Вариативность орфографий в идише

Д. Кирьянов, Б. Орехов, Т. Панова
НИУ ВШЭ

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Постановка проблемы

- Сейчас парсер корпуса идиша работает только с одной орфографией;
- она считается нормативной, но она введена в 1913 г.;
- огромное количество текстов (XIX-XX вв.) написано в других орфографиях;
- эти тексты пока невозможно включить в корпус;
- необходим орфографический нормализатор

Постановка проблемы

У пользователя должна быть возможность посмотреть выдачу корпуса и в латинице. Соответственно, наш нормализатор должен:

- уметь нормализовать орфографию
- транслитерировать
- показывать в выдаче корпуса исходный вариант в оригинальной орфографии, но при этом правильно размеченный
- выдавать транслитерированный в латиницу вариант – как оригинальный, так и нормализованный

Ход работы

- Обзор орфографических традиций свидетельствует об отсутствии чётких правил внутри каждой из них.
- Исчисление возможных отклонений от нормативной орфографии: единицы какого уровня имеют вариативность – только буквы? морфемы? слова?
- Что должна знать программа?
- Создание нормализатора
- Создание транслитератора
- Подключение их к процессу обработки текстов для корпуса.

Орфография идиша и проблемы нормализации

- Идиш пользуется еврейским квадратным письмом, в основном слова записываются фонетически. Где наблюдается вариативность?
- Исключения из фонетического принципа записи – заимствования из семитских языков (записываются консонантным письмом):
- מיר- mir - [mir] VS כּל- kl - [kol], но в некоторых издательских практиках записываются фонетически

Орфография идиша и проблемы нормализации

- Для различения некоторых букв в разных орфографических традициях используется либо диакритика (как в нормативной), либо «немые» буквы א- a, ה– h
- Произношение некоторых морфем отошло от их традиционного написания, В некоторых орфографиях отсутствует диакритика. Поэтому некоторые графемы становятся неразличимы:
- אַ/ אָ– a/o → א– a, o
- И некоторые другие

Технология и тестирование

- Алфавит со всеми вариантами букв
- Проблемные морфемы, буквосочетания и слова с вариантами
- модуль для гебраизмов
- n-граммы, HMM для букв без диакритики
- точность - 98% для текстов в нормативной орфографии (2% - заимствования из семитских языков), 94-97% для текстов в других орфографиях
- полнота - 100%