ПЕРИОДИЗАЦИЯ ТВОРЧЕСТВА ОТДЕЛЬНОГО АВТОРА С ОПОРОЙ НА КОЛИЧЕСТВЕННЫЕ МЕТОДЫ

Балуева Д. В. (baluevadaria@ya.ru)
НИУ ВШЭ, Москва

We introduce a new method of periodization of an author's oeuvre based on quantitative features. It is applied to single-author corpora taken from the Poetic subcorpus of the Russian National Corpus, texts have poetic and date markup. We use a multivariate stylometry approach with mixed features: character 4-grams, verse features and word embeddings. The chronologically ordered texts are divided into 2 or 3 subsets in all possible ways, obtained subsets are compared using mixed feature vectors and the cosine distance. Splits with maximal distances between subsets are seen as periodization variants, checked and compared with ones obtained using only 4-grams and suggested by literary scholars. Texts' size is considered.

Keywords: authors' corpora, automatic periodization, Russian poetry, Russian National Corpus, multivariate stylometry, mixed features, text similarity.

Введение

В работе [Балуева 2020] был предложен и проверен метод разбиения творчества отдельного автора на два периода с опорой на частоты символьных 4-грамм. Цель настоящего исследования – усовершенствовать этот метод, учитывая дополнительные признаки текстов, а также предложить новый способ проверки периодизаций.

Изучение развития авторского стиля – одна из задач, которыми занимается стилометрия, компьютерная стилистика [Neal 2017; Скоринкин 2018: 51]. Она основана на предположении о том, что авторский стиль выражается в уникальных количественных признаках текстов и все тексты, написанные одним автором, похожи друг на друга [Evert et al. 2017: 1]. Однако даже тексты одного автора могут отличаться, если они написаны в разные периоды его жизни и творчества [Stamou 2008]. В данной работе мы понимаем периодизацию как сравнение всех текстов одного автора между собой с учетом хронологии и разделение их на несколько максимально непохожих групп: раннее, зрелое, позднее творчество и т. п. Если авторское творчество действительно меняется со временем (меняется его «форма», то есть авторский стиль, и «содержание», то есть тематика творчества), значит его можно периодизовать по этим признакам. Среди исследований, близких нашему, например, [Reeve 2018; Salgaro, Rebora 2018], они посвящены периодизации художественной прозы.

В компьютерных науках тексты сравниваются количественно: вычисляется расстояние между векторами признаков этих текстов [Rayson et al. 2000; Burrows 2002; Gomaa, Fahmy 2013]. Чаще всего используется один языковой признак: частоты лемм слов [Holmes 1994], служебных слов [Burrows 2002] или символьных *n*-грамм [Kjell et al.1994], как в нашем предыдущем исследовании. Однако эти методы лучше работают для длинных прозаических текстов, чем для стихотворений. В исследованиях [Plecháč et al. 2018; Шеля и др. 2020] показано, что для определения авторства поэтических текстов эффективнее учитывать не один, а сразу несколько признаков:

частоты символьных *n*-грамм, лексики и форм стиха. Такой подход называется **многомерной стилометрией**. Смешанные векторы содержат больше информации об авторском стиле и позволяют более надежно сравнивать небольшие объемы текста. С точки зрения периодизации, мы предполагаем, что на основе смешанных признаков удастся поделить творчество поэта на группы, которые максимально не похожи друг на друга и стилистически, и тематически.

Материал и метод

Покажем работу метода на материале корпусов двух авторов из поэтического подкорпуса НКРЯ объемом 50 тысяч токенов и более[1]. Все тексты подкорпуса размечены по дате создания, и мы использовали только тексты размеченные точным годом, например, ‘1910’, а не ‘1910-1911’, даты вида ‘1965.08.05’ сокращались до года – ‘1965’.

Автор	Количество текстов	Текстов с точной датой	Объем корпуса (токены)	Объем текстов с точной датой (токены)
А. С. Пушкин	904	799	193440	134367
И. Г. Эренбург	710	674	70186	66749

Таблица 1: Выбранные авторы и объем их корпусов

Отталкиваясь от предыдущих исследований, мы выбрали три признака:

1) **Частоты символьных *n*-грамм длины 4.** Этот признак мы использовали и в предыдущем исследовании. При создании *n*-грамм учитывались буквы в нижнем регистре, цифры, знаки препинания, пробелы и знаки перевода строки, использовались все n-граммы, которые встретились в корпусе хотя бы 1 раз. Такие символьные сочетания могут содержать информацию о лексике (корни слов или служебные слова полностью), морфологии (морфемы) [Kestemont 2014: 60-62], эвфонии [Piperski 2019: 8] и структуре текста (пунктуация и деление на строки и строфы). Эффективность *n*-грамм длины 4 подтверждалась неоднократно [Kestemont 2014; Evert et al. 2017; Piperski 2019].
2) **Частоты метрических цепочек.** Такие цепочки мы создавали из стиховедческой разметки корпуса – иктов. Икт – сильное место в стихе, чаще всего совпадающее с ударением. [Гришина и др. 2009] Икты в виде знака грависа (`) в подкорпусе проставляются автоматически после того, как разметчик вручную припишет произведению метр и тип клаузулы, а затем разметка еще раз экспертно проверяется. Так, из стиха (стихотворной строки) *Я лю̀лий на̀рвала̀ прекра̀сных у̀ душѝстых* получается цепочка *0101010101010*, где 0 – слабое место в стихе (чаще всего – безударный слог), а 1 – икт. Как и в случае с n-граммами, мы использовали все цепочки, которые встретились в корпусе хотя бы 1 раз.[2] В отличие от ритмических цепочек, которые использовались в [Шеля и др. 2020], наши цепочки не отражают все реальные ударения и информацию о словоразделах, но отражают метр, стихотворный размер и тип клаузулы.

[1] Мы благодарим создателей подкорпуса и лично Д. В. Сичинаву за предоставление текстов с XML разметкой.

[2] Использовались не абсолютные частоты слов и цепочек, а относительные: *(абсолютная частота/полный объем словаря единиц)*1000*

3) **Дистрибутивная семантика.** В предыдущих исследованиях одновременно учитывались символьные *n*-граммы длины 2 и лексика (леммы слов). В нашем случае учет лемм был бы избыточен, так как символьные 4-граммы покрывают большие части слов и/или корней. Так как лексика чаще всего отражает именно тематику, мы тоже решили учитывать значения слов, используя эмбеддинги. Мы обучили модель fastText[3] на всем поэтическом подкорпусе НКРЯ (более 12 млн слов)[4] и для каждого отдельного автора получали эмбеддинги всех слов его корпуса. Затем создавали средний эмбеддинг для каждого года в корпусе (среднее арифметическое эмбеддингов всех слов за год), а из эмбеддингов отдельных лет получались средние эмбеддинги периода из нескольких лет. FastText обучается не на словах целиком, а на символьных *n*-граммах [Bojanowski et al. 2017], что позволяет учитывать морфологию.

Два частотных вектора и эмбеддинг объединялись ("склеивались" друг за другом) и получался единый вектор признаков. Размерность каждого вектора: *объем словаря n-грамм + объем словаря метрических цепочек + 100 (размерность эмбеддинга)*. Векторы n-грамм и цепочек нормированы, стандартное отклонения вектора постоянно и равно 1, значит, все три признака вносят равный вклад.

Способ периодизации

Как шаг периодизации мы выбрали один год, и поэтому сначала объединяли все тексты, написанные за год, в единый массив. Затем для каждого автора делили все годы его творчества (массивы текстов за определенный год) на два или три подмножества с учетом хронологии. Каждое подмножество – предположительно тексты раннего, зрелого или позднего творчества поэта. Задача – найти такое разбиение, где подмножества будут максимально отличаться друг от друга. Поэтому при каждом разбиении из текстов каждой группы создается по вектору признаков, и между векторами соседних групп вычисляется косинусное расстояние. Затем каждое разбиение сравнивается с предыдущим и следующим: если расстояние (или расстояния) при каком-либо разбиении больше, чем при двух соседних, то такое разбиение предположительно отражает периодизацию корпуса. Рассмотрим сначала, как мы разбиваем на 2 периода творчества поэта, который писал с 1910 по 1930 год:

Номер	Период 1	Период 2	Расстояние
1	1910–1911	1912–1930	0.005
2	1910–1912	1913–1930	**0.01**
3	1910–1913	1914–1930	0.006

Таблица 2: фрагмент разбиения на 2 периода

При втором разбиении корпуса (1910-1912 и 1913-1930) расстояние между частями оказалось больше, чем два соседних (0.01> 0.005, 0.01 > 0.006). Чем больше расстояние

[3] Характеристики модели: unsupervised, skipgram, 100-мерные векторы, ширина контекстного окна – 5, длина *n*-грамм 3–6.

[4] Тексты были приведены к нижнему регистру, пунктуация удалена. Благодаря использованию fastText лемматизация не понадобилась.

между частями, тем меньше они похожи. Это может значить, что после 1912 года стиль поэта изменился сильнее, чем после 1911 года или после 1913 года. Такой максимум расстояния на границе лет мы будем называть **локальным максимумом расстояния**, а разбиение №2 считать вариантом разбиения творчества на два периода. Мы решили отказаться от выбора глобального максимума, так как самые большие расстояния всегда при сравнении первого или последнего года со всеми остальными, что, судя по всему, связано с несбалансированностью сравниваемых подкорпусов.

Теперь рассмотрим фрагмент разбиения творчества какого-либо автора (автор писал в 1900– 1908 гг.) на три периода:

Номер	Период 1	Период 2	Период 3	Расстояние 1 (Период 1, Период 2)	Расстояние 2 (Период 2, Период 3)
1	1900–1901	1902–1904	1905–1908	0.0002	0.0001
2	1900–1901	1902–1905	1906–1908	**0.004**	**0.007**
3	1900–1901	1902–1906	1907–1908	0.00005	0.0002

Таблица 3: фрагмент разбиения на 3 периода

При втором разбиении корпуса (1900–1901, 1902–1905 и 1906–1908) сразу два расстояния (между подмножествами текстов 1900–1901 и 1902–1905) оказались больше, чем соответствующие расстояния в двух соседних разбиениях (0.004>0.0002, 0.004>0.00005 и 0.007>0.0001, 0.007>0.0002). Период 2 – “окно” между двумя крайними периодами, которое может расширяться или двигаться. Именно когда вторая граница этого “окна” оказалась перед 1906 годом, группа текстов Периода 2 стала значимо отличаться от Периода 1 и Периода 3. Разбиение №2 – вариант разбиения творчества на 3 периода.

Автоматическая проверка периодизаций

Для каждого автора мы получили несколько «локальных максимумов» и чтобы выбрать из них наилучшие варианты, мы все их дополнительно проверили. Каждый локальный максимум расстояния при разбиении на две части и каждое “окно” с двумя локальными максимумами при разбиении на три части мы считали вариантом периодизации: для одного поэта вариантов могло быть несколько. Все варианты мы проверили:

1) Перемешали тексты внутри каждого периода, приписав им случайные даты этого же периода.

2) Затем, как и на предыдущем этапе, для каждого года все тексты за этот год объединяются в единый массив. Например, если перемешать тексты 1903–1905 года, в массив текстов 1903 года могут попасть тексты и за 1903, и за 1904, и за 1905 год. Тексты разных периодов не смешиваются.

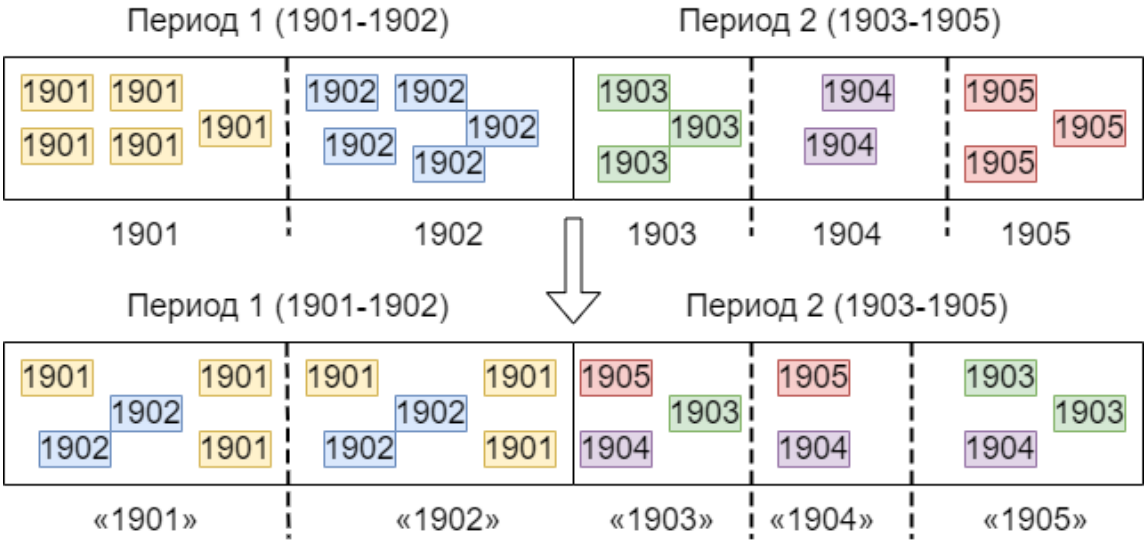


Рис. 1: Способ перемешивания текстов внутри одного периода

3) Применяли тот же способ периодизации, что и в начале, и смотрели, оказался ли самый большой локальный максимум расстояния на границе смешанных частей.
4) Повторили предыдущий шаг 100 раз и посчитали долю “успешных” случаев.

Таким образом, для всех вариантов периодизации мы проверяли, правда ли тексты каждого предполагаемого периода представляют собой единый массив и значимо отличаются от массивов других периодов. Тогда, даже если перемешать признаки отдельных лет, суммарный набор признаков не поменяется. Этот метод проверки для периодизации любой дробности (два, три, четыре и т. д. периода), пока минимальное количество лет в периоде не дойдет до двух.

Учет сбалансированности периодов и интерпретация результатов

После проверки, описанной в предыдущем разделе, каждый вариант периодизации мы оценили на сбалансированность по размеру. С одной стороны, чем больше объем сравниваемых текстов, тем надежнее работают методы стилометрии. С другой стороны, мы оценивали и общую правдоподобность результата: маловероятно, что можно выделить отдельный период творчества поэта длиной в один год или на основе нескольких коротких текстов. Кроме того, мы сопоставляли результаты, полученные автоматически, с теми, которые предлагали эксперты-литературоведы. Таким образом, наилучший вариант мы выбирали, произведя следующую последовательность действий: проверка с перемешиванием, повторенная сто раз, оценка сбалансированности, учет фактов из истории литературы.

Рассмотрим подробно варианты, которые оказались самыми подтвержденными, и сравним нынешний результат с периодизациями только на основе символьных *n*- грамм длины 4. Все варианты, результаты проверок, а также визуализации приводятся в Приложении.

1) А. С. Пушкин – 3 периода: 1813–1816, 1817–1832, 1833–1836 (24 т. слов и 89 текстов / 88 т. слов, 613 текстов / 22 т. слов и 97 текстов; 11/100 успешных проверок)

Судя по всему, между 1832 и 1833 годами проходит самая четкая граница в творчестве Пушкина. Она встречается и в самом подтвержденном из бинарных вариантов (см. Приложение) и этом в варианте, который наиболее сбалансирован. В [Баевский 2001] производится автоматическая периодизация творчества Пушкина и эта граница тоже упоминается. Баевский пишет, что литературоведы обычно ее не проводят, но она однозначно выделяется на основе кластерного анализа. В работе Баевского и во всех периодизациях, обсуждаемых в обзоре [Фомичев 1982], выделяется раннее, лицейское творчество поэта до 1817 года и позднее творчество – 30-е годы. С 1817 и до 1830 – срединный период, который разные исследователи делят по-разному. Границы для разбиения на 2 периода, которые мы выделяли на основе только символьных n- грамм (1813–1829 и 1830–1836, 1813–1820 и 1821–1836), в новых вариантах тоже встречались, но ни одна из них не входит ни в самое подтвержденное, ни в самое сбалансированное разбиение.

1) И. Г. Эренбург

- 2 периода: 1910–1945 и 1946–1958 (60 т. слов и 614 текстов / 7 т. слов и 60 текстов; 59/100 успешных проверок)
- 3 периода: 1910–1922, 1924–1941, 1942–1958 (45 т. слов и 449 текстов / 8 т. слов, 95 текстов / 12 т. слов и 130 текстов; 9/100 успешных проверок)

На основе одних 4-символьных *n*-грамм мы делили творчество Эренбурга на 1910–1924 и 1939–1958 годы. Этот вариант может отражать не столько изменение индивидуального стиля, сколько изменения в языке, произошедшие за 15 лет, когда поэт ничего не писал. Б. Я. Фрезенский [Фрезенский 2000] делает периодизацию творчества Эренбурга и выделяет такие периоды: 1910–1914 (до начала войны), 1914–1920 (война и революция), 1921–1922 (по сравнению с прозой стихотворений мало), 1924–1941 (1924 г., 15 лет перерыва и 1939–1941 гг.), 1941– 1967 (но после 1958 года стихи автор не писал). В нашей периодизации первый период охватывает три первых периода, выделенных Фрезенским, второй период – второй у Фрезенского (он тоже выделяет период, куда попадает пауза и несколько лет вокруг нее), третий период тоже совпадает с тем, который выделил Фрезенский. Вариант разделения на 2 периода менее сбалансирован, но подтвердился бóльшим количеством проверок, значит тоже может заслуживать рассмотрения экспертами-литературоведами.

Заключение

Итак, мы предложили и проверили новый метод периодизации творчества отдельного автора с опорой на количественные признаки текстов и продемонстрировали его работу на примере корпусов двух авторов. Сравнив результаты с экспертными вариантами периодизаций и результатом, полученным только на 4-граммах, мы провели их подробный качественный анализ. Продолжением работы может стать количественный анализ результатов и сравнение предлагаемого метода с другими. Например, можно собрать набор эталонных экспертных периодизаций для десятков или сотен поэтов и сравнивать с ними результаты применения метода. Кроме того, можно проверить, сравнить и проинтерпретировать, какие результаты дает применение всех трех признаков в отдельности или, например, применение отдельно стилистического вектора (n- граммы + метрика) и эмбеддингов.

Библиография

Баевский В.С. Лингвистические, математические, семиотические и компьютерные модели в истории и теории литературы. Языки славянской культуры, М., 2001

Балуева Д. В. Автоматическая периодизация авторских корпусов. По материалам ежегодной международной конференции «Диалог». Студенческие статьи, Электронная публикация на сайте конференции, 2020. http://www.dialog-21.ru/media/4918/baluevad.pdf

Гришина Е. А., Корчагин К. М., Плунгян В. А., Сичинава Д. В. Поэтический корпус в рамках НКРЯ: общая структура и перспективы использования // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009.

Скоринкин Д. А. Семантическая разметка художественных текстов для количественных исследований в филологии (на примере романа "Война и мир" Л.Н. Толстого): дис. канд. филол. наук. М., 2018.

Фрезинский Б. Я. Из слов остались самые простые. / Илья Эренбург. Стихотворения и поэмы. СПб.: Академический проект, 2000 г.

Фомичев С. А. Периодизация творчества Пушкина: (К постановке проблемы) // Пушкин: Исследования и материалы / АН СССР. Ин-т рус. лит. (Пушкин. Дом). Л.: Наука. Ленингр. отд-ние, 1982.

Шеля А., Плехач П., Зеленков В. Феномен Батенькова и проблема атрибуции авторства: многомерный статистический подход к нерешеному вопросу // Acta Slavica Estonica. XI. Пушкинские чтения в Тарту 6. Выпуск 2. Тарту, 2020.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135-146, 2017.

Burrows J. ‘Delta’: a measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, 17(3), 267-287, 2002.

Evert S., Proisl, T., Jannidis F., Reger I., Pielström S.,Schöch C., & Vitt, T. Understanding and explaining Delta measures for authorship attribution, Digital Scholarship in the Humanities, Volume 32, Issue suppl_2, December 2017, 24–6, 2017.

Gomaa W. H., Fahmy A. A Survey of Text Similarity Approaches International / Journal of Computer Applications, 68(13), April, pp. 13–18, 2013.

Holmes D.I. Authorship attribution. Computers and the Humanities, 28(2), 87– 106. 1994.

Juola P. Authorship attribution. Foundations and Trends in Information Retrieval, 1(3): 233–334, 2006.

Kestemont, M. Function Words in Authorship Attribution. From Black Magic to Theory? Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL), 59–66, 2014.

Kjell B., Woods W.A., & Frieder O. Discrimination of authorship using visualization. Information Processing & Management, 30(1), 141–150. 1994.

Kutuzov A. Distributional word embeddings in modeling diachronic semantic change. Thesis submitted for the degree of Philosophiae Doctor Department of Informatics The Faculty of Mathematics and Natural Sciences Language Technology Group. 2020. © Andrey Kutuzov, 2020.

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., & Woodard, D. Surveying Stylometry Techniques and Applications. ACM Comput. Surv., 50(6), 86:1– 86:36, 2017.

Piperski A. Authorship Attribution with a Very Naïve Bayes Model and What It Can Tell Us about Russian Poetry. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International. Conference “Dialogue”. Issue 18, 2019.

Plecháč P., Bobenhausen K., Hammerich B. Versification and authorship attribution. A pilot study on Czech, German, Spanish and English poetry // Studia Metrica et Poetica. Vol. 5. No. 2 P. 29–54. 2018.

Rayson P., & Garside R. Comparing corpora using frequency profiling. In Proceedings of the Comparing Corpora Workshop at ACL 2000. Hong Kong, 2000.

Reeve J. P. “Does ‘Late Style’ Exist? New Stylometric Approaches to Variation in Single-Author Corpora”, in DH2018 Book of Abstracts, ADHO, Mexico City, pp. 478- 481, 2018.

Salgaro M., & Rebora S. Is “Late Style” measurable? A stylometric analysis of Johann Wolfgang Goethe’s, Robert Musil’s, and Franz Kafka’s late works. Elephant&Castle, Systems, and Applications (pp. 77–86). Berlin, Heidelberg: Springer, 2018.

Stamou, C. Stylochronometry: Stylistic Development, Sequence of Composition, and Relative Dating. Literary and Linguistic Computing, 23(2), 181–199, 2007.

Приложение

1) **Разбор всех периодизаций корпуса А. С. Пушкина:**
    1. 1813–1821, 1822–1836
    2. 1813–1824, 1825–1836
    3. 1813–1829, 1830–1836
    4. 1813–1832, 1833–1836
    5. 1813–1815, 1816–1818, 1819–1836
    6. 1813–1816, 1817–1820, 1821–1836

7. 1813–1816, 1817–1832, 1833–1836
8. 1813–1817, 1818–1820, 1821–1836
9. 1813–1818, 1819–1820, 1821–1836
10. 1813–1818, 1819–1823, 1824–1836,
11. 1813–1832, 1833–1834, 1835–1836

4	1813–1832, 1833–1836	47
3	1813–1829, 1830–1836	37
2	1813–1824, 1825–1836	33
1	1813–1821, 1822–1836	31
5	1813–1815, 1816–1818, 1819–1836	22
11	1813–1832, 1833–1834, 1835–1836	13
7	1813–1816, 1817–1832, 1833–1836	11
10	1813–1818, 1819–1823, 1824–1836	9
6	1813–1816, 1817–1820, 1821–1836	7
8	1813–1817, 1818–1820, 1821–1836	2
9	1813–1818, 1819–1820, 1821–1836	0

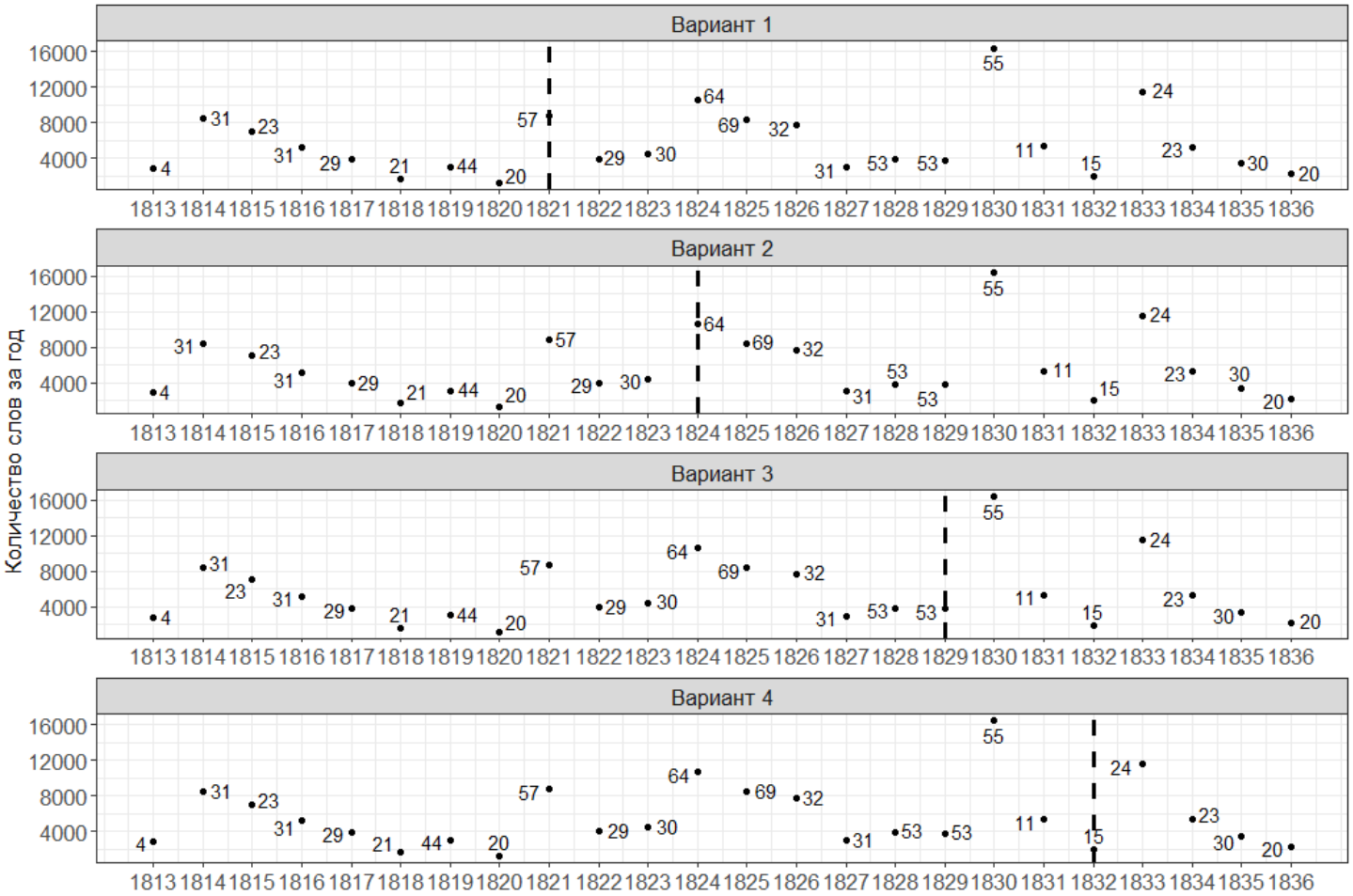
Таблица 4: Количество успешных проверок вариантов периодизации корпуса А. С. Пушкина, сортировка по убыванию количества успешных проверок

№	Слов 1	Текстов 1	Слов 2	Текстов 2	Слов 3	Текстов 3	Слов мин.	Текстов мин.
2	61321	383	73046	416			61321	383
3	88153	621	46214	178			46214	178
1	42258	260	92109	539			42258	260
4	111890	702	22477	97			22477	97
7	23584	89	88306	613	22477	97	22477	89
9	29177	139	21512	180	83678	480	21512	139
10	29177	139	21512	180	83678	480	21512	139
5	18374	58	10803	81	105190	660	10803	58
6	23584	89	9899	114	100884	596	9899	89
8	27498	118	5985	85	100884	596	5985	85
11	111890	702	16837	47	5640	50	5640	47

Таблица 5: Сбалансированность вариантов периодизации корпуса А. С. Пушкина, сортировка по убыванию кол-ва слов

Вот как читать графики ниже:

- каждая точка и число около нее – количество текстов, написанных в определенный год;
- на вертикальной оси отображается количество слов за год, на горизонталь;
- на горизонтальной оси отмечены года творчества поэта;
- пунктирной линией обозначены границы предполагаемых периодов, пограничный год включается в более ранний период.



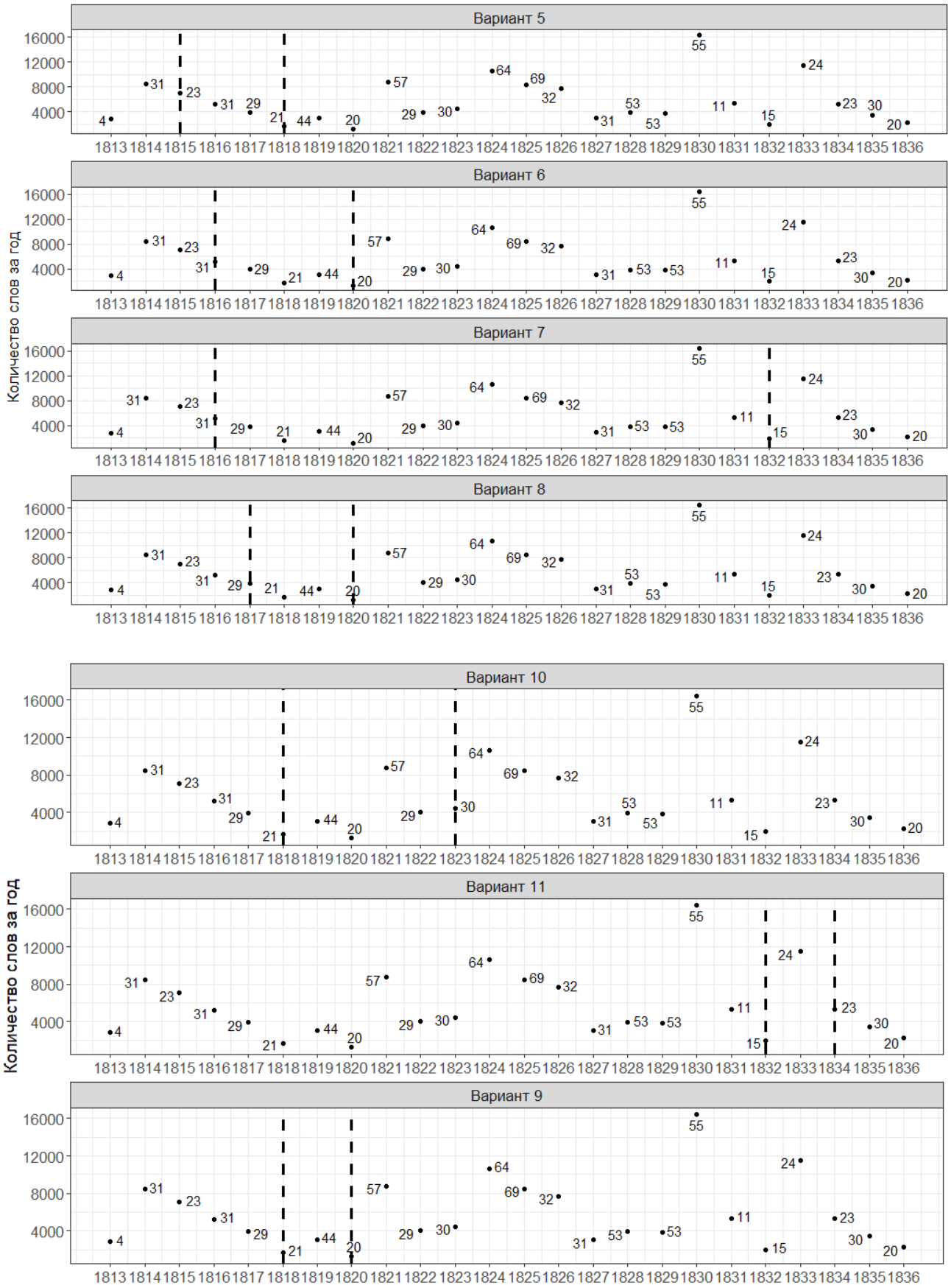


Рис. 2-4: Визуализация вариантов периодизации корпуса А. С. Пушкина

2) **Разбор всех периодизаций корпуса И. Г. Эренбурга:**

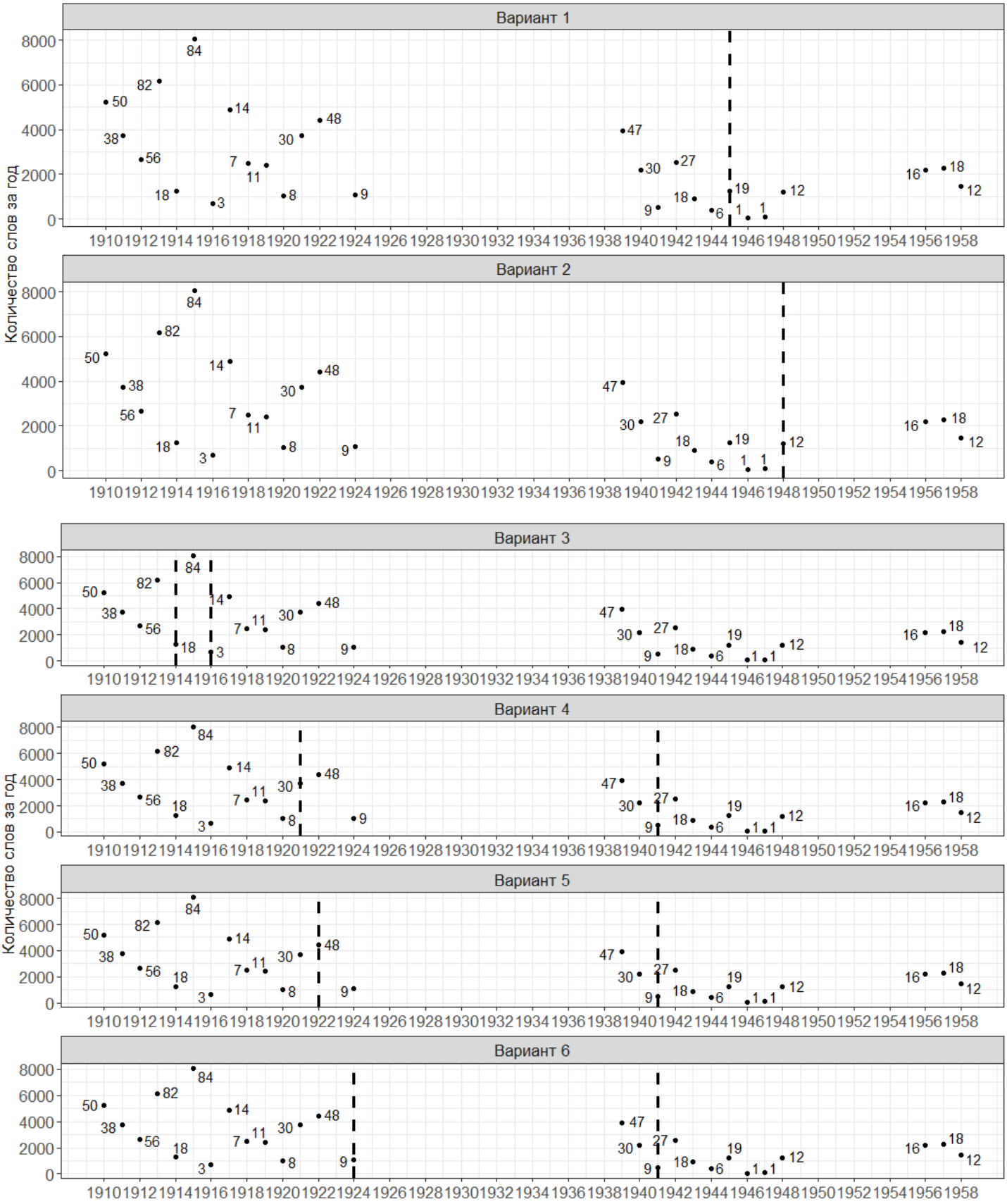
1. 1910–1945, 1946–1958
2. 1910–1948, 1956–1958
3. 1910–1914, 1915–1916, 1917–1958
4. 1910–1921, 1922–1941, 1942–1958
5. 1910–1922, 1924–1941, 1942–1958
6. 1910–1924, 1939–1941, 1942–1958
7. 1910–1939, 1940–1941, 1942–1958
8. 1910–1943, 1944–1946, 1947–1958
9. 1910–1944, 1945–1946, 1947–1958

1	1910–1945, 1946–1958	59
8	1910–1943, 1944–1946, 1947–1958	10
5	1910–1922, 1924–1941, 1942–1958	9
2	1910–1948, 1956–1958	3
4	1910–1921, 1922–1941, 1942–1958	3
7	1910–1939, 1940–1941, 1942–1958	3
3	1910–1914, 1915–1916, 1917–1958	0
6	1910–1924, 1939–1941, 1942–1958	0
9	1910–1944, 1945–1946, 1947–1958	0

Таблица 6: Количество успешных проверок вариантов периодизации корпуса И. Г. Эренбурга, сортировка по убыванию количества успешных проверок

№	Слов 1	Текстов 1	Слов 2	Текстов 2	Слов 3	Текстов 3	Слов мин.	Текстов мин.
4	252411	401	73582	143	69686	130	69686	130
3	113906	244	51975	87	229798	343	51975	87
5	279891	449	46102	95	69686	130	46102	95
1	354577	614	41102	60			41102	60
6	286164	458	39829	86	69686	130	39829	86
2	362389	628	33290	46			33290	46
7	309959	505	16034	39	69686	130	16034	39
8	345653	589	9273	26	40753	59	9273	26
9	347783	595	7143	20	40753	59	7143	20

Таблица 7: Сбалансированность вариантов периодизации корпуса И. Г. Эренбурга


Вариант 1
Вариант 2
Вариант 3
Вариант 4
Вариант 5
Вариант 6
Количество слов за год
0
2000
4000
6000
8000
1910 1912 1914 1916 1918 1920 1922 1924 1926 1928 1930 1932 1934 1936 1938 1940 1942 1944 1946 1948 1950 1952 1954 1956 1958
84 82 50 14 48 38 30 47 56 7 11 27 30 16 18 19 12 18 18 3 8 9 9 6 1 1 12

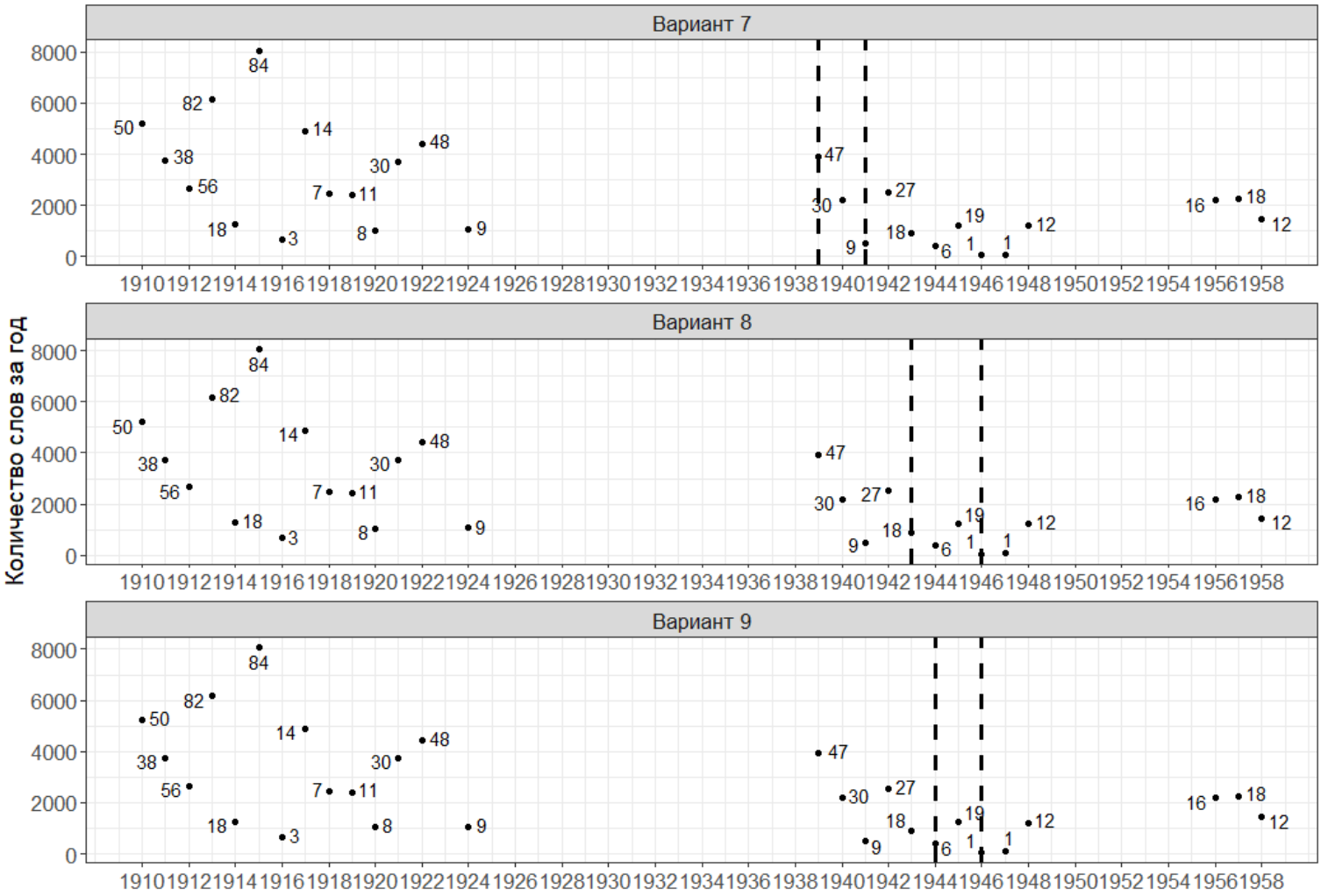


Рис. 5-7: Визуализация вариантов периодизации корпуса И. Г. Эренбурга