Введение в компьютерную лингвистику. Отчёт о практической работе.

Деткова Юлия, 493 группа

Информация о корпусе:

В качестве материала для исследования было выбрано творчество Аркадия и Бориса Стругацких. Корпус составлен из следующих произведений:

- Трудно быть богом
- Понедельник начинается в субботу
- Путь на Амальтею
- Улитка на склоне
- Обитаемый остров
- Сказка о тройке
- Волны гасят ветер
- Стажеры.
- Попытка к бегству
- Пикник на обочине

Суммарный объем корпуса составил 503223 слов.

Было составлено несколько частотных словарей: словарь лемм, словоформ, диграмм, триграмм и четыреграмм. Найдена статистика употребления времен и залогово глаголов, частей речи, падежей.
Для работы был использовая язык программирования python, библтотека pymorphy2.

Работа над корпусом:

1. Токенизация. Использованы регулярные выражения для очистки от знаков пунктуации.
2. Приведение всех словоформ у нормальному виду
3. Составление частотных словарей для лексем и словоформ
4. Составление частотных словарей для n-грамм (n = 2,3,4)
5. Нахождение статистики по использованию частей речи, времен глаголов, падежей и залогов.

Результаты работы:

1. Были построены графики зависимости количества употреблений слова от его номера в частотном словаре для проверки закона Ципфа: если все слова языка (или просто достаточно длинного текста) упорядочить по убыванию частоты их использования, то частота n-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру n.
Коэффициент пропорциональности для произведение А. И Б. Стругацких равен трём.

2. Частотные словари для 1000 самыч частых экземпляров представлены в приложении

3. Статистика по использованию частей речи (в процентах):
INFN 0.0239628560773
NOUN 0.25980402911
NUMR 0.00607697232355
INTJ 0.00385320332429
PRCL 0.0576484689899
PRED 0.00614877114326
ADJS 0.0100997006388
ADVB 0.0718008141188
GRND 0.00939567332335
VERB 0.152877636862
PRTF 0.0093258689153
COMP 0.00351215893068
CONJ 0.103204820094
ADJF 0.100785598752
PREP 0.0944174423266
PRTS 0.00280813161521
NPRO 0.0842778534546

4. Статистика по использованию времен:
past 0.670626274801
futr 0.0715626738365
pres 0.257811051363

5. Статистика по использованию залогов:
pssv 0.606344510191
actv 0.393655489809

6. Статистика по использованию падежей:
gen2 0.000706300372649
gent 0.242079036312
datv 0.0625184158073
nomn 0.442040038132
loc2 0.00218823121588
loct 0.0524178871653
voct 0.000437646243175
ablt 0.0801759251235
accs 0.117436519629

Заключение:
Для глаголов исследуемого корпуса характерны: время – прошедшее, залог – пассивный. Основная часть речи – существительное, падеж – именительный.

Резуьлтаты анализа корпуса Стругацких были сравнены с результатами анализа Грибоедова и Достоевского.

Автор	Часть речи	Падеж	Время	Залог
А. и Б. Стругацкие	существительное	именительный	прошедшее	пассивный
А. С. Грибоедов	существительное	именительный	настоящее	действительный
Ф. М. Достоевский	существительное	именительный	прошедшее	действительный