Анализ перевода авторских номинаций как маркера идиостиля переводчика (на примере романа Дж.Р.Р. Толкина “Властелин Колец”)

Курсы ДПО "Компьютерная лингвистика"

Учебный проект №2

Студеникина Д.

Основные понятия

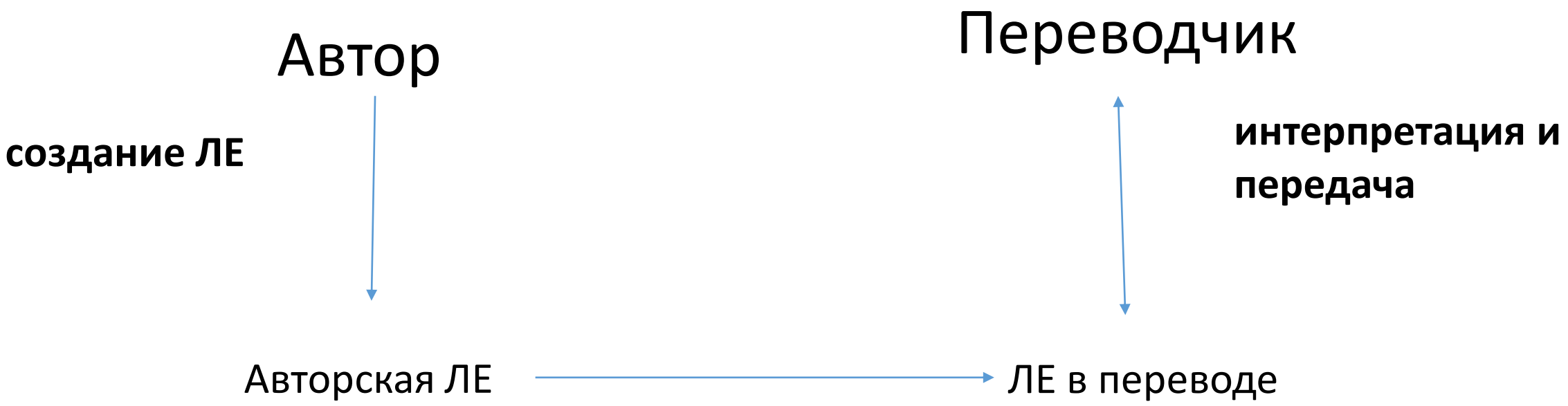
- *Номинация* — процесс и результат именования, то есть присвоения имени обозначаемому объекту.

- *Идиостиль* – совокупность характерных языковых и текстовых особенностей, которые отличают носителя языка

- *Языковая личность* — это носитель языка, обладающий уникальной совокупностью вербальных, когнитивных и мотивационных характеристик, которая проявляется в создаваемых им речевых произведениях.

Языковая личность переводчика в аспекте перевода авторских номинаций

Языковая личность в процессе перевода



Гипотеза

Выбор способа перевода авторской лексики является достаточно устойчивым и уникальным для каждого переводчика, чтобы служить основой для автоматической атрибуции перевода

Цель работы

Оценить степень уникальности идиостиля каждого переводчика на основе анализа перевода авторской лексики

Задачи:

- Подготовить корпус параллельных контекстов (оригинал + 3 перевода)
- Обучить модель узнавать переводчика на основе перевода авторских номинаций (на корпусе A)
- Проверить, узнает ли модель переводчика (на корпусе B с другим набором номинаций и на корпусе C с замаскированными номинациями)

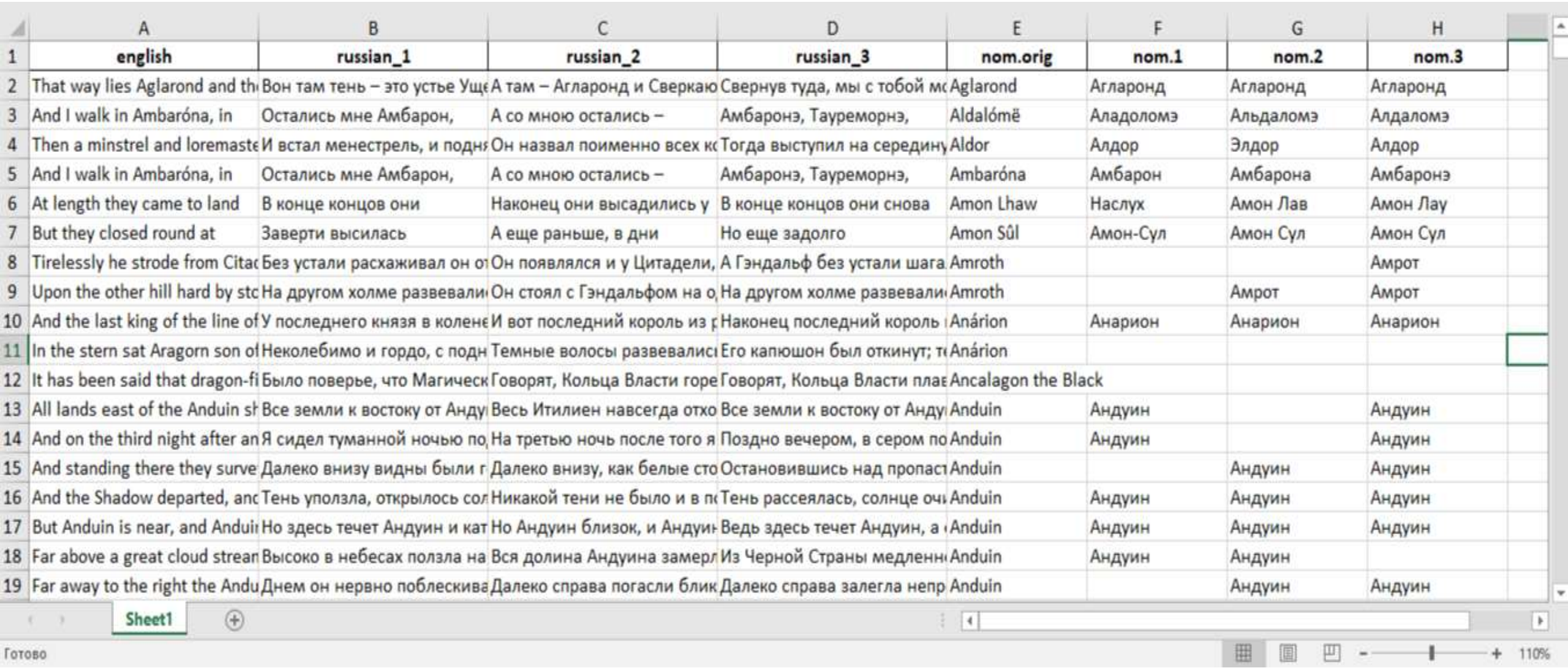
Этапы:

- Подготовка корпуса параллельных переводов (оригинал + переводы)
- Аннотировать параллельный корпус
- Разделить на корпусы
- Обучить модель на корпусе А, используя таблицу переводов номинаций
- Проверить способность предсказания переводчика по корпусу В
- Проверить способность предсказания переводчика по корпусу С

Таблица номинаций и их переводов

	A	B	C	D
1	original	1	2	3
2	Zirak-zigil	Зиракзигил	Зирак Зигиль	Зирак-Зигил
3	Zirak	Зирак		Зирак
4	Yule	Просечень		Юл
5	Yellow Face	Желтая Морда	Желтый Лик	Желтое Лицо
6	Yale, lowlands of the	Йельские Низины		Йельская низина
7	Wraith-lord			Король Призраков
8	Wraith-king			Король-Призрак
9	Woses	лешаки	Лесные Люди	Воосы
10	Wormtongue	Гнилоуст	Червослов	Червеуст
11	Worm	Гниль	Червяк	Червяк
12	World's End			край мира
13	Woody End	Лесной Угол	Залесье	Лесной Угол
14	Woodhall	Лесной Чертог	Задоры	Лесной Приют
15	Wood-elves	лесные эльфы		Лесные Эльфы
16	Wold of Rohan	раздолье Ристании		Волд, Стена Эмин Муйла
17	Wold			Волд
18	Wizard's Vale	Колдовская логовина	Колдовская Долина	Чародеева Долина
19	Withywindle	Ветлянка	Ивлинка	Ивий Вьюн
20	Withered Tree			Увядшее Дерево

Пример корпуса



	A	B	C	D	E	F	G	H
1	english	russian_1	russian_2	russian_3	nom.orig	nom.1	nom.2	nom.3
2	That way lies Aglarond and th	Вон там тень – это устье Уще	А там – Агларонд и Сверкаю	Свернув туда, мы с тобой мс	Aglarond	Агларонд	Агларонд	Агларонд
3	And I walk in Ambaróna, in	Остались мне Амбарон,	А со мною остались –	Амбаронэ, Тауреморнэ,	Aldalómë	Аладоломэ	Альдаломэ	Алдаломэ
4	Then a minstrel and loremaste	И встал менестрель, и подня	Он назвал поименно всех ко	Тогда выступил на середину	Aldor	Алдор	Элдор	Алдор
5	And I walk in Ambaróna, in	Остались мне Амбарон,	А со мною остались –	Амбаронэ, Тауреморнэ,	Ambaróna	Амбарон	Амбарона	Амбаронэ
6	At length they came to land	В конце концов они	Наконец они высадились у	В конце концов они снова	Amon Lhaw	Наслух	Амон Лав	Амон Лау
7	But they closed round at	Заверти высилась	А еще раньше, в дни	Но еще задолго	Amon Sûl	Амон-Сул	Амон Сул	Амон Сул
8	Tirelessly he strode from Cita	Без устали расхаживал он от	Он появлялся и у Цитадели,	А Гэндальф без устали шага	Amroth			Амрот
9	Upon the other hill hard by stc	На другом холме развевали	Он стоял с Гэндальфом на о	На другом холме развевали	Amroth		Амрот	Амрот
10	And the last king of the line of	У последнего князя в колене	И вот последний король из р	Наконец последний король	Anárion	Анарион	Анарион	Анарион
11	In the stern sat Aragorn son of	Неколебимо и гордо, с подн	Темные волосы развевалис	Его капюшон был откинут; т	Anárion			
12	It has been said that dragon-fi	Было поверье, что Магическ	Говорят, Кольца Власти горе	Говорят, Кольца Власти плав	Ancalagon the Black			
13	All lands east of the Anduin sh	Все земли к востоку от Анду	Весь Итилиен навсегда отхо	Все земли к востоку от Анду	Anduin	Андуин		Андуин
14	And on the third night after an	Я сидел туманной ночью по	На третью ночь после того я	Поздно вечером, в сером по	Anduin	Андуин		Андуин
15	And standing there they surve	Далеко внизу видны были г	Далеко внизу, как белые сто	Остановившись над пропаст	Anduin		Андуин	Андуин
16	And the Shadow departed, anc	Тень уползла, открылось сол	Никакой тени не было и в по	Тень рассеялась, солнце очи	Anduin	Андуин	Андуин	Андуин
17	But Anduin is near, and Anduir	Но здесь течет Андуин и кат	Но Андуин близок, и Андуин	Ведь здесь течет Андуин, а	Anduin	Андуин	Андуин	Андуин
18	Far above a great cloud strean	Высоко в небесах ползла на	Вся долина Андуина замерл	Из Черной Страны медленн	Anduin	Андуин	Андуин	
19	Far away to the right the Andu	Днем он нервно поблескива	Далеко справа погасли блик	Далеко справа залегла непр	Anduin		Андуин	Андуин

Sheet1

Готово

110%

Random Forest

Обучение на корпусе A определять переводчика

Проверка обобщения/запоминания на корпусе B

Проверка на корпусе C – насколько падает узнавание переводчика без номинаций

RF

Корпус А

Обучающая выборка: 3937 примеров

Тестовая выборка: 985 примеров

РАСПРЕДЕЛЕНИЕ В ТЕСТОВОЙ ВЫБОРКЕ:

Переводчик 1: 373 примеров (37.9%)

Переводчик 2: 282 примеров (28.6%)

Переводчик 3: 330 примеров (33.5%)

RF

Корпус А

Точность: 59.09%

(случайное угадывание: 33%)

	precision	recall	f1-score	support
Переводчик 1	0.51	0.99	0.67	373
Переводчик 2	0.77	0.18	0.29	282
Переводчик 3	0.86	0.49	0.62	330
accuracy		0.59	985	
macro avg	0.71	0.55	0.53	985
weighted avg	0.70	0.59	0.55	985

RF

Корпус А

МАТРИЦА ОШИБОК:

	предсказано		
	1	2	3
Реально 1:	[369	1	3]
Реально 2:	[207	51	24]
Реально 3:	[154	14	162]

RF

Корпус А

Точность: 58.05%

Большинство фрагментов приписано Переводчику 1

RF

Корпус B

	precision	recall	f1-score	support
Переводчик 1	0.42	0.99	0.59	506
Переводчик 2	0.87	0.13	0.22	506
Переводчик 3	0.91	0.42	0.58	506
accuracy		0.51	1518	
macro avg	0.73	0.51	0.46	1518
weighted avg	0.73	0.51	0.46	1518

RF

Корпус В

МАТРИЦА ОШИБОК (корпус Б):

	предсказано		
	1	2	3
Реально 1:	[502	0	4]
Реально 2:	[425	65	16]
Реально 3:	[282	10	214]

СРАВНЕНИЕ С КОРПУСОМ А:

Корпус А (известные имена): 58.05%

Корпус Б (новые имена): **51.45%**

Разница: -6.60%

RF

Корпус B

Точность 51.45%

Много фрагментов приписано Переводчику 1

Модель хорошо обобщает новые имена

RF

Корпус C

Маскировка номинации = [nom]

МАТРИЦА ОШИБОК (корпус C):

предсказано

	1	2	3
Реально 1:	[374	0	0]
Реально 2:	[374	0	0]
Реально 3:	[374	0	0]

RF

Итог

Точность на корпусе С без имен: **33.33%**

(корпус А с именами: 58.05%)

(корпус Б с новыми именами: 52.17%)

Вклад имен: 51.84%

Линейная регрессия

Обучение на корпусе A определять переводчика

Проверка обобщения/запоминания на корпусе B

Проверка на корпусе C – насколько падает узнавание переводчика без номинаций

LR

Корпус А

МАТРИЦА ОШИБОК:

предсказано

	1	2	3
Реально 1:	[302	46	25]
Реально 2:	[75	183	24]
Реально 3:	[64	85	181]

Точность: 67.61%

LR

Корпус B

МАТРИЦА ОШИБОК:

предсказано

	1	2	3
Реально 1:	[359	137	10]
Реально 2:	[128	357	21]
Реально 3:	[68	178	260]

Точность: 64.30% (корпус A: 67.61%)
Разница: -3.32%

LR

Корпус С

МАТРИЦА ОШИБОК:

	1	2	3
Реально 1:	9	365	0
Реально 2:	3	371	0
Реально 3:	3	371	0

Точность: 33.87%

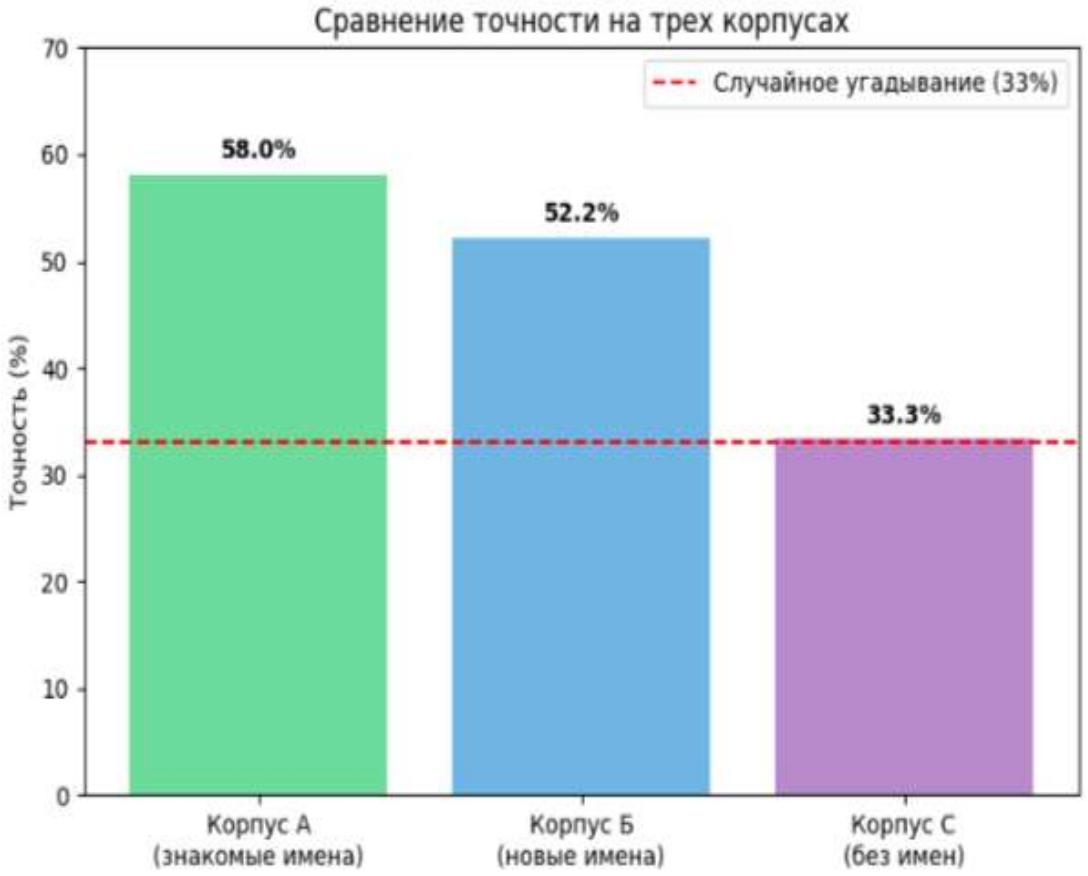
LR

Итог

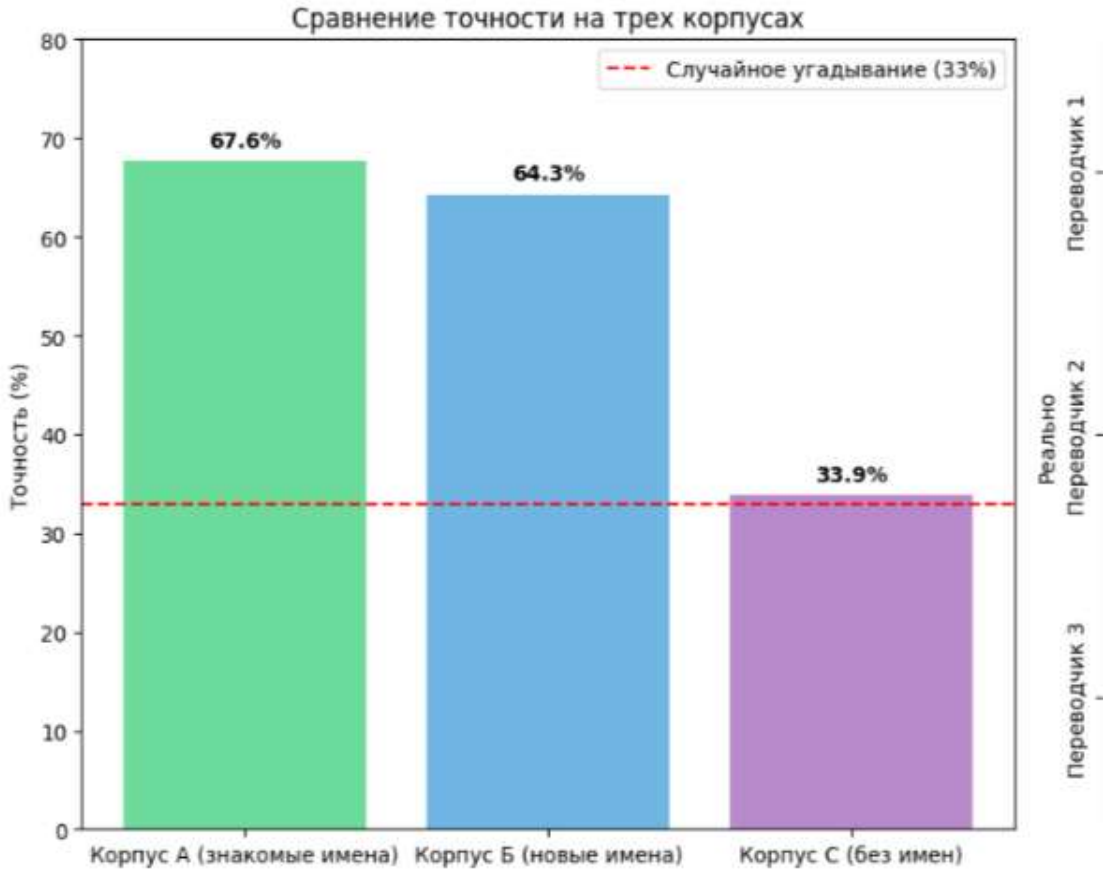
Корпус	**Точность %**
А (знакомые имена)	67.614213
Б (новые имена)	64.295125
С (без имен)	33.868093

Сравнение
Random Forest и Линейной регрессии

Random Forest



Linear Regression



Проблемы исследования

Отсутствует эксперимент по оценке стилистических различий (без фокуса на именах)

Решение:

1. Обучение на стилистических особенностях
2. Обучение на авторских номинациях
3. Сравнение

Сложности при исследовании

Проблемы с формированием корпуса

Проблемы с таблицей номинаций

Особенности авторских номинаций

составные имена + вторичная номинация

Original	1	2	3
accursed Tower	-	-	Проклятая Башня
Orthanc	Ортханк (Отрханская крепость)	Ортханк	Орфанк

Проблемы с поиском имен в корпусе

Вторичная номинация

Original	1	2	3
White Face	Белая Морда	Белый Лик	Белое Лицо

“Эовин стояла [...] с белым лицом…”

Проблемы с поиском имен в корпусе

king	конунг	повелитель	король
king	повелитель	повелитель	король
lord	конунг	король	король

Решение

Сосредоточиться на уникальных номинациях

Bag End, Under-Hill	Торба-на-Круче	Засумки	Холм под Котомкой
Éowyn	Эовин	Йовин	Эовейн
Amon Sûl	Амон-Сул	Амон Сул	Амон Сул

Разделить номинации на группы (антропонимы, топонимы, события и др.)

Инструментарий

Библиотеки:

Pandas

NumPy

Scikit-learn

razdel

Sentence Transformers

Matplotlib

Методы:

LinearRegression

Random Forest

Спасибо за внимание