УДК 159.9.07 + 37.015.3
https://doi.org/10.5281/zenodo.19351554

ЭМПИРИЧЕСКОЕ ОБОСНОВАНИЕ ТЕМАТИЧЕСКОГО ЯДРА ЛИЧНОСТНОГО САМООПРЕДЕЛЕНИЯ: ПРОТОКОЛ ОТКРЫТОГО ИССЛЕДОВАНИЯ

А.Ю. Смирнов
АНО «Центр научных исследований
в сфере профориентации и психологии труда»
Москва, Российская Федерация
smirnov@crcg.ru

Аннотация

В статье представлен протокол эмпирического исследования, направленного на формирование воспроизводимого тематического ядра для педагогического сопровождения личностного самоопределения (ЛСО) подростков. В отличие от предшествующих экспертных подходов, мы предлагаем метод выделения ключевых тем на основе высокочастотного анализа множества разнородных описаний художественных и документальных произведений (литература, кино, театр), полученных из открытых интернет-источников. Методология включает сбор текстов, дедупликацию, извлечение тем с помощью больших языковых моделей (LLM), частотный анализ, кластеризацию для выделения доменов и экспертную валидацию. Описан пилотный этап на русскоязычном корпусе, критерии перехода к масштабированию и перспективы межкультурного расширения. Все данные, код и промпты будут опубликованы под открытыми лицензиями. Результаты призваны обеспечить эмпирическую базу для построения тематического плана занятий по ЛСО и

могут быть использованы для адаптации методики в других культурных контекстах.

Ключевые слова: личностное самоопределение, тематический анализ, большие языковые модели, открытая наука, педагогическое сопровождение, культурные инварианты.

EMPIRICAL SUBSTANTIATION OF THE THEMATIC CORE OF PERSONAL SELF-DETERMINATION: AN OPEN RESEARCH PROTOCOL

A.U. Smirnov
ANO "Center for Scientific Research
in the Field of Career Guidance and Labor Psychology"
Moscow, Russian Federation
smirnov@crcg.ru

Abstract

This article presents a protocol for an empirical study aimed at forming a reproducible thematic core for the pedagogical support of adolescents' personal self-determination (PSD). In contrast to previous expert-based approaches, we propose a method for identifying key themes based on high-frequency analysis of a multitude of heterogeneous descriptions of artistic and documentary works (literature, cinema, theater) obtained from open internet sources. The methodology includes text collection, deduplication, theme extraction using large language models (LLMs), frequency analysis, clustering to identify domains, and expert validation. The pilot stage on a Russian-language corpus, the criteria for transitioning to scaling, and prospects for cross-cultural expansion are described. All data, code, and prompts will be published under open licenses. The results are intended to provide an empirical basis for constructing a thematic plan for PSD sessions and can be used to adapt the methodology in other cultural contexts.

Keywords: personal self-determination, thematic analysis, large language models, open science, pedagogical support, cultural invariants.

1. ВВЕДЕНИЕ

Концепция личностного самоопределения (ЛСО) как базового этапа становления субъектности в подростковом возрасте [1] выделяет три содержательные линии: формирование ценностного фундамента, рефлексию личностных качеств и построение адекватной картины мира. Ключевым элементом первой линии является работа с морально-нравственными темами, которые служат материалом для рефлексии и осознанного выбора. В предшествующих разработках (например, в рукописи Г.В. Резапкиной «Грани личности», а также в дидактических материалах «Беседы о самоопределении» [2, 3]) тематический перечень формировался экспертно, на основе анализа художественной литературы и педагогического опыта. Такой подход обеспечил высокую содержательную валидность, но оставляет открытым вопрос о воспроизводимости и возможности адаптации к разным культурным средам.

В настоящем исследовании мы предлагаем **эмпирически обоснованную и воспроизводимую процедуру** выделения тематического ядра, пригодного для педагогического сопровождения ЛСО. Вместо априорного задания тем мы извлекаем их из **множества разнородных описаний** одного и того же произведения, сделанных разными людьми в разных контекстах (пользовательские отзывы, аннотации книжных магазинов, рецензии, вики-статьи). Наша гипотеза: высокочастотные ключевые слова, устойчиво повторяющиеся в этих описаниях, образуют набор мета-понятий, которые могут служить «кирпичиками» для построения картины мира и ценностного фундамента.

Данный подход не отменяет экспертной оценки, а дополняет её, позволяя выявить темы, которые реально актуализируются в речевой практике при обсуждении произведений. Полученный тематический список будет

структурирован в домены с помощью кластеризации и послужит основой для создания тематического плана занятий по ЛСО. Кроме того, разработанный пайплайн может быть использован для адаптации методики к другим языкам и культурам.

2. ТЕОРЕТИЧЕСКИЕ ОСНОВАНИЯ И ПОСТАНОВКА ПРОБЛЕМЫ

2.1. Тема как единица содержания ЛСО

В рамках концепции ЛСО тема понимается как «круг жизненных явлений, событий или вопросов, которые авторы выбирают для изображения и исследования в своем произведении» [1]. Это определение близко к понятию «морально-нравственной категории» в этике и к понятию «темы» в литературоведении и киноведении. Однако для педагогических целей важно, чтобы темы были операционализируемы: их можно было обсуждать, иллюстрировать примерами, связывать между собой.

2.2. Экспертный vs. эмпирический подход к выделению тем

В гуманитарных и социальных науках существуют два основных подхода к выделению тем:

- Экспертный (дедуктивный): исследователи или практики на основе своего опыта и анализа источников формулируют список тем. Преимущество — высокая содержательная обоснованность. Недостаток — субъективность, трудность воспроизведения, возможная укоренённость в конкретной культурной традиции.
- Эмпирический (индуктивный): темы извлекаются из текстовых корпусов с помощью методов корпусной лингвистики, тематического моделирования (LDA, BERTopic) или краудсорсинга. Преимущество — воспроизводимость, возможность выявить неочевидные темы. Недостаток — зависимость от корпуса и методов.

Наш подход сочетает элементы обоих: мы используем эмпирическое извлечение из разнородных описаний, а затем проводим экспертную валидацию, чтобы отфильтровать артефакты и адаптировать список к педагогическим задачам.

2.3. Выбор источников: почему описания произведений?

Произведения искусства (литература, кино, театр) являются носителями культурных смыслов и моральных дилемм. Описания этих произведений, сделанные разными людьми (читателями, зрителями, библиотекарями, рецензентами), отражают, какие темы они считают важными для передачи. При этом:

- Формальные аннотации (в каталогах, на сайтах издательств) дают сжатое, часто нормативное представление о теме.
- Пользовательские отзывы (на книжных порталах, в соцсетях) отражают субъективное эмоциональное восприятие, выделение личностно значимых аспектов.
- Рецензии (в журналах, блогах) содержат аналитическое развёртывание темы.

Смешение этих типов описаний позволяет получить устойчивые темы, которые выделяются в разных коммуникативных контекстах.

3. ЦЕЛИ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Цель: разработать и апробировать воспроизводимую методику выделения тематического ядра для педагогического сопровождения ЛСО на основе высокочастотного анализа множества описаний произведений.

Задачи:

1. Создать пайплайн сбора и обработки разнородных описаний (русскоязычных, с перспективой мультиязычного расширения).

2. Оценить чувствительность метода к типу источника (пользовательские отзывы, формальные аннотации, рецензии) и к составу корпуса.
3. Получить эмпирически обоснованный список высокочастотных тем (мета-понятий) и их кластерную структуру.
4. Сравнить полученный список с экспертным списком «Грани личности» [рукопись] и задокументировать совпадения/расхождения.
5. Определить критерии перехода от пилотного этапа к полномасштабному сбору данных и к межкультурному расширению.

4. МЕТОДОЛОГИЯ

4.1. Общая логика исследования

Исследование разбито на три этапа:

- **Этап 0 (пилотный):** сбор ограниченного корпуса (≈10 000 текстов описаний) из 3–4 типов источников. Цель — отладка пайплайна, проверка чувствительности метода, определение оптимальных параметров.
- **Этап 1 (масштабирование на русском языке):** расширение корпуса до 100 000+ текстов, стратифицированных по типам источников, для получения устойчивого частотного списка.
- **Этап 2 (межкультурное расширение):** параллельные корпуса на английском, китайском и других языках (при наличии ресурсов) для выявления культурных инвариантов и специфических тем.

В данном протоколе детально описываются этапы 0 и 1, а этап 2 обозначается как перспектива, реализуемая после успешного завершения русскоязычного масштабирования.

4.2. Формирование корпуса данных (этап 0)

Источники описаний:

Для русского языка на пилотном этапе используются следующие открытые источники (выбор обоснован разнообразием типов текстов и доступностью):

Тип источника	**Конкретные ресурсы**	**Характеристика текстов**
Пользовательские отзывы	LiveLib, Knigogid, ReadRate	Субъективные, часто эмоциональные, разной длины
Формальные аннотации	Litres, MyBook, книжные магазины («Лабиринт», «Читай-город»), издательства («МИФ», «Альпина», «Эксмо», «АСТ»)	Краткие, преимущественно описательные
Рецензии и статьи	Журнальные и блогерские рецензии (поиск через открытые агрегаторы)	Аналитические, развёрнутые
Энциклопедические описания	Wikipedia (первые 2 абзаца статьи о произведении)	Нейтральные, фактологические

Для каждого произведения собирается не менее 5–10 описаний из разных источников (там, где это возможно). Сбор осуществляется через парсинг с соблюдением robots.txt и ограничений на частоту запросов. Все данные анонимизируются на уровне сбора: удаляются идентификаторы пользователей, сохраняется только текст описания и метаинформация о произведении (название, автор, год, тип источника).

Критерии включения текстов:

- Язык: русский.
- Длина текста: не менее 50 слов (для содержательности).
- Наличие явной связи с конкретным произведением (определяется по метаданным или по структуре URL).

4.3. Предобработка и дедупликация

1. Очистка: удаление HTML-тегов, приведение к нижнему регистру, удаление лишних пробелов, удаление явно маркетинговых фрагментов («купить книгу», «скидка» и т.п.) с помощью регулярных выражений.
2. Дедупликация текстов: для удаления полностью идентичных или почти идентичных записей (например, один и тот же отзыв, скопированный на разных платформах) используется метод minhash + LSH [4] с порогом сходства Жаккара 0,85. Реализация: библиотека datasketch.
3. Группировка текстов по произведениям: поскольку одно произведение может быть представлено разными записями с разными названиями (например, «Война и мир» и «Война и мир. Том 1»), мы используем кластеризацию метаданных. Строим эмбеддинги названий с помощью мультиязычной модели sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 [5], затем применяем HDBSCAN [6] с параметром min_cluster_size=2. Кластеры интерпретируются вручную для верификации.

4.4. Извлечение ключевых тем с использованием LLM

Для извлечения тем из каждого текста описания используется большая языковая модель. На пилотном этапе применяется GPT-4o (OpenAI API) с температурой 0 и максимальным числом токенов 200. Промпт унифицирован и приведён в Приложении А. Общая инструкция:

«*Выдели ключевые темы (морально-нравственные категории, жизненные явления, экзистенциальные вопросы), которые раскрываются в*

произведении на основе приведённого описания. Ответ дай в виде списка существительных или словосочетаний, приведённых к нормальной форме (именительный падеж, единственное число). Исключи жанры, формальные признаки, оценочные суждения. Если в описании нет явных тем, верни пустой список.»

Результат парсится и сохраняется в виде списка строк. Для повышения устойчивости каждый текст обрабатывается один раз; при необходимости последующей проверки используется тот же промпт.

4.5. Приведение понятий к единой форме и фильтрация

1. Нормализация: все извлечённые термины приводятся к нижнему регистру, удаляются знаки пунктуации. Для приведения к нормальной форме (именительный падеж, единственное число) используется библиотека pymorphy3 [7].
2. Фильтрация стоп-слов: применяется стоп-лист для русского языка (на основе nltk.corpus.stopwords), дополненный жанровыми, формальными и оценочными терминами: *роман, фильм, книга, автор, актёр, режиссёр, шедевр, интересно, скучно, рекомендую* и др. Список формируется вручную на основе предварительного анализа частотных слов.
3. Объединение синонимов (на этапе кластеризации): на этапе частотного анализа синонимы не объединяются, чтобы сохранить исходную лексическую вариативность. Семантическое объединение будет выполнено на этапе кластеризации через эмбеддинги.

4.6. Частотный анализ и определение высокочастотных тем

Для каждого понятия (нормализованного) подсчитывается количество текстов, в которых оно встречается (документная частота). Построение частотного списка производится:

- для всего корпуса;

- отдельно для каждого типа источника.

Определение порога высокочастотности на пилотном этапе будет проводиться двумя способами с последующим сравнением:

- фиксированный порог: топ-N понятий (N будет выбран после анализа распределения, предположительно N=200–500);
- динамический порог: все понятия с частотой выше медианы + 1,5×IQR (межквартильный размах).

Оптимальный способ будет выбран на основе экспертной оценки релевантности полученных тем (см. раздел 4.8).

4.7. Кластеризация для выделения доменов и микро-тем

1. Векторизация понятий: для каждого высокочастотного понятия вычисляется эмбеддинг с помощью модели text-embedding-3-large (OpenAI) или, для открытости, с помощью sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 [5]. Использование мультиязычной модели позволяет в перспективе унифицировать анализ с другими языками.
2. Кластеризация: применяется агломеративная кластеризация с косинусной метрикой и методом Уорда [8]. Число кластеров определяется автоматически на основе коэффициента силуэта [9] или задаётся после визуального анализа дендрограммы.
3. Интерпретация: каждый кластер получает название (домен), которое формулируется экспертами на основе входящих в него понятий. Например, кластер, содержащий «любовь», «романтика», «верность», может быть назван «Отношения и привязанность».
4. Выделение микро-тем: внутри каждого домена проводится вторичная кластеризация для выделения подгрупп понятий (подтем), которые могут использоваться для структурирования занятий.

4.8. Экспертная валидация и ручная обработка

Участники: 50–100 практиков (педагоги, психологи, профконсультанты), имеющие опыт работы по методике ЛСО. Они будут приглашены через сеть экспериментальных площадок АНО «Центр научных исследований в сфере профориентации и психологии труда».

Процедура: экспертам предоставляется:

- список доменов с входящими в них понятиями (после кластеризации);
- частотный список высокочастотных понятий (для контекста);
- краткое описание целей исследования.

Эксперты оценивают:

1. Релевантность каждого домена для педагогических задач ЛСО (по 5-балльной шкале).
2. Полноту: какие важные темы отсутствуют? (открытый вопрос).
3. Корректность отнесения понятий к доменам (возможность предложить перенос).
4. Общее впечатление и предложения по структуре.

Сбор данных осуществляется через онлайн-форму (Google Forms или Яндекс.Формы) анонимно. Статистический анализ: медианные оценки, коэффициент конкордации Кендалла для согласованности [10]. На основе экспертных комментариев проводится ручная корректировка финального списка тем и структуры доменов.

4.9. Критерии перехода к масштабированию (этап 1)

Пилотный этап считается успешным и даёт основание для масштабирования при выполнении следующих условий:

1. **Релевантность извлечённых тем:** не менее 80% высокочастотных понятий (топ-100) признаны экспертами релевантными для ЛСО (оценка ≥4 из 5).

2. **Устойчивость между типами источников:** коэффициент ранговой корреляции Спирмена между частотными списками из разных типов источников (например, пользовательские отзывы vs. формальные аннотации) ≥0,7.
3. **Техническая реализуемость:** пайплайн обработки работает без критических ошибок, время обработки 10 000 текстов не превышает 24 часов (с учётом лимитов API).

При невыполнении условий будут скорректированы параметры (промпт, стоп-лист, пороги кластеризации) и пилот повторён. Если после двух итераций условия не достигаются, методология будет пересмотрена (например, переход на гибридный подход с использованием тематического моделирования).

5. ПЛАН МАСШТАБИРОВАНИЯ И МЕЖКУЛЬТУРНОГО РАСШИРЕНИЯ

5.1. Масштабирование на русском языке

После успешного пилота корпус расширяется до 100 000+ текстов. Сбор данных автоматизируется с использованием тех же источников, но с увеличением числа произведений и глубины парсинга. На этом этапе:

- Проводится полный цикл обработки (дедупликация, извлечение тем, частотный анализ, кластеризация) на всём корпусе.
- Частотные списки и кластеры сравниваются с пилотными для проверки устойчивости.
- Финальный список тем и доменов фиксируется и публикуется.

5.2. Межкультурное расширение

Параллельно или после завершения русскоязычного этапа планируется создание аналогичных корпусов на английском, китайском, арабском и других языках. Для этого:

- Подбираются аналогичные типы источников (IMDb, Goodreads, Douban, местные библиотечные каталоги).
- Адаптируется пайплайн: перевод промпта, замена стоп-листа и морфологического анализатора на соответствующий язык, использование мультиязычных эмбеддингов для кластеризации (LaBSE [11]).
- Проводится сравнительный анализ: выделяются культурные инварианты (темы, входящие в топ во всех корпусах) и культурно-специфические темы.

Результаты межкультурного этапа позволят оценить универсальность тематического ядра ЛСО и подготовить рекомендации по адаптации методики для разных стран.

6. Открытость и воспроизводимость

Исследование следует принципам открытой науки:

- **Данные:** все собранные корпуса (очищенные, анонимизированные) будут опубликованы на Zenodo [12] под лицензией Creative Commons Attribution (CC BY 4.0).
- **Полный код обработки** (Python, Jupyter notebooks) предполагается разместить на GitHub [13] под лицензией GPL v3.
- **Промпты и параметры LLM:** точные формулировки промптов и параметры API будут размещены на сайте проекта [Центр научных исследований в сфере профориентации и психологии труда | Психология самоопределения](https://crcg.ru/so/lso-thematic-core).
- **Результаты:** промежуточные и финальные результаты (частотные списки, кластеры, оценки экспертов) публикуются в виде дополнительных файлов к препринту.

Статус материалов и версионность

На момент публикации данного препринта ведётся активная работа по сбору корпуса, отладке пайплайна и уточнению промптов. Актуальные версии материалов исследования поддерживаются в открытом доступе на сайте проекта [14]. Изменения документируются в логе исследования на этой странице.

По завершении пилотного этапа (после стабилизации методологии) все материалы будут заархивированы на Zenodo с присвоением DOI. Соответствующие ссылки будут добавлены в данный препринт при публикации финальной версии.

7. ЭТИЧЕСКИЕ АСПЕКТЫ

- Все данные собираются из общедоступных источников. Парсинг осуществляется в соответствии с robots.txt и условиями использования платформ (где это явно разрешено). Для ресурсов, запрещающих автоматический сбор, используются официальные API или данные не включаются.
- Персональные данные пользователей (имена, аватары) не сохраняются; тексты отзывов анонимизируются путём удаления любых идентификаторов.
- Экспертная оценка проводится с добровольного согласия участников; данные обрабатываются анонимно.
- Исследование не предполагает вмешательства в психику испытуемых, не требует одобрения этического комитета, так как работает исключительно с открытыми текстовыми данными и анонимными опросами специалистов.

8. ОГРАНИЧЕНИЯ И ВОЗМОЖНЫЕ РИСКИ

8.1. Смещение источников

Если после стратификации выяснится, что разные типы источников дают несопоставимые списки тем ($\rho < 0,5$), это будет означать, что единого «тематического ядра» не существует. В этом случае мы представим отдельные списки для разных контекстов (например, список для формальных аннотаций и список для пользовательских отзывов) и предложим педагогические рекомендации по их использованию.

8.2. Качество извлечения LLM

Если на пилотной выборке эксперты признают менее 60% извлечённых ключевых слов релевантными темами, мы:

- изменим промпт (уточним инструкцию, добавим примеры);
- сменим модель (например, на DeepSeek-V3 или локально развёртываемую модель);
- рассмотрим гибридный подход: сначала извлечение кандидатов через LLM, затем их фильтрацию с помощью тематического моделирования (BERTopic).

8.3. Технические ограничения

При масштабировании стоимость API (при использовании GPT-4o) может стать значительной. В качестве альтернативы мы рассмотрим использование локальных моделей (например, Llama 3, Qwen) или более дешёвых моделей (GPT-3.5-turbo) с сохранением качества извлечения. Решение будет принято после оценки точности на пилоте.

8.4. Несовпадение с экспертным списком

Если полученные темы сильно расходятся с экспертным списком «Грани личности», это не будет считаться неудачей, а станет важным результатом,

демонстрирующим различия между экспертным и коллективным выделением тем. Мы подробно проанализируем расхождения и обсудим, какие темы могут быть упущены в экспертной традиции, а какие, напротив, являются артефактами речевой практики.

8.5. УСЛОВИЯ ДОСРОЧНОГО ПРЕКРАЩЕНИЯ ИССЛЕДОВАНИЯ

Исследование будет остановлено, если:

- на пилотном этапе после двух итераций корректировок не удаётся достичь релевантности извлечённых тем ≥60%;
- стоимость обработки полного корпуса окажется более чем в 3 раза выше запланированного бюджета, и не будет найдено альтернативных технических решений;
- возникнут юридические препятствия (изменение условий использования источников, запрет на парсинг).

9. ОЖИДАЕМЫЕ РЕЗУЛЬТАТЫ И ИХ РОЛЬ В КОНЦЕПЦИИ ЛСО

9.1. Научные результаты

- Эмпирически обоснованный список высокочастотных тем (мета-понятий) для русского языка, структурированный в домены и подтемы.
- Сравнительный анализ с экспертным списком «Грани личности».
- Открытый пайплайн (код, данные, промпты), который может быть использован для воспроизведения и адаптации в других языках.
- (При реализации этапа 2) выявление межкультурных инвариантов и культурно-специфических тем.

9.2. Практическое применение в ЛСО

Полученная структура тем ляжет в основу **тематического плана занятий** по ЛСО:

- Каждый домен будет раскрыт через серию занятий, построенных по принципу цикличности (от 12 до 18 лет с углублением).
- Для каждой темы будут подобраны репрезентативные примеры из художественных произведений (с использованием рейтинга произведений, построенного на основе частоты упоминания тем в их описаниях).
- Карта связей между темами (на основе кластеров) позволит выстраивать последовательность занятий так, чтобы «поднимать» связанные темы одновременно.

Кроме того, разработанная методология может быть использована для построения аналогичных тематических ядер для других типов самоопределения (субкультурного, политического и т.д.) в рамках общей системы педагогического сопровождения.

10. ЗАКЛЮЧЕНИЕ

В данном протоколе представлен план эмпирического исследования, направленного на создание воспроизводимого тематического ядра для личностного самоопределения подростков. Предлагаемый метод сочетает современные методы корпусного анализа, большие языковые модели и экспертную валидацию, что обеспечивает баланс между объективностью и содержательной глубиной. Открытость всех этапов (данные, код, промпты) позволяет научному сообществу критически оценить результаты и использовать их в своей работе.

Реализация исследования будет проходить в три этапа, с чёткими критериями перехода. Мы надеемся, что полученные результаты не только укрепят эмпирическую базу концепции ЛСО, но и предложат практический инструмент для педагогов и психологов, работающих с подростками.

Благодарности

Автор выражает благодарность Галине Владимировне Резапкиной за предоставленную рукопись «Грани личности» и многолетнее сотрудничество в развитии методики ЛСО, а также участникам экспериментальных площадок АНО «Центр научных исследований в сфере профориентации и психологии труда» за готовность участвовать в экспертной валидации.

Использование систем искусственного интеллекта

При подготовке данного протокола использовались языковые модели ChatGPT (OpenAI), DeepSeek и Google Gemini для стилистической правки и структурирования текста. Автор несёт полную ответственность за содержание и методологические решения.

Список литературы

[1] Смирнов А.Ю. Концепция личностного самоопределения (ЛСО) как базового этапа становления субъектности в подростковом возрасте. Zenodo, 2026. DOI: 10.5281/zenodo.19051635.

[2] Резапкина Г.В. Беседы о самоопределении: книга для чтения учащихся 5–7 классов. М.: Академия, 2012.

[3] Резапкина Г.В. Беседы о самоопределении: книга для чтения учащихся 8–9 классов. М.: Академия, 2012.

[4] Leskovec J., Rajaraman A., Ullman J.D. Mining of Massive Datasets. Cambridge University Press, 2020.

[5] Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP 2019.

[6] McInnes L., Healy J., Astels S. hdbscan: Hierarchical density based clustering. Journal of Open Source Software, 2017.

[7] Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages. 2015.

[8] Murtagh F., Legendre P. Ward's hierarchical agglomerative clustering method. Journal of Classification, 2014.

[9] Rousseeuw P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 1987.

[10] Kendall M.G., Gibbons J.D. Rank Correlation Methods. Edward Arnold, 1990.

[11] Feng F., Yang Y., Cer D., et al. Language-agnostic BERT sentence embedding. ACL 2022.

[12] Zenodo. https://zenodo.org/

[13] GitHub. https://github.com/

[14] Центр научных исследований в сфере профориентации и психологии труда | Психология самоопределения. https://crcg.ru/so/lso-thematic-core/

Reference

[1] Smirnov, A.U. (2026). *The Concept of Personal Self-Determination (PSD) as a Foundational Stage in the Formation of Subjectivity in Adolescence*. Zenodo. DOI: 10.5281/zenodo.19051635.

[2] Rezapkina, G.V. (2012). *Conversations on Self-Determination: A Reader for Students in Grades 5–7*. Moscow: Akademiya.

[3] Rezapkina, G.V. (2012). *Conversations on Self-Determination: A Reader for Students in Grades 8–9*. Moscow: Akademiya.

[4] Leskovec, J., Rajaraman, A., & Ullman, J.D. (2020). *Mining of Massive Datasets*. Cambridge University Press.

[5] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP*.

[6] McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*.

[7] Korobov, M. (2015). Morphological analyzer and generator for Russian and Ukrainian languages.

[8] Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method. *Journal of Classification*.

[9] Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*.

[10] Kendall, M.G., & Gibbons, J.D. (1990). *Rank Correlation Methods*. Edward Arnold.

[11] Feng, F., Yang, Y., Cer, D., et al. (2022). Language-agnostic BERT sentence embedding. *ACL*.

[12] Zenodo. https://zenodo.org/

[13] GitHub. https://github.com/

[14] Center for Scientific Research in Career Guidance and Occupational Psychology | Psychology of Self-Determination. https://crcg.ru/so/lso-thematic-core/