

RUN-028. От источника к независимой разметке

KLT-RBDv6 / ARBD40. Проект «ЛОГИКА КУРПИШЕВА 2». Автор метода и научной постановки: Курпишев Иван Борисович. 13 сентября 2026 года.

Продолжение RUN-027 / ARBD39-LEARN. Цель выпуска - подготовить проверку авторской атрибуции и реперных ролей на русских и немецких источниках, а также явно учитывать неопределённость интерпретации в конечной предсказательной модели.

Результат выпуска

Восстановлены все пять английских файлов RUN-026: совпадают SHA-256 исходных байтов и 3 722 прежних сегмента. Добавлены восемь немецких страниц эссе «Beantwortung der Frage: Was ist Aufklärung?» (AA VIII, 35-42) и полный PDF-контейнер русского издания «Пролегомен» 1889 года в переводе Владимира Соловьёва: 387 сканированных страниц, 12 594 930 байтов.

Пилотный пакет содержит 21 фрагмент: девять немецких единиц (восемь основных блоков и одна сноска) и двенадцать русских страниц с непроверенным OCR. Это целевой малый пилот для настройки протокола, а не вероятностная выборка всего наследия Канта.

Доказательный статус

SCIENCE = ABSTAIN; truth_layer_promotion = 0. Канонический смысловой слой semantic_run024 не изменён: 55 реперов, 156 компонентов, 155 прямых текстовых оснований; KANT-SC-015-038-03 остаётся открытым. Новые исходники и программные проверки не увеличивают число семантически подтверждённых компонентов.

Два независимых читателя ещё не выполнили разметку. Эмпирическая точность предсказаний, независимая экспертиза и EXT-RUN остаются NOT_RUN. Этот выпуск подготавливает проверку и формализует её границы.

1. Источник, контейнер и авторство

Все восемь страниц Korpora и русский PDF были получены с HTTP 200; время, конечный URL, число байтов и SHA-256 зафиксированы в source-manifest.json. Немецкий слой содержит 251 нумерованную строку; номера страниц и строк сохраняются. HTML-сущности декодированы, пробелы внутри печатной строки сведены к одному, орфография и переносы не исправлялись. Основной блок страницы и сноска сохраняются отдельно.

В русском PDF нет встроенного текстового слоя: проверены все 387 страниц. OCR выполнено только для PDF-страниц 3, 5, 17, 18, 19, 20, 31, 32, 33, 215, 217, 218. Результаты Tesseract сохранены как производные без автоматического исправления орфографии. Контрольная сумма языковой модели, версия распознавателя и изображения страниц входят в реестр.

Выявленный риск атрибуции

Титул и оглавление показывают составной характер издания. Помимо «Пролегомен» Канта, переведённых Вл. Соловьёвым, оно включает предисловие Н. Я. Грота и приложение Куно Фишера «Критика Кантовской философии». На PDF-странице 215 находится титул приложения, на странице 217 начинается его текст. Приписывать весь файл одному Канту нельзя.

Основание этого вывода - визуальная проверка титулов, оглавления и границы приложения. Полная постраничная экспертиза авторства не проведена. В частности, наличие «предисловия» не означает автоматического исключения: предисловие самого Канта и предисловие редактора имеют разные статусы.

Связь русского контейнера с семейством «Пролегомен» служит правилом разделения данных. Она не означает, что каждый помещённый в контейнер фрагмент принадлежит Канту. В плане проверки русский контейнер, английский перевод и немецкая редакция «Пролегомен» удерживаются в одной проверочной группе.

Источники: [1] Korpora, AA VIII, страницы 35-42; [2] скан издания 1889 года, Wikimedia Commons. Издание 1889 года обозначено в каталоге источника как public domain.

2. Неопределённость интерпретации и множество исходов

Пусть Θ - конечное непустое множество допускаемых интерпретаций. Для $\theta \in \Theta$ задано непустое конечное множество состояний $\Omega_\theta(o)$, совместимых с доступными наблюдениями o , и целевая функция $f_\theta: \Omega_\theta(o) \rightarrow \mathbb{R}$. Все ограничения области, времени наблюдений и источников наследуются от RUN-024/025. Введение такой функции является явной модельной предпосылкой; разметка U/I/R/D сама по себе её не задаёт.

$$Y_\theta(o) = \{ f_\theta(\omega) : \omega \in \Omega_\theta(o) \}; Y^*(o) = \bigcup_{\theta \in \Theta} Y_\theta(o).$$

Объединение является множеством исходов при неизвестной, но допускаемой интерпретации. Это не распределение вероятностей. Если одна из заявленных интерпретаций не имеет корректно определённых исходов, программный интерфейс отказывает в расчёте; он не удаляет её молча.

Утверждение 2.1. Критерий однозначности

Единый точный числовой прогноз по наблюдению o существует тогда и только тогда, когда $Y^*(o)$ состоит из одного элемента. Эквивалентно: каждая $Y_\theta(o)$ одноэлементна и все эти элементы равны.

Доказательство. Если один прогноз a точен во всех допускаемых состояниях и интерпретациях, то каждый элемент каждого $Y_\theta(o)$ равен a , поэтому $Y^*(o) = \{a\}$. Обратно, из $Y^*(o) = \{a\}$ следует $f_\theta(\omega) = a$ для всех θ и ω . Непустота исключает формально пустую «точность». $\square$

Утверждение 2.2. Точная минимаксная граница

Обозначим $m = \min Y^*(o)$, $M = \max Y^*(o)$. При абсолютной потере и разрешении выбирать оценку a из всей $\mathbb{R}$ справедливо:

$$\inf_{a \in \mathbb{R}} \max_{y \in Y^*(o)} |a - y| = (M - m)/2; a^* = (M + m)/2.$$

Доказательство. Для любой a выполняется $M - m \leq |M - a| + |a - m| \leq 2 \max\{|M - a|, |a - m|\}$. Значит, худшая ошибка не меньше половины размаха. Средняя точка m и M достигает этой границы на обоих концах и не превосходит её внутри отрезка. $\square$

3. Что этот результат меняет в предсказательном движении

Расширенная область состояний имеет вид $\Omega^* = \{(\theta, \omega): \theta \in \Theta, \omega \in \Omega_\theta\}$. Наблюдение $\rho^*(\theta, \omega)$ и цель $f^*(\theta, \omega)$ задаются явно. Теорема факторизации из RUN-024 применяется к этой области без нового физического предположения: точный прогноз возможен тогда и только тогда, когда f^* постоянно на каждом слое ρ^* .

Поэтому проверять достаточность реперов требуется одновременно внутри интерпретаций и между ними. Если разные допустимые интерпретации дают разные исходы при одинаковом реперном представлении, одной структурной согласованности каждой ветви недостаточно.

Контрпример к преждевременному выбору ветви

Пусть интерпретация А допускает только исход 0, а интерпретация В - только исход 2. Каждая ветвь по отдельности имеет нулевую ошибку, но объединение $\{0, 2\}$ имеет минимаксный радиус 1. При $\epsilon = 0$ нужен ABSTAIN в рамках модели; при $\epsilon = 1$ условная граница достижима оценкой 1. При этом 1 не обязано быть физически возможным исходом: это центр оценки при указанной функции потерь.

Пересечение $\{0\} \cap \{2\} = \emptyset$ здесь не доказывает внутреннее противоречие ни А, ни В. Оно показывает несовпадение двух альтернатив. Выбор одной из них требует различающего основания. Простое усреднение меток читателей такого основания не создаёт.

Монотонность и границы применимости

Добавление допускаемой интерпретации расширяет Y^* , поэтому минимаксный радиус не может уменьшиться. Обоснованное исключение интерпретации может сузить множество, но не гарантирует строгого улучшения. Для $\epsilon \geq 0$ требуемая точность достижима тогда и только тогда, когда $M - m \leq 2\epsilon$.

Если допустимые решения ограничены, например $a \in Y^*$, формула для свободного а может перестать быть точной: для $\{0, 2\}$ при $a \in \{0, 2\}$ минимаксная ошибка равна 2. В этом выпуске функция `robustOutcomes` использует именно свободную вещественную оценку и абсолютную потерю. Выводы не переносятся автоматически на вероятностный, экономический или физический прогноз.

4. Раздельная разметка и описательное согласие

Единица пилота - зафиксированный блок текста или OCR одной страницы. Каждый читатель задаёт атрибуцию из пяти категорий: основной текст Канта; предисловие Канта; примечание Канта; переводчик, редактор или другой автор; пока не установлено. Роли U/I/R/D допускают несколько значений одновременно. Цитата должна точно встречаться в показанном тексте.

Пустая метка означает отсутствие решения. «Пока не установлено» означает состоявшееся решение о недостаточности основания. Эти состояния не смешиваются. Для чужого или неустановленного авторства авторские роли запрещены. Для OCR оба читателя дополнительно отмечают сверку с изображением; при ошибке распознавания требуется новая редакция текста или сохранение неопределённости.

Согласие атрибуции

Пусть n - число элементов, получивших обе непустые метки. Через n_{ij} обозначим таблицу парных частот, n_{i+} и n_{+i} - её суммы. Вычисляются наблюдаемое согласие p_o , маргинальное совпадение p_e и коэффициент k Коэна [3,4]:

$$p_o = \sum_i n_{ii}/n; p_e = \sum_i n_{i+}n_{+i}/n^2; k = (p_o - p_e)/(1 - p_e).$$

При $n = 0$ показатели не определены. При $p_e = 1$ k также не определён: код возвращает null и причину, а не фиктивную единицу. Например, два читателя могут одинаково оставить все 21 фрагмент неустановленными; наблюдаемое согласие равно 1, но пригодных авторских оснований нет.

Совпадение множеств U/I/R/D вычисляется отдельно для тех пар, где оба читателя выбрали авторскую категорию. Даже одинаковая атрибуция при разных реперных ролях создаёт расхождение, требующее разбора.

Это описательные показатели специально выбранного малого пакета. Доверительные интервалы, статистическая независимость произведений и точность генеральной совокупности не заявляются. Два кода читателя и две галочки декларации не удостоверяют независимость людей. Даже полное совпадение оставляет независимую научную проверку открытой.

5. Исполнимый протокол CP3-PREP

До начала

Куратор фиксирует packet.json, его SHA-256, рубрику CP3-v1, язык и компетенции читателей. Пакет не содержит готовых семантических меток или ответов модели. Источники остаются видимыми: скрывание авторских ответов не означает скрывание библиографического происхождения. Выбор 21 фрагмента объявлен целевым и ретроспективным.

Раздельная работа

Двум людям передаются одинаковый пакет и отдельные пустые формы. Они работают без обмена ответами. На странице annotation.html можно выбрать фрагмент, открыть источник, увидеть скан, указать атрибуцию, роли и цитату. Результат экспортируется в JSON; до экспорта работа существует только в текущей вкладке. Отправка писем, серверный приём анкет и приглашение экспертов в этом выпуске не выполняются.

Фиксация и сопоставление

Каждый ответ связывается с хешем пакета и текстов. Куратор сохраняет два ответа и их контрольные суммы до взаимного раскрытия. Только затем выполняется сопоставление. Статическая страница не удостоверяет порядок раскрытия, личность читателей или отсутствие доступа к старой базе; эти условия требуется подтвердить независимо.

Программа проверяет версию пакета, уникальность идентификаторов, полноту структуры, наличие точной цитаты и допустимость меток. Затем выдаёт список отсутствующих пар, неопределённых решений, расхождений и OCR без сверки. Запрещено считать совпадение достаточным для автоматического закрытия смыслового компонента.

Разбор и дальнейший выпуск

Расхождения обсуждаются после фиксации ответов с сохранением обеих исходных версий. Новая согласованная редакция получает отдельный идентификатор и основание изменения. Прогностическая проверка требует заранее задать р, цель, горизонт, момент доступности данных, базовый метод и критерий успеха. Она не входит в настоящий пилот.

6. Проверки, воспроизводимость и незакрытые вопросы

Пройдена 151 проверка ARBD: 125 унаследованных и 26 новых. Новые проверки включают изменение исходного текста, фальшивую цитату, повтор кода читателя, несоответствие версии, дубликаты элементов, пустые ответы, расхождение ролей, неопределённость к и недопустимое повышение научного статуса.

Минимаксная формула проверена сравнением с отдельным перебором решений для всех 961 пар непустых подмножеств множества $\{-2, -1, 0, 1, 2\}$. Решения перебирались по сетке с шагом $1/2$, содержащей все возможные средние точки. Это тест вычислительной реализации; общее утверждение доказано отдельно в разделе 2.

Пустые формы действительно дают 0 из 21 парного решения и статус INCOMPLETE_PAURED_REVIEW. Синтетические примеры вычислений явно обозначены как учебные; они не являются результатами разметки Канта. Прежний seed.js и corpus-catalog.json сохраняются побайтно.

Числовой контроль сравнивает точные двоичные значения IEEE 754 посредством целых чисел. Если минимаксный центр или радиус не представимы точно в binary64, выдаётся ABSTAIN_NUMERIC_PRECISION. Например, для $\{0, \text{минимальное положительное binary64}\}$ округление радиуса к нулю не принимается за точный прогноз.

Состав воспроизводимого пакета

В комплект входят первичные файлы, исходные HTTP-реестры, нумерованные немецкие строки, русский постраничный индекс, сохранённое OCR и изображения, формы читателей, математические примеры, код модуля, тесты и настоящая статья. OCR хранится как фиксированная производная; повторная сборка использует эти байты. Повторное распознавание с иной версией ПО может давать иной текст и требует новой редакции.

Остаётся открытым

Полный корпус Канта не собран. В RU-контуре обработаны OCR только 12 страниц составного скана; остальные 375 страниц имеют лишь реестр и доступные изображения в PDF. Полная сверка распознавания и независимая семантическая разметка не выполнены. CP2F-038 остаётся открытым. Эмпирический прогноз, слепое испытание и EXT-RUN не проведены.

Научный смысл приращения: неоднозначность источника и интерпретации стала явным входом проверки. Выход алгоритма не подменяет источник, согласие читателей не подменяет истинность, а программный PASS не подменяет независимое испытание.

7. Источники и точка продолжения

[1] Kant, I. Beantwortung der Frage: Was ist Aufklärung? (1784). Akademie-Ausgabe, Bd. VIII, S. 35-42. Электронная транскрипция Korpora. Получены все восемь указанных страниц; ссылки и время загрузки сохранены в source-manifest.json.

<https://www.korpora.org/Kant/aa08/035.html>

<https://www.korpora.org/Kant/aa08/042.html>

[2] Кант И. Пролегомены ко всякой будущей метафизике, могущей возникнуть в смысле науки. Перевод Вл. Соловьёва. Москва, 1889. В составе сборника: предисловие Н. Я. Грота и приложение Куно Фишера. Скан Wikimedia Commons: 387 PDF-страниц. Библиографическое описание сверено с изображениями титула и оглавления.

https://commons.wikimedia.org/wiki/File:Пролегомены_Канта.pdf

[3] Cohen, J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20 (1960), 37-46. DOI: 10.1177/001316446002000104. Библиографические сведения подтверждаются авторским ретроспективным пояснением [4]; прямое чтение журнального файла 1960 года в этом выпуске не заявляется.

<https://doi.org/10.1177/001316446002000104>

[4] Cohen, J. Citation Classic: A coefficient of agreement for nominal scales. Current Contents, 20 January 1986. Авторское пояснение наблюждённого и ожидаемого согласия, доступное в архиве University of Pennsylvania.

<https://garfield.library.upenn.edu/classics1986/A1986AXF2600001.pdf>

[5] Tesseract OCR project. tessdata_fast, русская модель. Локально зафиксированы версия распознавателя и SHA-256 модели. Использование OCR не означает ручной верификации дореформенной орфографии.

https://github.com/tesseract-ocr/tessdata_fast

[6] Внутренние основания проекта: RUN-024 / ARBD37 (множества исходов), RUN-025 / ARBD38 (ограниченная погрешность и доступность наблюдений), RUN-026 / ARBD39 (семейства произведений), RUN-027 / ARBD39-LEARN (учебный раздел). Текущая исходная точка отката: версия сайта 29.

Далее: независимая разметка CP3, проверка и исправление OCR отдельной редакцией, расширение RU/DE-корпуса и CP6-P2 с заранее фиксированным целевым каналом. Тома I-V и замороженные научные статусы не изменяются.