Ученые записки Крымского федерального университета имени В. И. Вернадского.
Серия «Филологические науки». Том 2 (68), № 3. 2016 г. С. 391–396.

УДК 811.222.1’06

ПРИКЛАДНАЯ ЛИНГВИСТИКА В ИСЛАМСКОЙ РЕСПУБЛИКЕ ИРАН: ДОСТИЖЕНИЯ И ТЕНДЕНЦИИ РАЗВИТИЯ

Фаткулин Б. Г.
ФГБОУ ВПО Южно-Уральский государственный университет (НИУ)
bfatkulin@gmail.com

Статья рассказывает о последних тенденциях в развитии прикладной лингвистики в Исламской Республике Иран. Исследователем подробно раскрываются задачи, решаемые иранской школой прикладной лингвистики, анализируются лингвистические ресурсы языка фарси. Автор статьи считает развитие лингвистических ресурсов и прикладных инструментов частью языковой политики государства, проводит обзор государственных и научных учреждений ИРИ, использующих достижения прикладной лингвистики и являющихся заказчиками научных исследований в этой сфере.

Ключевые слова: Иран, персидский язык, прикладная лингвистика, лингвистические ресурсы, электронные словари, тезаурусы, корпуса персидского языка.

ВВЕДЕНИЕ

Обладая высокой степенью политического, культурного и научного суверенитета, Иран в настоящее время является наиболее стабильной страной Среднего и Ближнего Востока Среди стран региона Иран отличается также высокими темпами развития науки. Международные санкции, объявленные Ирану рядом стран с целью изоляции страны, привели к тому, что Иран нарастил темпы собственных научных разработок и занял в регионе ключевые позиции во многих научных отраслях. По уровню информатизации Иран занимает ведущее место в ряду ближневосточных и средневосточных стран. Иранские специалисты в области информации работают в рамках операционных систем Windows, Unix, Linux (проекты в рамках открытого программного обеспечения), MacOS и Android, регулярно публикуют результаты своих исследований [10; 11; 12; 13; 14; 15]. Репозитории программного обеспечения (Github, sourceforge.net, PyPI – the Python Package Index и т. п.) содержат множество сценариев и скриптов для обработки текстов на персидском языке, включающей токенизацию, лемматизацию, парсинг и другие функции.

В изучении и преподавании языка фарси, а также в информационно-аналитической работе с иранскими источниками информации российскими иранистами пока в большинстве случаев используются классические, проверенные временем методы и приемы ручной выборки информации, требующие больших временных затрат. Однако время не стоит на месте, и российские иранисты постепенно овладевают современными методами и инструментарием, основанным на достижениях современной прикладной лингвистики.

В номенклатуре наук «прикладная лингвистика» – это деятельность по приложению научных знаний об устройстве и функционировании языка в нелингвистических научных дисциплинах и в различных сферах практической деятельности человека.

Исламская Республика Иран является одним из передовых центров по прикладной лингвистике на уровне региона. Наличие собственной научной школы прикладной лингвистики является частью языковой политики и залогом культурного суверенитета страны.

Персидский язык является одним из ключевых языков Востока и одновременно с этим входит в индоевропейскую языковую семью. Поэтому многие алгоритмы, используемые в обработке современных европейских языков (английского, немецкого, французского и др.), с успехом могут быть экстраполированы и на персидский язык. Письменность на основе арабского алфавита отличает персидский язык от других индоевропейских языков, однако современная прикладная лингвистика успешно справилась и с этой задачей, создав множество текстовых кодировок, учитывающих эту особенность персидского языка, например кодировку UTF-8. Овладение компетенциями прикладной лингвистики позволяет современному иранисту избежать множества рутинных операций, работать с большими массивами оцифрованной информации на персидском языке, обогатить свои исследования количественно-статистическими подходами. Исследователь, работающий на стыке иранистики и прикладной лингвистики должен обладать следующими знаниями и умениями:

1) знать основные лингвистические ресурсы персидского языка;

2) уметь работать с корпусами персидского языка;

3) понимать принципы основных алгоритмов по обработке естественных языков;

4) уметь пользоваться программным обеспечением для обработки текстов на персидском языке в различных операционных системах.

По месту расположения научные школы прикладной лингвистики в области иранских языков делятся на две группы – внешнюю и внутреннюю [2]:

1) внешние научные школы расположены вне пределов Ирана: на востоковедческих отделениях ведущих мировых университетов, в соответствующих силовых ведомствах крупных государств, а также в транснациональных корпорациях;

2) внутренняя иранская школа прикладной лингвистики в области иранских языков – прикладная лингвистика, которую иранские ученые сами развивают на своей территории. (Иранские прикладники обрабатывают и другие мировые языки (английский, русский, арабский, китайский и т. д.).

Иранское государство ставит перед прикладной лингвистикой следующие задачи:

1) сохранение культурного наследия персидского языка, создание электронных корпусов языка фарси, корпусов классических произведений текстов на персидском языке;

2) оцифровка источников по шиитскому направлению в исламе;

3) накопление и оцифровка лингвистических ресурсов персидского языка, создание алгоритмов и прикладных программ для обработки языка фарси и других языков;

4) создание и хранение лингвистических ресурсов на языке фарси;

5) разработка стандартов для электронного документооборота на языке фарси;

6) структурирование данных на языке фарси.

Проводниками государственной политики Ирана в области прикладных лингвистических исследований выступают специализированные организации, к примеру:

1) Научно-исследовательский институт информационных наук и технологий Ирана (IRANDOC) http://www.irandoc.ac.ir/;

2) Iranian Journal of Applied Linguistics Studies (IJALS). Этот журнал издается в сотрудничестве с Лингвистическим обществом Ирана, а также Университетом Уппсала, одним из главных центров исследований по иранским языкам в Европе;

3) Иранское лингвистическое общество (основано в 2001 году) было признано Министерством науки, исследований и технологий в качестве академического общества в 2004 году, ее цели и деятельность Ирана включают в себя развитие языкового и культурного исследования, сотрудничество с научно-исследовательскими центрами, проекты, связанные с лингвистикой и изучением языка, предоставление образовательных, научно-исследовательских и технических услуг на национальном и международном уровнях, организацию и проведение местных, региональных и всемирных конференций и издание книг, журналов и информационных бюллетеней.

М. Фаал-Хамеданчи в своей статье [Ошибка: источник перёкрестной ссылки не найден], написанной в 2010 году, подробно рассматривает развитие лексикографической деятельности в современном Иране с точки зрения прикладной лингвистики. Информация, данная М. Фаал-Хамеданчи, дополнена и актуализирована в диссертации выпускника Уппсальского университета Можгана Сираджи, в которой он проводит обзор современного программного обеспечения в области обработки персидского языка [Ошибка: источник перёкрестной ссылки не найден].

Лингвистические ресурсы языка фарси

Прикладная лингвистика оперирует таким понятием, как «лингвистические ресурсы». Лингвистические информационные ресурсы – это множество определенным образом организованных речевых и языковых данных, находящихся на машинных носителях информации и используемых в различных сферах практической деятельности (образовании, промышленности, экономике, культуре, искусстве и т. д.).

Выделяют активные и пассивные лингвистические ресурсы. К пассивным формам относят письменный лексикон, терминологические словари, письменные текстовые массивы (корпуса текстов), фонетические ресурсы, электронные библиотеки и т. д., к активным – алгоритмы, модели, программы, базы знаний.

Проблемам создания лингвистических ресурсов ежегодно посвящается большое количество научных конференций во всем мире. Создан ряд организаций, занимающихся разработкой лингвистических ресурсов: LDC (Linguistic Data Consortium, USA), ELRA (European Language Resources Association), TELRI (TransEuropean Language Resources Infrastructure). Перед этими организациями стоят следующие задачи:

1) разработка единых стандартов создания ресурсов;

2) разработка способов защиты от несанкционированного доступа;

3) создание единых экспертных требований;

4) планирование единой стратегии разработки лингвистических ресурсов;

5) создание многофункциональных лингвистических ресурсов большого объема для использования в разных странах.

Перечисленные выше задачи стоят и перед иранскими научными структурами. М. Фаал-Хамеданчи в своей статье [3] перечисляет примеры лингвистических ресурсов персидского языка, созданных в Иране. В качестве дополнения к списку М. Фаал-Хамеданчи мы можем привести Корпус Бижанхан (Bijankhan) – размеченный корпус, который содержит ежедневные новости и тексты на общую тематику (4300 различных тематических подразделов). Этот корпус был создан в Университете Тегерана (Database Research Group).

Программные продукты для обработки персидского языка могут быть представлены в различных видах:

1) программы с графическим интерфейсом (для пользователей ОС Windows);

2) программы для командной строки (для пользователей ОС UNIX и многочисленных разновидностей Linux);

3) модули для включения в скрипты на языках программирования (для продвинутых специалистов, использующих языки программирования, например Python).

В качестве примера результатов использования инструментов прикладной лингвистики для обработки текстов на языке фарси мы можем привести библиотеку hazm 0.5.2. Python (NLTK совместимая и поддерживает язык Python версий 2.7, 3.2, 3.3, 3.4 и 3.5) для обработки текстов на персидском языке, которая выполняет следующие функции:

1) нормализация текста (устранение лишних символов);

2) токенизация слов на уровне предложений (отделение слов друг от друга и представление их в виде списков);

3) лемматизация слов (отделение от слова словообразовательных частиц): сведение слова к его корню;

4) низкоуровневый анализ предложений;

5) синтаксический анализ и формирование диаграммы зависимостей;

6) предоставление интерфейса для просмотра корпусов текстов на персидском языке.

Мы апробировали эту библиотеку в командной строке в языке Python и получили следующий вывод:

```
>>> from __future__ import unicode_literals
>>> from hazm import *
>>> normalizer = Normalizer ()
>>> normalizer.normalize('اصلاح نويسه ها و استفاده از نيم‌فاصله پردازش را آسان مي كند')
'اصلاح نویسه‌ها و استفاده از نیم‌فاصله پردازش را آسان می‌کند'
```

```
>>> sent_tokenize('ما هم برای وصل کردن آمدیم! ولی برای پردازش، جدا بهتر نیست؟')
['ما هم برای وصل کردن آمدیم!', 'ولی برای پردازش، جدا بهتر نیست؟']
>>> word_tokenize('ولی برای پردازش، جدا بهتر نیست؟')
['ولی', 'برای', 'پردازش', '،', 'جدا', 'بهتر', 'نیست', '؟']
>>> stemmer = Stemmer()
>>> stemmer.stem ('کتاب‌ها')
'کتاب'
```

ВЫВОДЫ

Российские иранисты, работающие на стыке иранистики и прикладной лингвистики, должны установить и поддерживать контакты с иранскими коллегами-прикладниками, активно участвовать в научных конференциях на территории Ирана, печататься в иранских научных журналах и повышать уровень взаимоцитирования с иранскими специалистами. Прикладная лингвистика дает российским иранистам мощный инструментарий для повышения эффективности и качества исследований.

Список литературы

1. Гладкова Е. Л. Школы преподавания индоиранских и африканских языков // Вестник МГИМО Университета. 2014. Т. 5 (38).

2. Фаал-Хамеданчи М. Современная лексикографическая деятельность в Иране // Вестник Российского университета дружбы народов. Серия: Лингвистика. 2010. Т. 1.

3. Фаткулин Б. Г. Использование лингвистически ориентированных модулей на языке Python для обработки больших текстовых массивов на восточных языках в целях эффективного сбора и обработки данных по отраслям востоковедческой тематики (на примере NLTK) // Вестник Южно-Уральского государственного университета. Серия: Лингвистика. 2015. Т. 12, № 1. – С. 72–75.

4. Фаткулин Б. Г. Современные методы извлечения терминологии и представления выборки терминов в виде базы данных (на примере исламоведческой отрасли современного китайского языка) // Судьбы национальных культур в условиях глобализации: сборник материалов III Международной научной конференции / Под ред. Смирнова М. Челябинск: Энциклопедия, 2015. – С. 395–398.

5. Фаткулин Б. Г. Использование современных инструментов прикладной лингвистики для извлечения терминологических единиц (на примере обработки раздела «Афганистан» в китайской он-лайн энциклопедии Baidu) // Профессиональный проект: идеи, технологии, результаты. 2013. Т. 3, № 12. – С. 143–147.

6. Фаткулин Б. Г. Прикладная лингвистика на службе китаеведения: автоматизация загрузки контента на заданную тему из китайской энциклопедии «БАЙДУ БАЙКЕ» с помощью специального программного обеспечения в рамках операционной системы LINUX // Россия и Китай: история и перспективы сотрудничества. Материалы V Международной научно-практической конференции. Ответственные редакторы Д. В. Буяров, Д. В. Кузнецов, Н. В. Киреева. 2015. Благовещенский государственный педагогический университет (Благовещенск), 2015. – С. 374–378.

7. Esfahbod B., Pournader R. FarsiTeX and the Iranian TeX Community // TUG 2002. 2002. Т. 23, № 23-1. – С. 41–45.

8. Megerdoomian K., Parvaz D. Low-density language bootstrapping: The case of Tajiki Persian // Proceedings of LREC 2008 (Language Resources and Evaluation Conference). Marrakech, Morocco, 2008.

9. Dolamic L., Savoy J. Ad Hoc Retrieval with the Persian Language // Lect. Notes Comput. Sci. Springer, 2010. Т. 6241/2010. – С. 102–109.

10. Shamsfard M., Barforoush A. A. An Introduction to Hasti: An Ontology Learning System // Proceedings of the 6th Conference on Artificial Intelligence and Soft Computing. 2002.

11. Evaluation of Part of Speech Tagging on Persian Text // Proceedings of the 2nd Workshop on Computational Approaches to Arabic Script-based Languages / Под ред. Farghaly A., Megerdoomian K. Stanford, California, USA, 2007.

12. Seraji M. Morphosyntactic Corpora and Tools for Persian. Uppsala University, Department of Linguistics and Philology, 2015. № 16. – 191 с.

13. Bijankhan M. и др. Lessons from building a Persian written corpus: Peykare // Lang. Resour. Eval. 2011. Т. 45, № 2. – С. 143–164.