Политическая лингвистика. 2024. № 5 (107).
Political Linguistics. 2024. No 5 (107).

УДК 811.161.1'42+811.581'42+81'27+81'25
ББК Ш141.12-51+Ш171.1-51+Ш118+Ш100.621 *ГРНТИ 16.31.21; 16.31.41* *Код ВАК 5.9.8*

Чжу Хуэй[1]✉, Ольга Александровна Митрофанова[2]✉

[1] Даляньский университет иностранных языков, Далянь, Китай, zhuhui1230@qq.com✉, https://orcid.org/0009-0003-2922-8156
[2] Санкт-Петербургский государственный университет, Санкт-Петербург, Россия, o.mitrofanova@spbu.ru✉, SPIN-код: 4169-6068

Оценка сложности китайско-русского корпуса параллельных и сопоставимых текстов политической тематики

АННОТАЦИЯ. *Статья посвящена оценке сложности китайско-русского корпуса параллельных и сопоставимых текстов политической тематики. Цель исследования заключается в том, чтобы экспериментально сравнить и проанализировать сложность оригинальных и переводных политических текстов на китайском и русском языках с точки зрения количественных и качественных параметров. В статье применялись методы корпусной лингвистики и лингвостатистического анализа. В ходе проведения исследования был разработан новый китайско-русский корпус параллельных и сопоставимых политических текстов. В результате обработки данных о лингвистических признаках, связанных с лексическим разнообразием и индексами удобочитаемости, было установлено, что переводные русскоязычные тексты политической тематики сложнее оригинальных. Эксперименты по оценке сложности текстов корпуса с помощью индексов удобочитаемости позволили проследить изменения сложности текстов во времени. Данное исследование имеет высокую теоретическую и практическую ценность, поскольку оно способствуют более глубокому пониманию проблемы сложности оригинальных и переводных китайских и русских текстов политической тематики, а также стимулирует совершенствование новых инструментов для оценки сложности текстов с учетом их языка, стиля и тематики. Полученные результаты могут быть учтены в практике перевода политических текстов с китайского языка на русский.*

КЛЮЧЕВЫЕ СЛОВА: *сложность текста, политический дискурс, политические тексты, тема политики, параллельный корпус, сопоставимый корпус, переводные тексты, русский язык, китайский язык, переводоведение, переводческая деятельность, политические переводы, сопоставительное языкознание*.

ИНФОРМАЦИЯ ОБ АВТОРАХ: *Чжу Хуэй, аспирант, Даляньский университет иностранных языков; 116044, China, Liaoning, Dalian; email: zhuhui1230@qq.com.*

Митрофанова Ольга Александровна, кандидат филологических наук, доцент кафедры математической лингвистики, Санкт-Петербургский государственный университет; 199034, Санкт-Петербург, В.О., Университетская наб., д. 11, ауд. 193; email: o.mitrofanova@spbu.ru.

БЛАГОДАРНОСТИ. Работа выполнена при поддержке проекта № 202307130002, утвержденного Советом по стипендиям Министерства образования Китая, при поддержке СПбГУ, шифр проекта 124032900006-1. (本文受中国国家留学基金委№ 202307130002圣彼得堡国立大学项目和俄罗斯124032900006-1项目资助号).

ДЛЯ ЦИТИРОВАНИЯ: *Чжу, Хуэй.* Оценка сложности китайско-русского корпуса параллельных и сопоставимых текстов политической тематики / Чжу Хуэй, О. А. Митрофанова. — Текст : непосредственный // Политическая лингвистика. — 2024. — № 5 (107). — С. 218-227.

Zhu Hui[1]✉, Ol'ga A. Mitrofanova[2]✉

[1] Dalian University of Foreign Languages, Dalian, China, zhuhui1230@qq.com✉, https://orcid.org/0009-0003-2922-8156
[2] Saint Petersburg State University, St. Petersburg, Russia, o.mitrofanova@spbu.ru✉, SPIN code: 4169-6068

Assessment of the Complexity of the Chinese-Russian Corpus of Parallel and Comparable Political Texts

ABSTRACT. *This paper focuses on assessing the complexity of the Chinese-Russian corpus of parallel and comparable political texts. The study aims to experimentally compare and analyze the complexity of source-language and target-language political texts in Chinese and Russian from both quantitative and qualitative perspectives. The research employs the methods of corpus linguistics and linguo-statistical analysis. A new Chinese-Russian corpus of parallel and comparable political texts was developed in the course of the study. The analysis of linguistic features related to lexical diversity and readability indices has revealed that Russian target-language political texts are more complex than the source-language texts. The experiments assessing the complexity of the corpus texts using readability indices allowed tracking changes in text complexity over time. This study holds significant theoretical and practical value as it contributes to a deeper understanding of the issue of complexity of source-language and target-language political texts in Chinese and Russian and promotes the development of new tools for assessing text complexity, considering language, style, and topic. The findings can be used in practical translation of political texts from Chinese into Russian.*

KEYWORDS: *text complexity, political discourse, political texts, political topic, parallel corpus, comparable corpus, translated texts, Russian language, Chinese language, translation studies, translation, political translations, comparative linguistics.*

AUTHOR'S INFORMATION: *Zhu Hui, Post-Graduate Student, Dalian University of Foreign Languages, Dalian, China.*

Mitrofanova Ol'ga Aleksandrovna, Candidate of Philology, Associate Professor of Department of Mathematical Linguistics, Saint Petersburg State University, St. Petersburg, Russia.

ACKNOWLEDGMENTS. The work was supported by project No. 202307130002, approved by the Scholarship Council of the Ministry of Education of China with the support of SPbU, project code 124032900006-1.

FOR CITATION: *Zhu Hui, Mitrofanova O. A.* (2024). Assessment of the Complexity of the Chinese-Russian Corpus of Parallel and Comparable Political Texts. In *Political Linguistics*. No 5 (107), pp. 218-227. (In Russ.).

ВВЕДЕНИЕ

В настоящее время спектр лингвистических параметров, представляющих интерес для исследования порождения и восприятия текстов, расширяется. Среди информативных параметров текстов сложность (complexity) стала одной из важнейших областей прикладных лингвистических исследований, связанных с концепцией «легкого/понятного/ясного» языка. Интерес к оценке сложности текстов в реципиент-ориентированных проектах [Mustajoki et al. 2021] подтверждается тем, что арсенал лингвистов пополнился новыми инструментами, применимыми для анализа текстов на различных языках, в том числе и русском: это инструмент «Текстометр» [Лапошина, Лебедева 2021], предназначенный для изучающих русский как иностранный и взрослых носителей русского как родного, платформа-агрегатор «RuLingva» для учителей, школьников и их родителей [Солнышкина и др. 2024], сервис «PolyLing» [Никулина и др. 2023], инструмент анализа сложности юридических текстов [Чаплинский и др. 2024] и т. д.

Актуальность исследования обусловлена тем, что в оценке сложности текстов разных жанров и тематики остаются нерешенные вопросы, в частности, связанные с пригодностью различных способов оценки сложности и их применимостью в обработке многоязычных корпусов текстов. Исследования, связанные с оценкой сложности текстов в процессе их перевода, малочисленны [Нечаева и др. 2020], и наша работа призвана восстановить существующие пробелы. Важность обращения к материалу китайско-русского корпуса параллельных и сопоставимых текстов политической тематики состоит в том, что высокая сложность текстов данного типа определяется особенностями официально-делового стиля, в то же время в качестве адресатов политических текстов выступают как узкие специалисты — политологи, юристы, работники социальной сферы, так и самый широкий круг носителей языка, не имеющих профильного образования, затрагивающего тематику текстов. Исследование сложности политических текстов [Кучаков, Савельев 2018; Митрофанова, Атугодаге 2023 и др.] — это ключ к повышению их понятности для разных групп коммуникантов.

Новизна исследования определяется тем, что в нем впервые проведены эксперименты и получены эмпирические данные о сложности китайских и русских текстов из корпуса параллельных и сопоставимых текстов политической тематики, при этом применяется семейство метрик сложности, допускающих межъязыковой сопоставительный анализ и исследование изменений сложности текстов во времени. **Цель** исследования состоит в комплексной оценке сложности оригинальных и переводных политических текстов на китайском и русском языках с точки зрения количественных и качественных параметров, соотносимых с лексическими, морфологическими и синтаксическими характеристиками текстов. Для достижения поставленной цели необходимо решить ряд **задач**: 1) изучить различные трактовки понятия сложности и смежных с ней явлений; 2) определить лингвистические признаки, являющиеся значимыми для оценки сложности текстов; 3) проанализировать метрики сложности и определить их применимость к работе с исследовательским корпусом; 4) провести эксперименты по оценке сложности китайско-русского корпуса параллельных и сопоставимых текстов политической тематики; 5) обобщить полученные данные и сформулировать выводы о состоятельности используемых метрик и о сходствах и различиях между китайскими и русскими текстами с точки зрения сложности.

Объектом исследования являются лингвистические параметры, определяющие сложность текстов. **Предмет** исследования — метрики сложности, применяемые в анализе текстов на китайском и русском языках. В исследовании используются **методы** корпусной лингвистики, лингвостатистического анализа, метрики сложности, **алгоритмы**

автоматической обработки текстов, реализованные в библиотеках для языка *Python*. **Теоретическая значимость** исследования заключается в проверке гипотезы о связи между типологическими характеристиками языков, лингвистическими параметрами текстов и метриками оценки их сложности, а также о влиянии характеристик сложности на процесс перевода и об изменении сложности текстов определенного жанра и тематики во времени. **Практическая значимость** исследования состоит в возможности применения полученных эмпирических данных для контроля сложности текстов в процессе перевода политических текстов в китайско-русской языковой паре.

1. ТЕОРЕТИЧЕСКАЯ БАЗА

Понятия сложности, читабельности, трудности, понятности. Разнообразие подходов к выбору метрик оценки **сложности** определяется отсутствием единого мнения учёных по вопросу трактовки общего понятия сложности текста и его противопоставленности более специфичным и дополняющим друг друга понятиям **читабельности** (*readability*), **трудности** (*difficulty*), **понятности** (*comprehensibility*) [Кисельников 2015].

Оценка **читабельности** текста обычно дается на основе некоторых количественных параметров. По мнению ряда исследователей, понятие читабельности по своей сути неоднозначно, поскольку оно отражает, с одной стороны, характеристики сложности текста с точки зрения его построения, а с другой — характеристики его потенциальных читателей, процессов восприятия и понимания текстов. Сейчас для оценки читабельности текстов используется значительное число метрик, одними из первых были разработаны индексы Флеша (Flesch Reading Ease) [Flesch 1948] и Флеша — Кинкейда (Flesch — Kincaid Readability Test) [Kincaid et al. 1975], на их основе предложены индексы SMOG [McLaughlin 1969], Колман — Лиау [Coleman, Liau 1975]; Дейла — Челла [Chall, Dale 1995] и ряд других. Применительно к русскому языку были разработаны специфические метрики читабельности: индексы Тулдавы [Тулдава 1975], Оборневой [Оборнева 2006] и др. Для решения задач в области обучения родному и иностранным языках [Белый и др. 2023], журналистики, подготовки и редактирования официально-деловых текстов исследователями применяется свыше двух сотен формул для оценки удобочитаемости текста, пригодных для работы не только с английским, но и с другими языками, среди которых немецкий, французский, русский, китайский и другие.

Понятие читабельности текста уточняется понятием **трудности**, которое таким же образом учитывает объективные аспекты текста (количественные параметры) и субъективные аспекты, связанные с механизмами восприятия текста читателями и их трудозатратами при чтении текста [Казачкова, Галимова 2023]. Оценка трудности текста учитывает разные группы носителей языка с точки зрения их возраста, пола, образования, социального положения, подготовленности к восприятию информации, фоновых знаний, когнитивных способностей и других факторов. Поэтому один и тот же текст, имеющий инвариантную оценку лексической, морфологической, синтаксической сложности, может получить варьирующиеся значения оценки трудности в зависимости от выбора целевой аудитории.

Понятность текста также важна для осмысления категории сложности. Понятность — это «свойство текста содействовать пониманию» [Микк 1981], которое можно оценить не через комбинацию независимых признаков (например, можно трактовать читабельность и трудность), а как меру взаимовлияния объективных (количественных) оценок сложности и субъективных оценок, связанных с социолингвистическим и психологическим профилированием говорящих, с их мотивированностью и функциональными особенностями. Понятность текста коррелирует с длиной его предложений в словоупотреблениях, мерой знакомости слов для читателя, устанавливаемой по лексическим минимумам для разных уровней владения языком как родным или иностранным и по лексико-семантическим признакам словаря.

Сложность текста — это категория, которая характеризует текст в аспектах его языковой, предметной и логической структуры [Судина 2022]. Сложность текста представляется как комплексная мера [Ляшевская 2016], опирающаяся на языковые факторы (выбор лексико-семантических средств, специфичность морфологических форм, организация структуры предложений, сверхфразовых единств, варианты риторической структуры текста, его дискурсивных свойств и т. д.) и не сводимая к информационной сложности текста (знакомству читателя с темой текста), визуальной сложности текста (восприятию шрифтового, иллюстративного оформления текста, инфографики, поликодовости текста и т. д.). Помимо указанных выше факторов, важно учитывать ясность текста и абстрактность лексики [Федюченко 2010], поскольку от этих параметров зависит качество его восприятия и интерпретации. Обобщая информацию о категории сложно-

сти и его коррелятов, в нашем исследовании мы будем следовать представлению об оценке сложности как о комплексе количественных и качественных параметров, включающих лингвостатистические признаки текстов, а также интересы и личный опыт читателя.

Лингвистические параметры текстов, учитываемые в метриках сложности. Именно из-за различий в трактовке понятия сложности текста ученые расходятся в выборе лингвистических параметров, учитываемых при оценке сложности. Исследование Р. Рейнольдса показало, что традиционные метрики, которые используются для оценки удобочитаемости текста (лексические, морфологические, синтаксические, дискурсивные), также полезны для измерения сложности текста [Reynolds 2016]. Среди лексических признаков важны три группы: коэффициенты лексического разнообразия, лексической сложности и лексической знакомости. В экспериментах по машинному обучению наиболее высокий вес имеют морфологические признаки: количество букв, слогов, слов, предложений; средняя длина слова, предложения и т. д. [Там же]. Синтаксические признаки, коррелирующие со сложностью, были изучены подробно применительно к частным задачам оценки сложности текстов различных функциональных стилей, прежде всего это касается длины предложения [Абрамов и др. 2011]. Для русского языка И. В. Оборонева разработала адаптированную формулу для оценки читабельности текстов, основанную на формуле Флеша — Кинкейда [Оборнева, 2006]. Данная формула имплементирована в инструменте *LightReader*, реализованном в виде макроса для *Microsoft Word* и позволяющем рассчитывать такие параметры, как среднее число слогов в слове, число многосложных слов, среднее число слов в предложении и т. д. В исследованиях М. И. Солнышкиной и коллег, развивающих алгоритмы оценки сложности текстов на основе дискурсивных характеристик в модели Coh-Metrix [Солнышкина и др. 2022], отмечается, что наряду с параметрами поверхностного кода (состав слова, частеречные категории, синтаксические признаки и т. д.), необходимо принимать во внимание уровень текста (пропозициональные структуры, референциальные связи, фокус дискурса и т. д.), уровень ситуации (участники ситуации, категории темпоральности, пространства, связи между событиями и т. д.), уровень дискурса (логическая организация текста, эпистемиологический статус и т. д.), прагматический уровень (цели говорящего и слушающего и т. д.).

В нашем исследовании было проведено сочетание основных количественных и качественных параметров. Первая группа параметров, примененных в работе с русско–язычными текстами, — это традиционные морфологические признаки, такие как количество букв, слов, слогов и предложений, части речи и т. д. Вторая группа — это лексическая сложность и лексическое разнообразие. Для изучения лексического разнообразия мы выбрали классический коэффициент Type-Token Ratio (TTR), а также Moving-average TTR (MATTR) [Covington 2010], преимущество которого состоит в меньшей чувствительности к длине текста [Захарова 2020]. Исследование лексической сложности включает следующие параметры: средняя длина слова, количество и доля уникальных, длинных и сложных слов в тексте. Помимо этого, были применены стандартные индексы читабельности текстов для русского языка.

Поскольку русский и китайский языки типологически различны, это отражается в выборе информативных параметров оценки сложности китайского сегмента нашего корпуса и метрик. Китайский язык — это изолирующий язык, в связи с этим применение для оценки сложности обычных индексов читабельности может вызвать затруднения. По данным [Soh 2020], в китайской комплексологии до сих пор не выработано единое мнение ученых о том, какие признаки играют ведущую роль в оценке сложности текстов на китайском языке. В исследовании [Wang 2008] обобщены результаты экспериментов с 1970-х годов, свидетельствующие о том, что для китайского языка важными параметрами, связанными со сложностью текстов, являются среднее количество штрихов в иероглифах, частотность слов, количество слов, среднее количество символов в предложении, длина предложения. При сравнении различных подходов оказалось, что в наборах параметров, предлагаемых учеными, мало пересечений. Так, в исследованиях [Yang 1971] большое внимание уделяется доле трудных слов, а в экспериментах [Sung et al. 2013] по машинному обучению с использованием машины опорных векторов и логистической регрессии наряду с параметром трудных слов значимыми оказались такие параметры, как среднее количество штрихов в иероглифах, доля конкретной лексики и другие.

2. ОПИСАНИЕ ЭКСПЕРИМЕНТАЛЬНОГО КОРПУСА И ИССЛЕДОВАТЕЛЬСКИХ ИНСТРУМЕНТОВ

Эмпирический материал данного исследования представляет собой китайско-рус-

ский корпус параллельных и сопоставимых политических текстов, состоящий из двух подкорпусов: параллельный корпус «Докладов о работе правительства в 2012–2022 гг.» (далее — ДРП), включающий в себя исходный китайский текст (далее — ДРП-К, объем 11 текстов, 114 294 словоупотребления) и перевод на русский язык (далее — ДРП-Р, объем 11 текстов, 133 636 словоупотреблений); сопоставимый корпус «Послания Президента Российской Федерации Федеральному Собранию РФ 2011–2021 гг.» (далее — ППР, объем 11 текстов, 89 790 словоупотреблений). Общий объем корпуса составляет 33 текста, 337 720 словоупотреблений. ДРП-К и ППР — тексты, официально изданные Китаем и Россией, которые являются репрезентативными с точки зрения тематики и стиля. Подробное описание корпуса представлено в статье [Чжу, Захаров 2024]. Для изучения сложности политических текстов подкорпусы ДРП-К, ДРП-Р и ППР были сегментированы по годам.

В ходе экспериментов по анализу сложности текстов с русскоязычными подкорпусами ДРП-Р и ППР мы использовали библиотеку *Russian Texts Statistics* (*ruts*)[1] на языке Python. *ruts* — это многофункциональная библиотека для оценки лингвостатистических параметров текстов на русском языке, которая позволяет анализировать следующие морфологические признаки, получать данные о количестве предложений, уникальных и сложных словах и др., вычислять классические метрики оценки лексического разнообразия и индексы читабельности: индексы Флеша, Флеша — Кинкейда, Колман — Лиау, SMOG, индекс LIX. Автоматический индекс удобочитаемости, реализованный в *ruts*, нами не использовался, поскольку его значения сильно коррелируют со значениями индекса Колман — Лиау.

В силу особенностей отбора информативных параметров для китайского языка вместо стандартных метрик оценки читабельности были применены метрики, реализованные в библиотеке *AlphaReadability Chinese* (*ARC*)[2]. Исследование [Lei et al. 2024] демонстрирует широкие возможности инструмента ARC, который позволяет определять сложность текстов по девяти лингвистическим критериям, отражающим лексические, семантические и синтаксические особенности китайских текстов с точки зрения вероятностно-статистических моделей корпуса, учитывающих оценки энтропии и результаты тематического моделирования с помощью алгоритма LDA (Latent Dirichlet Allocation) [Lei et al. 2024]: лексическое богатство (lexical richness), синтаксическое богатство (syntactic_richness), семантическая точность (semantic accuracy (n, v, n_v, c), семантическое богатство (semantic richness), семантическая ясность (semantic clarity), семантический шум (semantic noise).

3. РЕЗУЛЬТАТЫ АНАЛИЗА

На первом этапе была проведена количественная обработка данных исследовательского корпуса (см. табл. 1). В сегментах ППР и ДРП-Р было определено количество букв, слогов, словоупотреблений, предложений для подкорпусов в целом (столбец «Общее количество») и для отдельных текстов в составе подкорпусов с последующим усреднением (столбец «Среднее значение по текстам»). В целом ДРП-Р по объему значительно больше, чем ППР, как по количеству букв и слогов, так и по количеству словоупотреблений и предложений. Подкорпус ДРП-Р представляет собой перевод китайских политических текстов подкорпуса ДРП-К на русский, соответственно, его количественные характеристики определяются не только объемом исходного текста, но и теми стратегиями перевода и лингвистическими средствами, применяемыми переводчиками. Анализируя полученную информацию, можно выдвинуть гипотезу о том, что с точки зрения сложности текста можно предположить, что чтение и понимание ДРП-Р требует от читателя больших усилий, чем ознакомление с текстами ППР. Кроме того, согласно результатам Чжу Хуэй и В. П. Захарова, процент знаменательных слов (существительных, глаголов и прилагательных) в ДРП-Р выше, чем в ППР [Чжу, Захаров 2024: 119], что является дополнительным свидетельством в пользу выдвинутой гипотезы.

[1] URL: https://pypi.org/project/ruts/
[2] URL: https://github.com/leileibama/AlphaReadabilityChinese

Таблица 1

Количественные показатели подкорпусов ППР и ДРП-Р

Параметры	ППР		ДРП-Р	
	Общее количество	Среднее значение по текстам	Общее количество	Среднее значение по текстам
Количество букв	564991	51362.82	992188	90198.91
Количество слогов	238988	21726.18	417820	37983.64
Количество словоупотреблений	91413	8310.27	135476	12316
Количество предложений	4298	389.91	6416	583.28

Таблица 2

Значения коэффициентов лексического разнообразия TTR и MATTR для ППР и ДРП-Р (Фрагмент)

Параметры		Значения по годам							Mean	RMSD
		2011	2012	2013	…	2019	2020	2021		
TTR	ППР	0.42	0.43	0.37		0.38	0.40	0.41	0.40	0.027
	ДРП-Р	0.36	0.39	0.38		0.42	0.33	0.34	0.37	0.025
MATTR	ППР	0.92	0.91	0.90		0.90	0.91	0.91	0.91	0.006
	ДРП-Р	0.89	0.88	0.89		0.90	0.88	0.88	0.89	0.007

На втором этапе работы для проверки выдвинутой гипотезы мы обратились к данным о лексическом разнообразии текстов, оценки которого получены с помощью TTR и MATTR. В табл. 2. представлены результаты оценки TTR и MATTR для подкорпусов ППР и ДРП-Р, а также среднего значения Mean и среднеквадратического отклонения RMSD. В целом по расчетам RMSD для TTR существенно выше (0.027, 0.025), чем для MATTR (0.006, 0.007). Было замечено, что в ППР средние значения TTR и MATTR несколько выше (0.40, 0.91), чем в ДРП-Р (0.37, 0.89). Однако в 2013 и 2019 годах значения TTR в ППР (0.37 и 0.38) ниже, чем в ДРП-Р (0.38 и 0.42), что связано со значительным уменьшением длины текста ДРП-Р в эти два года. TTR для текстов ППР и ДРП-Р в 2018 году одинаково низки (0.34). На основании статистических оценок TTR и MATTR можно сделать вывод, что повышение лексического разнообразия в текстах ППР по сравнению с текстами ДРП-Р отражает особенности перевода с китайского на русский. В ППР местоимения используются гораздо чаще, чем в ДРП-Р, то есть в оригинальных политических текстах частотные дейктические слова регулярно обеспечивают связи между фрагментами, в то время как в переводных текстах переводчики используют знаменательную лексику вместо местоимений, что может облегчить читателям понимание содержания документов и снизить их неоднозначность.

Аналогичные наблюдения позволяет сделать оценка параметров лексической сложности, представленных в табл. 3. В текстах ДРП-Р по сравнению с ППР несколько меньше уникальных, простых и односложных слов, значительно больше длинных, трудных и многосложных слов, что указывает на более высокую сложность подкорпуса ДРП-Р и меньшую сложность подкорпуса ППР.

Таблица 3

Лексическая сложность для ППР и ДРП-Р

Параметры	ППР		ДРП-Р	
	Среднее значение	Доля	Среднее значение	Доля
Уникальные слова	3285.00	39.53%	4476.18	36.34%
Длинные слова	4549.00	54.74%	8139.09	66.09%
Трудные слова	2313	27.83%	4860.56	39.47%
Простые слова	5415.46	65.17%	6605.27	53.63%
Односложные слова	1733.73	20.86%	1976.18	16.05%
Многосложные слова	5994.73	72.14%	9489.64	77.05%

Третий этап экспериментов направлен на получение значений метрик сложности с помощью библиотеки *ruts,* в которой реализованы основные метрики оценки сложности текстов, адаптированные для русского языка. По шкалам для индексов Флеша, Флеша — Кинкейда, Колман — Лиау, SMOG тексты корпуса характеризуются самой высокой сложностью (от 13.10 до 25.61 при контрольном диапазоне 0…30) и предназначены для носителей языка с академической подготовкой. Единственный индекс, получающий высокое значение в экспериментах, это индекс LIX, который при превышении порога 60 (ППР 76.36, ДРП-Р 87.29) указывает на то, что анализируемый текст обладает исключительной сложностью и скорее всего относится к официально-деловому стилю. Сравнение данных, полученных для подкорпусов ППР и ДРП-Р, указывает на следующие факты: значения метрик для ППР ниже соответствующих показателей для ДРП-Р, что окончательно подтверждает выдвинутую нами гипотезу о повышенной сложности переводных текстов по сравнению с оригинальными русскоязычными документами. Аналогично результатам в экспериментах по оценке лексического разнообразия, среднеквадратическое отклонение RMSD заметно выше для подкорпуса ППР (1.056…4.856), чем для подкорпуса ДРП-Р (0.817…4.289). Дополнительное замечание касается индекса Флеша, который демонстрирует высокое среднеквадратическое отклонение RMSD как в подкорпусе ППР, так и в подкорпусе ДРП-Р. Тем самым, сложность является варьирующимся параметром текстов ППР и относительно стабильным для текстов ДРП-Р. Поскольку оценки сложности были получены как для подкорпусов в целом, так и для отдельных документов в составе корпуса, нам удалось оценить динамику изменений в количественных характеристиках текстов. В наших данных прослеживается тенденция к повышению сложности политических текстов со временем. Для ППР наблюдается локальный максимум для основных индексов, приходящийся на 2017 год (исключение составляет индекс Флеша, достигающий максимума в 2019 году), тогда как для ДРП-Р повышение сложности происходит более равномерно, и значения индексов достигают наибольших значений к 2022 году.

В ходе анализа сложности подкорпуса китайских текстов ДРП-К с помощью инструмента *ARC* мы изучили две группы метрик. Результаты представлены в табл. 6. Группа I объединяет метрики, позволяющие оценить лексико-семантическое и синтаксическое богатство (lexical, semantic, syntactic richness) текста и семантический шум (semantic noise), при этом чем выше значения метрик, тем сложнее текст. Группа II представляет оценки семантической точности (semantic accuracy) и ясности (semantic

Таблица 4

Метрики оценки сложности текстов подкорпуса ППР (Фрагмент)

Метрики	Значения по годам							Mean	RMSD
	2011	2012	2013	…	2019	2020	2021		
Индекс Флеша	22.43	14.08	22.68		29.51	29.34	24.66	21.42	4.856
Индекс Флеша — Кинкейда	12.24	14.01	12.99		11.22	10.44	13.15	13.10	1.356
Индекс Колман — Лиау	13.41	15.28	13.41		11.85	12.11	13.27	13.76	1.056
Индекс SMOG	20.89	23.14	21.61		19.47	18.72	21.64	21.72	1.520
Индекс LIX	73.78	78.53	76.17		72	71.27	76.65	76.36	2.894

Таблица 5

Метрики оценки сложности текстов подкорпуса ДРП-Р (Фрагмент)

Метрики	Значение по годам							Mean	RMSD
	2012	2013	2014	…	2020	2021	2022		
Индекс Флеша	13.41	10.21	9.15		13.06	21.15	22.21	13.89	4.289
Индекс Флеша — Кинкейда	16.14	16.88	15.73		14.99	17	17.09	16.33	0.817
Индекс Колман — Лиау	20.59	19.91	19.69		20.1	22.1	22.34	20.75	0.872
Индекс SMOG	25.16	26.46	24.97		23.81	26.42	26.54	25.61	0.993
Индекс LIX	86.59	87.03	85.03		84.65	89.9	90.15	87.29	1.999

Таблица 6

Метрики оценки сложности текстов подкорпуса ДРП-К (Фрагмент)

Метрики		Значения ДРП-К						Mean	RMSD
	2012	2013	2014	…	2020	2021	2022		
I									
lexical_richness	6.30	6.27	6.43		6.38	6.41	6.43	6.43	0.080
syntactic_richness	2.14	2.12	2.14		2.13	2.11	2.11	2.13	0.011
semantic_richness_n	0.27	0.28	0.27		0.25	0.27	0.27	0.27	0.009
semantic_noise_n	57.81	59.32	55.83		35.76	34.45	52.79	50.81	10.274
II									
semantic_accuracy_n	3.67	3.69	3.72		3.62	3.74	3.65	3.70	0.042
semantic_accuracy_v	8.19	8.58	8.05		8.12	7.97	7.88	8.05	0.194
semantic_accuracy_n_v	6.04	6.12	6.02		6.18	6.03	6.00	6.04	0.064
semantic_accuracy_c	6.46	6.44	6.35		6.55	6.44	6.41	6.42	0.067
semantic_clarity_n	0.03	0.03	0.03		0.02	0.02	0.02	0.03	0.005

clarity), для них чем ниже значения, тем проще и понятнее текст. Заметим, что значения метрик обеих групп имеют довольно низкое значение среднеквадратического отклонения RMSD (0.005…0.194), исключение составляет метрика семантического шума (10.274). Полученные данные позволяют высказать наблюдение о том, что китайские тексты из подкорпуса ДРП-К имеют высокую сложность, на что явно указывают высокие показатели семантического шума (semantic noise) и низкие показатели семантической ясности (semantic clarity).

ЗАКЛЮЧЕНИЕ

В результате серии проведенных экспериментов по оценке сложности китайско-русского корпуса параллельных и сопоставимых текстов политической тематики были получены ценные данные о роли лингвистических параметров, представляющих разноуровневые явления в китайском и русском языках. Цель исследования была достигнута благодаря решению следующих задач.

Были проанализированы подходы к трактовке понятия сложности, его соотношения с читабельностью, трудностью и понятностью в трудах российских и зарубежных исследователей, что позволило определить соотношение количественных и качественных параметров, учитываемых при оценке текста в аспектах его структурной организации, коммуникативных целей, его восприятия носителями языка из различных групп и т. д.

В зависимости от целей оценки сложности текстов следует корректно проводить отбор лингвистических параметров, которые представляют явления лексического, морфологического, синтаксического, семантического уровней, а также внутритекстовые связи, ситуативную соотнесенность текста, дискурсивные и прагматические параметры текста. На отбор параметров сложности повлияли типологические различия между китайским и русским языками.

В ходе экспериментов мы выдвинули и проверили гипотезу о различиях в сложности оригинальных русскоязычных текстов и переведенных с китайского на русский. Данные о метриках сложности позволили подтвердить данную гипотезу и получить эмпирические данные, свидетельствующие в пользу того, что тексты подкорпуса ППР характеризуются меньшей сложностью по сравнению с подкорпусом ДРП-Р. При анализе китайского подкорпуса ДРП-К метрики показали свою состоятельность, при этом оказалось, что наиболее явными показателями высокой сложности оригинальных политических текстов на китайском языке следует считать метрики семантической ясности и семантического шума. Дополнительное исследование было проведено с целью выявить изменения сложности текстов со временем. Была установлена общая тенденция к усложнению политических текстов, при этом выявленные локальные максимумы сложности можно связать с событиями в общественно-политической жизни Китая и России.

Перспективы дальнейшего исследования связаны с расширением реестра лингвистических признаков, привлечением дополнительных индексов сложности для китайского и русского языков, с установлением корреляции между группами индексов, с привлечением дополнительных источников текстового материала, со сравнением сложности политических текстов и текстов разной тематической, стилевой принадлежности.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Абрамов, В. Е. Статистический анализ связности текстов по общественно-политической тематике / В. Е. Абрамов,

Н. Н. Абрамова, Е. В. Некрасова, Г. Н. Росс. — Текст : непосредственный // Труды 13-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». — 2011. — С. 127–133.

2. Белый, А. В. Автоматическая генерация лексико-грамматических заданий по русскому языку как иностранному с помощью предсказывающих языковых моделей / А. В. Белый, О. А. Митрофанова, Н. А. Дубинина — Текст : непосредственный // Мир русского слова. — 2023. — № 2. — С. 108–118.

3. Захарова, Е. Ю. Лексическое разнообразие текста и способы его измерения / Е. Ю. Захарова, О. Ю. Савина. — Текст : непосредственный // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. — 2020. — № 1 (21).

4. Казачкова, М. Б. Лексическое разнообразие как параметр сложности текстов учебников английского языка (на примере учебников «SPOTLIGHT» и «STARLIGHT») / М. Б. Казачкова, Х. Н. Галимова. — Текст : непосредственный // Ученые записки Крымского федерального университета имени В. И. Вернадского. Филологические науки. — 2023. — № 4. — С. 178–186.

5. Кисельников, А. С. К проблеме характеристик текста: читабельность, понятность, сложность, трудность / А. С. Кисельников. — Текст : непосредственный // Филологические науки. Вопросы теории и практики. — 2015. — № 11 (53). — С. 79–84.

6. Кучаков, Р. К. Сложность правовых актов в России: лексическое и синтаксическое качество текстов / Р. К. Кучаков, Д. А. Савельев. — Санкт-Петербург : ИПП ЕУ СПб, 2018. — Текст : непосредственный.

7. Лапошина, А. Н. Текстометр: онлайн-инструмент определения уровня сложности текста по русскому языку как иностранному / А. Н. Лопошина, М. Ю. Лебедева. — Текст : непосредственный // Русистика. — 2021. — № 3 (19). — С. 331–345.

8. Ляшевская, О. Н. Индексы удобочитаемости как мера оценки сложности текста / О. Н. Ляшевская. — Текст : электронный // Доклад НУГ ВШЭ. — 2016. — URL: https://ling.hse.ru/data/2016/12/15/1111563794/Readability%20talk.pdf.

9. Микк, Я. А. Оптимизация сложности учебного текста: в помощь авторам и редакторам / Я. А. Микк. — Москва : Просвещение, 1981. — 119 с. — Текст : непосредственный.

10. Митрофанова, О. А. Динамическое тематическое моделирование русскоязычного корпуса юридических документов / О. А. Митрофанова, М. М. Атугодаге. — Текст : непосредственный // Terra Linguistica. — 2023. — Т. 14. — № 1. — С. 70–87.

11. Нечаева, Н. В. Перевод на ясный и простой языки: зарубежный опыт и перспективы в России / Н. В. Нечаева, К. С. Хельмле, Э. М. Каирова. — Текст : непосредственный // Вестник Пермского национального исследовательского политехнического университета. Проблемы языкознания и педагогики. — 2020. — № 3. — С. 8–24.

12. Никулина, Е. Р. Разработка сервиса для оценки удобочитаемости текста с применением технологий машинного обучения / Е. Р. Никулина, А. В. Черкас, Е. Д. Козина, А. В. Бойко, Л. А. Дмитриева. — Текст : электронный // SAEC. — 2023. — № 2. — URL: https://cyberleninka.ru/article/n/razrabotka-servisa-dlya-otsenki-udobochitaemosti-teksta-s-primeneniem-tehnologiy-mashinnogo-obucheniya.

13. Оборнева, И. В. Автоматизированная оценка сложности учебных текстов на основе статистических параметров : автореф. дис. … канд. пед. наук / Оборнева И. В. — Москва, 2006. — Текст : непосредственный.

14. Солнышкина, М. И. Лингвистическое профилирование текста: приключенческая повесть vs. учебник / М. И. Солнышкина, Р. В. Куприянов, Г. Н. Шоева. — Текст : непосредственный // Научный результат. Вопросы теоретической и прикладной лингвистики. — 2024. — Т. 10. — № 1. — С. 115–132.

15. Солнышкина, М. И. Обработка естественного языка и изучение сложности дискурса / М. И. Солнышкина, Д. Макнамара, Р. Р. Замалетдинов. — Текст : непосредственный // Russian Journal of Linguistics. — 2024. — № 2 (26). — С. 317–341.

16. Судина, И. И. Трудности текстов для чтения при обучении иностранному языку / И. И. Судина. —Текст : электронный // Символ науки. — 2022. — № 7. — С. 29–32.

17. Тулдава, Ю. А. Об измерении трудности текстов / Ю. А. Тулдава. — Текст : непосредственный // Ученые записки Тартуского университета. — Тарту : Изд-во Тарт. ун-та. — 1975. — № 4. — С. 102–120.

18. Федюченко, Л. Г. Когнитивный подход в описании уровня сложности текста на перевод / Л. Г. Федюченко. — Текст : непосредственный // Lingua mobilis. — 2010. — № 2 (21). — С. 170–175.

19. Чаплинский, А. В. Понятность языка правосудия: опыт эмпирического исследования содержания и синтаксиса судебных решений / А. В. Чаплинский, А. В. Кнутов, Д. Р. Алимпеев. — Текст : непосредственный // Закон. — 2024. — № 2. — С. 159–177.

20. Чжу, Хуэй. Корпусное сравнение языка китайских и российских политических текстов / Чжу Хуэй, В. П. Захаров. — Текст : непосредственный // Политическая лингвистика. — 2024. — № 1 (103). — С. 115–128.

21. Chall, J. S. Readability revisited : the new Dale-Chall readability formula / J. S. Chall, E. Dale. — Text : unmediated // Brookline Books. — 1995. — P. 159.

22. Coleman, M. A. computer readability formula designed for machine scoring / M. A. Coleman, T. L. Liau — Text : unmediated // Journal of Applied Psychology. —1975. — № 60 (2). — P. 283–284.

23. Covington, M. A. Cutting the Gordian knot: The moving-average type—token ratio (MATTR) / M. A. Covington., J. D. McFall. — Text : unmediated // Journal of Quantitative Linguistics. — 2010. — № 2. — P. 94–100.

24. Flesch, R. A new readability yardstick / R. Flesch. — Text : unmediated // Journal of Applied Psychology. — 1948. — P. 221–233.

25. Kincaid, J. P. Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel / J. P. Kincaid, R. P. Fishburne, R. L. Rogers, B. S. Chissom. — Text : unmediated // Chief of Naval Technical Training: Naval Air Station Memphis. —1975.

26. Lei, L. AlphaReadabilityChinese: A tool for the measurement of readability in Chinese texts and its applications / Lei L., Wei Y., Liu K. — Text : unmediated // Foreign Languages and Their Teaching. — 2024. — № 46 (1). — P. 83–93.

27. McLaughlin, G. H. SMOG grading: A new readability formula / G. H. McLaughlin. — Text : unmediated // Journal of Reading. — 1969. —№ 12 (8). — P. 639–646.

28. Mustajoki, A. Easy language in Russia / A. Mustajoki, Zh. Mihienko, N. Nechaeva, E. Kairova, A. Dmitrieva. — Text : unmediated // Handbook of Easy Languages in Europe / Eds. C. Lindholm, U. Vanhatalo. — Berlin : Frank & Timme, 2021. — P. 439–466.

29. Reynolds, R. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories / R. Reynolds. — Text : unmediated // Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. — 2016. — P. 289–300.

30. Soh, K. Readability Formula for Chinese as a Second Language / K. Soh. — Text : unmediated // Teaching Chinese Language in Singapore. — 2020. — P. 51–62.

31. Sung, Y. T. An exploration of readability of Chinese texts: Selection of indicators, model construction, and verification of validity / Y. T. Sung, J. L. Chen, Y. S. Lee, J. H. Cha, H. C. Tseng, K. E. Chang. — Text : unmediated // Chinese Journal of Psychology. — 2013. — № 55 (1). — P. 75–106.

32. Wang, L. The connotation and research model of readability formulas: Readability for Chinese as a foreign language and its research task / L. Wang. — Text : unmediated // Language Teaching and Linguistic Studies. — 2008. — № 6. —P. 46–53.

33. Yang, S. J. A readability formul a for Chinese language : Unpublished doctoral dissertation / S. J. Yang. — Madison : University of Wisconsin-Madison, 1971. — Text : unmediated.

REFERENCES

1. Abramov, V.E., Abramova, N.N., Nekrasova, E.V., & Ross, G.N. (2011). Statistical analysis of the cohesion of texts on sociopolitical topics. *Proc. of the 13th All-Russian Scientific Conference*, 127–133. (In Russ.)

2. Belyi, A.V., Mitrofanova, O.A., & Dubinina, N.A. (2023). Automatic generation of lexico-grammatical tasks in Russian as a foreign language with the help of predictive language models. *The World of Russian Word*, *2*, 108–118. (In Russ.)

3. Zakharova, E.Y., & Savina, O.Y. (2020). Lexical diversity of the text and ways of its measurement. *Vestnik of Tyumen State University*, *1*(21). (In Russ.)

4. Kazachkova, M.B., & Galimova, H.N. (2023). Lexical diversity as a parameter of complexity of texts of English language textbooks (on the example of textbooks "SPOTLIGHT" and "STARLIGHT"). *Scientific Notes of the V. I. Vernadsky Crimean Federal University*, *4*, 178–186. (In Russ.)

5. Kiselnikov, A.S. (2015). To the problem of text characteristics: readability, comprehensibility, complexity, difficulty. Philological Sciences. *Questions of theory and practice*, *11*(53), 79–84. (In Russ.)

6. Kuchakov, R.K., & Savelyev, D.A. (2018). *Complexity of legal acts in Russia: Lexical and syntactic quality of texts*. St. Petersburg: IPP EUSPb, 20p. (In Russ.)

7. Laposhina, A.N., & Lebedeva, M.Y. (2021). Textometer: an online tool for determining the level of text complexity in Russian as a foreign language. *Rusistika*, *3*(19), 331–345. (In Russ.)

8. Lyashevskaya, O.N. (2016). Readability indices as a measure of text complexity assessment. *Doklad NUG VSHE*. Retrieved from https://ling.hse.ru/data/2016/12/15/1111563794/Readability%20talk.pdf?ysclid=m013y6lkcl788928979 (In Russ.)

9. Mikk, Y.A. (1981). *Optimization of the complexity of the educational text: To help authors and editors*. Moscow: Prosveshchenie, 119 p. (In Russ.)

10. Mitrofanova, O.A., & Atugodage, M.M. (2023). Dynamic thematic modeling of the Russian-language corpus of legal documents. *Terra Linguistica*, *1*(14), 70–87. (In Russ.)

11. Nechaeva, N.V., Helmle, K.S., & Kairova, E.M. (2020). Translation into clear and simple languages: foreign experience and prospects in Russia. *Vestnik PNIPU. Problems of linguistics and pedagogy*, *3*, 8–24. (In Russ.)

12. Nikulina, E.R., Cherkas, A.V., Kozina, E.D., Boyko, A.V., & Dmitrieva, L.A. (2023). Developing a service for text readability assessment using machine learning technologies. *SAEC. 2*. (In Russ.)

13. Oborneva, I.V. (2006). *Automated estimation of complexity of educational texts on the basis of statistical parameters* (Cand. ped. sciences). Moscow. (In Russ.)

14. Solnyshkina, M.I., Kupriyanov, R.V., & Shoeva, G.N. (2024). Linguistic profiling of the text: adventure story vs. textbook. *Scientific Result. Problems of theoretical and applied linguistics*, *1*(10), 115–132. (In Russ.)

15. Solnyshkina, M.I., McNamara, D., & Zamaletdinov, R.R. (2022). Natural Language Processing and the Study of Discourse Complexity. *Russian Journal of Linguistics*, *2*(26), 317–341. (In Russ.)

16. Sudina, I.I. (2022). Difficulties of reading texts in teaching a foreign language. *The symbol of science*, *7-2*, 29–32. (In Russ.)

17. Tuldava, J.A. (1975). About measuring the difficulty of texts. *Scientific Notes of the University of Tartu*, *4*, 102–120. (In Russ.)

18. Fedyuchenko, L.G. (2010). Cognitive approach in describing the level of text complexity per translation. *Lingua mobilis*, *2*(21), 170–175. (In Russ.)

19. Chaplinsky, A.V., Knutov, A.V., & Alimpeev, D.R. (2024). The comprehensibility of the language of justice: experience of empirical study of the content and syntax of court decisions. *Law*, *2*, 159–177. (In Russ.)

20. Zhu, Hui, & Zakharov, V.P. (2024). Corpus comparison of the language of Chinese and Russian political texts. *Political linguistics*, *1*(103), 115–128. (In Russ.)

21. Chall, J.S., & Dale, E. (1995). *Readability revisited: the new Dale-Chall readability formula*. Cambridge, Mass. : Brookline Books, 159 p.

22. Coleman, M.A., & Liau, T.L. (1975). Computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *2*(60), 283–284.

23. Covington, M.A., & McFall, J.D. (2010). Cutting the Gordian knot: The moving-average type—token ratio (MATTR). *Journal of Quantitative Linguistics*, *2*(17), 94–100.

24. Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221–233.

25. Kincaid, J.P., Fishburne, R.P., Rogers, R.L., & Chissom, B.S. (1975). *Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel*. Chief of Naval Technical Training: Naval Air Station Memphis.

26. Lei, L., Wei, Y., & Liu, K. (2024). AlphaReadability Chinese: A tool for the measurement of readability in Chinese texts and its applications. *Foreign Languages and Their Teaching*, *1*(46), 83–89.

27. McLaughlin, G.H. (1969). SMOG grading: A new readability formula. *Journal of Reading*, *8*(12), 639–646.

28. Mustajoki, A., Mihienko, Z., Nechaeva, N., Kairova, E., Dmitrieva, A. (2021). Easy language in Russia. In *Handbook of Easy Languages in Europe* (pp. 439–466).

29. Reynolds, R. (2016). Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 289–230.

30. Soh, K. (2020). Readability Formula for Chinese as a Second Language. *Teaching Chinese Language in Singapore*, 51–62.

31. Sung, Y.T., Chen, J.L., Lee, Y.S., Cha, J.H., Tseng, H.C., & Chang, K.E. (2013). An exploration of readability of Chinese texts: Selection of indicators, model construction, and verification of validity. *Chinese Journal of Psychology*, *1*(55), 75–106.

32. Wang, L. (2008). The connotation and research model of readability formulas: Readability for Chinese as a foreign language and its research task. *Language Teaching and Linguistic Studies*, *6*, 46–53.

33. Yang, S.J. (1971). *A readability formul a for Chinese language* (Unpublished doctoral dissertation). University of Wisconsin-Madison, 107 p.