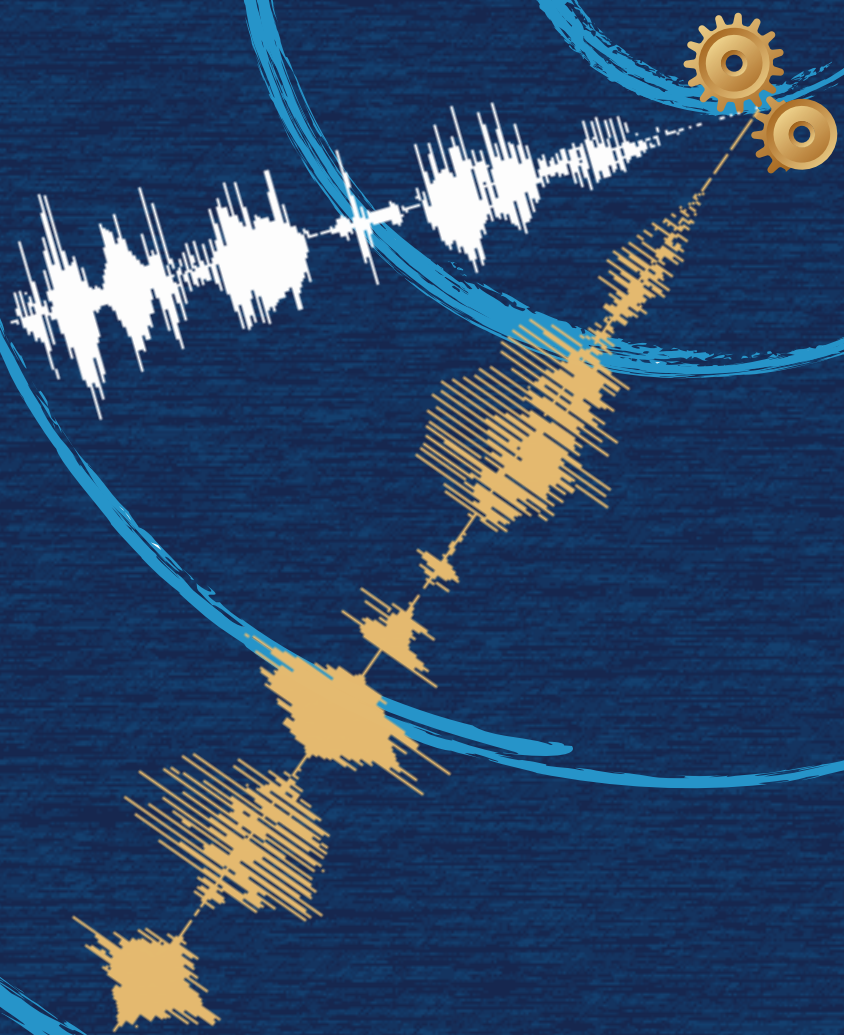
Научно-практический журнал

РЕЧЕВЫЕ ТЕХНОЛОГИИ



ISSN 2305-8129

Speech technology

2 2024

Речевые технологии

2/2024

Содержание

Редакция:

Редактор: *Татьяна Иванова*
Корректор: *Людмила Асанова*
Дизайн: *Анна Ладанюк*
Вёрстка: *Евгения Соболь*

Адрес редакции: 109341, Москва, ул. Люблинская, д. 157, корп. 2.
Тел.: +7 495 345 52 00
Электронная почта: narob@yandex.ru
Сайт: www.narodnoe.org

Подписано в печать 23.12.2024. Формат 60×90⅛. Бумага офсетная. Печать офсетная.
У.-п.л. 9,47. Заказ № 24С28. Тираж 100 экз. Издательский дом «Народное образование».
Отпечатано в типографии ИД "Народное образование".
109341, Москва, ул. Люблинская, д. 157, корп. 2. Тел.: +7 495 345 52 00/59 00.

Текстовый анализ и алгоритм численной аппроксимации стохастических дифференциальных уравнений в задачах финансовой математики

Габдрахманова Н. Т., *доцент, Российский университет дружбы народов им. Патриса Лумумбы, Москва, Россия*

В статье рассматривается одна из задач финансовой математики, а именно задача прогнозирования процентных ставок ценных бумаг. Разработана гибридная модель на основе стохастических дифференциальных уравнений и текстового анализа новостных сообщений. При построении модели используется алгоритм аппроксимации решений стохастических дифференциальных уравнений (СДУ) с помощью нейронных сетей. Проводится анализ результатов прогнозирования, полученных с помощью решения стохастических дифференциальных уравнений. Показано, что методы NLP новостных сообщений позволяют классифицировать экономическую ситуацию в будущем, а композиция моделей NLP и СДУ позволяет повысить точность решения задачи прогнозирования. Анализ адекватности построенной гибридной модели проведён на данных процентных ставок.

стохастическое дифференциальное уравнение • нейронная сеть • процентная ставка • прогноз • текстовый анализ

1. ВВЕДЕНИЕ

В условиях нестабильности финансовых рынков стохастические математические модели являются стандартными моделями для финансовых величин, например процентных ставок, цен активов. В статье рассматривается проблема построения прогноза для процентных ставок. В настоящее время разработано достаточно много методов построения математических моделей для прогнозирования. Однако, ввиду сложности задачи, проблема прогнозирования финансовых рядов до сих пор актуальна. Цель работы: разработать новые подходы в построении прогноза финансовых рядов, позволяющие снизить риск финансовых операций.

В данной работе в качестве стохастической модели использована диффузионная модель эволюции процентных ставок — модель Васичека [1]. Финансовые ряды обобщают всю информацию экономической ситуации в предыдущие моменты

времени, но в них нет информации о новых ситуациях в экономике. С целью прогноза возникновения новых ситуаций в экономике предлагается использовать текстовый анализ новостных сообщений. Использование анализа текстовых ситуаций позволяет классифицировать поведение экономической ситуации в будущем. Спрогнозированный класс ситуации позволяет более точно определить соответствующую ему траекторию движения процентной ставки.

1.1. Data

Материалом для исследования послужили новостные данные (табл. 1) и данные курса доллара в период с 01.01.2000 00:00 по 30.09.2020 23:59 (рис. 1).

Таблица 1

Количественные параметры набора данных

Parameter	Value
Новости	800 000
USD/RUB	1259

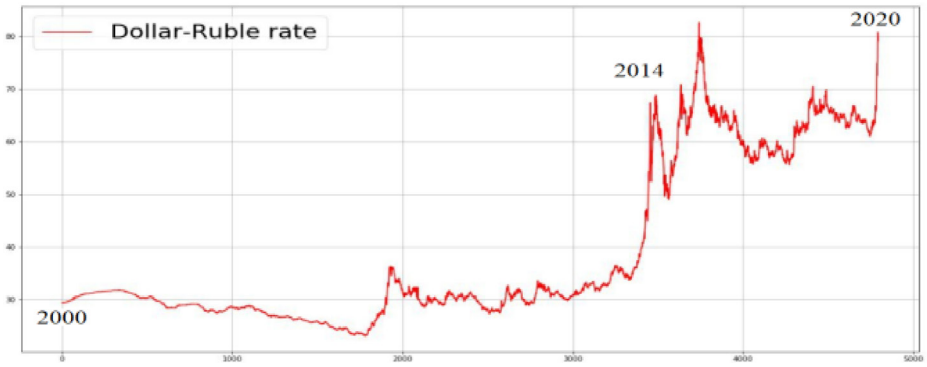


Рис. 1. Динамика USD/RUB

2. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

2.1. Стохастические дифференциальные уравнения

Стоит заметить, что обыкновенные дифференциальные уравнения имеют единственное решение для каждого своего начального условия, в то время как решения стохастических дифференциальных уравнений (СДУ) являются стохастическими процессами с непрерывным временем. Стохастические дифференциальные уравнения представляют собой мощную математическую основу для моделирования сложных явлений, на которые влияют как детерминированная динамика, так

и случайные флуктуации. Стохастические дифференциальные уравнения являются важными моделями при решении множества практических задач. СДУ находят применения в квантовой теории поля [2], гидродинамике [3, 4], стохастических моделях эволюции биологических или химических величин [5, 6], моделях процентных ставок [7, 8].

СДУ стали стандартными моделями финансовых величин, таких как цены активов, процентная ставка. Методы численного решения СДУ основаны схожей техникой решения обыкновенных дифференциальных уравнений, но включают элемент случайности, например стохастический процесс Винера, для обеспечения учета стохастичности процесса.

Приведем основные понятия стохастического исчисления, которые необходимы в работе. Множество случайных величин обозначим через X_t, где индекс t – действительные числа. Каждый пример или *реализация* стохастического процесса есть выбор из случайных величин X_t для каждого t, поэтому его можно считать функцией от t.

X_t называется *стохастическим процессом с непрерывным временем*.

Определение. Процесс Орнштейна — Улембека представим в виде стохастического дифференциала:

$$dX_t = aX_t dt + bdW_t$$

В интегральном виде:

$$X_t = X_0 + e^{at}\int_0^t be^{-as}dW_s.$$

Процесс широко применяется в непрерывных моделях динамики процентных ставок.

В 1977 году на основе модели Орнштейна — Улембека была предложена другая модель для короткосрочной процентной ставки.

Определение. Модель Васичека выглядит следующим образом:

$$dr = a(b-r)dt + \sigma dW_t, \quad (1)$$

где a — скорость возврата к среднему значению; b — средний уровень процентной ставки; σ — волатильность, не зависящая от значения ставки на текущий момент.

Каждый из этих параметров является константой, и их можно найти методом максимального правдоподобия. W_t — винеровский процесс.

Модель является однофакторной математической моделью, то есть зависит только от процентной ставки и может не учитывать влияние других параметров.

Решение уравнения (1) имеет вид:

$$r(t) = r_0 e^{-at} + b\left(1 - e^{-at}\right) + \sigma e^{-at} \int_0^t e^{as} \mathrm{d}\, W_s.$$

При $t \to +\infty$ можем найти:

$$M(r) = b,$$

$$D(r) = \frac{\sigma^2}{2a}.$$

Существенной особенностью данной модели является то, что здесь учтен фактор того, что процентная ставка, в отличие от цен, не может расти до бесконечности и колеблется вокруг какого-то среднего значения. Если ставка начнет расти до бесконечности, то это сильно повлияет на экономическую деятельность.

Для решения стохастического уравнения использован алгоритм численной аппроксимации. Реализация алгоритма осуществлена с помощью построения нейросети.

2.2. Модели NLP

Текстовый анализ основан на преобразовании текстовых данных в вид, удобный для выполнения с ним математических операций. Изначально NLP был сильно ограничен мощностями компьютеров и нехваткой эффективных архитектур, способных обрабатывать полученные векторные представления текстов. Бурное развитие NLP началось с 2010 г. Далее, в связи с огромным скачком компьютерных мощностей и широким распространением нейронных сетей, NLP начал широко применяться в задачах машинного перевода, сравнения текстов на схожесть, классификации текстов по темам и в многих других задачах [9, 10, 11, 12]. В данной работе рассматривается анализ текстовых данных с целью поиска оценки корреляции между рядом помеченных слов в новостных сообщениях и изменением курса доллара в последующие дни. Известно, что главный вклад в принятии решений на финансовых рынках всегда вносят люди, анализирующие информационный фон. В данной работе сделаны попытки построить модель, способную анализировать экономическую ситуацию с помощью использования современных алгоритмов машинного обучения и алгоритмов векторизации слов. Целью данной модели является классификация информационного фона посредством текстового анализа новостей для предсказания финансового временного ряда.

3. ОСНОВНЫЕ РЕЗУЛЬТАТЫ РЕШЕНИЯ ЗАДАЧИ

3.1. Построение модели Васичека

Постановка задачи. Пусть $\left(\Omega, \mathcal{F}, F\left(\mathcal{F}_n\right), \mathrm{P}\right)$ — фильтрованное вероятностное пространство. На этом вероятностном пространстве даны

стохастические последовательности: $\{S_n\}$, $n \geq 1$, где S_n — цена доллара в рублях в момент n. Необходимо, на основе $n = 1, \ldots, N$ этих показателей, построить модель для прогноза S на несколько шагов вперед.

Решение. Согласно исследованиям в области финансовой математики [1], для построения математических моделей необходимо использовать преобразованные данные:

$$r_n = \ln\left(S_n / S_{n-1}\right) \quad n \geq 1.$$

Для построения прогноза использована модель Васичека, которая выглядит следующим образом:

$$dr = a\left(b - r\right)dt + \sigma dW_t. \tag{2}$$

В данной задаче поставлена цель: подобрать значения параметров уравнения модели, построить модель и сделать прогноз на несколько шагов вперед.

Алгоритм подбора параметров модели описан ниже.

Шаг 1. На рисунке 2 представлен фрагмент данных. В первом столбце цены актива, во втором — приращение стоимости актива за один шаг: $r\left(t+1\right) - r\left(t\right)$.

Десятич. вид	r(t+1)-r(t)
0,0715	
0,0708	-0,0007
0,0688	-0,002
0,07	0,0012
0,0719	0,0019
0,0707	-0,0012
0,0698	-0,0009
0,0694	-0,0004
0,0701	0,0007
0,0706	0,0005
0,0707	0,0001

Рис. 2. Фрагмент данных

Шаг 2. Для нахождения оценок параметров a, b, σ модели Васичека вначале их значения выбираются произвольным образом (рис. 3).

a	0,02
b	0,3
σ	0,0002

Рис. 3. Значения параметров на первом шаге

Шаг 3. Используя коэффициенты *a*, *b*, рассчитываются $a(b-r(t))$. Результаты показаны на рисунке 4. Заметим, что выражение $a(b-r(t))$ является реверсией — показатель скорости возвращения к среднему значению в долгосрочной перспективе.

Десятич. вид	r(t+1)-r(t)	a(b-r(t))
0,0715		
0,0708	-0,0007	0,00457
0,0688	-0,002	0,004584
0,07	0,0012	0,004624
0,0719	0,0019	0,0046
0,0707	-0,0012	0,004562
0,0698	-0,0009	0,004586
0,0694	-0,0004	0,004604
0,0701	0,0007	0,004612
0,0706	0,0005	0,004598
0,0707	0,0001	0,004588

Рис. 4. Расчет $a(b - r(t))$

Шаг 4. Вычисляются ошибки: $\left(r\left(t+1\right)-r\left(t\right)\right)-\left(a\left(b-r\left(t\right)\right)\right)$ (рис. 5).

Десятич. вид	r(t+1)-r(t)	a(b-r(t))	(r(t+1)-r(t))-(a(b-r(t)))
0,0715			
0,0708	-0,0007	0,00457	-0,00527
0,0688	-0,002	0,004584	-0,006584
0,07	0,0012	0,004624	-0,003424
0,0719	0,0019	0,0046	-0,0027
0,0707	-0,0012	0,004562	-0,005762
0,0698	-0,0009	0,004586	-0,005486
0,0694	-0,0004	0,004604	-0,005004
0,0701	0,0007	0,004612	-0,003912
0,0706	0,0005	0,004598	-0,004098
0,0707	0,0001	0,004588	-0,004488

Рис. 5. Расчет ошибок модели

Шаг 5. Предполагая, что ошибки подчиняются нормальному закону распределения с математическим ожиданием, равным нулю, найдем плотность распределения σdf и $ln(\sigma df)$ и сумму всех получившихся чисел (рис. 6).

σdf	ln(σdf)		$\sum \ln(\sigma\, df)$
3,3858E-148	-339,5629953		-129319,3
9,3688E-233	-534,2649453		
4,52105E-61	-138,9489453		
5,30634E-37	-83,52674534		
1,1592E-177	-407,4097953		
8,2648E-161	-368,6041953		
2,3207E-133	-305,4019453		
1,6624E-80	-183,6985453		
1,35757E-88	-202,3217953		

Рис. 6. Вычисление интегральной суммы логарифмов

Шаг 6. С помощью метода максимального правдоподобия максимизируем интегральную сумма логарифмов и находим оценки параметров модели (рис. 7).

b	0,070822854
a	0,033888699
σ	0,001346681

Рис. 7. Оценки параметров модели Васичека

Подставляя найденные оценки параметров в модель (1), мы можем найти прогнозные значения процентных ставок.

Поставленную задачу можно решить численными методами с использованием аппроксимации Эйлера — Маруяма. Алгоритм можно реализовать с помощью нейросетей. В данной работе алгоритм Эйлера — Маруяма был реализован на Python, с использованием TensorFlow и Keras.

На рисунке 8 реализация стохастического процесса с помощью нейросети. Представлено прогнозное поведение стохастического процесса на несколько шагов вперед. Из рисунка видно, что поведение процентной ставки может пойти по любой траектории. Для того чтобы более точно определить поведение траектории в будущем, необходима дополнительная информация. Предлагается разбить все множество траекторий на несколько кластеров. Кластеризацию предлагается осуществить на основе NLP новостных сообщений.

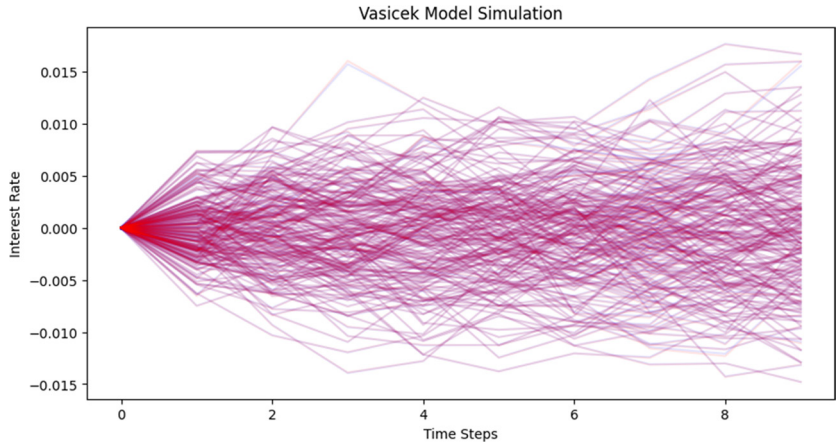


Рис. 8. Реализация стохастического процесса с помощью нейросети

3.2. Модель NLP для классификации информационного фона

В качестве корпуса новостных сообщений взят архив известий редакции lenta-ru, содержащий 800,000+ новостей и включающий данные: заголовок, текст новости, теги, дата. В итоге были отобраны данные о курсе доллар-рубль, включающие в себя информацию об изменении курса валюты по сравнению с прошлым днем.

Далее данные маркированы в зависимости от того, упало значение курса на следующий день или выросло. Маркер тренда 0 соответствует ситуации, когда курс упал, маркер 1 соответствует ситуации, когда курс поднялся. В итоге получен датасет, где каждому токенизированному тексту соответствует маркер 0 или 1.

Далее необходимо подобрать хорошее сочетание алгоритма векторизации и алгоритма машинного обучения. Эксперименты показали, что наилучшей является модель Word2Vec из python-библиотеки gensim.

На следующем шаге данные разбиты на обучающее и тестовое множество: 3600 образцов — для обучения, 946 — для теста. Для решения задачи выбрана нейросеть с архитектурой LSTM (long-short-term-memory). В качестве оптимизатора выбран Adam.

После обучения нейросети мы загружаем в нейросеть новостные сообщения и получаем на выходе номер кластера. Номер кластера указывает, какую реализацию траектории на рис. 8 использовать для прогноза процентной ставки.

ВЫВОДЫ

Эксперименты показали, что прогнозирование временных рядов процентных ставок можно осуществлять с помощью нейросетевых моделей новостных сообщений и нейросетевых моделей стохастических процессов. Каждая из моделей имеет свои недостатки, однако композиция

этих моделей может скорректировать ошибки каждой из них и дать более точное решение. Построение нейросетевых моделей новостных сообщений позволяет найти новые ситуации, которые невозможно обнаружить в числовых временных рядах процентных ставок.

Список использованных источников

1. *Ширяев А. Н.* Основы стохастической финансовой математики.
2. *Jean-Christophe Mourrat and Hendrik Weber*. Convergence of the two-dimensional dynamic. Ising-Kac model to Φ 4 2. *Communications on Pure and Applied Mathematics*, 4:717–812, 2017.
3. *Bjorn Birnir*. The Kolmogorov-Obukhov theory of turbulence: A mathematical theory of turbulence. SpringerBriefs in Mathematics. Springer, New York, 1st edition, 2013.
4. *Franco Flandoli and Eliseo Luongo*. Stochastic Partial Differential Equations in Fluid Mechanics. Lecture Notes in Mathematics Series; Volume 2330. Springer, Singapore, 1st edition, 2023.
5. *Gopinath B. Kallianpur and Jinbo Xiong*. Stochastic models of environmental pollution. *Advances in Applied Probability*, 26(2):377–403, 1994.
6. *Michael A. Kouritzin and Hongwei Long*. Convergence of Markov chain approximations to stochastic reaction-diffusion equations. *The Annals of Applied Probability*, 12(3):1039–1070, 2002.
7. *Damir Filipovic, Stefan Tappe, and Josef Teichmann*. Term structure models driven by Wiener ´ processes and Poisson measures: Existence and positivity. *SIAM Journal on Financial Mathematics*, 1(1):523–554, 2010.
8. *Philipp Harms, David Stefanovits, Josef Teichmann, and Mario V. Wüthrich*. Consistent recalibration of yield curve models. *Mathematical Finance*, 28(3):757–799, 2018.
9. *Benjamin Bengfort*. Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning. O'Reilly Media, Incorporated, 2019.
10. *Charu C*. Aggarwal "Neural Networks and Deep Learning", 2018.
11. *Tomas Mikolov*. Distributed Representations of Words and Phrases and their Compositionality, 2013.
12. Gabdrakhmanova N., Fedin V., Matsuta B., Pilgun M. The modeling of forecasting new situations in the dynamics of the economic system on the example of several financial indicators, Procedia Computer Science, Volume 186, 2021, Pages 512-520, ISSN 1877-0509. https://doi.org/10.1016/j.procs.2021.04.211; https://www.sciencedirect.com/science/article/pii/S1877050921010589

TEXT ANALYSIS AND NUMERICAL APPROXIMATION ALGORITHM OF STOCHASTIC DIFFERENTIAL EQUATIONS IN FINANCIAL MATHEMATICS PROBLEMS

***Gabdrakhmanova N. T.**, Associate Professor, Peoples' Friendship University of Russia, Moscow, Russia*

The article discusses one of the problems of financial mathematics, namely the problem of forecasting interest rates on securities. A hybrid model based on stochastic differential equations and text analysis of news reports has been developed. When constructing the model, an algorithm for approximating solutions of stochastic

differential equations (SDES) using neural networks is used. The analysis of forecasting results obtained by solving stochastic differential equations is carried out. It is shown that NLP methods of news reports make it possible to classify the economic situation in the future, and the composition of NLP and SDU models makes it possible to increase the accuracy of solving the forecasting problem. The adequacy analysis of the constructed hybrid model is based on interest rate data.

stochastic differential equation • neural network • interest rate • forecast, text analysis

American English & British English Main differences between the two most influencing linguistic varieties of the English language

David Garcia Rodriguez-Izquierdo, *junior research assistant, Department of French and English Philology, Research Group HUM-485, Foreign language teaching: materials for a new curriculum design, University of Cádiz, Spain*

Давид Гарсиа Родригес-Искьердо*, младший научный сотрудник, кафедра французской и английской филологии, исследовательская группа HUM-485v (Преподавание иностранных языков: материалы для разработки новой учебной программы), Университет Кадиса, Испания*

Статья посвящена изучению наиболее значительных различий между американским (AmE) и британским английским (BrE). Основное внимание уделяется различиям в лексике, произношении, правописании, грамматике и культурных особенностях. Рассматривается влияние исторических, социальных и лингвистических факторов на развитие обоих вариантов языка. Приведены примеры словарных и фонетических различий, а также обсуждается их влияние на взаимопонимание носителей AmE и BrE. Особое внимание уделено глобальному распространению американского английского и его роли как глобального языка.

американский английский • британский английский • лексика • произношение • грамматика • правописание • культурные различия

The article examines the most significant differences between American English (AmE) and British English (BrE). It focuses on lexical, pronunciation, spelling, and grammatical variations, as well as cultural distinctions. The paper explores the historical, social, and linguistic factors influencing the development of both varieties. Examples of lexical and phonetic differences are provided, highlighting their impact on mutual understanding between AmE and BrE speakers. Special attention is given to the global spread of American English and its role as a global language.

American English • British English • lexicon • pronunciation • grammar • spelling, cultural differences

1. INTRODUCTION

The purpose of this paper is not to mention or study all the differences between American English (AmE henceforward) versus British English (BrE henceforward), in full detail, but to

mention the most important features of AmE in contrast with BrE. These features are connected with lexical aspects, pronunciation, being one of the most noticeable differences to draw the speaker's attention, spelling, vocabulary and grammar.

A few people think that the major difference between AmE and BrE concerns spelling or vocabulary, but in fact that is not true. Indeed, these differences can be understood from the context. It is other differences, such as pronunciation or cultural concepts, that require more effort and time to be understood. I can illustrate this with a well-known example as follows:

There are many famous jokes that exploit the cultural use of words in both AmE and BrE. I like one of them, which I often tell my students in the classroom, whenever I refer to this topic. Indeed, I think this joke is really good, as it explains how AmE and BrE differ from each other. This is the joke I am referring to:

An American arrived at the beginning of the 90s at Victoria Station in London and met an Englishman. He waved to him and said: Hi! The Englishman looked at him and replied: Low!

Obviously, the American person meant *hi* as a greeting, not as an adverb or an adjective. However, British people did not understand *hi* as a way of greeting. So, in this case, the word *high* /hal/ is pronounced in the same way as *hi* /hal/. They are different words, morphologically talking, but they have the same pronunciation. This is so-called *homonymy in linguistics*.

In this case, we can see that this word is pronounced in the same way, but culture plays an important role too in this example, that is to say, in England and in the United Kingdom (UK henceforward) the most common greeting is *hello*, but in America it is *hi*. In contrast, the word *hi* is pronounced as *high*. Hence, the confusion is clear. This is one of a thousand examples I could mention, but this is one of my favourite ones. However, the greeting *hi* is really common in the UK nowadays as consequence of the influence of AmE in the English-speaking world.

On the other hand, the pronunciation of some words in both varieties is really difficult to understand. For instance, the word *schedule* is pronounced in a different way and the difference is really noticeable and outstanding: AmE /□sked□u:l/ and BrE /□□edju□l/. These variants in speech are really important and demand much more time, since the context does not always help.

Everybody knows that when an American and a British person meet, the most noticeable difference is the pronunciation of words like the word *schedule,* among many others. We can notice other differences having to do with: vocabulary, (that is to say, different words with same meaning or the same word with a complete different meaning), syntax, spelling and punctuation. However, there are many other differences which are equally important, such as cultural, political and educational ones, among others, but we are going to focus on the previous ones.

1.1. British English and American English as standard varieties

There are many varieties of English other than BrE (here the English of the United Kingdom) and AmE (here the English of the United States). All those other varieties are intrinsically just as worthy of study. However, these two varieties are the ones that are spoken by the greatest number of speakers of English and studied by the majority of foreign learners. They have a special status as the two main varieties of the language because of their influence in the English-speaking world.

But what are AmE and BrE exactly? What is standard English? (SE henceforward). Firstly, I will start talking about the second term.

The Cambridge International Dictionary of English (1995) defines *standard language* as: "The form of that language which is considered acceptable and correct by most educated users of it".

So, SE is to be understood as the language that is written and spoken by educated users as the correct form of the language in most types of public speech, on television, on the radio and in all types of media, above all, formal media. For example, in broadcasting, lectures and books.

However, there is no clear consensus on defining SE. For instance, Swann (2003) defines SE as follows:

> A prestige variety of the language; a variety with agreed norms and conventions that can be used as a model for education and for public use (i.e. in the media). The term may also refer to a variety that has been codified; that is, represented in dictionaries and grammars. This suggests that the forms of
>
> Standard English should show little if any variation. (Swann 2003:25)

Clearly, SE is connected with lexicon, orthography, grammar and discourse, but it is not really associated with pronunciation.

In addition to this definition, Yule defines SE as:

> [...] the variety which forms the basis of printed English in newspapers and books, which is used in the mass media and which is taught in schools. It is a variety we normally try to teach to those who want to learn English as a second language. It is really associated with education and broadcasting in public contexts and is more easily described in terms of the written language (i.e. vocabulary, spelling, grammar) than the spoken language. (Yule 1997:227)

1.2. The influence of American English as a global language

To understand better the influence of AmE as a global language, we must look back in time and remember some important facts, which took place when the British Empire ruled many countries in every continent. The British Empire spread over many parts of the world in the five continents. It expanded in the Pacific, China, India, Africa and Australia. It was during the reign of Queen of Victoria (1819-1901) that the British Empire

was at its peak of territorial expansion. In 1837, Victoria got the throne and became the Queen of the UK and Ireland. During her reign, the UK became an Empire, which lasted until the First World War.

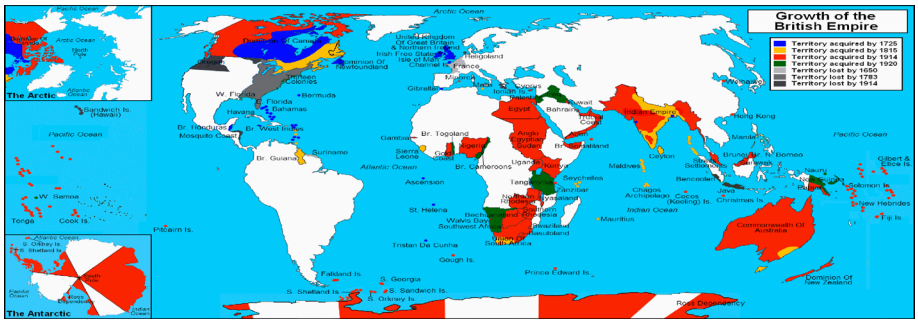


Figure 1. The British Empire from 1725 up to 1920 (http://www.atlasofbritempire.com/Commonwealth.php)

There are three main reasons for the expansion of British colonialism throughout the world:

1. The social, economic and demographic consequences of the Industrial Revolution

2. The nationalistic feeling based on the supremacy of the British race

3. The idea of cultural supremacy

By the end of the century, the British Empire comprised nearly a quarter of the world's land surface and more than a quarter of the world's population. However, the US started to emerge as one of the most powerful countries in the world, especially after the Second World War. It was a glorious era, which is best-known as *The Roaring Years* [1](1920-1945). It was an era marked by some discoveries and inventions of far-reaching consequences. Nevertheless, some important inventions such as the electric light system, the phonograph and motion pictures were created in the previous years in the US, that is to say, at the end of the 19th century and at the beginning of the 20th century.

It was also a time of amusement and leisure time: *the Jazz Age*. Everything was possible to be achieved thanks to these new technologies. The US industry mass-produced. There were changes in lifestyle, such as make-up for women, shaving razors for men and jeans. It was the time when Henry Ford[2] started mass-producing cars.

---

1 The Roaring Years were one of the most prosperous times in the history of the US. It was a time of economic growth, although the country suffered a critical period of economic crisis from the late of the 1920s until the end the 1930s. It was also a time of military power, a reference in fields such as technology, leisure time and culture in the whole world. The US became the reference for all the countries of the globe. It was in this era that words such as mass-produce were heard by first time.

2 Henry Ford (1863-1947) was an important American industrialist who was the founder and owned Ford Motor Company. Nowadays, Ford is one of the most well-known car makes in the US and in the world.

After WWII, the US became the most powerful country in the world and emerged as a military, industrial and political superpower. It was TV boom, leisure time, cinema and military power, among others, that made the US a symbol in the English-speaking countries and in the world. As a result, the US took a global role in politics, economy and military matters, and what is more, technological advances in the country fostered American influence in many parts of the world.

Crystal (2003b) explains that the main reason why AmE became a global language, in this case, a global variety, is not a question of linguistic or aesthetic criteria. This author confirms the connection that there is between linguistic supremacy and economic, technological and cultural power.

There are other reasons why AmE is the most widely spoken English dialect in the world. These are as follows.

1. Population: Nowadays, the US population comes to 300 million people. As a result of that, the US counts with at least seventy per cent of first-language speakers of English in the world. In other words, the number of AmE speakers of English is greater than in any other nation.

2. Publishing: Most of advertising and literature come from the US.

3. Mass media and technology media: The US has the most innovative technology in the world and its mass media has influenced almost the whole world.

4. American culture has crossed frontiers worldwide, above all, at the beginning of the 20th century, and has been the centre of the film industry and music in the world.

In conclusion, AmE became a source of change in the English of the UK and influenced other English-speaking countries. So, AmE had the privilege to head the English-speaking world as the most widely used and spoken in the world. I will deal with this point in the following paragraphs.

It is clear to everyone that both AmE and BrE are used worldwide and it is obvious the advantages these speakers have in comparison with other speakers of different languages or dialects. In other words, the English language has exerted power for many years and it still exerts it. In fact, around a quarter of the world's population speaks some variety of English, as we can see in figure 2. According to this information, I have to say that AmE has had much more influence than BrE in the English-speaking world due to the fact that it is a powerful nation and its language represents that. Moreover, the US has become the most powerful nation in the whole of the world as a consequence of its political, economic, technological and cultural strength. For instance, the language of big companies, the computer industry, air traffic control, the film industry, television and the scientific community is American English. We have clear examples such as Microsoft, IBM, Hollywood, CNN, NBC, among many others.

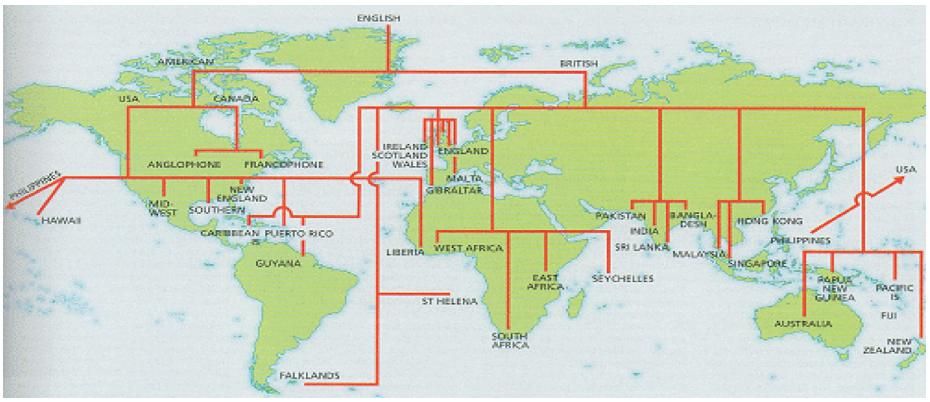


Figure 2 English spoken in the world (Crystal 2003: 107)

These reasons enable AmE to be called global language, since it is used around the world. However, there are other reasons why a global language is developed. For instance, an important factor is the people who speak it. In countries such Australia, South Africa, Nigeria, New Zealand and others English is learned as a native language, that is to say, L1[3]. English is L1 for almost 400 million people in the world. In other countries, it is used as a L2[4]. (See Quirk *et al.* 2000:128)

For example, in Spain some businessmen study English just because it is a tool for their jobs, that is to say, a tool to communicate and interact with other businessmen from different countries where Spanish is not spoken at all. Indeed, almost a billion people use English as a L2. Obviously, this stresses the fact that users rely on English for their specific aims. A study by David Graddol (2005) proves this and he suggests that the use of English as a L1 is decreasing. In fact, his study reveals that English is still privileged thanks to the L2 speakers. The author believes that Chinese will be the predominant language in the whole of Asia in the year 2050, as shown in the next table.

Table 1

Projected changes in global language communities: 1995-2050 (Graddol 2005)

Top ten global language communities of native speakers in 1995	Number of native speakers in millions	Top ten global language communities of native speakers aged 15-25 in 2050	Number of native speakers in millions
Chinese	1,113	Chinese	166
English	372	Hindi/Urdu	74
Hindu/Urdu	316	Arabic	72
Spanish	304	English	65
Arabic	201	Spanish	63
Portuguese	165	Portuguese	33

[3] L1 is used in Linguistics to refer to first language, native language, mother tongue, primary language and stronger language.

[4] L2 is a term in Linguistics that is used to mean second language, non-native language, foreign language, secondary language and weaker language.

Top ten global language communities of native speakers in 1995	Number of native speakers in millions	Top ten global language communities of native speakers aged 15-25 in 2050	Number of native speakers in millions
Russian	155	Bengali	32
Bengali	125	Russian	15
Japanese	123	Japanese	11
German	102	Malay	10

According to this table, the number of L1 English speakers will be decreasing little by little. In fact, estimates reveal that the number of native English speakers will decline from 316 million speakers in 1995 to only 65 million speakers by the year 2050.

In conclusion, AmE has grown up with the country. It began to diverge from BrE during its colonial beginnings, around 1600, and gained regional differences and ethnic flavor during the settlement of the continent. Indeed, the English colonization of North America resulted in the creation of a distinct American variety of the English language. Some expressions that the British call "Americanisms" are in fact original British expressions that were preserved in the colonies while lost for a time in Britain (for example *trash* for *rubbish*, *loan* as a verb instead of *lend*, and *fall* for *autumn*; another example, *frame-up*, was re-imported into Britain through Hollywood gangster movies). Spanish also has influenced on American English (and subsequently British English), with words like *canyon*, *ranch*, *stampede* and *vigilante*, among others. These are examples of Spanish words that entered English through the settlement of the American West. French words (through Louisiana) and West African words (through the slave trade) also influenced American English.

Today, American English is particularly the most powerful variety of English, due to its supremacy over the cinema industry, achievements in medicine, social media, television, popular music, trade and, above all, technology (including the Internet).

2. FEATURES OF AMERICAN ENGLISH

2.1. Lexical aspects of American English

Whenever we mention AmE, many of us think of the differences in the lexicon, pronunciation, grammar, among other linguistic aspects in comparison with BrE. As I have already mentioned in the introduction of this paper, an American greets differently as a Briton does. Americans normally use phrases and greetings such as *Hey! What's up, man!*, which are not usually used by Britons. This difference in the lexicon is due to the exposure of AmE to different languages existing in the US. In fact, when the first settlers arrived in the US words started suffering changes of meaning and this affected the vocabulary of AmE. For example, the word *barn* entered AmE in the British settlement era in North America. It meant the place where grains are stored in AmE, but along time this meaning changed. *Barn* in AmE means not only the place where grains are stored, but also where any other farm items are stored, as well as animals and machinery. This is called **broadening**, that is to say, when a word enters a language and gradually broadens its meaning.

On the other hand, **narrowing** is the antonym of broadening. For instance, the word *meat* was used to refer to food in general, but now it refers to only one type of food. Both broadening and narrowing affected AmE, but not BrE. For example, *corn* still means any type of grain in BrE, but its meaning has narrowed to mean a specific type in the US.

In addition, there is a word which is much used to explain the change in meaning in both varieties. That word is *bluff*. *Bluff* entered AmE around 1735, when the first settlers started complaining about how Americans used that word to mean a bank or cliff. Once it entered AmE, this word suffered a change in meaning. This is due to the use of the word to describe the American landscape better. (See American English: Dialects and Variation Walt Wolfram and Natalie Schilling-Estes).

Another example is the word *creek*, which meant originally "small saltwater inlet" in BrE, where it keeps that meaning today. However, in AmE it is different and is used to refer to "small river or stream", particularly freshwater stream.

It is really important to mention here that when the first immigrants settled in North America, they found many things and objects without names, that is to say, they found plants, animals, objects which they had not seen before and consequently there were no words for them in English. Many of the names were borrowed from other languages, such as Dutch, Spanish, German and French. For instance, in the first settlement periods many words were related to the local flora and fauna and food such as *raccoon*, *moose*, *hickory, hominy, skunk* or were words borrowed from native Americans such as *moccasin*, *totem*, *tomahawk*, *wigwam, terrapin*.

New waves of migration introduced new words from other languages such as, Spanish (*bonanza*, *rodeo, stampede, marijuana, burro, armadillo*, *bonito, coyote*), French (*leeve, saloon, pumpkin, chute, caribou)*, German (delicatessen, kindergarten, dumb). Moreover, some names of states of the US are examples of borrowings that entered AmE such as: Colorado, Florida and Nevada (from the Spanish words meaning "red", "land of flowers" and "snowy" respectively). There are two examples from French such as Louisiana (from King Louis) and Vermont (big mountain).

In relation to word formation, the AmE lexicon was affected by coinage, blending, compounding, clipping, derivation, conversion, acronyms and recutting. It is true that these changes affected BrE as well, even though most of these alterations in lexical features affected AmE mainly. For example, *seaboard*, *underbrush* and *backwoods* were compounds that were created in America. Other factors influenced AmE and as a consequence new words appeared. For instance, coinages were used to refer to cultural, social, technical and political changes in the country. So, the contact with native Americans, the culture of gambling, the railroad and the gold rush influenced the development of different processes of word formation, as a I already mentioned above. New words appeared such as: *bootleg, cowboy*, *steamboat, maverick, congress*.

Now, let's focus on the different process of word formation that I have just mentioned above. It's important to make clear that this formation process has influenced to higher extent to AmE, but also to BrE as well.

1. **Coinage**: It is also known as eponym. It is a process of involving a specific use of a product name to more general use. In other words, they are brands, which are used as generic names. For instance, *Kleenex*, *Xerox*, *Levis'*, *Kodak*, *Jello*, *Zipper* and *Hoover* are clear examples of coinage. These brands are used as generic names for different brands. There are coinages referring to proper names, geographical areas, mythology and literature as follows: *Boycott, cardigan, diesel, bikini, china, chimera* and *morphine*.

2. **Blending**: It consists of two parts of existing words to form a new one. For example, words such as *smog* (smoke/fog), *brunch* (breakfast/lunch), *sitcom* (situation/comedy), *motel* (motor/hotel) are combinations of two words.

3. **Compounding**: It is the mixture of two existing words to form a new one. The difference between blending and compounding, is that blending deals with parts of words instead of words. Examples of compounding are: *Postcard* (post+card), *bookcase* (book+case), honeysuckle (honey+suckle), *fatback* (fat+back: *Fatback* is the strip of fat from the upper part of a side of pork, usually dried and salt-cured).

4. **Clipping**: It is a word that is formed shortening existing words. For example, *gas* (gasoline), *hi-fi* (high fidelity), *'za* (pizza), *lab* (laboratory).

5. **Derivation**: It is a word that is formed of derivational morphemes. For example, *be-twitched* (betwich+ed) and *rearrangement* (rearrange+ment).

6. **Conversion**: It is a word that its grammatical category is extended to another. For instance, the modal verb *must* is used as a noun in sentences such as: *Attendance* is a must in this university. Other examples are: They *treed* (as a verb) a cat, they scored a *run* (as a noun).

7. **Acronym**: It is a word formed form the initial letters or sounds of several words. For example, *usb* (Universal Serial Bus), *MOT* (Ministry of Transport), VAT (value added tax), *UN* (United Nations), among many other examples.

8. **Recutting**: It is the reanalysis of component parts which are different from the original ones. For example, the word *workaholic* comes from *alcohol + ic* and the word an *apron* comes from a *napron* (napron is an out-of-date and more original form of apron).

2.2. Pronunciation

AmE pronunciation arouses considerable interest among linguists. In fact, it is one of the main branches of study. However, it is lexis by far that is the language point of most interest rather than the way words are pronounced. Obviously, this paper cannot present a deep analysis of the pronunciation of AmE, as there is no enough space for that. Therefore, I will mention some aspects in which AmE differs from BrE in pronunciation, above all, the pronunciation of /r/ and differences in stress, but not in full detail. In others words, I will explain how RP differs from General American[5] (GenAm

[5] General American (GenAm) is the main pattern of study in pronunciation in the US today.

henceforth). Both RP and GenAm are the two general patterns of pronunciation in the UK and the US respectively.

Many students wonder why AmE and BrE differ so much from each other in pronunciation. There are many factors to explain this noticeable contrast, but probably the settlement in the East of different peoples, most of them were English immigrants, who brought with them their languages, is one of the answers. As a consequence, new words entered AmE and these kept their original pronunciation.

Another reason is found in history. When the American Revolution took place, Americans wanted to abolish any symbols of Great Britain, that is to say, they did not want anything related to Britain, and the language, i.e. BrE, was one of the things to do away with. Noah Webster[6] was one of the linguists or politicians who wanted to change the language spoken in the US in the 19th century and create a new language completely different from the language the British brought to North America. In fact, Webster was one of the most important people to break with all BrE rules.

The word *schedule*[7] is one of those words, which differs so much in pronunciation from BrE as follows: AmE /□sked□u:l/ and BrE /ˈʃedju:l/. In fact, this word seems to keep the Old French pronunciation and as a consequence it kept its original pronunciation in BrE today.

Both *cafe* and *ballet* are other examples of French influence. However, it is not clear that the word *schedule*, using the British pronunciation, has been affected by French influence. Indeed, words older than seventeenth century normally fit into modern pronunciation and spelling models. This modern model is words that are spelt *s-c-h* and pronounced (s-sound + k sound), above all, if they come from Greek. For example, *school, scholar, scheme and schematic* follow this pattern of pronunciation.

Webster's Third New International Dictionary of the English Language, Unabridged (1971)

According to Webster, the different pronunciation of *schedule* lies on the fact that the pronunciation in AmE /ˈskedʒu:l/ maintains its Greek roots and was not influenced by French at all as happened in BrE /ˈʃedju:l/.

Webster's Third New International Dictionary of the English Language, Unabridged (1971)

---

6 Noah Webster was a lawyer, teacher and writer who published in 1828 the first American dictionary. He was really more famous for changes on American spelling.

7 According to Oxford Etymology Dictionary (OED), late 14c., sedule, cedule "ticket, label, slip of paper with writing on it," from O.Fr. cedule, from L.L. schedula "strip of paper," dim. of L. schida "one of the strips forming a papyrus sheet," from Gk. skhida "splinter," From stem of skhizein "to cleave, split" (see shed (v.) and cf. schism). The notion is of slips of paper attached to a document as an appendix (a sense maintained in U.S. tax forms). The specific meaning "printed timetable" is first recorded 1863 in railway use (the verb in this sense is from 1862). Modern spelling is 15c., in imitation of Latin; the modern British pronunciation /□□edju□l/ is from French influence, while the U.S. pronunciation /□sked□u:l/ is from the practice of Webster, and is based on the Greek original.

A. Rhoticity

Whenever students of English complained about how an American pronounces differently, the focus of attention is, above all, the pronunciation of /r/, especially in mid position and end position, such as *inform /ɪn'fɔ:rm/ and number /'nʌmbər/. Oxford Advanced Learner's Dictionary (2010)*

They were able to notice that /r/ is not pronounced in RP, unless it is followed by a vowel, but they realized that almost every American pronounces it. This is due to rhoticity, as I mentioned in the section 1.2 of this essay. Rhoticity, also known as postvocalic /r/, is the pronunciation /r/ whenever /r/ is featured in spelling (Chambers 1992, 23-28).

At the end of the 17th century, the migrants from the East of England, who settled in Boston and New York, brought with them an accent feature which it is kept today, that is to say, these settlers brought non-rhoticity accent. Meanwhile, other settlers brought rhoticity with them, above all, migrants from the West of England. Moreover, in the 18th century, Northern Irish and Scottish immigrants arrived in the US and they also brought with them non-rhotic features. This affected first Philadelphia and then moved south and westwards. This explains why most of Americans speak with rhoticity, since non-rhoticity accents has been found, above all, in the Boston area, in Massachusetts. The answer was given in the year 1776. Both American accents and British accents were largely rhotic. It was around this time that non-rhotic speech took off in southern and east of England, especially among the upper class. This "prestige" non-rhotic speech was standardized, and has been spreading in Britain ever since. Most American accents, however, remained rhotic. There are a few fascinating exceptions: Boston and New York accents became non-rhotic, as new migrants form the southern England settled in Boston and New York (Algeo 1993, 72-76).

In figure 3, I show in which areas non-rhotic accent is used:

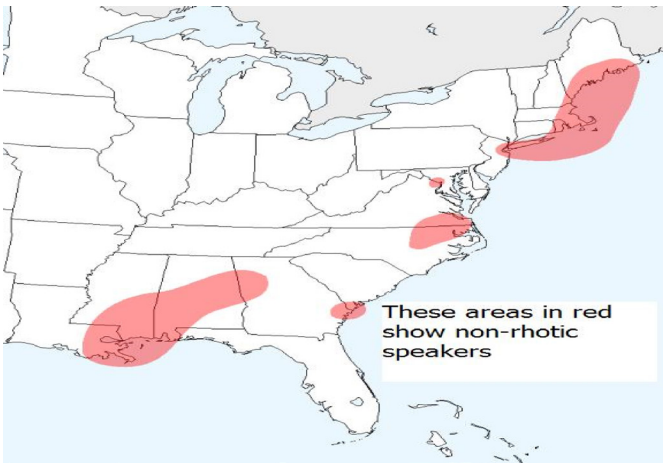


Figure 3. Non-rhotic areas in the US (Labov, Ash and Buberg 2006:48)

Consequently, the Americans /r/ is pronounced before vowels, consonants and at the end of words. For example, *lord /lɔ:rd /,arm /*ɑ:rm/ and *hear /*hɪr/.

B. Stress

Stress is to pronounce (a word or syllable) with greater force than other words in the same sentence or other syllables in the same word (Procter 1995:1438).

Both AmE and BrE are stress-timed dialects. Indeed, incorrect stress often results in misunderstanding many times. But what is a stress-timed language or dialect? Well, linguists call languages stress-timed when stresses tend to occur at regular intervals, with the result that the remaining unstressed syllables, no matter how many they are, have to put in between the stresses to fit the regular beat of the stress. For instance, English, Swedish, Russian and Arabic are examples of stress-timed languages.

On the other hand, there are other languages, such as Spanish, Italian, French and Portuguese, which are syllable-timed languages, that is to say, syllables tend to follow each other at regular intervals, with an equal amount of time being allocated for each syllable.

It is important to mention that the stress in English is not indicated in writing as happens in other languages, such as Spanish, French or Italian. It is only marked in dictionaries with an apostrophe or in transcriptions, except one syllable words. These are the physical properties of stress in English:

A. The intensity of the sound.

B. The length of the sound.

C. The pitch: the faster the cords vibrate, the higher the sound produced, like the notes on a guitar.

D. The "carrying power" of the sound.

AmE and BrE differ from the stress point of view too. In fact, an American puts the stress on different syllable in many words. Next, I will talk about the different positions of stress in both BrE and AmE.

1. Different position for primary stress

One of things we have to take into account in relation to stress is that it depends on the category of the word. All words have stressed syllables, but we do no words at all are marked for stress in writing. This is marked in dictionaries with an apostrophe.

The stressed syllables are in the words that give the main message: nouns, verbs, adjectives, adverbs. These are known as **content words**. On the other hand, there are words known as **function words**, i.e. conjunctions, auxiliary verbs, which are not stressed and serve to link content words, and these convey less relevant information in comparison with content

words. For example, in this sentence only content words are stressed. E.g. **Ka**ren has **bro**ken her **new gla**sses /kəˈren həz ˈbroʊkən (h)ər ˈnuː ˈglæsɪz/ (Bastida & Prieto 2008, 223-224).

In most stressed words, there is a syllable where the stress lies on noticeably. This is called primary stress. In both AmE and BrE, the position of primary stress is located in same position. However, there some exceptions as follows:

Table 2

Position of primary stress in some words in AmE (Wells 2008)

AmE	adult	debris	address	magazine
	əˈdʌlt	**dəˈbri**ː	**ˈædres**	**ˈmægəziːn**
BrE	adult	debris	address	magazine
	ˈædʌlt	**ˈdebri**ː	**əˈdres**	**ˌmægəˈziːn**

2. Different position for secondary stress

In some words containing three or more syllables, there are two stresses: one which is the main stress and the other which is the secondary stress. However, BrE never follows this rule and, in these cases, it keeps only the main stress and abolishes the secondary one.

In other words, some words in BrE have only one main stress, but in AmE have two: main and secondary, particularly when the primary stress is on the first or secondary syllable. As a consequence of that, we have different pronunciation in words such as: *Laboratory, secretary, cemetery, auditory, matrimony, antimony, accumulative*, among many others.

Table 3

Secondary stress in AmE vs BrE (Wells 2008)

AmE	laboratory	secretary	auditory	cemetery
	ˈlæbrəˌtɔːri	**ˈsekrəˌteri**	**ˈɔːdəˌtɔri**	**ˈseməˌteri**
BrE	laboratory	secretary	auditory	cemetery
	ləˈbɑrətri	**ˈsekrətri**	**ˈɔːdətri**	**ˈsemətri**

2.3. Spelling

The main spelling differences found in both AmE and BrE are found after The Revolution War of 1776. A number of Americans[8] wanted a national feeling and identity, i.e. sense of nationalism, to differ from the British at any level, that is to say, culture, spelling, grammar and lexicon. In 1828, Noah Webster (1758-1843) published *An American Dictionary of the English Language*, which was and still is the pattern to follow by all Americans. According to Webster, AmE had to be completely different from BrE. The aim of one of these changes in spelling, such us *color, defense, miter*, etc, was to create

[8] I am referring to a number of Americans who wanted a reform in spelling related to political idealism. For example, in 1768, Benjamin Franklin wrote "A Scheme and a Reformed Mode of Spelling". In 1854, Brigham Young supported a different system of alphabet, i.e. "Deseret Alphabet" consisting of the representation of the sounds of English with different graphic symbols. In 1934, "The Chicago Tribune" stated the intention to reform some spellings. This tendency lasted until 1955. As a result of it, some new spellings were accepted such as: hiccup instead of hiccough. (The American Language Mecken H.L. 1985).

the Americanization of English spelling (Crystal 1988, 145). He introduced some changes in spelling, grammar and lexicon, which were both accepted and successful. However, Webster introduced some spelling changes which were not ever accepted. For example, he wanted to change the spellings of *tongue, women, medicine* and *examine* into *tung, wimmin*, *medicin* and *examin*. What follows are the main spelling reforms suggested by Webster.

1. AmE *-or* / BrE *-our*

In BrE, nouns and verbs ending in *-our* are spelt *-or* in AmE. These changes also affected derivatives and inflected forms such as *favourite* and *savoury*. These are spelt *favorite* and *savory* in AmE. However, this difference did not affect to words with the suffix *-or*. In other words, there are words with the same spelling *-or* in both AmE and BrE used such as *governor, senator, ancestor, progenitor*, etc. In this case, both varieties spelled are alike, since it is used for a designation of a person.

Table 4

-or /-our group

-or / -our group	
AmE	BrE
color	colour
favorite	favourite
honor	honour

2. AmE *-se* / BrE *-ce*

Nouns ending with *-ce* in BrE are spelled *-se* in AmE. As a consequence of this difference, there is a pair of homographs such as *practise* (verb) / *practise / practice* (noun) and *license* (verb) / *license* (noun) in AmE. However, in BrE there is a clear difference between *practise* (verb) and *practice* (noun). However, both AmE and BrE preserve the distinction, both in spelling and pronunciation, between the noun and the verb in words such as *advice/advise* and *device/devise*. Apart from these examples, most of the words ending in *-ce* in BrE are spelt *-se* in AmE as shown in the next table.

Table 5

–se /-ce group

-se / -ce group	
AmE	BrE
defense	defence
license	licence
offense	offence

3. AmE *-er* / BrE *-re*

The origin of the words plays a key role in this group. This is important because depending on the origin of the word, words are spelt with *-er* or *–re*. In BrE, words coming from French or Greek end with *p* or *b* and an unstressed *–re* (Wolfram and Schilling 2001, 234-237). In AmE, the spelling

of these words changes to *–er* such as *luster / lustre*, *miter / mitre*, *fiber / fibre*. However, the suffix *–cre* is kept in both varieties such as *massacre, lucre, acre, mediocre*, etc.

On the other hand, BrE distinguishes between *metre* (the unit of measurement such *nanometre*) and *meter* (an instrument used for measuring, such as a *thermometer*, a *barometer*,etc.)

Table 6

–er / -re group

-er / -re group	
AmE	BrE
miter	mitre
theater	theatre
specter	spectre

4. Final *–l*

In BrE, *–l* is doubled before inflectional and derivational suffixes, such as *-ed, -ing, -er, -ery, -or, -ist*. AmE does not follow this rule and it is never doubled, as is shown in Table 7.

Table 7

–l group

-l / -ll group	
AmE	BrE
jeweler-jewelry	jeweller-jewllery
traveled-traveling	travelled-travelling
modeled-modeling	modelled-modelling

Nevertheless, we can find *-ll* in AmE only when the stress falls on the second or last syllable of the verb. In that case, double *-l* is compulsory. For example, propel /prə□pel/ has its stress on the second syllable and consequently *-ll* is spelt in this case: *propelled-propelling, propeller*, etc.

There some exceptions and the spelling of *-l* is the same in both dialects. For instance, *excel- excelled-excelling* is doubled in both varieties and *paralleled-paralleling* is spelt in the same way, because of the stress.

On the other hand, verbs ending in *-l* which are spelt with only one *-l* in BrE are spelt with a double *-l* in AmE. As a consequence, the inflectional and derivational suffixes are spelt in this way as well. We have to take into account that verb forms in BrE have double *-ll*. This is shown in the Table 8.

Table 8

Double –l in verb endings

verbs ending in -l / -ll	
AmE	BrE
enroll-enrollment	enrol-enrolment
fulfill-fulfillment	fulfil-fulfilment
install-installing	install-installing

5. *Ae* and *oe* group:

In BrE, borrowed words derived from Greek and Latin with *ae* and *oe* keep the same spelling. For example, *encyclopaedia, anaesthesia, diarrhoea*. However, this is simplified in AmE to *e* as shown in Table 9.

The case of *manoeuvre*, which is spelt *maneuver* in AmE, is an interesting exception, since *manoeuvre* seems not to derive from either Greek or Latin, but it was simplified in AmE to *maneuver*.

Table 9

ae / oe group

ae / oe group	
AmE	BrE
encyclopedia	encylycopaedia
medieval	mediaeval
airplane	aeroplane
anemia	anaemia

2.4. Vocabulary

Vocabulary or lexicon draws many linguists' attention. I will talk about three different levels of lexicon in this paper. Firstly, I will mention different words with the same meaning in both varieties. Secondly, I will talk about identical words with different meaning and finally words that do not have a full lexical equivalent in the other variety.

Many speakers of AmE are conscious of BrE words as *lift, tube, aubergine, sweet*, etc. Obviously, Americans never ever use them. Other words such as *driving licence* or *swimming costume* are not similar to them, but they are able to guess their meaning. However, words such as *semidetached house* or *naff* (lack of taste, style or quality) are weird and a little confusing to understand. On the contrary, BrE speakers are familiarized with AmE words such as, *candy, mall* and *gas*, as a consequence of the American dominance media. In addition to this, it is interesting to mention the phrase *two percent milk*, which completely lacks meaning in BrE. This American term is used to refer to *semi-skimmed milk*.

Level 1: Different words with the same meaning

A porter in a British hotel comes upon an American tourist impatiently jabbing at the button for the lift.
"Sir, the lift will be here in a moment."
"Lift? Lift?" replies the American. "Oh, you mean the elevator."
"No sir, here we call it a lift."
"Well, as it was invented in the United States, it's called an elevator."
"Yes sir, but as the language was invented here, it's called a lift." (Johnson 1996, 25).

This reference shows how much AmE and BrE differ in vocabulary. The term *elevator* is not used in BrE and the term *lift* is not used in AmE. However, there is feedback, as both the tourist and the porter understand different terms. This is

level 1, that is to say, both *lift* and *elevator* have the same referent. There is no lack of communication or understanding. The following table shows other common examples of this.

Table 10

Different words with the same meaning in AmE and BrE

AmE	BrE
elevator	lift
eggplant	aubergine
candy	sweet
truck	lorry
to stand in line	to queue
fall	autumn
store	shop
mall	shopping centre
wrench	spanner
to call collect	to make a reverse collect
bill	note
sidewalk	pavement
faucet	tap
one-way ticket	single ticket
basement	cellar
dead end	cul-de-sac
detour	diversion

Level 2: Identical words with different meanings

The following table shows words with the same spelling in both AmE and BrE but with different meaning. I have chosen only some of them, which were difficult for my students to understand the differences in both dialects.

Table 11

Same words in AmE and BrE with different meaning

Word	Meaning in BrE	Meaning in AmE
bathroom	a room with a toilet	a room where there is a toilet and sometimes a bath
sink	it is where you do the washing-up in	washbasin
fag	in slang language, it is a cigarette. *Faggot* is accepted in slang language, but not *fag*	homosexual, guy. It only refers to male
purse	a small bag used for keeping coins, cards and notes	handbag
vest	garment without sleeves under a shirt	garment without sleeves over a shirt
pants	underpants	trousers

Level 3: Words without a lexical counterpart in AmE and BrE

Culture and other factors cause words to develop different meanings. As a consequence, some words are related to culture and there is no lexical equivalent for it in the other variety. Table 12 shows some examples of this (Wolfram &Schilling-Estes 2001, 289-291).

Table 12

Words without a lexical counterpart in AmE and BrE

Word in BrE or AmE	Similar equivalent or meaning in the other variety
AmE busboy (also busser)	a person working in a restaurant whose job is to clear the table.
BrE lollipop man/ lady	a person who helps students to cross a busy road with a hand traffic sign.
AmE diner	a small and cheap restaurant.
AmE cookout	similar to a barbacue
AmE kindergarten	the first year of primary school in the USA
AmE freshman	a student in the first year at high school in the USA

2.5. Grammar

In 2003, a high school teacher in Maryland objected to an answer to one of the questions in the English Writing Skills section of Preliminary Scholastic Aptitude Test (PSAT), an academic assessment tool issued by Educational Testing Service (ETS).

In one of the questions, there were two possible correct answers, according to the teacher. The questions consisted of 4 answers and only one was supposed to be correct. However, the teacher did not agree with the only answer given by ETS[9] and stated that in that case there were two correct answers according to Standard American English grammar. As a consequence, many newspapers printed the question and the four answers. It drew many people's attention, such as teachers, graders, retirees, presidents of language societies, etc. Some agreed with the teacher and others did not. However, ETS decided to accept both answers and consequently around 500,000 tests were rescored.

This case reinforces the idea that AmE grammar accepts changes, which are accepted as correct. However, some changes, verb forms and others are accepted in AmE but are declined in BrE, as for example *gotten*.

Next, I will talk about the main grammar differences between the two varieties.

The improper use of a pronoun referring to a possessive led Kevin Keegan, a Maryland journalism teacher, to object to Educational Testing Service's scoring of the PSAT. ETS did not admit the mistake, but said it would re-score the test to accept two answers.

[9] Grammar question results in PSAT rescore (http://www.smartbrief.com/servlet/ArchiveServlet?issueid=67D95DE5-A037-4830-8161-195D53FC42AA&lmid=archives)

1. Use of the Present Perfect

AmE usually makes use of the Past Simple tense, where the Present Perfect is normally used in BrE.

This use takes place in two different situations: recent actions and with adverbs such as *ever, never, already, just* and *yet*, as shown in Table 13. In AmE, both the Past Simple and the Present Perfect tenses are accepted, but the first is certainly preferred (Bastida & Prieto 2008, 94-95).

Table 13

Past and perfect tenses in AmE and BrE

past simple / present perfect / yet, already, ever, never, already, just	
AmE	**BrE**
I am not able to find my mobile Did you see it?	I am not able to find my mobile Have you seen it?
Did you ever play basketball before?	Have you ever played basketball before?
I just had dinner / I've just had dinner	I've just had dinner
Did you do the washing-up yet? / Have you done the washing-up yet?	Have you done the washing-up yet?

2. Collective nouns and verb agreement

Some collective nouns such as *audience, class, club, committee, company, government, jury*, etc can be used with singular or plural verb (Swan 2002, 526-528).

In BrE, these nouns are used with singular verbs when the group is treated as a single entity and with plural verbs when it is treated as many individuals. However, AmE only accepts the singular verb.

Table 14

Collective nouns in AmE and BrE

AmE	BrE
The committee is deciding new policies on foreign matters	The committee is/are deciding new policies on foreign matters
Is the American team winning the game?	Is/are the American team winning the game?
The present government is deciding to raise VAT	The present government is/are deciding to raise VAT

3. *Have* or *take*

Usually, AmE uses *take* with a noun to express actions. On the contrary, BrE uses *have*. Both *have* and *take* are known as *delexical* verbs, i.e. they are often used in contexts where they have little meaning, but are used with a noun to express an action (Svejcer 1985, 35-37).

Table 15

Take and Have in AmE and BrE

AmE	BrE
I will take a shower after dinner	I will have a shower after dinner
I take a nap every day	I have a nap every day
I will take a drink after class	I will have a drink after class

4. Auxiliaries and modal verbs

Auxiliary and modal verbs, such as *do, needn't, don't need to, shall* and *have*, are used in a different way in AmE and BrE (Bastida & Prieto 2008, 95-96).

1. The auxiliary verb *do* is normally used in BrE to refer to an action mentioned already. AmE does not use *do* in this way. Moreover, *do* is used in BrE as an imperative form in order to express emphasis. These two uses are not common in AmE (Bastida & Prieto 2008, 95-96).

   BrE: Are you coming with us? I might *do*.

   AmE: Are you coming with thus? I might

2. In BrE, both *needn't* and *don't need to* are accepted. However, the first is much more used than the latter. On the contrary, in AmE is *needn't* is unusual and *don't have* is frequently used (Bastida & Prieto 2008, 96-97).

   BrE: You *needn't* finish your housework now or You *don't have/need to* finish your housework now.

   AmE: You *don't have to* finish your housework now.

3. In AmE, *shall* is not often used. In fact, it uses only *will* to express future. In contrast, *shall* was used in the both first singular and first plural person in BrE to express future.

   On the other hand, *shall* is used in BrE to ask for advice, suggest and offer to do something. In AmE, *shall* is never used in this way. *Shall* is replaced by *should* or *how about?* (Bastida & Prieto 2008, 97).

   BrE: Where *shall* I put the shopping? / *Shall* we go for a walk?

   AmE: Where *should* I put the shopping? / How about a walk?

4. Negatives question with *have*. BrE accepts both *haven't got* and *don't have*. However, the latter is not frequently used in everyday English. However, AmE prefers "don't have" instead (Bastida & Prieto 2008, 97).

   BrE: He *hasn't got* two children / *Has* she *got* your keys?

   AmE: He *doesn't jave two children* / Does she have your keys?

5. *Got* and *gotten*

Gotten is only used in AmE. It is also used to express "to become", but only in AmE, never in BrE. In AmE, *have gotten* is only used to express the result of an action (Bastida & Prieto 2008, 98).

BrE: He's got much better at playing tennis / She has *got* fatter this year

AmE: He has *gotten* much better at playing tennis / She *has gotten* fatter this year

6. Past tense forms of regular verbs

In SE, the past tenses of some verbs are regular and they take *-ed* endings. For instance, BrE prefers *-t* endings and AmE *-ed* endings. This happens in the verbs shown in Table 16. It is important to mention that the verbs *burn, dream* and *spoil* take both endings to form past tenses. In BrE, the verb *lean* can be formed with either regular or irregular ending (Bastida & Prieto 2008, 99).

Table 16

Past tense forms of regular and irregular verbs

Infinitive	*simple past (AmE)*	*simple past (BrE)*	*past participle (AmE)*	*past participle (BrE)*
burn	burned/burnt	burned/burnt	burned/burnt	burned/burnt
dream	dreamed/ dreamt	dreamed/dreamt	dreamed/ dreamt	dreamed/dreamt
lean	leaned	leaned/leant	leaned	leaned/leant
learn	learned	learned/learnt	learned	learned/learnt
smell	smelled	smelled/smelt	smelled	smelled/smelt
spill	spilled	spilled/spilt	spilled	spilled/spilt
spoil	spoiled/spoilt	spoiled/spoilt	spoiled/spoilt	spoiled/spoilt

7. Questions tags

Both AmE and BrE use question tags at the end of a statement in order to turn it into a question or check that the statement is correct. However, it is really more common to see question tags in BrE than in AmE. AmE uses expressions such as *right* or *ok* instead (Bastida & Prieto 2008, 100).

BrE: You are coming with us, *aren't you*? / You ate the pie, *didn't you*?

AmE: You are coming with us, *right*/*ok*? / You ate the pie, *right*?

8. Prepositions

There are some clear differences in the use of prepositions in AmE and BrE. One of the most noticeable ones is used to tell time, but there are other examples that are shown below.

1. Prepositions to tell time

BrE: It's ten to five / It's half past nine / It's twenty past ten

AmE: It's ten till five / It's half nine / It's twenty after ten

2. Omission of *to* / use of *with*

BrE: I think I will write *to* Kathryn / I never talk *to* her / I often *meet* him at the shopping centre

AmE: I think I will *write* Kathryn / I never talk *with* her / I often meet *with* her at the mall

3. *Through*

BrE: The meeting will last *from* Monday *to* Friday

AmE: The meeting will last *from* Monday *through* Friday

4. Different use of the prepositions of time

BrE: We are planning to fly to Toronto *at the weekend* / I often go cycling *at weekends*

AmE: We are planning to fly to Toronto *on the weekend* / I often go cycling *on weekends*

5. In hospital or in the hospital

In BrE, there is a significant difference between to *be in hospital* and to *be in the hospital, and* to *be at university* and *to be at the university*. The meaning of the article is subtly different. The article shows that a person is just a visitor. On the contrary, the lack of article means that the person is a patient, who is having an operation or a test. However, this difference is not found in AmE.

Consequently, neither *hospital* and *university* are used without articles in AmE. On the other hand, AmE sometimes uses *at the hospital* to specify that someone is visiting a patient, rather than being the patient itself. Note that this phrase is also accepted in BrE (Amberg & Vause 2004, 79).

BrE: I met your brother *in the hospital* / I am still *in hospital* / I met your brother *at the hospital*

AmE: I met your brother *in the hospital* / I am still *in the hospital* / I met your brother *at the hospital*.

Conclusion:

Taking into account that this paper is only a brief summary of both linguistic varieties' features and some other aspects concerning the difference between both, we must into account that even AmE and BrE are the same language, they differ in grammar, lexical aspects, pronunciation, spelling, and of course, others aspects such as politics, education, etc., though these are more common and logical, as we are talking about two different countries. Indeed, it's clear that between speakers of AmE and BrE, the

speech may be misunderstood, above all, in some specific contexts, where the context is the only way to "unravel the enigma". For instance, the counterparts *bill/note.* The first is used in a restaurant context in BrE meaning *check* in AmE, but in AmE bill is currency, that is to say, note. In other words, as I have mentioned in this paper, both dialects use different words for the same concept.

On the other hand, and more particularly, AmE and BrE contrasts are worldwide recognized. For instance, the pronunciation models differ from each other noticeably, British Received Pronunciation versus General American. Indeed, the accents are absolutely different, so it may lead to difficulty to mutual comprehension at times.

To conclude, both linguistic varieties, the two most influencing ones at the moment, are two variants of the English language, with many similarities to get understood between both speakers; however, some important differences spring up, which make the speakers not to get what is being uttered, explained or said, enriching both culture and the language itself.

References:

1. Algeo, John & Pyles, Thomas 1993. *The origins and development of the English language*. Wadsworth: Harcourt Brace Jovanovich College Publishers.
2. Amberg, Julie S & Vause, Deborah J. 2004. *American English: History, Structure and Usage.* Pennsylvania: Cambridge University Press.
3. Bastida Rodríguez, Patricia & Prieto Arranz, José Igor 2008. *On Cultural Diversity: Britain and North America*. Palma de Mallorca: Ediciones UIB (Universidad de las Islas Baleares).
4. Castañeda, Perry 2001. Maps collection. http://www.lib.utexas.edu/maps/europe.html
5. Chambers, J.K. 1992. "Dialect Acquisition" Language 68(4): 673-705
6. Crystal, David 1998. The Cambridge Encyclopedia of English. London: Cambridge University Press.
7. Crystal, David 2003. English Spoken in the World. Holyhead: Cambridge University Press.
8. G. & C. Merriam Co. 1971. *Webster's Third New International Dictionary of the English Language*. Springfield, Mass: Merriam Webster.
9. Graddol, David 2005 "The Future of Language". *Science* 303:1329-1331.
10. Johnson, Samuel 1996. 'American English vs British English'. *The Reader's Digest* 25:2.
11. Labov, Ash and Buberg. The Atlas of North American English Phonetics, Phonology and Sound Change. AtlasofNorthAmericanEnglish_WdG_PartA1470967896980.pdf
12. Oxford Advanced Learner's Dictionary 8th Edition. 2010. Oxford: Oxford University Press.

13. Procter, Paul 1995. *Cambridge International Dictionary of English*. Cambridge: Cambridge University Press.

14. Svejcer, Aleksandr 1985. *Standard English in the United States and England*. Volume 159 in the series Janua Linguarum. Series Minor https://doi.org/10.1515/9783110814484

15. Swan, Michael 2008. *Practical English Usage*. Oxford: Oxford University Press.

16. Swann, Joan 2003. English voices. In D. Graddol, D. Leith and J. Swann 2003. *English: History, Diversity and Change*. London and New York: Routledge.

17. Trudgill, Peter & Hannah, Jean 1994. *International English. A Guide to the Varieties of Standard English*. 3rd ed. London: Edward Arnold.

18. Trudgill, Peter and Jean, Hannah 2002. *International English*. 4th ed. Oxford University Press: Edward Arnold.

19. Warfel, Harry 1966 2nd Edition. *Letters of Noah Webster*. New York: Literary Publishers.

20. Wells, John 2008. *Pronunciation Dictionary* 3rd ed. London: Longman.

21. Wolfram, Walt & Schilling-Estes, Natalie 2001. *American English: Dialects and Variation*. Raleigh, NC: Blackwell.

22. Yule, George 1997 (1985). *The Study of Language*. Cambridge: Cambridge University Press.

Перспективы использования анализатора RuLingva для выявления сложности публицистического текста

***Иванова Т. К.**, доцент, доктор филологических наук, профессор, Казанский (Приволжский) федеральный университет, Tatiana.ivanova@kpfu.ru*

В статье описываются возможности объективации сложности публицистического текста как фактора, влияющего на понимание идейного содержания, а также возможности применения данной программы применительно к текстам медиапространства. Фактор сложности текста может быть использован для повышения адекватности и эффективности восприятия информации, а также на степень заинтересованности читателей в информации.

сложность текста • речевое воздействие • параметры сложности • индекс Флеш — Кинкейда • программа RuLingva

Наша повседневность сегодня напрямую связана с информационным пространством: успешность и рискогенность человеческой деятельности находится в прямой зависимости от скорости получения, обработки и владения информации. Среднестатистический человек очень часто взаимодействует с печатным словом, проверяя почтовый ящик в браузере или листая новостную ленту. Неслучайно по силе воздействия на общественное мнение медиасфера прочно занимает лидирующие позиции. И это воздействие, а также передача, хранение и транслирование информации осуществляется через текст.

Особенности медиатекса достаточно часто находились и находятся в фокусе внимания ученых-лингвистов как в России, так и за рубежом (Т.Г. Добросклонская, Н.Г. Нестерова, В.В. Богуславская, Т.В. Шмелева, Л.Р. Дускаева, А.Г. Пастухов, Алан Бел, Мартин Монтгомери, Норман Фейрклаф, Кристин Финк и др.). Однако большинство исследователей сходятся во мнении, что медиатексту свойственны такие функции, как информирование и воздействие через убеждение или манипулирование общественным сознанием (Дзялошинский И.М., Пильгун М.А., 2011, с. 82–113).

С возрастанием роли социальных сетей в медиасфере появляются новые жанры и типы текстов, как, например, тексты блогов, которые также имеют свои особенности, язык и адресатов)Bushkanets et al., 2021, 2022; Ivanova et al., 2022). Нередко такие тексты обладают так называемым индексом дочитывания, который, в свою очередь, зависит от множества факторов. При этом особое значение приобретает не только сама информация, но и формирование оценки различных явлений действительности как средства воздействия на общественное мнение: «медиатекст рассматривается как инструмент информационного воздействия, регулирующий его восприятие членами языкового сообщества» (Шпильная, Н.Н., 2020. С. 23).

Однако некоторые тексты легко воспринимаются и читаются, другие трудны для восприятия адресатом. Нахождение исследователем объективных средств измерения подобных трудностей является поэтому первоочередной задачей исследователя. Один из подобных параметров был сформулирован как фактор сложности текста (Солнышкина М.И., Кисельников А.С., 2015).

Как свидетельствуют многочисленные исследования, сложность понимания многих текстов слишком высока для значительной части населения (Conrad & Schober, 2007; Graesser, Cai, Louwerse, & Daniels, 2006). Это объясняется наличием в текстах сложных слов, многозначных конструкций, сложным синтаксисом или общей структурой построения логики повествования. Все это сказывается на когнитивных механизмах человеческого восприятия, и, как следствие, текст сообщения может быть не понят адресатом (Казачкова и др., 2021).

История изучения сложности текста началась более ста лет назад (см. Рубакин (1898), Чолдин М.Т. (1979)), но до начала XXI в. исследования были изолированными и очень ограниченными вычислительными инструментами, а также отсутствовали соответствующие базы данных (например, отсутствие корпусов). В настоящее время наблюдается всплеск исследований, связанных с наличием современных вычислительных инструментов и методик, находящихся в открытом доступе, таких как seo-analyses, Coh-Metrix, Core of Text Ease Readability Assessor (TERA); RuLingva; Text Inspector, Reader Bench и т.д.

В 2020 году Н.А. Левковская опубликовала статью, посвященную проблеме интерпретации политического текста. В статье речь шла об опосредованном влиянии политической речи через средства массовой информации на аудиторию. Исследователь утверждала, что любой медиатекст ориентирован на адресата, поэтому в медийном пространстве автор и адресат осуществляют постоянную коммуникацию через своеобразный диалогический монолог. В работе отмечалось существование трех моделей коммуникации: 1) информационно-когнитивной, 2) эмоционально-экспрессивной и 3) риторической и декларативной (Левковская Н.А., 2020). Каждая модель выполняет свою функцию, которая призвана повысить результативность воздействия на аудиторию. Что же текст делает более понятным или легким для восприятия читателем?

Как было установлено рядом исследователей, сложность текста, влияющая на процесс его понимания, определяется как качественными (сюжет, замысел, цель текста и др.), так и количественными параметрами (длина слов, грамматические структуры, частотность встречаемых словоформ, терминология и т.д.). При этом большую роль в понимании текста играют также значимость темы для читателя, его мотивированность и заинтересованность (Fisher et al., 2012). Так, к количественным параметрам сложности текста чаще всего относят четыре параметра: лексическую плотность; лексическое разнообразие (TTR); индекс читабельности (FKGL) и количество юридических терминов.

За последние десять — пятнадцать лет в лингвистике был проведен ряд исследований, посвященных понятию сложности текста и возможности

использования этого критерия в работах, связанных с переводческой деятельностью, медиапространством.

В 2017 году М.И. Солнышкина рассмотрела параметры сложности на примере англоязычных текстов по подготовке к сдаче ЕГЭ по учебнику «В центре внимания 11» (2009). Для анализа данных использовался автоматизированный текстовый процессор Coh-Metrix Common Core Text Ease Readability Assessor (TERA) с использованием индекса Флеш — Кинкейда, показавшего, что тексты данного учебника обнаруживают тенденцию к повышению уровня сложности текста для мотивации обучаемых и выполнения педагогических задач (Солнышкина, 2017).

Сложность текста является неотъемлемым свойством текста и одновременно результатом осознанного выбора со стороны говорящего или автора тех или иных лингвистических средств. С другой стороны, сложность текста всегда связана со слушателем или с читателем, и на нее легко могут повлиять их опыт и знания. В то время как сложность текста может колебаться в зависимости от аудитории или читателей, сложность текста относительно стабильна. Таким образом, исследование показало, что различия между переводами объясняются целями и методами перевода, а также вниманием переводчика к присущей ему неоднородности исходного текста.

Как уже говорилось выше, для подобного рода исследований сегодня используются широкие технические возможности. Платформа-агрегатор RuLingva, например, предназначена для учителей, исследователей, занимающихся проблемами текста, а также широкого круга пользователей, интересующихся проблемами современной филологии. RuLingva помогает оценить сложность текстов и проверить ее соответствие уникальному уровню читательских навыков. Данная программа была использована в рамках деятельности научной лаборатории «Текстовая аналитика» Казанского федерального университета. С помощью RuLingva проводятся измерения достижений обучающихся в процессе обучения, позволяющих регистрировать процесс восприятия и понимания текста, а также процесс составления текстов для контроля достижений, так как RuLingva производит автоматизированное оценивание словарного запаса и читательского уровня. Таким образом, на базе данной программы становится возможным подобрать текст соответствующего уровня сложности для различной целевой аудитории.

В завершение хотелось бы отметить, что параметр сложности текста как объективный показатель может применяться и в медиалингвистике как для оценки качества подготовленных к публикации сообщений, так и для моделирования и улучшения качества самих публикаций. Кроме того, объективность данного критерия позволяет применять его как к текстам на родном (русском) языке, так и на языке перевода (например, английском). Возможности применения данного параметра к различным видам публицистических текстов еще не исследованы и открывают новую страницу в исследованиях текстов медиадискурса.

Список использованных источников

1. *Дзялошинский И. М., Пильгун М. А.* Медиатекст: особенности создания и функционирования. — М.: НИУ-ВШЭ, 2011. — 377 с.
2. *Казачкова М. Б., Галимова Х. Н.* Лексическое разнообразие текста как параметр сложности текстов // Вестник Марийского государственного университета. — 2021. — Т. 15. — № 3. — С. 384–390. https://doi.org/10.30914/2072-6783-2021-15-3-384-390

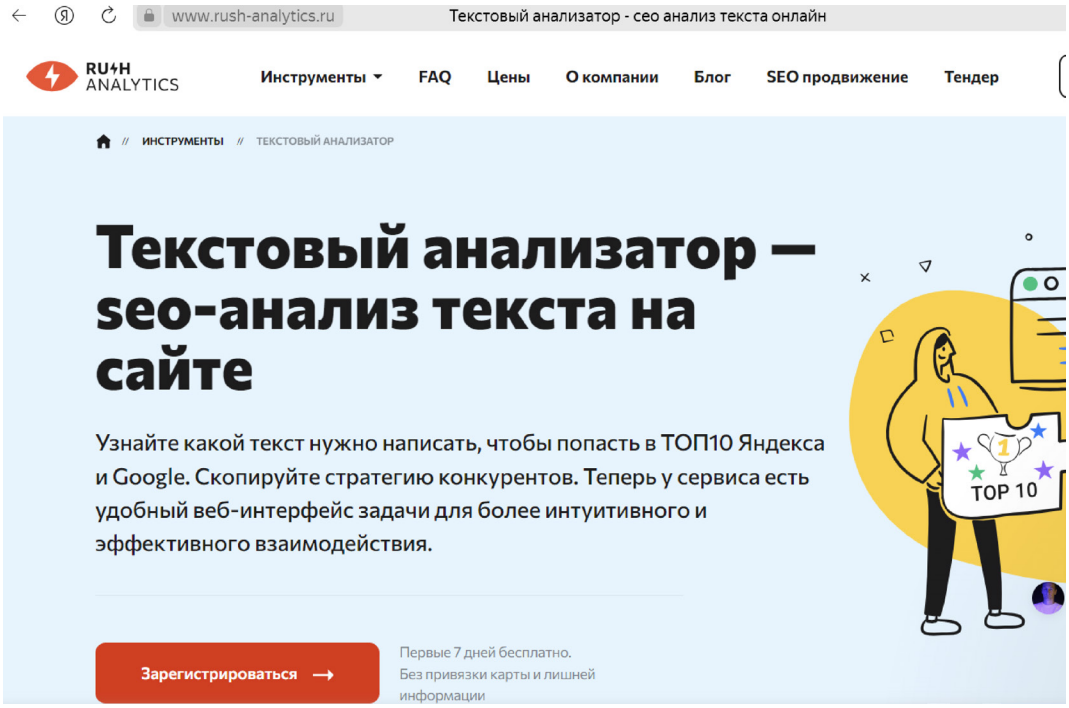
3. *Солнышкина М. И., Кисельников А. С.* Сложность текста: этапы изучения в отечественном прикладном языкознании // Вестник Томского государственного университета. Филология. — 2015. — № 6 (38). — С. 86–99.
4. *Солнышкина М. И, Кисельников А. С., Габитов А. И.* Сравнительная оценка сложности текста // Информационные технологии в гуманитарных науках. — 2017. — С. 118–119.
5. *Шпильная, Н. Н.* Модели интерпретации политического медиасобытия в профессиональном и обыденном медиадискурсе / Н. Н. Шпильная // Политическая лингвистика. — 2020. — № 5 (83). — С. 22–28.
6. *Bushkanets, L. E. Bushkanets, O. S. Zubkova, E. A.* Zvereva and Anastasia Fakhrutdinova (2021). Tolerant Language Personality as the Basis of Communicative Interaction // Entrelínguas Journal. Vol. 8 No. esp.4 (2021): (2021) v. 7, n. esp. 4.
7. *Bushkanets, L. E., Borisovich, S. L., Khairutdinov, R. R., Nurgalievich, M. R.* (2022). Situational Centers of the Republic of Tatarstan and Analysis of Social Networks (Instagram)(Article) Res Militaris Volume 12, Issue 3, 2022, Pages 859–868.
8. *Ivanova, T. K., Arzhantseva, N. V., Shaikhaidarovna, A. R., Borisovich, S. L.* (2022). Blogs of Civil Servants of the Republic of Tatarstan in Social Networks(Article) Res Militaris Volume 12, Issue 3, November 2022, Pages 75–87.
9. *Levkovskaya N. A.* Diological Aspect of political Discourse. Linguistics & Polyglot Studies. 2020; 21(1):28-33. (In Russ.). https://doi.org/10.24833/2410-2423-2020-1-21-28-33
10. *Fisher D., Frey N., Lapp D.* (2012). Text Complexity: Raising Rigor in Reading. International Reading Association. 212 p // Joanna L. Anglin Rockdale Career Academy, Conyers, GA Journal of Language and Literacy Education Vol. 10 Issue 2, 143–145 pp. — Fall 2014 Michelle M. Falter, Editor. http://jolle.coe.uga.edu
11. *Solnyshkina M. I, Vishnyakova O. D.*, English textbooks for Russian students: Problems and specific features//Journal of Social Studies Education Research. — 2017. — Vol. 8, Is. 3. — P. 215–226.

PROSPECTS FOR USING THE RULINGUA ANALYZER TO IDENTIFY THE COMPLEXITY OF A JOURNALISTIC TEXT

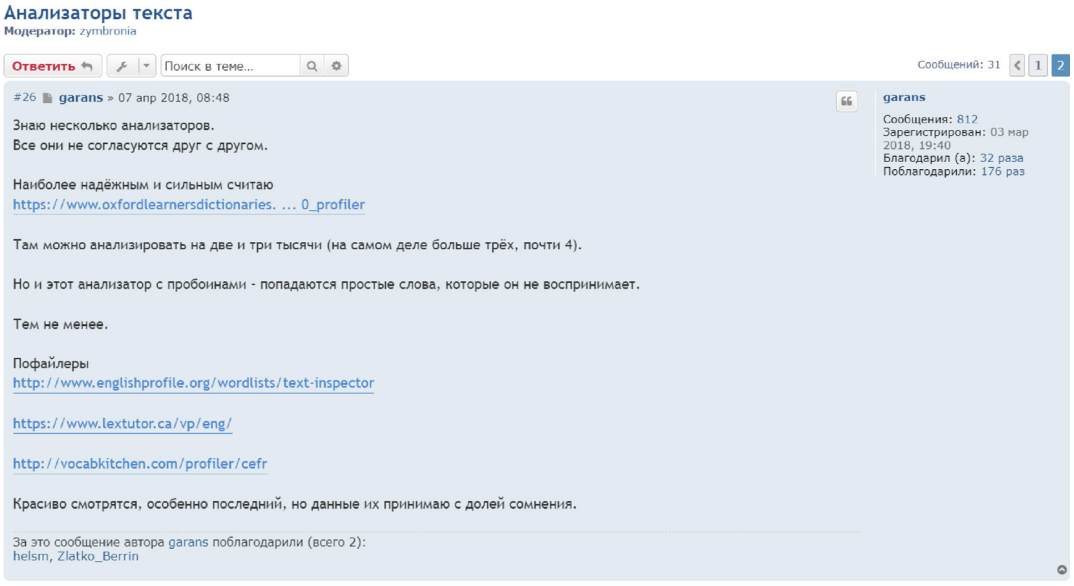
***Ivanova T. K.**, Associate Professor, PhD, Professor, Kazan (Volga Region) Federal University, Tatiana.ivanova@kpfu.ru*

The article describes the possibilities of objectifying the complexity of a journalistic text as a factor. The complexity can affect the understanding of the ideological content, as well as the possibilities of applying this program in relation to media texts. The complexity factor of the text can be used to increase the adequacy and effectiveness of information perception, as well as the degree of readers' interest in information.

text complexity • speech impact • text complexity parameters • Flash-Kincaid-index • RuLingua analyzer



https://rvb.ru/soft/catalogue/catalogue.html год 2001, обновление 2024
https://neuro-texter.ru/?yclid=17531462777230065663
https://aiwiz.ru
https://smm-trends.ru



Алгоритмы применения больших языковых моделей в образовательном процессе и деловом общении

***Комарова Е. В.**, кандидат филологических наук, доцент кафедры английского языка № 8, Московский государственный институт международных отношений (университет) МИД России, elvakom@rambler.ru*

В настоящий момент в современных технологиях большую роль играют системы моделирования языка (LLM). Эти технологии не только меняют способы взаимодействия людей, но и открывают новые перспективы для развития множества индустрий. В настоящей статье рассматриваются виды воздействия LLM на разные сферы деятельности человека, в частности образование и деловую коммуникацию. Приводится анализ языковых особенностей эффективных промтов для решения профессиональных и технических задач, а также изучаются возможности применения LLM в образовательном процессе.

LLM • большие языковые модели • промт • образовательный процесс, деловое общение

Технологический прогресс неизбежно приводит к социальным переменам, так как новые технологии влияют на все аспекты жизни общества, включая экономику, культуру, образование, науку, способы и формы социального взаимодействия. Технологические изменения влияют на производительность труда и конкурентоспособность бизнеса [3, 11]. Доступ к информации, образовательным платформам и онлайн курсам способствует росту уровня грамотности и образования в целом [4, 8]. Отсутствие доступа к современным технологиям приводит к усугублению существующих различий в обществе и росту неравенства [10].

Очевидно, что наличие доступа и адаптация человека к цифровой среде и современным технологиям становятся ключевыми факторами, определяющими социально-экономическое развитие общества.

Среди современных технологий особое место занимают технологии моделирования языка, которые лежат в основе множества систем и приложений. В результате использования больших данных и метода глубокого обучения LLM эффективно справляется с такими задачами, как генерация текста, перевод, реферирование, ответы на вопросы и сентимент анализ [16]. Системы перевода, основанные на моделях языка, становятся все более точными и контекстуальными. Использование таких систем для делового общения способствует повышению

эффективности международного сотрудничества. Языковые модели делают информацию более доступной: они обрабатывают огромные объемы данных, извлекая из них определенную информацию и представляя ее в удобном формате. Это позволяет оперативно находить необходимые данные и принимать более обоснованные решения. Кроме того, технологии моделирования языка позволяют автоматизировать множество рутинных задач, таких как обработка запросов клиентов, создание отчетов и анализ данных [1].

Системы генеративного искусственного интеллекта на основе моделей языка открывают новые возможности для создания текста [9]. Журналисты, писатели и маркетологи могут использовать эти технологии для генерации идей, написания статей или создания рекламных материалов. Появляются новые формы искусства и развлечений, такие как интерактивные истории или игры, где текст генерируется в реальном времени в зависимости от действий пользователя [5].

В образовательной сфере такие технологии могут быть использованы для создания адаптивных учебных материалов, которые подстраиваются под уровень знаний учащихся, автоматизированного тестирования, оценки письменных работ. Такие возможности отвечают сегодняшнему запросу на индивидуализацию образования и позволяют повысить академическую успеваемость обучающихся [2, 6].

К настоящему моменту системы языкового моделирования прошли четыре значительных этапа развития, включая: статистические языковые модели, нейронные языковые модели, предварительно обученные языковые модели и большие языковые модели (рис. 1) [14].

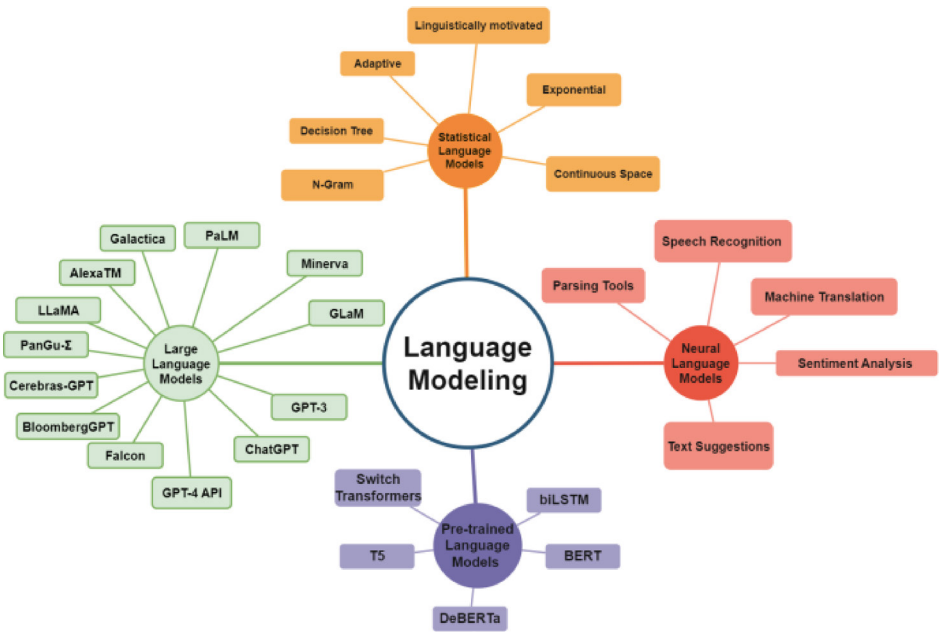


Рис. 1. Типы систем языкового моделирования. Источник: Hadi M. U. et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects //Authorea Preprints. 2024

Появление больших языковых моделей (LLM) обеспечило прорыв в области автоматической обработки естественного языка и последующих исследованиях искусственного интеллекта [17].

История развития LLM начинается с появления первых языковых моделей и нейронных сетей. Первые попытки построить языковые модели основывались на статистических методах и *n*-грамм моделях [13]. Однако эти модели были ограничены проблемами, связанными с долговременной персистентностью (LRD) и языковым контекстом. С появлением нейронных сетей и доступностью больших дата-сетов исследователи начали изучать более сложные подходы. Важным этапом стала разработка рекуррентной нейронной сети (RNN) [17, 18], которая позволяла моделировать последовательные данные, включая язык. Однако RNN также столкнулись с проблемами исчезающих градиентов и дальнодействующих зависимостей, что ограничивало их эффективность.

Прорыв в LLM произошел после того, как в 2017 г. в работе Васвани «Внимание — все, что вам нужно» была представлена архитектура Transformer [12]. Вместо рекуррентности модель Transformer использует механизм, который позволяет сосредоточиться на наиболее важных элементах входной последовательности (self-attention mechanism), что обеспечивает параллелизацию и эффективную работу с долговременной персистентностью. Она создала основу для таких моделей, как серия GPT (Generative Pre-trained Transformer) от OpenAI и BERT (Bidirectional Encoder Representations from Transformers) [15] от Google, которые достигли новаторских результатов в широком спектре языковых задач.

Впоследствии LLM прошли несколько этапов развития, при этом модели увеличивались в размере и сложности. Серия GPT, начиная с GPT-1 и продолжая GPT-2 и GPT-3, последовательно увеличивала количество параметров, что обеспечило соответствующую обработку языка и последующее создание текста.

Среди LLM российского производства можно выделить GigaChat, YandexGPT, ruGPT-3.5, Saiga 2.

Целью нашего исследования было изучение возможности применения ChatGPT как дополнительного инструмента для повышения эффективности учебно-методической работы преподавателя иностранного языка, а также анализ языковых особенностей промтов для использования в деловой коммуникации.

В процессе подготовки учебного курса по аудированию для студентов, изучающих английский язык как иностранный, мы использовали ChatGPT для составления вопросов на понимание аудиотекста, задания на «верные-неверные утверждения», вопросов для дальнейшего обсуждения проблематики аудиотекста. Несмотря на некоторые стилистические и смысловые ошибки, данный чат-бот быстро генерирует

необходимый текст, что значительно ускоряет процесс подготовки учебного материала.

В итоге мы пришли к выводу, что ChatGPT при правильном использовании является полезным инструментом для методиста, при условии, что последний принимает во внимание такие особенности работы с чат-ботом, как: языковой регистр и набор языковых средств; организация текста в зависимости от необходимого стиля; смысловая разбивка текста; возможность допущения нейросетью смысловых и грамматических погрешностей; контекст и адаптация; соблюдение авторских прав; правильное построение запроса [7].

Одним из определяющих успешное взаимодействие с любой LLM условий является правильное построение запроса или промта. Для анализа языковых особенностей промтов мы собрали текстовые данные популярных шаблонов на английском языке, имеющиеся в свободном доступе на веб-ресурсе Github (https://github.com/f/awesome-chatgpt-prompts), общий объем которых составил 90 014 знаков. Определив типичные сценарии их использования, мы разделили их на несколько категорий: профессиональные, технические, образовательные, коммуникативные, творческое письмо, личностное развитие, развлечения.

Среди языковых особенностей профессиональных и технических промтов можно выделить употребление отраслевых терминов и профессиональных жаргонизмов (*cyber security, encryption, firewalls, implementing policies*), глаголов в повелительной форме (act).

Например: I want you to act as a cyber security specialist. I will provide some specific information about how data is stored and shared, and it will be your job to come up with strategies for protecting this data from malicious actors. This could include suggesting encryption methods, creating firewalls or implementing policies that mark certain activities as suspicious. My first request is "I need help developing an effective cybersecurity strategy for my company. (см. https://github.com/f/awesome-chatgpt-prompts).

На синтаксическом уровне отмечено преобладание побудительных предложений, коммуникативной задачей которых является дать четкие инструкции чат-боту для выполнения необходимого действия.

К стилистическим особенностям анализируемых промтов относятся использование высокого регистра языка, формальной тональности, отсутствие изобразительных средств языка. Наличие таких характеристик обусловлено ситуативным контекстом.

Подводя итог, следует отметить, что мы находим целесообразным внедрение LLM в образовательный процесс и деловое общение с целью упрощения подготовки учебного материала и повышения эффективности взаимодействия. Несмотря на то что сгенерированный LLM текст нуждается в доработке и корректировке, применение этой технологии значительно облегчает процесс составления учебного пособия, а также предлагает многогранный подход к рассмотрению проблемы и позволяет выявить альтернативные точки зрения. При правильно сформированном запросе LLM быстро выдает необходимый ответ и предлагает разные варианты решения проблемы.

Таким образом, технологии моделирования языка представляют собой мощный инструмент, способствующий социальным изменениям. Они открывают новые возможности для получения и обработки информации, позволяют повысить эффективность образовательного процесса и межличностной коммуникации, в частности делового общения. Необходимо продолжать исследовать их потенциал, разрабатывать алгоритмы их использования для решения различных задач и минимизации возможных рисков, чтобы обеспечить их положительное воздействие на общество.

Список использованных источников

1. *Балдин К., Уткин В.* Информационные системы в экономике. — Litres, 2022.
2. *Винокуров И. В.* Особенности эффектов цифровой трансформации экономики Российской Федерации // Дискуссия. — 2024. — № 1 (122). URL: https://cyberleninka.ru/article/n/osobennosti-effektov-tsifrovoy-transformatsii-ekonomiki-rossiyskoy-federatsii (дата обращения: 27.11.2024).
3. *Игнатьева Г. А., Поначугин А. В.* Сквозные технологии научно-методического сопровождения педагогов в условиях цифровой трансформации образования // Педагогическое образование в России. — 2024. — № 4. URL: https://cyberleninka.ru/article/n/skvoznye-tehnologii-nauchno-metodicheskogo-soprovozhdeniya-pedagogov-v-usloviyah-tsifrovoy-transformatsii-obrazovaniya (дата обращения: 27.11.2024).
4. *Казакова Н. Ю.* Парадигма развития гейм-дизайна как вида проектной деятельности на основании этапов развития аппаратной части // Вестник Кемеровского государственного университета культуры и искусств. — 2016. — № 35. URL: https://cyberleninka.ru/article/n/paradigma-razvitiya-geym-dizayna-kak-vida-proektnoy-deyatelnosti-na-osnovanii-etapov-razvitiya-apparatnoy-chasti (дата обращения: 27.11.2024).
5. *Киселев Г., Бочкова Р.* Информационные технологии в педагогическом образовании. — Litres, 2022.
6. *Комарова Е. В.* Вовлечение чат-ботов в процесс обучения иностранным языкам. Филология и культура. Philology and Culture. 2024;(1):149–158. https://doi.org/10.26907/2782-4756-2024-75-1-149-158
7. *Мельникова Д. А., Лопаткин Д. С., Кожева А. А.* Искусственный интеллект как способ создания нового контента // Успехи в химии и химической технологии. — 2023. — № 1 (263). URL: https://cyberleninka.ru/article/n/iskusstvennyy-intellekt-kak-sposob-sozdaniya-novogo-kontenta (дата обращения: 27.11.2024).
8. *Напсо М. Д.* Цифровое неравенство и сфера образования //Человеческий капитал. — 2024. — № 2 (182). — С. 150.
9. *Петрова Н. П., Соковикова А. В.* Современные цифровые технологии в формировании и поддержке образовательной экосистемы университета // МНКО. — 2024. — № 4 (107). URL: https://cyberleninka.ru/article/n/sovremennye-tsifrovye-tehnologii-v-formirovanii-i-podderzhke-obrazovatelnoy-ekosistemy-universiteta (дата обращения: 27.11.2024).
10. *Петров П. Ю.* Социальные последствия технологического прогресса // Экономика и социум. — 2020. — № 8 (75). URL: https://cyberleninka.ru/article/n/sotsialnye-posledstviya-tehnologicheskogo-progressa (дата обращения: 27.11.2024).
11. *Vaswani A., Shazeer N., Parmar N., et al.* "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

12. *Arisoy E., Sainath T. N., Kingsbury B., and Ramabhadran B*. "Deep neural network language models," in Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT, pp. 20–28, 2012.
13. *Hadi M. U*. et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects // Authorea Preprints. — 2024.
14. *Devlin J., Chang M.-W., Lee K., and Toutanova K*., "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
15. *Muhammad Usman Hadi, Qasem Al Tashi, Rizwan Qureshi*, et al. A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. TechRxiv. July 10, 2023.
16. *Hochreiter S*. and *Schmidhuber J*. "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
17. *Mikolov T., Karafia t M., Burget L., Cernocky` J*., and *Khudanpur S*., "Recurrent neural network based language model.," in Interspeech, vol. 2, pp. 1045–1048, Makuhari, 2010.

ALGORITHMS FOR THE APPLICATION OF LARGE LANGUAGE MODELS IN THE EDUCATIONAL PROCESS AND BUSINESS COMMUNICATION

***Komarova E. V.**, PhD, Assistant Professor of English Language Department № 8, Faculty of International Law, Moscow State Institute of International Relations (University) The Russian Foreign Ministry Moscow, Russia, elvakom@rambler.ru*

Currently, language modeling systems (LLMs) play a significant role in modern technologies. These technologies not only transform human interactions but also open new prospects for the development of numerous industries. This article examines the impact of LLMs on various fields of human activity, particularly education and business communication. It provides an analysis of linguistic features of effective prompts for solving professional and technical tasks and explores the potential applications of LLMs in the educational process.

LLM • large language models • prompt • educational process • business communication

Прагматические аспекты влияния определённых дескрипций на восприятие имплицитной темпоральной информации

***Ларина О. В.**, кандидат филологических наук, доцент, Московский государственный институт международных отношений (университет) МИД России, Москва, Россия, olga21.06larina@gmail.com*

Данная статья представляет собой обзор влияния важнейших прагматических факторов (пресуппозиций, определённых дескрипций и контекста) на процесс создания говорящим и последующей обработки адресатом высказываний с имплицитной темпоральной семантикой. Рассматривается, в частности, особенность функционирования определённых дескрипций как маркеров прагматической пресуппозиции и их роль в процессе извлечения и интерпретации имплицитной информации. В работе также представлена уровневая классификация косвенных способов передачи времени.

прагматика • пресуппозиция • инференция • имплицитность • определённая дескрипция • контекст.

Современная лингвистика, по сути своей, призвана ответить на два фундаментальных вопроса: как в рамках успешной коммуникации говорящий закладывает в сообщение именно ту информацию, которую он желает донести до собеседника, и как именно собеседник понимает сообщение именно так, как это было задумано говорящим. Данные вопросы особенно важны сейчас, когда в сферу интересов лингвистов входит не только «живое» человеческое общение, но и разработка и совершенствование способов взаимодействия с искусственным интеллектом. В этой связи задачи, которые ставит перед собой прагмалингвистика уже почти сотню лет, актуальны как никогда.

В данной работе мы сосредоточимся на нескольких ключевых прагмалингвистических понятиях (а именно пресуппозициях, определенных дескрипциях, контексте и инференции) и попытаемся проанализировать их взаимосвязь и роль в процессе порождения и интерпретации скрытых темпоральных смыслов.

В качестве материала для анализа мы выбрали нашумевшее эссе The Case against Travel профессора философии Чикагского университета Агнес Каллард, опубликованное в июне 2023 г. в журнале The New Yorker.

В процессе работы над проблемой имплицитной темпоральности в английском языке нами была разработана уровневая классификация способов косвенной передачи временных значений. Нам представляется возможным выделить три основные группы конструкций.

1. *Лексические*. Косвенное указание на время содержится в значении отдельных лексических единиц. Чаще всего это достигается через упоминание деятеля, места действия либо других сопутствующих действию обстоятельств:

E.g. *Pessoa, Emerson, and Chesterton* **believed** that travel, far from putting us in touch with humanity, divorced us from it [3].

В приведенном примере время действия выделенного глагола связано с годами жизни упомянутых писателей.

2. *Грамматические*. К данной группе относятся грамматически обусловленные случаи нарушения привычного согласования времен, свернутые конструкции, а также определенные/неопределенные дескрипции:

E.g. One common argument for *travel* is that it **lifts** us into an enlightened state, educating us about the world and connecting us to its denizens [3].

Существительное пропозициональной семантики *travel* может быть развернуто до полной пропозиции **when we travel,* содержащей темпоральную информацию.

3. *Контекстуальные.* Имплицитная темпоральная референция в данном случае осуществляется через отсылки к узкому лингвистическому и/или широкому ситуационному контексту:

E.g. For example, *a decade ago, when I was in Abu Dhabi*, I went on a guided tour of a falcon hospital. [...] I **entered** the hospital knowing that my post-Abu Dhabi life would contain exactly as much falconry as my pre-Abu Dhabi life–which is to say, zero falconry [3].

В данном примере второе предложение отстоит от первого в тексте эссе на несколько страниц. Таким образом, глагол *enter* не имеет непосредственной временной локализации и получает ее только с учетом знания читателем всей истории.

Последний пример подводит нас непосредственно к теме нашего исследования. Использование определенного артикля со словом *hospital* призвано сигнализировать адресату, что он уже знаком с ситуацией и с ее обстоятельствами из предыдущего контекста, а следовательно, может восстановить в памяти и время действия глагола *enter*.

Контекст является одной из ключевых категорий в прагматике. Под контекстом принято понимать: а) непосредственное вербальное окружение высказывания; б) ситуацию общения в целом, включающую также фоновые социокультурные знания коммуникантов.

Если мы вернемся к примеру с больницей для соколов в Абу-Даби, то мы увидим, что само по себе существительное *hospital* не несет никакой темпоральной информации. В рамках высказывания "I entered the hospital knowing that my post-Abu Dhabi

life would contain exactly as much falconry as my pre-Abu Dhabi life…" время действия постепенно проясняется. Мы можем прийти к выводу, что говорящий был в больнице в Абу-Даби, и именно тогда произошло действие глагола. И наконец, широкий контекст всей ситуации помогает четко установить время действия. Широкий контекст в данной ситуации может включать в себя не только упоминание *a decade ago* за несколько страниц до этого, но и тот факт, что эссе было написано в 2023 г., и даже то, что Соколиная больница открылась в Абу-Даби в 1999 г., что позволяет точно восстановить имплицитно заложенную в сообщении информацию о времени действия.

Таким образом, соотношение «слово — контекст», по всей видимости, выглядит следующим образом:

лексическая единица >>вербальный контекст >>прагматический контекст

Понятие определенных дескрипций напрямую связано с понятием контекста. В рамках когнитивно-прагматического подхода к языку и мышлению определенные дескрипции неизменно получают отдельное внимание со времен Г. Фреге и Б. Рассела, поскольку «…весь процесс абстрактного мышления, который получил в языке звуковую и буквенную фиксацию, основан на диалектике восприятия одного и того же объекта, явления, действия или признака как отдельного, конкретного и как обобщённого, абстрактного» [2]. Учитывая, что в английском языке существуют формальные служебные дейктические слова (артикли), которые представляют определенные/неопределенные дескрипции, можно сделать вывод, что данная категория имеет особое влияние на организацию вербального мышления, а также на порождение и восприятие английской речи.

Определенные дескрипции, в отличие от имен собственных, получают референциональный потенциал (т.е. способность указывать на конкретный предмет или явление) только в контексте. Другими словами, в их случае «понятие референции не играет никакой роли в семантике высказывания как таковой, а скорее в специфике использования предложения в конкретной ситуации» [4][1]. Тем не менее, согласно Р. Брэндому, «не существует инференции без референции» [7], т.е. любое информативное высказывание должно быть привязано к ситуации и миру в целом.

В современной интерпретации фрегевская концепция определенных дескрипций гласит: «Значением определенной дескрипции является объект, который удовлетворяет предикату, следующему за определенным артиклем […]. Однако это выражение также несет пресуппозицию […], согласно которой существует только один объект, удовлетворяющий данному предикату» [1]. Смысл данного определения для нашего исследования заключается в том, что использование определенного артикля в контекстах с имплицитным временем сужает область темпоральной инференции и заменяет общее значение времени, заключающееся во временной семантике глагола личной формы, на конкретное.

[1] Здесь и далее перевод автора.

В предыдущем параграфе появилось сразу два термина, на наш взгляд, требующих отдельного комментария. Под пресуппозициями мы понимаем контекстообразующие переменные, отражающие лингвистическую и экстралингвистическую ситуацию общения (уровень владения языком коммуникантов, их фоновые знания, прагматические аспекты высказывания и его модус). На основании заложенных адресантом пресуппозиций адресат осуществляет иференцию — вывод о заложенной в сообщении информации о «языке, ситуации и мире», который будет тем точнее, чем в большей степени совпадают знания коммуникантов о контексте (узком (языковом, вербальном) и широком (ситуативном)) высказывания.

Тён Ван Дейк выделяет следующие пресуппозиции высказывания:

- [знание] структурных характеристик высказывания;
- [знание] паралингвистических (фонетических и просодических) особенностей высказывания;
- непосредственное наблюдение/восприятие коммуникативного контекста;
- предшествующее знание личных особенностей говорящего;
- знание/предположение о характере вербального взаимодействия;
- знания/предположения, извлеченные из предыдущих речевых актов или частей дискурса на микро- (узкий контекст) и макро- (широкий контекст) уровнях;
- знание общепринятых семантических и прагматических законов коммуникации;
- прочие виды фоновых знаний [6].

На нашем материале маркерами пресуппозиции являются, в частности, определенные дескрипции. Так, в следующем примере:

E.g. "During *my Paris wanderings*, I **would stare** at people, intently inspecting their clothing, their demeanor, their interactions [3].

Выражение *my Paris wanderings* содержит определенную дескрипцию, которая является маркером пресуппозиции адресанта о том, что его читатель знаком с предыдущим контекстом, в котором содержится указание в т.ч. на время этой истории. Таким образом, определенная дескрипция, в частности, запускает процесс инференции адресатом значения данного выражения как части предшествующего контекста и, как следствие, извлечение темпоральной информации.

Чтобы проиллюстрировать место и значение использования определенных дескрипций в процессе порождения и восприятия высказывания с имплицитной темпоральной семантикой, можно обратиться к схеме интеграционной обработки высказывания на естественном языке, предложенной М. Бирвишем [5].

$$F(s) = < P, S, LF(the), C(Pr/the) > \rightarrow <P', S', LF'(the), C'(I/*PR*the) >.$$

Мы внесли изменение в нотацию Бирвиша с учетом когнитивно-прагматического представления о порождении и интерпретации высказывания. Данная модель отражает создание высказывания (s) одновременно на фонетическом (P), синтаксическом (S), семантико-морфологическом (LF/logical formula) и контекстуальном (C) уровнях с учетом появления определенных дескрипций (the) на семантическом уровне, их зависимости от пресуппозиций и актуализации на уровне контекста. Второй этап предполагает интегральное извлечение заложенной говорящим

информации и инференцию адресатом всех или части пресуппозиций адресанта, в том числе с опорой на распознавание определенных дескрипций.

E.g. I **walked** from one end of *the city* to the other, over and over again, in a straight line; if you plotted my walks on a map, they would have formed a giant asterisk [3].

В приведенном примере определенная дескрипция *the city* появляется на семантика-морфологическом уровне как маркер субстантива, однако актуализуется и приобретает референциональные характеристики на уровне контекста, отмечая собой пресуппозицию знания коммуникантами вербального и прагматического контекста высказывания. На этапе извлечения заложенной в предложении информации адресат при адекватном знании языка непременно распознает семантическую и грамматическую функцию артикля, но при отсутствии осведомленности о предыдущем контексте «потеряет» прагматическую пресуппозицию, при этом часть информации (в особенности о времени действия) останется за пределами инференции.

Таким образом, на основе проведенного анализа представляется возможным сделать следующие выводы:

- прагматические факторы, такие как наличие в высказывании определенных дескрипций, способны влиять на восприятие имплицитной темпоральной информации. Однако для реализации их референциального потенциала требуется знание коммуникантами контекста сообщения (как узкого вербального, так и широкого прагматического);

- определенные дескрипции способствуют процессу инференции и являются со стороны адресанта маркерами прагматической пресуппозиции, при распознании которой общая темпоральная информация, заложенная в форме глагола, получает конкретизацию и точечную локализацию на оси временного дейксиса;

- иногда, однако, инференция временных смыслов, основанная на интерпретации определенных дескрипций, не совпадает с заложенными адресантом пресуппозициями. В этом случае адресат «теряет» прагматическую пресуппозицию знания контекста, замещая ее общими знаниями структурно-семантических особенностей языка, и, как следствие, не «считывает» информацию о времени высказывания.

Список использованных источников

1. *Вострикова Е. В.* Семантика определенных дескрипций: новые перспективы // Эпистемология & философия науки. — 2014. — С. 240.
2. *Тымчук Е. В.* Артиклевый дейксис и его роль в выражении концептуальной категории общего и единичного. 2018. URL: https://cyberleninka.ru/article/n/artiklevyy-deyksis-i-ego-rol-v-vyrazhenii-kontseptualnoy-kategorii-obschego-i-edinichnogo (дата обращения: 8.12.2024).

3. *Callard A*. The Case Against Travel. URL: https://www.newyorker.com/culture/the-weekend-essay/the-case-against-travel (дата обращения: 8.12.2024).
4. *Horn L. R*. and *Ward G*. The handbook of pragmatics. Wiley-Blackwell, 842 p., 2006.
5. *Bierwish M*. "How on-line is language processing?", in The process of language understanding/ Edited by Flores D'Arcais G.B., Jarvella R.J. New York: John Wiley, P. 113–168, 1983.
6. *Van Dijk T. A*. "Context and cognition: knowledge, frames and speech act comprehension," in Journal of Pragmatics Volume 1, Issue 3, p. 211–231, 1977.
7. *Zarębski T*. and Kublikowski R. Deixis, Reference and Inference, 2021. URL: https://journals.openedition.org/ejpap/2530 (дата обращения: 08.12.2024).

PRAGMATIC ASPECTS OF THE INFLUENCE OF CERTAIN DESCRIPTIONS ON THE PERCEPTION OF IMPLICIT TEMPORAL INFORMATION

***Larina O. V.**, Candidate of Philological Sciences, Associate Professor,*
Moscow State Institute of International Relations (University)
The Ministry of Foreign Affairs of Russia, Moscow, Russia,
olga21.06larina@gmail.com

This article provides an overview of the influence of key pragmatic factors (presuppositions, definite descriptions, and context) on the process of utterance generation by the speaker and subsequent processing by the addressee, focusing on implicit temporal semantics. Special attention is given to the function of definite descriptions as markers of pragmatic presupposition and their role in extracting and interpreting implicit information. The paper also presents a multi-level classification of indirect methods for conveying temporal information.

pragmatics • presupposition • inference • implicitness • definite description • context

Мера определённости/неопределённости текста как проблема научно-технического перевода (на материале переводов с английского языка на русский)

***Крюков А. Н.**, доктор филологических наук, доцент, профессор, Военный университет Министерства обороны; Московский авиационный институт (НИУ), Москва, anatolikryukov@yandex.ru*

В статье сделана попытка описать и систематизировать многолетний опыт преподавания автором практических курсов научно-технического перевода в комбинации английского и русского языков. Сформулирован ряд закономерностей, которые пока ещё, на взгляд автора, не получили широкого освещения в теории перевода, но незнание которых, как правило, приводит к многочисленным и подчас грубым ошибкам в переводе. Приведённые закономерности, по мысли автора, восходят к неким речевым универсалиям, характерным для каждого из контактирующих языков: большей мерой определённости для русского языка и меньшей для английского.

пресуппозиции • фоновые знания • мера определённости/неопределённости • трудности перевода • редукция многокомпонентных терминов • замена местоимений антецедентами • редукция промежуточных смысловых выводов

ПРОБЛЕМЫ ПЕРЕВОДА

Проблемы научно-технического перевода в различных аспектах представлены в различных научных и методических работах [1–16].

Предлагаемая работа явилась плодом первичного теоретического осмысления продолжительного опыта преподавания автором курсов основ технического перевода и перевода спецтекста на кафедре перевода и переводоведения Института иностранных языков Московского авиационного института.

Для целей настоящей работы под мерой определенности текста будем понимать степень вербальной выраженности в нем всех сем (элементарных

смыслов) и пресуппозиций (суждений о мире), на которые опирается любое высказывание (фраза) и которые делают высказывание осмысленным. Причем под пресуппозициями нами будут пониматься не только суждения (или знания) о мире, существующие вне конкретного переводимого текста (фоновые знания), но и те, которые были сформированы в ходе развертывания смыслового содержания конкретного переводимого текста (актуальные знания).

Можно, по-видимому, считать, что научно-технические тексты как на русском, так и английском языках стремятся к большей, по сравнению с текстами других жанров и стилей, определенности. Это подтверждается, в частности, широким использованием специальной терминологии, стремлением к однозначности выражения мысли и исключению двусмысленности и инотолкования. Разумеется, мы не можем при этом говорить о 100 % определенности научно-технического текста в понимаемом выше смысле. Абсолютная экспликация всех релевантных для коммуникации сем и пресуппозиций — недостижимый идеал научно-технического текста. Кроме того, вряд ли такой гипотетический текст можно было бы считать в полном смысле идеалом текста. Любая коммуникация — в том числе и научно-техническая — стремится не только к максимальной определенности, но и максимальной эффективности. А эффективной может быть только информативная, а не тавтологичная коммуникация, построенная по типу известного стихотворного текста «Дом, который построил Джек».

Таким образом, в научно-технической коммуникации (равно как и всякой другой) сталкиваются две противоположные тенденции — тенденция к определенности и тенденция к информативности. Всякий научно-технический текст является воплощением уникальной равнодействующей этих противоборствующих тенденций при относительном, по сравнению с другими жанрами и стилями, доминировании тенденции к определенности.

Данное положение распространяется на научно-технические тексты как на русском, так и на английском языках. Проблема, однако, заключается в **различной мере определенности** научно-технического текста на русском и английском языках. По нашим наблюдениям, научно-технический текст на русском языке отличается **большей** мерой определенности по сравнению с аналогичным текстом на английском языке.

Этот, казалось бы, теоретический вывод влечет за собой целый ряд чисто практических закономерностей научно-технического перевода в комбинации английского и русского языков. Остановимся вкратце на некоторых из них.

1. Редуцирование многокомпонентных научно-технических терминов на английском языке до одно-двухкомпонентных по мере их употребления в тексте, с одной стороны, и сохранение в переводе многокомпонентной структуры аналогичных терминов на русском языке, с другой стороны.

Проиллюстрируем сказанное на конкретных примерах:

'Even if the windmills are built, they will not in themselves plug the **generation gap**'. Словарное значение "generation gap" –«разрыв между поколениями, конфликт поколений, проблема отцов и детей». Совершенно очевидно, что словарный эквивалент здесь бесполезен. Если же мы, имея в виду феномен редукции

многокомпонентный терминов, обратимся к английскому тексту, то несколькими абзацами выше обнаружим, от какого многокомпонентного термина он произошел — это не что иное, как "power generation". Тогда перевод всей фразы приобретает следующий вид: «Даже если будут построены ветряные генераторы, они не покроют дефицит **генерирующих мощностей»**.

Рассмотрим еще один пример редукции многокомпонентного термина: "Making ethanol and burning it in a car (**without land changes**) emits 20% less carbon dioxide than refining and burning petrol". Ни один из студентов четырех языковых групп не справился с переводом фрагмента текста, заключенного в скобки. А между тем знание тенденции к редуцированию многокомпонентных терминов в английском языке вначале позволяет выдвинуть гипотезу о том, что мы как раз и имеем дело с проявлением такой тенденции, а затем найти в тексте и подтверждение данной гипотезы — исходной формой терминологического словосочетания land changes была использованная ранее форма land use changes. Восстановление многокомпонентного термина позволяет корректно перевести всю фразу: «Производство этилового спирта и его сжигание в автомобиле (**без учета изменений в землепользовании**) приводит к выбросу в атмосферу углекислого газа на 20% меньше, чем при производстве и сжигании бензина».

2. Замена (по мере их употребления в тексте) терминологических словосочетаний в английском языке на слова общелитературного английского языка при обязательном или преимущественном сохранении в переводном тексте на русском языке изначального терминологического сочетания.

Приведем только один, но показательный пример: "The researchers found about 190 genes that passed all the tests. Encouragingly, these included the previously known **culprits** among the two cancers in question". Словарное значение существительного culprits — «преступники», «виновники». Это слово общелитературного языка использовано в качестве аналога терминологического словосочетания «гены, вызывающие что-либо» или «гены, ответственные за что-либо». Перевод всей же фразы принимает следующий вид: «Исследователи обнаружили порядка 190 генов, которые прошли все тесты. К счастью, в их число вошли и ранее известные **гены, вызывающие** две рассматриваемые разновидности рака».

3. Широкое использование в исходном научно-техническом тексте на английском языке местоимений и их преимущественная последующая замена на соответствующие антецеденты в переводном тексте на русском языке.

Например, в следующей фразе на английском языке используются два местоимения — притяжательное местоимение its и указательное местоимение it , а в переводе на русский язык в силу большей меры определенности русскоязычного текста вместо указательного местоимения используется его антецедент: "The idea would be to stare at a distant

star and report instances when its light had been bent by the gravity of a planet passing in front **of it"** — «Идея состояла бы в том, чтобы наблюдать за удаленной звездой и регистрировать случаи, когда свет от нее искривляется за счет силы гравитации планеты, проходящей перед **звездой**».

Реконструкция антецедента местоимения зачастую оказывается непростой интеллектуальной задачей и требует от переводчика осмысления всей описываемой фразой ситуации.

4. Использование в научно-технических текстах на английском языке количественных и неопределенных числительных без характеризуемых ими существительных, если ранее эти существительные уже были названы, при непременном повторении таких существительных в переводе на русский язык:

"For one point about anticancer genes is that **several** have to go wrong in the same cell before a tumour can grow" — «Одним из факторов, характеризующих антираковые гены, является то, что должен произойти сбой в работе сразу **нескольких таких генов** в пределах одной и той же клетки, чтобы начала развиваться раковая опухоль».

"In the immediate postwar era, up to 1970, it was normal practice not to introduce an engine to the airliners until **hundreds** or even **thousands** had gained experience in fighters and bombers." — «Сразу же в послевоенный период и вплоть до 1970 года обычной практикой было не устанавливать новые двигатели на воздушные суда гражданской авиации до тех пор, пока **сотни и даже тысячи таких двигателей** не пройдут испытания на истребителях и бомбардировщиках».

5. Редуцирование промежуточных смысловых выводов в обобщающем (итоговом) умозаключении на английском языке с последующим их восстановлением в переводе на русский язык.

Данная закономерность представляет собой особый интерес и одновременно особую трудность при переводе научно-технических текстов с английского языка на русский. С нею связаны наиболее радикальные трансформации или компенсации (в терминах лингвистической теории перевода, хотя на самом деле здесь имеет место реконструкция опущенных смысловых звеньев) исходного текста в переводе, которые, насколько нам известно, пока еще не описаны теорией перевода. Восстановление в переводе на русский язык промежуточных смысловых выводов, опущенных в научно-техническом тексте на английском языке, отдаленно может напоминать разве что такой известный в теории и практике перевода раздел, как перевод реалий. Однако эта параллель касается лишь глубины и масштабности перевыражения исходного текста, но не его причин (мотивированности) и используемого для его описания понятийного аппарата.

Проиллюстрируем данную закономерность научно-технического перевода и соответствующий ей прием перевода, который мы назвали в рабочем порядке как восстановление промежуточных смысловых выводов, на конкретном примере:

"Acid rain and old age explain the shortage of capacity. Britain's coal and nuclear plants together account for just under 45% of all power generation. But most of the nuclear plants, and around half of the coal plants are due to close in the near future." Хотя

нас будет интересовать перевод только первой фразы, мы специально привели целый абзац, чтобы продемонстрировать, что в нем нет прямого ответа на то, как надо переводить первую фразу абзаца (заметим, что данный абзац встретился нам примерно в середине статьи, посвященной анализу электроэнергетического сектора экономики Великобритании). В буквальном, но литературном переводе первая фраза приведенного выше абзаца звучит следующим образом: **«Кислотные дожди и продолжительный срок эксплуатации объясняют дефицит генерирующих мощностей»**. Подчеркнем, что это хотя и буквальный (в смысле подстрочный), но при этом уже литературно обработанный перевод: так, английский термин capacity правильно идентифицирован в нем как редуцированный дублет многокомпонентного термина power generation capacity, ранее встречавшийся в тексте, что нами уже описано в пункте 1 выше; словосочетание old age терминологически переведено как «продолжительный срок эксплуатации», а не «преклонный возраст».

И тем не менее данный перевод должен быть отвергнут как абсолютно алогичный исходя из существующих норм построения русскоязычного текста вообще и научно-технического в частности, т.к. в нем отсутствуют исходные посылки для итогового умозаключения. Нам не остается ничего другого, как попытаться восстановить их, опираясь на сведения, содержащиеся в предшествующем тексте: **«Борьба с кислотными дождями, вызываемыми работой тепловых электростанций, и закрытие объектов электроэнергетики в связи с продолжительным сроком их эксплуатации являются причиной дефицита генерирующих мощностей».** Возможен и другой вариант перевода первой посылки, содержащейся в данной фразе: **«Закрытие тепловых электростанций из-за кислотных дождей, вызываемых их работой, а также закрытие объектов электроэнергетики в связи с продолжительным сроком их эксплуатации являются причиной дефицита генерирующих мощностей».** Предложенные варианты перевода отличаются друг от друга тем, что в первом из них смысловой вывод имплицируется, но поскольку он легко и однозначно вытекает из эксплицированной посылки, то такой перевод, на наш взгляд, не нарушает допустимой для русскоязычного научно-технического текста меры определенности и вполне может быть признан приемлемым. Во втором же варианте необходимый смысловой вывод эксплицирован переводчиком.

Заметим, что ни один студент из всего потока с переводом этой фразы, к сожалению, не справился (а на наш вопрос, каким образом кислотные дожди приводят к дефициту генерирующих мощностей, в лучшем случае звучал ответ: «наверное, кислотные дожди вызывают коррозию электростанций и их выход из строя»), что мы можем объяснить тем, что подобного рода явления, во-первых, реально представляют собой особую, повышенную, трудность для перевода, а во-вторых, они пока еще не стали объектом теории и практики перевода.

Таким образом, нами рассмотрены пять закономерностей научно-технического перевода с английского языка на русский, которые восходят к одной стоящей за ними тенденции — стремлению русскоязычного

научно-технического текста к большей мере определенности по сравнению с аналогичным англоязычным текстом. При этом мы понимаем, что данная тенденция реализуется не только в описанных закономерностях, но существуют и другие, не описанные в данной работе. Кроме того, даже те пять закономерностей, которые мы выявили в ходе обучения практике перевода, нуждаются в более углубленном изучении. Так, было бы, на наш взгляд, весьма плодотворно привлечение количественных методов исследования вышеприведенных закономерностей. Одно дело прийти к выводу, что, например, замена местоимений в англоязычном научно-техническом тексте на их антецеденты в русскоязычном переводном тексте является закономерной, другое дело — исследовать статистику употребления местоимений в тех и других текстах и придать данной закономерности количественное измерение. То же самое касается и остальных закономерностей. Нам представляется, что это плодотворные и интересные темы для будущих курсовых и выпускных квалификационных работ наших студентов.

Список использованных источников

1. *Агузарова К. К.* Особенности перевода английского научно-технического текста // Науки о земле: Труды молодых учёных. — 2010. — № 1. — С. 171–175.
2. *Гринев-Гриневич С. В.* Введение в терминоведение. — М. : Московский лицей, 1993. — 309 с.
3. Жидков А. В. Понятие «научно-технический перевод» // Science Time. — 2014. — № 4. — С. 99–102.
4. *Климзо Б. Н.* Ремесло технического переводчика. Об английском языке, переводе и переводчиках научно-технической литературы. — М.: «Р. Валент», 2006.
5. *Комиссаров В. Н.* Лингвистические основы научно-технического перевода. Пособие по научно-техническому переводу / Под. ред. В. Н. Комиссарова. — М.,1980. — Ч. I.26.
6. *Коняева Л. А.* О некоторых трудностях научно-технического перевода // Перевод и сопоставительная лингвистика. — 2015. — № 11.
7. *Лейчик В. М.* Терминоведение: Предмет, методы, структура. — М.: ЛИБРОКОМ, 2009. — 256 с.
8. Мельникова А. И. Научно-технический перевод и его виды // Язык и культура (Новосибирск). — 2015. — № 17. — С. 116–120.
9. *Тараканов Г. В.* Основные термины в нефтегазопереработке. Краткий справочник. — Астрахань: Изд-во АГТУ, 2011. — 100 с.
10. *Филатов О. С.* Проблема перевода научно-технических текстов:
11. Особенности и переводческие стратегии // Вестник Челябинского государственного университета. — 2023. — № 6 (476). — С. 144–153.
12. *Чистюхина С. Н.* Межотраслевая полисемия в терминологической системе современного английского языка : автореф. дис. ... канд. филол. наук. — М., 2011. — 28 с.
13. *Felber H.* Terminology manual. Paris : Unesco and Infoterm, 1984. 426 p.
14. *Fu, L., Liu, L.* What are the differences? A comparative study of generative artificial intelligence translation and human translation of scientific texts. Humanit Soc Sci Commun 11, 1236 (2024). https://doi.org/10.1057/s41599-024-03726-7
15. *Goncharova, L. A., Fedorinova, Z. V., Kositskaya, F. L.* (2022). Characteristics of Scientific and Technical Texts in the Engineering Field of Modern English. In: Anikina, Z. (eds) Integration of Engineering Education and the Humanities: Global Intercultural Perspectives. IEEHGIP 2022. Lecture Notes in Networks and Systems, vol 499. Springer, Cham. https://doi.org/10.1007/978-3-031-11435-9_57
16. *Li, J.* Effectiveness of SmartCat Technology in Artistic, Scientific-Technical and Social-Publicist Translations: How Modern Software Influence the Specificity of Different Types of Translation. J Psycholinguist Res 52, 1841–1854 (2023). https://doi.org/10.1007/s10936-023-09973-0

THE MEASURE OF CERTAINTY/UNCERTAINTY OF THE TEXT AS A PROBLEM OF SCIENTIFIC AND TECHNICAL TRANSLATION (BASED ON THE MATERIAL OF TRANSLATIONS FROM ENGLISH INTO RUSSIAN)

Kryukov A. N., *PhD, Associate Professor, Professor, Military University of the Ministry of Defense; Moscow Aviation Institute (NRU), Moscow, anatolikryukov@yandex.ru*

The article attempts to describe and systematize the author's long-term experience of teaching practical courses of scientific and technical translation in a combination of English and Russian languages. A number of patterns have been formulated, which, in the author's opinion, have not yet been widely covered in translation theory, but ignorance of which, as a rule, leads to numerous and sometimes gross errors in translation. The above patterns, according to the author, go back to certain speech universals characteristic of each of the contacting languages: a greater measure of certainty for the Russian language and a lesser one for English.

presuppositions • background knowledge • measure of certainty/uncertainty, translation difficulties • reduction of multicomponent terms • replacement of pronouns with their antecedents • reduction of intermediate semantic conclusions

Зарубежные дискурсивные школы: немецкие дискурсивные практики и перспективы их использования

Ахметов А. А., *аспирант, Казанский федеральный университет, г. Казань*

Тагиров Р. М., *старший преподаватель, кафедра теории и практики преподавания иностранных языков, Институт филологии и межкультурной коммуникации, Казанский федеральный университет, г. Казань*

В данной статье речь идёт об определении понятия «дискурс» в лингвистической практике немецких школ дискурс-анализа и перспективах использования данных подходов с учётом современного развития технологий. Авторы дают определение понятия «дискурс» в рамках указанных школ (Гайдельбергской, или Маннгеймской, Дюссельдорфской, Ольденбургской лингвистических школ), перечисляют основных представителей данных направлений XX в. и критически освещают основные идеи, которые кажутся перспективными на настоящий момент.

дискурс-анализ • немецкие лингвистические школы • компьютерная лингвистика • корпус • аргументация

ВВЕДЕНИЕ

Обращение в этой статье к немецким подходам в дискурсивных исследованиях не является случайностью: лингвистические подходы XX в. немецких лингвистов отличались глубиной и степенью разработанности параметров, определяющих сущностные характеристики того или иного лингвистического явления, тем более такого многостороннего, как дискурс. Современные лингвистические исследования, по сути своей, являются междисциплинарными, что позволяет обратиться к накопленному опыту и по-новому взглянуть на существующее.

При анализе дискурса перед лингвистом сразу возникает проблема многосторонней связи лингвистического воплощения и идейного содержания конкретного дискурса. Возникает необходимость понимания и различения исходного и отраженного текста в общественной практике. План идейного содержания сегодня, по сути, является определяющим компонентом проведения подобного анализа. В этой связи хотелось бы остановиться на опыте зарубежных коллег, которые в 80–90-е годы прошлого столетия достаточно активно занимались выявлением факторов, влияющих на интерпретацию содержания исходного текста как элемента отраженного дискурса, его письменной фиксации. Остановимся на немецких школах и направлениях, которые, как нам кажется, наиболее тесно связаны с проблематикой исследования.

ГАЙДЕЛЬБЕРСКАЯ, ИЛИ МАНГЕЙМСКАЯ, ГРУППА

Одной из немецких школ дискурс-анализа является Гайдельберская, или Мангеймская, группа, которая занимается исторической семантикой. Для выявления изменений семантики определенных концептов как отражения явлений действительности представителями данной школы применяются традиционные лингвистические методы: при семантическом анализе учитываются изменения семантики отдельных лингвистических единиц в процессе исторического развития, дистрибуция и валентность. Внимание исследователей привлекает и художественная литература, которая создает историю научно-философского сознания. В противовес этому Дитрих Буссе и Фриц Германнс предлагают программу «всеобъемлющей истории сознания в исторические времена» (Busse, с. 11) и «историю языка как историю менталитета» (Herrmans).

В этой связи обнаруживается концепция создания корпусов на основе текстов как отражения дискурсов, разработанная Вольфгангом Тойбертом. Дискурсы представляют собой тематически организованные специфические корпусы, состоящие из отдельных текстов, которые связаны между собой (Teubert, с. 148). Совокупность устных или письменных текстов по определенной теме является в этом случае потенциальным дискурсом (Hermanns, с. 89). В понимании Буссе и Тойберта дискурсы — это виртуальный корпус, состоящий из сохранившихся отдельных текстов (Busse, с. 14; Teubert, с. 148).

В качестве критерия для отбора текстов в корпус Тойберт указывает на значимость текста в рамках всего дискурса, определяемой частотой перекрестных ссылок (Teubert, с. 148, 190).

Определение репрезентативности корпуса вызывает больше сложностей, чем оценка важности текста. В лексикографии репрезентативность, по сути, представляет собой статистическую задачу, тогда как в семантике дискурса она возникает только как результат интерпретации данных. В определенных случаях, когда речь идет о качестве исследуемых текстов, при создании корпуса важно также учитывать их однородность. Например, при изучении изменения значения слов следует следить за сферой функционирования данной лексики. Критерий однородности для создания корпуса предполагает, что такие тексты должны быть тематически связаны между собой. Кроме того, корпус становится однородным, когда включает функционально схожие тексты (Teubert, с. 190).

ДЮССЕЛЬДОРФСКАЯ ГРУППА

Дюссельдорфская школа дискурс-анализа была основана Георгом Штетцелем в середине 80-х годов прошлого столетия и известна своими многочисленными публикациями по истории языка, связанной с языковыми изменениями в немецком языке после 1945 г. Изначально

понятие дискурса не играло значительной роли для анализа языка публичных дебатов, которые изучал Штетцель. Оно было введено в его антологии «Концепции противоречий», изданной в 1995 г. В данной работе Штетцель упоминает «лингвистическую историю как ... историю тем публичного дискурса» и «конфликтные дискурсы», которые необходимо изучить (Stoetzel, с. 14), не углубляясь при этом в детали концепции дискурса.

Другие представители Дюссельдорфской школы, такие как Матиас Юнг, Карин Бёке и Мартин Венгелер, в частности, сделали попытку конкретизации концепции дискурса. Например, в 1996 г. Юнг предпринял попытку дальнейшего развития понятия лингвистического анализа дискурса, разработанного Буссе и Тойбертом, через введение дополнительных подразделов, чтобы выделить в общих дискурсах (например, дискурс экологической политики) отдельные тематики социально-политических сообществ или поддискурсов (например, в экологическом дискурсе отдельной тематикой является «энергия ветра»).

Юнг противопоставляет концепцию дискурса как корпуса текстов, представленную Буссе, Германом и Тойбертом, своему собственному пониманию дискурса как совокупности или сложной цепочке высказываний. Причем комментарии и высказывания должны быть иерархически организованы. Он определяет дискурс как «систему отношений между тематически связанными комплексами высказываний» (Jung, с. 463). В ходе дискурс-анализа исследование может опираться на взаимосвязи отдельных высказываний (включая утверждения или комментарии) и охватывать значительные объемы текста, что позволяет достичь высокой степени объективности (например, контент-анализ комментариев в прессе по конкретной тематике за год).

Юнг ранжирует единицы анализа по следующей схеме: «звук — слово — предложение — текст — дискурс» (Jung, с. 459), что, как и у Буссе, подтверждает иерархическую структуру дискурса. В то же время дискурс, в отличие от отдельного текста, не является замкнутой единицей и его временные и пространственные формы, а также границы являются относительно размытыми. Юнг акцентирует внимание на изучении лексической стороны дискурса и сосредотачивается на ключевых или «популярных» словах. Он подчеркивает необходимость «общего исследования лексического запаса, связанного с дискурсом» (Jung, с. 465) и более глубокого анализа ключевых слов в «контексте моделей аргументации» (Jung, с. 153). По его мнению, «изменения в использовании тематических моделей в дискурсе должны анализироваться вместе с изменениями в лексических единицах, таких как метафоры и ключевые слова ..., чтобы выявить доминирующие социальные тенденции и понимание коллективного мышления и мнений того времени» (Wengeler, с. 101). Кроме того, особое внимание исследователя вызывают метафоры и ассоциации (К. Бёке), а также анализ мотивированности высказывания (М. Венгелер).

Исторически ориентированный метафорический анализ дискурса, предложенный Карин Бёке, также представляет собой исследование лексических единиц, подобно анализу ключевых слов. Однако метафора, обладая особыми функциями, может служить «индикатором и фактором социокультурных условий, подходов и событий, что, проще говоря, предоставляет ценную информацию о политической истории и социальном менталитете» (Jung, с. 448). Метафоры могут быть изучены с точки зрения их семантико-прагматического содержания и социокультурной роли. В этом случае дискурсивно-исторический подход позволяет проследить

изменения в типах метафор и сделать выводы о метафорических областях, которые преобладают в дискурсе.

Для М. Венгелера важно выявление «повторяющихся в текстах, относящихся к дискурсу, моделей и утверждений, а задача ... заключается в реконструкции типичных базовых моделей для различных дискурсов и этих моделей для сравнительного анализа» (Jung, с. 123). Иначе говоря, для исследователя важно не только то, что используется, но намного важнее, почему было использовано какое-либо явление. Венгелер осознаёт, что его работа носит субъективный характер, особенно когда дело касается «топосов, лежащих в основе убедительности аргументации» (Jung, с. 125). Он стремится найти баланс между контекстуальной абстрактностью моделей аргументации и их актуальностью для отдельных случаев. М. Венгелер пытается разработать собственную категорию моделей аргументации для конкретных дискурсов, различая языковые и дискурсивные топосы (Jung, с. 154).

Исследования дискурса, представленные в рамках Дюссельдорфской школы, направлены на улучшение качества и объективности методов интерпретации дискурса, для чего используются описательные методики и диахронический семантический анализ, а также отдельные фигуры речи (метафоры, тропы и пр.), что, несомненно, облегчает выявление глубинных структур внутри дискурса.

ОЛЬДЕНБУРГСКАЯ ГРУППА

Группа исследователей дискурса во главе с Клаусом Глойем (университет Ольденбурга) опирается на социально-философскую концепцию дискурса Фуко и может быть отнесена к критической лингвистике, ориентированной на содержание (Gloy, с. 7). Дискурс рассматривается ими как совокупность взаимосвязанных текстов, которые ссылаются друг на друга. В рабочей группе Ольденбурга дискурс трактуется как супертекст или гипертекст — категория, предназначенная для описания и анализа диалогических и интертекстуальных связей между отдельными текстами. Поскольку такие связи не являются линейными, тексты образуют многослойную и запутанную сеть (Busse Teubert, 1994; Hermanns, 1995). В этом Ольденбургское определение дискурса тесно связано с концепциями дискурса Гайдельбергской, или Маннгеймской, группы, в частности Буссе, Тойберта, Германса и Юнга. Обе группы немецких исследователей солидарны в том, что исследованию подлежат полные тексты, даже если в рамках дискурса один текст содержит лишь ссылки на отдельные утверждения из другого текста.

Однако, согласно пониманию рабочей группы Глойа, дискурс представляет собой не языковую единицу, находящуюся на вершине цепочки «звук — слово — предложение — текст», а представляет собой «динамическое образование коммуникативной практики» (Gloy, с. 8), формируемое

в каждом конкретном случае. Ссылаясь на работы Йегера (1993), Ольденбургская группа определяет дискурс как совокупность переплетенных дискурсивных связей, которые организованы тематически и образуют сложную иерархию.

Представители Ольденбургской группы утверждают, что корпус не может быть определен на основе субъективной оценки исследователя, поскольку не учитывает позиции других участников дискурса и в этом случае теряется фактор процессуальности. Для снятия данного противоречия понятие дискурса должно быть уточнено с точки зрения категории рецепции и рефлексивности для более детального объяснения референций текстов друг к другу. Рефлексивность дискурса проявляется в том, что текст связывает или противопоставляет определенные противоречия, т.е. она служит определением внутренней сложности и нелинейности дискурса.

Рецепции понимаются как реакции на предшествующее текстовое смыслообразование или «отклики на нечто ранее ... выраженное» (Gloy, с. 17). Триггером рецепции являются ядерные структуры текста, которые предполагают определенную реакцию с точки зрения реципиента, т.е. сперва текст должен быть воспринятым, а потом вызывать определенные реакции.

Ольденбургская группа исследователей оперирует также понятиями «(синтаксическая) когезия» и «(семантическая) когерентность», что является относительно новым в эмпирических исследованиях дискурса. Кроме того, в эмпирических исследованиях дискурса ими был рассмотрен ряд лингвистических концептов, например с помощью исследования фразеологизмов, идиом, метафор, а также семантики прототипов и стереотипов было предложено выявлять общественно значимые идеи (концепты) и нормативность с точки зрения дискурса. Конкретный анализ в таком случае базируется на выявлении аллюзии, вариативного заимствования, коннотации и интертекстуальности. Особенно продуктивны в этом смысле эксплицитные наблюдения над предшествующими текстами, поскольку они позволяют выявить то, что было воспринято в тексте (Gloy, с. 15).

Таким образом, дискурс — это и объект, полученный в результате анализа, и объект, конституируемый в самой рецепции, а референции, конституирующие дискурс, представляют собой «практики исключения, включения, субституции и формирования оппозиций» (Gloy, с. 16). Таким образом, рецепции отражают реакцию на нечто ранее высказанное, но в то же время несут в себе стратегическое послание, с помощью которого они стремятся воздействовать на других. Референтный текст, таким образом, обращен несколько раз: он обращается к исходному тексту и к публике. Однако рецепции также фиксируют работу над имиджем, отражение общих ценностей, социальную идентичность, индивидуальность и преобладание одного над другим (Gloy, с. 16–18).

Тем не менее остается сомнительным, что детальное доказательство связи текстов, проведенное исследователями Ольденбургского проекта, может быть применимо в глобальном масштабе: в этом случае необходимо, чтобы каждый реципиент был способен проводить лингвистический и семантический анализ отдельных фрагментов дискурса самостоятельно. Это особенно актуально, поскольку накопление знаний о предмете отходит на второй план по сравнению с его лингвистической реализацией.

ВЫВОДЫ

Понятие дискурса многократно пересматривалось на протяжении всего своего существования. Письменный дискурс с точки зрения немецких ученых определяется как некий виртуально-существующий конструкт или совокупность тематически или иначе объединенных текстов, иерархически связанных между собой. Например, группа Ольденбургского университета подчеркивает важность внутренней структуры дискурса и вводит категорию рефлексивности. Дискурс представляется не как эвристическая конструкция, а как жизненная реальность. Дискурс рассматривается как сеть высказываний, отраженных в текстах. Одновременно каждый текст может содержать несколько дискурсов. Такое понимание дискурса смешивает объектный уровень (дискурс как реальная сущность) и уровень наблюдателя (дискурс как объект изучения лингвиста.

Это подводит к двум основным проблемам.

1. Научные работы с историческим материалом сталкиваются с разрывом между воображаемым и виртуальным миром, создаваемых внутри дискурса.

2. Дискурс является тематически обусловленной конструкцией, и ограничение формальными критериями не является объективным. Многие металингвистические дискурсы не могут быть поняты при анализе лишь одного типа текста, т.к. каждое высказывание может повлиять на формирование дискурса. Ограничение исследования только одним видом текста (например, публицистическим текстом без учета комментариев) заведомо будет неполным.

Одним из методологических подходов к лингвистическому анализу дискурса является моделирование дискурса как гипертекста в компьютерных системах. Гипертекст представляет собой сеть текстовых элементов, связанных электронными ссылками, где узлы обозначают текстовые блоки, ссылающиеся на другие элементы. В отличие от традиционных линейных форм представления информации, гипертекст формирует сложную и многослойную структуру.

Дискурс может быть смоделирован с помощью гипертекста, если дискурс, состоящий из множества текстов, может быть определен как гипертекст. Интертекстуальность, превращающая корпус в дискурс, отображается через ссылки в гипертексте. Обычно гипертекст демонстрирует сложную синхронную информацию для пользователей и при лингвистическом анализе дискурса позволяет объединить текстовую связанность различных дискурсивных тематик с хронологической последовательностью текстов внутри системы.

Можно выделить два типа ссылочных данных.

1. Фиксирующие диахронические рецепции.

2. Отражающие синхронные тематические связи на разных уровнях дискурса.

Эти данные полезны для сравнительного анализа форм и содержания дискурса. Однако важнейшим является моделирование интертекстуальности как гипертекста. Интертекстуальность становится не просто теоретической концепцией, а важным компонентом анализа текстов. Генезис семантических фигур, моделей аргументации или метафор можно изучать на основании их первоначального появления и восприятия, что позволяет проследить их распространение и функции в дискурсе.

Изучая интертекстуальные ссылки, которые усиливаются в модели рецепции, можно высказывать гипотезы о социальных условиях создания текста. Положительные ссылки и исключение противоположных позиций помогают делать выводы о социальных группах через общие структуры установок. Кроме того, моделирование генезиса дискурса через гипертекст облегчает аргументативно-аналитический подход к исследованию, систематизируя формы интертекстуальной референции. Гипертекстовые системы не только создают и используют сетевую лингвистическую текстовую информацию, но и связывают ее с другими формами представления, такими как изображения и видео. Этот метод репрезентации соответствует концепции культуры как текстовой сети, разработанной в современной культурной антропологии.

Список использованных источников

1. *Арутюнова, Н. Д.* Язык и мир человека / Н. Д. Арутюнова. — М.: Языки русской культуры, 1999. — 905 с.
6. *Busse, D.* Historische Semantik : Analyse eines Programms. Stuttgart, Klett-Cotta, 1987. — 334 S. (Sprache und Geschichte).
7. *Busse, D., Teubert, W.* Ist Diskurs ein sprachwissenschaftliches Objekt? Zur Methodenfrage der historischen Semantik. In: Busse, Hermanns, Teubert (Hg.), 1994. — S. 10–28.
8. *Busse, D.* Das Eigene und das Fremde. Annotationen zu Funktion und Wirkung einer diskurssemantischen Grundfigur. In: Jung, Wengeler, Böke (Hg.), 1997. — S. 17–35.
12. *Gloy, K.*1998: Ethik-Diskurse. Praktiken öffentlicher Konfliktaustragung. Skizze eines Forschungs vorhabens. (Ethik-Diskurse. Praktiken öffentlicher Konfliktaustragung. Arbeitspapier Nr. 1). Universität Oldenburg.
13. *Hermanns, F.* 1991: "Umwelt". Zur historischen Semantik eines deontischen Wortes. In: Dietrich Busse (Hg.): Diachrone Semantik und Pragmatik. Untersuchungen zur Erklärung und Beschreibung des Sprachwandels. Tübingen. S. 235–257.
14. *Hermanns, F.* 1995a: Sprachgeschichte als Mentalitätsgeschichte. Überlegungen zu Sinn und Form und Gegenstand historischer Semantik. In: Andreas Gardt, Klaus Mattheier, Oskar Reichmann (Hg.): Sprachgeschichte des Neuhochdeutschen. Gegenstände, Methoden, Theorien. Tübingen. S. 69–101.
15. *Hermanns, F.* 1995b: Kognition, Emotion, Intention. Dimensionen lexikalischer Semantik. In: Gisela Harras (Hg.): Die Ordnung der Wörter. Kognitive und lexikalische Strukturen. Berlin, New York. S. 138–178.
16. *Jung, M.* 1996: Linguistische Diskursgeschichte. In: Böke, Jung, Wengeler (Hg.), S. 453–472.
17. *Jung, M., Wengeler, M.* 1999: Wörter — Argumente — Diskurse. Was die Öffentlichkeit bewegt und was die Linguistik dazu sagen kann. In: Gerhard Stickel (Hg.): Sprache — Sprachwissenschaft — Öffentlichkeit. Berlin, New York, S. 143–171.
19. *Stötzel, G.* 1995: Einleitung. In: Stötzel, Wengeler u.a., S. 1–17.
20. *Teubert, W.* 1998a: Eigentum, Arbeit, Naturrecht. Schlüsselwörter der Soziallehre im Wandel. In: Heidrun Kämper, Hartmut Schmidt (Hg.): Das 20. Jahrhundert. Sprachgeschichte — Zeitgeschichte. Tübingen. S. 188–224.

21. *Teubert, W.* 1998b: Korpus und Neologie. In: Wolfgang Teubert (Hg.): Neologie und Korpus. Tübingen. S. 129–170.
25. *Wengeler, M.* 1997a: Argumentation im Einwanderungsdiskurs. Ein Vergleich der Zeiträume 1970–1973 und 1980–1983. In: Jung, Wengeler, Böke (Hg.), S. 121–149.
26. *Wengeler, M.* 1997b: Vom Nutzen der Argumentationsanalyse für eine linguistische Diskursgeschichte. Konzept eines Forschungsvorhabens. In: Sprache Und Literatur In Wissenschaft Und Unterricht 80, S. 96–109.

FOREIGN DISCURSIVE SCHOOLS: GERMAN DISCURSIVE PRACTICES AND PROSPECTS FOR THEIR USE

***Akhmetov A. A.**, postgraduate student, Kazan Federal University, Kazan*

***Tagirov R. M.**, senior lecturer, Department of Theory and Practice of Teaching Foreign Languages, Institute of Philology and Intercultural Communication, Kazan Federal University, Kazan*

This article addresses the concept of discourse in the linguistic practices of German schools of discourse analysis and explores the prospects of applying these approaches in light of modern technological advancements. The authors provide definitions of the concept of discourse within the framework of these schools (Heidelberg or Mannheim, D sseldorf, and Oldenburg linguistic schools), list the key representatives of these 20th-century movements, and critically highlight the main ideas that appear promising at present.

discourse analysis • German linguistic schools • computational linguistics • corpus • argumentation.

Основные корпуса немецкого и русского языков: обзор и перспективы

***Агеева В. Е.**, аспирант, МГУ им. М. В. Ломоносова, Москва*

В статье представлен анализ одиннадцати основных корпусов текстов для немецкого и русского языков, включая шесть немецкоязычных и пять русскоязычных корпусов. Рассматриваются их структура, особенности разметки, объёмы данных и области применения в лингвистических исследованиях и разработке речевых технологий. Внимание уделяется значимости корпусов для задач обработки естественного языка (NLP). Делается акцент на репрезентативности данных и потенциале корпусов для сопоставительных исследований.

корпуса • немецкий язык • русский язык • речевые технологии • обзор

ВВЕДЕНИЕ

Современные речевые технологии стремительно развиваются благодаря достижениям в области обработки естественного языка (Natural Language Processing, NLP) и искусственного интеллекта. Одним из ключевых элементов, способствующих этому прогрессу, являются текстовые, звуковые, устные корпусы (Богданова-Бегларян и др. 2013: Dobrushina at al. 2022), которые служат основой для обучения моделей, анализа языковых структур и разработки инновационных решений.

Корпусы немецкого и русского языков играют важную роль в создании и совершенствовании речевых технологий, предоставляя данные для исследования особенностей различных языков и разработки приложений, таких как системы машинного перевода, голосовые помощники, инструменты для анализа текста и автоматического синтеза речи.

Немецкий и русский языки существенно различаются по грамматическим, морфологическим и фонетическим характеристикам, что создает уникальные вызовы и возможности при их изучении и обработке. Для эффективного использования этих языков в речевых технологиях необходимы качественные и репрезентативные корпуса, включающие письменные и устные данные, отражающие различные стили, жанры и контексты использования.

Корпусы текстов играют ключевую роль в современном лингвистическом исследовании, предоставляя исследователям мощный инструмент для анализа языковых данных в различных аспектах — от морфологии и синтаксиса до прагматики и семантики. Особенно важным является использование корпусов для изучения и сопоставления языков с различными структурными особенностями, таких как немецкий, русский языки и др. (Sharoff at al., 2023; Muneer, Nawab, 2022).

В современной лингвистике для решения значительного числа задач, связанных с употреблением слов, и других используются электронные корпуса текстов (Melikyan at al., 2023).

Чтобы определить, что такое корпус, приведем определение, взятое из учебника М.В. Копотева «Введение в корпусную лингвистику»: «Корпус в современной лингвистике в отличие от любого набора текстов может быть более точно определен как ограниченный по объему набор электронных текстов, собранных с целью максимально точно представлять исследуемый вариант языка» (определение взято автором учебника из книги McEnery T., Wilson A. Corpus linguistics. Edinburgh: Edinburgh University Press, 1996) (Копотев, 2014).

Для лингвистических исследований корпусу необходима разметка или аннотация, т.е. сведения о текстах, свойствах входящих в них слов и другие параметры. Самые частые виды разметки — метаданные и морфологическая.

В своем учебнике М.В. Копотев также приводит основные свойства корпуса, на список которых мы будем опираться при создании обзора избранных корпусов.

1. Репрезентативность — какие тексты входят в корпус, каких жанров, сбалансированность по жанрам.

2. Объем корпуса — обычно выражается в количестве словоупотреблений.

3. Наличие и вид разметки (морфологическая, синтаксическая и др.).

В данном обзоре будут задействованы 11 одноязычных корпусов (шесть на немецком языке, пять — на русском). Ниже приведен список корпусов, ссылки на которые даны в разделе источников.

Немецкий язык

- Deutsches Referenzkorpus (COSMAS)
- Digitale Wörterbuch der deutschen Sprache (DWDS)
- Deutsche Textarchiv (DTA)
- Wörterbuchnetz
- Wortschatz. Leipzig
- Datenbank für Gesprochenes Deutsch (DGD)

Русский язык

1. Национальный корпус русского языка
2. Корпус русского литературного языка
3. Russian Corpora in Tübingen
4. RuTenTen - Corpus of Russian Web
5. Корпус с параллельной синтаксической разметкой

КОРПУСА НЕМЕЦКОГО ЯЗЫКА

Deutsches Referenzkorpus (COSMAS)

Известен также как Мангеймский корпус немецкого языка и DeReKo (сокращение от Deutsches Referenzkorpus). Основной немецкоязычный корпус, ориентированный на максимальный охват текстов, включает в себя более ста подкорпусов и более 1846 млн словоупотреблений. Корпус постоянно обновляется, поскольку предназначен для изучения синхронного немецкого языка. В корпусе представлены различные типы текстов: научная и научно-популярная литература, большое количество газетных текстов, художественная проза и др. Представляет собой самое большое в мире электронное собрание современных немецкоязычных текстов начиная с 1956 г.

Данный корпус содержит шесть основных подкорпусов:

- Archiv der geschriebenen Sprache — корпус письменной речи (1.370.766.704 сл.-уп.), включает в себя подкорпуса по странам и жанрам литературы (например, подкорпус беллетристики);
- N-öffentlich — alle öffentlichen Neuakquisitionen — корпус новых поступлений (181.319.744 сл.-уп.);
- Archiv der phasengegliederten Wendecorpora — корпус речи времен объединения Германии (3.813.688 сл.-уп.), поделен на две части, относящиеся соответственно к ФРГ (Запад) и ГДР (Восток) — 1989–1990 гг.;
- Kartei der Gesellschaft für deutsche Sprache — картотека общества немецкого языка;
- Archiv der historischen Korpora — исторический корпус (6.296.361 сл.-уп.), содержит также корпуса произведений значимых авторов, таких как Карл Маркс, Фридрих Энгельс и братья Гримм;
- Archiv morphosyntakt. annotierter Korpora — морфосинтаксически размеченный корпус (29.695.096 сл.-уп.), включает также подкорпуса газет.

Исследователю, только начинающему работать с корпусом, стоит прочитать FAQ, где в формате ответов на вопросы перечисляется, чем является COSMAS, чем он не является (поисковиком) и как с ним работать.

Например, выдержка из первого же раздела, Was ist COSMAS II?, что такое COSMAS II:

Suchanfragen werden u.a. gebildet aus

- Wörtern, Teilwörtern, Wortgrundformen,
- Wortklassen (z.B. Verb, Artikel) und grammatikalischen Mustern,
- Angaben zu Wort- und Satzabstand,
- Angaben zu Textbereichen (z.B. Überschriften) und Position (z.B. erstes Wort eines Satzes).

Ergebnisse können u.a.

- nach Entstehungszeit, Erscheinungsland und Thematik sortiert werden,
- durch Statistiken auf häufig verwendete Gebrauchsmuster hin analysiert werden.

Запросы могут состоять из:

- слов, частей слов, лексем,
- частей речи (например, глагол, артикль) и грамматических моделей,
- информации о расстоянии между словами и предложениями,
- информации о текстовых областях (например, заголовках) и позиции (например, первое слово предложения).

Результаты могут быть:

- отсортированы по времени, стране и тематике,
- проанализированы с помощью статистики часто используемых шаблонов.

Чтобы пользоваться поиском по корпусу, необходимо зайти на стартовую страницу (https://cosmas2.ids-mannheim.de/cosmas2-web/) и зарегистрироваться в системе. После этого последовательно нужно выбрать архив и подкорпус (не все из них доступны) и можно искать необходимое. Поиск расширенный, и поэтому сначала придется разобраться в обозначениях: например, тег "direkt nacheinander" — слово стоит сразу после заданного, тег "direkt /+w1 nacheinander" — добавляет между ними еще одно слово, запрос со звездочкой после заданного слова выдает различные формы и производные этой лексемы. Далее можно посмотреть сам текст, частоту вхождений по годам и коллокации (Kookkurenz).

Digitale Wörterbuch der deutschen Sprache (DWDS)

DWDS — это огромный словарь, объединенный с корпусом текстов, который охватывает период с 1600 г. по настоящее время. Этим сервисом можно пользоваться как и обычным словарем, и как корпусом, поскольку в словарных статьях присутствуют ссылки на корпус. Разработчиком корпуса является Берлинско-Бранденбургская академия наук. База данных постоянно пополняется. Часть корпусов доступна сразу, часть — после регистрации. На главной странице можно зайти в отделы словаря или корпуса, посмотреть статистику, поиграть в слова и даже почитать отдельный словник про футбол. Кроме того, там имеется статья дня.

Ввиду большого количества корпусов (около полусотни) они разбиты на группы. Сюда входят:

- метакорпуса (объединение нескольких для удобства использования, например, общий современный со свободным доступом (Gegenwartskorpora mit freiem Zugang) и общий исторический

(Historische Korpora). Приведем примеры подкорпусов: Archiv der Gegenwart (1931–2000) (современный), A. v. Humboldts Publizistik (dt., 1790–1859) (исторический);
- справочные корпуса — корпуса, сбалансированные по времени и типологическому распределению;
- газетные корпуса — тексты из крупных ежедневных и еженедельных газет, оцифрованные или созданные исключительно в цифровом формате. Например, Berliner Zeitung (1945–1993);
- веб-корпуса;
- специальные корпуса — корпуса, составленные для конкретных языковых областей, например по времени (ГДР).

Система поиска такая же, как в COSMAS, необходимо знать специальные обозначения, упомянутые выше. Все корпуса содержат метаданные о типах текстов, авторе, названии, издателе и дате публикации. Содержащиеся в них слова лемматизируются и снабжаются частью речи.

Deutsche Textarchiv (DTA)

DTA предоставляет на своей платформе междисциплинарные и межжанровые коллекции и корпуса текстов на немецком языке. Основной корпус DTA включает около 1500 наименований и служит основой справочного корпуса нововерхненемецкого языка.

Для корпуса характерны: сбалансированный выбор типов текста, период публикации с начала XVII в. и до начала XX в., использование первых изданий при оцифровке, адекватное структурирование текста при транскрипции.

С главной страницы архива можно перейти на справочную страницу (Hilfe zur Suche in den Korpora des Deutschen Textarchivs mit der linguistischen Suchmaschine DDC), где подробно описаны приемы лингвистического поиска, аналогичного поиску в предыдущих корпусах. Примеры: ‘слово’ без дополнительных знаков — лексема, поиск выдает все формы и производные слова; @‘слово’ — точная форма слова. Искать можно как в общем архиве, так и фильтровать по корпусам, времени, авторам и отдельным произведениям. Тексты в DTA автоматически анализируются по частям речи.

Ниже приведем небольшой обзор корпусов, входящих в DTA.

- DTA-Kernkorpus — основной корпус, в который входят произведения немецких авторов за указанный выше период (XVII–XX вв.);
- DTAE-Einzeltexte — Erweiterungen des Deutschen Textarchivs — отдельные тексты и небольшие сборники, дополняющие основной корпус;
- например, в расширения входит собрание лекций Александра фон Гумбольдта (Alexander von Humboldts Kosmos-Vorträge) и сборник текстов немецких феминисток (Frauenstudium).

Wörterbuchnetz

Wörterbuchnetz — это собрание различных словарей немецкого языка, оцифрованных и снабженных перекрестными ссылками. На странице поиска можно легко переключаться между словарями.

Словари объединены по разделам:

- Enzyklopädien, например Oekonomische Encyklopädie von Johann Georg Krünitz (1773 bis 1858);
- Wörterbücher zur Sprachgeschichte (исторические словари), например Deutsches Wörterbuch von Brüder Grimm (словарь братьев Гримм);
- Althochdeutsches Wörterbuch — словарь древневерхненемецкого;
- Goethe-Wörterbuch — словарь Гёте;
- Regionalsprachliche Wörterbücher — диалектные словари, например Rheinisches Wörterbuch (словарь рейнского диалекта).

Wortschatz. Leipzig

Собрание более тысячи моноязычных словарей различных языков. В словарь немецкого языка входят тексты новостей за 2008–2023 гг., вебтексты из различных немецкоговорящих стран (2002–2019) и статьи из немецкого сегмента Википедии (2018 и 2021).

На странице с результатами можно посмотреть следующие сведения о заданном слове:

- Artikel (артикль);
- Wortart (часть речи);
- Grundform von (начальная форма от);
- Kompositum (состав, если составная словоформа);
- Beschreibung (толкование);
- Synonym (синонимы);
- Dornseiff-Bedeutungsgruppen (предметные группы, в которые входит слово);
- Beispiele (примеры употребления из корпуса);
- Signifikante Nachbarn (частотный правый и левый контекст);
- Wortgraph (семантическое поле в виде графа).

Datenbank für Gesprochenes Deutsch (DGD)

Последний немецкоязычный корпус в этой подборке отличается от предыдущих каналом восприятия. Этот корпус содержит аудио- и видеозаписи, по тематике подкорпус делится на «диалектные» и «разговорные», последние, как правило, определяются областью исследования (например, научный разговорный немецкий) и периодом записи разговорной речи. Тексты в корпусе расшифрованы, содержат транскрипты, которые синхронизированы с записью, а также списки слов и лемм.

Для использования корпуса необходима бесплатная авторизация, где обязательно нужно указать область исследования, для которой нужен корпус. После авторизации можно увидеть список подкорпусов таблицей, с перечислением числа записей, числа говорящих и периодов, к которым относятся записи. Например, в этом списке есть Australiandeutsch (австралийский немецкий) — 1966–1973 гг., небольшой подкорпус Mehrsprachige Kinder im Vorschulalter (мультиязычные

дети дошкольного возраста) — 2005–2006 гг., и Slawische Mundarten in Ruhrgebiet (славянские диалекты в Руре) — 1969 г.

Для общих исследований в области морфологии и синтаксиса могут быть полезны следующие подкорпуса DGD.

1. Gesprochene Wissenschaftssprache Kontrastiv — корпус научного разговорного немецкого: создан для сравнительных исследований, содержит 276 записей в двух жанрах — лекция и экзамен, полностью расшифрован, есть список лемм и теги.

2. Forschungs- und Lehrkorpus Gesprochenes Deutsch — корпус разговорного немецкого для исследований и преподавания: имеет множество жанров, поскольку перед проектом стояла задача достичь максимального распределения по параметру «тип взаимодействия» — 485 аудиозаписей и 278 видео, полностью расшифрован, ведется с 2008 по настоящее время, помимо лемм и тегов есть частеречная разметка.

3. Grundstrukturen: Freiburger Korpus — фрайбургский корпус основных структур: предназначен для стилистических исследований и содержит 222 записи стандартного или близкого к стандартному разговорного языка, охвачены различные, в т.ч. достаточно сложные грамматические конструкции, такие как сослагательное наклонение (конъюнктив) и пассивный залог, в 82 стенограммах имеется интонационная разметка в дополнение к стандартной.

4. Deutsche Standardsprache: König-Korpus — корпус немецкого стандартного языка, созданный фрайбургским ученым Вернером Кёнигом, содержит записи произношения (начитанные тексты и списки слов) говорящих из 62 городов, распределенных по территории ФРГ, по юридическим причинам доступны только 43 записи 1975 г. с расшифровкой и лемматизацией.

5. Deutsch heute — немецкий сегодня: очень большой корпус (более 6 тыс. элементов), содержит записи чтения и интервью говорящих со всего немецкоязычного региона (Германия, Австрия, Швейцария, Южный Тироль, Люксембург, Восточная Бельгия, Лихтенштейн), в дополнительных материалах также приведена методика опроса, результаты исследований публикуются в Интернете в атласе произношения немецкого стандарта с 2011 г.

КОРПУСА РУССКОГО ЯЗЫКА

Национальный корпус русского языка

Национальный корпус русского языка, или НКРЯ, — как говорится на главной странице сервиса, это «представительная коллекция текстов на русском языке общим объемом более 2 млрд слов, оснащенная лингвистической разметкой и инструментами поиска».

На данный момент НКРЯ состоит из 16 подкорпусов, некоторые из которых имеют дальнейшие подуровни:

- основной корпус: содержит письменные прозаические тексты с 1700 г., сбалансированные по жанрам;

- газетные корпуса: центральные СМИ и региональные, корпус центральных СМИ самый большой в НКРЯ, с 1983 по 2021 г.;
- синтаксический корпус: содержит специальную подробную морфосинтаксическую разметку, состоит из журнальных статей научно-популярной и общественной тематики и художественной прозы второй половины XX — начала XXI в.;
- социальные сети: включает в себя тексты из открытых источников (кроме заблокированных, разумеется) — блоги и мессенджеры;
- корпус устной речи: содержит расшифровки магнитофонных записей, а также транскрипты фильмов и спектаклей, помимо стандартной имеется социологическая разметка со сведениями о говорящих, звуковой материал недоступен;
- акцентологический корпус: размечен с точки зрения постановки ударения и интонаций;
- мультимедийный корпус: включает в себя кинофильмы и записи речи (публичной, непубличной, театральной), прилагающиеся к ним транскрипты, а также небольшой фрагмент корпуса содержит отметки о типе речевых действий и жестикуляции;
- мультиПАРКи — два мультимедийных параллельных корпуса, в русском сегменте можно сравнить различные постановки одной и той же пьесы (например, «Вишневый сад» А.П. Чехова в четырех вариантах), в англо-русском — сравнить оригинальный текст и дубляж;
- диалектный — записи диалектной речи из различных регионов России, полностью сохранена морфологическая, синтаксическая и лексическая специфика текстов, большая их часть представлена в фонологизированной записи с ударениями;
- обучающий — разметка адаптирована к школьной программе;
- от 2 до 15 — детское и подростковое чтение (отмечен возраст);
- поэтический — содержит дополнительную стиховедческую разметку (метр и ритм);
- русская классика — XIX — начало XX в.;
- исторические: древнерусский, эпиграфика, берестяные грамоты, старорусский, церковнославянский;
- панхронический — объединение основного корпуса с историческим, позволяющее проследить историю лексем.

Разметка: все тексты основного корпуса содержат:

- метаразметку (информацию о тексте, вплоть до жанра, аудитории и типе разметки (автоматическая или ручная));
- морфологическую разметку со снятием омонимии;
- синтаксическую разметку, позволяющую строить дерево зависимостей;
- семантическую разметку с системой тегов.

Кроме примеров употребления на странице поиска можно также посмотреть портрет слова, включающий в себя:

- начальную форму (лемму);
- грамматические признаки;
- семантические теги;
- частотность слова;

- морфемный разбор;
- однокоренные слова;
- формы слова;
- похожие слова.

Корпус русского литературного языка

Корпус русского литературного языка — более узконаправленный корпус, чем НКРЯ, поскольку содержит только официальные письменные тексты и ориентируется на литературную норму. Доходит до примерно 100–150 млн слов. Все тексты имеют проставленные ударения и букву «ё». Содержит морфологическую разметку. На базе корпуса создан частотный словарь словоформ русского языка. В систему лексико-грамматических классов добавлены такие, как вводные слова, аналитические прилагательные (*интернет*-кафе), связанные слова (*тёмно*-красный), служебные слова (самый в суперлативе). В корпус включены тексты со сбалансированным жанровым составом с 50-х годов XX в. по настоящее время.

Russian Corpora in Tübingen

Тюбингенский корпус основан на Уппсальском корпусе отделения славистики шведского Уппсальского университета (по которому также был создан частотный словарь) и дополнен текстами интервью. Часть текстов содержат морфологическую разметку. К сожалению, на странице корпуса большинство ссылок не работает, но корпус можно скачать по ссылке «Другие корпуса» (https://ruscorpora.ru/page/corpora-other/) на сайте НКРЯ.

RuTenTen — Corpus of Russian Web

Корпус RuTenTen принадлежит к семейству корпусов TenTen и содержит интернет-тексты на русском языке объемом более 10 млрд слов.

В корпусе используется особая система морфологических тегов, где каждый атрибут обозначается одной буквой и стоит на определенном месте в коде. Например, код Vmip3p-m-e- можно расшифровать буква за буквой следующим образом: Verb, main, indicative, present, third person, plural, прочерк на месте рода, media, прочерк на месте определенности/неопределенности, perfective, прочерк на месте падежа. В итоге по тегу можно найти такой вариант, как «(фильмы) смотрятся».

Кроме того, в корпусе доступные такие инструменты, как коллокации, синонимы, частотные списки слов, конкорданс и метаданные.

Корпус с параллельной синтаксической разметкой

Это проект кафедры теоретической и прикладной лингвистики МГУ, он содержит банк синтаксического разбора и деревьев 64 800 предложений (1 млн словоупотреблений) автоматическими системами для синтаксического разбора. Кроме того, 800 случайных предложений были размечены вручную. Корпус содержит научные, художественные и публицистические тексты. Ресурс создан для отработки синтаксической разметки, поиска сложных мест и позволяет выявить наиболее надежно устанавливаемые синтаксические связи.

Роль корпусов в развитии речевых технологий

Следует отметить потенциал использования корпусов прежде всего в сфере машинного перевода. Корпусы параллельных текстов, такие как OpenSubtitles или Europarl: A Parallel Corpus, позволяют обучать модели для автоматического перевода между немецким и русским языками. Использование нейросетевых архитектур, таких как трансформеры, требует больших объемов данных, предоставляемых этими корпусами.

Значительную роль играют корпуса разных языков в синтезе и распознавании речи. Звуковые корпуса, содержащие записи речи, аннотации и транскрипции, необходимы для обучения моделей синтеза и распознавания речи. Например, использование устных корпусов помогает повысить точность голосовых помощников, таких как Alexa или Siri, в понимании и генерации речи на немецком и русском языках.

Корпусы с лексической и синтаксической разметкой позволяют разрабатывать модели для анализа текста, включая определение тональности, извлечение информации и кластеризацию данных. Это особенно актуально для создания чат-ботов и инструментов анализа социальных медиа.

Корпусы текстов способствуют разработке технологий для людей с ограниченными возможностями, например автоматических субтитров или систем преобразования текста в речь.

ЗАКЛЮЧЕНИЕ

В данной статье были описаны различные корпуса русского и немецкого языков. Количество информации и величина обзора на каждый корпус зависели от степени разработанности корпуса, например большинство приведенных корпусов русского языка содержат мало данных и не все ссылки доступны, хотя эти корпуса относятся к числу самых известных. В больших корпусах, таких как НКРЯ и все немецкоязычные, помимо разметки, объема и жанрового состава были описаны также подкорпуса (целиком или частично).

Каждый из описанных корпусов подходит для разных типов исследований в зависимости от формата (письменный или устный), наличия разметки разного рода и репрезентативности.

Корпуса немецкого и русского языков представляют собой неоценимый ресурс для развития речевых технологий. Их использование позволяет преодолевать языковые барьеры, улучшать взаимодействие человека и машины, а также расширять доступность технологий для пользователей по всему миру. Будущие исследования должны быть направлены на расширение и улучшение существующих корпусов, а также на разработку новых подходов к их интеграции с современными алгоритмами.

Список использованных источников

1. *Богданова-Бегларян Н. В., Баева Е. М., Бродт И. С., Павлова О. В.* Звуковой корпус как материал для анализа русской речи. Отв. ред. — д. ф. н., профессор Наталья Викторовна Богданова-Бегларян. Часть 1. Чтение. Пересказ. Описание (СПб.: Филологический факультет СПбГУ, 2013. — 532 с.); 2) Часть 2. Теоретические и практические аспекты анализа. Том 1. О некоторых особенностях устной спонтанной речи разного типа. Звуковой корпус как материал для преподавания русского языка в иностранной аудитории (СПб.: Филологический факультет СПбГУ, 2014. — 396 с.); 3) Часть 2. Теоретические и практические аспекты анализа. Том 2. Звуковой корпус как материал для новых лексикографических проектов (СПб.: Филологический факультет СПбГУ, 2015. — 364 с.)
2. *Копотев, М.* Введение в корпусную лингвистику. — Прага : Animedia Company, 2014. — 195 с. [эл. изд.]
3. *Dobrushina, N., Sokur, E.* Spoken Corpora of Slavic Languages. *Russ Linguist* 46, 77–93 (2022). https://doi.org/10.1007/s11185-022-09254-9
4. *Melikyan, A., Nikolaeva, E., Plotnikova, N., Babloyan, A.* (2023). Language Corpora: Obtaining Data for Modern Linguistic Research and Solving Translation Problems. In: Beskopylny, A., Shamtsyan, M., Artiukh, V. (eds) XV International Scientific Conference "INTERAGROMASH 2022". INTERAGROMASH 2022. Lecture Notes in Networks and Systems, vol 574. Springer, Cham. https://doi.org/10.1007/978-3-031-21432-5_77
5. *Muneer, I., Nawab, R.M.A.* Develop corpora and methods for cross-lingual text reuse detection for English Urdu language pair at lexical, syntactical, and phrasal levels. *Lang Resources & Evaluation* **56**, 1103–1130 (2022). https://doi.org/10.1007/s10579-022-09613-4
6. *Sharoff, S., Rapp, R., Zweigenbaum, P.* (2023). Building Comparable Corpora. In: Building and Using Comparable Corpora for Multilingual Natural Language Processing. Synthesis Lectures on Human Language Technologies. Springer, Cham. https://doi.org/10.1007/978-3-031-31384-4_3

Интернет-ресурсы

1. Deutsches Referenzkorpus (COSMAS) — https://www2.ids-mannheim.de/cosmas2/
2. Digitale Wörterbuch der deutschen Sprache (DWDS) — https://www.dwds.de/r
3. Deutsche Textarchiv (DTA) — https://www.deutschestextarchiv.de/
4. Wörterbuchnetz — https://www.woerterbuchnetz.de/
5. Wortschatz. Leipzig — https://corpora.uni-leipzig.de/de?corpusId=deu_news_2023
6. Datenbank für Gesprochenes Deutsch (DGD) — https://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome
7. Национальный корпус русского языка — https://ruscorpora.ru/
8. Корпус русского литературного языка — https://narusco.ru/
9. Russian Corpora in Tübingen — https://www.lingexp.uni-tuebingen.de/sfb441/b1/en/korpora.html
10. RuTenTen — Corpus of Russian Web — https://www.sketchengine.eu/rutenten-russian-corpus/
11. Корпус с параллельной синтаксической разметкой — http://otipl.philol.msu.ru/~soiza/testsynt/

THE MAIN CORPORA OF GERMAN AND RUSSIAN LANGUAGES: OVERVIEW AND PROSPECTS

Ageeva Valeriia Evgenievna, *PhD Student, Lomonosov Moscow State University, Moscow*

The article provides an analysis of eleven key corpora for the German and Russian languages, including six German-language and five Russian-language corpora. It examines their structure, annotation features, data volumes, and areas of application in linguistic research and the development of speech technologies. Special attention is given to the importance of corpora for natural language processing (NLP) tasks. The focus is on the representativeness of the data and the potential of the corpora for cross-linguistic comparison.

corpora • German language • Russian language • speech technologies • review

CONTENTS

ISS представляет уникальное решение, позволяющее автоматически выявлять нарушения производственных регламентов и другие инциденты, связанные с влиянием человеческого фактора.

Система анализа поведения людей на соответствие установленным правилам и регламентам SecurOS SBA (Semantic Behavior Analyzer)

Система собирает, обрабатывает и анализирует видеоданные, полученные от стереокамер; выдает оповещения при детекции действий или бездействии, нарушающих регламенты, а также опасных ситуаций; преобразует результаты анализа в интерактивные карточки событий и табличные отчеты, а также отображает статистику инцидентов в виде дашбордов (графиков и диаграмм).

Интерактивные карточки событий дают возможность специалистам служб безопасности и оперативного контроля незамедлительно реагировать на происшествия. Отчеты и дашборды предназначены для управляющих сотрудников и позволяют делать выводы о ситуации на конкретных участках или на предприятии в целом.

Архитектура решения

Решение создано на базе технологий стереозрения, нейросетевых алгоритмов и экспертных систем (семантического анализа).

Ключевые преимущества решения

- SecurOS SBA дает возможность учитывать поведенческие характеристики людей с привязкой к элементам сцены, а также причинно-следственные связи во взаимодействиях между людьми или человеком и объектом (-ами). Таким образом, глубина и точность анализа при использовании технологий SecurOS SBA значительно превышает возможности традиционной видеоаналитики, что существенно расширяет круг применений решения.
- Лингвистический процессор преобразует текст регламентов в работающие правила на встроенной экспертной системе, т.е. без необходимости создания специализированного программного обеспечения под каждый проект.

ISS — Интеллектуальные Системы Безопасности
Россия, Москва, ул. Мневники, дом 17
+7 (495) 645-2121
info@iss.ru | https://iss.ru