Приложения[1]

(к статье С. А. Коротаева «Историческая семантика концепта “класс” в академической литературе: опыт количественного анализа»)

Приложение 1

Список вошедших в корпус журналов

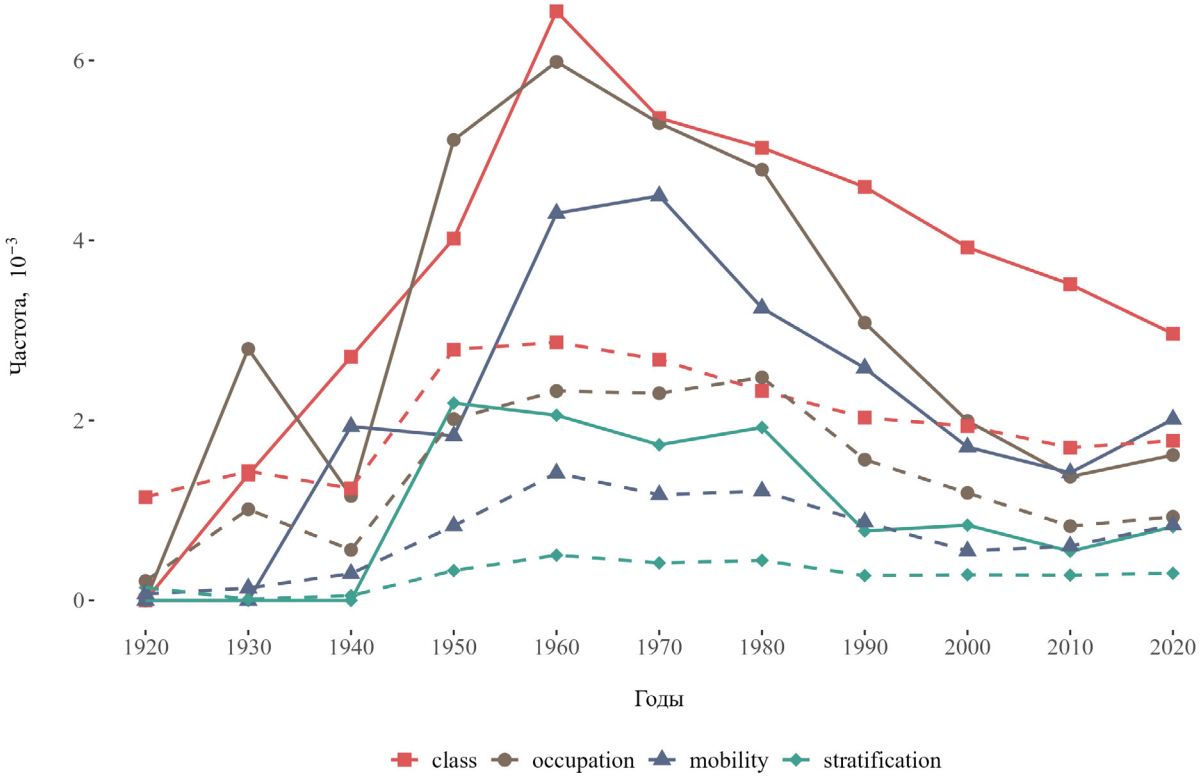
1. American Journal of Sociology;
2. American Sociological Review;
3. Sociology;
4. Social Problems;
5. Sociological Inquiry;
6. Social Forces;
7. Acta Sociologica;
8. The Sociological Quarterly;
9. Social Science Research.

[1] Библиографические отсылки даются к пристатейному списку литературы. — *Примеч. ред*.

Приложение 2

Аннотации и заголовки

На рисунке П2.1 представлена динамика частот четырёх слов в корпусе заголовков и аннотаций. Для того, чтобы частоты были сопоставимы, они нормированы на общее число токенов до удаления редких токенов и токенов из стоп-листа. Направления трендов в двух корпусах достаточно точно соответствуют друг другу, однако представленные слова примерно в 2–3 раза чаще встречаются в текстах заголовков, чем в аннотациях. Если одни слова встречаются чаще, то какие-то другие — реже. Для более систематического анализа мы отобрали относительно частотные токены, чтобы не тратить время на интерпретацию случайных флуктуаций, и рассмотрели те из них, для которых отношение частот в заголовках к частотам в аннотациях максимально, и те, для которых оно минимально.



Примечание: Сплошная линия — частотность в заголовках; пунктирная — в аннотациях. Цифры по горизонтальной оси соответствуют середине периода (например: 1950: 1945–1954 гг.).

Рис. **П2.1**. Относительные частоты встречаемости слов по декадам

Наиболее характерны для аннотаций слова, позволяющие описать последовательность этапов исследования, например: *finding* (вывод), *result* (результат), *increase* (увеличение), *examine* (исследовать), *develop* (развивать) — пять токенов с наименьшим среди всех отношением частот в корпусе заголовков и корпусе аннотаций; методологические и статистические термины, например: *sample* (выборка), *measure* (мера), *regression* (регрессия), *term* (термин), *data* (данные); слова, используемые для презентации теоретической позиции, например: *literature* (литература), *framework* (рамка), *hypothesis* (гипотеза); все перечисленные токены встречаются в 2–12 раз чаще в аннотациях, чем в заголовках. Такие слова, как представляется, характеризуют общий стиль написания аннотаций в академических жур-

налах, но не их содержательное наполнение, потому они не играют большой роли при рассмотрении семантики интересующих нас концептов[2].

Для заголовков мы также можем выделить частотные слова, служащие, вероятно, по большей части, решению риторических задач, однако их доля невелика: *versus* (против), *paradox* (парадокс), *dilemma* (дилемма), *back* (назад), *toward* (вперед). Основная масса токенов, чья частотность в заголовках в разы превосходит частотность в аннотациях, характеризует предмет исследования: это названия стран и регионов, а также слова, описывающие конкретные феномены, институты, группы и т. п. Такие слова часто вполне распространены в повседневном языке, однако в социологии они наполнены специфическим концептуальным смыслом: *ethnic* (этнический), *adulthood* (взрослость), *disability* (недееспособность), *globalization* (глобализация), *democracy* (демократия) — несколько примеров среди токенов с наибольшим отношением частотностей (более 3,5)[3]. Фамилии классиков также оказались в числе слов, встречающихся в заголовках чаще, чем в аннотациях: *Durkheim* (Дюркгейм) — 2,8 раза, *Parsons* (Парсонс) — 2,6, *Weber* (Вебер) — 2, *Marx* (Маркс) — 1,9.

Как видно, заголовки содержат б*о*льшую долю концептуально насыщенных слов, чем аннотации. Заголовок в более сжатой, чем аннотация, форме должен соответствовать ключевой идее статьи, отражать специфическую ставку автора; согласно этой же логике, аннотация в компактной форме отражает содержание статьи (со всеми оговорками относительно специфических стилистических особенностей). Если экстраполировать отмеченные закономерности, можно заключить, что корпус аннотаций в большей степени состоит из слов, имеющих в рамках социологии концептуальное наполнение, чем корпус текстов, и в меньшей степени из относительно слабо релевантных слов. Контекст аннотаций более информативен с точки зрения семантического значения терминов, тогда как в полных текстах термины часто могут оказываться в случайном и не характерном для них контексте. Следовательно, корпус аннотаций позволяет получить более качественное векторное семантическое пространство, чем аналогичный по размеру (в токенах) корпус полных текстов. Хотя это предположение нельзя назвать строго сформулированным, как и подтвердить на наших данных непосредственно, мы расцениваем его как аргумент в пользу достоверности предпринятого в работе анализа.

[2] Возможны следующие возражения: (a) «статистические термины» также являются социологическими концептами и играют большую роль в исследовательской практике, этот вопрос будет рассмотрен далее; (b) сама по себе организация повествования с помощью обращения к теории, создания гипотез и описания эмпирических шагов может рассматриваться как характеристика определённой философской позиции (например, позитивизма). По всей видимости, это верно, однако мы сомневаемся, что избранная оптика позволила бы нам выявить непозитивистские подходы к классовому и (или) стратификационному анализу.

[3] Также «риторические» слова отличаются высокой вероятностью того, что они встречаются в заголовке статьи, но при этом не встречаются в ее аннотации (напр., 90% — versus), тогда как для «содержательных» слов эта вероятность значительно ниже (напр., 8% — «класс»).

Приложение 3

Частотность слов и популярность тем

О попытке отделить изменения частоты, связанные с семантическими изменениями, от тех, что обусловлены колебаниями популярности тем[4], в обсуждении которых используется рассматриваемое слово, см.: [Karjus et al. 2020]. Авторы упомянутой работы исходят из допущения, что при отсутствии семантических сдвигов варьирование популярности слова будет синхронизировано с изменениями частотностей совместно встречающихся с ним слов. Предложенный алгоритм состоит из следующих шагов:

— выявляются слова, часто расположенные в тексте в пределах некоторого небольшого расстояния (окна) от рассматриваемого слова; мерой близости является позитивная поточечная взаимная информация (*positive pointwise mutual information*, PPMI) — нормированная частота совместной встречаемости слов (см. приложение 5); отбирается некоторое заранее определённое число слов (~50–100) с наибольшим PPMI по отношению к целевому слову;

— для каждого из отобранных на предыдущем шаге токенов рассчитывается логарифм отношения частотностей этого слова в подкорпусах, соответствующих двум сравниваемым моментам времени;

— вычисляется взвешенное среднее логарифмов отношений частотностей, полученных на втором шаге, где в качестве весов используются PPMI, полученные на первом шаге.

Таким образом, рассчитывается среднее изменение частотности слов, близких к рассматриваемому. Причём чем ближе слово к интересующему нас (больше PPMI), тем выше его вклад в это среднее. Шаги повторяются для каждой пары сравниваемых временны*х* отрезков. Сильное отклонение изменения частотности исследуемого слова от среднего его соседей указывает на его семантическую трансформацию, то есть, вычитая из наблюдаемых изменений рассчитанные средние, мы получаем «очищенное» изменение частотности, обусловленное только семантикой. Результат[5] этой манипуляции представлен на рисунке П3.1 (см. панели *d* — *f*). Как видно, общий тренд на панелях *d* и *f* остался неизменным; особенно выражен нисходящий тренд у «профессии». Таким образом, имеющиеся данные свидетельствуют в пользу сокращения контекстов употребления рассматриваемых слов. Значительно более существенны изменения на центральных панелях, рост на панели *b* сменился горизонтальным трендом для «дохода»; скорее, ниспадающим для «образования» при сохранении положительного тренда для «неравенства».

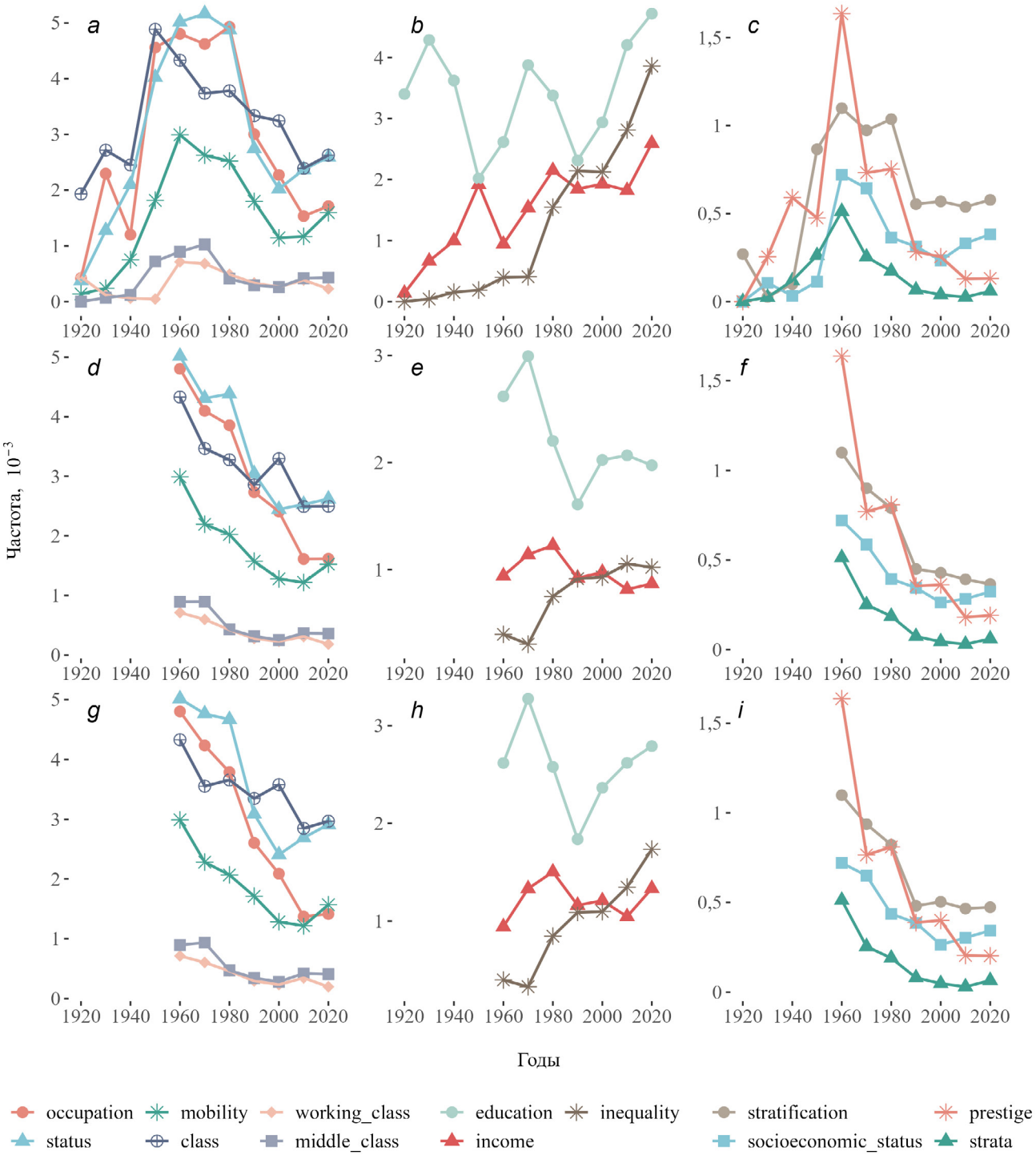
Также мы повторили указанный анализ, используя в качестве меры близости не PPMI сами по себе, но косинусные расстояния между векторами, образованными профилем значений PPMI сравниваемых слов по всем словам корпуса, то есть брали строки таблицы со значениями PPMI (строки и столбцы таблицы соответствуют типам токенов корпуса) как векторы, вычисляли косинусное расстояние между ними и использовали его в качестве меры близости. Такую манипуляцию можно считать простейшим способом создания семантического векторного пространства (см. приложение 5).

---

4 Колебания популярности тем могут быть связаны и с изменением выборки, то есть с включением новых журналов, имеющих отличные от других редакторскую политику и направленность.

5 Корректировка для периода до 1954 г. не проводилась ввиду малочисленности соответствующих подкорпусов. Алгоритм анализирует только отношения частотностей, но не их абсолютные значения, поэтому за исходную точку взят первый рассматриваемый период. На рисунке П3.1 представлены бутстрепные оценки. Предварительный анализ показал, что картинка слабо зависит от размера окна и числа учитываемых соседей.

Результат этого варианта процедуры оказался весьма близок к предыдущему (см рис. П3.1, панели *g* — *i*).



Примечание: Цифры по горизонтальной оси соответствуют середине периода (например: 1950: 1945–1954 гг.). Принадлежность токена к типу эксклюзивна: в число упоминаний класса не включены средний класс и рабочий класс; статуса — социоэкономический статус.

Рис. **П3.1.** Относительные частоты встречаемости слов и словосочетаний (типов токенов) по декадам: исходные (*a* — *c*), скорректированные с помощью процедуры А. Карджуса (A. Karjus) и его коллег (*d* — *f*), скорректированные с помощью модифицированной процедуры А. Карджуса и его коллег (*g* — *i*)

Приложение 4

Дополнительные рисунки

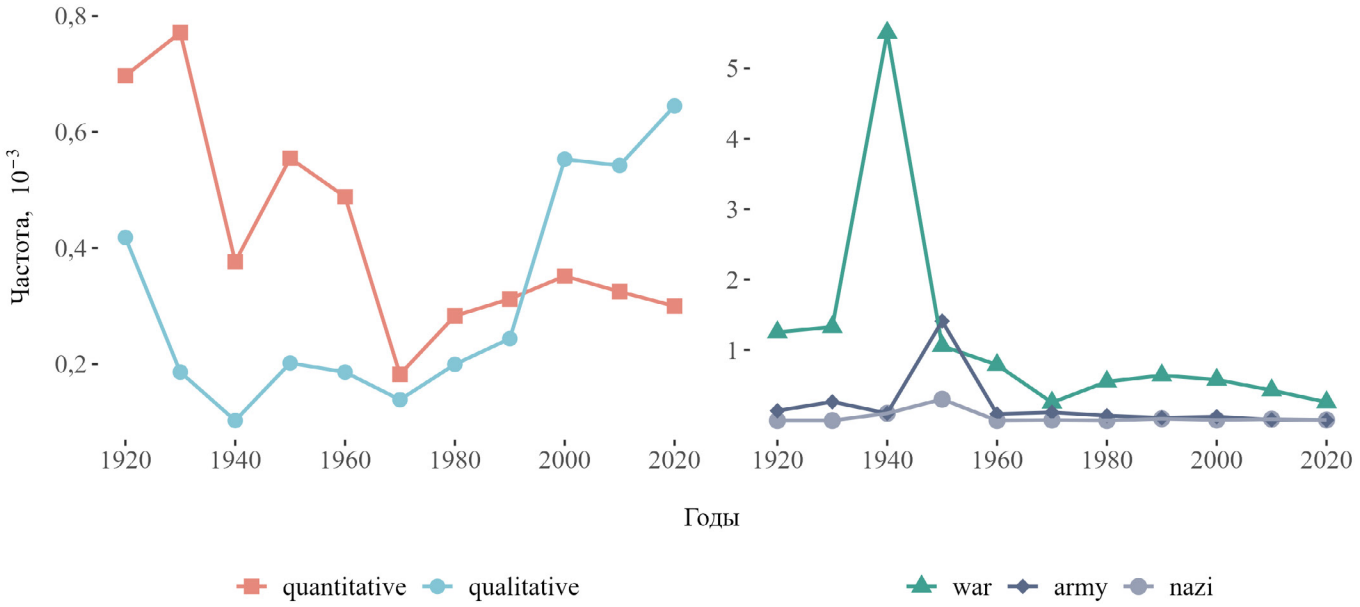


Рис. **П4.1.** Относительные частоты встречаемости слов по декадам: левая панель — *qualitative* (качественный), *quantitative* (количественный); правая — *war* (война), *army* (армия), *nazi* (нацист)

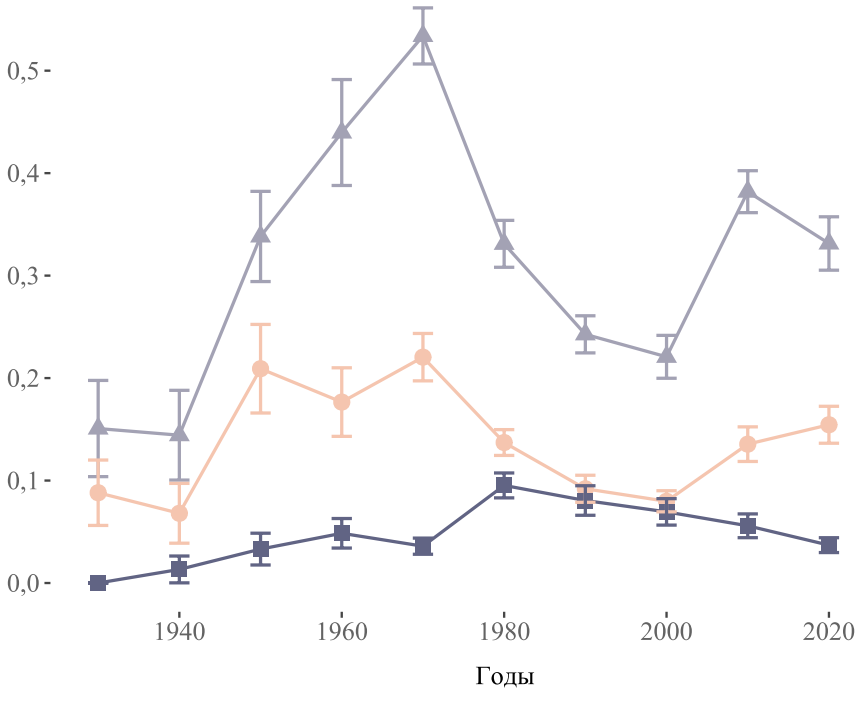
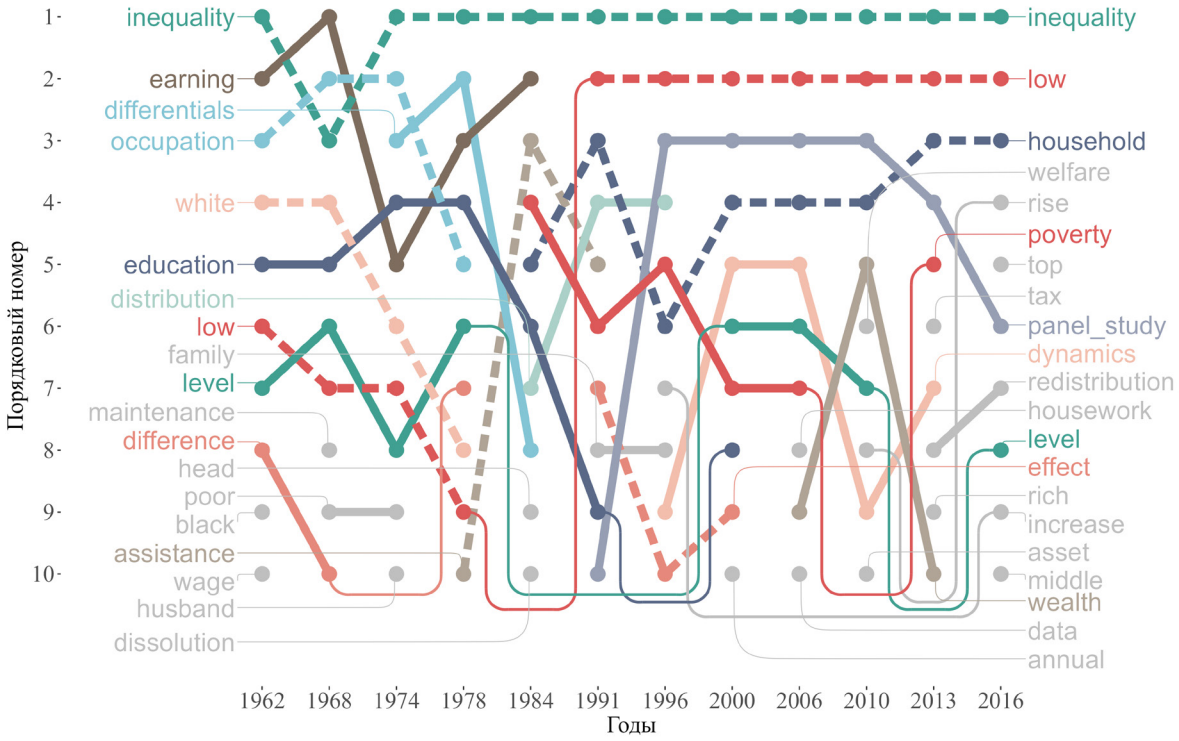
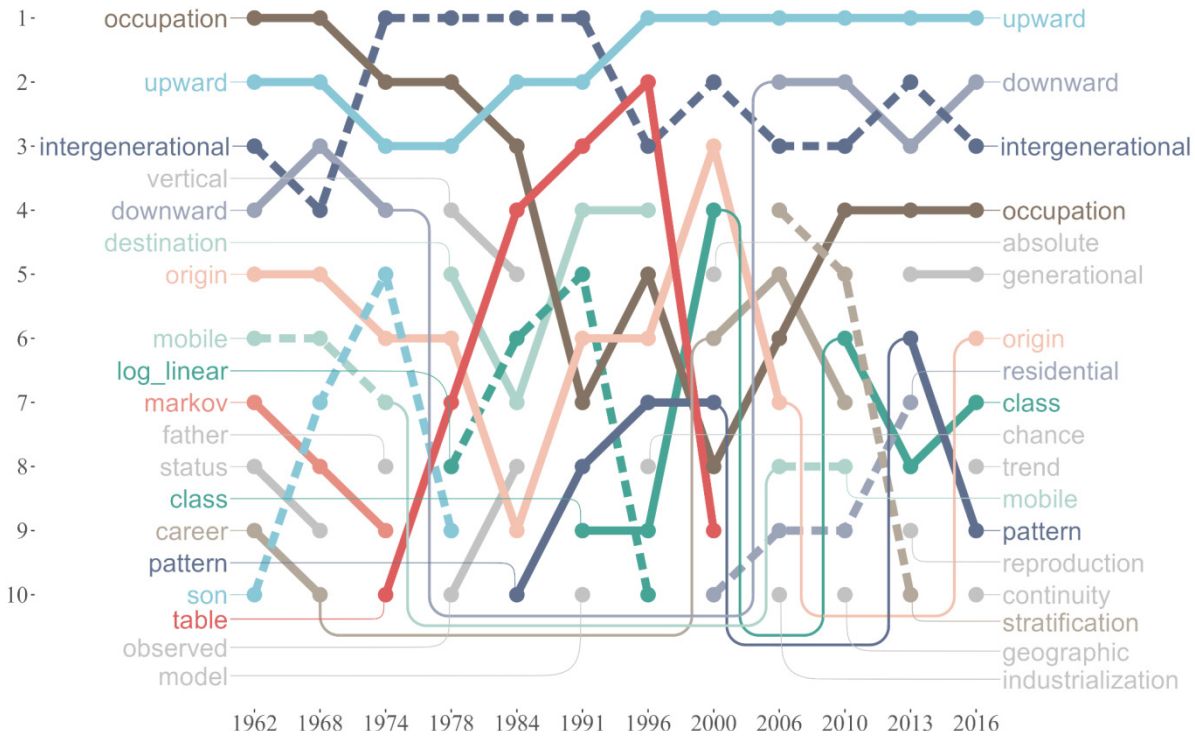


Рис. **П4.2.** Отношение числа упоминаний *class* (класса) после слов *social* (социальный), *working* (рабочий), *middle* (средний), а также перед словами *structure* (структура), *system* (система), *approach* (подход), *scheme* (схема), *position* (позиция), *model* (модель) к числу всех случаев упоминания



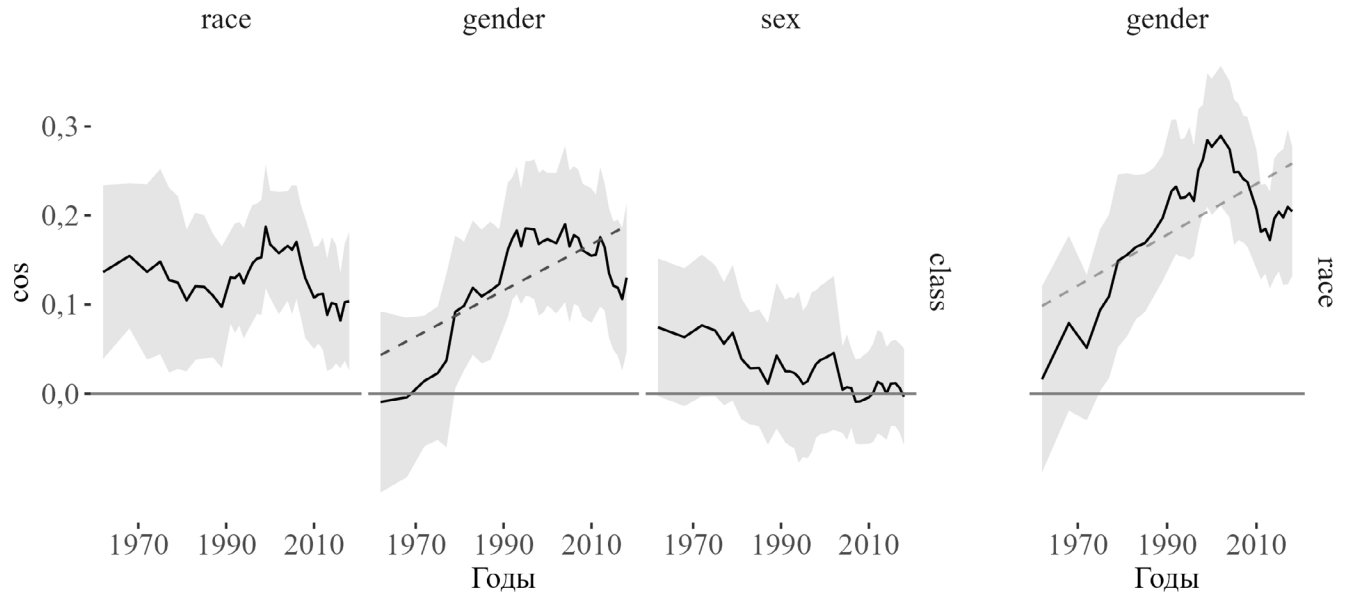
Примечание: Координаты по горизонтальной оси соответствуют средневзвешенному году публикации текстов, составляющих подкорпус. Серым цветом обозначены токены, которые встречаются менее трёх раз на диаграмме. Тонкие линии соединяют отметки токенов, покинувших на некоторый период десятку ближайших соседей.

Рис. **П4.3.** Ближайшие соседи токена *income* (доход)



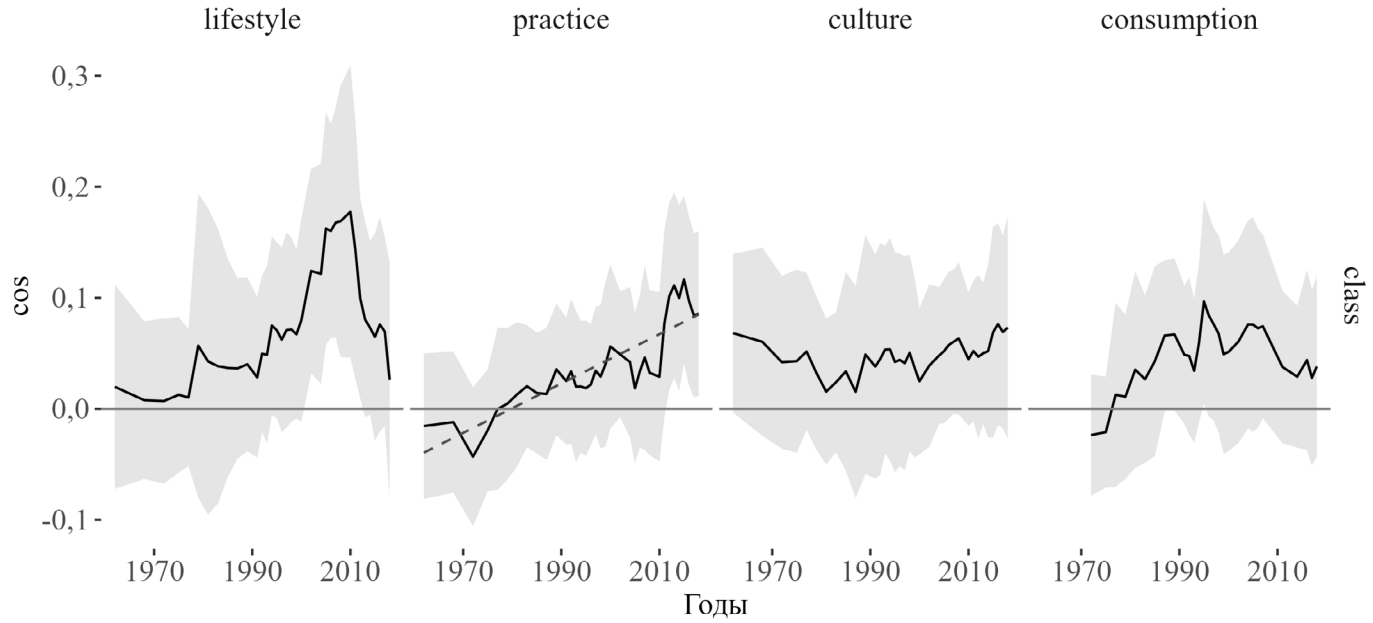
Примечание: Координаты по горизонтальной оси соответствуют средневзвешенному году публикации текстов, составляющих подкорпус. Серым цветом обозначены токены, которые встречаются менее трёх раз на диаграмме. Тонкие линии соединяют отметки токенов, покинувших на некоторый период десятку ближайших соседей.

Рис. **П4.4.** Ближайшие соседи токена *mobility* (мобильность)



Примечание: Координаты по горизонтальной оси соответствуют средневзвешенному году публикации текстов, составляющих подкорпус. Сплошные линии — бутстрепная оценка величины для каждого из 36 диахронных подкорпусов, серая заливка — 95%-ный доверительный интервал этой величины. Пунктирная линия отмечает аппроксимацию тренда линейной регрессией, представлена только в случае значимо отличного от 0 наклона.

Рис. **П4.5.** Косинус угла между вектором токена *class* (класс) и векторами токенов *race* (раса), *gender* (гендер), *sex* (пол), а также между *gender* и *race* в 36 диахронных подкорпусах 1930–2023 гг.



Примечание: Координаты по горизонтальной оси соответствуют средневзвешенному году публикации текстов, составляющих подкорпус. Сплошные линии — бутстрепная оценка величины для каждого из 36 диахронных подкорпусов, серая заливка — 95%-ный доверительный интервал этой величины. Пунктирная линия отмечает аппроксимацию тренда линейной регрессией, представлена только в случае значимо отличного от 0 наклона.

Рис. **П4.6.** Косинус угла между вектором токена *class* (класс) и векторами токенов *lifestyle* (стиль жизни), *practice* (пактика), *culture* (культура), *consumption* (потребление) в 36 диахронных подкорпусах. 1930–2023 гг.

Приложение 5

Методология построения векторного пространства

Векторное пространство строится на основе таблицы встречаемости токенов (соответствующих строкам) в различных контекстах (столбцах). Числа в таблице определяются не только взаимной сочетаемостью слов, но и распространённостью слов в корпусе. Более частотные слова будут иметь в среднем бóльшие значения в каждой ячейке соответствующих строк или столбцов. Для выявления действительной связи между словами нужна нормализация. Распространённой мерой является поточечная взаимная информация (*pointwise mutual information, PMI*):

$$PMI\,(x, y) = \log_2 \frac{P\,(x, y)}{P(x)\,P(y)},$$

где $P\,(x, y) = n\,(x, y) / N$ — наблюдаемая вероятность встретить x в контексте y;
$P\,(x) = n\,(x) / N$ и $P\,(y) = n\,(y) / N$ — вероятность встретить x и y соответственно;
$n\,(x, y)$ — число случаев, когда x оказался в контексте y;
$n\,(x)$ и $n\,(y)$ — число токенов типов x и y в корпусе;
N — размер корпуса.

Если токены встречаются вместе чаще, чем при полностью независимом распределении, то числитель дроби больше знаменателя, а логарифм положителен; если реже — отрицателен. Отрицательные значения *PMI* малоинформативны, так как оценка частоты совместного присутствия редко встречающихся слов требует очень большого корпуса. На практике большинство ячеек рассматриваемой таблицы заполнены нулями; следовательно, для них *PMI* стремится к минус бесконечности. По этой причине отрицательные значения *PMI* зануляются, такая мера называется положительной *PMI* (*PPMI*).

Один из недостатков *PPMI* заключается в том, что редкие слова в силу случайных причин периодически получают высокие значения. Скажем, если слово встречается лишь несколько раз в корпусе, вполне возможно, что эти несколько раз придутся на один фрагмент какого-то текста, в который также попал интересующий нас термин. Таким образом, из ближайших соседей какого-либо токена a, то есть контекстных токенов b с наибольшими *PPMI* (a, b), значительная часть окажется малочисленными токенами, оказавшимися рядом с a случайно. Решением является модифицированная мера:

$$PPMI_{\alpha}\,(x, y) = \max\,(0, \log_2 \frac{P\,(x, y)}{P(x)P_{\alpha}P\alpha(y)}),$$

$$\text{где } P_{\alpha}(y) = \frac{n\,(y)^{\alpha}}{\sum_{z\varepsilon c} n(z)^{\alpha}},$$

где $0 < \alpha < 1$;
C — словарь (множество типов токенов) рассматриваемого корпуса.

Для редких y величина $(y) > P(y)$, следовательно, значение $< PPMI$. Конвенционально принимается $\alpha = 0{,}75$ — результат, полученный в ходе ряда численных экспериментов [Levy, Goldberg, Dagan 2015]. Эта мера использована в работе.

Представленные в таблице данные уже можно рассматривать как векторы типов токенов (строки) в координатном пространстве контекстов (столбцов) (см., например, приложение 3, где представлен пример такого использования *PPMI*-таблицы). Преимущество такого представления — его наглядность и

прозрачность: координаты не являются результатом работы какого-то сложного алгоритма, но получены непосредственно из данных о сочетаемости слов. Это преимущество некоторые авторы признают принципиальным [Bolla et al. 2019]. Тем не менее такое векторное представление имеет и недостатки: очень большая размерность, соответствующая числу типов токенов, то есть размеру словаря корпуса; качество полученного пространства уступает альтернативным вариантам [Levy, Goldberg, Dagan 2015]. По этой причине предпринимается следующий шаг: сокращение размерности с помощью сингулярного разложения (*singular value decomposition*, SVD).

Сингулярное разложение заключается в представлении исходной матрицы M в виде $M = U*\Sigma*V^T$, где U, V — унитарные матрицы; Σ — диагональная матрица; ^{T} — знак транспонирования. Лежащие на диагонали элементы Σ называются сингулярными значениями; они упорядочены по убыванию от верхнего левого угла к нижнему правому. Если исходные данные были стандартизированы, то SVD эквивалентно хорошо знакомому социологам методу главных компонент[6] (*principal component analysis*, PCA). В этом случае произведение $U*\Sigma$ является матрицей координат исходных точек в новом пространстве, V — матрица поворота (нагрузок), диагональ Σ соответствует главным компонентам (*eigenvalue*s). Тем не менее SVD не требует непременной стандартизации данных и в этом смысле является обобщением PCA. Для построения векторного пространства используются первые d столбцов U, где d — достаточно большое число (конвенционально — 300). Матрица Σ при этом отбрасывается. Такое решение выглядит контринтуитивным, тем не менее его существенный положительный эффект подтверждён экспериментально [Levy, Goldberg, Dagan 2015].

[6] Точнее говоря, SVD является наиболее распространённым в современном программном обеспечении методом расчёта PCA.

Приложение 6

Сравнение мер семантических изменений

Для рассмотрения изменений конкретных токенов необходим какой-то способ сравнения семантических пространств, построенных на подкорпусах, которые относятся к различным временны*ы*м интервалам. Сами по себе координаты векторов в семантическом пространстве не несут какого-либо смысла, значение имеет лишь положение векторов относительно друг друга, а оно инвариантно к любому повороту или отражению координатных осей. Наиболее очевидным способом преодоления этого затруднения является прокрустово вращение. Суть процедуры заключается в подборе такого поворота и (или) отражения одного пространства, которое минимизирует сумму квадратов расстояний между точками, соответствующих одними и тем же типам токенов, в первом и втором пространствах. Возникает, однако, определённое логическое противоречие при использовании этого метода: мы, с одной стороны, предполагаем, что слова изменили свой смысл, а с другой — пытаемся максимально точно свести старые и новые значения. Возможно, например, что алгоритм совместит слова, изменившие своё значение, и из-за этого слова со стабильным смыслом окажутся ошибочно отдалёнными от самих себя в старом корпусе [Gonen et al. 2020].

Х. Дубоссарским и соавторами предложена альтернативная процедура [Dubossarsky et al. 2019]. Оси исходного пространства векторов (до сокращения), то есть столбцы таблицы значений *PPMI*, не являются произвольными, а соответствуют конкретным контекстным токенам. Таким образом, две *PPMI*-таблицы, соответствующие двум корпусам, могут быть совмещены в общую матрицу с числом строк, равным сумме чисел строк исходных, и числом столбцов, равным числу типов токенов в объединённом корпусе. Данная особенность делает удобной работу с векторным пространством, построенным непосредственно на таблицах *PPMI* (см., например: [Hamilton, Leskovec, Jurafsky 2016b]. Однако нет препятствий для сокращения размерности объединённой таблицы, то есть вместо объединения сокращённых пространств с помощью прокруста можно осуществить сокращение предварительно объединённого пространства [Dubossarsky et al. 2019; Aida et al. 2021].

Принципиальная альтернатива заключается в отказе от совмещения пространств и использовании локальной меры изменения, то есть изменения положения относительно ближайших соседей [Hamilton, Leskovec, Jurafsky 2016a]. Метод предполагает отбор некоторого числа k ближайших соседей слова в одном и другом пространствах. Составление общего списка из соседей длиной (g), лежащей в диапазоне от k, если списки совпадают, до $2k$, если они не имеют пересечений. После этого вектор слова в каждом из пространств может быть представлен g-мерным вектором, состоящим из косинусов слова с каждым из g его соседей. Изменение слова будет измерено как косинус угла между этими двумя векторами размерности g. Как утверждается, глобальные меры, предполагающие выравнивание всего пространства, в большей степени схватывают общие синтагматические сдвиги, а локальная мера — парадигматические [Hamilton, Leskovec, Jurafsky 2016a].

Наконец, можно просто отказаться от сравнений координат в любой форме, а вместо этого сравнивать списки ближайших соседей слова [Antoniak, Mimno 2018; Gonen et al. 2020]. Традиционно оценивается степень пересечения этих списков. Для множеств слов-соседей A и B размера k мерой близости будет пропорциональное пересечение этих списков $|A \cap B| / k$ или мера Жаккарда $|A \cap B| / |A \cup B|$, где $|.|$ — количество членов множества. Эти и подобные им меры имеют существенный недостаток, а именно чувствительность к изменению произвольного параметра k. Одно из предлагаемых решений — значительное увеличение k с 10 до нескольких сотен [Gonen et al. 2020]. Но такое решение не устраняет главного упущения традиционного подхода — игнорирования упорядоченности списков соседей. Рассмотрим простой пример: есть два упорядоченных списка из четырёх элементов: *ABCD* и *ABDC*. Если

сравнивается состав четырёх первых элементов, то списки идентичны, мера близости равна 1; если три, то идентичны только две трети элементов, мера близости — ⅔, жаккардова мера — 0,5.

Таблица П6.1

Пример расчёта среднего пересечения последовательностей

k	Список 1	Список 2	A$k = \|A \cap B\| / k$	Среднее пересечение
1	*A*	*A*	1	1
2	*AB*	*AB*	1	1
3	*ABC*	*ABD*	$^{2}/_{3}$	$^{8}/_{9}$
4	*ABCD*	*ABDC*	1	$^{11}/_{12}$
5	*ABCDE*	*ABDCF*	$^{4}/_{5}$	$^{67}/_{75}$

Для сравнения таких рядов была предложена специальная мера [Webber, Moffat, Zobel 2010]. Продолжим пример (см. табл. П6.1), если для всех значений k от 1 до d посчитать пересечение (A_k) и усреднить, полученная величина (среднее пересечение) обладает желательными свойствами: она меньше зависит от выбора d, элементы, стоящие впереди, вносят больший вклад в итоговую величину. Было также предложено ввести дополнительные веса, позволяющие контролировать скорость снижения влияния членов списка при движении от начала к концу, в качестве же весов взять члены геометрической прогрессии:

$$RBO = (1 - p) * p^{k-1} \qquad (*)$$

Веббер с соавторами рассматривают два ряда длиной d как первые d членов двух бесконечных последовательностей, чьей мерой схожести является (*). Суммирование в этом случае ведётся от 1 до ∞. На основе знания первых d членов может быть получена следующая оценка меры схожести этих последовательностей:

$$= A_d * p^d + (1 - p) * p^{k-1}.$$

Эта величина принята нами в качестве показатели схожести списков ближайших соседей. Коэффициент p взят как 0,92. При таком значении первые 10 соседей задают 80% веса общей величины *RBO* (*rank-biased overlap*) [Webber, Moffat, Zobel 2010: формулы 7, 23, 21].

В совокупности нами применены четыре метода сравнения векторов двух семантических пространств в отношении подкорпусов, соответствующих периодам 1945–1980 и 2005–2023 гг.:

— косинусная схожесть на пространствах, выровненных с помощью прокрустова вращения;

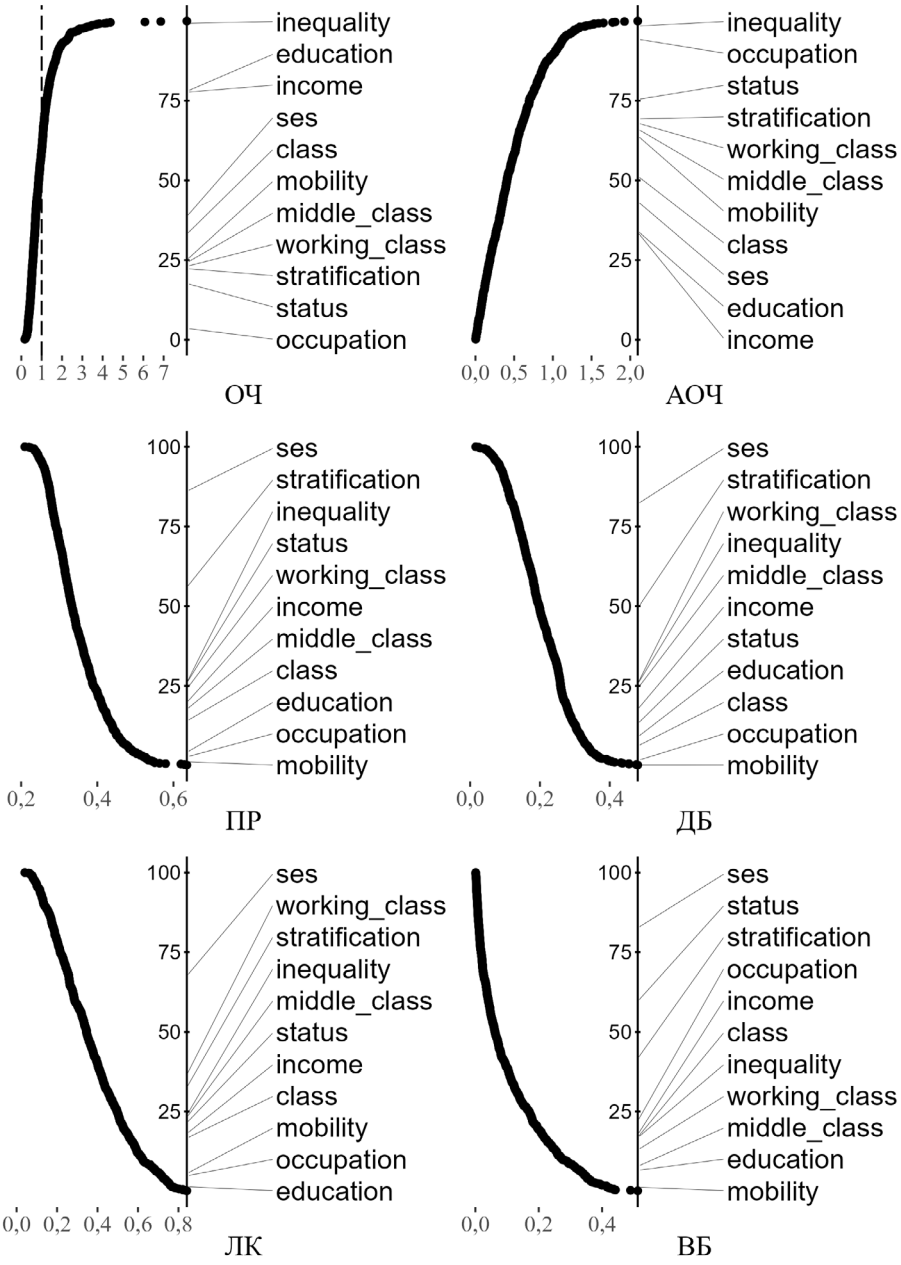
— косинусная схожесть в общем пространстве, полученном на основе соединения двух таблиц *PPMI*;

— локальная мера схожести;

— *RBO* Веббера и соавторов.

Результаты представлены на рисунке П6.1, корреляция Спирмена между ними показана в таблице П6.2. Также приведены отношения частотностей (выбранные периоды отличаются от рассмотренных в основном тексте; см. раздел «Частотный анализ») и абсолютное отношение частотностей, то есть отношения без учёта направления изменения (уменьшение или увеличение). Абсолютные отно-

шения частотностей получено как модуль логарифма отношения частотностей: увеличение в два раза и уменьшение в два раза дадут одно и то же её значение.



Примечания:

Условные обозначения: ОЧ — отношения частотностей; АОЧ — абсолютное отношение частотностей; ПР — выравнивание прокрустовым вращением; ДБ (Дубоссарский) — сокращение размерности объединённой PPMI-таблицы; ЛК — локальная косинусная близость; ВБ (Веббер) — сравнение списков соседей по Вебберу и коллегам; ses — socioeconomic status (социально-экономический статус).

Для всех индикаторов, кроме ОЧ, нижние значения указывают на стабильность, верхние — на изменение. Для ОЧ отсутствие изменений соответствует 1 (отмечено пунктирной линией). На вертикальной шкале обозначены перцентили: 75 означает, что токен, находящийся в этой точке, изменился сильнее, чем 75% анализируемых токенов. Представлены результаты только для токенов, встретившихся не менее 100 раз в каждом из сравниваемых подкорпусов.

Рис. **П6.1.** Показатели семантического изменения между периодами 1945–1980 и 2005–2023 гг.

Таблица П6.2

Корреляция индикаторов семантического изменения слов

	АОЧ	ПР	ДБ	ЛК	ВБ
ОЧ	– 0,35	0,19	0,12	0,22	0,17
АОЧ		– 0,04	– 0,02	– 0,09	– 0,08
ПР			0,91	0,89	0,78
ДБ				0,81	0,71
ЛК					0,85

Примечание. Условные обозначения: ОЧ — отношения частотностей; АОЧ — абсолютное отношение частотностей; ПР — выравнивание прокрустовым вращением; ДБ (Дубоссарский) — сокращение размерности объединённой PPMI-таблицы; ЛК — локальная косинусная близость; ВБ (Веббер) — сравнение списков соседей по Вебберу и коллегам.

Как видно из таблицы П6.1, все векторные меры довольно сильно связаны между собой. Корреляция с отношением частотностей тоже есть, но несравнимо меньшая. Абсолютное отношение частотностей неожиданным образом оказалось отрицательно связано с прочими показателями. Очевидный вывод: только увеличение частотности, а не всякое её изменение, позволяет ожидать, что слово модифицировало семантическое значение. То, что корреляции отрицательны, является, по всей видимости, артефактом процедуры: сравнивались только слова, встречающиеся не менее 100 раз в каждом из подкорпусов, при этом объём второго корпуса ощутимо больше; следовательно, у относительно редких слов, чья частотность увеличилась, было меньше шансов попасть в рассматриваемую выборку, чем у частых слов, чьё употребление сократилось.

Несмотря на сильную корреляцию метрик, расположение слов из нашего списка существенно варьируется между методами. Значительное семантическое изменение, согласно всем индикаторам, осуществил *ses* (*socioeconomic_status)* — социально-экономический статус. Большинство же терминов укладываются в 50% наиболее стабильных по всем показателям. Можно ли считать на основе этих наблюдений значение слов, принадлежащих к рассматриваемому семантическому полю, неизменным, за исключением социоэкономического статуса? Анализ усреднённого по разным метрикам списка наиболее изменившихся токенов показывает, что подавляющее большинство из них никак нельзя отнести к социологическим терминам. Их изменение, таким образом, относится к лингвистическим фактам, лежащим далеко за пределами наших интересов. Строгий утвердительный ответ был бы возможен при сравнении целевых слов с другими научными терминами, но такого словаря у нас нет под рукой, хотя его создание представляется возможным на основе социологических справочников, словарей, энциклопедий, а также анализа ключевых слов публикаций.