ББК 81.1
Т78

Ответственный редактор издания
В. П. Захаров

Т78 **Труды международной конференции «Корпусная лингвистика–2021».** — СПб.: Издательство Скифия-принт, 2021. — 396 с.

ISSN 2412-9623
ISBN 978-5-98620-557-1

Сборник содержит материалы докладов, представленных на научной конференции «Корпусная лингвистика-2021» 1–3 июля 2021 г. в Санкт-Петербурге.

Создание и использование корпусов текстов является одним из приоритетных направлений в современной лингвистике. Проведение конференции по данной тематике знакомит ученых с современными разработками и новыми технологическими решениями в этой области, а также способствует обобщению опыта научных исследований по корпусной лингвистике.

ББК 81.1

ISSN 2412-9623
ISBN 978-5-98620-557-1

© Авторы, 2021

О. В. Блинова, Н. А. Тарасов

O. V. Blinova, N. A. Tarasov

СЛОЖНОСТЬ РУССКИХ ПРАВОВЫХ ТЕКСТОВ: МЕТОДЫ ОЦЕНКИ И ЯЗЫКОВЫЕ ДАННЫЕ[1]

COMPLEXITY OF RUSSIAN LEGAL TEXTS: ASSESSMENT METHODS AND LANGUAGE DATA

Аннотация. Для создания модели автоматического определения сложности русских правовых текстов было необходимо собрать коллекцию таких текстов, разметить их, выделить параметры оценки сложности в применении к выбранному формату разметки. Эти шаги описываются в настоящей работе. Обозначается состав корпусов современных русских юридических текстов CorRIDA, CorDec, CorCodex общим объемом 8,5 млн токенов. Описываются основания выбора инструментов лингвистической разметки (UDPipe, pymorphy2). Кратко характеризуются языковые признаки оценки сложности, среди которых: простейшие базовые метрики; пять формул читабельности; параметры оценки лексической сложности (значения TTR, Yule's K, количество гапаксов, аббревиатур, абстрактных слов и мн. др.); параметры оценки морфосинтаксической и дискурсивной сложности (значения Noun-Verb Ratio; количество граммем генитива, среднего рода, пассива; относительных предложений, аппозитивных модификаторов, лексических средств дискурсивной связности и пр.).
Ключевые слова. Языковая сложность, правовые документы, читабельность, лексическая сложность, морфосинтаксическая сложность, дискурсивная сложность, русские синхронные юридические корпусы.

Abstract. Our goal is to create a model for the automatic assessment of Russian legal texts complexity. To achieve this goal, it is necessary to create a text collection; perform linguistic markup; highlight the parameters for measuring the complexity, oriented on the selected markup format. These steps are described in this paper. We briefly describe three corpora of modern Russian legal texts "CorRIDA", "CorDes", "CorCodex" with a total size of 8.5 million tokens. We justify the choice of linguistic markup tools (UDPipe, pymorphy2). Then we characterize the linguistic features of the complexity assessment, including: the simplest basic metrics; five readability formulas; parameters for assessing lexical complexity (TTR values, Yule's K, the number of hapaxes, abbreviations, abstract words, etc.); parameters for assessing morphosyntactic and discursive complexity (Noun-Verb Ratio values; the number of grammemes of genitive, neuter, passive; relative sentences, appositive modifiers, lexical devices of discursive connectivity, etc.).
Keywords. Linguistic complexity, legal documents, lexical complexity, morphosyntactic complexity, discursive complexity, synchronous corpora of legal Russian.

[1] При поддержке гранта РНФ № 19-18-00525 «Понятность официального русского языка: юридическая и лингвистическая проблематика».

1. Введение

Языковой сложности посвящена обширная научная литература, как теоретическая, напр., [Dahl 2004; Бердичевский 2014], так и охватывающая конкретные методы оценки текстов (см., напр., обзоры в [Collins-Thompson 2014; Солнышкина, Кисельников 2015]). Современные методы подразумевают далеко не только использование классических формул читабельности. Как указано в [Crossley et al. 2019], «классические формулы читабельности в меньшей степени предсказывают понимание текста, чем формулы, разработанные с использованием лингвистических параметров, основанные на словесных признаках <…> и на признаках, которые оценивают вхождения лексических и синтаксических конструкций, связность текста, тональность, тематику и семантику». Таким образом, при оценке сложности используется все большее количество сложных для извлечения текстовых признаков, не только морфосинтаксических и лексических, но и дискурсивных (отражающих содержательную и структурную связность).

Юридический язык издавна критикуется за многословие, избыточность, длинноты, синтаксическую переусложненность, архаичную лексику и пр. (см., напр., [Tiersma 1999]).

Неудивительно, что и русские правовые тексты привлекли внимание исследователей, которые, во-первых, сконцентрировались в основном на оценке сложности текстов законов, во-вторых, использовали для оценки сложности либо только формулы читабельности (в [Дмитриева 2017] используется известная формула И. В. Оборневой), либо другие достаточно простые и немногочисленные метрики. Так, в [Кучаков, Савельев 2018] используется одна лексическая метрика (TTR, значение которой зависит от длины текста) и одна синтаксическая (расстояние между главным и зависимым по синтаксическому дереву зависимости, вычисляемое так: «для каждого конкретного текста взято одно значение, которое является максимальным для всех предложений текста» [Там же]).

В новейшей работе [Кнутов и др. 2020] использовано большее количество метрик (девять), среди которых: «доля глаголов в страдательном залоге», «доля глаголов от общего количества слов в тексте», «среднее количество слов в субстантивных именных словосочетаниях», «среднее количество причастных оборотов, расположенных в предложениях после определяемого слова, на одно предложение»,

«среднее количество деепричастных оборотов на одно предложение», «среднее количество слов в предложениях», «среднее расстояние между зависимыми словами в предложении», «среднее количество грамматических основ (предикативных основ, предикативных ядер) предложения (подлежащее, сказуемое или одно из них) в одном предложении», «среднее количество слов в абзаце».

2. Состав и объем юридических корпусов

Нашей целью является создание модели автоматического определения сложности русских правовых текстов, учитывающей значительное количество разнообразных параметров. Для достижения этой цели на начальном этапе было необходимо собрать коллекцию таких текстов, разметить их, выделить параметры оценки сложности в применении к выбранному формату разметки. Эти шаги описываются в разделах настоящей работы.

Мы собрали, предобработали и разметили три коллекции современных русских юридических текстов общим объемом около 8,5 млн токенов.

Во-первых, это коллекция русских локальных документов и актов CorRIDA, содержащая документы, с которыми периодически сталкиваются носители языка — неюристы (формы информированных согласий, договоров и пр., скачанные с сайтов государственных учреждений). Корпус CorRIDA состоит из 1546 документов и содержит 1 млн 784 тыс. токенов.

Во-вторых, это коллекция решений Конституционного Суда РФ CorDec, включающая 584 документа, 3427 тыс. токенов. Решения пишутся высокопрофессиональными юристами и адресованы широкому кругу граждан, описание см. в [Blinova et al. 2020a].

В-третьих, это коллекция нормативных документов CorCodex, содержащая 279 текстов кодексов, федеральных законов и постановлений Правительства РФ (в общей сложности 3 млн 227 тыс. токенов). Такие тексты вынуждены читать прежде всего профессиональные юристы.

Размеченные файлы корпусов в формате *.json будут полностью опубликованы на сайте plaindocument.org в конце 2021 г.

3. Разметка корпусов

Известно, что синтаксические признаки хорошо предсказывают языковую сложность, см., например, [Ivanov et al. 2018]. Корпусы, размеченные в формате UD (Universal Dependencies), в последнее время все более активно используются при оценке морфосинтаксической сложности как при межъязыковом сопоставлении, так и при сравнении текстов (коллекций текстов) на одном языке, см., например, [Berdicevsks et al. 2018; Çöltekin, Rama 2018; Yan, Kahane 2018; Dyer 2018].

Поэтому в качестве базового инструмента разметки наших трех коллекций выбран UDPipe. Как инструмент подробного морфологического анализа взят pymorphy2 [Korobov 2015].

Отдельной задачей стал выбор между доступными моделями UDPipe (существуют модели ru-syntagrus, ru-gsd, ru-taiga). Основанием для принятия решения стала статистика метрик, показывающая аккуратность работы моделей, представленная М. Стракой [Universal Dependencies 2.5 Models for UDPipe; Straka 2017]. Согласно этой статистике, где даны значения метрик аккуратности по параметрам UAS, LAS, MLAS и BLEX [CoNLL 2018 Shared Task], модель russian-syntagrus-ud-2.5 работает лучше, поэтому нами выбрана именно она.

После предобработки выполнена автоматическая лемматизация, морфологическая и синтаксическая разметка корпусов. Каждой словоформе присвоена двойная частеречная помета — в терминах UDPipe и в терминах pymorphy2. Частеречная разметка pymorphy2 позволяет различать ADJF — полные прилагательные, ADJS — краткие прилагательные, VERB — глаголы в личной форме, INFN — инфинитивы, PRTF — полные причастия, PRTS — краткие причастия и GRND — деепричастия. Это удобно для оценки морфосинтаксической сложности, в частности, потому, что наблюдается положительная корреляция между количеством полных прилагательных (а также причастий и деепричастий) и сложностью и отрицательная корреляция между количеством глаголов в личной форме и сложностью текстов [Дружкин 2016].

4. Выбранные параметры оценки языковой сложности

С целью оценки языковой сложности юридических текстов в составе собранных корпусов отобрано более 50 параметров. Значения каждого из параметров будут записаны в состав метаданных к текстам корпусов.

Среди параметров — значения простых (базовых) метрик текстов и индексы сложности, вычисленные при помощи формул читабельности. Кроме того, более содержательным образом будут оцениваться в значительной степени условно различаемые лексическая, морфосинтаксическая и дискурсивная сложность.

В качестве базовых метрик решено использовать (среди прочих): ASL — среднюю длину предложения в словах, ASW — среднюю длину слова в слогах. Из формул читабельности, адаптированных для русского языка, выбраны FRE (GL), SMOG, ARI, DCI, CLI, см. [Бегтин 2016; Solovyev et al. 2018].

В качестве параметров оценки лексической сложности выбраны: значения простого TTR; значения метрик лексического разнообразия Yule's K и Yule's I, не зависящие от длины текстов, см. [Blinova et al. 2020a]; количество гапаксов; количество лемм со значениями индексов частотности Zipf Value, в том числе лемм с низкой общеязыковой частотностью [Blinova et al. 2020b]; количество аббревиатур и сокращений; количество слов с абстрактным значением; количество юридических терминов. При оценке сложности будут использованы пользовательские словари, в том числе словарь абстрактных слов [Solovyev et al. 2020].

Для оценки морфосинтаксической сложности выбраны, кроме прочего: количество неслужебных слов, в частности, существительных; значения Noun-Verb Ratio; количество подчинительных и сочинительных союзов; количество тегов граммем родительного падежа, среднего рода, пассива; количество относительных предложений, аппозитивных модификаторов и предложений, содержащих отношения parataxis.

Среди учитываемых дискурсивных признаков, например, количество лексических средств дискурсивной связности, в частности, дискурсивных маркеров (в UD ter discourse); повтор именных групп в соседних предложениях (эта информация полезна для оценки референциальной связности); повтор значений граммем времени и вида у глаголов в личной форме.

5. Перспективы

Корпусы собраны и размечены, параметры оценки сложности текстов отобраны. Ближайшим шагом станет автоматическое вычисление значений этих параметров. Затем с помощью методов машинного

обучения мы создадим модель автоматического определения сложности русских текстов.

После определения численных значений (итоговых индексов) сложности будут использованы средства машинного обучения. Применение моделей трансферного обучения позволит не только с большой точностью определять сложность текстов, но и установить, какие параметры текстов в наибольшей степени коррелируют с целевой сложностью. В качестве примера нейросети с подобной архитектурой можно привести модель Universal Sentence Encoder [Cer et al. 2018]. Эта модель кодирует текст в многомерные векторы, которые можно использовать для классификации, определения семантического сходства, кластеризации и других задач обработки естественного языка. Задача определения сложности текстов в данном контексте является задачей регрессии, и метод USE в ней можно использовать для первоначального кодирования текстов и дальнейшего обучения модели методами глубокого обучения или градиентного бустинга.

Литература

1. *Бегтин И. В.* (2016), Оценка читабельности текста. URL: https://github.com/ivbeg/readability.io/wiki/API (дата обращения: 01.07.2021).
2. *Бердичевский А.* (2012), Языковая сложность (Language Complexity). Вопросы языкознания. № 5, с. 101–124.
3. *Дмитриева А. В.* (2017), «Искусство юридического письма»: количественный анализ решений Конституционного Суда Российской Федерации. Сравнительное конституционное обозрение. Т. 118, № 3, с. 125–133.
4. *Дружкин К. Ю.* (2016), Метрики удобочитаемости для русского языка: выпускная квалификационная работа магистра. М.: НИУ ВШЭ.
5. *Кнутов А. В., Плаксин С. М. и др.* (2020), Сложность российских законов. Опыт синтаксического анализа. М.: ИД Высшей школы экономики.
6. *Кучаков Р., Савельев Д.* (2018), Сложность правовых актов в России: лексическое и синтаксическое качество текстов. СПб.: Институт проблем правоприменения Европейского университета в Санкт-Петербурге.
7. *Солнышкина М. И., Кисельников А. С.* (2015), Сложность текста: этапы изучения в отечественном прикладном языкознании. Вестник Томского государственного университета. Филология. № 6(38), с. 86–99.

References

1. *Begtin I. V.* (2016), Ocenka chitabel'nosti teksta [Text Readability Assessment], URL: https://github.com/ivbeg/readability.io/wiki/API (date of access: 01.07.2021).
2. *Berdicevskij A.* (2012), Jazykovaja slozhnost' [Language Complexity]. In: Voprosy Jazykoznanija [Topics in the study of language]. No. 5, pp. 101–124.

3. *Berdicevskis A., Çöltekin Ç., Ehret K., Prince K., Ross D., Thompson B., Yan C., Demberg V., Lupya G., Rama T., Bentz C.* (2018), Using Universal Dependencies in cross-linguistic complexity research. In: Proceedings of the Second Workshop on Universal Dependencies (UDW 2018).
4. *Blinova O. V., Belov S. A., Revazov M. A.* (2020a), Decisions of Russian Constitutional Court: Lexical Complexity Analysis in Shallow Diachrony. R. V. Bolgov, A. V. Chugunov, A. E. Voiskounsky (eds.). In: CEUR Workshop Proceedings. Vol. 2813. Proceedings of the International Conference "Internet and Modern Society" (IMS-2020), St. Petersburg, Russia, 17–20 June 2020, pp. 61–74.
5. *Blinova O. V., Tarasov N. A., Modina V. V., Blekanov I. S.* (2020b), Modeling Lemma Frequency Bands for Lexical Complexity Assessment of Russian Texts. In: Komp'juternaja Lingvistika i Intellektual'nye Tehnologii [Computational Linguistics and Intellectual Technologies]. Vol. 19(26), pp. 76–92.
6. *Cer D., Yang Y., Kong S., Hua N., Limtiaco N. et al.* (2018), Universal Sentence Encoder. ArXiv: 1803.11175.
7. *Collins-Thompson K.* (2014), Computational assessment of text readability: a survey of current and future research. In: Th. François, D. Bernhard (eds.). Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics. Vol. 165(2), pp. 97–135.
8. *Çöltekin Ç., Rama T.* (2018), Exploiting universal dependencies treebanks for measuring morphosyntactic complexity. In: Proceedings of the First Shared Task on Measuring Language Complexity, pp. 1–8.
9. CoNLL 2018 Shared Task. URL: http://universaldependencies.org/conll18/evaluation.html (date of access: 01.07.2021).
10. *Crossley S. A., Skalicky S., Dascalu M.* (2019), Moving beyond classic readability formulas: new methods and new models. In: Journal of Research in Reading. No. 42(3-4), pp. 541–561.
11. *Dahl Ö.* (2004), The growth and maintenance of linguistic complexity. Amsterdam.
12. *Dmitrieva A. V.* (2017), «Iskusstvo juridicheskogo pis'ma»: kolichestvennyj analiz reshenij Konstitucionnogo Suda Rossijskoj Federacii. Sravnitel'noe konstitucionnoe obozrenie ["The art of legal writing": A quantitative analysis of Russian Constitutional Court rulings]. In: Sravnitel'noe konstitucionnoe obozrenie [Comparative Constitutional Review]. No. 118(3), pp. 125–133.
13. *Druzhkin K. Ju.* (2016), Metriki udobochitaemosti dlja russkogo jazyka: vypusknaja kvalifikacionnaja rabota magistra [Readability metrics for Russian: master's theses]. Moscow: Higher School of Economics.
14. *Dyer W.* (2018), Integration complexity and the order of cosisters. In: Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pp. 55–65.
15. *Ivanov V. V., Solnyshkina M. I., Solovyev V. D.* (2018), Efficiency of text readability features in Russian academic texts. In: Komp'juternaja Lingvistika i Intellektual'nye Tehnologii 2018 [Computational Linguistics and Intellectual Technologies]. Vol. 17, pp. 277–287.
16. *Knutov A. V., Plaksin S. M. et al.* (2020), Slozhnost' rossijskih zakonov. Opyt sintaksicheskogo analiza [The complexity of Russian laws. Syntactic analysis experience]. Moscow: Publishing house of the Higher School of Economics.

17. *Korobov M.* (2015), Morphological Analyzer and Generator for Russian and Ukrainian Languages. In: Analysis of Images, Social Networks and Texts, pp. 320–332.
18. *Kuchakov R., Saveliev D.* (2018), Slozhnost' pravovyh aktov v Rossii: leksicheskoe i sintaksicheskoe kachestvo tekstov [The complexity of legal acts in Russia: Lexical and syntactic quality of texts: analytic note]. European University at Saint Petersburg, St. Petersburg.
19. *Solovyev V., Solnyshkina M., Andreeva M., Danilov A., Zamaletdinov R.* (2020), Text Complexity and Abstractness: Tools for the Russian Language. In: R. V. Bolgov, A. V. Chugunov, A. E. Voiskounsky (eds.). CEUR Workshop Proceedings. Vol. 2813. Proceedings of the International Conference "Internet and Modern Society" (IMS-2020), St. Petersburg, Russia 17–20 June 2020, pp. 75–87.
20. *Solnyshkina M. I., Kisel'nikov A. S.* (2015), Slozhnost' teksta: jetapy izuchenija v otechestvennom prikladnom jazykoznanii [Text complexity: study phases in Russian linguistics]. In: Vestnik Tomskogo Gosudarstvennogo Universiteta, Filologiya [Tomsk State University Journal of Philology]. No. 6(38), pp. 86–99.
21. *Tiersma P.* (1999), Legal Language. Chicago: University of Chicago Press.
22. Universal Dependencies 2.5 Models for UDPipe. URL: https://github.com/jwijf-fels/udpipe.models.ud.2.5/blob/master/inst/udpipe-ud-2.5-191206 (date of access: 01.07.2021).
23. Universal Dependencies. URL: https://universaldependencies.org.

Блинова Ольга Владимировна
Санкт-Петербургский государственный университет (Россия)
Национальный исследовательский университет
«Высшая школа экономики» (НИУ ВШЭ СПб) (Россия)
Blinova Olga
Saint Petersburg State University (Russia)
National Research University Higher School of Economics (Russia)
E-mail: o.blinova@spbu.ru

Тарасов Никита Андреевич
Санкт-Петербургский государственный университет (Россия)
Tarasov Nikita
Saint Petersburg State University (Russia)
E-mail: tarasovn2468@yandex.ru