Электронный языковой корпус как фактор сохранения мордовских (мокшанского и эрзянского) языков

Мария З. Левина

Национальный исследовательский
Мордовский государственный университет имени Н. П. Огарёва,
г. Саранск, Российская Федерация,
Lev.Mariya@mail.ru

Аннотация. Систематизации языкового материала и создания электронного корпуса способствует не только сохранению мордовских (мокшанского и эрзянского) языков, но и имеет особую значимость для финно-угорского языкознания в целом. Электронные ресурсы создадут более широкие возможности для проведения сравнительно-исторических, сопоставительных и типологических исследований.

Ключевые слова: мордовские (мокшанский и эрзянский) языки, языковой корпус, подкорпус, корпусная лингвистика.

Abstract. The systematization of the linguistic material and the creation of an electronic corpus contributes not only to the preservation of the Mordvin (Moksha and Erzyan) languages, but also has a special significance for the Finno-Ugric linguistics in general. Electronic resources will create wider opportunities for conducting comparative-historical, contrastive and typological research.

Keywords: Mordvin (Moksha and Erzya) languages, language corpus, subcorpus, corpus linguistics.

Электронные языковые корпуса в настоящее время являются одним из самых важных инструментов в сохранении языков малых народов и служат основой раздела прикладной лингвистики, известного как корпусная лингвистика.

Корпусная лингвистика – раздел языкознания, занимающийся разработкой, созданием и использованием текстовых корпусов [https://ru.wikipedia.org/wiki/Корпусная_лингвистика]. В современном ее понимании как наука, занимающаяся созданием и анализом текстовых корпусов, зародилась в США и Западной Европе в конце 1960-х гг. С ростом возможностей современных компьютерных технологий, с середины 1980-х гг. корпусная лингвистика получает бурное развитие, стали активно появляться корпусные проекты различных масштабов на разных языках и для разнообразных целей.

Достижения в области корпусной лингвистики находят широкое применение в процессе изучения и преподавания языкознания. В ведущих вузах мира становится повседневной практикой использование корпусных данных в качестве эмпирической составляющей проектов и отдельных курсов.

Корпусный подход оптимален для наглядного представления таких аспектов языка, как историческая, географическая и социальная вариация и изменения в языковой системе, параллельно давая живую возможность овладеть базовыми принципами корпусных методов лингвистического анализа.

В лингвистике *корпус* – совокупность текстов, собранных и обработанных в соответствии с определенными принципами, размеченных по определённому стандарту, обеспеченных специализированной поисковой системой, используемых в качестве базы для исследования языка [https://ru.wikipedia.org/wiki/Корпус_текстов].

К настоящему времени для большинства крупных языков мира уже созданы так называемые электронные языковые ресурсы. Общеизвестным образцом является, в частности, Британский национальный корпус (British National Corpus (BNC)), на который ориентированы и многие другие корпуса, Чешский национальный корпус, Национальный корпус русского языка.

Ведется создание электронных ресурсов и отдельных финно-угорских языков. Так, *Corpus of Estonian Written Texts* содержит эстонские письменные тексты 1983–1987 гг., художественную литературу, научные и научно-популярные тексты и др. [Некрасова 2010 : 14] .

Для венгерского, например, уже к началу 2000-х гг. были созданы первая версия сбалансированного Национального корпуса [Váradí 2002]; меньший по объёму, но синтаксически аннотированный Сегедский корпус [Csendes et al. 2004]; корпус веб-страниц на венгерском объёмом около 2 млрд словоформ [Halácsy et al. 2004]; исторический корпус [Pajzs 2000] и др.

В 2014-2015 годах были созданы первые версии общедоступных письменных корпусов коми (<http://komicorpo.ru>), созданный командой FU-Lab под руководством М. С. Фединой, удмуртского, созданный М. Медведевой и Т. Архангельским и марийских языков [Bradley 2015]. Каждый язык представлен в виде двух корпусов: корпус текстов социальных сетей и корпус всех остальных текстов. Оба вида текстов подверглись автоматическому морфологическому анализу; кроме того, тексты из соцсетей прошли дополнительные фильтрацию и анонимизацию, снабжены поисковым веб-интерфейсом и являются общедоступными [Архангельский 2019 : 528 <http://volgakama.web-corpora.net/>].

Возможности использования письменного корпуса коми языка для лингвистических исследований рассматриваются в работе Фединой М. С. «Корпус коми языка как база для научных исследований» [Федина 2019 : 45-48].

Следует отметить, что над мордовскими (мокшанским и эрзянским) языками также ведется работа по созданию электронных корпусов. Были разработаны корпуса интернет-текстов мокшанского и эрзянского языков https://www.researchgate.net/publication/325447073_Towards_an_open-source_universal-dependency_treebank_for_Erzya, созданные Джеком Рютером [Rueter, Tyers 2018], которые способствуют сохранению не только языкового

богатства мордвы-мокши и мордвы-эрзи, но культуры мордовского народа. Как отмечает сам ученый, главной задачей является сбор и систематизация текстов на данных языках.

Языковой материал по мордовским языкам представлен в базе GiellaLT http://gtweb.uit.no/u_korp/ для лингвистического, сравнительно-исторического сопоставительного аспекта изучения данных языков.

Осуществляется разработка онлайн-словаря по мокшанскому и эрзянскому языкам с открытым исходным кодом на основе Mediawiki в работе «Synchronized Mediawiki based analyzer dictionary development» <https://www.aclweb.org/anthology/W17-0601/> [Rueter, Härmäläinen 2017 : 1–7]. Разработка FST с открытым исходным кодом для уральских языков и языков других меньшинств демонстрирует огромный потенциал для дополнительных разработок, связанных с историей развития и сохранения языков.

Сервер korp.csc.fi в Финляндии предоставляет множество разновидностей текстовых корпусов для больших и малых языков. Онлайн-корпуса мокшанского и эрзянского языков представлены в работе «On New Text Corpora For Minority Languages On The Helsinki korp.csc.fi Server» <https://researchportal.helsinki.fi/en/publications/on-new-text-corpora-for-minority-languages-on-the-helsinki-korpcs> [Rueter, Partanen 2019 : 32–36].

Как было отмечено авторами, мокшанский и эрзянский языки находятся под угрозой исчезновения языков меньшинств, поэтому следует пристальное внимание обратить на сохранение языкового материала. Разнообразие двух родственных языков — мокшанского и эрзянского, морфологическое единство данных языков описаны в работе «Open-Source Morphology for Endangered Mordvinic Languages» <https://www.aclweb.org/anthology/2020.nlpss-1.13/>. Характерные особенности этих двух языков мотивируют для совместной разработки FST с открытым исходным кодом [Rueter, Härmäläinen, Partanen 2020 : 94–100].

Электронный языковой корпус мордовских (мокшанского и эрзянского) языков — это коллекция электронных текстов, собранных в соответствии с определенными принципами, снабженных лингвистической информацией, размеченных по определенному стандарту и обеспеченных специализированной поисковой системой. Его основная задача заключается в представлении каждой разновидности языка.

Электронный корпус является одним из главных проектов по сохранению и развитию мордовских (мокшанского и эрзянского) языков, включающей в себя два основных вида электронных ресурсов — корпус текстов и словарь. Данная проблема обосновывается тем, что с течением времени носителей мордовских (мокшанского и эрзянского) языков становится все меньше и языковой материал может остаться не исследованным в полном объеме. Это связано со стремительными процессами глобализации, которые размывают этнические особенности, в связи с чем, современные языки все более сближаются, одновременно парадоксально взаимообогащаются и беднеют.

Электронная форма представления языкового материала повысит сохранность собранного уникального материала, обеспечивает условия для лингвистов

различной специализации более свободного доступа к языковому материалу, позволяет наблюдать реальные отношения между языковыми единицами в потоке речи, при минимальных затратах усилий самостоятельно создавать полные базы данных в соответствии со своими исследовательскими задачами, классифицировать материал на основании отдельных параметров и их комплексов. Электронные ресурсы должны стать источником исследовательских проектов, программно-методических комплексов в общеобразовательной и вузовской системах преподавания. Они создадут более широкие возможности для проведения сравнительно-исторических, сопоставительных и типологических исследований по финно-угорским языкам.

Таким образом, создание электронного корпуса мордовских (мокшанского и эрзянского) языков открывает широкие возможности для исследований, так как современные информационные технологии упрощают и ускоряют процедуры лингвистической обработки больших текстовых материалов. Электронный корпус мордовских языков – это реальный инструмент сохранения всего письменного наследия мордовского народа в виде цифровых копий, передачи будущим поколениям культуры народа.

Список литературы

1. Агафонова Н. А. Диалектологический атлас как результат исследований эрзянских диалектов методами лингвистической географии / Н. А. Агафонова, И. Н. Рябов // Финно-угорский мир. – 2013. – № 2. С. 48–53.
2. Архангельский Т. А. Интернет-корпуса финно-угорских языков России / Т. А. Архангельский // Ежегодник финно-угорских исследований. Ижевск, 2019. Т.13, №3. С. 528–537. <http://volgakama.web-corpora.net/>
3. Левина М. З. Из истории классификации диалектов мокшанского языка / М. З. Левина // Финно-угорский мир. Саранск, 2015, № 3. С. 24–35.
4. Левина М. З. Морфология форм неопределённого склонения имени существительного в мокшанских диалектах Поволжья (лингвогеографический аспект исследования) / М. З. Левина // Вопросы языкознания. Москва, 2015. – № 5. С. 104–119.
5. Некрасова А. Г. Электронный диалектный корпус как ресурс сохранения и изучения коми диалектов / А. Г. Некрасова // Финно-угорский мир. Саранск, 2010. №1. С. 13–16.
6. Федина М. С. Финно-угорские языки Российской Федерации в электронном информационном пространстве: опыт, проблемы и перспективы / М. С. Федина // Финно-угорский мир. 2016. Т. 3. № 28. С. 111–121.
7. Федина М. С. Корпус коми языка как база для научных исследований // Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы : материалы II Международной научной конференции. Уфа: госуд.автон.учр.науки Республики Башкортостан "Башкирская энциклопедия", 2019. С. 45–48.
8. Феоктистов А. П. Диалекты мордовских языков / А. П. Феоктистов // Paasonens H. Mordwinisches Wörterbuch. Bd. I: A—J. Suomalais-ugrilainen seura. 1990. С. LX–LXXXVI.
9. Halácsy P., Kornai A., Németh L., Rung A., Szakadát I., Trón V. Creating open language resources for Hungarian. LREC 2004 Proceedings. 2004. Pp. 203–210.

10. Rueter J., Härmäläinen M. Synchronized Mediawiki based analyzer dictionary development // Третьего семинара по компьютерной лингвистике для уральских языков. СПб.: Ассоциация компьютерной лингвистики, 2017. С 1–7. <https://www.aclweb.org/anthology/W17-0601/>
11. Rueter J., Partanen N. On New Text Corpora For Minority Languages On The Helsinki korp.csc.fi Server // Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы. Уфа, 2019. С. 32–36.
12. Rueter J., Tyers F. Towards an open-source universal-dependency treebank for Erzya // Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages. 2018. С. 108–120.
https://www.researchgate.net/publication/325447073_Towards_an_open-source_universal-dependency_treebank_for_Erzya
13. Rueter J., Härmäläinen M., Partanen N. Open-Source Morphology for Endangered Mordvinic Languages // Труды второго семинара по программному обеспечению с открытым исходным кодом NLP (NLP-OSS): Ассоциация компьютерной лингвистики, 2020. С. 94–100 <https://www.aclweb.org/anthology/2020.nlposs-1.13/>
14. Zeman, Daniel; Nivre, Joakim; Abrams, Mitchell, Levina, Maria et al. Universal Dependencies 2.6, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. 2020-05-15. Type corpus <http://hdl.handle.net/11234/1-3226>
15. Váradi T. The linguistic relevance of corpus linguistics. Proceedings of the Corpus Linguistics 2001 Conference. 2001. Pp. 587–593.
http://gtweb.uit.no/u_korp/
17. http://komicorpora.ru/Корпус_коми_языка
18. <http://korp.csc.fi/>
19. https://ru.wikipedia.org/wiki/Корпусная_лингвистика
20. https://ru.wikipedia.org/wiki/Корпус_текстов/
21. http://corpus.nytud.hu/mnsz/index_eng.html / Венгерский национальный корпус (HNC)