Отчет по созданию жанровых корпусов

Для исследования особенностей функционирования и распределения предлогов в текстах различной жанрово-стилевой направленности были собраны равновесные корпуса по основным типам функциональных стилей: научного, делового, публицистического, литературно-художественного и разговорного. При составлении корпусов мы придерживались ряда принципов с целью создания языковой модели, предельно приближенной к реальному языку:

1) жанрово-стилевое соответствие языкового материала: для обеспечения достоверности получаемых статистических данных не допускалось явное смешение функциональных стилей и жанров внутри одного корпуса;

2) репрезентативность с точки зрения современного русского языка: был отобран языковой материал начиная с 1960-х годов для фиксации только актуальных языковых тенденций;

3) равновесность корпусов во избежание искажения репрезентации того или иного стиля: объём каждого корпуса – около 1 млн 250 тысяч токенов;

4) тематическое разнообразие, отражающее многообразие реалий языковой культуры: языковой материал корпусов относится к различным сферам жизни и соответствует значительному множеству языковых ситуаций;

5) разнообразие авторов: поскольку служебные слова являются средством атрибуции текста авторскому стилю, был ограничен допустимый объём языкового материала, принадлежащего одному автору;

6) равновесность текстов: в каждый корпус были включены тексты сопоставимых объёмов с целью избежать излишнего влияния стиля одного автора.

В дополнение к следованию вышеописанным принципам в целях сохранения естественности языкового материала было принято решение максимально ограничить включение в корпуса текстов, являющихся переводами с иностранных языков.

Тексты, вошедшие в корпуса, были собраны с применением сервисов SketchEngine (технология WebBootCat) и ручного отбора.

По итогам работы мы получили пять корпусов:

1. Наука и техника: корпус текстов научного и научно-популярного стилей, относящихся к следующим областям: астрономия, космонавтика, биология, геология, география, информатика, математика, физика, химия, медицина, сельское хозяйство, архитектура, строительство, технологии, транспорт, радиотехника, металлургия, машиностроение, лингвистика, социология, экология, философия, экономика.

2. Корпус законодательных текстов: собрание текстов федеральных законов с 1994 по 2009 год.

3. Публицистика: корпус, включающий в себя тексты газетных и журнальных статей (новостные тексты, заметки, аналитические статьи, интервью, обзоры), опубликованных в печатных и интернет-изданиях между 1999 и 2019 годами.

4. Художественная литература: корпус произведений русскоязычных авторов второй половины XX — начала XXI века. В корпус были включены преимущественно тексты объёмом от 20 до 40 тысяч токенов.

5. Разговорный корпус: собрание расшифрованных записей радиопередачи «Эхо Москвы» с 2000 по 2020 год, субтитров русскоязычных фильмов с 1960 года, а также фрагмент «Сбалансированной аннотированной текстотеки» (корпус спонтанных монологических текстов, методика Н.В. Богдановой-Бегларян), включающий в себя пересказы текстов, описания изображений и рассказы на

заданную тему трёх профессионально-ориентированных групп носителей языка (медики, юристы, студенты).

Принципы:

- репрезентативность с т.з. современного русского языка (последние 50 лет, преимущественно с 2000 г)
- жанрово-стилевое соответствие
- равновесность корпусов
- равновесность текстов
- разнообразие авторов (служебные слова как средство атрибуции/примера стиля)
- тематическое разнообразие
- практически полное отсутствие переводовю

Технология/методика сбора текстов:
ручной отбор / кроулер –> ручной отбор

Количественные характеристики:
5 подкорпусов, каждый объёмом около 1 млн 250 тыс токенов
Научный – по 78 тыс на область
Художественный – только русские писатели, тексты по ~ 20-40 тыс
Деловой: законы с 1994 по 2009
Разговорный: «Эхо Москвы» – разнообразие спикеров; примерно по полмесяца за каждый год с 2000 по 2020 г, 771 тыс токенов; субтитры – русские фильмы с 1960 по 2019 г за каждый год, 367,5 тыс; CAT – пересказ/описание/рассказ, юристы/медики/студенты, средний возраст – 33/41/22, 98 тыс; прямая трансляция открытия/закрытия Олимпиады-2014 — 24 тыс.

Публицистика: 625 тыс — корпус статей различных изданий с ~1999 по ~2006 год, 625 тыс — публикации на тему Петербурга за осень 2019 года: