ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа

«Фундаментальная и компьютерная лингвистика»

Панасюк Анастасия Васильевна

КОЛИЧЕСТВЕННЫЕ ХАРАКТЕРИСТИКИ ЯЗЫКА ХУДОЖЕСТВЕННЫХ ПЕРЕВОДОВ С ЯПОНСКОГО НА РУССКИЙ

QUANTITATIVE FEATURES IN JAPANESE-RUSSIAN TRANSLATIONS

Выпускная квалификационная работа студентки 4 курса бакалавриата группы 182

Академический руководитель
образовательной программы
канд. филологических наук, доц.
Ю.А. Ландер

_____

« » _____ 2022 г.

Научный руководитель

канд. филологических наук, доц.
Б.В. Орехов

_____

Москва 2022

ОГЛАВЛЕНИЕ

1. Введение

Язык в переводных текстах отличается от языка, которым написаны тексты, созданные на нем изначально: статистически различается употребление грамматических конструкций, сочетаемость функциональных и лексических слов настолько, что язык переводов даже может считаться отдельным диалектом (Gellerstam 1986). Это явление преимущественно описывают через термин translationese (Gellerstam 1986), но термины third-code (Baker 1993), interlanguage (Toury 1985) также в ходу.

На отличие диалекта переводов влияют как переводческие универсалии (‘translational universals’) — черты, которые проявляются вне завсимости от языка, с которого создают перевод, обоснованные самим процессом переноса смыслов на другой язык: симплификация, эксплиситация, нормализация, конкретизация и другие (Baker 1993) — так и черты, обусловленные языком, с которого переводят (lingvo-specific items).

Существовавние особого диалекта подтверждает множество работ. Его описывают теоретические работы из русла переводоведения (напр. Gellerstam 1986; Baker 1993; Laviosa-Braithwaite 1998), описывают через методологию корпусной лингвистики и машинного обучения (напр. Puurtinen 2003; Baroni, Bernardini 2006). В последнее время актуальными стали исследования переводческих диалектов при помощи нейросетей (напр. Pylypenko и др., 2021) и, в частности, исследования языка переводов, которые были сгенерированы нейросетями при помощи языковых моделей (machine translation) (Parthasarathi и др., 2021). Тем не менее, каждая языковая пара языка-оригинала (source language, SL) и целевого языка (target laguage, TL) заслуживает отдельного рассмотрения, потому что в каждом конкретном случае взаимодействие проявляется уникальным образом (Kunilovskaya, Lapshinova-Koltunski 2020).

Задача этой работы — установить, какие особенности у языка переводов с японского на русский в художественных текстах. Цель работы: определить, какое влияние морфосинтаксические особенности японского оказывают на русский язык в художественных текстах, и, если оно наблюдается, описать его черты.

Методологически я следую анализу языка переводов посредством корпусного анализа и машинного обучения (Baroni, Bernardini 2006, Kunilovskaya Lapshinova-

Koltunkski 2006). Главная задача исследования — выделить интепретируемые через лингвистическую оптику черты переводческого диалекта художественных текстов, переведенных с японского на русский, а не добиться высшего качества в классификации. По этой причине нейросетевые подходы, показывающие высшие результаты, но трудно интерпретируемые, не применяются (Pylypenko и др., 2021).

Многие работы описывают язык переводов с английского на русский (среди прочих Kunilovskaya и др., 2021; Kunilovskaya, Kutuzov 2017), русский также исследуют на ряду с другими европейскими языками в больших компаративных исследования (Baroni, Bernardini, 2006; Pourdamghani и др, 2019). Относительно японского, описывается язык переводов с английского и шведского на японский (Meldrum 2009, Svanberg 2017).

Стрижак пишет в работах о японской языковой картине мира об особенностях русско-японских и японско-русских переводов на основе корпусного подхода (Strizhak 2018, 2020). Моя работа призвана посмотреть на природу переводов с японского на русский через автоматическую разметку и подсчет квантитативными методами.

В результате работы:

- обучен machine learning-классификатор, является ли текст переводом с японского или нет,
- получен список черт диалекта переводов художественной литературы с японского на русский
- добавлено их описание с точки зрения лингвистики.

2. Обзор литературы

Язык, с которого осуществляется перевод (SL), оказывает систематическое влияние на язык перевода (TL). Это проявляется в статистической разнице в частоте грамматических конструкций, частей речи, функциональных слов (дискурсивных маркеров, слов-связок). Это различие указывает на разницу в структуре двух языков, а не на низкое качество выполнения перевода (Gellerstam, 1986). Такое систематическое влияние может даже привести к тому, что изменится целевой язык — его грамматика, контексты употребления лексем сблизятся с языком-источником. Геллерстам описывает случай влияния английского на шведский: в обоих языках были одинаковые латинские заимствования (local,

massive, dramatic и т.д.), но контекст их употребления различался. В переводах часто оставляют одни и те же слова, даже если сфера их употрбления различается (local в шведском со временем начало употребляться в более широких контекстах под воздействием английского субстрата). Таким образом, со временем слова целевого языка могут изменить значение под воздействием переводов (Gellerstam 1986).

Следующая закономерность в том, что переводные тексты тяготеют к “словарному” переводу слов, даже если в целевом языке они не употребляется в таком контексте. *Anlända* — словарный перевод лексемы *arrive* на шведский. В шведском это слово официального стиля, которое используется в ограниченном контексте, тогда как английское *arrive* употребляется в намного более широких контекстах, поэтому частота встречаемости различается в переводных и оригинальных текстах (Gellerstam 1986).

В то же время, распределение лексики, обоснованное различиями в культуре, не является особенностью языка переводов (если в переводах с английского чаще öl `*пиво*` встречается чаще, а в текстах, изначально написанных на шведском, чаще встречается kaffe *`кофе`* из этого ничего не следует (Gellerstam 1986). Translationese — структурное, грамматическое различии между языками.

В 1980-1990е разрабатывалось описание универсальных характеристик переводов, независимых от языковых пар. Среди переводческих универсалий выделяют:

Эксплиситация: переводчик стремится выразить скрытые смыслы более явно (“...regardless of the languages concerned, the interpreter tends to render implicit forms more explicitly” (Shlesinger, 1995))

Симплификация: сложный синтаксис передается более простыми способами (“ …difficult syntax is made easier” (Baker, 1993))

Нормализация/консерватизм: тенденция использовать более конвенциональные грамматические структуры (a strong preference for conventional grammatical structures (Baker, 1993))

Levelling out (“сглаживание”): если в оригинале использованы контрастные, полярные формы, в переводе они скорее всего будут “сглажены” и более

нейтральны (“...steering a middle course between any two extremes, converging towards the centre” (Baker, 1996))

В то же время, переводческие универсалии могут нарушаться. Пууртинен (2003) описала нефинитные конструкции в финской детской переводной литературе: их количество было значимо выше, чем в текстах, изначально написанных на финском. Отличалась и их семантика, силистика и прагматика: они употреблялись в контекстах, не характерных для финского. Смысл этих нефинитных конструкций (стяженные формы глаголов, причастия, предшествующие определяемому слову, второй инфинитив в инструктиве, -minen- номилизации перед определяемым словом) можно выразить также через финитные, для этого нужно сделать предлолежние сложным, но в переводе выбираются именно более тяжелые для парсинга нефинитные конструкции. Они передают большое количество информации компактно, что увеличивает лексическую плотность (противоречит универсалии сиплификации), при этом не всегда явно устанавливаются актанты (противоречие универсалии эксплицитности). В целом такие конструкции редкие для финского, что противоречит универсалии нормализации. Тем интереснее, что они частотны в переводах именно детских книг, в которых преподчитают писать более явно и конкретно, а не метафорично, а эти конструкции являются довольно сложными поэтому кажутся избыточными, когда целевой аудиторией текстов являются дети. Пууртинен предполгает, что нефинитные конструкции могут использоваться благодаря их особому ритму, более мелодичному при чтении вслух, чем в подчиненных клаузах. Также сложность нефинитных конструкций может нивелироваться тем, что они передают смыслы более сжато, компенсируя тенденцию переводов быть длинее, чем оригинальный текст.

В этих исследованиях язык переводов описывался корпусными методами, черты подсчитывались и описывались вручную. Бернардини и Барони одни из первых использовали машинное обучение (метод опорных векторов) для выделения черт языка переводов на материале итальянских переводов геополитических статей (Baroni & Bernardini 2006). Они составили список черт (‘feature engineering’) и обучили на нем классификатор, после чего выделили список самых значимых черт.

Наиболее важными в классификации переводов и оригинальных текстов оказалось распространение н-грам функицональных слов и морфосинтаксические категории. В 86.7% случаев классификатор определял текст верно, что выше, чем качество, с которым отличали переводы от оригинальных текстов специалисты.

В связи с развитием нейросетевых подходов появились новые способы для детекции translationese. (Pylypenko и др., 2021) показали, что:

1) подходы с выучиванием репрезентаций эмбеддингов (representation-learning-based) работают лучше, чем выделение лингвистических черт вручную (hand-crafted linguistically inspired feature-selection)
2) черты, выученные при помощи репрезентаций эмбеддингов, лучше обобщают на мультиязычном материале
3) мультиязычные эксперименты свидететсльвтуют о существовании независимых от конкретных языков черт переводческого языка

Тем не менее, пускай нейросетевые методы добиваются высокой степени точности в выделении переводческого диалекта (90% и более), они не помогают в объяснении этих различий и могут быть использованы только для того, чтобы определить, что текст является переводом с другого языка.

Считается, что использование неодушевленных существительных в качестве субъектов переходных глаголов является неестественным для японского и поэтому является характеристикой переводческого диалекта (Morioka, 1988). Yoshioka (1973) указывал, что в японском такое использование уже стало общепринятым из-за влияния европейских языков, но работа 1995 года утверждает, что оно все еще достаточно маргинально (Suzuki, 1995).

Икэгами указал, что японский является скорее языком-“становящимся” (naru-no gengo, BECOME-language), чем языком-“делающим” (suru-no gengo, DO-language) в отличие от английского или русского (Ikegami, 1991). Языки-“становящиеся” скорее описывают ситуацию в целом и ее изменения из одного состояния в другое, тогда как в фокусе языков-“делающих” субъект, который совершает действия. По этой причине японский язык скорее всего будет соблюдать единство субъекта в предложении и описывать происходящее с ним через пассивный залог там, где язык-“делающий” скорее сменил бы субъекты. Стрижак (2017) продемонстрировала, что переводы русских предложений с переходными

глаголами на английский сохраняют структуру, в которой субъектом является неодушевленное существительное (*жизнь, запах, письмо, семейная жизнь*), тогда как переводы на японский переводят предложение в пассив и субъектом становится агенс-человек; или же японский выбирает менее переходный глагол:

(1a) Жизнь презренная, жизнь обывательская затянула нас. (Rus)

жизнь-NOM затянула-TR нас-ACC

(1b) Geretsu-na seikatsu-no o-kage-de, wareware mo kono naka-he hikizurikomarete ni shimatta ja nai ka.

из-за жизни-GEN мы-NOM затянуты-PASS

(1c) This miserable life has sucked us under.

жизнь-NOM затунять-TR нас-ACC

[`Дядя Ваня` by А. П. Чехов]

(2a) Запах этот начал преследовать прокуратора с рассвета.

запах-NOM преследовать-TR прокуратор-ACC

(2b) Sono nioi-ga yoake-kara hana-ni tsuite hanarenakatta.

запах-NOM пристать к-INTR

(2c) This smell had been pursuing the procurator since down.

запах-NOM преследовать-TR прокуратор-ACC

[`Мастер и Маргарита` М. А. Булгаков]

Исследование Мельдрум (2009) демонстрирует, что в современной японской художественной литературе (1980-2006) неодушевленные существительные в качестве аргументов переходных глаголов встречаются чаще, чем в переводах за этот же временной период. Она связывает это с тем, что такая синтаксическая структура считается признаком плохого перевода, поэтому переводчики осознанно избегают ее, тогда как японские авторы таких ограничений не ощущают и поэтому пишут свободнее.

Янасэ (2000) и другие исследователи указывали, что в японском личные местоимения преимущественно опускаются, и поэтому их выражение — признак переводческого диалекта. Мельдрум подсчитала (2009), что в ее корпусе местоимения третьего лица действительно встречаются чаще в переводах, чем в оригинальных текстах.

Ихара (2008) указал, что в японских художественных текстах прямая передача речи встречается чаще, чем непрямая, и поэтому такое распределение также может считать чертой переводческого диалекта в японском.

3. Данные

Я провожу исследование на основе корпуса русскоязычных художественных текстов. В корпусе два подкорпуса: половина текстов изначально написана на русском языке в период с 1956 по 2014 год, половина текстов является переводами с японского на русский, осуществленными с 1955 по 2014 год профессиональными переводчиками. Объем корпуса составляет 1 миллион словоупотреблений. Файлы книг скачаны из публичных интернет-библиотек, в русскоязычном подкорпусе использованы ознакомительные версии магазина электронных книг Литрес. В японской части подкорпуса в исследовании использована первая четверть полных текстов книг. Всего 49 книг.

Публикация перевода	Публикация оригинала	Название книги	Автор	Переводчик
1955	1905	Нарушенный завет	Тосон Симадзаки	Н. И. Фельдман-Конрад
1957	1902	Куросиво	Токутоми Рока	И. Львова
1962	1919	Женщина	Такэо Арисима	А. Г. Рябкин
1969	1964	Чужое лицо	Абэ Кобо	В. С. Гривнин
1977	1968	Сны о России	Иноуэ Ясуси	Б. В. Раскин
1978	1973	Объяли меня воды до души моей	Оэ Кэндзабуро	В. С. Гривнин
1982	1970	Свет без тени	Ватанабэ Дзюнъити	Г. Б. Дуткина, В. Вениаминов
1986	1924	Любовь глупца	Танидзаки Дзюнъитиро	Г. Г. Иммерман
1989	1966	Молчание	Эндо Сюсаку	И. Львова, Г. Б. Дуткина

1989	1925	Человек-кресло	Рампо Эдогава	Г. Б. Дуткина
2001	1937	Снежная страна	Ясунари Кавабата	З. Рахим
2004	1988	Кухня	Ёсимото Банана	А. М. Кабанов
2005	1990	Соль жизни	Исихара Синтаро	А. Н. Мещеряков
2005	1990	Синева небес	Соно Аяко	Т. И. Бреславец
2005	1982	Узорчатая парча	Миямото Тэру	Г. Б. Дуткина
2005	1978	До заката	Ёсиюки Дзюнноскэ	Ю. Окамото
2006	1917	Соперницы	Кафу Нагаи	И. В. Мельникова
2006	1996	Остров мертвых	Бандо Масако	Ю. Чинарева
2008	1965	Цитадель	Энти Фумико	Г. Б. Дуткина
2008	1993	Мелодия неба	Сэридзава Кодзиро	И. Мотобрывцева
2011	1981	Супермаркет	Адзути Сатоси	А. А. Долин
2013	1999	Божественная лодка	Экуни Каори	И. Пурик
2013	1999	Мой любимый sputnik	Мураками Харуки	Н. Куникова
2014	1985	Остров мечты	Хино Кэйдзо	Т. И. Редько-Добровольская
2014	1979	Приговор	Отохико Кага	Т. Л. Соколова-Делюсина

Таблица 1. Состав корпуса переводных текстов

Год публикации	Название	Автор
1956	Алешкино сердце	Михаил Шолохов
1958	Незнайка в солнечном городе	Николай Носов
1959	Динка	Валентина Осеева
1965	Понедельник начинается в субботу	Аркадий Стругацкий,

		Борис Стругацкий
1967	Крутой маршрут	Евгения Гинзбург
1972	Летчик для особых поручений	Владислав Крапивин
1973	Школа для дураков	Саша Соколов
1974	Сердце хирурга	Федор Углов
1977	Сто лет тому вперед	Кир Булычев
1981	Чучело	Владимир Железников
1983	У войны не женское лицо	Светлана Алексиевич
1986	Завтра была война	Борис Васильев
1990	Гигиена	Людмила Петрушевская
1995	Смерть и немного любви	Александра Маринина
1998	Ночной дозор	Сергей Лукьяненко
1999	Generation P	Виктор Пелевин
2000	Казус Кукоцкого	Людмила Улицкая
2000	Кысь	Татьяна Толстая
2000	Настя	Владимир Сорокин
2004	Темная сторона	Макс Фрай
2010	Страж	Алексей Пехов, Елена Бычкова
2012	Лавр	Евгений Водолазкин
2012	Часовая башня	Наталья Щерба
2014	Обитель	Захар Прилепин

Таблица 2. Состав корпуса художественных текстов, написанных по-русски.

Корпус состоит из книг разных жанров: от классической литературы до массовой. Бо́льшая часть корпуса — романы, но небольшое количество повестей и рассказов также включены в него.

Вторым корпусом в этом исследовании являются тексты Григория Чхартишвили, более известного в русскоязычной литературной среде под псевдонимом Борис Акунин. Чхартишвили — японист, который перевел многие книги на русский (в частности, многие работы Мисима Юкио). Помимо этого он написал серию романов, и события некоторых из них происходят в Японии. Этот корпус по объему составляетт 500 тысяч словоупотреблений и является основой для валидационной выборки для классификатора.

4. Методы

Эта работа выполнена в русле корпусной лингвистики. Тексты лемматизуются и разбираются морфологически и синтаксически во фреймворке Universal Dependencies (Nivre et al., 2016). После этого подсчитываются значения лингвистических фичей, выбранных релевантными для выделения переводческого языка на основе теоретических наблюдений. Список фичей наследует списку (Kunilovskaya & Lapshinova-Koltunski, 2020)[1] из 45 фичей, к которому добавлены фичи, основанные на теоретических положениях об иерархии агенса в японском и большей представленности прямой речи как черты японских текстов.

4. 1. Описание добавленных черт

4. 1. 1. Доля прямой речи в тексте

Чтобы выделить из всего массива текста прямую речь, я использовала правила, написанные для приложения hseling-api-direct-speech[2]. После этого предложения, обнаруженные правилами, подсчитывались, и число предложений с прямой речью делилось на общее число предложений — переменная *directs*.

— Так, значит, ты хотела узнать разницу между “знаком” и “символом”, поэтому позвонила мне. В воскресенье утром, до рассвета. Хм-м… [Мураками Харуки. Мой любимый sputnik (1999)]

Выделенная цитата содержит три предложения, и в общий счетчик добавляется цифра 3.

[1]Kunilovskaya, Maria & Ekaterina Lapshinova-Koltunski. translationese45: code to extract 45 translationese indicators for English, German and Russian, most of which were used in the research presented at LREC 2020 url (https://github.com/kunilovskaya/translationese45) Дата обращения: 11.05.2022.

[2] HSE School of Linguistics. hseling/hseling-api-direct-speech: HSE Linguistics API: Direct speech url: https://github.com/hseling/hseling-api-direct-speech) Дата обращения: 11.05.2022.

4.1.2. Доли одушевленных/неодушевленных субъектов (и местоимений)

Так как теория показывает, что неодушевленные существительные реже становятся субъектами транзитивных глаголов, логично протестировать это положение и подсчитать доли 1) одушевленных субъектов (Anim в Universal Dependencies), 2) неодушевленных субъектов (Inan), 3) местоимений (PRON).

В большинстве своем художественные тексты пишут о людях, и поэтому самыми частотными субъектами являются местоимения вне зависимости от того, является ли текст переводом. Из разметки UD я извлекла субъекты переходных и непереходных глаголов в активном залоге, после чего вручную вычитала разметку для всех субъектов, так как автоматическая ошибается, и подсчитала две метрики:

- inanpronanim: inanimate / (pronouns + animate)
- inananim: inanimate / animate

Все же, тексты действительно в первую очередь пишутся о людях, часто от первого лица, поэтому доля местоимений может быть менее показательной, чем соотношение неодушевленных и одушевленных субъектов.

4.1.3. Доля причастий

Японский язык предпочитает сохранять единый субъект, кроме того, это язык с левым ветвлением, определения в нем предшествуют определяемому. По этой причине кажется логичным проверить долю причастий в переводных текстах, так как при сохранении единого субъекта могут использоваться как причастия, так и определения с вложенной клаузой тоже могут переводиться через причастия, хотя их скорее переведут через относительное предложение, если составляющие длинные (сделавшая снимок девушка купила этот цветок vs девушка, которая сделала снимок, купила этот цветок - shashin-wo to-tta shoujo-ga ano hana-wo ka-tta - photo-ACC make-PST girl-NOM that flower-ACC buy-PST).

Исходя из этих посылок я добавила в датасет две фичи:

- *partsent* — количество предложений с причастиями / все предложения
- *totalpart* — количество причастий / все предложения

Я также поставила эксперименты с извлечением субъектов пассивов и посчитала конструкции с pro-drop — опущением субъекта, но они не показались репрезентативными в рамках работы.

5. Описание самых значимых фичей, качество модели

Я обучила несколько классификаторов на предобработанных данных, лучшим себя показал многослойный перцптрон. Но, так как выделить самые значимые фичи в решениях перцептрона затруднительно, я использовала значения magnitude coefficent логистической регрессии.

	classifier	f1	accuracy
0	LogisticRegression	0.705882	0.705882
1	MLPClassifier	0.763072	0.764706
2	GaussianNB	0.705882	0.705882
3	KNeighborsClassifier	0.526144	0.529412
4	SVC	0.301176	0.470588
5	DecisionTreeClassifier	0.647059	0.647059

Таблица 3. Качество классификаторов в определении, является ли текст переводом с японского или нет.

cconj	-1.334638e-01
simple	-2.500981e-02
finites	-1.314326e-02
…	
pasttense	1.463939e-02
mdd	2.013139e-02
mhd	2.242815e-02
inanpronanim	2.639556e-02
sconj	2.888708e-02
attrib	8.484602e-02
sentlength	1.022403e-01
directs	1.111841e-01
numcls	1.314940e-01

inananim	5.821506e-01

Таблица 4. Первые 13 самых значимых фичей: негативные (чем больше значение в негативную сторону, тем скорее текст не является переводом с японского) и положительные (чем больше значение в положительную сторону, тем скорее текст является переводом с японского).[3]

Наиболее значимыми фичами с определении, является ли текст переводом или нет, себя показали:

- доля неодушевленных субъектов от одушевленных (inananim),
- число клауз (numcls),
- количество прямой речи (directs),
- длина предложения (sentlength) — во многих работах по языку переводов, применяющих машинное обучение, пишут, что именно длина предложений хорошо разделяет оригиналы от переводов,
- количество прилагательных и причастий в функции amod (предшествующие определяемому) (attrib)
- количество подчинительных союзов (sconj)
- доля неодушевленных субъектов от местоимений и одушевленных (inanpronanim)
- mean hierarchical distance — средняя длина всех путей от корня до узлов (Jing, Liu, 2015) (mhd)
- mean dependecy distance — расстояние между словами и словами, с которыми они связаны в количестве слов в самом предложении (не в дереве) (Hudson, 1995) (mdd)
- количество глаголов в прошедшем времени (pasttense)

После этого я проверила фичи с расхождением в значении медиан:

feature	переводы	оригиналы
inanimate/animate	0.9206731	0.6711449
number of clauses	1.0669208	0.8719073
attributives	0.7910314	0.5641281
inanimate/(animate+pronoun)	0.3789869	0.2927617

[3] Описание всех фичей смотрите в репозитории (Kunilovskaya, Lapshinova-Koltunski 2020): github.com/kunilovskaya/translationese45/blob/master/lrec20_45featureset_description.pdf

mean hierarchical distance	3.0764281	2.9078395
mean dependency distance	1.055062	0.9698634
past tense	1.4230769	1.2452122
sentence length	15.9426087	13.045853

Таблица 5. Медианы значений фичей для книг из корпусов

По-видимому, б*о*льшее количество неодушевленных субъектов действительно играет роль в классификации, по крайней мере, на материале, на котором обучалась моя модель. Интуиция, что смыслы переносятся через более длинные адъективные группы (причастия, клаузы) тоже может быть верной, но возможно еще одно объяснение. Тексты переводов имеют тенденцию быть более длинными, чем тексты оригиналов, что обусловлено универсалией эксплиситации. Переводчик стремится более полно передать грани емко упакованного смысла, который в языке перевода часто нет языковых средств передать столь же емко. Отсюда, возможно, и увеличение средних расстояний в переводных текстах.

Неожиданным было б*о*льшее медианное количество прошедшего времени в переводах. Возможно, это связано с тем, что язык переводов более стандартизован.

5.1. Ошибки классификаторов

В тестовой выборке было 17 книг. Многослойный перцептрон не справился с классификацией 4 книг, логистическая регрессия и гауссовский наивный байес неправильно классифицировали по 5 книг. Тем интереснее, что ошибки были сделаны примерно в одних и тех же книгах, но распределены они были неоднородно между классификаторами.

Книга	Классификаторы
1958. Носов. Незнайка в Солнечном город	Логистическая регрессия, Многослойный перцептрон, Гауссовский наивный байес
1973. Соколов. Школа для дураков	Логистическая регрессия
1982. Ватанабэ. Свет без тени	Логистическая регрессия, Гауссовский наивный байес

1999. Пелевин. Generation P	Логистическая регрессия, Гауссовский наивный байес
2004. Фрай. Темная сторона	Многослойный перцептрон
2005. Ёсиюки. До заката	Логистическая регрессия, Многослойный перцептрон, Гауссовский наивный байес
2010. Пехов, Бычкова. Страж	Многослойный перцептрон, Гауссовский наивный байес

Таблица 6. Какие книги были неправильно классифицированы классфикаторами.

Все классификаторы сделали ошибку на отрывке из книги о Незнайке. Я могу попробовать объяснить это тем, что в начальном отрывке многие субъекты предложений — неодушевленные (машины, комбайны). Такое же объяснение может подойти к «Стражу» Пехова и Бычковой: в начале книги речь идет об ожившем пугале, которое, конечно, размечается как неодушевленное. «Школа для дураков» Соколова — постмодернистский текст со сложным языком, и поэтому, как кажется, сложность его устройства (mhd, mdd, attributives, numclauses) на поверхностном уровне сближает его с языком переводов, сложностье количеста зависимостей которого обусловлена универсалией эксплиситации.

5.2. Валидационный датасет

Для того, чтобы протестировать работу классификатора, я использовала датасет, собранный из текстов Григория Чхартишвили (Борис Акунина), всего 14 книг. Указана дата публикации книги или перевода.

Книга	Автор
1998. Азазель	Борис Акунин
1998. Турецкий гамбит	Борис Акунин
1998. Смерть Ахиллеса	Борис Акунин
1999. Пиковый валет	Борис Акунин
1999. Статский советник	Борис Акунин
2002. Алмазная колесница	Борис Акунин

2006. Любовник Смерти	Борис Акунин
1989. Испытание зверя	Моримура Сэйити
1991. Сердцебиение	Маруяма Кэндзи
1993. Золотой храм	Мисима Юкио
1993. Патриотизм	Мисима Юкио
1993. Исповедь маски	Мисима Юкио
1996. Собачья невеста	Тавада Ёко
2003. Царь Армадилл	Симада Масахико

Таблица 7. Валидационный датасет

Все три классификатора неправильно классифицируют две книги: «Любовницу смерти» Акунина и «Собачью невесту» Тавада. Наивный байес делает ошибку в «Сердцебиении» Маруяма, но в остальном классификаторы справляются с задачей и, как кажется, можно говорить о том, что выделенные фичи отделяют переводные тексты.

6. Особенности лексики

Еще одна важная черта диалекта переводов в том, что лексика может использоваться в иных контекстах, чем принято в естестественном языке. Я, следуя методологии (Gellerstam 1986), выделила:

1) Слова, которые встречаются как минимум в половине текстов;
2) При этом, 70% и больше употреблений — в переводах.

В результате было отсеяно 64 слова: *буквально, виднеться, возможно, возникать, волна, вот-вот, выражение, глубина, гора, дерево, душевный, единый, желание, житель, здание, испытать, испытывать, исходить, компания, летний, луч, магазин, малейший, маска, местный, метр, младший, море, наверное, название, напоминать, невольно, облик, одиночество, окружать, охватить, ощутить, ощущать, ощущение, парень, поспешно, постепенно, поток, прекрасно, прекрасный, приготовить, придавать, примерно, приобрести, природа, пристально, приятель, район, слой, супруга, существование, течение, тоска, тревога, уверенность, университет, учитель, фотография, школа*.

Из этих слов можно выделить несколько групп. Первая из них — наречия, вторая — глаголы.

Наречия/частицы: *буквально, возможно, вот-вот, наверное, невольно, поспешно, постепенно, прекрасно, примерно, пристально*.

Глаголы: *виднеться, возникать, испытать, испытывать, исходить, напоминать, охватить, ощутить, ощущать, приготовить, придавать, приобрести*.

Можно выделить и некоторые другие семантические группы:

указания на людей: житель, компания, местный, младший, парень, приятель, супруга;

чувства: душевный, желание, одиночество, тоска, тревога.

Но если в первой группе можно попробовать найти объяснения для некоторых из слов: ассиметрия в употреблении лексики/выбор конкретных переводчиков:

Не знаю, зачем им этот *парень*, но он достаточно важен, коль уж они так суетятся. [Пехов, Бычкова. Страж (2010)]

Какой-то *парень*, пьяный и кудлатый, узнавший стороной об измене подруги, в отчаянии: Валю любили трое. [Соколов. Школа для дураков (1973)]

Когда дело доходит до управления амобилером, этот фантастически шустрый *парень* ничем не отличается от прочих обитателей Соединенного Королевства. [Фрай. Темная сторона (2004)]

— в русском *парень* — это что-то, чему нужен ограниченный контекст: так говорят либо о ком-то незнакомом/малознакомом, либо о чьем-то партнере, либо в переносных контекстах: машина — шустрый парень.

В переводах же возможно называть на протяжении всего рассказа главного героя *парнем*[4]. В повести Маруяма Кэндзи «Сердцебиение» нарратор пытается понять, как ему называть главного героя:

Мужчина пересел так, что теперь был хорошо виден в зеркальце. Обменявшись с ним парой слов, я чувствовал себя немного свободнее и впервые внимательно взглянул на него. Совсем молодой парень, намного моложе меня. [Маруяма Кэндзи. Сердцебиение (1982)]

[4] В данный момент моего знания японского недостаточно, чтобы описать, какие лексические единицы в японском соответствуют их русскоязычным эквивалентам в переводах, поэтому в тексте диплома я только высказываю предположения о причинах различия в семантике (хотя я и могу уловить разницу в значении). Вероятно, семантическое поле устроено иначе, некоторым словам не хватает эквивалентов, и они вносят японский контекст в русский.

И из-за возраста рассказчик делает выбор в пользу референции через *парень*, а не *мужчина*, называя героя на всем протяжении истории так, что в русском тексте оставляет ощущение отчуждения[5], и, если бы текст был изначально на русском, автор скорее бы подобрал другой способ указывать на героя. Очевидно, это происходит из ассиметрии полей и того, что в русском нет другого более нейтрального средства указать, что кто-то старше, чем *мальчик*, но младше, чем *мужчина*[6].

Такие случаи ассиметрии кажутся вероятными. Что касается превалирования в переводном корпусе чувств или же таких слов, как *фотография, школа, университет*, это скорее связано с темами переводных текстов, чем с признаками диалекта.

6.1. Наречия/частицы

Из списка наречий можно выделить несколько типов:

1) модальные: возможно, наверное
2) по характеру действия: буквально, вот-вот, невольно, поспешно, постепенно, пристально

Вероятно, первая группа «словарным» образом соответствует некоторым японским наречиям, которые так и переводятся, а в тексте, изначально писавшемся на русском, было бы больше вариантов для того, чтобы показать степень уверенности, поэтому эти наречия «высвечиваются».

Но самое интересное, как передается грамматическая семантика японских глаголов в русском в совокупности с частотноми японскими наречиями. Это становится видным из примеров второй группы.

6.1.1. Буквально

И в русском, и в японском корпусах *буквально* в 50% случаев модифицирует глагол. Но употребления в переводном корпусе (78% всех употреблений) интересны тем, что *буквально* модифицирует существительные не так, как принято в русском.

---

[5] Из концовки: И тут я увидел, что в машине кто-то есть. Через залитое дождем стекло было не видно лица, но на переднем сиденье явственно вырисовывался силуэт. Я огляделся по сторонам — не следят ли за мной — и приблизился. Глубоко вздохнув, рывком открыл дверцу. В машине, откинувшись назад, сидел парень. [Маруяма Кэндзи. Сердцебиение (1982)]
В конце было бы более естественно указать «тот парень» или обратиться иначе, так как читатель уже давно познакомился с героем.

[6] У *юноши* тоже не совсем нейтральные коннотации.

Буквально кожей в Национальном корпусе русского языка встречается всего 9 раз. В небольшом корпусе переводов оно встретилось трижды в разных текстах разных переводчиц.

Я буквально кожей почувствовала, что жизнь изрядно поломала Вас, и Ваше существование отнюдь нельзя назвать безмятежным. [Миямото Тэру. Узорчатая парча (1982)]

Фудзио *буквально* кожей ощущал напряженность ситуации. [Соно Аяко. Синева небес (1990)]

Что-то едва уловимо шевельнулось на дне запавших глазниц, и Тикаки *буквально* кожей ощутил исходящую от корейца волну враждебности. [Отохико Кага. Приговор (1979)]

В непереводных примерах *буквально* — это интенсификатор, который одновременно несет оттенок эвиденциальности: сам автор высказывания пережил это либо видел это. В переводах с японского области интенсификации становятся шире.

Я познакомилась с ним в 1929 году, и он овельможивался *буквально* на моих глазах. [Евгения Гинзбург. Крутой маршрут (1967)]

Меня часто не только что восхищало – *буквально* поражало сознательное отношение детей к своей болезни, понимание того, что все медицинские манипуляции, нередко болезненные, которые мучительно переносили даже взрослые, – необходимы, без них не обойдешься. [Федор Углов. Сердце хирурга (1974)]

6.1.2. Вот-вот

Вот-вот эвиденциально окрашено, автор высказывания видит (чувствует), что на его глазах что-то случится, этого ждут, и оно действительно скорее всего произойдет (71% всех употреблений в корпусе переводов).

Это еще что такое? — воскликнул он поспешно, точно боясь, что Бой *вот-вот* выстрелит. [Оэ Кэндзабуро. Объяли меня воды до души моей (1978)]

В примере выше нет семантики зафиксированности события: она снимается глаголом *бояться*.

Я радовался, что у меня *вот-вот* будет ребенок. [Соно Аяко. Синева небес (1990)]

В примере выше акт говорения и акт рождения разнесены по времени, хотя и ощущаются говорящим как близкое.

6.1.3. Поспешно, постепенно, пристально, невольно

Эти наречия указывают на то, какие модификаторы, способы действия частотны в японском. И поспешно (70%), постепенно (73%) выражают аспектуальную характеристику глаголов. Невольно (77%) указывает на ненамеренность действия. Эта семантика может добавляться формой глагола или сопутствующими наречиями[7]. Пристально (71%) — интенсификатор.

В оригинальных текстах *невольно* говорят про однократное неконтролируемое короткое действие, часто — неконтролируемую эмоциональную реакцию (*невольно улыбнулся*). В переводных текстах *невольность* действия маркируется чаще, в менее контролируемых контекстах, чем в русском.

Я посмотрела на её спящее лицо и *невольно* прилегла рядом — так, чтобы чувствовать её дыхание. [Экуни Каори. Божественная лодка (1999)]

6.2. Глаголы

Из глаголов также можно выделить несколько групп:

1) *испытать* (70%)*, испытывать* (80%)*, напоминать* (73%)*, ощутить* (72%)*, ощущать* (73%) — чувства
2) *исходить* (75%)*, охватить* (83%) — распространяться (тоже: часто о чувствах)
3) *виднеться* (76%)

Частотность первых двух групп может быть обусловлена выборкой текстов. Выборка текстов составлялась на основе списка издательства Эксмо[8], поэтому, либо же на русский язык переводят тексты про ощущения, чувства, переживания,

[7] У многих форм глагола есть имлицитная семантика, например, у японского пассива есть имплицитная семантика негативногой вовлеченности субъекта. “It follows that the negative interpretation of the examples (1) through (4) must be explained from the fact that John’s desire is of no consequence to an event which involves someone or something connected with him. As soon as it is clear why the affected person’s desire is irrelevant to the situation, the negative implication disappears.” (Kortlandt, 1992)

(1) John wa Mary ni kare no piano o hikareta.
John TOP Mary DAT he GEN piano ACC play-PSV-PAST
‘Mary played John’s piano; John was negatively affected by it.’

[8] 50 лучших японских романов XX века [Электронный ресурс]. URL: https://eksmo.ru/selections/50-luchshikh-yaponskikh-romanov-xx-veka-ID3258958/ (дата обращения: 27.05.2022).

либо же японской литературе свойственно рассказывать нарратив через такие средства.

6.2.1. Виднеться

В японском что-то *виднеется* — намного чаще, чем в русском. Так передается непереходный глагол 見える(mieru `виднеться`). Частоту его использование фиксирует и диалект переводов с японского на русский.

7. Заключение.

Распределение неодушевленных и одушевленных существительных в качестве субъектов глаголов, длина определений, глубина зависимостей в дереве зависимостей, количество прямой речи помогают определить автоматически, является ли текст переводом с японского или нет. Чем выше значение, тем с бо́льшей вероятностью текст является переводом.

Для диалекта переводов с японского характерна особая лексика: наречия *буквально, вот-вот, поспешно, постепенно, пристально, невольно*, глаголы *испытать, испытывать, напоминать, ощутить, ощущать, исходить, охватить, виднеться*, через которые структура и грамматика японского «просвечивается» в русском.

.

Приложение

Репозиторий с кодом: github.com/mjolnika/japanese-russian-translationese

Примеры из корпуса с лексикой: github.com/mjolnika/japanese-russian-translationese/blob/main/EXAMPLES.docx

Источники

Стрижак, У. (2018). *Одушевленность как грамматическая и понятийная категория языка: Когнитивный подход к описанию*. Opuscula Iaponica et Slavica. Т. 5. https://publications.hse.ru/chapters/223997743

Стрижак, У. (2020). *Культурнонагруженная грамматика: иерархия агенса в японском*. 210–223. https://elibrary.ru/item.asp?id=42371236&

Baker, M. (1996). Corpus-Based Translation Studies: The challenges that lie ahead. In *Researching Translation in the Age of Technology and Global Conflict* (pp. 44–54). Routledge.

Baroni, M., & Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, *21*(3), 259–274.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. *Translation Studies in Scandinavia*, *1*, 88–95.

Ihara, N. (2008). Nichi-Ei Shōsetsu no katari ni arawareru “koe”–Jiyū kansetsu wahō to sono hon‟ yaku [Narrative Voices in English and Japanese novels: Free Indirect Discourses and their translation]. *The Japanese Journal of Language in Society*, *11*(1), 151–163.

Ikegami, Y. (1987). Source’vs.‘goal’: A case of linguistic dissymmetry. *Concepts of Case*, *122*, 146.

Ikegami, Y. (1991). *"Do-language and Become-language": Two Contrasting Types of Linguistic Representation*. https://doi.org/10.1075/fos.8.14ike

Jing, Y. and Liu, H. (2015). Mean Hierarchical Distance Augmenting Mean Dependency Distance. In Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pages 161–170.

Kortlandt, F. (1992). On the meaning of the Japanese passive. *Acta Linguistica Hafniensia*, *24*(1), 97–108. https://doi.org/10.1080/03740463.1992.10412271

Kunilovskaya, M., & Kutuzov, A. (2017). Universal Dependencies-based syntactic features in detecting human translation varieties. *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, 27–36. https://aclanthology.org/W17-7606

Kunilovskaya, M., & Lapshinova-Koltunski, E. (2020). Lexicogrammatic translationese across two targets and competence levels. *Proceedings of the 12th Language Resources and Evaluation Conference*, 4102–4112. https://aclanthology.org/2020.lrec-1.505

Kunilovskaya, M., Lapshinova-Koltunski, E., & Mitkov, R. (2021). Translationese in Russian Literary Texts. *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 101–112. https://doi.org/10.18653/v1/2021.latechclfl-1.12

Meldrum, Y. F. (2009a). *Contemporary Translationese in Japanese Popular Literature*. 255.

Meldrum, Y. F. (2009b). *Translationese-Specific Linguistic Characteristics:* 28.

Morioka, K. (1988). *Gendaigo kenkyū shirīzu: Buntai no hyōgen*. Meiji shoin.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D.,

McDonald, R., Petrov, S., Pyysalo, S., & Silveira, N. (2016). Universal dependencies v1: A multilingual treebank collection. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1659–1666.

Puurtinen, T. (2003). Nonfinite constructions in Finnish children's literature: Features of translationese contradicting translation universals? In *Corpus-based approaches to contrastive linguistics and translation studies* (pp. 141–154). Brill.

Pylypenko, D., Amponsah-Kaakyire, K., Chowdhury, K. D., van Genabith, J., & España-Bonet, C. (2021). Comparing Feature-Engineering and Feature-Learning Approaches for Multilingual Translationese Classification. *ArXiv:2109.07604 [Cs]*. http://arxiv.org/abs/2109.07604

Shlesinger, M. (1995). Shifts in cohesion in simultaneous interpreting. *The Translator*, *1*(2), 193–214.

Strizhak, U. (2017). Cognitive linguistics in education: The new paradigm of Japanese language teaching. *Mir Nauki, Kul'tury, Obrazovanija [The World of Science, Culture, Education], 6 (67)*, 581–583.

Suzuki, C. (1995). *Watashi no Hon'yaku Dangi [My Translation Monologue]*. Tokyo: Kawade Shobō Shinsha.

Svanberg, J., Inose, H., & Jonsson, H. (n.d.). *A corpus-based translation study on Japanese translationese by Swedish to Japanese translation*. *2017*, 50.

Toury, G. (1985). *Aspects of translating into minority languages from the point of view of translation studies*.

Yanase, N. (2000). *Honyaku wa ikani subekika*. Iwanami Shoten.

Yoshioka, M. (1973). Gendai Nihongo ni okeru ōbunmyaku no eikyō: Hon‟ yakutai no nihongoka [Influences of European style Japanese in modern Japanese: Japanization of translation style]. *Gengo Seikatsu, 259*, 62–69.

Kunilovskaya, Maria & Ekaterina Lapshinova-Koltunski. translationese45: code to extract 45 translationese indicators for English, German and Russian, most of which were used in the research presented at LREC 2020 url (https://github.com/kunilovskaya/translationese45) Дата обращения: 11.05.2022.

HSE School of Linguistics. hseling/hseling-api-direct-speech: HSE Linguistics API: Direct speech url: https://github.com/hseling/hseling-api-direct-speech) Дата обращения: 11.05.2022.