Портфолио

Михаил Кириллович Воронов
Email: mikivo@list.ru

29 июля 2019 г.

Содержание

1. Образование

2015-2019: Факультет гуманитарных наук НИУ ВШЭ, программа «Фундаментальная и компьютерная лингвистика»

2. Академическая деятельность

2.1. Краткое описание научных интересов

В мои научные интересы в данный момент входит в первую очередь разработка программных библиотек и приложений для теоретического исследования языка. В разное время я также занимался аспектологией, экспериментальной фонетикой и лексикостатистикой.

В рамках магистратуры хочу лучше освоить NLP и заняться им тоже.

2.2. Дипломная работа по теме «Программная библиотека для лингвистической типологии на языке Python»

В рамках дипломной работы под руководством Бориса Орехова я занимался разработкой программной библиотеки для исследований в области лингвистической типологии.

Библиотека LingTypology предоставляет API (интерфейс прикладного программирования) для ряда лингвистических баз данных (WALS, PHOIBLE и т.д.), а также инструмент для создания интерактивных лингвистических карт. Такой функционал библиотеки имеет смысл, так как существует много типологических исследований, которые опираются на данные из лингвистических баз и используют визуализацию в виде карт (например, (Blasi и др. 2019)).

Схожая библиотека уже существует для языка R, однако, в связи с тем, что среди лингвистов чаще используется язык Python, и, в отличие от R, Python является универсальным языком, существование аналога на этом языке выглядит целесообразным. К тому же, моя библиотека обладает рядом преимуществ в сравнении с библиотекой на R: возможность сохранять карты в формате PNG, расширенные возможности работы с базами данных и ряд других новых возможностей.

В ходе работы я также привожу несколько демонстрационных исследований, проведённых при помощи данной библиотеки. Это короткие квантитативные исследования, связанные с обработкой данных из баз WALS, Autotyp и PHOIBLE, подробнее о них можно посмотреть в тексте работы, который доступен по следующей ссылке (на английском языке):

https://github.com/OneAdder/lingtypology_research/blob/master/work/diplom.pdf

Одной из целей создания библиотеки было упростить вопроизводимость типологических исследований. Очевидно, что в таком случае код самой библиотеки должен быть открытым. Он опубликован с лицензией GNU GPL и доступен по следующей ссылке:

https://github.com/OneAdder/lingtypology

Проект снабжен документацией, доступной по этой ссылке:

https://oneadder.github.io/lingtypology/html/index.html

Библиотека представляет собой уже готовый продукт, который оформлен в виде пакета. На момент написания этого резюме последней версией LingTypology является релиз 0.8.5.

2.3. Участие в грантовом проекте

С 2018 года участвую в проекте Института языкознания РАН по созданию и поддержке базы семантических переходов в роли программиста и исследователя. Работа проекта поддержана грантом РФФИ (№ 17-29-09124 «Когнитивные механизмы семантической деривации в свете типологических данных»). Руководитель проекта: Анна А. Зализняк. Должность руководителя: ведущий научный сотрудник. С 2019 года я официально состою в этом гранте.

Моя роль заключается в создании программных компонентов, необходимых для проекта, в частности, поддержка сайта базы и, на более ранних этапах, его создание.

2.4. Участие в экспедициях

- Экспедиции в горномарийский язык с темой «Ретроспективный сдвиг в горномарийском языке» (июль 2016, февраль 2017 и июль 2017). Руководитель экспедиций: С.Ю. Толдова, научный руководитель исследования: А.А. Козлов.
- Экспедиция в горномарийский язык с темой «Лексикостатистический анализ базисной лексики марийских языков» (февраль 2018). Руководитель экспедиции: С.Ю. Толдова, научный руководитель исследования: Г.С. Старостин.

2.5. Доклады на конференциях

- М. К. Воронов, Д. Д. Мордашова. Горномарийские глагольные конструкции с ə̑ l’ə̑ / ə̑ lə̑ n в зоне «сверхпрошлого» и за её пределами. XIII конференция по типологии и грамматике для молодых исследователей. ИЛИ РАН, Санкт-Петербург, 24-26 ноября 2016 г.
- Алексей Кошевой, Михаил Воронов. Исследование вариативности посессивных суффиксов в СРЯ. «АНТРОПОЛОГИЯ. ФОЛЬКЛОРИСТИКА. СОЦИОЛИНГВИСТИКА». Европейский университет в Санкт-Петербурге, Санкт-Петербург, 22-24 марта 2018.

2.6. Тезисы конференций

См. ниже в списке литературы (ссылки снабжены гиперссылками):

- (М.К.Воронов 2016)
- (Михаил Воронов 2018)

3. Проектная и практическая деятельность

3.1. Опыт работы

03.04.2017-25.09.2018: работал на должности стажёра-исследователя в Международной лаборатории языковой конвергенции (Факультет гуманитарных наук НИУ ВШЭ). Моя работа заключалась в подготовке данных для создания корпуса устных текстов для лугового марийского языка.

3.2. Участие в проектах

3.2.1. Проекты, связанные с НИУ ВШЭ

- 2019: участие в проекте LingCorpora

  (https://github.com/lingcorpora/lingcorpora.py).

  LingCorpora – это программная библиотека на языке Python, предоставляющая унифицированный интерфейс для текстовых корпусов. Я участвовал в подготовке к релизу 2.0. Среди нововведений этого релиза: добавление 18 новых корпусов, значительное изменение API (с утратой обратной совместимости), улучшение качества кода, создание новой документации и многочисленные исправления ошибок и мелкие улучшения качества работы библиотеки.

- 2018-2019: участие в проекте по исследованию шугнанского языка. Моя работа заключалась в создании единого интерфейса к двум словарям шугнанского языка (словари Карамшоева и Зарубина). Сложность заключается в том, что словари используют разные графические системы, а также в том, что один из словарей не полностью оцифрован, а просто распознан автоматически и частично вычитан. В результате я получил файл JSON с двумя словарями и простой графический интерфейс на Qt для работы с ним.

- 2019: создание алгоритма унификации графики древнерусских текстов (в соавторстве с Анной Сорокиной). В данный момент алгоритм находится в разработке. Как известно, в древнерусских текстах встречается вариативность. Она связана с историей развития древнерусского языка, диалектными особенностями писца и престижа некоторых вариантов написании в конкретные эпохи. Это создаёт сложности для поиска по словарям и корпусам. Существуют некоторые алгоритмы унификации, однако, на мой взгляд, они далеки от идеала. На данный момент мы разработали алгоритм на основе статей (Баранов и др. 2007) и (Гаврилова, Шалганова и Ляшевская 2017). Пока алгоритм нацелен на словарные формы и имеет ряд других недостатков. Однако, я планирую продолжить работу над ним в ближайшее время. Код оформлен в виде библиотеки и доступен здесь:

  https://github.com/OneAdder/Old_Russian_graphics_unification

- 2016: помощь в переносе фонетических данных из грамматик языков Кавказа в цифровой вариант (несколько таблиц).

3.2.2. Проекты, не связанные с НИУ ВШЭ

- 2019: контрибуция в проект Folium. В ходе работы над дипломом я участвовал в обсуждении в рамках проекта Folium, в ходе которого мы пришли к выводу, что неизвестно, как с помощью функционала данной библиотеки добиться того, что мне требуется для функционала миниграфиков в LingTypology. После чего я нашёл способ добиться нужного мне результата, и получил просьбу от разработчиков Folium создать пример для их галереи примеров. Что и было сделано.

- 2019: разработка графического приложения для утилиты libinput-gestures (https://github.com/OneAdder/libinput_gestures_qt). Окружение рабочего стола для UNIX-систем KDE Plasma при использовании сервера окон Xorg лишено поддержки жестов тачпада, однако их можно обрабатывать при помощи утилиты libinput-gestures. Однако утилита не имеет графического интерфейса и управляется через терминал и файл конфигурации. Это неудобно, поэтому я разработал графический интерфейс на основе фреймворка Qt, который поддерживает маппинг жестов тачпада на комбинации клавиш, команды UNIX и ярлыки оконного менеджера KWin.

- 2017: разработка (в соавторстве с Анастасией Тимошиной) бота в сети Telegram, воспроизводящего так называемую хуй-редупликацию. На данный момент в боте используется упрощённая схема редупликации без задействования акцентного словаря, который необходим для полного воспроизведения явления хуй-редупликации (например, здесь (Пиперски 2017)). В дальнейшем планируется имплементировать данную возможность. Бот доступен в сети Telegram под именем *@huyach_bot*.

- 2019: создание небольшой библиотеки на языке JavaScript, которая умеет переводить тексты в русской кириллической графике в научную транслитерацию (ГОСТ 7.79-2000, ISO 9:1995).

  Код библиотеки доступен здесь:

  https://github.com/OneAdder/scientific_cyrillic_transliteration

3.3. Прочая деятельность

- Летняя лингвистическая школа. В июле 2017 и июле 2018 я участвовал в ЛЛШ в качестве «исполняющего обязанности студента на ЛЛШ».

- Московская традиционная олимпиада по лингвистике. 14 февраля и 6 марта 2016 года и 26 февраля и 19 марта 2017 года я участвовал в организации Московской традиционной олимпиады по лингвистике. 14 февраля и 6 марта 2016 года и 19 марта 2017 года я также участвовал в её проверке.

- Зимняя олимпиадная школа МФТИ. В январе 2019 я участвовал в ЗОШ МФТИ по компьютерной лингвистике в роли «вожатого-ассистента».

- Зимняя экологическая школа. В январе 2018 я участвовал в ЗЭШ в роли куратора.

4. Награды и достижения

- Признание в сетях Reddit и Github за создание приложения libinput-gestures-qt (32 лайка, 47 лайков и 12 звёздочек).

- Победы в олимпиадах в школьное время. В частности победа в Турнире Ломоносова по лингвистике (2014-2015). Также награждён медалью от Южного округа Москвы «Победитель олимпиад, смотров, конкурсов» (2013).

5. Мотивационное письмо

Этим летом я закончил программу бакалавриата ФиКЛ в Высшей школе экономики. Многие курсы программы, как оказалось, не только интересные, но также полезные для моей дальнейшей деятельности.

Достаточно очевидно, что курсы по теории лингвистики всегда оказываются полезны для всех. Однако моё образование также включало в себя ряд других курсов, которые оказались не менее полезными. Стоит отметить курсы иностранных языков (в моём случае: итальянский, монгольский и тундровый ненецкий), так как изучение иностранных языков значительно расширяет кругозор лингвиста. Курсы по древним языкам (в моём случае: латынь, древнецерковнославянский, древнерусский и древнегреческий) не только расширяют кругозор, но и позволяют прочувствовать то, как языки меняются, что очень ценно.

Я участвовал в лингвистических экспедициях в горномарийский язык. Опыт полевого исследования языка, на мой взглядт, даёт возможность почувствовать предмет науки лингвистики наиболее полно и непосредственно. Когда я впервые был в такой экспедиции, меня больше всего впечатлила вариативность. Мы редко задумываемся над тем, что все люди говорят на немного разном языке, даже если знаем в теории, что это так. Что ещё более важно, полевая работа стимулирует читать теоретическую литературу, чтобы понимать языковые явления, с которыми сталкиваешься.

На третьем курсе бакалавриата я выбрал теоретический профиль подготовки. Однако за это время я понял, что меня больше интересует компьютерная лингвистика. Как я писал выше, образование дало мне познания в области теории языка и расширило мой линвистический кругозор. Теперь, я хочу получить фундаментальное образование в области компьютерной лингвистики. Для меня это логичное продолжение. Собственно, мне больше всего интересны курсы по прикладной лингвистике в рамках данной магистратуры.

В бакалавриате я посещал майнор по математематике, а также НИСы по статистике и базам данных, которые помогли заметно продвинуться в информационных технологиях. В свободное время я изучал некоторые технологии языка Python (библиотеки threading, multiprocessing, фреймворки Flask и Qt и т.д.), а также язык C. Таким образом, мне хотелось бы дальше продвинуться в этих областях, в чём мне могут помочь курсы по машинному обучению, вебдевелопменту и квантитативному анализу лингвистических данных.

Теория компьютерной лингвистики и прикладные знание в информационных технологиях суть то, что я называю фундаментальным образованием в области компьютерной лингвистики, и это то, что я хочу получить.

Помимо учёбы, я планирую продолжать акдемическую и проектную деятельность. В частности, я хочу продолжить работу над алгоритмом унификации древнерусской графики, о котором я упомянул выше. Также я хочу продолжать поддержку и разработку библиотек LingTypology и LingCorpora. Пока есть план добавить функционал визуализации семантических карт в LingTypology, так как кажется, что это нужно. Я уверен, что магистратура откроет для меня возможности для участия в новых интересных проектах. Также я смогу получше разобраться в NLP, чтобы принимать участие в проектах в этом направлении знания.

В бакалавриате я участвовал в проектах, связанных с ВШЭ (секция выше). В магистратуре я собираюсь продолжать участие в этих проектах и прини-

мать участие в новых. Таким образом, какие-то из них могут значиться среди презентационных достижений магистратуры и Школы лингвистики в целом. Иными словами, на мой взгляд моё поступление в магистратуру будет полезно не только мне, но и магистратуре.

В заключение, хотелось бы сказать, что в все курсы в рамках бакалавриата так или иначе оказались полезными для моей дальнейшей деятельности. Судя по программе магистратуры по компьютерной лингвистике, такое же должно получиться и здесь. Проекты и исследования, которыми я занимаюсь, по своей тематике также наиболее близки именно к этому направлению знания. Таким образом, я уверен, что в магистратуре по компьютерной лингвистике я смогу получить важные для меня знания и в полной мере вносить свой вклад в учебные и научные проекты.

6. Список литературы

Blasi, D. E. и др. (2019). «Human sound systems are shaped by post-Neolithic changes in bite configuration». В: *Science* 363.6432. ISSN: 0036-8075. DOI: 10.1126/science.aav3218. eprint: https://science.sciencemag.org/content/363/6432/eaav3218.full.pdf. URL: https://science.sciencemag.org/content/363/6432/eaav3218.

Баранов, ВА и др. (2007). «Автоматический морфологический анализатор древнерусского языка: лингвистические и технологические решения». В: *«EVA 2007 Москва»: 10-я юбилейная междунар. конф.*

Гаврилова, Татьяна Сергеевна, Татьяна Александровна Шалганова и Ольга Николаевна Ляшевская (2017). «Взіалъ, възялъ, вьзял: обработка орфографической вариативности при лексико-грамматической аннотации старорусского корпуса XV-XVII вв». В: *Вестник Православного Свято-Тихоновского гуманитарного университета. Серия 3: Филология* 51.

М.К.Воронов, Д.Д.Мордашова (2016). «Горномарийские глагольные конструкции с ə̑ l'ə̑ / ə̑ lə̑ n в зоне «сверхпрошлого» и за её пределами». В: *Тринадцатая Конференция по типологии и грамматике для молодых исследователей*. Институт лингвистических исследований РАН, с. 36—39. URL: https://www.youngconfspb.com/application/files/7414/7999/6755/Tezisnik_2016.pdf.

Михаил Воронов, Алексей Кошевой (2018). «Исследование вариативности посессивных суффиксов в современном русском языке». В: *Антропология. Фольклористика. Социолингвистика*. Европейский университет в Санкт-Петербурге, с. 19—24. URL: https://eu.spb.ru/images/et_dep/afs7/AFS2018_tezisy.pdf.

Пиперски, Александр (2017). *Редупликация в русском языке*. URL: https://www.youtube.com/watch?v=QoOmgz1A5fo.