного и страдательного залога соответственно: rasavəl–доставлять, rasedəl – прибывать, teravəl – проводить, teredəl – проходить, portakavəl – поднимать, portakedəl – подниматьсяи т.п.

№	1 лицо ед.ч.	2 лицо ед.ч.	3лицо ед.ч.	1 лицо мн.ч.	2 лицо мн.ч.	3 лицо мн.ч.	Инфинитив	Перевод
1	(k)avəm	(k)ave	(k)avi	(k)avu	(k)avəy	(k)avi	kavəl	делать
2	(k)eẓəm	(k)eẓe	(k)eẓi	(k)eẓu	(k)eẓəy	(k)eẓi	kedəl	делаться

Количество одноосновных и двуосновных глаголов в языке пушту примерно одинаково с небольшим преобладанием последних. Для обеспечения точности при поиске слов глаголы обоих классов выписываются вручную, и каждый из них образует свой подкласс. Приведём ещё по одному примеру глаголов каждого класса:

№	1 лицо ед.ч.	2 лицо ед.ч.	3лицо ед.ч.	1 лицо мн.ч.	2 лицо мн.ч.	3 лицо мн.ч.	Инфинитив	Перевод
3	likəm	like	liki	liku	likəy	liki	likəl	писать
4	dzəm	dze	dzi	dzu	dzəy	dzi	tləl	идти

Так прописываются все морфологические формы, каждая из которых выступает лексической единицей в электронном словаре.

Следующим этапом составления компьютерного словаря станет работа с синтаксисом, которая будет заключаться во введении корпуса текстов пушту, начиная с отдельных предложений.

Список литературы

1. Зубов А. В., Зубова И. И. Основы искусственного интеллекта для лингвистов. Москва: РГГУ, 2013. 320 с.
2. База данных по русской и английской морфологии [Электронный ресурс] – Режим доступа: http://www.solarix.ru/sql-dictionary-sdk.shtml (дата обращения 15.05.2018 г.).
3. Селегей В. П. Компьютерная лексикография [Электронный ресурс] – Режим доступа: https://www.abbyy.com/ru-ru/science/technologies/lexicography/ (Дата обращения 15.05.2018 г.).
4. Лингвистический процессор естественного языка [Электронный ресурс] – Режим доступа: https://studfiles.net/preview/972381/page:5/ (Дата обращения 16.05.2018 г.)

Сведения об авторах:

Лалетин Юрий Павлович – кандидат исторических наук, доцент кафедры индоиранских и африканских языков МГИМО. Сфера интересов: язык пушту, лингвострановедение. E-mail: yupl@mail.ru.

Сорвёнков Владислав Олегович – студент 4-го курса факультета Международных отношений МГИМО. Сфера интересов: язык пушту, лингвострановедение. E-mail: vladislavsorvyonkov@yandex.ru.

Тимофеев Михаил Алексеевич – студент 4-го курса факультета Международных отношений МГИМО. Сфера интересов: язык пушту, лингвострановедение. E-mail: tmishela@gmail.com.

ELECTRONIC PASHTO DICTIONARY: CREATING A MORPHOLOGY DATABASE

Yu.P. Laletin, V.O. Sorvyonkov, M.A. Timofeev

Moscow State Institute of International Relations (University),
76, Prospect Vernadskogo, Moscow, 119454, Russia.

Abstract: *The article deals with the creation of an electronic dictionary based on the "Pushto-Russian Dictionary" by M. G. Aslanov, which is currently the most comprehensive Pashto dictionary. However, paper dictionaries inevitably become outdated, while electronic dictionaries have a number of indisputable advantages over traditional ones. The work on an electronic dictionary includes three stages: 1) compiling a vocabulary (or using an already completed one), 2) creating a morphology base, 3) working with syntax, which consists in creating a corpus of texts that will allow revealing non-free compatibility of words,*

starting with a word combination. The article focuses mainly on creation of a Pashto morphology database for an electronic dictionary, on nouns and adjectives, which has never been done before. According to grammatical variables, the paradigmatic classes of the indicated parts of speech are distinguished, as well as all possible forms of the word (morphemes) within the classes. Twenty-six paradigmatic classes are distinguished for nouns, and eight for adjectives. Some classes are divided into subclasses, each of which includes one word. This refers to the so-called exceptions to the rules. For each class, the most characteristic word is given as a model. Each morpheme (as well as each meaning of a word) appears as a separate dictionary unit, which allows the user to easily find the desired word, as well as to make a reverse translation. This article is intended for Afghans who speak Pashto, as well as Russian speakers who deal with Pashto. Of particular interest are the results of the study for those who compose or intend to compile electronic dictionaries, especially of rare languages.

Key Words: *the Pashto language, computer lexicography, electronic dictionary, morphology database, noun, adjective, paradigmatic class, grammatical variables*

References

1. Zubov A. V., Zubova I. I. Osnovy iskusstvennogo intellekta dlia lingvistov. [Basics of artificial intelligence for linguists]. Moskva: RGGU, 2013. 320s.
2. Baza dannykh po russkoi I angliiskoi leksike I morfologii [Database on Russian and English vocabulary and morphology]. Available at: http://www.solarix.ru/sql-dictionary-sdk.shtml (accessed 15 May 2018).
3. Selegej V. P. Komp'iuternaia leksikografiia [Computer lexicography]. Availableat: https://www.abbyy.com/ru-ru/science/technologies/lexicography/ (accessed 15 May 2018).
4. Lingvisticheskii protsessor estestvennogo iazika [Natural Language Linguistic Processor]. Availableat: https://studfiles.net/preview/972381/page:5/ (accessed 16 May 2018).

About the authors:

Yuriy Pavlovich Laletin – PhD (History), Assistant Professor of the Department of IndoIranian and African Languages, MGIMO. Spheres of interest: the Pashto Language, linguistics. E-mail: yupl@mail.ru.

Vladislav Olegovich Sorvyonkov – fourth-year student of the International Relations Faculty, MGIMO. Spheres of interest: the Pashto Language, linguistics. E-mail: vladislavsorvyonkov@yandex.ru.

Mikhail Alekseevich Timofeev – fourth-year student of the International Relations Faculty, MGIMO. Spheres of interest: the Pashto Language, linguistics. E-mail: tmishela@gmail.com.

* * *