

УДК 519.876.2

**ЧАСТОТНЫЙ АНАЛИЗ УНИГРАММ АНГЛОЯЗЫЧНЫХ БУКТРЕЙЛЕРОВ**Р.Р. Исмагилов<sup>1</sup>, В.А. Михайлова<sup>2</sup>Научный руководитель: доцент, к.ф.-м.н., М.Е. Семёнов<sup>1</sup><sup>1</sup>Национальный исследовательский Томский политехнический университет,

Россия, г. Томск, пр. Ленина, 30, 634050,

<sup>2</sup>Московский городской педагогический университет,

Россия, г. Москва, 2-ой Сельскохозяйственный проезд, 4, к. 1, 129226

E-mail: rri10@tpu.ru

**FREQUENCY ANALYSIS OF TRANSCRIBED ENGLISH BOOK TRAILERS UNIGRAMS**R.R. Ismagilov<sup>1</sup>, V.A. Mikhailova<sup>2</sup>Scientific Supervisor: Assoc. Prof., PhD. M.E. Semenov<sup>1</sup><sup>1</sup>Tomsk Polytechnic University, Russia, Tomsk, Lenin str., 30, 634050,<sup>2</sup>Moscow City University, Russia, 2nd Selskhozjajstvenny dr., 4, building. 1, 129226

E-mail: rri10@tpu.ru

**Abstract.** *The state of book trailers is getting pervasive nowadays as the digital era provides publishers with new tools of the advertisement. The frequency analysis using transcripts of the English book trailers' unigrams was conducted to set the most commonly used words and word groups subsequently divided into categories according to qualification and age. The results revealed that the most frequently used word is "one" (333 mentions) as both a numeral and a pronoun and the second is "book" with 278 mentions. The categories' data featured significant information that may be further leveraged in the areas of teaching the English language and book marketing.*

**Введение.** Буктрейлер - это короткий видеоролик по мотивам книги, основная задача которого рассказать о книге и заинтересовать читателя. Коммерческий аспект подобных видеороликов был закреплён Ведомством по патентам и товарным знакам США в 2003 году. В литературе крайне мало описано результатов исследований специфики воздействия буктрейлеров на потенциальных читателей. В частности, эксперимент [1] нацелен на исследование взаимосвязи книжной видеорекламы и кратких аннотаций в повышении спроса на неё. Опрос [2] продемонстрировал как интерпретации режиссёра влияет на ценность буктрейлера в качестве фактора стимулирования сбыта. В статье [3] представлено исследование книжного спроса посредством риторической теории. На основании данных исследований можно сделать вывод, что содержимое буктрейлеров, включая лексику и выражения, используемые автором, оказывают влияние на потенциального потребителя. Однако до сих пор не была приведена информация о семантической составляющей и характерной лексике аудиосопровождения буктрейлера.

Цель работы – разработка программы для проведения частотного анализа униграмм англоязычных буктрейлеров для выявления востребованных речевых конструкций в книгоиздательской среде. Для достижения цели необходимо провести сбор данных, транскрибировать аудиодорожки с помощью

Automatic Speech Recognizer (ASR) модели, провести предобработку текстовых данных и вычислить частоты встречаемости каждого слова относительно всех слов в наборе данных.

**Экспериментальная часть.** Мы использовали аудиодорожки, извлеченные из англоязычных буктрейлеров, размещенных на портале YouTube (410 записей: 110 профессиональных и 100 любительских записей, ориентированных на взрослую аудиторию; 100 профессиональных и 100 любительских записей – детскую аудиторию). Для профессиональных буктрейлеров, ориентированных на взрослую аудиторию, использовались видеоролики издательств Dutton Books, Penguin Books, Harvest Books, для профессиональных детских буктрейлеров - Scholastic. Полученные аудиофайлы были переведены в текст с помощью модели NVIDIA Conformer-Transducer [4], обученной на датасете NeMo ASRSET. Текстовые файлы были очищены от служебных слов и местоимений. Над каждым словом в документах была проведена процедура лемматизации с помощью модели WordNet [5].

**Модель Conformer.** В работе [6] представлена нейросетевая модель для распознавания речи, сочетающая достоинства свёрточных нейронных сетей и трансформеров. Архитектура трансформеров, содержащих механизм self-attention, позволяет обнаружить глобальную закономерность между элементами экземпляра входных данных исходя из их контекста. В свою очередь, свёрточные нейронные сети позволяют извлечь локальные признаки в данных, при этом плохо справляясь с оценкой признаков в глобальном плане. В основе энкодера модели Conformer лежит модель Masagon-Net [7], заменяя слой прямого распространения в блоке трансформера на два слоя прямого распространения с полуостаточной связью. Между модулями прямого распространения расположены модули механизма self-attention и свёрточный модуль. Второй модуль прямого распространения подаётся на вход слоя нормализации.

Свёрточный модуль состоит из оператора нормализации по слою; двух свёрточных слоев Pointwise [8] – свёртка с ядром размером 1x1, которая проходит по каждому элементу экземпляра данных; функции активации GLU; свёрточного слоя Depthwise [9] – свёртка проходит по каждому из каналов отдельно, после чего результаты фильтрации конкатенируются; оператора нормализации по батчу и функции активации Swish. В качестве метода регуляризации используется Dropout. В основе модуля self-attention лежит слой множественного внимания с относительными позиционными эмбедингами. Модуль прямого распространения состоит из оператора нормализации по слою; двух линейных слоёв; функции активации Swish; двух регуляризаторов Dropout. Результаты анализа приведены в таблице 1.

Таблица 1

Топ-5 частот униграмм англоязычных буктрейлеров по категориям

№	Профессиональные				Любительские			
	Взрослые		Детские		Взрослые		Детские	
1	book	135	book	59	one	113	one	62
2	one	119	friend	40	know	71	dog	49
3	like	118	one	39	book	70	new	47
4	know	105	new	36	like	65	life	44
5	really	92	life	32	story	56	like	44

**Заключение.** Данное исследование позволило выявить лексику, используемую в буктрейлерах с целью привлечения внимания аудитории к прочтению книг. В зависимости от возрастной группы, профессиональные буктрейлеры либо уделяют большее внимание апелляции к авторитету (например,

частое использование предлог «like»), знаниям, самой информации («know») и усилению упоминаемых утверждений посредством наречия «really», либо обращаются к жизненным ценностям, таким как дружба («friend») и жизнь («life»), а также к новизне происходящего («new»).

Любительские буктрейлеры подтверждают, какой именно спрос существует в книгоиздательской сфере – чему посвящены прочитанные авторами видео произведения и что именно они выделяют в них, когда стремятся поделиться избранной историей с другими людьми. Таким образом, видеоролики, нацеленные на детскую аудиторию, часто упоминают наличие в историях собак («dog»).

Все буктрейлеры, независимо от навыков автора и возрастной группы, на которую направлена реклама, объединяет частое употребление и числительного, и местоимения «one». Важно также отметить, что лексика, используемая в подобных рекламных видеороликах, по функциональности близка не только к художественному стилю речи, но и к разговорному. Данная характеристика позволяет использовать буктрейлеры в методических целях – например, в обучении английскому языку на начальном и среднем уровнях.

### СПИСОК ЛИТЕРАТУРЫ

1. Basaraba, N. The Rhetoric of Transmedia Storytelling in Book Trailers Published on YouTube // Msc, Diss. Edmonton. – 2015. – 86 p.
2. Schafer-Hock, C., & Hartmann, D. (2013). How do book trailers affect the recipients? Results of the first online experiment in Germany. *Communication Today*, 2, pp. 66-79.
3. Davila, D. (2010). Not so innocent: Book trailers as promotional text and anticipatory stories. *The ALAN Review*, no. 38(1), pp. 32-42.
4. NVIDIA Conformer-Transducer X-Large [Электронный ресурс]: <http://huggingface.co/nvidia/stt-en-conformer-transducer-xlarge> (Дата обращения: 15.03.2023).
5. Khyani, Divya & B S, Siddhartha (2021) An Interpretation of Lemmatization and Stemming in Natural Language Processing. *Shanghai Ligong Daxue Xuebao. Journal of University of Shanghai for Science and Technology*, no. 22, pp. 350-357.
6. Gulati A. et al. (2020) Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100.
7. Lu, Y., Li, Z., He, D., Sun, Z., Dong, B., Qin, T., Wang, L., & Liu, T. (2019). Understanding and Improving Transformer From a Multi-Particle Dynamic System Point of View. ArXiv, abs/1906.02762.
8. Hua, Binh-Son & Tran, Minh-Khoi & Yeung, Sai-Kit. (2018). Pointwise Convolutional Neural Networks, pp. 984-993.
9. Yunhui, G., Li, Y., Wang, L. & Rosing, T. (2019). Depthwise Convolution Is All You Need for Learning Multiple Visual Domains. *Proceedings of the AAAI Conference on Artificial Intelligence*, no. 33. pp. 8368-8375.