# Автоматическое распознавание новообразований в "русском английском" текстов разных жанров

*Анастасия Буракова, БКЛ181*
*Научный руководитель: А. В. Виклова*

Для начала определим основные понятия, используемые в работе:

- Новообразование — слово или группа слов, образованные с ошибкой.
- Лингвистическая интерференция — «взаимодействие языковых систем в условиях двуязычия, ... которое выражается в отклонениях от нормы и системы второго языка под влиянием родного».

Данная работа посвящена создания программы автоматического распознавания новообразований в русском английском.

Интерес данной работы заключается в том, что на данный момент системы автоматического распознавания новообразований не существует, и даже системы распознавания неправильных слов, доступные на  некоторых сайтах и сервисах, пока имеют заметный процент ошибок (см. Таблица 1).

Итак, первым этапом работы был сбор данных для будущего корпуса. Программой, написанной на Python с опорой на пакет pyspellchecker, было обработано 5393 текста из корпуса REALEC (состоящего из эссе на английском языке, написанных носителями русского языка). Из полученных 20042 слов с ошибками вручную были удалены слова с опечатками, слова, неправильно распознанные как ошибочные, и все повторения. В результате осталось 192 уникальных новообразования, из которых и был составлен корпус.

Следующим этапом была разметка получившегося корпуса. Всем новообразованиям были присвоены теги по каждому из следующих параметров: значение, лингвистический процесс, локация ошибки, тип ошибки, вероятность лингвистической интерференции, источник, и количество вхождений в корпусе enTenTen18 (см. пример в Таблица 2).

По параметру лингвистической интерференции было создано три тега: 2 —  для случаев, когда интерференция очевидна, 1 —

для случаев, когда интерференция возможна, и 0 — для случаев, когда вероятность интерференции мала или отсутствует

Один из тегов параметра «лингвистический процесс» — **калька**. Новообразования, размеченные этим тегом, делятся на 4 вида:

1) Слова, полученные транслитерацией, такие, как *reglament*, *abonement*.
2) Слова, полученные в результате дословного перевода по частям, например, *not-spoiling* (*непортящийся*), *timespending* (*времяпрепровождение*).
3) Новообразования, полученные в результате кальки, с ошибкой в префиксе, *disbalanse* вместо *imbalance*.
4) И, наконец, слова, образованные с ошибкой в суффиксе: *immunitet* вместо *immunity*, *tendention* вместо *tendency*.

   Всем новообразованиям с тегом «калька» был присвоен тег 2 по параметру лингв интерференц; при этом интересно отметить, что в корпусе есть слова, образованные как будто вопреки нормам русского языка, например, слово *acupuncthurism*, получившее лишний суффикс -ism, на русском звучит как «акупунктура», то есть, созвучно правильному английскому варианту *acupuncture*.

Другой интересный тег — вариант параметра «тип ошибки» **compounding**, обозначающий составление слова из более чем одной основы. Размеченное этим тегом новообразование *drugselling* на первый взгляд кажется похожим на *timespending*, но на самом деле не является результатом дословного перевода, и по параметру возможности интерф размечено тегом «0».

Кроме тега, обозначающего слияние двух слов, есть и тег разъединения — **splitting**. Пока единственный пример этого типа ошибки — коллокация *out-of the day* (употребленная вместо слова *outdated*). У этого слова вероятность лингв. инт. также равна нулю.

Еще один тег параметра «тип ошибки» — добавление **лишней морфемы**. Лишние префиксы встречаются только в новообразовании *imprisoners*, бóльшая часть «ненужных» морфем — суффиксы. В первую очередь они встречаются у существительных: -ing служит суффиксом для новообразований, полученных от глаголов (*challengings*, *travellings*), *-ness* и *-ment* — для образования абстрактных существительных, таких, как *beautiness* (вместо *beauty*) и *decreasement* (вместо *decrease*). К

этому же тегу относятся формы мн. числа неисчисляемых существительных (*knowledges*, *advices*, *sportsmans*).

Еще один тип ошибки — использование **неправильной морфемы**. Неправильные корни встречаются преимущественно в кальках, а неправильные префиксы — в отрицаниях, например, у прилагательных (*discomfortable* вместо *uncomfortable*) или существительных (*unstability* вместо *instability*). Интересный вариант этой ошибки — новообразования, в которых неправильной является не присоединяемая в процессе словообразования морфема, но основа, к которой она присоединяется, как в случае со словами *destroyment* (вместо *destruction*) и *qualificated* (вместо *qualified*). Отметим, что в некоторых случаях при таком образовании происходит наращение основы (как в слове *huntering* вместо *hunting*), а в других случаях — наоборот усечение (как в случае *summaring* вместо *summarizing*). Кроме того, в случае новообразования descripts (вместо *describes*) можно предположить, что слово образовано с оглядкой на существительное *description* и буква *t* в нем является рудиментом номинализационного суффикса *-tion*.

На основе собранного и размеченного корпуса работает программа автоматического распознавания новообразований. Она получает на вход файл-корпус в формате csv и файл с обрабатываемым текстом в формате txt. Все найденные в тексте новообразования, их параметры, а также контекст употребления в обрабатываемом тексте записываются в файл результатов (в формате txt). Если в тексте не было найдено новообразований, программа выводит сообщение об этом в консоль.

Код программы с подробным описанием, все данные и примеры работы обеих программ доступны в открытом доступе в GitHub репозитории проекта.

Созданная программа работает без ошибок, и в дальнейшем возможно дополнительно проверить ее на новых поступлениях в корпус REALEC. Кроме того, предлагается обрабатывать новые поступления и программой сбора данных, тем самым постепенно увеличивая корпус новообразований.

Возможно добавление написанной программы на сайт корпуса REALEC и к коду находящейся в разработке системы ADWISER.