# ПРЕДСТАВЛЕНИЕ ЦЕРКОВНОСЛАВЯНСКОГО ЯЗЫКА В "СОВРЕМЕННОМ" ВИДЕ И ЕГО ОБРАБОТКА ПРИ ПОМОЩИ ПАРСЕРА, ОБУЧЕННОГО НА МАТЕРИАЛЕ СОВРЕМЕННОГО РУССКОГО ЯЗЫКА / "MODERN" REPRESENTATION OF OLD CHURCH SLAVONIC AND PROCESSING IT WITH THE PARSER TRAINED WITH MODERN RUSSIAN

**Плугарёв Матвей Романович** (plugaryov@yandex.ru)

**Сорокин Алексей Андреевич** (alexey.sorokin@list.ru)

Московский Государственный Университет им. М.В. Ломоносова, Москва, Россия

The purpose of this work is to find a way to represent texts in Old Church Slavonic so they could be processed with the parser preliminarily trained with texts in Modern Russian. Although the difference between these two languages is rather big, it is possible to modify Old Church Slavonic so that it would be similar to Modern Russian in many ways and able to be processed. This can show that sometimes it is possible to process ancient languages using the models trained with the material in one of the descendant or genetically cognate languages.

**Key words**: processing, Old Church Slavonic, Russian, Python, BERT, DeepPavlov.

## 1. Введение

Целью данной работы было изменение церковнославянского языка таким образом, чтобы тексты на данном языке могли быть обработаны при помощи модели BERT на основе DeepPavlov, заранее обученной на материале русского языка. Данная модель является более доступной в использовании, чем UDPipe, который основывается на другом принципе работы. Таким образом, мы будем производить анализ на посимвольном уровне. В конце работы данные нашего анализа будут сравниваться в том числе с данными, полученными при работе с UDPipe. Корпус современного русского языка, используемый для обучения, намного больше, чем соответствующий корпус церковнославянского, поэтому проведение подобного эксперимента представляется целесообразным, если мы примем во внимание удачный опыт подобных исследований. Удачный эксперимент подтвердил бы возможность использования языков-потомков или языков, родственных языкам потомкам, для обучения моделей, обрабатывающих древние языки,

## 2. Основная часть. Создание правил трансформации текста.

### 2.1 Подготовительный этап

Для первого этапа работы было необходимо создание кода для того, чтобы переводить церковнославянский в тот вариант, который был бы максимально близок к современному русскому на разных уровнях языка. В качестве языка разработки кода был выбран

Python (версия 3.8). В основу кода легли регулярные выражения и функция re.sub, которая позволяет заменять регулярные выражения найденные в тексте на необходимые пользователю. В качестве исходных церковнославянских текстов были выбраны материалы с сайта https://lindat.mff.cuni.cz, (корпус церковнославянского языка: 3 текста общим объёмом в 10 000 слов, отрывки из Мариинского Евангелия/Евангелие от Матфея, датируемые XI веком). Для работы с текстом было выделены основные этапы работы с текстами, в определённой степени соответствующие уровням языка.

1. Фонетико-графический
2. Стилистико-лексический
3. Морфологический
4. Синтаксический
5. Заключительный (Корректирующий) этап

**2.2 Фонетико-графический этап**

Для фонетико-графического этапа работы были произведены замены букв, которые не используются в современном русском алфавите на их созвучные (или близкие к ним/ общепринятые) аналоги (например, ѫ -> у). Также были произведены замены некоторых буквенных сочетаний в соответствии с принципами современной русской орфографии (чѧ - ча, а не чя). Таким образом, для парсера, обученного на материале русского будут исключены сложность с распознаванием букв и известных корней, что предотвратит ошибки на самом первом, примитивном этапе обработки. Важно понимать, что подобные замены имеют некоторую степени неточности, и мы получим изменения в тех словах, которых мы не хотим, но эти ошибки не столь значительны и могут быть исправлены на 3-м этапе. Главная и самая большая часть кода из этого этапа была посвящена редуцированным «ъ» (ер) и «ь» (ерь). Так как эти графемы встречаются в тексте очень часто, а в современном – намного реже, и к тому же имеют совершенно иную функцию, то их изменение было одной из важнейших задач работы. Поэтому было написано несколько функций, которые практически воспроизвели исторический процесс падения редуцированных. Так как данное явление (как и любой языковой процесс в принципе) не является автоматическим изменением, а результатом многовековых естественных изменений, то некоторая непоследовательность и отклонение от правил естественно будут возникать. Тем не менее воссозданная модель довольно точно воспроизводит формы, которые получились в результате падения редуцированных. Хотя важно заметить, что данные правила не являются контекстно свободными, так как основаны только на регулярных выражениях.
Пример**:** *домаштьнѧѧ* -> домащьняя

**2.3 Стилистико-лексический этап**

На следующем этапе были изменены словоформы, написанные под титлами. Для корней, которые сокращались таким образом, были подобраны близкие полные варианты замены. (йерслм –> иерусалим) Эти операции были проведены, так как в современном русском отсутствует запись под титлами, и это могло бы вызвать проблемы в распознавании текста.

Также были проведены замены частотных функциональных единиц («егда», «аще» и т.д.), которые способствуют созданию правильной синтаксической структуры отдельно взятого предложения.
Пример: *аще* -> если

### 2.3 Морфологический этап

Далее были проведены изменения, касающиеся морфологии. Этот этап включал в себя наибольшее количество материла, так как отдельные правила должны были быть написаны для разных частей речи. Для местоимений были написаны их современные варианты, как для личных, так и для некоторых указательных и вопросительно-относительных форм. (тебе –> тебя). Для существительных были осуществлены замены падежных окончаний в тех случаях, где это возможно. Проблема заключается в том, что многие падежные флексии омонимичны (иногда показателям других частей речи) и при их изменении могут пострадать другие части речи в большом количестве, поэтому правки затронули только некоторые склонения и их показателя в разных падежах для существительных и определённых финитных формах глагола Некоторые изменения были вынесены до правил осуществляющих «падение» редуцированных, чтобы избежать омонимии и добиться большей точности результата. Во многом это касается и прилагательных, хотя полные формы всё же удалось преобразовать достаточно серьезно. Одной из главных проблем стало наличие кратких прилагательных и в том, что их склонение совпадает со склонением существительных. Без перевода и понимания значения не всегда можно установить частеречную принадлежность, поэтому эта проблема требует дальнейшего рассмотрения в будущих работах и исследованиях. Также были осуществлены изменения в глагольных формах. Причастия после фонетических изменений могут быть распознаны правильно, поэтому данные правила затрагивают финитные формы глагола и инфинитив. Многие формы также остались без изменений, потому что их изменения повлекло бы за собой нарушение работы со всем остальным текстом, поэтому здесь изменения точно так же касались только тех показателей, которые уникальны для данного класса. Были исключены аналитические формы, которые не употребляются в современном русском (перфект, плюсквамперфект, будущее II). Также была произведена замена окончаний в неаналитических формах (имперфект, аорист, презенс). Были изменены показатели инфинитивов.
Пример: *прокаженъıѧ* -> прокаженые

### 2.4 Синтаксический этап

На этапе синтаксиса были рассмотрены 2 главные проблемы: постфикс «-ся» и порядок расположения согласованного определения относительно определяемого слова. В результате постфикс «-ся» присоединился к финитному глаголу (или другой глагольной форме). Согласованные определения (прилагательные и причастия в полной форме, некоторые местоимения), если они находились в постпозиции относительного определяемого слова (чаще всего существительного), перешли в препозицию, как при нейтральном порядке слов в предложении современного русского.
Пример: *не оубоите сѧ* -> не убоитеся

**2.5 Заключительный (Корректирующий) этап**

При проверке работы предыдущих блоков были найдены некоторые неточности. Таким образом на данном этапе были внесены корректировки для самых частых корней и слов, которые были ошибочно модифицированы.

Пример: *мозда* -> мзда

**2.6* Итоговый список изменений**

– замена букв *ѣ, ѕ, ѧ, ѫ, ѭ, ѧ, ѩ, ѡ, ѱ, ѯ, ѳ, ѵ, і,* **ѭ**, *ѥ* на современные аналоги

– замена сочетаний шипящих с *ѧ, я* и *ю* на *ча/ша/ща* и *чу/шу/щу* и т.д.

– замена сочетаний под титлами и сокращений *о҃, б҃, бж҃, гл҃, г҃, быс, йерслм, ерслм, нбс (небс, нб), цр, цс, благслв, сн, с҃, отц, члв, чкв, члк, дш, язц, цркв, евнѧ, спс, крст, ст, срдц, пр, анг* на полные варианты

– правило «падения редуцированных» (замена ъ на *о* и ь на *е* или ъ на « » и ь на ь)

– для дательного, творительного и местных падежей для u и i склонений всех чисел

– правила заменяющие окончания полных прилагательных на современные

– замена всех личных и йотовых местоимений на современные аналоги

– замена окончаний форм имперфекта и аориста на современные окончания прошедшего времени

– замена аналитических форм второго будущего, перфекта и плюсквамперфекта на современные синтетические формы

– замена частотных функциональных единиц «егда», «аще», «токмо», «еко» и «ако» на современные аналоги

– присоединение *-ся* к глаголу (переход от частицы к постфиксу)

Пример изменения предложения:

– *приближи сѧ цсрствие нбское болѧщѧѧ цѣлите мрътвыѧ вьскрѣшаите прокаженыѧ* (отрывок из оригинального текста) ->

– приближися цесарствие небесекое болящие целите мротвые вескрешаите прокаженые

**3. Проверка полученного текста и анализ результатов**

| | **Morph. accuracy** | **UAS** | **LAS** |
|---|---|---|---|
| **BERT (before)** | 71,97 | 71,86 | 60,15 |
| **BERT (before) + pre-train** | 73,26 | 73,89 | 61,89 |
| **BERT (after)** | 87,15 | 86,7 | 79,95 |

| | | | |
|---|---|---|---|
| **BERT (after) + pre-train** | 87,64 | 86,87 | 80,55 |
| **UDify** | 71.30 | 76.71 | 66.67 |
| **UDPipe** | 90.66 | 89.66 | 85.04 |

В первых двух строках таблицы представлены данные обработки русским парсером оригинального церковнославянского текста и данные, когда модель была предобучена на 5000 тысячах предложений на русском. (BERT/DeepPavlov) В 3-4 строках представлены данные, полученные при обработке текста русским парсером (в 4-й строке также с предобучением в 5000 предложений) после трансформации текста. Мы видим, значительный рост качества данных, что говорит об успешности произведённых трансформаций. Если сравнить с данными UDify и UDPipe из исследования (Dan Kondratyuk, Milan Straka «75 Languages, 1 Model: Parsing Universal Dependencies Universally» [Электронный ресурс] // arXiv:1904.02099: [сайт]. [2019]. URL:https://arxiv.org/abs/1904.02099), то можно отметить, что данные UDify, который имеет сходный принцип работы даже хуже, чем полученные в ходе эксперимента, однако результаты UDPipe более точные, и данный алгоритм лучше подходит для анализа церковнославянского языка, о чём и говорят авторы статьи. Возможно, если изначально использовать алгоритм с более качественной базовой точностью, то после трансформации можно добиться лучших результатов, и данное утверждение может являться темой дальнейших исследований в данной теме.

**4. Заключение**

Эта работа представляет собой попытку изменения оригинального текста на церковнославянском и приближения её к современному русскому варианту с использованием моделей основанных только на посимвольном анализе. При использовании алгоритма с более качественными базовыми результатами обработки (например, UDPipe), результаты, полученные после преобразования текста, также будут выше. Однако, полученные данные позволяют продолжать работу в данном направлении по улучшению качества результатов работы программы, используя другие методы и инструментарий. Можно сделать вывод о том, что этот способ действительно можно применять для мертвых языков с ограниченным корпусом текстов или для редких языков, для которых нет достаточного объёма корпусов, используя при этом модель, обученную на родственном языке, с большим количеством корпусных данных.

**Список литературы**

1. Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding» [Электронный ресурс] // arXiv:1810.04805v2: [сайт]. [2018]. URL:https://arxiv.org/abs/1810.04805v2

2. РЕМНЁВА М.Л. СТАРОСЛАВЯНСКИЙ ЯЗЫК. - 2 изд., испр. - М.: Акад. проект, 2004. - 352 с. Прил.: Ремнёва М.Л., Дедова О.В. Старославянский язык: Электронный курс. СБ.
3. Python : [Электронный ресурс] 2020. URL:https://docs.python.org/3/library/re.html
4. LINDAT/CLARIAH-CZ:[Электронный ресурс] 2020. URL:https://lindat.mff.cuni.cz
5. Сорокин А. А. Автоматический морфологический анализ на основе нейронных моделей с использованием языковых моделей // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018” 2018. URL: http://www.dialog-21.ru/media/4530/sorokinaa.pdf
6. Dan Kondratyuk, Milan Straka «75 Languages, 1 Model: Parsing Universal Dependencies Universally» [Электронный ресурс] // arXiv:1904.02099: [сайт]. [2019]. URL:https://arxiv.org/abs/1904.02099
7. Milan Straka, Jana Straková, Jan Hajič «Evaluating Contextualized Embeddings on 54 Languages in POS Tagging, Lemmatization and Dependency Parsing» [Электронный ресурс] // arXiv:1908.07448 [сайт]. [2019]. URL:https://arxiv.org/abs/1908.07448