Spelling Error Patterns in Brazilian Portuguese

Priscila A. Gimenes (EACH/USP), Norton T. Roman (EACH/USP), Ariadne M. B. R. Carvalho (Institute of Computing/Unicamp)

Согласно Дамеро (1964), 80% ошибок попадают в 4 категории: 1) вставка лишней буквы; 2) пропуск буквы; 3) замена одной буквы на другую; 4) перестановка двух букв местами. И хотя эта работа была опубликована более пятидесяти лет назад, эти данные до сих пор лежат в основе всех state-of-the-art исследований и разработок в проверке орфографии благодаря её лёгкой применимости на основе статистики и вероятностей.

Целью авторов статьи было проверить, насколько данные Дамеро применимы к языкам, отличным от английского, т.к. сам Дамеро использовал статистику только родного языка, но при этом во всех (или в большинстве) работ они применяются без адаптации для рассматриваемого языка. Таким образом, авторы поставили себе задачу собрать необходимую статистику по несловесным (non-word) опечаткам в текстах на бразильском португальском языке, сравнить полученные данные с подобной статистикой для испанского и затем — с данными Дамеро.

Авторы исследования использовали корпус Романа (2013), состоящий из 1808 текстов (62858 слов), напечатанных 452 участниками онлайн-эксперимента без помощи спеллчекеров (корпус C_1). Кроме того, для проверки данных корпуса C_1 был собран корпус C_2 (192 текста, 26418 слова) на основе сообщений в блогах. Сравнивая два этих корпуса, авторы хотели избавиться от возможных biases на основании стиля письма или бэкграунда авторов. Затем полученные тексты были загружены в текстовый редактор, который выделил все возможные опечатки, после чего один из исследователей вручную их просмотрел, после чего они были распределены по группам: помимо 4 оригинальных групп, описанных выше, они выделили три дополнительных **а)** использование дикратических знаков; **б)** использование седилей (из-за особенностей добавления седиля к букве с на португальской клавиатуре могут возникать опечатки); **в)** использование пробелов.

Чтобы выдержать стиль Дамеро, авторы делят ошибки из группы **а** на следующие: 1) отсутствие диакритического знака; 2) наличие лишнего

диакритического знака; 3) использование верной диакритики на неверной букве; 4) использование неверной диакритики на верной букве. Кроме того, они выделили ещё одну группу прочих — недостаточно частотных опечаток, которые не вошли ни в одну из групп.

Результаты были представлены в виде двух таблиц, в которых перечислялись все виды опечаток, их частоты встречаемости, количество опечаток в слове (одна, две или три) и процентное соотношение от общего числа опечаток. В обоих корпусах было примерно одинаковое количество ошибок, но их пропорция отличается: в C_1 средняя доля ошибок 1,81%, а в C_2 - 4,77%, что было вызвано, скорее всего, более высоким уровнем образования авторов в случае первого корпуса или с “разговорностью” стиля письма в блогах. Относительно распределения ошибок по категориям, согласно критерию Колмогорова-Смирнова, значительных различий в корпусах не было.

Исследование показало, что почти половина опечаток относилась к использованию диакритик, а вместе с ошибкой в употреблении седиля — даже больше (58,84% для C_2). Если рассматривать эти два типа ошибок как 3 ошибку в классификации Дамеро (замену), то вместе с прочими заменами общее количество таких опечаток составит почти 90% для обоих корпусов, что подтвердит данные Дамеро. Были подтверждены наблюдения Дамеро, что свыше 85% всех слов с опечатками имеют только одну опечатку (для корпусов авторов 92,3% и 86,3%). В целом все статистические наблюдения Дамеро были подтверждены (с учтением седилей и диакритик).

Кроме того, авторы сравнили свои результаты с результатами подобного исследования для испанского языка (*Bustamante, Arnaiz, and Gines (2006)*). Их анализ показал, что нет значительных различий в распределении опечаток с учётом диакритик, существующих в испанском. Авторы предполагают, что существующее распределение ошибок вообще не связано с культурой или языком, а с набором букв (и соответствующих диакритических знаков), “разрешенным” этим языком. Более того, это приводит их к выводу, что их данные показывают важность учета диакритических знаков при разработке систем проверки правописания ввиду их одинаково большой доли среди опечаток как в испанском, так и в португальском языках (>40%).

На мой взгляд, это достаточно полезное исследование, т.к. позволяет улучшить системы проверки правописания для тех языков, которые ещё не так хорошо разобраны в плане компьютерно-лингвистических инструментов и страдают от “англоцентричности” современного состояния науки (а эта работа, вероятно, способна подтолкнуть исследователей и разработчиков куда более маленьких языков, чем португальский). Однако, на мой взгляд, сравниваемые корпуса были не очень удачно составлены: во-первых, один из них был намного больше другого (в 2,4 раза), при этом количество опечаток примерно одинаковое; авторы оправдывают это уровнем образования авторов, однако это не отменяет того, что выборки для их исследования не были одинаково хорошо сбалансированы, что могло повлиять на результат исследования (и в некоторых случаях виден очень большой разрыв в результатах по корпусам, хотя в целом и примерно одинаковое); однако отчасти эта проблема снимается тем, что данные совпадают с данными других исследователей для испанского и исходными данными по английскому языку.