Отчет по комплексному лингвистическому анализу текстов

Валерия Лелик, кандидат на позицию «Психолингвист/исследователь»

июль 2025

Введение

Цель данного анализа — провести комплексное лингвистическое и психолингвистическое исследование отрывков из двух интервью, используя ряд метрик, оценивающих различные аспекты речи. Анализ направлен на выявление различий между текстами по списку параметров и формирование психолингвистических особенностей речи говорящих. Представленные данные могут быть полезны при разработке инструментов автоматического анализа текстов.

Предоставленные тексты интервью были сначала проанализированы с точки зрения содержания. В результате анализа мы сформулировали следующие гипотезы о данных текстах:

1. Первый текст был порожден сотрудником, не занимающим руководящую должность, второй текст – интервью руководителя. Таким образом, в полном датасете могут содержаться как минимум две группы текстов, основанных на иерархии говорящих в структуре компании;
2. Перед началом интервью говорящим давались две разные инструкции: в первом случае время было более ограничено, требовался сжатый рассказ, затрагивающий определенные пункты, озвученные интервьюером. Второй текст содержит более свободное повествование, следовательно, мог порождаться при менее строгой инструкции.

В результате мы можем получить возможность связывать особенности речи респондентов с их личностными характеристиками и определять по фрагментам речи психологический портрет человека. Так как метаинформация о говорящих неизвестна, далее в анализе мы будем опираться на лингвистические и психолингвистические параметры и делать предположения об интерпретации полученных результатов для метрик.

Часть 1: Расширенный лингвистический анализ

В данной секции мы представим отчет по базовым метрикам для текстов. Для каждой из метрик приведено описание, теоретическое обоснование выбранной метрики, а также визуализация результатов.

1.1 Базовые метрики

- Сложность текста (индексы читабельности)

Оценка читабельности текста является важным этапом лингвистического анализа, позволяющим количественно определить, насколько легко воспринимается текст читателем. В данном исследовании мы использовали три распространённых показателя: *Flesch Reading Ease, Gunnung Fog Index* и *Coleman Liau Index*.

Индекс *Flesch Reading Ease* (Flesch, 1948) рассчитывается на основе средней длины предложения и среднего количества слогов на слово. Индекс приобретает значения от 0 до 100: чем выше значение индекса, тем проще текст для восприятия. Тексты, получающие

значения выше 60 считаются легко читаемыми, а ниже 30 — трудными для восприятия. Индекс рассчитывается по следующей формуле:

$$FRE = 206.835 – 1.015 \times (\text{среднее число слов на предложение}) – 84.6 \times (\text{среднее число слогов на слово})$$

Flesch Reading Ease остается надежной и валидной широко используемой метрикой для оценки читабельности текстов, поэтому его применение релевантно для нашего анализа.

Gunning Fog Index (Индекс туманности Ганнинга) показывает, сколько лет формального образования требуется для понимания текста (Gunning, 1952). Рассчитывается по следующей формуле, где *ASL* – средняя длина предложения, *Complex Words* – слова с тремя и более слогами, *Total Words* – общее количество слов в тексте:

$$GFI = 0.4 \cdot (ASL + 100 \cdot (\text{Complex Words} / \text{Total Words}))$$

Результаты интерпретируются следующим образом: текст, набирающий значения 6-8 считается легко читаемым (так как требует 6-8 лет образования, соответствующие уровню средней школы). Чем выше значение индекса – тем больше лет образования требуется для восприятия текста. Этот индекс может быть особенно полезен при анализе научных, деловых и формальных текстов, где большое количество длинных слов может существенно повышать когнитивную нагрузку.

Coleman–Liau Index (Coleman & Liau, 1975) оценивает читаемость на основе количества букв и предложений, в отличие от двух предыдущих метрик, опирающихся на слоги. Индекс рассчитывается по формуле ниже, в которой L — среднее число букв на 100 слов, S — среднее число предложений на 100 слов:

$$CLI = 0.0588 \cdot L - 0.296 \cdot S - 15.8$$

Результат индекса также показывает количество лет образования, необходимое для понимания текста. Также общее количество букв может быть более информативным, чем большее количество слогов, так как помогает учитывать скопления согласных в специализированных словах, что может дать более надежный результат для определения сложности текста.

Приведем количественные результаты по индексам читабельности текстов для отрывков из двух интервью (Таблица 1), их визуализацию (Рисунок 1), а также статистику (Таблица 2):

Таблица 1. Результаты применения метрик читабельности к текстам двух интервью

text_id	**Text_1**	**Text_2**
flesch_reading_ease	100,0177778	109,7140185
gunning_fog_index	8,755555556	4,686956522
coleman_liau_index	9,742639594	10,75992579

Рисунок 1. Результаты применения метрик читабельности к текстам двух интервью

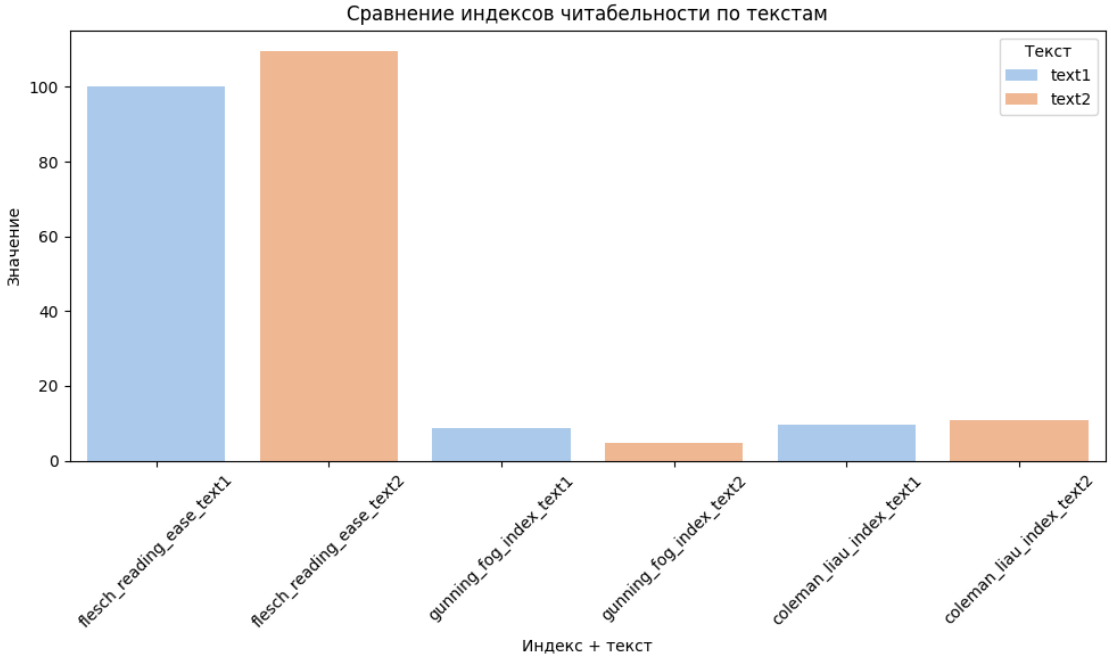

Таблица 2. Результаты применения метрик читабельности к текстам двух интервью, значения округлены до сотых долей (статистика)

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
flesch_reading_ease	2.0	104.87	6.86	100.02	102.44	104.87	107.29	109.71
gunning_fog_index	2.0	6.72	2.88	4.69	5.70	6.72	7.74	8.76
coleman_liau_index	2.0	10.25	0.72	9.74	10.00	10.25	10.51	10.76

По шкале *Flesch Reading Ease*, более высокое значение указывает на большую простоту восприятия текста. В нашем случае text_2 демонстрирует более высокий показатель (109.7 vs 100.0 у text_1), что говорит о его большей доступности для восприятия. Это может означать, что второй текст может быть более понятным для широкой аудитории.

Gunning Fog Index, отражающий количество лет образования, необходимое для понимания текста, показал, что text_1 имеет более высокое значение (8.76 vs 4.69 у text_2), что также указывает на большую сложность: такой текст потребует более высокого уровня образования для понимания.

Coleman–Liau Index опирается на среднюю длину слов и предложений в символах. Значения обоих текстов находятся в пределах около 10, однако text_2 показывает более высокий индекс (10.76 vs 9.74 у text_1), что может говорить о его большей сложности для восприятия.

Таким образом, совокупность двух показателей позволяет заключить, что text_2 — более прост для чтения, в то время как text_1 может восприниматься как более сложный для восприятия. Это отражается и в интуитивной логике при прочтении текста: второй текст относится скорее к разговорному стилю, а первый выстроен более формально.

- Метрики частотности слов

Для оценки частотности используемых слов были применены следующие показатели: *TTR (Type-Token Ratio), average frequency* (средняя частотность слов по корпусу) и *MTLD (Measure of Textual Lexical Diversity)*.

Type-Token Ratio (TTR) — это классический показатель, отражающий степень лексического разнообразия (Richards, 1987). Рассчитывается как отношение количества уникальных слов (types) к общему числу слов в тексте (tokens). Чем выше *TTR*, тем разнообразнее лексика текста.

TTR = Number of types / Number of tokens

Средняя частотность слов (*average word frequency*) — отражает, насколько часто встречающиеся слова используются в тексте. Частота слов в нашем исследовании определялась на основе выборки из Национального корпуса русского языка (НКРЯ). Выборка представляет собой таблицу, в которой указаны леммы и их частотность. Тексты с высокой средней частотностью содержат более частотную лексику, а низкие значения указывают на использование редких, специфичных или более профессиональных слов. Этот показатель является важным при оценке доступности текста для широкой аудитории.

avg_freq = среднее значение частот всех лемм, найденных в списке

MTLD (Measure of Textual Lexical Diversity) — современная метрика лексического разнообразия, предложенная McCarthy & Jarvis (2010). В отличие от *TTR, MTLD* менее чувствителен зависит к длине текста, так как измеряет среднюю длину последовательностей, в которых *TTR* остаётся выше заданного порога. *MTLD* считается одной из наиболее надёжных метрик лексического разнообразия в прикладной лингвистике и психолингвистике.

MTLD = количество завершённых фрагментов (моменты, когда TTR становится меньше определенного порога) / общее число слов

Приведем количественные результаты по метрикам лексического разнообразия текстов для отрывков из двух интервью (Таблица 3), их визуализацию (Рисунок 2), а также статистику (Таблица 4):

Таблица 3. Результаты применения метрик частотности к текстам двух интервью

text_id	**Text_1**	**Text_2**
TTR	0,441260745	0,451403888
average word frequency	5879,765973	4501,492484
MTLD	45,78603529	64,53744799

Рисунок 2. Результаты применения метрик частотности к текстам двух интервью

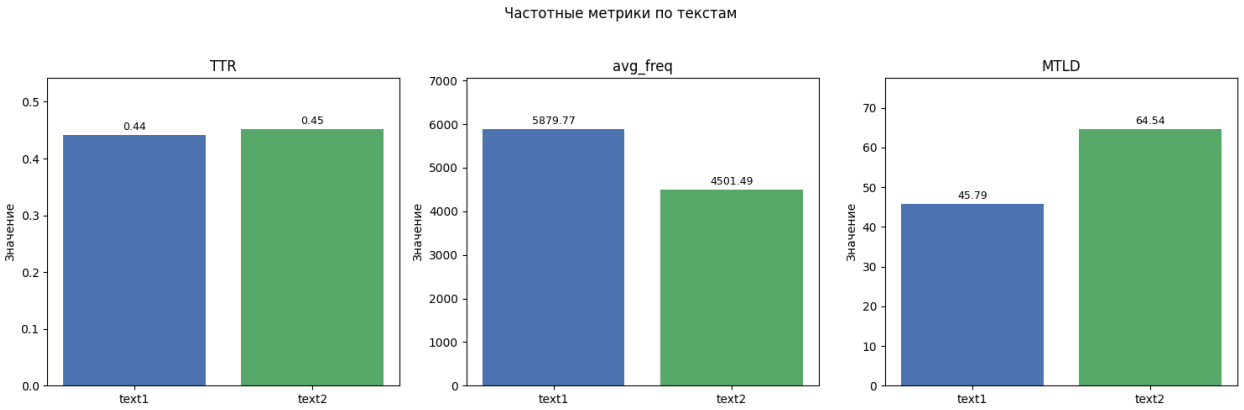

Таблица 4. Результаты применения метрик частотности к текстам двух интервью, значения округлены до сотых долей (статистика)

	count	mean	std	min	25%	50%	75%	max
TTR	2.0	0.45	0.01	0.44	0.44	0.45	0.45	0.45
avg_freq	2.0	5190.63	974.59	4501.49	4846.06	5190.63	5535.20	5879.77
MTLD	2.0	55.16	13.26	45.79	50.47	55.16	59.85	64.54

Таким образом, несмотря на схожие значения TTR, более высокая MTLD и более низкая средняя частота слов указывают на то, что text_2 обладает большим лексическим разнообразием. В то же время text_1 использует более частотную и, возможно, более знакомую лексику, что может делать его проще для восприятия.

- Self-focus индекс

Self-Focus Index — это метрика, отражающая степень самореференциальности текста, может измеряться как доля местоимений 1-го лица (например, я, мы, мой, наш) от общего числа слов. Высокие значения индекса могут интерпретироваться в контексте нашего исследования как маркер повышенного внимания автора к себе, личного опыта или вовлеченности в обсуждаемую тему. Предыдущие исследования, например, (Brockmeyer et al., 2015) исследовали взаимосвязь данной метрики в тексте с тревогой и депрессией.

В нашем анализе для подсчета мы использовали списки личных, возвратных и притяжательных местоимений первого лица в единственном и множественном числе, а затем делили количество найденных местоимений на общее количество слов текста.

Приведем количественные результаты по *Self-Focus Index* текстов для отрывков из двух интервью (Таблица 4), их визуализацию (Рисунок 3), а также статистику (Таблица 5):

Таблица 4. Результаты применения метрики Self-Focus Index к текстам двух интервью

text_id	Text_1	Text_2
self_focus	0,063037249	0,051835853

Рисунок 3. Результаты применения метрики Self-Focus Index к текстам двух интервью

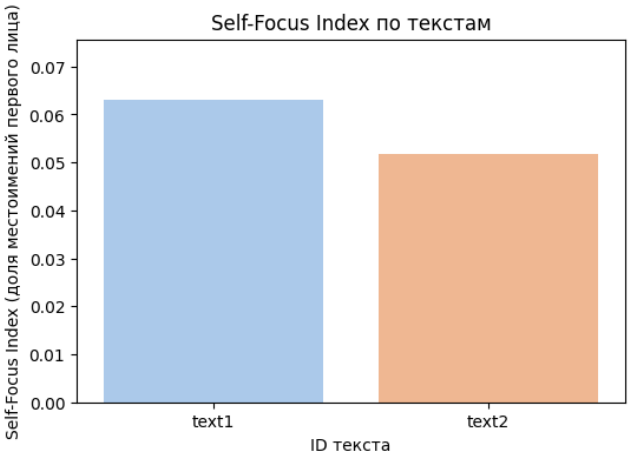

Таблица 4. Результаты применения метрики Self-Focus Index к текстам двух интервью, значения округлены до сотых долей (статистика)

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
self_focus	2.0	0.06	0.01	0.05	0.05	0.06	0.06	0.06

Text_1 демонстрирует более высокий индекс самофокусировки (6.3% местоимений 1-го лица vs 5.2% в text_2), что может свидетельствовать о большей степени личной вовлеченности автора, более субъективном стиле повествования. В контексте интервью это может указывать на то, что интервьюируемый в text_1 активнее ссылается на собственный опыт или позиционирует себя как участника описываемых событий. Напротив, text_2 может быть более нейтральным или ориентированным на внешние процессы, а не на внутренние переживания. Однако различия в проценте личных местоимений первого лица небольшое, что не дает возможность уверенно различать их по данному параметру.

- Соотношение абстрактных/конкретных понятий

Для анализа абстрактности/конкретности лексики использовался подход, основанный на частотном сопоставлении слов текста с эмпирическими оценками конкретности. Этот метод был предложен Brysbaert et al. (2014) для английского языка, где каждое слово получало балл конкретности по шкале от 1 (максимально абстрактное) до 5 (максимально конкретное), основываясь на восприятии носителей языка. Мы адаптировали этот подход для русского языка, используя словарь, разработанный коллективом авторов (Соловьев и др., 2022) В нем содержатся оценки конкретности/абстрактности для широкого набора существительных и прилагательных по шкале от -5 (более абстрактные слова) до 5 (более конкретные). Далее для каждого текста мы рассчитывали среднюю конкретность слов по списку лемм.

Приведем количественные результаты по *индексу конкретности* текстов для отрывков из двух интервью (Таблица 5), их визуализацию (Рисунок 4), а также статистику (Таблица 6):

Таблица 5. Результаты применения метрики конкретности к текстам двух интервью

text_id	Text_1	Text_2
concreteness_index	-0,863864072	-1,08062869

Рисунок 4. Результаты применения метрики конкретности к текстам двух интервью

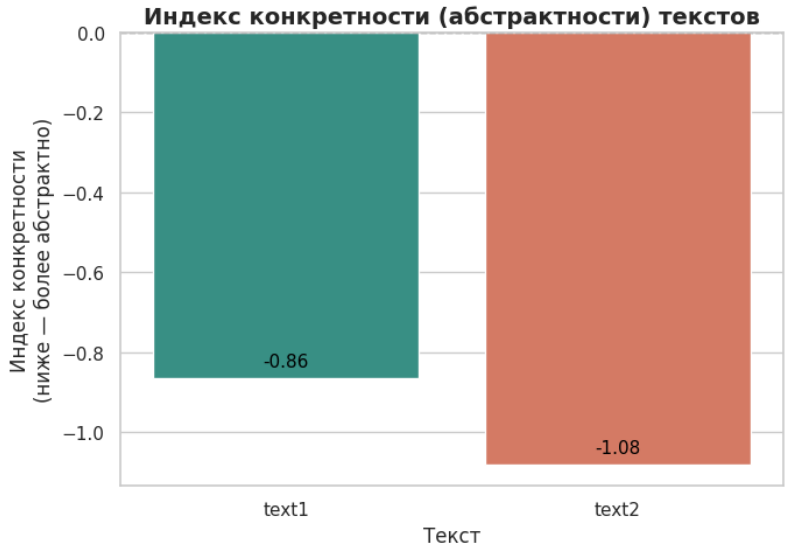

Таблица 6. Результаты применения метрики конкретности к текстам двух интервью, значения округлены до сотых долей (статистика)

	count	mean	std	min	25%	50%	75%	max
concreteness_index	2.0	-0.97	0.15	-1.08	-1.03	-0.97	-0.92	-0.86

Оба текста демонстрируют низкий индекс конкретности, что по нашей шкале означает большее использование авторами более абстрактных слов. Однако значение для первого текста ближе к нейтральному, что может указывать на более сбалансированный стиль между абстрактным и конкретным описанием в нем. Большая абстрактность слов второго текста может указывать на более «живое» и эмоциональное описание, так как абстрактные слова часто описывают чувства и эмоции.

1.2 Дополнительные метрики для анализа

В данной секции мы представим отчет по дополнительным метрикам для текстов. Для каждой из метрик приведено описание, теоретическое обоснование выбранной метрики, а также визуализация результатов.

- Сложность синтаксических конструкций

Для оценки синтаксической сложности текстов мы использовали две метрики: среднюю длину предложений и глубину вложенности.

Средняя длина предложений отражает среднее количество слов в одном предложении. Более длинные предложения, как правило, требуют от читателя большего когнитивного усилия, поскольку предполагают более сложную структуру высказывания и увеличивают

нагрузку на кратковременную память. Метрика является классическим индикатором синтаксической сложности и активно используется в психолингвистических исследованиях (Davis, 1937). Мы вычисляли среднюю длину предложений как сумму слов во всем тексте, деленную на количество предложений в нем.

Максимальная глубина вложенности синтаксических конструкций основана на анализе синтаксических деревьев зависимостей и отражает количество уровней вложенности (то есть глубину дерева). Высокая глубина может указывать на наличие вложенных придаточных предложений и других сложных конструкций, требующих дополнительных усилий для обработки. Для её оценки мы использовали синтаксический парсер библиотеки spaCy (Honnibal et al., 2020), вычисляя максимальную глубину вложенности зависимостей в каждом предложении.

Приведем количественные результаты по синтаксической сложности текстов для отрывков из двух интервью (Таблица 6), их визуализацию (Рисунок 5), а также статистику (Таблица 7):

Таблица 6. Результаты применения метрик оценки синтаксиса к текстам двух интервью

text_id	Text_1	Text_2
avg_sentence_length	349	154,3333333
syntax_depth	17	12

Рисунок 5. Результаты применения метрик оценки синтаксиса к текстам двух интервью

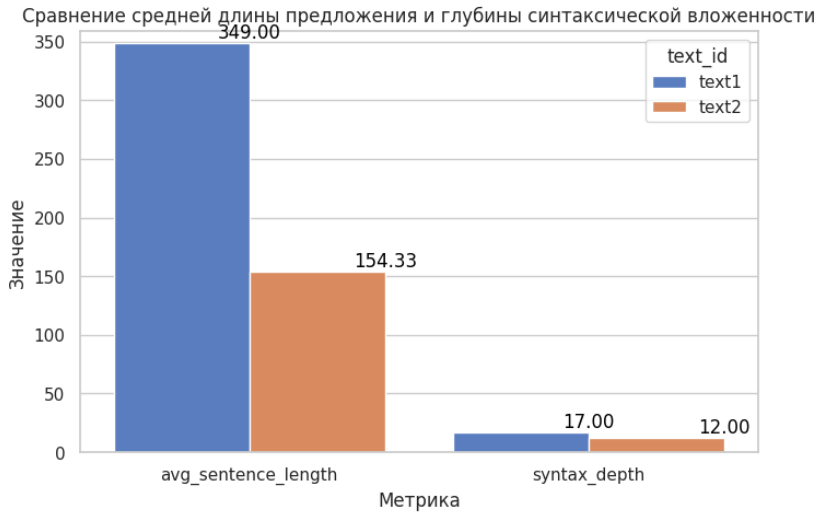

Таблица 7. Результаты применения метрик оценки синтаксиса к текстам двух интервью, значения округлены до сотых долей (статистика)

	count	mean	std	min	25%	50%	75%	max
avg_sentence_length	2.0	251.67	137.65	154.33	203.00	251.67	300.33	349.00
syntax_depth	2.0	14.50	3.54	12.00	13.25	14.50	15.75	17.00

Таким образом, text_1 демонстрирует значительно более высокую синтаксическую сложность как по длине предложений, так и по глубине вложенности. Однако длинные предложения часто могут восприниматься сложнее, что снижает внимание слушателя. Возможно, респондент, породивший первый текст, имеет меньше опыта публичных выступлений и менее развитые ораторские навыки.

- Temporal focus (прошлое, настоящее, будущее)

Temporal focus — это характеристика текста, отражающая направленность речевой деятельности говорящего на определённый временной план: прошлое, настоящее или будущее. В рамках данного анализа использован комбинированный подход: морфологический анализ глаголов производился с помощью библиотеки pymorphy2 (Korobov, 2015), которая позволяет извлекать грамматические признаки слов, в том числе вид, время, лицо и число. Далее мы искали временные наречия по заранее составленным спискам, отражающим принадлежность к прошлому, настоящему или будущему. Количество и доля таких наречий приведены отдельными значениями.

Приведем количественные результаты по временному фокусу текстов для отрывков из двух интервью (Таблица 7), их визуализацию (Рисунок 6), а также статистику (Таблица 8):

Таблица 6. Результаты применения метрик оценки temporal focus текстов двух интервью

text_id	Text_1	Text_2
past_verb_ratio	0,574468085	0,095238095
present_verb_ratio	0,170212766	0,53968254
future_verb_ratio	0,021276596	0,015873016
past_adverb_count	0	1
present_adverb_count	1	1
future_adverb_count	0	0
total_temporal_adverbs	1	2

Рисунок 5. Результаты применения метрик оценки temporal focus глаголов двух интервью

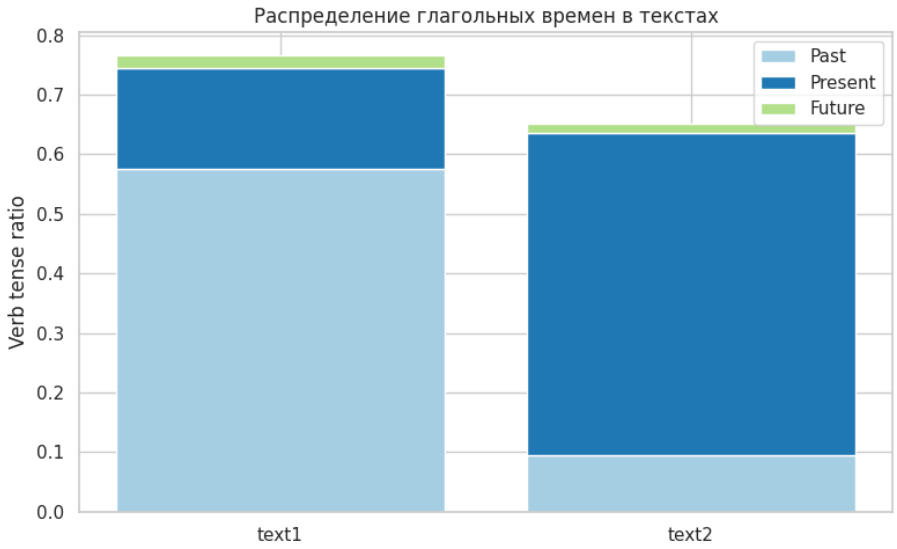

Таблица 7. Результаты применения метрик оценки temporal focus глаголов двух интервью, значения округлены до сотых долей (статистика)

	count	mean	std	min	25%	50%	75%	max
past_verb_ratio	2.0	0.33	0.34	0.10	0.22	0.33	0.45	0.57
present_verb_ratio	2.0	0.35	0.26	0.17	0.26	0.35	0.45	0.54
future_verb_ratio	2.0	0.02	0.00	0.02	0.02	0.02	0.02	0.02
past_adverb_count	2.0	0.50	0.71	0.00	0.25	0.50	0.75	1.00
present_adverb_count	2.0	1.00	0.00	1.00	1.00	1.00	1.00	1.00
future_adverb_count	2.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00
total_temporal_adverbs	2.0	1.50	0.71	1.00	1.25	1.50	1.75	2.00

Таким образом, text_1 можно охарактеризовать как ретроспективный, то есть описывающий уже завершённые события. Это может свидетельствовать о характере текста, где автор осмысляет прошлое. В свою очередь, text_2 демонстрирует актуальный фокус, описывая происходящее здесь и сейчас. Связывая данные характеристики с говорящими, можно предположить, что второй говорящий рассказывает о работе, которой обладает в момент взятия интервью, в то время как первый говорящий говорит о своей прошлой работе, вспоминая прошедшие события.

- Certainty vs uncertainty markers

Лексические маркеры уверенности, неуверенности и хезитации — показатели прагматической модальности, отражающие степень выраженной уверенности/сомнения в тексте. Для оценки выраженности модальных установок (уверенности и неуверенности) реализован подход, основанный на лексическом анализе, предложенный в работе Rubin et al. (2006). Метод основан на подсчёте частоты появления в тексте специализированных слов и выражений — маркеров: уверенности (например, "точно", "очевидно", "безусловно"), неуверенности (например, "возможно", "наверное"). В нашем анализе дополнительно

проводится оценка количества хезитаций (например, "э-э"). Для русского языка маркеры были адаптированы с учётом грамматических и семантических особенностей, а также проанализированных текстов. Частотность маркеров нормализуется на общее количество слов в тексте, что позволяет получить сравнимые количественные показатели для разных текстов. Таким образом, данные метрики позволяют оценить, насколько автор текста склонен выражать уверенность, осторожность или нерешительность в суждениях.

Приведем количественные результаты по маркерам уверенности и неуверенности текстов для отрывков из двух интервью (Таблица 8), их визуализацию (Рисунок 7), а также статистику (Таблица 9):

- Таблица 8. Результаты применения метрик оценки certainty vs uncertainty markers текстов двух интервью

text_id	Text_1	Text_2
certainty_count	9	14
uncertainty_count	1	3
hesitation_count	12	30
certainty_ratio	0,02189781	0,023688663
uncertainty_ratio	0,00243309	0,005076142
hesitation_ratio	0,02919708	0,050761421

Рисунок 7. Результаты применения метрик оценки certainty vs uncertainty markers текстов двух интервью

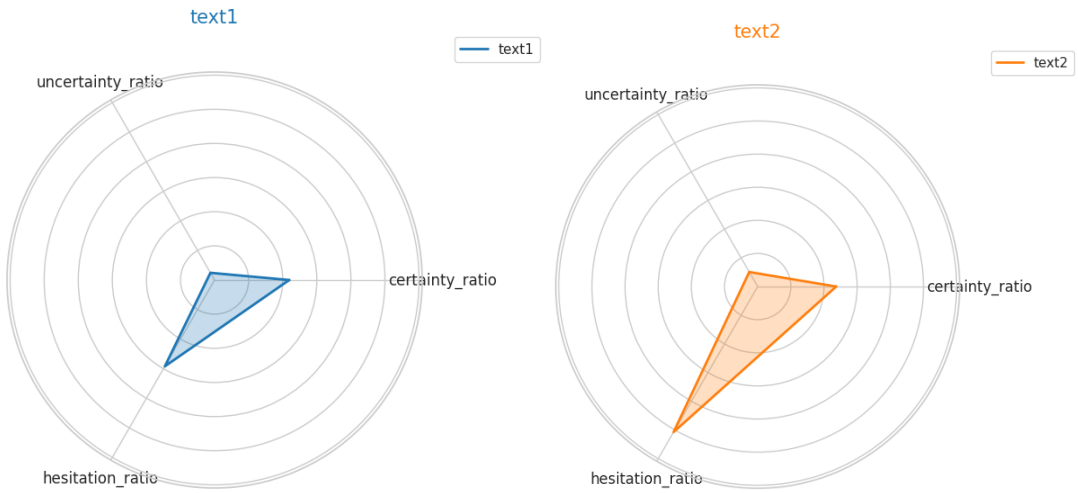

Таблица 9. Результаты применения метрик оценки certainty vs uncertainty markers текстов двух интервью, значения округлены до сотых долей (статистика)

	count	mean	std	min	25%	50%	75%	max
certainty_count	2.0	11.50	3.54	9.00	10.25	11.50	12.75	14.00
uncertainty_count	2.0	2.00	1.41	1.00	1.50	2.00	2.50	3.00
hesitation_count	2.0	21.00	12.73	12.00	16.50	21.00	25.50	30.00
certainty_ratio	2.0	0.02	0.00	0.02	0.02	0.02	0.02	0.02
uncertainty_ratio	2.0	0.00	0.00	0.00	0.00	0.00	0.00	0.01
hesitation_ratio	2.0	0.04	0.02	0.03	0.03	0.04	0.05	0.05

Повышенное количество маркеров уверенности и неуверенности в text_2 может свидетельствовать о более выраженной субъективной позиции автора. Преобладание хезитаций — может указывать на меньшую степень подготовленности текста, его импровизированный или разговорный характер, а также на эмоциональное или неформальное интонирование. Напротив, text_1 выглядит более формальным и сдержанным, с меньшим числом эксплицитных высказываний уверенности или сомнений и менее частыми хезитациями. Таким образом, text_2 выглядит более эмоционально насыщенным, гибким и потенциально менее формальным по сравнению с text_1, что находит отражение как в частоте маркеров уверенности и неуверенности, так и в высокой доле хезитаций.

- Тип локуса контроля (внутренний/внешний)

Для анализа типа локуса контроля в тексте был реализован метод, основанный на распределении глагольных форм в активном и пассивном залоге. Исходная гипотеза базируется на психологических исследованиях, связывающих преобладание активного залога с внутренним локусом контроля — установкой на личную ответственность, инициативу и контроль над ситуацией (Wood, 1973). Напротив, пассивные конструкции чаще ассоциируются с внешним локусом контроля, когда действия объясняются внешними обстоятельствами, случайностью или воздействием других людей. Для определения залога глаголов использовалась библиотека spaCy, обеспечивающая морфосинтаксический разбор и извлечение информации о каждом токене. В процессе обработки извлекались все глаголы из текста, классифицировались по типу залога (active / passive), подсчитывалось количество вхождений каждого типа и, наконец, вычислялась доля активных и пассивных форм от общего числа глаголов. Это позволяет судить о преобладающем типе локуса контроля в тексте и количественно оценить речевые проявления установки на контроль и ответственность.

Приведем количественные результаты по типам локуса контроля текстов для отрывков из двух интервью (Таблица 10), их визуализацию (Рисунок 9), а также статистику (Таблица 11):

Таблица 10. Результаты применения метрик оценки локуса контроля текстов двух интервью

text_id	Text_1	Text_2
active_voice_ratio	0,857142857	0,966101695
passive_voice_ratio	0,142857143	0,033898305

Рисунок 8. Результаты применения метрик оценки локуса контроля текстов двух интервью

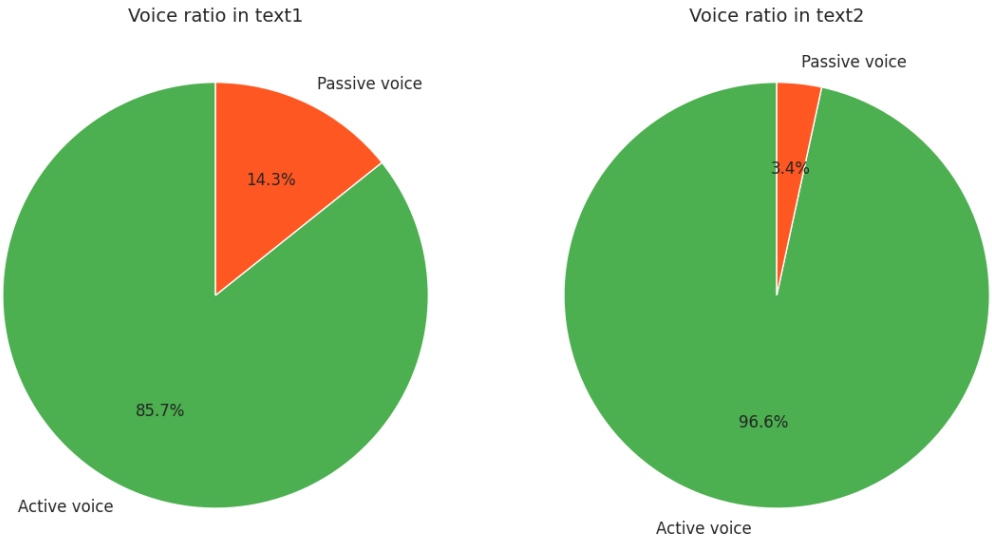

Таблица 11. Результаты применения метрик оценки локуса контроля текстов двух интервью, значения округлены до сотых долей (статистика)

	count	mean	std	min	25%	50%	75%	max
active_voice_ratio	2.0	0.91	0.08	0.86	0.88	0.91	0.94	0.97
passive_voice_ratio	2.0	0.09	0.08	0.03	0.06	0.09	0.12	0.14

В обоих текстах наблюдается значительное преобладание глаголов в активном залоге, что говорит о склонности авторов брать на себя инициативу и демонстрировать контроль над ситуацией. Такая структура речи указывает на доминирование внутреннего локуса контроля, особенно выраженного во втором тексте. Автор text_2, в котором доля глаголов в активном залоге оказалась выше, чем в text_1, вероятно, воспринимает себя как основного агента происходящих событий, что может отражать уверенность в собственных действиях, инициативность и стремление к управлению ситуацией, что потенциально может быть связано с более выраженными лидерскими установками или проактивной позицией в жизни.

- другие метрики

Мы реализовали подсчет меры MSP (Mean size of paradigm), обозначающую морфологическое разнообразие речи и впервые предложенную в статье (Xanthos et al., 2011). Метод заключался в расчёте меры MSP – среднего размера парадигмы, эта мера определялась авторами как количество уникальных словоформ, деленные на количество уникальных лемм. Большая мера MSP означает большее морфологическое богатство речи.

Мы также рассчитали классическую меру для оценки длины высказываний MLU (Mean Length of Utterance), предложенную в работе Brown, R. (1973). A First Language: The Early Stages. Эта мера отражает среднюю длину высказываний в словах или морфемах и используется для оценки языкового развития и сложности речевых конструкций. В нашем исследовании мы рассчитываем MLU в словах и слогах, что позволяет получить представление о среднем размере высказываний в текстах.

Приведем количественные результаты по данным метрикам для отрывков из двух интервью (Таблица 12), их визуализацию (Рисунок 10), а также статистику (Таблица 13):

Таблица 12. Результаты применения дополнительных метрик оценки текстов двух интервью

text_id	Text_1	Text_2
MSP	1,239263804	1,159292035
mlu_words	19,57142857	10,94444444
mlu_syllables	38,0952381	21,7962963

Рисунок 10. Результаты применения дополнительных метрик оценки текстов двух интервью

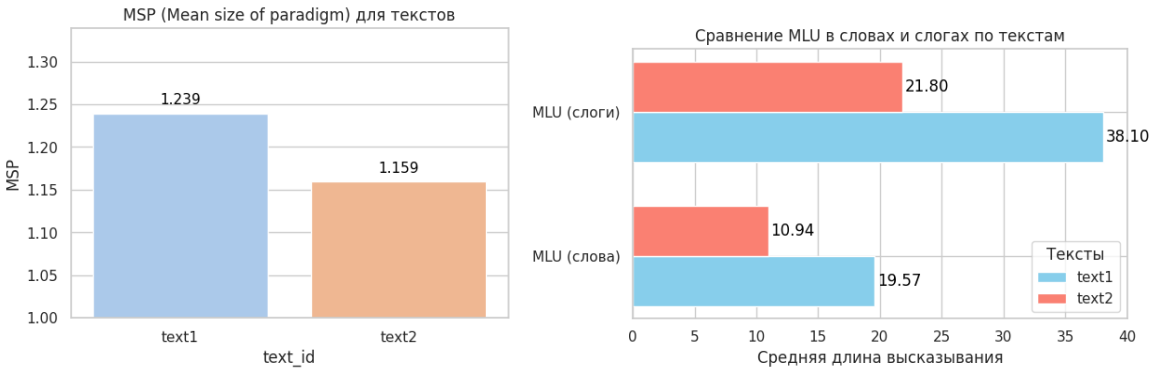

Таблица 13. Результаты применения дополнительных метрик оценки текстов двух интервью, значения округлены до сотых долей (статистика)

	count	mean	std	min	25%	50%	75%	max
MSP	2.0	1.20	0.06	1.16	1.18	1.20	1.22	1.24

mlu_words	2.0	15.26	6.10	10.94	13.10	15.26	17.41	19.57
mlu_syllables	2.0	29.95	11.53	21.80	25.87	29.95	34.02	38.10

Результаты показали, что text_1 характеризуется большим разнообразием форм слов, а text_2 — более стабильной формой лексики. Средняя длина высказывания в словах (MLU) в text_1 значительно выше, чем в text_2, что указывает на более длинные предложения. Аналогично, MLU в слогах для text_1 почти вдвое превышает показатель text_2, подтверждая более сложную структуру высказываний в первом тексте. Таким образом, первый текст более морфологически и синтаксически сложен, что указывает на его более формальный стиль, а второй текст относится к более «живому» разговорному стилю.

Часть 2: Статистический анализ

2.1 Описательная статистика

Описательная статистика по всем вычисленным метрикам представлена в соответствующих секциях с описанием метрик.

Дополнительно:

Число наблюдений: 2 текста для всех метрик.

Читаемость и сложность текста

flesch_reading_ease: 100.0 — 109.7 (легкая читабельность)

gunning_fog_index: 4.7 — 8.8 (умеренная сложность)

coleman_liau_index: 9.7 — 10.8 (относительно простая читабельность)

Лексическое разнообразие и частотность

TTR: 0.44 — 0.45 (средняя лексическая вариативность)

avg_freq: 4501 — 5879 (более частотная лексика в первом тексте)

MTLD: 45.8 — 64.5 (второй текст более лексически разнообразен)

self_focus: 0.05 — 0.06 (умеренный личностный фокус)

concreteness_index: -1.08 — -0.86 (высокая абстрактность текстов)

Синтаксические характеристики

avg_sentence_length: 154 — 349 (значительная разница в длине предложений)

syntax_depth: 12 — 17 (умеренная синтаксическая сложность)

Временные показатели (глаголы и наречия)

past_verb_ratio: 0.10 — 0.57 (доля глаголов в прошедшем времени в одном из текстов значительно выше, чем в другом)

present_verb_ratio: 0.17 — 0.54 (доля глаголов в настоящем времени в одном из текстов значитклно выше, чем в другом)

future_verb_ratio: ~0.02 (доля глаголов в будущем времени небольшая в обоих текстах)

past_adverb_count: 0 — 1 (в одном из текстов нет наречий-маркеров прошедшего времени, в другом одно)

present_adverb_count: 1 (в каждом тексте по 1 наречию-маркеру настоящего времени)

future_adverb_count: 0 (в двух текстах нет наречий-маркеров будущего времени)

total_temporal_adverbs: 1 — 2 (в одном из текстов 1 наречие-маркер времени, в другом 2)

Маркеры уверенности и сомнений

certainty_count: 9 — 14 (маркеры уверенности - в одном из текстов маркеров уверенности больше)

uncertainty_count: 1 — 3 (маркеры неуверенности - в одном из текстов маркеров екуверенности больше)

hesitation_count: 12 — 30 (маркеры хезитаций - в одном из текстов маркеров хезитаций значительно больше)

certainty_ratio: ~0.022 (доля слов-маркеров уверенности в тексте - относительно небольшое количество маркеров уверенности в обоих текстах)

uncertainty_ratio: 0.002 — 0.005 (доля слов-маркеров неуверенности в тексте - относительно небольшое количество маркеров уверенности в обоих текстах)

hesitation_ratio: 0.03 — 0.05 (доля слов-маркеров колебаний или сомнений - относительно небольшое количество хезитаций в обоих текстах (за счет одного из текстов)

Локус контроля (залог)

active_voice_ratio: 0.86 — 0.97 (преобладание активного залога в обоих текстах)

passive_voice_ratio: 0.03 — 0.14 (пассивный залог встречается реже, однако в одном из текстов доля значительно больше)

Длина высказывания и паузы

MSP: 1.16 — 1.24 (средний размер парадигмы - среднее значение морфологического разнообразия речи)

mlu_words: 11 — 20 (средняя длина высказывания в словах - относительно длинные высказывания (за счет пауз, самопрерываний))

mlu_syllables: 22 — 38 (средняя длина высказывания в слогах - относительно длинные высказывания (за счет пауз, самопрерываний))

2.2 Сравнительный анализ

- Проведите статистическое сравнение двух текстов

Поскольку мы имеем по одному значению на каждую метрику для каждого текста, классические статистические тесты неприменимы. Вместо этого мы делаем осмысленное сравнение по каждой метрике, акцентируя внимание на различиях, которые могут быть значимы с точки зрения восприятия, читаемости, стиля и когнитивной нагрузки. В коде для подсчета статистики приведен алгоритм подсчета на большем количестве данных при возможном расширении выборки.

2.3 Корреляционный анализ

- Найдите взаимосвязи между различными лингвистическими метриками
- Интерпретируйте найденные корреляции

Анализ корреляций между лингвистическими метриками показал идеальные значения коэффициентов корреляции ±1, что указывает на сильную линейную зависимость некоторых признаков или их обратную взаимосвязь. Однако при тестировании статистической значимости связи ($p < 0.05$) ни одна из пар признаков не продемонстрировала значимой корреляции, что, вероятно, связано с небольшим размером выборки. Таким образом, на данном этапе нельзя уверенно говорить о статистически значимых взаимосвязях между показателями. Необходимо расширение выборки.

Часть 3: Методология

3.1 Критическая оценка

- Какие ограничения есть у предложенных метрик?

Читабельность текстов – не вполне применима для устного дискурса. Нужно применять специальные метрики для устных текстов. Индексы читаемости считают сложными для восприятия слова, содержащие большее количество слогов (например, индекс Флеша, туманности Ганнинга), при этом не всегда такие слова могут быть действительно специализированным. При этом, короткие слова также могут быть терминами и требовать определенного уровня образования для понимания.

Метрики оценки лексического разнообразия текста также имеют определенные ограничения. Например, TTR чувствительна к длине текста: при увеличении объема текста мера обычно снижается. Это ограничение делает инструмент менее надёжным при сравнении текстов разного размера (Koizumi, & In'nami, 2012).

Расчет индекса конкретности основан на словаре, а в настоящий момент такой словарь для русского языка нуждается в пополнении (содержит только существительные и прилагательное, а также ограниченное количество слов). В этом случае можно применять подход, основанный на извлечении семантических векторов слов, выделяя компонент абстрактности. Таким образом, мы могли бы учесть все слова текстов и оценивать их близость к конкретным или абстрактным понятиям. Предлагается применить подход автоматической семантической кластеризации слов с помощью алгоритма машинного обучения k-means. Таким образом, мы можем выделить кластеры абстрактных и конкретных понятий на основе всех слов текстов.

Метрики для оценки синтаксиса устного дискурса имеют ограничение в плане определения границ предложения: если в письменной речи мы можем дробить текст по соответствующим знакам препинания, в устном рассказе подход к делению предложений

может отличаться, что сказывается на подсчете средней длины предложений и глубины вложенности.

- Как обеспечить валидность и надежность измерений?

Необходимо давать определенное задание для порождения, чтобы все участники находились в равных условиях (либо несколько разных видов условий для групп). Например, согласно одной из наших гипотез, участникам давалось два вида инструкций: в одной респондентам необходимо было породить сжатый рассказ, упомянув ответы на определенные вопросы, в другой фигурировал свободный рассказ о работе. В контексте устного дискурса может быть важным ограничивать время ответа и либо унифицировать его для всех участников, либо делать разным и дробить выборку по условиям. Важно формировать выборку по критериям пола, возраста, стажа работы, занимаемой должности. Для связи лингвистических характеристик текста с психологическими особенностями говорящего рекомендуется проводить ряд психологических тестов перед или после интервью.

- Какие дополнительные данные помогли бы улучшить анализ?

Расширение выборки текстов, метаинформация о говорящих (возраст, стаж, количество лет образования, психологические характеристики (можно измерять с помощью тестирования).

- Как можно автоматизировать предложенный подход?

Данный подход уже обладает высокой степенью автоматизации. Однако многие метрики были подсчитаны с применением словарного подхода: ручного пополнения массивов данных или поиска существующих словарей. Для некоторых метрик, например, для подсчета абстрактных и конкретных понятий возможно применение подхода, основанного на извлечении семантических векторов слов и определения их косинусной близости.

3.2 Технические требования

В этой секции приведено описание инструментов, использованных в процессе анализа.

- Для предобработки текста

- pandas – для работы с табличными данными (загрузка и организация корпусов текстов, словарей)

- re – регулярные выражения для очистки текста и извлечения паттернов

- nltk – токенизация

- string – работа со знаками препинания

Для извлечения лингвистических признаков

- textstat – подсчеты метрик читабельности

- nltk – извлечение характеристик слов и предложений

- lexical_diversity – подсчет MTLD

- natasha – извлечение характеристик слов и предложений

- pymorphy2 – лемматизация, определение времени глаголов,

- pymystem3 – лемматизация, извлечение характеристик слов

- spacy – извлечение синтаксических характеристик, залога глаголов

- Для статистического анализа
  - numpy – работа с числовыми массивами

  - scipy.stats – статистические тесты

  - sklearn.preprocessing.StandardScaler, MinMaxScaler – стандартизация

  - collections.Counter – «счетчик» количества токенов

- Для визуализации результатов
  - matplotlib – графическое представление

  - seaborn – графическое представление

Обсуждение и выводы

Сравнительный анализ двух текстов выявил существенные различия в их лингвистических характеристиках, отражающих как стилистические, так и потенциально психологические особенности авторов. Возвращаясь к гипотезам, наиболее вероятная из которых относится к разделению текстов по статусу говорящих (руководитель – не руководитель), мы выявили возможные лингвистические характеристики речи лидера.

Метрики читабельности текста показали лучшие значения для интервью руководителя. Это может означать способность лидера хорошо структурировать текст, способный быть воспринятым широкой аудиторией, а также ораторские способности.

Лексическое разнообразие второго текста, предположительно принадлежащего руководителю, также оказалось выше, чем в первом тексте. Это характеризует лидера как человека с большим словарным запасом.

Для руководителя также более выражен внутренний локус контроля и временной фокус на настоящем времени. Если первый фактор характеризует лидера, как человека, способного брать личную ответственность, то второй может указывать лишь на метаданные (первый респондент рассказывал о предыдущем месте работы, второй – о текущем).

В отношении метрик оценки абстрактности слов и фокусировки на себе, тексты показали схожие значения. Это не позволяет разграничить два типа текстов и требует дальнейшего изучения. В дополнение, оба текста содержали большое количество маркеров определенности, а также хезитаций, что указывает на разговорный стиль текстов, для которых такая ситуация типична. При этом, в тексте предполагаемого руководителя количество слов, выражающих уверенность было выше.

Таким образом, представленный в отчете свод метрик и методов для их подсчета может быть полезным при оценке корпуса текстов интервью в целях выявления лингвистических коррелятов психологических характеристик сотрудников или клиентов. Это потенциально может помогать в адаптации продуктов компании для клиентов, а также при подборе персонала или разработке обучающих и мотивационных программ для сотрудников.

Источники

Соловьев, В. Д., Вольская, Ю. А., Андреева, М. И., & Заикин, А. А. (2022). Словарь русского языка с индексами конкретности/абстрактности. *Russian Journal of Linguistics*, *26*(2), 515-549.

Brown, R. (1973). *A first language: The early stages*. Harvard University Press.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, *46*(3), 904-911.

Brockmeyer, T., Zimmermann, J., Kulessa, D., Hautzinger, M., Bents, H., Friederich, H. C., ... & Backenstrass, M. (2015). Me, myself, and I: self-referent word use as an indicator of self-focused attention in relation to depression and anxiety. *Frontiers in psychology*, 6, 1564.

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

Davis, E. A. (1937). Mean sentence length compared with long and short sentences as a reliable measure of language development. *Child Development*, 69-79.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221.

Gunning, R. (1952). The technique of clear writing.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in python.

Koizumi, R., & In'nami, Y. (2012). WITHDRAWN: Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, *40*(4), 522-532.

Korobov, M. (2015, April). Morphological analyzer and generator for Russian and Ukrainian languages. In *International conference on analysis of images, social networks and texts* (pp. 320-332). Cham: Springer International Publishing.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2), 381-392.

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of child language*, *14*(2), 201-209.

Rubin, V. L., Liddy, E. D., & Kando, N. (2006). Certainty identification in texts: Categorization model and manual tagging results. In *Computing attitude and affect in text: Theory and applications* (pp. 61-76). Dordrecht: Springer Netherlands.

Wood, G. B. (1973). Effect of locus of control and instructions on speech performance under delayed auditory feedback (Doctoral dissertation, Virginia Tech).

Xanthos, A., Laaha, S., Gillis, S., Stephany, U., Aksu-Koç, A., Christofidou, A., ... & Dressler, W. U. (2011). On the role of morphological richness in the early development of noun and verb inflection. *First Language*, *31*(4), 461-479.