Психология. Журнал Высшей школы экономики. 2024. Т. 21. № 1. С. 202–225.
Psychology. Journal of the Higher School of Economics. 2024. Vol. 21. N 1. P. 202–225.
DOI: 10.17323/1813-8918-2024-1-202-225

Обзоры и рецензии

АВТОМАТИЧЕСКАЯ ОЦЕНКА ТЕСТОВ НА ВЕРБАЛЬНУЮ КРЕАТИВНОСТЬ: ОТ ЛЕКСИЧЕСКИХ БАЗ ДАННЫХ К БОЛЬШИМ ЯЗЫКОВЫМ МОДЕЛЯМ

Е.А. ВАЛУЕВА[a,b], А.С. ПАНФИЛОВА[a], А.С. РАФИКОВА[a,c]

[a] *ФГБУН Институт психологии РАН, 129366, Москва, ул. Ярославская, д. 13, к. 1*
[b] *ФГБОУ ВО «Московский государственный психолого-педагогический университет», 127051, Россия, Москва, ул. Сретенка, д. 29*
[c] *Государственный академический университет гуманитарных наук, 119049, Москва, Мароновский пер., д. 26*

Automatic Scoring of Verbal Divergent Thinking Tests: From Lexical Databases to Large Language Models

E.A. Valueva[a,b], A.S. Panfilova[a], A.S. Rafikova[a,c]

[a] *Institute of Psychology, Russian Academy of Sciences, 13 build. 1 Yaroslavskaya Str., Moscow, 129366, Russian Federation*
[b] *Moscow State University of Psychology & Education, 29 Sretenka Str., Moscow, 127051, Russian Federation*
[c] *State Academic University for the Humanities, 26 Maronovskiy Pereulok, Moscow, 119049, Russian Federation*

Резюме

В статье рассматривается эволюция методов автоматической оценки вербальных тестов на дивергентное мышление. Основным предметом внимания исследователей становится возможность оценить оригинальность ответов испытуемых с помощью подсчета их семантической удаленности от стимульной задачи. В период с

Abstract

The article explores the evolution of methods for automatically assessing verbal divergent thinking tests. Researchers have increasingly focused on the ability to evaluate the originality of respondents' answers by calculating their semantic distance from the stimulus task.

Исследование выполнено при поддержке РНФ, проект № 23-18-01059.
The research was supported by the Russian Science Foundation, project # 23-18-01059.

2009 по 2019 г. главным методом оценки семантических расстояний стал латентно-семантический анализ. В целом, с точки зрения внутренней согласованности и корреляции с экспертными оценками, его применение давало удовлетворительные результаты, позволяя сохранить допустимый баланс качества и затраченных усилий. Однако выявились проблемы (зависимость оценок от используемого корпуса, нестабильность результатов, систематические искажения, связанные с длиной анализируемых ответов), которые заставили исследователей перейти к более продвинутым моделям дистрибутивной семантики (GloVe, Word2Vec), большим языковым моделям и обучению с учителем. Большие языковые модели (особенно дообученные на материале тестов креативности) показали более высокую эффективность, чем модели, оценивающие семантические расстояния, и приблизились к оценкам, которые дают эксперты. Помимо оценки оригинальности, в статье рассматриваются работы, в которых предлагаются методы автоматической оценки разработанности, гибкости, ассоциативного потока и дивергентной семантической интеграции. Приводятся ссылки на онлайн-платформы, позволяющие получать автоматические оценки оригинальности ответов на дивергентные тесты. Обсуждается проблема интерпретации полученных с помощью больших языковых моделей результатов. Недостатком применения этих моделей является отсутствие понимания, на каких основаниях выносится суждение об оригинальности творческих продуктов. Обсуждаются перспективы применения объяснимого искусственного интеллекта для оценки результатов вербальных и невербальных тестов творческого мышления.

Ключевые слова: дивергентные тесты, оригинальность, автоматическая обработка, семантические расстояния, дистрибутивная семантика, латентно-семантический анализ, большие языковые модели.

Валуева Екатерина Александровна — научный сотрудник, лаборатория психологии и психофизиологии творчества, ФГБУН «Институт психологии Российской академии наук»; старший научный сотрудник, лаборатория

From 2009 to 2019, latent semantic analysis became the primary method for assessing semantic distances. Overall, in terms of internal consistency and correlation with expert ratings, its application yielded satisfactory results, maintaining an acceptable balance of quality and effort expended. However, issues emerged (dependence on the corpus used, result instability, systematic distortions related to the length of analyzed responses), prompting researchers to transition to more advanced models of distributional semantics (GloVe, Word2Vec etc.), large language models, and supervised learning. Large language models, especially those fine-tuned on creativity test materials, demonstrated higher effectiveness compared to models assessing semantic distances and approached expert evaluations. In addition to evaluating originality, the article considers works proposing methods for automatic assessment of elaboration, flexibility, associative flow, and divergent semantic integration. References to online platforms that allow for automatic assessments of originality in responses to divergent tests are provided. The issue of interpreting results obtained through large language models is discussed. A drawback of using these models is the lack of understanding of the basis on which judgments of the originality of creative products are made. The perspectives of applying explainable artificial intelligence for evaluating results of verbal and non-verbal tests of creative thinking are being discussed.

Keywords: divergent thinking tests, originality, automatic processing, semantic distances, distributional semantics, latent semantic analysis, large language models.

Ekaterina A. Valueva — Research Fellow, Institute of Psychology, Russian Academy of Sciences; Senior Researcher, Laboratory for the Study of Cognitive and Communicative Processes in

исследования когнитивных и коммуникативных процессов у подростков и юношей при решении игровых и учебных задач в цифровых средах, ФГБОУ ВО «Московский государственный психолого-педагогический университет» (ФГБОУ ВО МГППУ), кандидат психологических наук.
Сфера научных интересов: когнитивная психология, интеллект, творчество.
Контакты: ekval@list.ru

Панфилова Анастасия Сергеевна — научный сотрудник, лаборатория психологии и психофизиологии творчества, ФГБУН «Институт психологии Российской академии наук», кандидат технических наук.
Сфера научных интересов: когнитивная психология, методы машинного обучения.
Контакты: panfilova87@gmail.com

Рафикова Антонина Семеновна — научный сотрудник, Государственный академический университет гуманитарных наук; научный сотрудник, лаборатория психологии и психофизиологии творчества, ФГБУН «Институт психологии Российской академии наук», кандидат психологических наук.
Сфера научных интересов: психолингвистика, когнитивная психология.
Контакты: antoninaraf@yandex.ru

Adolescents and Young Adults while Solving Game and Educational Problems using Digital Environments, Moscow State University of Psychology & Education, PhD in Psychology.
Research Area: cognitive psychology, intelligence, creativity.
E-mail: ekval@list.ru

Anastasia S. Panfilova — Research Fellow, Institute of Psychology, Russian Academy of Sciences, PhD in Engineering.
Research Area: cognitive psychology, machine learning methods.
E-mail: panfilova87@gmail.com

Antonina S. Rafikova – Research Fellow, State Academic University for the Humanities; Research Fellow, Institute of Psychology, Russian Academy of Sciences, PhD in Psychology.
Research Area: psycholinguistics, cognitive psychology.
E-mail: antoninaraf@yandex.ru

In our view, one main methodological bottleneck that limits the productivity and impact of creativity research has historically been the time- and resource-intensiveness of human-rated DT tasks (Dumas, Organisciak, Doherty, 2021).

Наиболее распространенным методом диагностики креативности в эмпирических исследованиях являются тесты дивергентного мышления. Дивергентное мышление — это процесс порождения разнообразных идей, в отличие от поиска единственного правильного решения. Тесты на дивергентное мышление включают задачи, в которых людей просят предложить как можно больше решений в ответ на данную проблему. Например, задачи могут включать в себя создание как можно большего количества предметов, удовлетворяющих определенным критериям, нахождение сходства между различными объектами, выявление новых способов использования объекта, предсказание как можно большего количества последствий гипотетической ситуации, завершение предложенных рисунков различными способами и т.д.

Оценка результатов тестирования — главная проблема в исследованиях креативности. Беглость (количество предложенных идей), оригинальность

(качество предложенных идей) и гибкость (разнообразие предложенных идей) являются основными параметрами, по которым оценивают работу испытуемых. Подсчет количества предложенных идей является самым простым способом оценки (чем больше придумал идей, тем более креативный), но вместе с тем оказывается наименее надежным с точки зрения конструктной валидности (Reiter-Palmon et al., 2019).

Оригинальность можно трактовать, учитывая три аспекта: редкость, удаленность, остроумность (Wilson et al., 1953). Последний аспект, как правило, специально не оценивается, но может быть учтен экспертами при выставлении оценок. Оригинальность как редкость может быть оценена на основе частотного принципа: ответы испытуемых объединяются в группы по принципу схожести, после чего подсчитывается количество ответов в определенной категории и выставляются оценки оригинальности, обратно пропорциональные частотности категории. Сложность здесь заключается в том, что не всегда очевидно, можно ли считать два ответа одинаковыми или нужно отнести их к разным группам (а от этого сильно зависит подсчет частотности). Недостатками этого подхода также являются высокая трудоемкость, необходимость иметь большую базу ответов для более надежных оценок (Forthmann et al., 2020) и сильная зависимость частоты встречаемости ответов от особенностей испытуемых (возраст, уровень образования, профессиональная группа и т.д.).

Оригинальность как удаленность — это непохожесть идеи на другие, что часто связано с выявлением неочевидного свойства стимула, предложенного в задаче. До недавнего времени оценка необычности могла быть проведена лишь экспертами. Для экспертной оценки оригинальности привлекаются специально обученные люди (которые, как правило, являются непосредственно организаторами эксперимента, их студентами или коллегами). Эксперты оценивают либо каждый ответ испытуемого в отдельности, либо весь набор ответов одного испытуемого целиком (Silvia et al., 2009). Каждый ответ должны оценить как минимум два эксперта, чтобы минимизировать субъективность оценок. Иногда добиться согласованности экспертов бывает сложно: они могут иметь разные критерии креативности или изменять свои собственные критерии по мере приобретения опыта. Все это ставит под сомнение надежность полученных результатов (критический обзор техники консенсусной оценки см.: Cseh, Jeffries, 2019).

Гибкость является важным с теоретической точки зрения измерением креативности, но крайне редко используется исследователями как реальная переменная. Скорее всего, это связано с большой трудоемкостью подсчетов. Гибкость может трактоваться либо как количество категорий, к которым относятся идеи испытуемых, либо как количество переключений между категориями в процессе порождения идей (Grajzel et al., 2023). Для подсчета гибкости также требуется разделение ответов на категории. Дополнительная проблема при подсчете гибкости связана с тем, что оценки гибкости напрямую связаны с широтой категорий: чем у́же категории, тем выше показатели гибкости и тем больше они коррелируют с беглостью. Определение широты категорий оказывается самым важным процессом, который также очень сильно

подвержен субъективности. Как правило, гибкость действительно довольно высоко коррелирует с оригинальностью и беглостью, поэтому предпочтение отдается последним.

Таким образом, главными ограничениями, возникающими при диагностике креативности, являются субъективность и потраченные усилия и время. Автоматический подсчет баллов позволяет преодолеть перечисленные выше сложности, поэтому неудивительно, что развитие информационных технологий сразу же привело к попыткам их применения в области исследования творческого мышления. На настоящий момент огромное количество работ связано с оценкой вербальной креативности, а подходы, позволяющие оценить рисуночные тесты, только начинают разрабатываться (Cropley, Marrone, 2022; Patterson, Barbot et al., 2023). Как мы увидим ниже, при разработке методов автоматической оценки вербальной креативности оригинальность (в аспекте удаленности) стала трактоваться как семантическая дистанция ответов по отношению к предъявленному стимулу. Именно на вычислении семантических расстояний сосредоточены основные усилия исследователей.

Что легко посчитать?

Дополнительные усилия и хитрые алгоритмы искусственного интеллекта не требуются, чтобы посчитать показатель беглости — для этого нужно лишь оценить общее количество небессмысленных ответов у испытуемого. Иногда исследователи обращаются к еще одному показателю — разработанности. Новизна и полезность (применимость, пригодность) — основные критерии творческого продукта (Runco, Jaeger, 2012), и именно разработанность связана с применимостью. Разработанность определяется как степень детализированности ответа (насколько подробно испытуемый объясняет или изображает свою идею). Разработка и конкретизация идеи повышают ее качество, делают идею более реалистичной или применимой на практике.

Основные силы исследователей, как мы увидим ниже, были брошены на совершенствование методов оценки новизны идей (что неудивительно, так как новизна, по всей видимости, является ведущим критерием креативности (см., например: Diedrich et al., 2015)). Вместе с тем разработанность идеи в вербальном тесте довольно легко оценить по количеству использованных слов. Д. Дюма с соавт. (Dumas, Organisciak, Maio, Doherty, 2021) усовершенствовали простой подсчет количества слов и сравнили четыре метода оценки разработанности вербальных ответов: невзвешенный подсчет слов, метод учета частей речи, метод учета стоп-слов и метод взвешивания по обратной частоте документа (IDF weighting).

Преимуществом невзвешенного подсчета слов является простота интерпретации: каждое слово в ответе испытуемого относится к латентному конструкту разработанности, и его оценка линейно увеличивается с каждым использованным словом. Проблема заключается в том, что каждому слову придается равный вес при подсчете, а это не всегда может быть справедливо. В отличие от невзвешенного подсчета слов, метод учета частей речи направлен на исключение очень распространенных или менее концептуально значимых

слов, т.е. идентифицируются части речи, которые включаются в анализ, и части речи, которые исключаются из анализа. Авторы включили в свой анализ существительные, глаголы, прилагательные и наречия. Метод учета стоп-слов позволяет исключить из анализа определенные слова; обычно это очень распространенные слова (например, if, is, it) или слова, не несущие семантической нагрузки, а призванные определять структуру предложения (например, although, meanwhile, whereas). Как правило, списки стоп-слов довольно универсальны и находятся в свободном доступе. Дюма с соавт. воспользовались одним из них (https://github.com/explosion/spaCy/blob/a741de7cf658ce9a90d7afe67c88face8fb658ad/spacy/lang/ru/stop_words.py). Методы учета стоп-слов и частей речи представляют собой схемы взвешивания, при которых включенные слова получают вес, равный единице, а исключенные слова получают нулевой вес. Метод взвешивания по обратной частоте документа основан на том, что каждое слово получает определенный вес, обратно пропорциональный частотности слов: общеупотребительные слова получают меньшие веса, а реже употребляемые слова получают большие веса. Веса слов были взяты из предварительно подготовленного одним из авторов корпуса художественных текстов (https://sense.porganized.com/a-dataset-of-term-stats-in-literature-e95d39bd5959).

Авторы оценили результаты теста «Необычное использование» четырьмя обозначенными выше способами и сопоставили баллы с другими показателями — с беглостью, оригинальностью, с данными по опросникам креативности, а также посчитали средние для разных профессиональных групп (не артисты, студенты-артисты, профессиональные артисты). Четыре оценки разработанности высоко коррелировали между собой (от 0.93 до 0.99), сходным образом коррелировали с другими переменными (негативно — с беглостью, незначимо позитивно — с оригинальностью, положительно — с самооценкой в сфере литературного творчества), а также были способны дифференцировать группы актеров и не актеров. Вместе с тем чуть более выраженные результаты по всем видам сравнений были получены с помощью метода исключения стоп-слов, поэтому авторы рассматривают его как наиболее перспективный и рекомендуют к использованию (Dumas, Organisciak, Maio, Doherty, 2021).

Лексические базы данных

С теоретической точки зрения наиболее важным параметром оценки креативности является оригинальность ответов. Как было отмечено выше, в эпоху ручных подсчетов оригинальность понималась либо как редкость («античастотность»), либо как согласованность экспертного мнения в отношении необычности (креативности) ответов. В контексте автоматической оценки вербальной креативности оригинальность стала интерпретироваться как семантическая удаленность, т.е. как расстояние на ассоциативной сети, которое отделяет ответ испытуемого от слова-стимула. Способы оценки семантического расстояния эволюционировали от использования лексических баз данных (например, WordNet) до оценки взаимоотношений между словами с помощью больших языковых моделей. Оценка оригинальности с помощью

семантических расстояний концептуально близка к оценкам частотности, так как вероятность извлечения семантически удаленного понятия оказывается существенно меньше, чем вероятность использования семантически близкого (однако иногда проводится разделение аспектов частотности (uncommonness) и удаленности (remoteness) (Forthmann et al., 2017)).

В одной из первых работ было предложено использовать лексическую базу WordNet для категоризации ответов испытуемых в тесте «Необычное использование» (Bossomaier et al., 2009). WordNet — это лексикон, в котором слова группируются в синонимические множества (синсеты) и связаны различными семантическими отношениями, такими как гиперонимия (более общий термин) и гипонимия (более специфический термин) и т.д. Авторы предложили распределять ответы по семантическим категориям, опираясь на показатели семантической близости. Таким образом можно оценивать разнообразие идей, не опираясь на субъективные оценки близости понятий. Эта работа имела теоретический характер, эмпирических данных для оценки эффективности предложенного подхода не было приведено. Однако в другой работе (Acar, Runco, 2014) было проведено сравнение эффективности использования разных лексических баз (WordNet, Word Associations Network и IdeaFisher) для подсчета показателей креативности. Word Associations Network — это семантическая база данных, содержащая информацию о том, какие слова ассоциируются друг с другом в языке. IdeaFisher — это программное обеспечение, разработанное для стимулирования и улучшения творческого мышления и генерации идей и содержащее также ассоциативную базу слов. Близкими ассоциациями считались ответы, которые присутствовали в ассоциативном списке соответствующей базы, далекими — те, которые отсутствовали. Индексы креативности, полученные с использованием трех баз, в значительной степени коррелировали между собой (что часто не наблюдается при экспертном оценивании), при этом более согласованными оказались оценки далеких ассоциаций по сравнению с близкими. В более позднем исследовании авторы увеличили количество баз, на основе которых вычисляется семантическое расстояние (Beketayev, Runco, 2016), и сравнили полученные оценки с традиционным способом подсчета. Выяснилось, что два способа подсчета близки в оценке гибкости (корреляция составила 0.74), но довольно сильно различаются в оценке оригинальности ($r = 0.36$).

Дистрибутивная семантика: латентно-семантический анализ

Направление, связанное с использованием статических баз данных, довольно быстро зашло в тупик, так как применение ограниченных по объему словарей часто не позволяет полноценно проанализировать все ответы испытуемых. Более перспективной оказалась линия, связанная с использованием дистрибутивной семантики. Основная идея дистрибутивных моделей заключается в том, что слова, встречающиеся в похожих контекстах, с большой вероятностью будут иметь сходное значение. Модели призваны описать все слова с помощью векторов (эмбеддингов) определенной размерности.

Семантическую близость слов при этом можно оценить как расстояние между векторами, с помощью которых слова представлены в моделях.

Метод латентно-семантического анализа (LSA) позволяет количественно выразить семантическую близость слов путем подсчета частоты слов в большом объеме документов. Первым шагом LSA является построение матрицы, где каждое слово представлено строкой, а каждый документ — столбцом. Ячейки этой матрицы заполняются числом вхождения определенного слова в соответствующий документ. Затем матрица преобразуется так, чтобы реже встречающиеся слова имели больший вес (так как менее частотные слова обычно сообщают более конкретные смыслы), а более часто встречающиеся слова — меньший вес. После применения к матрице сингулярного разложения (сходного с анализом главных компонент) размерность пространства уменьшается (обычно до 300–400 измерений). В полученном семантическом пространстве слова представлены как векторы, расстояние между которыми можно измерить, взяв обратный косинус угла между векторами (значения, близкие к единице, интерпретируются как несвязанные слова, нулевые значения — как одинаковые слова). Также можно измерять расстояния не просто между отдельными словами, а между фразами (двумя ответами испытуемого, например) путем объединения векторов слов в центроиды.

И. Форстер и К. Данбар впервые применили LSA для анализа ответов на тест «Необычное использование» в 2009 г. (Forster, Dunbar, 2009). Интересно отметить, что они рассчитывали семантическое расстояние не между словом-стимулом и ответами испытуемого (как это будет делаться в последующих исследованиях), а сравнивали ответы испытуемого с ответами других людей или с наиболее типичным использованием предмета. Именно последняя оценка (сравнение с наиболее типичным использованием) коррелировала с экспертными оценками креативности ($r = 0.60$), а включение в регрессионную модель беглости (количество идей) и разработанности (количество использованных слов) наряду с оценкой LSA позволило объяснить 64% дисперсии экспертных оценок креативности. Так как экспертные оценки сильно подвержены искажениям из-за вариаций в согласованности экспертов, самым большим преимуществом LSA является стандартизация процедуры присваивания баллов.

Получив в руки мощный инструмент, исследователи стали применять его для тестирования теоретических моделей. Так, например, Д. Дюма и К. Данбар выяснили, что два измерения креативности — беглость и оригинальность (оцененная с помощью LSA) — представляют собой связанные (0.38), но отдельные факторы (Dumas, Dunbar, 2014). Более того, оценка конструктной валидности с помощью факторного анализа продемонстрировала более высокие показатели именно для оригинальности (H = 0.816) по сравнению с беглостью (H = 0.631). Д. Дюма и М. Ранко (Dumas, Runco, 2018) показали, что показатели оригинальности (по LSA), очищенные от влияния беглости (т.е. взятые как остатки регрессии оригинальности на беглость), обладают достаточной внутренней согласованностью, что свидетельствует о существовании фактора оригинальности, независимого от беглости. В еще одной статье Дюма использовал LSA для проверки пороговой теории креативности — теории,

согласно которой связь интеллекта и креативности наблюдается до определенного уровня (порога) интеллекта, а потом эти показатели становятся независимыми (Dumas, 2018). Дюма показал, что пороговая теория справедлива для оригинальности, но не для беглости. При этом важным теоретическим продвижением стал вывод о том, что переменной, на которой располагается порог, должна быть креативность (оригинальность), а не интеллект (т.е. до определенного уровня креативности интеллект связан с оригинальностью, а потом эти две способности становятся независимыми). Р. Бити с соавт. продемонстрировали, что семантические расстояния между ответами в тесте на ассоциативную беглость совместно с управляющими функциями значимо предсказывают результаты теста «Необычное использование». Тем самым авторы показали роль как восходящих, так и нисходящих процессов в творческом мышлении (Beaty et al., 2014).

Также было изучено значительное влияние инструкции на проявление творческих способностей. Р. Прабхакаран с соавт. предложили новый инструмент диагностики креативности — придумывание глагола в ответ на существительное (Prabhakaran et al., 2014). Испытуемым либо давалась, либо не давалась инструкция быть креативными. С помощью LSA считалась оригинальность предложенного ответа. Полученные оценки семантической дистанции высоко коррелировали с фактором, образованным традиционными мерами креативности (дивергентное мышление, невербальные задания теста Торренса, сочинение историй, опросник креативности) ($\beta = 0.5$), при этом значимая корреляция сохранялась даже при контроле интеллекта ($\beta = 0.3$). Однако эти связи наблюдались только в условиях, когда испытуемым специально давалась установка на креативность. В продолжение этого исследования Д. Хайнен и Д. Джонсон задались вопросом о том, какой аспект креативности (новизна или пригодность) и в какой мере отражается в измерении семантического расстояния (Heinen, Johnson, 2018). Испытуемым также требовалось придумывать глагол в ответ на существительное в соответствии с тремя инструкциями: 1) дать творческий ответ; 2) дать оригинальный ответ; 3) дать наиболее пригодный ответ. Ответы испытуемых были оценены экспертами с точки зрения креативности, новизны и пригодности, а также были вычислены семантические дистанции между существительными и сгенерированными глаголами. Во-первых, семантические расстояния на базе LSA высоко коррелировали с оценками как креативности (0.47–0.71), так и новизны (0.69–0.97). Во-вторых, на основе сопоставления условий с разными инструкциями авторы приходят к выводу, что участники, руководствовавшиеся инструкцией давать творческий ответ, жертвовали некоторой новизной, чтобы соответствовать критерию пригодности и, таким образом, достигнуть действительно творческих ответов (учитывающих оба аспекта креативности). Авторы считают, что именно семантическая дистанция (в условиях установки на творческий продукт) позволяет наиболее точно оценить истинную креативность, включающую в себя как новизну, так и практическую применимость.

LSA также использовался для изучения установок и межгрупповых различий в проявлениях креативности. Исследование Д. Дюма и К. Данбар (Dumas,

Dunbar, 2016) было посвящено тому, как стереотипы влияют на дивергентное мышление. Участникам эксперимента было предложено при выполнении теста «Необычное использование» представить себя либо «эксцентричным поэтом», либо «строгим библиотекарем». Результаты показали, что участники, проходившие тестирование в роли «эксцентричного поэта», продемонстрировали более высокие результаты (как по семантической дистанции, так и по беглости), чем участники из другой «стереотипной» группы и контрольной группы. Авторы делают важный теоретический вывод о том, что, возможно, представление о креативности как устойчивой индивидуальной способности неверно. Дюма и Стрикленд обнаружили, что вредоносная креативность предсказывается полом (мужчины более склонны придумывать зловредные идеи) и (независимо от пола) оригинальностью, но не беглостью (Dumas, Strickland, 2018).

Проблемы с LSA

LSA оказался весьма привлекательным инструментом для исследования дивергентного мышления, позволял получить оценки оригинальности (удаленности), как правило, имеющие преимущество перед экспертными оценками в отношении внутренней согласованности (см., например: Dumas, Runco, 2018; Forster, Dunbar, 2009), давая возможность исследовать важные теоретические проблемы. Однако по мере накопления опыта в использовании LSA для обработки тестов дивергентного мышления стали возникать вопросы к его валидности и универсальности.

В первую очередь, исследователь принимает важное решение о том, какой корпус текстов использовать для построения семантического пространства. В большинстве первых исследований применялся так называемый корпус TASA (Touchstone Applied Science Associates), включающий около 37 тыс. документов и более 11 млн слов и составленный для приблизительного охвата опыта чтения типичного англоговорящего студента («general-reading-up-to-the-first-year-in-college»). Как мы увидим далее (Forthmann et al., 2019), по всей видимости, использование этого корпуса текстов может занижать корреляции с внешними критериями за счет недостаточного объема. Другой аспект связан с содержанием текстов. В своем исследовании Р. Хасс изучал кластеризацию ответов на тест «Необычное использование» (Hass, 2017a). Он показал, что экспертные оценки сходства между ответами в кластере (например: «использовать кирпич в качестве веса», «использовать кирпич, чтобы держать дверь открытой», «использовать кирпич, чтобы держать дверь закрытой») гораздо более высокие, чем оценки, полученные с помощью LSA. Вывод Хасса таков: «То, что LSA не может сделать, это предоставить меру сходства этих ответов в контексте возможных применений для кирпича. Иными словами, рассматриваемые по отдельности, слова “вес” и “дверь” не встречаются достаточно часто, чтобы иметь сходство по LSA. Тем не менее, как подсказывает здравый смысл, ответ “удерживание двери, подпертой кирпичом”, основан на факте, что кирпичи обычно имеют значительный вес» (Ibid., p. 354). При этом в том же исследовании Хасс показал, что динамика оценок семантической

близости между вопросом и последовательными ответами испытуемых соответствует теоретически ожидаемой. Таким образом, для принятия решения о возможности использования LSA для конкретной задачи следует учитывать, насколько планируемое к использованию семантическое пространство адекватно исследуемым вопросам.

Другая проблема связана с нестабильностью результатов: не во всех исследованиях и не для всех тестов дивергентного мышления оценки семантической дистанции достаточно высоко коррелируют с экспертными оценками. Так, например, Хасс обнаружил, что для задачи, где надо придумать примеры определенной категории (например, «Круглые вещи»), корреляции экспертных оценок с оценками LSA крайне низкие. Для теста «Необычное использование» корреляционные связи в исследовании Хасса также были заметно ниже, чем в других исследованиях (около 0.2) и существенно зависели от размерности пространства LSA (Hass, 2017b). В работе Н. Ла Вуа с соавт., напротив, были получены очень высокие корреляции оценок оригинальности по LSA с экспертными оценками для теста «Последствия» ($r = 0.94$), но, видимо, это связано с особенностями метода, который был применен для подсчета оригинальности (LaVoie et al., 2020). В исследовании С. Ачара с соавт. методы дистрибутивной семантики также показали себя хуже в отношении теста «Просто представь», по сравнению с тестом «Необычное использование» (Acar et al., 2023). Б. Фортман с соавт. продемонстрировали, что по крайней мере часть этой проблемы может быть связана с тем, что семантические расстояния, подсчитанные с помощью LSA, искажаются длиной ответа (разработанностью). Они обнаружили, что корреляция экспертных оценок креативности с семантической дистанцией повышается с 0 до 0.553 при контроле длины ответа. Авторы считают, что это связано с артефактом оценки семантических расстояний: взятые случайным образом косинусы имеют тенденцию быть выше (а расстояния, соответственно, ниже) для более длинных фраз по сравнению с более короткими (Forthmann et al., 2017). В другом исследовании Б. Фортман с соавт. явным образом продемонстрировали этот эффект с помощью компьютерной симуляции, а также предложили варианты для коррекции нежелательных смещений (Forthmann et al., 2019). Для каждого потенциально возможного количества слов в ответе (одно, два, три и т.д.) авторы отобрали из используемого пространства LSA по 10 тыс. случайных наборов слов, посчитали семантические расстояния между ними и фразой, используемой в качестве стимула, и показали, что семантические расстояния систематически уменьшаются по мере увеличения количества слов в ответе. Такое смещение, по всей видимости, является следствием метода сложения векторов для репрезентации фраз. Авторы предложили, во-первых, исключать стоп-слова из ответов испытуемых (что в первую очередь сокращает количество слов в ответе), а во-вторых, использовать коррекцию смещения (вычитать из косинуса реальных ответов косинус случайных ответов). В качестве критерия для оценки качества LSA-оценок оригинальности авторы использовали экспертные оценки и обратные показатели частотности ответов (редкость ответа). Наиболее высокие корреляции между семантическим расстоянием и внешними критериями

(экспертные оценки — $r = 0.477$; редкость — $r = 0.507$) были обнаружены при одновременном удалении стоп-слов, коррекции смещения и использовании более крупного семантического пространства (EN 100k vs TASA). Также примечательно, что и в этой работе конвергентная валидность была выше для теста «Необычное использование», чем для теста «Последствия».

За пределами LSA: предсказательные модели и GloVe

Описанные выше сложности с LSA, а также эволюция искусственного интеллекта в области обработки естественного языка логично привели к попыткам использовать другие модели для анализа результатов по тестам креативности.

В первую очередь исследователи стали изучать возможности моделей Word2Vec (Mikolov et al., 2013) и GloVe (Global Vectors for Word Representation) (Pennington et al., 2014). Word2Vec является нейросетью неглубокого обучения. В ее основе лежит представление о том, что похожие слова встречаются в похожих контекстах. На выходе слова, имеющие похожий смысл, будут иметь близкие числовые векторы, а связи между не встречающимися совместно словами будут минимизированы. В Word2Vec существует два основных подхода: предсказание слова на основе его контекста (Continuous Bag of Words, CBOW) и предсказание контекстных слов на основе целевого слова (Skip-Gram). Модель GloVe, так же как и LSA, использует матричное разложение и создает векторные представления слов, основываясь на том, насколько часто слова встречаются вместе в текстовом корпусе. Модель настраивает параметры так, чтобы полученные векторы наилучшим образом соответствовали статистике встречаемости слов.

В исследовании Д. Дюма с соавт. изучались возможности четырех моделей (по сравнению с экспертными оценками) в отношении внутренней согласованности получаемых оценок оригинальности для теста «Необычное использование», а также в отношении предсказательной валидности (корреляции с беглостью, разработанностью, самооценками креативности и личностных черт) (Dumas, Organisciak, Doherty, 2021). В исследовании использовались две модели на основе LSA (TASA и EN 100k), GloVe и Word2Vec (Skip-Gram). Лучшую внутреннюю согласованность (0.94) продемонстрировали экспертные оценки, модель GloVe (0.80) и модель TASA (0.81) следовали за ними. Остальные модели показали более низкую согласованность (0.73–0.74). Ни одна из оценок оригинальности значимо не коррелировала с самоотчетными методиками креативности, но оценки по GloVe при этом коррелировали (хоть и незначимо) наиболее похожим на экспертные оценки образом. Также оценки GloVe наименьшим образом (0.23), по сравнению с другими семантическими дистанциями (0.24–0.24), коррелировали с разработанностью, что тоже было приближено к экспертным оценкам (0.22). Авторы делают вывод о преимуществе модели GloVe для оценки оригинальности ответов на дивергентные тесты и связывают это преимущество с несколькими причинами: с очень большим размером и неспецифичностью содержания обучающего корпуса, а

также с применением вероятностного моделирования. Сходные результаты о преимуществе GloVe перед моделями LSA получили и С. Ачар с соавт. (Acar et al., 2023). По результатам исследования Д. Дюма с соавт. представили онлайн-платформу Open Creativity Scoring (OCS) (Organisciak, Dumas, Acar, de Chantal, 2023). С помощью этой платформы К. Гражзель с соавт. (Grajzel et al., 2023) провели дополнительный анализ данных из исследования Ачара с соавт. (Acar et al., 2023). Были проанализированы показатели гибкости, полученные в результате обработки тестов «Необычное использование» и «Просто представь». Если оригинальность понимается как семантическая удаленность каждого ответа испытуемого от слова-стимула, то гибкость — это семантические расстояния между последовательными ответами испытуемых (для агрегирования при этом применялись разные методы — суммирование, усреднение или максимальная оценка). Сопоставлялись оценки оригинальности, беглости и гибкости, посчитанные согласно руководству к тесту Торренса и вычисленные с помощью модели GloVe на платформе OCS. Оценки гибкости на основе семантических расстояний значимо коррелировали с «традиционными» оценками (r от 0.25 до 0.72 в зависимости от способа агрегирования). С помощью конфирматорного факторного анализа для теста «Необычное использование» было показано, что усреднение семантических расстояний максимизирует корреляцию гибкости с оригинальностью ($r = 0.79$) и минимизирует корреляцию гибкости с беглостью ($r = 0.32$). Напротив, использование максимальных оценок максимизирует корреляцию гибкости с беглостью ($r = 0.63$) и минимизирует — гибкости с оригинальностью ($r = 0.12$). Дополнительные преимущества использования максимальной оценки (вместо усреднения семантических расстояний для всех слов в ответе) показаны в работе Ю. Ю с соавт. (Yu et al., 2023).

Р. Бити и Д. Джонсон (Beaty, Johnson, 2021) оценивали оригинальность для теста «Необычное использование» и «Теста творческих ассоциаций» (придумывание глаголов в ответ на существительное) в пяти семантических пространствах. Три из них построены как модели CBOW, предназначенные для предсказывания слов исходя из окружающего их контекста, аналогично Word2Vec, и различающиеся размерностью и объемом обучающих корпусов. Еще два — TASA на основе LSA и GloVe. Представленный в статье анализ обладал следующими особенностями: 1) применялся метод умножения векторов (а не сложения) в случае вычисления вектора для фраз; 2) автоматические оценки оригинальности были получены не для каждого пространства в отдельности, а в виде латентного фактора всех этих пространств, вычисленного в результате конфирматорного факторного анализа. Латентный фактор отражает общую дисперсию семантических расстояний, вычисленных разными методами. В пяти исследованиях было обнаружено, что латентный фактор коррелировал с экспертными оценками креативности высоко и значимо (коэффициенты корреляции варьировали от 0.45 до 0.91), часто (но не всегда) превосходя корреляции для отдельных семантических пространств. Было показано, что использование метода умножения векторов (а не сложения) позволяет устранить проблему, связанную с отрицательной корреляцией

между семантическими расстояниями и длиной ответа (разработанностью). В этом исследовании было показано, что латентный фактор семантического расстояния (так же как и экспертные оценки) положительно коррелирует с разработанностью (на уровне около 0.4). Последний результат особенно примечателен, так как обычно семантические расстояния коррелируют с количеством слов отрицательно. Оценки семантической дистанции также положительно коррелировали с показателями когнитивных и самооценочных мер креативности (придумывание метафор, творческая самоэффективность), но не с когнитивными и личностными факторами (флюидный интеллект и открытость опыту). Результатом исследования авторского коллектива также стала онлайн-платформа — SemDis.

Затем Р. Бити с соавт. изучали, как влияют особенности стимульного материала теста «Необычное использование» на надежность и валидность оценок семантических расстояний (Beaty et al., 2022). По результатам двух исследований авторы разработали рекомендации для повышения достоверности оценок оригинальности с использованием семантического расстояния в тесте «Необычное использование». В частности, авторы отобрали из 46 объектов-стимулов 13, по которым оценки семантических расстояний получаются наиболее согласованными, рекомендуют использовать инструкцию «быть креативным» и избегать многословных стимулов (т.е. не использовать словосочетания типа «гитарная струна»). Также было показано, что при использовании нескольких семантических пространств способ агрегации результатов для подсчета итогового балла (простое усреднение, вычисление факторных оценок и т.д.) существенно не влияет на валидность и надежность полученных оценок (Forthmann et al., 2023).

Бити с соавт. (Beaty et al., 2021) исследовали понятие потока, направленного вперед (Gray et al., 2019), для описания динамики ассоциативных связей, последовательно возникающих в процессе блуждания мыслей. Здесь они использовали тот же подход, что и в предыдущей работе (латентный фактор семантических расстояний для описания потока, объединение семи семантических пространств (четыре модели CBOW, две модели LSA и модель GloVe)). Было показано, что поток, направленный вперед, предсказывает дополнительную дисперсию в экспертных оценках дивергентного мышления, сверх дисперсии, которую предсказывают интеллектуальные способности. С теоретической точки зрения это свидетельствует о том, что ассоциативные способности являются уникальным предиктором креативности, который не является избыточным по отношению к общим когнитивным способностям (см. также: Beaty et al., 2014).

Большие языковые модели

Несмотря на существенное продвижение в области автоматической оценки творческих продуктов, ни одной из предложенных моделей дистрибутивной семантики не удалось в полной мере имитировать оценки креативности, выставляемые экспертами, особенно если речь шла не об однословных ответах,

а об идеях, выраженных в виде целых текстов (Wang et al., 2023). Конкурентами контекстно-независимых моделей дистрибутивной семантики в исследованиях творчества очень быстро стали модели, основанные на нейросетях глубокого обучения, позволяющие не только учитывать частоту встречаемости слов, но и распознавать значение слов в зависимости от контекста. Эти глубокие нейронные сети требуют обучения на огромных объемах текста и имеют много слоев, каждый из которых создает различные контекстно-зависимые эмбеддинги, кодирующие различные типы морфосинтаксической и семантической информации. Среди моделей, испробованных на сегодняшний день исследователями творчества, находятся BERT, RoBERTa, GPT-2, GPT-3, T5.

Д. Джонсон с соавт. (Johnson et al., 2022) предложили понятие дивергентной семантической интеграции (ДСИ) для оценки креативности текстов (нарративов), создаваемых испытуемыми. ДСИ характеризует степень, в которой текст объединяет не связанные или отдаленно связанные идеи. Показатель ДСИ — это средняя семантическая дистанция между всеми словами текста. Авторы сравнили эффективность шести моделей (три модели CBOW, две модели LSA, модель GloVe и BERT) на материале разных исследований, в которых испытуемым надо было придумывать короткие истории. Во всех исследованиях модель BERT превосходила другие модели, демонстрируя корреляции с экспертными оценками вплоть до 0.85. В своем исследовании Джонсон с соавт. в том числе провели реанализ данных К. Зеделиус с соавт. (Zedelius et al., 2019). Изначально Зеделиус с соавт. оценивали написанные испытуемым истории с помощью лингвистических инструментов — Coh-Metrix (оценка связанности и согласованности текстов, синтаксической простоты и конкретности слов) и LIWC (Linguistic Inquiry and Word Count — лингвистический анализ на основе частоты слов, попадающих в различные лингвистические и психологические категории). Полученные ими лингвистические оценки слабо коррелировали с экспертными оценками разных аспектов креативности текстов, особенно с оригинальностью. При этом модель BERT (Johnson et al., 2022) смогла предсказать оригинальность полученных текстов на уровне $r = 0.35$. Вместе с тем, как отмечают авторы, одна из проблем, которая может быть связана с диагностикой креативности по ДСИ, — это то, что ДСИ будет велика как для действительно креативных текстов, так и для случайных наборов слов.

Модели глубокого обучения позволили сделать следующий шаг в понимании возможностей искусственного интеллекта оценивать человеческую креативность. Дж. Паттерсон с соавт. провели кросс-культурное исследование, собрав базы ответов с экспертными оценками для теста «Необычное использование» на 12 языках (включая русский) (Patterson, Merseal et al., 2023). Они использовали многоязычные версии BERT и RoBERTa — MBERT и XLMR. Так как эти модели содержат множество слоев, задача состояла в том, чтобы для каждого языка выявить модель и слои, которые наилучшим образом предсказывают экспертные оценки. На половине языков (включая английский) лучшей моделью была MBERT, и в 4 из 12 случаев усреднение двух наиболее

высоко коррелировавших с экспертными оценками слоев модели давало наилучший результат. Самые высокие корреляции с экспертными оценками получены для данных на английском языке (r = 0.52), а самые низкие — для иврита, китайского и французского (0.23–0.24). Корреляции с внешними критериями (самооценками креативности, творческими достижениями и открытостью опыту) были невысокими (в основном в пределах от 0 до 0.3), что несколько ниже корреляций этих показателей с экспертными оценками, но согласуется с полученными ранее данными. В целом, исследование не продемонстрировало стабильности результатов на материале разных языков. Подавляющее число описанных в данной статье исследований было проведено на английском. Напрашивается вывод о том, что прямая экстраполяция результатов и выводов на другие языковые контексты по меньшей мере не очевидна.

Обучение с учителем

Еще одним направлением развития в области автоматической оценки креативности стало использование обучения с учителем и дообучения имеющихся больших языковых моделей. На обучающей выборке модель учится предсказывать интересующую переменную исходя из совокупности имеющихся признаков. Эффективность обучения (насколько хорошо удается предсказать значение переменной) оценивается на тестовой выборке. Для континуальных переменных используются регрессионные модели, для дискретных — модели классификации. При дообучении стандартные языковые модели учатся на примерах экспертных оценок распознавать оригинальность предложенных стимулов.

К. Стивенсон с соавт. (Stevenson et al., 2020) использовали гибридный подход, сочетающий в себе обучение без учителя (кластеризацию) и обучение с учителем в виде получения прогнозов на основе наблюдаемых примеров. На первом этапе была создана база ответов (более 70500 ответов) на тест «Необычное использование» и их оценок, данных экспертами. Ответы были кластеризованы на основе семантических расстояний (использовалась модель Word2Vec). В результате усреднения всех экспертных оценок, специфичных для кластера, получалась репрезентативная средняя оценка креативности для каждого кластера. Новым ответам присваивался балл по креативности семантически ближайшего кластера. Почти по всем показателям валидности и надежности автоматически полученные оценки на высоком уровне коррелировали с оценками экспертов.

П. Бюжак с соавт. (Buczak et al., 2023) использовали обучение с учителем для прогнозирования оценок результатов диагностики дивергентного мышления. Они сравнили три алгоритма машинного обучения — Random Forest, XGBoost и Support Vector Regression. Признаки для обучения включали в себя формальные характеристики ответов (количество слов, средняя длина слова, максимальная длина слова), характеристики, вычисленные на основе эмбеддингов (с использованием GloVe и Word2Vec), и сами эмбеддинги. Модели,

включающие все три типа данных, обладали наилучшей предсказательной способностью. Наиболее значимыми предикторами экспертных оценок оказались семантические расстояния и количество слов в ответе. Сравнение трех алгоритмов машинного обучения показало, что Random Forest и XGBoost, как правило, немного превосходили Support Vector Regression.

П. Органайсчак с соавт. (Organisciak, Acar, Dumas, Berthiaume, 2023) дообучили нейросети GPT-3 и T5 на основе 27000 ответов на тест «Необычное использование» из девяти предыдущих исследований и получили впечатляющие результаты. Авторы сравнили разные модели по способности предсказывать экспертные оценки креативности. Лучшей моделью оказалась GPT-3 (корреляция с экспертными оценками достигала 0.81), а корреляции с семантическими расстояниями, посчитанными с помощью платформ SemDis (Beaty, Johnson, 2021) и OSC (Dumas, Organisciak, Doherty, 2021), на этих данных составили всего 0.12 и 0.26 соответственно. Авторы также обучили GPT-3 на примерах определенных заданий (использование кирпича, веревки, ножа и т.д.) и в качестве теста предложили модели оценить креативность ответов на другие слова-стимулы (ложка, бутылка и т.д.). Даже в этом случае лучший вариант модели демонстрировал корреляция с экспертными оценками на уровне 0.63. Более того, авторы проверили способность разных версий GPT оценивать креативность без дообучения. Например, они пробовали обучение с нулевой разметкой, где модели просто предлагалось ответить на вопрос об оригинальности (по шкале от 10 до 50) предложенных способов использования предмета. В случае обучения на небольшом количестве примеров модели предлагались 5 или 20 способов использования предмета с экспертными оценками и ряд других ответов (без оценок), которые модель должна была оценить сама. GPT-4 превзошла в этих заданиях свои предыдущие версии (GPT-3 и GPT-3.5): корреляция с экспертными оценками составила 0.56 для обучения с нулевой разметкой, 0.66 — для обучения на пяти примерах и 0.70 — для обучения на двадцати примерах. Сходным образом в исследовании П. ДиСтефано с соавт. (DiStefano et al., 2023) и в исследовании С. Лучини с соавт. (Luchini et al., 2023) дообученные модели RoBERTa и GPT-2 научились оценивать креативность метафор и ответов в «Тесте творческого решения повседневных проблем» так, что корреляция с экспертными оценками достигала значений 0.70–0.83.

Можно констатировать, что специальным образом дообученные большие языковые модели способны оценивать оригинальность ответов на творческие задания на уровне, мало отличающемся от оценок людей. Остается, однако, вопрос, являются ли внешне сходные (высоко коррелирующие) оценки одинаковыми по природе, ориентируется ли искусственный интеллект на те же паттерны, на которые ориентируется человек, принимая решение о креативности оцениваемого материала, — иными словами, «думает» ли нейронная сеть подобным человеку образом? Для исследователей творчества ответ на этот вопрос имеет большое значение, потому что от него зависит возможность построения моделей творческого мышления человека на основе данных, полученных с помощью искусственного интеллекта.

Онлайн-сервисы для оценки вербальной креативности

Открытый код и доступ к данным играют ключевую роль в современном научном исследовании и технологическом развитии. Они обеспечивают репродуцируемость, содействуют совместной работе и обмену идеями между учеными, способствуют инновациям и развитию открытого и коллективного подхода к науке. Хорошим тоном в работах последнего времени стало предоставление свободного доступа не только к данным, но и к программному коду для их обработки (в основном на платформе https://osf.io/). Исследователи творчества, осознавая важность обмена идеями, стали создавать онлайн-системы, облегчающие работу коллег.

Подавляющее большинство работ, оценивающих семантические расстояния для тестов дивергентного мышления на основе LSA и Word2Vec, использовали семантические пространства, представленные на сайте https://sites.google.com/site/fritzgntr/software-resources/semantic_spaces (Günther et al., 2015). Можно скачать модели для английского, немецкого, французского, итальянского, испанского и хорватского языков.

Платформа Open Creativity Scoring (https://openscoring.du.edu/), разработанная на основе результатов работы Дюма с соавт. (Dumas, Organisciak, Doherty, 2021; Organisciak, Acar, Dumas, Berthiaume, 2023), дает возможность оценить семантические расстояния между вербальным стимулом и ответами испытуемых на него с использованием модели GloVe (OCS) или GPT-3 (Ocsai). И в том, и в другом случае можно выбрать разные варианты моделей, а также при желании автоматически удалить стоп-слова или настроить взвешивание для лучшего контроля количества слов в ответе. Предусмотрены введение данных в специальное поле или загрузка файла для обработки. Результат выдается в виде таблицы, которую можно выгрузить в файл. Для модели GloVe вычисляются семантические расстояния и разработанность, а для GPT-3 — балл оригинальности (от 1 до 5) и разработанность. Важно отметить, что подсчеты на основе модели GPT-3 можно делать не только на английском, но и на русском языке (хотя качество и валидность этих подсчетов не ясны).

Р. Бити и Д. Джонсон (Beaty, Johnson, 2021) предлагают свой вариант онлайн-платформы для обработки вербальных тестов — SemDis (semdis.wlu.psu.edu/). Эта платформа предназначена для обработки результатов тестов на дивергентное мышление и словесные ассоциации, а также для подсчета дивергентной семантической интеграции для текста или поиска ближайших ассоциаций к определенному слову. Можно выбрать между тремя семантическими пространствами: CBOW, моделью LSA (TASA) или моделью GloVe. Также можно выбрать способ, которым объединяются вектора слов в фразах (сложение или умножение), имеется возможность построения графиков. На сайте приведены подробные инструкции для работы с системой и даны ссылки на OSF, где выложены материалы по расчету оценок ДСИ с использованием BERT.

К. Грей с соавт. (Gray et al., 2019) представили платформу «Поток, направленный вперед» (Forward Flow, http://www.forwardflow.org/). Это мера, подсчитанная на основе LSA и характеризующая динамику ассоциативных связей, последовательно возникающих в процессе блуждания мыслей. В своих исследованиях авторы продемонстрировали, что мера потока значимо коррелирует как с результатами дивергентных тестов мышления, так и с профессиональными творческими достижениями. На сайте можно проанализировать список слов (свободных ассоциаций) или загрузить целую таблицу с данными (последовательность слов-ассоциаций) и получить в качестве результата матрицу семантических расстояний между ответами, характеризующую поток мыслей.

Перечисленные выше платформы в основном предназначены для исследователей, использующих англоязычный стимульный материал. На русском языке существует сайт RusVectōrēs (https://rusvectores.org/ru/). В отличие от англоязычных сервисов, он не был разработан исследователями креативности, но может быть использован для расчета семантических расстояний между словами, подбора ближайших ассоциаций к заданным словам, визуализации семантических связей между словами, скачивания моделей, как контекстуализированных (ELMo), так и статических (Word2Vec, FastText), обученных на разных текстовых корпусах.

Заключение

Оценка результатов тестирования творческого мышления до недавнего времени являлась трудоемким занятием, требующим человеческих ресурсов и занимающим много времени. Автоматизированная оценка результатов диагностики творческого мышления представляет собой актуальную тему в исследованиях креативности, изучение ее возможностей в основном проводятся на материале тестов дивергентного мышления и в подавляющем большинстве случаев (на сегодняшний день) — на вербальном материале.

Первоначально усилия исследователей были направлены на изучение потенциала дистрибутивной семантики. В первую очередь стали применять латентно-семантический анализ, позволяющий вычислять семантические расстояния между словами. Выяснилось, что применение LSA дает высокосогласованные оценки, но иногда они обладают низкой валидностью (низкие корреляции с критериями — например, с экспертными оценками). В некоторых работах были показаны способы улучшить валидность получаемых оценок (исключение стоп-слов, коррекция смещений, использование более крупных корпусов текстов), однако постепенно акцент сместился на другие подходы. Из контекстно-независимых моделей наилучшим образом проявила себя модель GloVe; в целом ряде исследований было показано, что оценки оригинальности, полученные на ее основе, обладают психометрическими преимуществами. Однако на текущий момент самых впечатляющих результатов удалось добиться с помощью дообученных контекстно-независимых больших языковых моделей. Они позволяют обрабатывать длинные ответы и демонстрируют

впечатляющую согласованность с экспертными оценками. Недостатком применения этих моделей, на наш взгляд, является отсутствие понимания, на каких основаниях выносится суждение об оригинальности творческих продуктов. В области оценки продуктов творческого мышления наблюдается характерная для всех сфер картина: искусственный интеллект добился потрясающей возможности прогнозировать, «понимать» и имитировать человеческое поведение, но при этом более далек от постижения его механизмов, чем, например, психология. С этой точки зрения в области изучения креативности более простые методы, основанные на подсчете семантических расстояний, дают более прозрачный и интерпретируемый результат. При этом наблюдается характерное смещение акцента в работах по мере перехода от LSA к нейросетям глубокого обучения — с содержательного (исследование вопросов теоретического характера, касающихся механизмов креативности) на методический (исследование нюансов подсчетов, позволяющих улучшить надежность и валидность оценок). Логичным направлением дальнейшей работы в этой области является применение объяснимого искусственного интеллекта.

Для русскоязычных исследователей творчества большим ограничением является то, что практически все работы в этой области проводятся на английском языке, а те немногие, которые включают русскоязычный материал (Patterson, Merseal et al., 2023), используют не очень большую выборку. Одним из препятствий к широкомасштабным исследованиям на русском языке является отсутствие доступа к качественным и обширным корпусам данных по тестам креативности, на которых можно было бы тестировать и сравнивать разные модели. Следует также отметить, что важно развивать кооперацию психологов и специалистов по искусственному интеллекту: первые могут задавать осмысленные вопросы в соответствующей предметной области, а вторые — искать на них ответы с помощью наиболее современных методов работы с данными.

Другим направлением должна стать разработка систем автоматической оценки тестов невербальной (рисуночной) креативности. На момент публикации данной статьи описаны только два инструмента для автоматической оценки результатов диагностики невербальной креативности с использованием искусственного интеллекта. Первый был предложен для оценивания результатов «Рисуночного теста творческого мышления» (TCT-DP) К. Урбана (Cropley, Marrone, 2022; Urban, 2005). В качестве экстрактора признаков использовалась предварительно обученная модель MobileNets. Она представляет собой сверточную нейронную сеть, состоящую из 53 слоев и обученную на более чем миллионе изображений, взятых из датасета ImageNet. В исследовании использовалось всего 414 изображений для решения задачи классификации результатов теста по нескольким уровням креативности (низкий, средний, высокий). Точность работы модели составила 94.2%. Одним из недостатков ее является тот, что она не представлена в открытом доступе. Единственной моделью для автоматической оценки образной креативности, находящейся в открытом доступе, является платформа для автоматизированной оценки рисунков (Automated Drawing Assessment, AuDrA) (Patterson,

Barbot et al., 2023). Данная модель также была обучена с помощью сверточной нейронной сети ResNet на основе более одиннадцати тысяч рисунков и их оценок, данных пятьюдесятью экспертами. Результаты этих исследований показали высокую корреляцию между баллами, полученными автоматически, и оценками, данными экспертами.

В одной из наших работ (Panfilova et al., 2023) мы также показываем возможность использования нейросетей для предсказания результатов по тесту TCT-DP, используя при этом алгоритмы объяснимого искусственного интеллекта. Применение методов объяснимого искусственного интеллекта к обученной модели продемонстрировало соответствие выявленных зон активации определенным критериям для экспертной оценки. Было показано, что четкость формулировки критерия оказывает влияние на результат работы как эксперта, так и модели.

Подводя итоги, можно отметить, что использование методов автоматической оценки тестов на креативность представляет собой важный шаг в развитии психологии творчества. Автоматизированные методы позволяют быстро и эффективно оценивать большие объемы данных, снижают затраты на обучение и работу экспертов, исключают субъективное влияние экспертов, тем самым обеспечивая более консистентные, надежные и воспроизводимые результаты. Проведенная на большом объеме англоязычных данных валидизация автоматических методов оценки свидетельствует о достаточной надежности и теоретической обоснованности получаемых результатов (по крайней мере для тестов дивергентного мышления). Наши надежды связаны с тем, что, получив столь мощный инструмент, исследователи смогут использовать его для более точного и глубокого понимания механизмов творческого мышления.

References

Acar, S., Berthiaume, K., Grajzel, K., Dumas, D., Flemister, C. "Tedd", & Organisciak, P. (2023). Applying automated originality scoring to the verbal form of Torrance Tests of Creative Thinking. *Gifted Child Quarterly, 67*(1), 3–17. https://doi.org/10.1177/00169862211061874

Acar, S., & Runco, M. A. (2014). Assessing associative distance among ideas elicited by tests of divergent thinking. *Creativity Research Journal, 26*(2), 229–238. https://doi.org/10.1080/10400419.2014.901095

Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods, 53*(2), 757–780. https://doi.org/10.3758/s13428-020-01453-w

Beaty, R. E., Johnson, D. R., Zeitlen, D. C., & Forthmann, B. (2022). Semantic distance and the alternate uses task: Recommendations for reliable automated assessment of originality. *Creativity Research Journal, 34*(3), 245–260. https://doi.org/10.1080/10400419.2022.2025720

Beaty, R. E., Silvia, P. J., Nusbaum, E. C., Jauk, E., & Benedek, M. (2014). The roles of associative and executive processes in creative cognition. *Memory & Cognition, 42*(7), 1186–1197. https://doi.org/10.3758/s13421-014-0428-8

Beaty, R. E., Zeitlen, D. C., Baker, B. S., & Kenett, Y. N. (2021). Forward flow and creative thought: Assessing associative cognition and its role in divergent thinking. *Thinking Skills and Creativity, 41*, Article 100859. https://doi.org/10.1016/j.tsc.2021.100859

Beketayev, K., & Runco, M. A. (2016). Scoring divergent thinking tests by computer with a semantics-based algorithm. *Europe's Journal of Psychology, 12*(2), 210–220. https://doi.org/10.5964/ejop.v12i2.1127

Bossomaier, T., Harre, M., Knittel, A., & Snyder, A. (2009). A semantic network approach to the Creativity Quotient (CQ). *Creativity Research Journal, 21*(1), 64–71. https://doi.org/10.1080/10400410802633517

Buczak, P., Huang, H., Forthmann, B., & Doebler, P. (2023). The machines take over: A comparison of various supervised learning approaches for automated scoring of divergent thinking tasks. *Journal of Creative Behavior, 57*(1), 17–36. https://doi.org/10.1002/jocb.559

Cropley, D., & Marrone, R. (2022). Automated scoring of figural creativity using a convolutional neural network. *Psychology of Aesthetics Creativity and the Arts*. Advance online publication. https://doi.org/10.1037/aca0000510

Cseh, G. M., & Jeffries, K. K. (2019). A scattered CAT: A critical evaluation of the consensual assessment technique for creativity research. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 159–166. https://doi.org/10.1037/aca0000220

Diedrich, J., Benedek, M., Jauk, E., & Neubauer, A. C. (2015). Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts, 9*(1), 35–40. https://doi.org/10.1037/a0038688

DiStefano, P. V., Patterson, J. D., & Beaty, R. (2023). Automatic scoring of metaphor creativity with large language models [Preprint]. *PsyArXiv*. https://doi.org/10.31234/osf.io/6jtxb

Dumas, D. (2018). Relational reasoning and divergent thinking: An examination of the threshold hypothesis with quantile regression. *Contemporary Educational Psychology, 53*, 1–14. https://doi.org/10.1016/j.cedpsych.2018.01.003

Dumas, D., & Dunbar, K. N. (2014). Understanding fluency and originality: A latent variable perspective. *Thinking Skills and Creativity, 14*, 56–67. https://doi.org/10.1016/j.tsc.2014.09.003

Dumas, D., & Dunbar, K. N. (2016). The creative stereotype effect. *PLoS ONE, 11*(2), Article e0142567. https://doi.org/10.1371/journal.pone.0142567

Dumas, D., Organisciak, P., & Doherty, M. (2021). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts, 15*(4), 645–663. https://doi.org/10.1037/aca0000319

Dumas, D., Organisciak, P., Maio, S., & Doherty, M. (2021). Four text-mining methods for measuring elaboration. *Journal of Creative Behavior, 55*(2), 517–531. https://doi.org/10.1002/jocb.471

Dumas, D., & Runco, M. (2018). Objectively scoring divergent thinking tests for originality: A re-analysis and extension. *Creativity Research Journal, 30*(4), 466–468. https://doi.org/10.1080/10400419.2018.1544601

Dumas, D., & Strickland, A. L. (2018). From book to bludgeon: A closer look at unsolicited malevolent responses on the alternate uses task. *Creativity Research Journal, 30*(4), 439–450. https://doi.org/10.1080/10400419.2018.1535790

Forster, E. A., & Dunbar, K. N. (2009). Creativity evaluation through latent semantic analysis. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (pp. 602–607). Austin, TX: Cognitive Science Society.

Forthmann, B., Beaty, R. E., & Johnson, D. R. (2023). Semantic spaces are not created equal — How should we weigh them in the sequel?: On composites in automated creativity scoring. *European Journal of Psychological Assessment, 39*(6). https://doi.org/10.1027/1015-5759/a000723

Forthmann, B., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2017). Typing speed as a confounding variable and the measurement of quality in divergent thinking. *Creativity Research Journal, 29*(3), 257–269. https://doi.org/10.1080/10400419.2017.1360059

Forthmann, B., Oyebade, O., Ojo, A., Gьnther, F., & Holling, H. (2019). Application of latent semantic analysis to divergent thinking is biased by elaboration. *Journal of Creative Behavior, 53*(4), 559–575. https://doi.org/10.1002/jocb.240

Forthmann, B., Paek, S. H., Dumas, D., Barbot, B., & Holling, H. (2020). Scrutinizing the basis of originality in divergent thinking tests: On the measurement precision of response propensity estimates. *British Journal of Educational Psychology, 90*(3), 683–699. https://doi.org/10.1111/bjep.12325

Grajzel, K., Acar, S., Dumas, D., Organisciak, P., & Berthiaume, K. (2023). Measuring flexibility: A text-mining approach. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.1093343

Gray, K., Anderson, S., Chen, E. E., Kelly, J. M., Christian, M. S., Patrick, J., Huang, L., Kenett, Y. N., & Lewis, K. (2019). 'Forward flow': A new measure to quantify free thought and predict creativity. *American Psychologist, 74*(5), 539–554. https://doi.org/10.1037/amp0000391

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun — An R package for computations based on latent semantic analysis. *Behavior Research Methods, 47*(4), 930–944. https://doi.org/10.3758/s13428-014-0529-0

Hass, R. W. (2017a). Semantic search during divergent thinking. *Cognition, 166*, 344–357. https://doi.org/10.1016/j.cognition.2017.05.039

Hass, R. W. (2017b). Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications. *Memory & Cognition, 45*(2), 233–244. https://doi.org/10.3758/s13421-016-0659-y

Heinen, D. J. P., & Johnson, D. R. (2018). Semantic distance: An automated measure of creativity that is novel and appropriate. *Psychology of Aesthetics, Creativity, and the Arts, 12*(2), 144–156. https://doi.org/10.1037/aca0000125

Johnson, D. R., Kaufman, J. C., Baker, B. S., Patterson, J. D., Barbot, B., Green, A. E., Van Hell, J., Kennedy, E., Sullivan, G. F., Taylor, C. L., Ward, T., & Beaty, R. E. (2022). Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods, 55*(7), 3726–3759. https://doi.org/10.3758/s13428-022-01986-2

LaVoie, N., Parker, J., Legree, P. J., Ardison, S., & Kilcullen, R. N. (2020). Using latent semantic analysis to score short answer constructed responses: Automated Scoring of the Consequences Test. *Educational and Psychological Measurement, 80*(2), 399–414. https://doi.org/10.1177/0013164419860575

Luchini, S., Maliakkal, N. T., DiStefano, P. V., Patterson, J. D., Beaty, R., & Reiter-Palmon, R. (2023). Automatic scoring of creative problem-solving with large language models: A comparison of originality and quality ratings [Preprint]. *PsyArXiv*. https://doi.org/10.31234/osf.io/g5qvf

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*. https://doi.org/10.48550/arXiv.1301.3781

Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity, 49*, Article 101356. https://doi.org/10.1016/j.tsc.2023.101356

Organisciak, P., Dumas, D., Acar, S., & de Chantal, P. L. (2023). Open Creativity Scoring [Computer software]. University of Denver. https://openscoring.du.edu

Panfilova, A. S., Valueva, E. A., & Ilyin, I. Y. (2023). *The application of explainable artificial intelligence methods to models for automatic creativity assessment* [Manuscript under review].

Patterson, J. D., Barbot, B., Lloyd-Cox, J., & Beaty, R. E. (2023). AuDrA: An automated drawing assessment platform for evaluating creativity. *Behavior Research Methods*. Advanced online publication. https://doi.org/10.3758/s13428-023-02258-3

Patterson, J. D., Merseal, H. M., Johnson, D. R., Agnoli, S., Baas, M., Baker, B. S., Barbot, B., Benedek, M., Borhani, K., Chen, Q., Christensen, J. F., Corazza, G. E., Forthmann, B., Karwowski, M., Kazemian, N., Kreisberg-Nitzav, A., Kenett, Y. N., Link, A., Lubart, T., ... Beaty, R. E. (2023). Multilingual semantic distance: Automatic verbal creativity assessment in many languages. *Psychology of Aesthetics, Creativity, and the Arts, 17*(4), 495–507. https://doi.org/10.1037/aca0000618

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Prabhakaran, R., Green, A. E., & Gray, J. R. (2014). Thin slices of creativity: Using single-word utterances to assess creative cognition. *Behavior Research Methods, 46*(3), 641–659. https://doi.org/10.3758/s13428-013-0401-7

Reiter-Palmon, R., Forthmann, B., & Barbot, B. (2019). Scoring divergent thinking tests: A review and systematic framework. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 144–152. https://doi.org/10.1037/aca0000227

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal, 24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

Silvia, P. J., Martin, C., & Nusbaum, E. C. (2009). A snapshot of creativity: Evaluating a quick and simple method for assessing divergent thinking. *Thinking Skills and Creativity, 4*(2), 79–85. https://doi.org/10.1016/j.tsc.2009.06.005

Stevenson, C., Smal, I., Baas, M., Dahrendorf, M., Grasman, R., Tanis, C., Scheurs, E., Sleiffer, D., & van der Maas, H. (2020). *Automated AUT scoring using a Big Data variant of the Consensual Assessment Technique: Final technical report*. Modeling Creativity Project, Universiteit van Amsterdam.

Urban, K. K. (2005). Assessing creativity: The Test for Creative Thinking — Drawing Production (TCT-DP). *International Education Journal, 6*(3), 272–280.

Wang, K., Dong, B., & Ma, J. (2023). Testing computational assessment of idea novelty in crowdsourcing. *Creativity Research Journal*. Advance online publication. https://doi.org/10.1080/10400419.2023.2187544

Wilson, R. C., Guilford, J. P., & Christensen, P. R. (1953). The measurement of individual differences in originality. *Psychological Bulletin, 50*(5), 362–370. https://doi.org/10.1037/h0060857

Yu, Y., Beaty, R. E., Forthmann, B., Beeman, M., Cruz, J. H., & Johnson, D. (2023). A MAD method to assess idea novelty: Improving validity of automatic scoring using Maximum Associative Distance (MAD). *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication. https://doi.org/10.1037/aca0000573

Zedelius, C. M., Mills, C., & Schooler, J. W. (2019). Beyond subjective judgments: Predicting evaluations of creative writing from computational linguistic features. *Behavior Research Methods, 51*(2), 879–894. https://doi.org/10.3758/s13428-018-1137-1