

**Галюченко Анастасия Сергеевна**

преподаватель

Одинцовский филиал

МГИМО МИД России

Одинцово, Россия

**Anastasiya Galyuchenko**

Teacher

Odintsovo branch of MGIMO University

Odintsovo, Russia

a.galyuchenko@odin.mgimo.ru

**НЕЙРОННЫЕ ТЕХНОЛОГИИ И КОРПУСНЫЕ ИНСТРУМЕНТЫ  
ДЛЯ ИССЛЕДОВАНИЯ ЮРИДИЧЕСКОГО МЕДИАДИСКУРСА  
(на примере термина smart-contract)**

В данной работе рассматривается использование инструментов корпусной лингвистики для исследования юридического медиадискурса, а нейронные технологии как вспомогательный инструмент для анализа материала. Целью работы является описание алгоритма работы с корпусом и нейронной сетью в лексикографических целях, в частности, для формирования дефиниций. Автором был сформирован авторский корпус в размере 48 текстов (научные и научно-популярные тематические тексты по теме цифрового права), что составляет 120 тысяч слов. В статье описывается стратегия работы с авторским корпусом и нейронной генеративной моделью на примере термина *smart-*

*contract*, который входит в область цифрового права. Термин анализируется как отдельный термин, а также в контексте юридического медиадискурса (в корпусе). Дана исчерпывающая дефиниция, полученная в результате исследования использования термина в речи: был осуществлен поиск в существующих словарях, сформирован запрос в генеративной модели Copilot, конкордансный поиск в авторском корпусе. В заключении формулируется дефиниция термина на базе проанализированных данных.

К л ю ч е в ы е с л о в а: *корпусная лингвистика; AntConc; юридический медиадискурс; OpenAI; Copilot; дефиниция; лексикография.*

## NEURAL TECHNOLOGIES AND CORPUS TOOLS FOR LEGAL MEDIA DISCOURSE RESEARCH ("SMART-CONTRACT" AS AN EXAMPLE)

This paper considers the use of corpus linguistics tools for the study of legal media discourse, and neural technologies as an auxiliary tool for analysing the material. The author has generated an author corpus of 48 texts (scientific and popular science thematic texts on the topic of digital law), which amounts to 120 thousand words. The article describes the strategy of working with the author corpus and neural generative model on the example of the term smart-contract, which is part of the digital law domain. The term is analysed as a separate term and also in the context of legal media discourse (in the corpus). An exhaustive definition is given, obtained as a result of the study of the use of the term in speech: a search in existing dictionaries, a query in the Copilot generative model, and a concordance search in the author's corpus were performed. In conclusion, the definition of the term is formulated on the basis of the analysed data.

К е y w o r d s: *corpus linguistics; AntConc; legal media discourse; OpenAI; Copilot; definition; lexicography.*

В настоящее время применение корпусных и нейронных технологий активно возрастает как в области исследований, так и в языковой подготовке, особенно когда речь идет об обучении переводу. Корпусы – это наборы текстов, которые можно использовать для поиска слов, фраз и закономерностей в языке. Нейронные технологии чаще всего используются в виде чат-ботов, которые являются генеративными моделями человеческой речи и заранее обучены. Благодаря таким современным технологиям для пользователя становится возможно проводить как собственные исследования языка и речи, так и подготавливать, например, обучающий материал. Корпусы уже давно признаны ценным ресурсом в языковой педагогике. [1; 2; 3; 4; 5; 6; 7; 8; 9; 10]. Во многочисленных книгах, журнальных статьях и докладах на конференциях предлагаются различные варианты применения корпусов: от более точного описания определенного целевого термина до создания всеобъемлющих материалов для преподавания и изучения языка при самостоятельном обучении. [11; 12]. Данный вопрос активно рассматривается в работах следующих отечественных ученых: С. А. Маник, А. В. Дмитриев, М. С. Коган, И. В. Дерябина, О. В. Садовникова и др. Среди зарубежных ученых Ник Эллис, Амадор Морено, Йе Джинтин и др. среди зарубежных. Помимо лингвистических целей данные инструменты могут применяться и для создания глоссариев. По мере развития компьютерных технологий существующие бумажные словари не всегда успевают дать актуальную лексику в своем составе

ве. [13; 14; 15; 16]. Так, представляется целесообразным создать алгоритм работы, который бы позволил ускорить данный процесс, например, с использованием корпуса текстов и нейронных сетей. Помимо скорости сбора материала данные инструменты позволяют совершить выборку наиболее актуального материала и проанализировать нюансы употребления.

Существует ряд компьютерных программ, которые можно использовать для изучения языка. AntConc [17] – одна из компьютерных программ, которая применяется для анализа больших объемов текстовых данных. Ее используют лингвисты, исследователи и другие специалисты для оперативного анализа больших объемов текста. С помощью AntConc для пользователей появляется возможность искать слова, фразы и модели в тексте, а также наглядно представлять результаты. Кроме этого, в данный момент набирает популярность использование разработок компании OpenAI, которая занимается машинным обучением, в частности, чат-ботами. Они функционируют с помощью языковой модели под названием ChatGPT. Эта языковая генеративная модель позволяет обрабатывать запрос, и генерировать соответствующие ответы.

Цель исследования заключается в выделении сильных сторон упомянутых инструментов и их использования при обучении студентов-юристов английскому языку для специальных целей на примере термина smart-contract. Исследование построено по следующей модели: 1) поиск дефиниций в существующих словарях; 2) формирование и анализ запроса в чат-боте; 3) конкордансный поиск в авторском корпусе. Алгоритм выстроен таким образом, чтобы: 1) выяснить, что уже известно о термине (в словаре); 2) проанализировать информацию, которую нашел чат-бот, и иметь примерное представление о термине (все значения, в том числе и новые, которых не будет в словаре); 3) источники авторского корпуса строго отбирались по тематике и его достоверность выше, чем в чат-боте. Кроме этого, корпус позволяет найти нюансы употребления термина.

Таким образом, при анализе термина в первую очередь представляется целесообразным обратиться к информации в словарях. В силу того, что данный вид договоров стал активно появляться после запуска Ethereum в 2015 г<sup>1</sup>, необходимо воспользоваться словарями, которые обновлялись хотя бы раз после 2015 г, или же онлайн-словарями. Были проанализированы ряд словарей по запросу “smart-contract”<sup>1,2,3,4,5</sup>, однако термин не был найден.

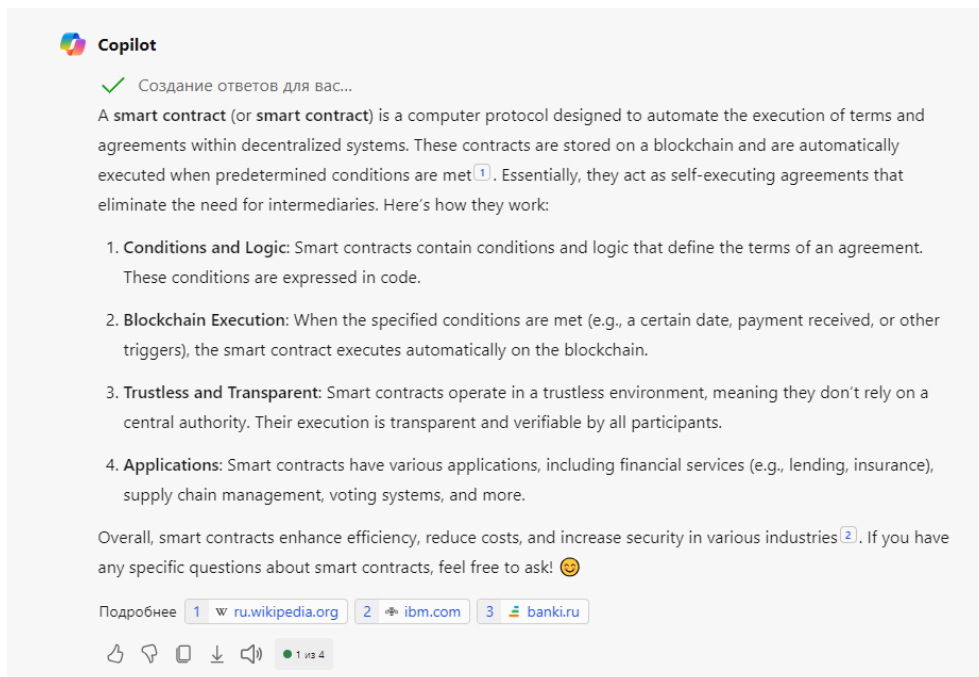
---

<sup>1</sup> What is ChatGPT and why does it matter? Here’s what you need to know // ZDNET URL: <https://www.zdnet.com/article/what-is-chatgpt-and-why-does-it-matter-heres-everything-you-need-to-know/> (дата обращения: 28.06.2024).

<sup>2</sup> Search Legal Terms and Definitions // LAW. COM LEGAL DICTIONARY URL: <https://dictionary.law.com/Default.aspx?typed=smart%20contract&type=1> (дата обращения: 28.06.2024).

<sup>3</sup> Search Results for: smart contract // The Law Dictionary URL: <https://thelawdictionary.org/?s=smart+contract> (дата обращения: 28.06.2024).

Затем, термин был обработан с помощью инструмента Microsoft Copilot – этот инструмент создан на базе искусственного интеллекта GitHub и OpenAI (рисунок).



Результат запроса в системе Copilot

Очевидно, что нейронная генеративная модель позволяет собрать информацию из сети Интернет и сформулировать объяснение. Однако представляется важным отметить, что данный результат не является словарной дефиницией. Согласно С. В. Гринев-Гриневичу, дефиниция должна быть точной и указывать на ближайшее родовое понятие. Стоит отметить и преимущества: дефиниция не содержит логический круг, не противоречит сама себе и не является отрицательной. Учитывая, что данный термин, скорее всего, будет запрошен юристом (отдельная область знаний от программирования), при обработке стоит сократить использование информации, которая может оказаться непонятной (computer protocol). При этом благодаря корпусу возможно дополнить данную дефиницию. Авторский корпус состоит из 48 текстов или 120 тыс. слов. В качестве источников текстов использованы научные и научно-популярные статьи юристов в области цифрового права. Таким образом, данные тексты принадлежат к юридическому медиадискурсу. Юридический медиадискурс представляет собой гибрид юридического дискурса и медиадискурса. В него входят и медиа, и право [18]. В данном виде дискурса пересекаются разные отрасли: само право и другие, например, экономика.

<sup>4</sup> “smart contract” // Merriam-Webster URL: <https://www.merriam-webster.com/legal/smart%20contract> (дата обращения: 28.06.2024).

<sup>5</sup> Search suggestions for smart contract // Cambridge Dictionary URL: <https://dictionary.cambridge.org/spellcheck/english/?q=smart+contract> (дата обращения: 28.06.2024).

Так, благодаря анализу результатов кажется целесообразным дополнить в дефиниции нейронной сети следующие признаки: *smart contract is a form of transaction to exercise some particular rights or promises. It may perform one or several conditions. It is not guaranteed by law.*

В итоге, финальная дефиниция может звучать следующим образом: *Smart contract is a form of transaction, where the execution of particular terms, agreements, rights or promises is automated due to blockchain technology. It is performed within a decentralized system, so it is not guaranteed by law. However, smart contract is transparent for all parties.* Так, были даны функция «умного договора», его состав и перечислены несколько свойств, а также то, как он формируется.

Безусловно, нейронные технологии являются отличным инструментом в исследованиях языка и речи, позволяют ускорить процесс создания глоссария и сделать информацию полной и актуальной. Однако их комбинирование с другими инструментами, например, с корпусом дает наиболее полные результаты.

## ЛИТЕРАТУРА

1. Дерябина И. В. Характеристики педагогической работы с английским национальным корпусом // Вестник ТГУ. 2012. № 95.
2. Садовникова О. Э. Прямое и косвенное использование корпусов в зарубежной лингводидактике // Magister Dixit. 2013. № 2.
3. Apresjan V. Russian constructions with syntactic reduplication of colour terms: a Corpus study // Вестник РУДН. Серия: Лингвистика. 2018. № 3.
4. Vydrin V. Where corpus methods hit their limits: the case of separable adjectives in Vambara // Rhema. Рема. 2018. № 4.
5. Дмитриев А. В., Коган М. С. Потенциал корпусной лингвистики в подготовке специалистов в области компьютерной лингводидактики // Terra Linguistica. 2019. № 4.
6. Голомазова М. В. Возможные трудности в использовании лингвистического корпуса в обучении английскому языку // Вестник науки. 2020. № 5 (26)3.
7. Горина О. Г., Царакова Н. С. Корпусные инструменты, маршруты и эксперименты в современной лингводидактике // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2021. № 24.
8. Ye J. Form-frequency correspondence in adjectives: a cross-linguistic corpus approach // Acta Linguistica Petropolitana. Труды института лингвистических исследований. 2021. № 17.
9. Савина О. Ю. Влияние типа ввода на эффективность усвоения коллокаций (корпус vs словарь) // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2022. № 3.

10. Колотаева А. Ю. Обучение студентов извлечению языковых данных о сложносоставных существительных немецкого языка с помощью корпусной поисковой системы Sketch Engine // Вестник ТГГПУ. 2022. № 1 (67).
11. Абдурахманова Л. А., Алиев З. Г. Использование информационных технологий в процессе изучения английского языка на материале корпусной лингвистики // Кронос. 2022. № 4 (66)1.
12. Ашхарава А. Т., Баранов А. Г. Формирование дидактической базы для преподавания языка для специальных целей с использованием приемов компьютерной лингвистики // Современное педагогическое образование. 2019. № 102.
13. Гринев-Гриневиц С. В. Терминоведение : учебное пособие. М. : Академия, 2008. 304 с. (Высшее профессиональное образование).
14. Карпова О. М. Новые вызовы современной английской лексикографии // Вестник Воронежского государственного университета. Серия: Лингвистика и межкультурная коммуникация. Воронеж, 2018. № 3. С. 24–28.
15. Иконникова В. А. Теория терминологического знака В. М. Лейчика и исследование лингвокультурологической составляющей отраслевых терминологий // Вестник МГУ. Теория перевода. 2018. № 4. С. 68–74.
16. Маник С. А. Политическая лексикография: перспективы развития ивановской лексикографической школы // Теория и практика иностранного языка в высшей школе : сб. науч. статей. Иваново, 2018. С. 162–170.
17. AntConc [Electronic resource]. URL: <https://laurenceanthony.net/software/antconc/> (accessed: 26.06.2024).
18. Чеметева Ю. В. Юридический медиадискурс как гибридный феномен [Электронный ресурс] // Вестник Волгоградского государственного университета. Серия 2, Языкознание. 2022. Т. 21, № 1. С. 110–118. (На англ. яз.). DOI: <https://doi.org/10.15688/jvolsu2.2022.1.9>