УДК 811.511.1:004.738.5 (045)

Т. А. Архангельский

ИНТЕРНЕТ-КОРПУСА ФИННО-УГОРСКИХ ЯЗЫКОВ РОССИИ[1]

Электронные языковые корпуса давно стали одним из самых важных инструментов в работе лингвиста и послужили основой для нового методологического направления, известного как корпусная лингвистика. В то время как для крупных европейских финно-угорских языков (венгерского, финского, эстонского) были созданы представительные корпуса, ситуация с финно-угорскими языками России до недавних пор была значительно хуже. В данной статье рассматриваются электронные корпуса, недавно разработанные автором для наиболее крупных финно-угорских языков России: удмуртского, коми-зырянского, лугового марийского, эрзянского и мокшанского. Тексты, доступные в электронном виде в Интернете, были собраны автором и специальным образом обработаны. Для каждого языка было создано два корпуса: корпусов 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. В данной работе будет рассмотрен процесс разработки корпусов и будут описаны их характеристики и возможности применения. Все описанные здесь корпуса снабжены поисковым веб-интерфейсом и являются общедоступными (http://volgakama.web-corpora.net/).

Ключевые слова: языковой корпус, корпусная лингвистика, социальные сети, финно-угорские языки, удмуртский, марийский, коми, эрзянский, мокшанский.

DOI: 10.35634/2224-9443-2019-13-3-528-537

В лингвистике под языковым корпусом в самом широком смысле понимается собрание текстов на каком-либо языке, специально предназначенное для изучения этого языка. Помимо собственно текстов, такие собрания обычно содержат дополнительную информацию, важную для исследований, – разметку, или аннотацию [McEnery, Hardy 2011, 13]. В зависимости от предназначения корпуса и от наличия технических инструментов для данного языка, корпус может содержать множество разных видов аннотации, однако в большинстве случаев среди них есть текстовые метаданные и морфологическая разметка. Дополнительными требованиями, часто предъявляемым к корпусам, являются сбалансированность и репрезентативность (см. дискуссию об этом вопросе в [Biber 1993; Váradi 2001; Leech 2007]), однако для большинства языков, кроме самых крупных, они трудновыполнимы.

Вероятно, первым языковым корпусом в современном смысле, хотя и не в современном техническом исполнении, стал так называемый Брауновский корпус американского английского [Francis, Kučera 1967]. За прошедшее с тех пор время корпусами, в том числе общедоступными, обзавелись многие языки. При этом по крайней мере в случае крупных языков для каждого из них существует множество разных корпусов, отличающихся содержанием, типами имеющейся разметки, способом доступа и т. п. Такова ситуация и с крупными европейскими финно-угорскими языками: венгерским, финским и эстонским. Для венгерского, например, уже к началу 2000-х гг. были созданы первая версия сбалансированного Национального корпуса [Váradi 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. Ситуация с финно-угорскими языками России, однако, совершенно иная. До середины 2010-х гг. создавались отдельные электронные текстовые коллекции небольшого размера на некоторых из этих языков (например, [Suihkonen 1998]) и небольшие устные корпуса, собранные в экспедициях. Однако общедоступных письменных корпусов в современном понимании ни для одного из этих языков не существовало. Такая же ситуация характерна для большинства языков России [Arkhangelskiy, Medvedeva 2016]. В 2014-2015 годах были созданы первые версии общедоступных письменных корпусов коми (http://komicorpora.ru/, созданный командой FU-Lab под руководством М. С. Фединой), удмуртского (созданный М. Медведевой и Т. Архангельским) и марийских языков [Bradley 2015]. Кроме того, были разработаны корпуса эрзянского и мокшанского языков, в которых на сегодняшний день нет мор-

[1] Работа была поддержана стипендией фонда Александра фон Гумбольдта.

фологической разметки[2] (созданы Джеком Рютером), а также ряд корпусов меньшего объёма в рамках проекта Universal Dependencies (см., например, [Rueter, Tyers 2018]).

В данной статье будут описаны корпуса интернет-текстов, разработанные автором для удмуртского, коми-зырянского, лугового марийского, эрзянского и мокшанского языков. Работа над этими корпусами проводилась в 2017–2019 гг. (за исключением удмуртского, работа над которым началась в 2014 г.).

Создание корпуса, предназначенного для лингвистических исследований, включает в себя несколько этапов: сбор и первоначальная обработка текстов, разметка текстов и размещение готового корпуса в поисковой системе (корпусной платформе). Ниже будут рассмотрены все эти этапы.

Сбор текстов и состав корпусов

Сбор текстов является наиболее времязатратным шагом при создании крупных корпусов. Составление репрезентативных корпусов требует включения в корпус разных жанров художественной литературы и публицистики разных временных периодов, а это, в свою очередь, требует гигантской работы по сканированию, оптическому распознаванию и вычитке текстов. В своей работе я изначально ориентировался на тексты, которые уже существуют в электронном виде в Интернете. Этот подход соответствует методологии «веб как корпус» [Kilgarriff, Grefenstette 2003], активно применяющейся для крупных языков. С одной стороны, такой подход делает получившиеся корпуса заведомо нерепрезентативными и ограниченными по объёму, поскольку в Интернете представлены только определённые жанры текстов, в основном написанные в 2000-х и 2010-х годах. С другой стороны, на сбор таких текстов требуется существенно меньше времени. Это позволило мне за ограниченное время подготовить однотипные корпуса для пяти разных языков. Кроме того, интернет-корпус даёт возможность прицельно изучать «цифровую жизнеспособность» языка (в терминах [Kornai 2016]) и на этих основаниях прогнозировать уровень сохранности языка в ближайшие десятилетия. Помимо данных о наличии какой-либо группы электронных ресурсов для каждого из этих языков (представленных для финно-угорских языков России в работе [Федина 2016]), интернет-корпуса позволяют получить более точную количественную и качественную оценку этих ресурсов.

Тексты для корпусов соцсетей и основных корпусов собирались по-разному. В обоих случаях первым шагом был поиск подходящих источников с помощью поисковых систем. Для каждого языка были отобраны несколько словоформ, которые очень частотны в этом языке, но не встречаются или не являются частотными в других языках. Например, для удмуртского такими словоформами были «ӧвӧл» и «сярысь», но не «со», «но» (совпадающие с частотными русскими словами) или «отын» (совпадающее с казахским словом). Этот подход ранее применялся Б. В. Ореховым и его коллегами для сбора интернет-текстов на разных языках России [Зайдельман и др. 2016], а до этого (в несколько другом виде) – К. Сканнеллом [Scannell 2007] в проекте «Crúbadán». Отличие моего подхода состоит в том, что все найденные источники я проверял вручную, отфильтровывая те, которые оказались в списке результатов поиска по ошибке. Ссылки на другие страницы, находящиеся в обнаруженных источниках, проверялись на наличие материалов на финно-угорских языках и при наличии таковых также включались в корпус. Тексты, которые были явно получены в результате некачественного оптического распознавания отсканированных изданий, в корпус не включались.

Загрузка самих текстов осуществлялась по-разному в разных случаях. В случае с текстами основного корпуса для каждого сайта писалась специальная программа-загрузчик, которая извлекала текст и другую необходимую информацию из HTML-кода страницы с помощью регулярных выражений. Загрузчики для разных страниц незначительно различались, поскольку их HTML-код и положение полезного текста в нём устроены по-разному. Поскольку новостные издания нередко публикуют сообщения на двух языках (на финно-угорском и на русском), каждая загружаемая статья проверялась простым определителем языка, разработанным ранее И. Ю. Самойленко на основе n-граммного метода, описанного в [Canvar, Trenkle 1994]; статьи на русском языке отсекались. В отдельных случаях тексты страниц копировались вручную. Для загрузки текстов из «Вконтакте» был написан скрипт[3], получающий на вход составленный вручную список адресов групп и пользователей и скачивающий

[2] Доступны через платформу Korp по адресу https://korp.csc.fi/?mode=other_languages#?lang=en&stats_reduce=word&cqp=%5B%5D&corpus=erme_mdf,erme_myv.

[3] Скрипт имеет открытый код и доступен по адресу https://bitbucket.org/timarkh/vk_texts_harvester.

все необходимые данные в формате JSON с использованием официального API ВКонтакте. Этот скрипт загружает все открытые посты и комментарии со страниц групп и со «стен» пользователей, независимо от того, на каком языке они написаны.

Состав основных корпусов разных языков оказался похожим. Основную часть корпуса занимает современная пресса (одно или несколько крупных периодических изданий; в некоторых случаях также несколько мелких). Существенно меньший объём имеют тексты с сайтов органов государственной власти, переводы Библии, научные статьи, публицистика и художественная литература. В основные корпуса были также включены блоги. Язык блогов далеко не всегда соответствует литературному стандарту и может содержать переключение кодов, большее количество русских заимствований, диалектизмы, незаконченные предложения и т. п. (см., например, оценку состояния удмуртской блогосферы в [Векшина 2016]). Однако как было показано ранее Т. Болдуином и др. [Baldwin et al. 2013], тексты блогов по разным характеристикам занимают промежуточное положение между редактированными текстами без отклонений от стандарта и текстами соцсетей. Мои предварительные наблюдения показывают, что это верно и в финно-угорском случае: по количеству переключений кодов, заимствований, ненормативной орфографии и т. п. блоги даже находятся ближе к «обычным» текстам, чем к соцсетям. Этим было обусловлено их включение в основные корпуса. Статьи из разделов Википедии на языках России, к сожалению, содержат множество некачественных (в лингвистическом смысле) статей, например, коротких однотипных автоматически порождённых заготовок [Орехов, Решетников 2014]. Поэтому только некоторое количество предварительно отфильтрованных по размеру и другим параметрам статей было включено в рассматриваемые корпуса.

Объём и жанровый состав основных корпусов по состоянию на июнь 2019 г. приведён в табл. 1.

Таблица 1

Объёмы и состав «основных» корпусов

Язык	объём	пресса (%)	блоги (%)	другое
удмуртский	9,57 млн	91,3 %	5,1 %	3,6 %
коми-зырянский	1,75 млн	100 %	0 %	0 %
луговой марийский	2,63 млн	84 %	0 %	16 %
эрзянский	2,3 млн	67,4 %	6 %	26,6 %
мокшанский	1,74 млн	86,4 %	0,7 %	12,9%

Разница в объёме и в количественном составе объясняется здесь не только разной представленностью языков в Интернете, но и разным количеством усилий, которые я прилагал в каждом случае. Например, в коми-зырянском и луговом марийском корпусах на сегодняшний момент отсутствуют блоги, несмотря на то, что блоги на этих языках существуют. Поскольку для этих языков существуют крупные корпуса, работа над которыми ведётся упомянутыми выше командами и сегодня, я предпочёл уделить больше внимания удмуртскому, эрзянскому и мокшанскому языкам.

Что касается текстов соцсетей, то здесь моей задачей было приблизиться к исчерпывающему списку источников, доступных на платформе «ВКонтакте» для каждого из этих языков. Помимо текстов из «ВКонтакте», я также загрузил тексты двух форумов, общение на которых в основном происходило на эрзянском языке (erzianj.borda.ru и erzianraske.forum24.ru).

Большую проблему при обработке текстов соцсетей представляло постоянное переключение кодов: очень часто в рамках одного сообщения предложения на русском языке чередуются с предложениями на одном из финно-угорских языков. Как следствие, мне было необходимо определять язык не отдельных текстов, а отдельных предложений. Поскольку традиционные n-граммные методы дают плохие результаты при определении языка таких коротких отрезков текста, я разработал более точные определители, основанные в основном на словарном методе (подробнее см. в [Arkhangelskiy 2019]). Точность определения языка, измеренная в предложениях, составляет в среднем 97 %. Корпуса соцсетей содержат и финно-угорские, и русские предложения, поскольку опустить русскоязычную часть было невозможно без ущерба для понимания текста. Однако поскольку каждое предложение содержит информацию о языке, пользователь может искать только среди финно-угорских или только среди русских предложений. Объёмы корпусов соцсетей, отдельно для русскоязычной и финно-угорских частей, представлены ниже в табл. 2.

Таблица 2

Объём и состав корпусов соцсетей

Язык	объём (ФУ)	объём (рус.)	группы	пользователи[4]
удмуртский	2,66 млн	9,83 млн	335	979
коми-зырянский	2,14 млн	16,12 млн	87	408
луговой марийский	3,59 млн	15,1 млн	177	588
эрзянский	0,83 млн	5,23 млн	20 (+ форумы)	111 (+ форумы)
мокшанский	14 тыс.	0,17 млн	17	17

В абсолютных цифрах лидером является луговой марийский, однако при подробном рассмотрении оказывается, что более чем половину текстов в этом корпусе написал один и тот же пользователь. Без учёта текстов этого пользователя объём корпуса уменьшается до 1,32 млн словоупотреблений. В остальных корпусах тексты более равномерно распределены по авторам. Удмуртский язык, таким образом, номинально занимает второе место по объёму, но лидирует по другим важным параметрам – количеству групп и пользователей, которые пишут на этом языке. Этот факт уже отмечался Б. В. Ореховым на основе данных его проекта «Языки России», куда не были включены личные страницы пользователей; как видно, более полные данные его подтверждают. Тем не менее, коми-зырянский и луговой марийский показывают сравнимые с удмуртским результаты. Динамика изменений позволяет предположить, что в будущем эти языки могут обогнать удмуртский, поскольку в последние годы они демонстрируют существенный прирост в объёме (в отличие от 2007–2014 гг., когда удмуртский сегмент «ВКонтакте» рос значительно быстрее остальных). Данные о распределении количества словоупотреблений по годам для каждого из языков, кроме мокшанского, представлены в табл. 3.

Таблица 3

Распределение текстов корпусов соцсетей по годам, в тысячах словоупотреблений

Год	удмурты	коми	эрзя (вк)	эрзя (форумы)	мари[5]
2006	0	0	0	15,9	0
2007	1,0	0,7	0,01	70,7	0,07
2008	15,1	1,9	0,7	23,1	12,7
2009	14,3	6,0	2,6	64,3	39,6
2010	42,7	5,9	3,8	**105,6**	26,4
2011	101,7	14,3	11,3	79,0	39,9
2012	273,1	33,0	29,2	40,8	59,0
2013	424,1	55,4	28,3	15,8	53,6
2014	473,6	140,6	79,2	20,4	101,8
2015	429,8	251,4	**96,5**	11,3	150,7
2016	350,6	259,0	70,8	1,4	217,3
2017	**505,2**	**660,6**	44,5	0,01	**462,7**
2018	?[6]	?	?	?	414,3

В целом три вышеупомянутых языка показывают схожую положительную динамику: количество словоупотреблений в соответствующих сегментах стабильно растёт из года в год, за исключением спада в удмуртском сегменте в 2015–2016 гг. и небольшого уменьшения в марийском сегменте в 2018 г. К сожалению, ситуация с эрзянским и мокшанским языками иная. Мокшанский язык в соцсетях практически не представлен. Эрзянский представлен в значительно меньшем объёме, чем пермские и луговой марийский; кроме того, количество текстов на нём снижается, а не растёт. Пик актив-

[4] Указаны пользователи, на стенах которых есть записи на данном языке. Если учесть также пользователей, которые оставили хотя бы один комментарий на финно-угорском языке, но не имеют таких записей на стенах, их количество вырастет в 2-3 раза.

[5] Без учёта текстов самого продуктивного автора.

[6] Поскольку тексты большинства корпусов были загружены в 2018 г., информация о количестве словоупотреблений за полный год отсутствует.

ности в эрзянском сегменте пришёлся на 2010 год в части форумов (которые сейчас практически не функционируют) и на 2015 год в платформе «ВКонтакте». Анализ текстов позволяет предположить, что активность в начале 2010-х гг. была вызвана обсуждением идеи создания единого литературного мордовского языка (см [Зайц 1995; Keresztes 1995; Мосин 2014]), против которой выступали многие языковые активисты. То же можно сказать об эрзянских блогах: после пика в 2012 г. (26 тыс. словоупотреблений) количество записей снижается, опустившись до 12 тыс. словоупотреблений в 2017 г. Всё это говорит о том, что положение эрзянского и мокшанского языков с точки зрения их представленности в цифровой сфере заметно хуже, чем у пермских языков и лугового марийского (хотя из анализа [Федина 2016] и [Kornai 2016] может показаться, что это не так).

Тексты основных корпусов и корпусов соцсетей сопровождаются разными наборами метаданных. В основных корпусах метаданные представляют собой несколько наиболее часто используемых полей: автор, заглавие текста, год (или годы) создания, жанр (ср., например, с более сложной системой, используемой в Национальном корпусе русского языка, [Савчук 2005]). В некоторых корпусах при текстах также указывается, в какой орфографии они набраны: в стандартной или с отсутствующими диакритиками. Цель включения всех этих метаданных в корпус – дать возможность пользователю ограничить поиск, например, только текстами блогов или только текстами, написанными после 2015 г.

Метаданные в соцсетях содержат больше информации. Помимо автоматически определённого языка, для каждого предложения доступны следующие данные:

– тип поста: оригинальный пост (post) / комментарий (comment) / репост (repost);
– вид аккаунта: пользователь (user) / группа (group);
– год написания;
– год рождения автора;
– пол автора: мужской (M) / женский (F);
– место рождения автора;
– место проживания автора.

Некоторые из этих параметров записаны отдельно для автора данного сообщения (в поисковом интерфейсе они имеют помету «пост» в скобках) и для владельца страницы, на которой происходит обсуждение.

Год рождения, место проживания и место рождения пользователей имеются в корпусе только в том случае, если пользователь сам указал их на своей странице. Кроме того, в целях анонимизации эти данные представлены в корпусе в обобщённом виде. Вместо точного года указывается пятилетний промежуток (например, «1990–1995» вместо 1992), а вместо точного места рождения или проживания – район или субъект РФ. Указание обобщённого места требует большой работы по установлению соответствий между населёнными пунктами и районами/субъектами и поэтому было пока реализовано только для удмуртского и лугового марийского корпусов. Настоящие имена или ники пользователей в корпусе не даются.

Морфологический разбор

Все корпуса подверглись автоматической морфологической обработке с помощью созданных мной анализаторов, основанных на словарях и правилах. Качество морфологической разметки разнится; уровень покрытия составляет от 80 % до 96 % словоформ. В корпусах соцсетей из-за частого использования ненормативной орфографии и других особенностей процент размеченных словоформ в среднем ниже, чем в «основных» корпусах. Использованный метод морфологического анализа имеет два важных последствия. Во-первых, слова, отсутствующие в словаре анализатора, остаются неразобранными. Несмотря на то, что словари регулярно пополняются частотными неразобранными лексемами (в основном русскими заимствованиями или недавно появившимися неологизмами), немало словоформ не имеют разбора. Современные методы решения этой проблемы, основанные на использовании методов машинного обучения (например, нейронных сетей) для угадывания разборов, в настоящий момент неприменимы к финно-угорским языками России, поскольку требуют наличия достаточно большой обучающей выборки текстов, размеченных вручную. Во-вторых, каждая словоформа анализируется независимо от контекста, и ей приписываются все потенциально возможные разборы. Например, удмуртская словоформа «пуны» может быть разобрана как немаркированная форма существительного «собака» или как одна из отрицательных форм глагола «вить, плести». В удмурт-

ском корпусе возникающая таким образом омонимия частично разрешается с помощью небольшого набора правил в формате Constraint Grammar [Bick, Didriksen 2015]. Например, разбор «отрицательная форма глагола *плести*» в данном случае будет отсечён, если этой словоформе в тексте не предшествует один из подходящих отрицательных глаголов (за которым, возможно, следует одна или несколько клитик). Однако вся омонимия в остальных корпусах и часть омонимии в удмуртском корпусе остаётся. Опять же, решение этой проблемы с помощью нейронных сетей потребовало бы наличия вручную размеченной выборки.

В результате морфологического анализа большинству словоформ приписывается один или несколько потенциальных вариантов разбора. Каждый из них содержит лемму (начальную, словарную форму), набор грамматических помет (часть речи, падеж, число и т. п.), глоссирование (разбиение на морфемы и сокращённые обозначения этих морфем) и краткий перевод на русский язык. Регулярные словообразовательные морфемы для простоты описания и поиска отнесены к грамматике и не затрагивают лемм; в первую очередь это касается глагольных дериваций и атрибутивов, характерных для всех рассматриваемых языков. Например, удмуртская словоформа *гожтйське* «пишется» будет разобрана как форма глагола *гожтыны* «писать», а не *гожтйськыны* «писаться», а факт наличия пассивного суффикса будет отражён в грамматических пометах и глоссах. Тем не менее, некоторые идиоматические комбинации корней и продуктивных суффиксов сопровождаются дополнительными леммой (поле «Лемма 2») и переводом на русский (поле «Перевод 2»). Помимо словоизменительных и регулярных словообразовательных категорий, в корпусах размечены некоторые лексико-семантические классы (например, одушевлённые существительные и русские заимствования) и отдельные словообразовательные суффиксы (например, диминутив в эрзянском). Грамматические пометы и глоссы записываются латинскими сокращениями, например, множественное число обозначается пометой *pl*. Расшифровка помет приводится на стартовой странице каждого из корпусов; также полные списки используемых помет доступны в поисковом интерфейсе по щелчку на кнопку в правой части полей «Грамматика» и «Глоссы». Многие грамматические категории известны под разными названиями в литературе, например, «фреквентатив» или «итератив». В корпусах для каждой категории выбрано одно из возможных названий (возможно, не всегда наилучшее); согласно использованному при этом «принципу приоритета широты поисковых возможностей» [Ляшевская и др. 2005], важно, чтобы каждую грамматическую категорию можно было найти в корпусе, а как она в нём называется, не так важно.

Поиск в корпусе

Тексты, входящие в корпуса, были размещены в Интернете с помощью свободно распространяемой корпусной платформы «tsakorpus»[7]. Она предоставляет пользователю широкие возможности поиска и отображения результатов. В частности, возможны следующие виды поиска:

– поиск отдельных словоформ, например, «пуныослы»;

– поиск всех форм какой-либо лексемы, например, всех форм лексемы «пуны»;

– поиск по грамматике с использованием логических функций, например, «все существительные в дательном или родительном падеже» («N,(dat|gen)»);

– поиск по глоссам с возможностью указать их взаимный порядок, например, «все словоформы, у которых суффикс пассива предшествует суффиксу каузатива» («PASS-CAUS»);

– поиск словоформ и лемм с использованием звёздочек или регулярных выражений, например, «все слова, начинающиеся на *пу*» («пу*»);

– поиск с одновременным заданием нескольких из вышеперечисленных параметров;

– поиск нескольких слов в рамках одного предложения с заданием расстояния, например, «слово *пуны* на расстоянии не более двух слов от глагола»;

– поиск только в некотором подкорпусе текстов, например, только в блогах или только среди текстов пользователей соцсетей, родившихся в Алнашском районе.

Результатом поиска может быть либо список предложений, содержащих искомые слова, либо список словоформ или лемм, подходящих под запрос. Для любого найденного предложения можно подгрузить его контекст – некоторое (ограниченное) число соседних с ним предложений. По умолчанию результаты поиска сортируются в случайном порядке. Такая сортировка гарантирует, что в случаях, когда поиск возвращает тысячи примеров, просмотр первых нескольких страниц даст исследо-

[7] https://bitbucket.org/tsakorpus/tsakorpus.

вателю довольно точное приблизительное представление о распределении изучаемого явления во всём корпусе. В противном случае – например, при сортировке по дате создания или по названию текста – могло бы оказаться, что изучаемое явление ведёт себя по-разному на разных страницах поисковой выдачи, и пользователю пришлось бы в обязательном порядке просматривать их все, чтобы сделать обоснованные выводы. Текст предложения вместе с глоссированием и другими слоями информации можно скопировать в буфер обмена нажатием на синюю кнопку слева от него, что может быть удобно при вставке примеров из корпуса в лингвистические статьи.

Заключение

В этой статье были представлены корпуса нескольких финно-угорских языков России, содержащие тексты из Интернета. Для каждого языка было создано два корпуса: один состоит из текстов соцсетей и форумов (в основном «ВКонтакте»), другой – из всех остальных текстов (в основном из новостных изданий). Корпуса оснащены поисковым веб-интерфейсом и доступны по ссылкам с общей стартовой страницы http://volgakama.web-corpora.net/ .

У представленных здесь корпусов имеется ряд недостатков, которые затрудняют их использование в ряде исследований. В первую очередь это касается текстового состава корпусов: они включают в себя тексты определённых жанров, созданные в последнее двадцатилетие. Таким образом, на материале этих корпусов невозможно проводить диахронические или литературоведческие исследования. Тем не менее, в большом количестве других видов лингвистической работы эти корпуса будут полезным инструментом. Помимо изучения лексики и грамматики финно-угорских языков России, представленные корпуса особенно хорошо подойдут для изучения современной ситуации в свете социолингвистики и многоязычия, как это предлагалось в [Пишлёгер 2017]. Использование корпусов соцсетей должно значительно упростить изучение переключения и смешения кодов (которое уже проводилось на материале финно-угорских соцсетей, см., например, [Гаврилова 2019] о марийско-русском двуязычии и [Деци 2019] об эстонско-русском двуязычии). То же можно сказать об изучении влияния русского языка в области лексики и грамматики и об исследованиях из области языкового планирования (например, с помощью корпусов можно выяснить, насколько широко используются на практике неологизмы). Отдельный интерес такие корпуса могут представлять для диалектологии, поскольку в текстах многих авторов заметны диалектные черты, особенно в лексике и морфологии. Можно надеяться, что наличие корпусов для финно-угорских языков России усилит интерес к этим языкам и приведёт к росту количества посвящённых им исследований.

ЛИТЕРАТУРА

Векшина М. М. Особенности языка удмуртоязычной блогосферы // Татарское языкознание в контексте Евразийской гуманитарной науки: Материалы Международной научно-практической конференции. Казань, 2016. С. 83–87.

Гаврилова В. Г. Русско-марийское переключение и смешение кодов в интернет-коммуникации // Ежегодник финно-угорских исследований. 2019. Т. 13. № 1. С. 6–13.

Деци А. Эстонские вкрапления в интернет-дискурсе русскоязычных жителей Эстонии // Ежегодник финно-угорских исследований. 2019. Т. 13. № 2. С. 331–342.

Зайдельман Л. Я., Крылова И. В., Орехов Б. В. Технология поиска и сбора в Интернете текстов на малых языках России // Труды Международной научной конференция CPT2015. Институт физико-технической информатики, 2016. С. 179–181.

Зайц Г. Сколько языков нужно эрзе и мокше? // Zur Frage der uralischen Schriftsprachen. Linguistica, Series A, Studia et Dissertationes. Будапешт: Az MTA Nyelvtudományi Intézete, 1995. С. 41–46.

Ляшевская О. Н., Плунгян В. А., Сичинава Д. В. О морфологическом стандарте Национального корпуса русского языка // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М.: Индрик, 2005. С. 111–135.

Мосин М. В. Создавать ли единые литературные языки для уральских народов? // Труды Карельского научного центра РАН. 2014. № 3. С. 76–82.

Орехов Б. В., Решетников К. Ю. К оценке Википедии как лингвистического источника: сравнительное исследование // Современный русский язык в интернете / под ред. Я. Э. Ахапкиной, Е. В. Рахилиной. М.: Языки славянской культуры, 2014. С. 309–321.

Пишлёгер К. Удмуртский язык в социальной сети «ВКонтакте»: Квантитативные и (возможные) квалитативные исследования. Электронная письменность народов Российской Федерации: Опыт, проблемы и перспективы. Сыктывкар: ГОУ ВО КРАСГСиУ, 2017. С. 154–162.

Савчук С. О. Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. М.: Индрик, 2005. С. 62–88.

Федина М. С. Финно-угорские языки Российской Федерации в электронном информационном пространстве: опыт, проблемы и перспективы // Финно-угорский мир. 2016. Т. 3. № 28. С. 111–121.

Arkhangelskii T. Corpora of social media in minority Uralic languages // Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages. Tartu, Estonia: Association for Computational Linguistics, 2019. С. 125–140.

Arkhangelskii T., Medvedeva M. Developing Morphologically Annotated Corpora for Minority Languages of Russia // Proceedings of Corpus Linguistics Fest 2016. Bloomington, IN. 2016. С. 1–6.

Baldwin T. и др. How Noisy Social Media Text, How Diffrnt Social Media Sources // International Joint Conference on Natural Language Processing. Nagoya, Japan. 2013. С. 356–364.

Biber D. Representativeness in Corpus Design // Literary and Linguistic Computing. 1993. Т. 8, № 4. С. 243–257.

Bick E., Didriksen T. CG-3 – Beyond Classical Constraint Grammar // Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015). Linköping University Electronic Press, 2015. С. 31–39.

Bradley J. Corpus.mari-language.com: A Rudimentary Corpus Searchable by Syntactic and Morphological Patterns // Proceedings of the First international workshop on computational linguistics for Uralic languages. Septentrio Conference Series. Septentrio Academic Publishing, 2015.

Canvar W. B., Trenkle J. M. N-Gram-Based Text Categorization // Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval. 1994. С. 161–176.

Csendes D., Csirik J., Gyimóthy T. The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus // Text, Speech and Dialogue / под ред. P. Sojka, I. Kopeček, K. Pala. Springer Berlin Heidelberg, 2004. С. 41–47.

Francis W. N., Kučera H. Frequency Analysis of English Usage: Lexicon and Grammar. Boston: Houghton Mifflin, 1982.

Halácsy P. и др. Creating open language resources for Hungarian // LREC 2004 Proceedings. 2004. С. 203–210.

Keresztes L. On the Question of the Mordvinian Literary Language // Zur Frage der uralischen Schriftsprachen. Linguistica, Series A, Studia et Dissertationes / под ред. G. Zaics. Budapest: Az MTA Nyelvtudományi Intézete, 1995. С. 47–55.

Kilgarriff A., Grefenstette G. Introduction to the Special Issue on the Web as Corpus // Computational Linguistics. 2003. Т. 29. № 3. С. 333–347.

Kornai A. Computational linguistics of borderline vital languages in the Uralic family // Proceedings of the Second International Workshop on Computational Linguistics for Uralic Languages. Szeged: 2016.

Leech G. New resources, or just better old ones? The Holy Grail of representativeness // Corpus Linguistics and the Web / под ред. M. Hundt, N. Nesselhauf, C. Biewer. Brill, 2007. С. 133–149.

McEnery T., Hardie A. Corpus linguistics: method, theory and practice. Cambridge: Cambridge University Press, 2011.

Pajzs J. Making Historical Dictionaries with the Computer // Proceedings of EURALEX 2000. 2000. С. 249–259.

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. 106–118.

Scannell K. P. The Crúbadán Project: Corpus building for under-resourced languages // Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop. 2007. С. 5–15.

Suihkonen P. Documentation of the Computer Corpora of Uralic Languages at the University of Helsinki. Helsinki: Department of General Linguistics, University of Helsinki, 1998.

Váradi T. The linguistic relevance of corpus linguistics // Proceedings of the Corpus Linguistics 2001 Conference. 2001. С. 587–593.

Váradi T. The Hungarian National Corpus // Proceedings of LREC 2002. 2002. С. 385–389.

Поступила в редакцию 22.07.2019

Архангельский Тимофей Александрович,
кандидат филологических наук, научный сотрудник
Университет Гамбурга (Германия),
Институт финно-угроведения/уралистики
E-mail: timarkh@gmail.com

T. A. Arkhangelskii

INTERNET-CORPORA FOR FINNO-UGRIC LANGUAGES OF RUSSIA

DOI: 10.35634/2224-9443-2019-13-3-528-537

Digital language corpora have long become one of the most important tools in linguistic research; a new methodological approach, known as corpus linguistics, has been based on corpora. While comprehensive corpora exist for the major European Uralic languages (Hungarian, Finnish, Estonian), the smaller Uralic languages of Russia did not have comparable resources until recently. In this paper, I present digital corpora recently developed for the largest Uralic languages of Russia: Udmurt, Komi-Zyrian, Meadow Mari, Erzya and Moksha. The corpora comprise digital texts available on the internet, which were collected and processed by the author. Two corpora were created for each language: a social media corpus and a non-social-media ("main") corpus. Both kinds of texts were automatically morphologically analyzed; the social media texts were additionally filtered and anonymized. I will outline the development process of these corpora, as well as present their features and possible applications. All corpora described in the paper are equipped with a web-based user interface and are publicly available at http://volgakama.web-corpora.net/.

Keywords: language corpus, corpus linguistics, social media, Uralic languages, Udmurt, Mari, Komi, Erzya, Moksha.

Citation: Yearbook of Finno-Ugric Studies, 2019, vol. 13, issue 3, pp. 528–537. In Russian.

REFERENCES

Vekshina M. Osobennosti yazyka udmurtoyazychnoi blogosfery [Language features of the Udmurt-language blogs]. *Tatarskoe yazykoznanie v kontekste Evraziiskoj gumanitarnoi nauki: Materialy Mezhdunarodnoi nauchno-prakticheskoi konferencii* [Tatar linguistics in the context of Eurasian humanities: Proceedings of the international conference]. Kazan. 2016. Pp. 83–87. In Russian.

Gavrilova V. Russko-mariiskoe perekl'uchenie i smeshenie kodov v internet-kommunikatsii [Russian and Mari Code-Switching and Mixing Codes in the Internet Communication]. *Ezhegodnik finno-ugorskikh issledovanii* [Yearbook of Finno-Ugric studies]. 2019. Vol. 13. No. 1. Pp. 6–13. In Russian.

Dezi A. Estonskie vkrapleniya v internet-diskurse russkoyazychnykh zhitelei Estonii [Estonian items in the internet discourse of the Russian speaking population of Estonia]. *Ezhegodnik finno-ugorskikh issledovanii* [Yearbook of Finno-Ugric studies]. 2019. Vol. 13. No. 2. Pp. 331–342. In Russian.

Zaydelman L., Krylova I., Orekhov B. Tekhnologiya poiska i sbora v Internete tekstov na malykh yazykakh Rossii [The technology of web-texts collection of Russian minor languages]. *International conference CPT2015*. Institut fiziko-tekhnicheskoi informatiki, 2016. Pp. 179–181. In Russian.

Zaics G. Skol'ko yazykov nuzhno erze i mokshe? [How many languages do the Erzya and the Moksha need?]. *Zur Frage der uralischen Schriftsprachen* [Questions of Uralic literary languages] Linguistica, Series A, Studia et Dissertationes. Budapest: Az MTA Nyelvtudományi Intézete, 1995. Pp. 41–46. In Russian.

Lyashevskaya O., Plungian V., Sitchinava D. O morfologicheskom standarte Natsional'nogo korpusa russkogo yazyka [About the morpholoigcal standard of the Russian National Corpus]. *Natsional'nyi korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy* [Russian National Corpus: 2003-2005. Results and prospects]. Moscow: Indrik, 2005. Pp. 111–135. In Russian.

Mosin M. Sozdavat' li edinye literaturnye yazyki dlya ural'skikh narodov? [Should unified literary languages be created for Uralic peoples?] *Trudy Karel'skogo nauchnogo tsentra RAN* [Proceedings of the Karelian scientific center of the Russian Academy of Sciences]. 2014. No. 3. Pp. 76–82. In Russian.

Orekhov B., Reshetnikov K. K otsenke Vikipedii kak lingvisticheskogo istochnika: sravnitel'noe issledovanie [Evaluating Wikipedia as a linguistic source: A comparative study]. *Sovremennyi russkii yazyk v internete* [Contemporary Russian language on the Internet]. Edited by Y. Akhapkina, E. Rakhilina. Moscow: Jazyki slavyanskoi kul'tury, 2014. Pp. 309–321. In Russian.

Pischlöger C. Udmurtskii yazyk v sotsial'noi seti "VKontakte": Kvantitativnye i (vozmozhnye) kvalitativnye issledovaniya [Udmurt in the Vkontakte Social Network: Quantitative and (Possible) Qualitative Research]. *Finno-ugorskie yazyki Rossiiskoi Federatsii v elektronnom informatsionnom prostranstve: opyt, problemy i perspektivy* [Finno-Ugric languages of Russia in the digital information space: experiences, challenges and prospects]. Syktyvkar: GOU VO KRASGSiU, 2017. C. 154–162. In Russian.

Savchuk S. Metatekstovaya razmetka v Natsional'nom korpuse russkogo yazyka: bazovye principy i osnovnye funktsii [Metadata in the Russian National Corpus: Basic principles and main functions]. *Natsional'nyi korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy* [Russian National Corpus: 2003-2005. Results and prospects]. Moscow: Indrik, 2005. Pp. 62–88. In Russian.

Fedina M. Finno-ugorskie yazyki Rossiiskoi Federatsii v elektronnom informatsionnom prostranstve: opyt, problemy i perspektivy [Finno-Ugric languages of Russia in the digital information space: experiences, challenges and prospects]. *Finno-ugorskii mir* [Finno-Ugric world]. 2016. Vol. 3. No. 28. Pp. 111–121. In Russian.

Arkhangelskii T. Corpora of social media in minority Uralic languages. *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*. Tartu, Estonia: Association for Computational Linguistics, 2019. Pp. 125–140. In English.

Arkhangelskii T., Medvedeva M. Developing Morphologically Annotated Corpora for Minority Languages of Russia. *Proceedings of Corpus Linguistics Fest 2016*. Bloomington, IN, 2016. Pp. 1–6. In English.

Baldwin T., Cook P., Lui M., MacKinlay A., Wang L. How Noisy Social Media Text, How Diffrnt Social Media Sources. *International Joint Conference on Natural Language Processing*. Nagoya, Japan, 2013. C. 356–364. In English.

Biber D. Representativeness in Corpus Design. *Literary and Linguistic Computing*. 1993. Vol. 8. No. 4. Pp. 243–257. In English.

Bick E., Didriksen T. CG-3 - Beyond Classical Constraint Grammar. *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*. Linköping University Electronic Press, 2015. Pp. 31–39. In English.

Bradley J. Corpus.mari-language.com: A Rudimentary Corpus Searchable by Syntactic and Morphological Patterns. *Proceedings of the First international workshop on computational linguistics for Uralic languages*. Septentrio Conference Series. Septentrio Academic Publishing, 2015. In English.

Canvar W.B., Trenkle J.M. N-Gram-Based Text Categorization. *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*. 1994. Pp. 161–176. In English.

Csendes D., Csirik J., Gyimóthy T. The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. *Text, Speech and Dialogue*. Editeb by P. Sojka, I. Kopeček, K. Pala. Springer Berlin Heidelberg, 2004. Pp. 41–47. In English.

Francis W.N., Kučera H. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin, 1982. In English.

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. In English.

Keresztes L. On the Question of the Mordvinian Literary Language. *Zur Frage der uralischen Schriftsprachen* [Questions of Uralic literary languages] Linguistica, Series A, Studia et Dissertationes. Edited by G. Zaics. Budapest: Az MTA Nyelvtudományi Intézete, 1995. Pp. 47–55. In English.

Kilgarriff A., Grefenstette G. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*. 2003. Vol. 29. No. 3. Pp. 333–347. In English.

Kornai A. Computational linguistics of borderline vital languages in the Uralic family. *Proceedings of the Second International Workshop on Computational Linguistics for Uralic Languages*, Szeged, 2016. In English.

Leech G. New resources, or just better old ones? The Holy Grail of representativeness. *Corpus Linguistics and the Web*. Edited by M. Hundt, N. Nesselhauf, C. Biewer. Brill, 2007. Pp. 133–149. In English.

McEnery T., Hardie A. *Corpus linguistics: method, theory and practice.* Cambridge: Cambridge University Press, 2011. In English.

Pajzs J. Making Historical Dictionaries with the Computer. *Proceedings of EURALEX 2000.* 2000. Pp. 249–259. In English.

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. Pp. 106–118. In English.

Scannell K.P. The Crúbadán Project: Corpus building for under-resourced languages. *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop.* 2007. Pp. 5–15. In English.

Suihkonen P. *Documentation of the Computer Corpora of Uralic Languages at the University of Helsinki.* Technical Reports TR-2. Helsinki: Department of General Linguistics, University of Helsinki, 1998. In English.

Váradi T. The linguistic relevance of corpus linguistics. *Proceedings of the Corpus Linguistics 2001 Conference.* 2001. Pp. 587–593. In English.

Váradi T. The Hungarian National Corpus. *Proceedings of LREC 2002.* 2002. Pp. 385-389. In English.

Поступила в редакцию 22.07.2019

Arkhangelskii Timofei Aleksandrovich,
Candidate of Philology, Research associate
University of Hamburg (Germany),
Institute of Finno-Ugric and Ural studies
E-mail: timarkh@gmail.com