УДК 811.411.21 Вестник СПбГУ. Сер. 13. 2014. Вып. 1

*О. И. Редькин*

# ФОРМИРОВАНИЕ КОРПУСА ТЕКСТОВ И ОПРЕДЕЛЕНИЕ ЧАСТОТНОСТИ СЛОВ В АРАБСКОМ ЯЗЫКЕ: ПРОБЛЕМЫ И РЕШЕНИЯ[1]

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

Хотя проблема формирования корпуса текстов на материале индоевропейских языков, включая русский, сравнительно хорошо разработана, в отношении других языков, прежде всего арабского, она далека от своего окончательного решения. В статье рассматриваются проблемы и возможные решения при построении арабского корпуса текстов на базе материала из Интернета и других доступных источников, а также принципы отбора данных. В статье также приведены результаты формирования частотного словаря арабского языка, список наиболее распространенных арабских слов с их частотной индексацией. Библиогр. 6 назв. Табл. 1.

*Ключевые слова*: Арабский язык, корпус, компьютер, данные, обработка, частотность, словарь.

## FORMATION OF TEXT CORPUS AND FREQUENCY DEFINITION FOR THE WORDS IN THE ARABIC LANGUAGE: PROBLEMS AND SOLUTIONS

*O. I. Redkin*

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

Although the problem of formation of corpus on the material of the Indo-European languages, including Russian, is comparatively developed in relation to other languages and particularly Arabic, it is far from its final solution. The article deals with the problems and solutions for building the Arabic corpus, based on the material from the Internet and other available sources, and identifies the principles of data selection. The article also considers the results of formation of frequency dictionary of Arabic, as well as peculiarities of the Arabic phonology, morphology and script. Besides, the article studies some peculiarities of the stress in Arabic. The article is supplied with a list of the most common Arabic words with their frequency indexing. Refs 6. Tables 1.

*Keywords*: Arabic, corpus, computer, data, proceeding, frequency, dictionary.

Корпусная лингвистика является одним из наиболее актуальных направлений общего языкознания, о чем свидетельствуют публикации, посвященные данной проблематике, и результаты состоявшихся в последнее десятилетие научных конференций, а также публикаций в периодических журналах и сети Интернет [1–3]. Проблема формирования корпусов текстов давно перестала носить чисто теоретический характер, но имеет также и большое практическое значение в таких областях, как социолингвистика, лексикография, историческая лингвистика, психолингвистика [4], а также в разработке поисковых систем в сети Интернет [5] и ряде других областей.

Хотя методика формирования корпусов текстов на материале индоевропейских языков, в том числе и русского, в значительной мере разработана в отношении других языков, в частности арабского, она далека от окончательного решения. Несмотря на формирование в последние годы корпусов текстов на различном материале, в частности на базе текста Корана, что имеет большое научное значение, как спра-

[1] Публикация подготовлена в рамках поддержанного РГНФ научного проекта № 13-04-00425.

ведливо отмечает М. А. Мансур, арабский мир «не располагает арабским лингвистическим корпусом, сопоставимым с английским» [1, p. 81].

В значительной мере это объясняется тем, что многие решения, эффективные при обработке материала европейских языков с алфавитным письмом, богатой парадигмой словообразования и четко выраженными на графике границами слов, оказываются малопригодны для восточных языков с иероглифической письменностью, а также языков, письменность которых хотя и построена на алфавитном принципе, однако не всегда позволяет четко маркировать границы между словами, где критерии выделения слов могут трактоваться по-разному.

Хотя в корпусной лингвистике, как правило, рассматриваются базы данных, созданные на основе письменных текстов, существует и принципиальная возможность формирования корпусов на базе устного материала, что важно для бесписьменных языков и весьма актуально, например, при исследовании арабских диалектов. Использование в указанных целях звучащей речи осложняется тем, что обработка акустического материала затрудняется в силу наличия помех (так называемых шумов), а также редуцированным характером реализации акустических элементов, например, опущением окончаний и т. д., что является следствием действия общелингвистического принципа экономии артикуляционных усилий; при этом неполнота реализации устных текстов компенсируется синхронным экстралингвистическим контекстом, использовать который и учитывать при анализе записи в дальнейшем в большинстве случаев не представляется возможным. Меньшая степень использования акустического материала объясняется, скорее, недостаточным развитием методики регистрации таковых и сложностью их дальнейшей обработки, нежели игнорированием такого рода источника.

В целом причины, обусловливающие сложности сегментации письменных и устных текстов, во многом сходны, и в их основе лежат особенности структуры арабского языка, уже первое знакомство с которыми показывает если не невозможность использования методов, применяемых при анализе индоевропейских языков, то необходимость серьезной корректировки имеющегося инструментария или разработки новых подходов.

Как и любой другой, арабский текст представляет собой сложно структурированную линейную последовательность графических единиц или, в случае его устной реализации — последовательность фонем, а также суперсегментных элементов. Сегментация компонентов осуществляется как на основе формальных признаков — пауз в основном тоне в случае обработки устного текста, а также интервалов, которые являются формальными маркерами границ между словами письменного текста, — так и на основе чисто лингвистических критериев. Хотя существует набор критериев, позволяющих на первый взгляд успешно сегментировать текст, вместе с тем, как показывает практика, при обработке арабского материала все не выглядит столь однозначным.

Так, в случае акустического текста единство слов как произносительных комплексов в значительной мере определяется наличием единого ударения. Однако даже предварительный инструментальный анализ показывает, что в арабском языке демилитативная функция ударения не столь ярко выражена, как в других языках, в арабском, скорее, имеет место комбинация двух типов ударения — квантитативного и силового. При этом гласный ударенного слога, являющийся своего рода его

ядром, выделяется как за счет пролонгации, так и динамических характеристик (амплитуды колебаний), на что указывают, в частности, данные осциллографического анализа.

С проблемами, связанными с сегментацией текста, приходится иметь дело и при анализе письменных текстов, где использование таких формальных показателей, как интервалы между графическими элементами, оказывается недостаточно эффективным. Среди факторов, затрудняющих сегментацию арабского письменного текста, следует назвать наличие значительного числа слитных предлогов, частиц и трансфиксов, равно как и слитное написание определенного артикля, различные варианты написания одних и тех же графем в зависимости от их положения в слове.

В случае оптической обработки письменных текстов и их распознавания исследователь сталкивается и с такими проблемами, как наличие лакун в графике, а также шумов (например, дефектов бумаги, значительной дисперсии диакритических знаков, степени наклона графических символов, их нечеткой реализации, повреждений).

К сказанному следует добавить значительные отличия, существующие между диалектными текстами и текстами на арабском литературном языке, что является частным проявлением ситуации, получившей более полувека тому назад название диглоссии [6].

Имеют место также и различия между текстами из различных регионов арабского мира, равно как и между текстами, относящимися к определенным хронологическим периодам развития арабского языка, например, доисламскому, классическому, современному. Следует учитывать и тематическую маркированность арабского лингвистического материала, спектр которой весьма широк и простирается от сакральных текстов до предельно функционального и упрощенного стиля электронной коммуникации, например, текстов электронной почты или sms-сообщений. Наконец, реальность такова, что исследователи нередко имеют дело со «средним» языком, представляющим собой смесь местной разновидности диалекта и литературного языка. Проблему осложняет и нерешенность ряда теоретических положений общей лингвистики, в частности критериев понятия слова.

В упрощенном виде алгоритм создания корпуса выглядит следующим образом.

На первом этапе осуществляется морфологическое аннотирование текстов с последующим формированием тезауруса для данного текстового массива; следующий блок текстов получает аннотирование уже в автоматизированном режиме путем соотнесения словоформ текста со словоформами тезауруса; новые словоформы вручную заносятся в текст с необходимой разметкой и аннотированием. Последовательность данных операций повторяется при обработке последующих массивов текстов.

Реализация каждого из данных действий требует решения ряда как теоретических, так и сугубо практических задач. Так, одним из основных условий является перевод корпуса текстов в цифровой формат с их последующей обработкой; маркированность текста в зависимости от его происхождения, тематики и иных формальных признаков, реализация действия поисковых запросов по грамматическим глоссам и т. д.

Новые компьютерные технологии и разработанные на их базе методики обработки лингвистического материала позволяют по-новому подойти к решению проблемы формирования корпусов текстов, а также создавать массивы текстов, кото-

рые можно использовать в ходе как теоретических изысканий, так и практических разработок. В частности, корпус текстов может быть подвержен различным видам дальнейшей цифровой обработки, одной из целей которой является частотный анализ языковых единиц, позволяющий выявить их разноуровневое актуальное распределение.

### Методы формирования частотных словарей на примере материала арабского языка

Частотный анализ текстового материала, отобранного по определенному признаку, или иначе, корпуса текстов, широко применяется в самых разных отраслях научного знания, таких как лексикография; прикладная лингвистика; компьютерная лингвистика; социо- и психолингвистика, литературоведение и ряде других.

Анализ лексического материала предполагает учет таких показателей, как частотность вхождений словоформ и соответствующих им корней в текстовый материал, т. е. возведение словоформ к корневой основе и выделение списка приоритетных слов в зависимости от их индекса частотности.

Как и при формировании корпусов текстов, при определении индекса частотности используются получившие широкое распространение современные технологии цифровой обработки материалов естественных языков, что облегчает задачу формирования максимально репрезентативной выборки текстов для анализа частотного вхождения в них лексических единиц, однако многие из числа такого рода технологий оказываются не вполне эффективны применительно для арабского материала.

### Компьютерные технологии и обработка арабского текста

Разработка и оптимизация принципов определения частотности вхождения лексических единиц в письменные тексты на арабском литературном языке включают в себя решение следующих задач:

1) сбор лексического материала в виде письменных текстов с их дальнейшей оцифровкой или использованием текстов в цифровом формате;

2) разработка принципов формализации лексических единиц и сведения их к словарной форме и корневой морфеме;

3) определение частотности вхождения лексических единиц в состав анализируемых текстов.

Рассмотрим методику составления частотного словаря словоформ на основе компьютерного процессинга арабского текста с целью определения частотности вхождения языковых единиц.

Определение индекса частотности арабской лексики было проведено на базе сформированного корпуса текстов на литературном языке, результатом проведенной работы было создание арабского частотного словаря. В анализируемый корпус текстов были включены фрагменты, имеющиеся в цифровом формате, либо переведенные в таковой после сканирования. Корпус включал различные по тематике тексты (общественно-политические, художественные, научные, учебные и т. д.), что позволило учесть максимально возможный инвентарь лексики и объективно оценить частотность ее использования, равно как и региональную маркированность.

В результате был сформирован корпус текстов объемом около одного миллиона лексических единиц, с оформлением в виде тематических разделов, что в дальнейшем может быть использовано для формирования специализированных тематических словарей.

Даже предварительный анализ показывает степень сложности классификации слов в зависимости от индекса частотности на базе значительных объемов лексики, что связано как с особенностями арабской графики, так и богатством арабской морфологической парадигмы, в рамках которой возможна реализация большого числа словоформ, производных от отдельно взятого корня.

Помимо возможностей слитного написания ряда частиц и предлогов, в арабском языке существует большое количество факультативно реализуемых графических символов — знаков для гласных, знаков геминации, знаков слитного произношения — васлы (ا вместо ٱ), знаков для долготы произношения алефа — мадды (ا вместо آ), а также некоторых согласных, например, хамзы (ا вместо أ, و вместо ؤ, ى вместо ئ, ا вместо ء), опущение в ряде случаев точек у конечного ي или «добавление» точек у конечного алифа максура — ي вместо ى. Список факультативных вариантов еще более расширится, если принять во внимание написание знаков, используемых при таджвидном чтении Корана, например, ۜ, ۧ, ۫, ۙ, ۚ, ۛ, указывающих на необходимость пауз и реализации других особенностей чтения.

Столь широкий спектр вариативности еще более увеличивает и без того значительное число потенциально возможных форм — дериватов от одного корня, количество которых нередко превосходит несколько тысяч. Так, для имен (масдаров, причастий, прилагательных, имен места и времени действия) это обусловлено наличием, как это было отмечено, элементов, пишущихся слитно в пре- или постпозиции к слову, а также инфиксов. С учетом факультативных вариантов написания, даже без учета различий в огласовках, число графических вариантов словоформ весьма велико. Так, например, насчитывается (без учета огласовок) 208 графических репрезентаций производных форм слова باب «дверь», а с наличием факультативных вариантов написания, но без учета омографов, соответствующих различным формам склонения, их число достигает 121.

Что касается глагольных форм, то число их еще более велико, чем именных, и обусловлено развитой системой глагольной парадигмы. В глагольной, как и в именной парадигме, помимо единственного и множественного числа, присутствуют также формы двойственного числа, а также пять наклонений. Так, только с учетом форм действительного залога общее число форм парадигмы правильного глагола достигает 166, а с учетом производных пород количество глагольных форм может доходить до нескольких тысяч единиц.

В ходе работы по определению индекса частотности словоформы возводились к словарной ячейке, для чего были созданы алгоритмы и на их основе макросы, позволяющие производить данную операцию. Наиболее успешно данная операция проходила с незначительным числом слов (до ста единиц), более значительные объемы лексики требовали большего объема времени и мощностей используемого оборудования.

В результате в ходе работы была использована комбинированная технология. Часть лексики возводилась к корневым морфемам в автоматическом режиме, а часть вручную. Как и следовало ожидать, вспомогательные слова и частицы ока-

**Перечень ста наиболее употребительных лексем в арабском языке**

| № п.п. | Лексема | Количество вхождений |
|---|---|---|
| 1 | و | 43501 |
| 2 | في | 31047 |
| 3 | من | 24603 |
| 4 | على | 10767 |
| 5 | أن | 8716 |
| 6 | التي | 6999 |
| 7 | إلى | 5256 |
| 8 | هذا | 5001 |
| 9 | ما | 4851 |
| 10 | عن | 4818 |
| 11 | ذي | 4257 |
| 12 | شمس | 4095 |
| 13 | هذه | 4095 |
| 14 | ان | 3825 |
| 15 | لا | 3813 |
| 16 | مع | 3525 |
| 17 | ذلك | 3288 |
| 18 | علي | 3267 |
| 19 | قمر | 2856 |
| 20 | هو | 2799 |
| 21 | الى | 2580 |
| 22 | كان | 2484 |
| 23 | كل | 2463 |
| 24 | خلال | 2349 |
| 25 | بين | 2283 |
| 26 | لم | 2262 |
| 27 | الله | 2166 |
| 28 | بعد | 2139 |
| 29 | قال | 2091 |
| 30 | كما | 2073 |
| 31 | أو | 2046 |
| 32 | عام | 1274 |
| 33 | أمير | 3822 |
| 34 | حيث | 1830 |
| 35 | هي | 1773 |
| 36 | إن | 1659 |
| 37 | قد | 1659 |
| 38 | كانت | 1608 |
| 39 | إلي | 1566 |
| 40 | مليون | 1497 |
| 41 | قبل | 1422 |
| 42 | لكن | 1326 |
| 43 | غير | 1314 |

*Продолжение таблицы*

| № п.п. | Лексема | Количество вхождений |
|---|---|---|
| 44 | ولا | 1287 |
| 45 | مجلس | 1230 |
| 46 | بعض | 1185 |
| 47 | هناك | 1155 |
| 48 | تم | 1149 |
| 49 | شركة | 1131 |
| 50 | يا | 1125 |
| 51 | اليوم | 846 |
| 52 | يمكن | 843 |
| 53 | شركات | 843 |
| 54 | لها | 831 |
| 55 | مثل | 819 |
| 56 | أنا | 810 |
| 57 | شركات | 807 |
| 58 | قطاع | 807 |
| 59 | منها | 804 |
| 60 | أميركية | 804 |
| 61 | عامة | 792 |
| 62 | أنت | 786 |
| 63 | ليس | 783 |
| 64 | مشروع | 777 |
| 66 | خاصة | 774 |
| 67 | دول | 762 |
| 68 | مما | 759 |
| 69 | اقتصادية | 753 |
| 70 | جميع | 747 |
| 71 | سوق | 744 |
| 72 | يكون | 741 |
| 73 | شركة | 738 |
| 74 | نعم | 735 |
| 75 | إلا | 732 |
| 76 | فيه | 729 |
| 77 | إذا | 726 |
| 78 | أنه | 720 |
| 79 | ثم | 720 |
| 80 | عراقية | 720 |
| 81 | أكثر | 708 |
| 82 | بما | 708 |
| 83 | وزارة | 708 |
| 84 | بن | 696 |
| 85 | عم | 696 |
| 86 | تي | 693 |
| 87 | عدد | 690 |
| 88 | به | 681 |

*Окончание таблицы*

| № п.п. | Лексема | Количество вхождений |
|---|---|---|
| 89 | الي | 681 |
| 90 | منذ | 678 |
| 91 | قطاع | 669 |
| 92 | مساعد | 669 |
| 93 | كذلك | 660 |
| 94 | نحو | 660 |
| 95 | مشروع | 654 |
| 96 | زيادة | 651 |
| 97 | عربية | 648 |
| 98 | ماضي | 648 |
| 99 | ثاني | 645 |
| 100 | تابع | 639 |

зались наиболее частотными. Одними из самых частотных оказались лексические единицы, имеющие нейтрально-стилистическую семантику, такие как شمس «солнце» и قمر «луна»[2], что весьма неожиданно. Отмечена и корреляция индекса частотности лексики с текущими политическими событиями (частотность топонимов и этнонимов, повторяемых в новостях: مبارك «(президент) Мубарак»; إسرائيل «Израиль» и т. д.). Пропорциональное соотношение между наиболее часто встречавшимися лексическими единицами и наиболее редко составляло 43 501 к одному.

Проведенная работа позволила идентифицировать у каждой лексической единицы ее корневую морфему и частотный показатель вхождения в состав рассмотренного текстового материала. Зависимости числа вхождений наиболее частотных лексем от тематики текстов отмечено не было.

При рассмотрении анализируемого массива текстов и спектра их тематики основной список наиболее частотных слов оставался практически неизмененным. Таким образом, удалось выделить наиболее характерную лексику письменных текстов литературного арабского языка, имеющую универсальный характер, часть ее возвести к корневой основе.

Суммируя сказанное выше, следует отметить, что успешная разработка технологий автоматизированной обработки материала арабского, равно как и других восточных языков, требует объединения усилий как лингвистов-востоковедов, так и специалистов в области программного обеспечения.

[2] Показатели частотности вхождений ста наиболее употребительных словоформ представлены в таблице.

**Литература**

1. *AbdelRaouf A., Higgins C. A., Pridmore T., Khalil M.* Building a multi-modal Arabic corpus (MMAC) // International Journal on Document Analysis and Recognition. 2010. Vol. 13 (Dec., 2010), N 4. P. 285–302.

2. *Haslina H., Mat D. N., Atwell E. S.* Connectives in the World Wide Web Arabic Corpus // World Applied Sciences Journal. 2013. Vol. 21 (Special Issue of Studies in Language Teaching and Learning). P. 67–72.

3. *Kilgarriff A., Rundell M., Dhonnchadha E. U.* Efficient corpus development for lexicography: building the New Corpus for Ireland // Language Resources and Evaluation. Vol. 40, N 2 (May, 2006). P. 127–152.

4. *Mansour M. A.* The Absence of Arabic Corpus Linguistics: A Call for Creating an Arabic National Corpus // International Journal of Humanities and Social Science. Vol. 3, N 12 (Special Issue — June 2013). P. 83–84.

5. *Hammo B., Abuleil S., Lytinen S., Evens M.* Experimenting with a Question Answering System for the Arabic Language // Computers and the Humanities. Vol. 38, N 4 (Nov., 2004). P. 397–415.

6. *Ferguson Ch.* Diglossia // Word. 1959. N 15. P. 325–340.

Статья поступила в редакцию 24 октября 2013 г.

Контактная информация

*Редькин Олег Иванович* — доктор филологических наук, профессор; oleg_redkin@mail.ru

*Redkin Oleg I.* — Doctor of philological sciences, Professor; oleg_redkin@mail.ru