УДК 008

Константин Антоненко

КОЛИЧЕСТВЕННЫЙ АНАЛИЗ КУЛЬТУРНОГО СЛЕДА ТЕХНИЧЕСКОГО ПРОГРЕССА

Висвітлено один зі способів застосування досягнень сучасної техніки в якості інструменту дослідження історії техніки методом аналізу її культурного сліду (impact). Звертається увага на появу нового джерела соціометричних даних. Наведені приклади і описані обмеження, які пов'язані з використанням російської мови. Здійснена пробна оцінка динаміки змін суспільного інтересу до технічного прогресу в російськомовному світі.

Ключові слова: *культуроміка, історія техніки, кількісний аналіз, квантитативні методи в історії, оцифровані тексти.*

Показан один из способов применения достижений современной техники в качестве инструмента исследования истории техники методом анализа ее культурного следа (impact). Обращается внимание на появление нового источника социометрических данных. Приведены примеры и описаны ограничения, накладываемые особенностями русского языка. Сделана пробная оценка динамики изменения общественного интереса к техническому прогрессу в русскоязычном мире.

Ключевые слова: *культуромика, история техники, количественный анализ, квантитативные методы в истории, оцифрованные тексты.*

One of the means of applying achievements of present-day technology is shown as a history of technology research instrument for analysis of its cultural impact. Attention is drawn to appearance of the new source of sociometric data. Examples are provided and the limits imposed by peculiarities of Russian language are given. The trial estimate is made to present the dynamics of changes of public interest to technical progress in Russian speaking world.

Key words: *culturomics, history of technology, quantitative analysis, quantitative methods in history, digitized texts*

«Прогресс в науке
обратно пропорционален
числу выходящих журналов»
6-й закон Паркинсона

Работа с отдельными тщательно отобранными документами позволяет исследователям делать выводы о тенденциях развития человеческой мысли. Однако такой подход не всегда дает возможность точно измерить явления, лежащие в основе рассматриваемых процессов. Попытки использования квантитативных методов в гуманитарных областях знания предпринимаются давно, однако обычно они сопряжены с трудоемкостью сбора или недостаточностью исходных данных для анализа.

В таких областях как социология и маркетинговые исследования предпринимаются специальные действия по сбору необходимых данных, однако вряд ли возможно говорить о возможности подобных действий в прошлом, находясь во времени настоящем.

16 декабря 2010 года группа исследователей опубликовала статью "Количественный анализ культуры на основе миллионов оцифрованных книг" [1] и открыла для публичного доступа описанный в этой статье набор данных [2].

Авторы указанной работы на основе крупнейшей из когда-либо существовавших коллекций книг – проекта Google Books – сформировали корпус из 5195796 текстов, составляющий около четырех процентов всех книг, напечатанных за всю историю человечества. По критерию качества распознавания текста и качества описания, из 15 миллионов оцифрованных книг (12% когда-либо существовавших) были отобраны упомянутые пять миллионов. Сформированный корпус содержит 500 миллиардов слов на семи языках: английском, французском, испанском, немецком, китайском, русском и иврите, и покрывает период с 1500 по 2009 годы. Под словами у авторов указанной работы, а также здесь и далее понимаются любые сочетания символов, включая слова, аббревиатуры, числа и опечатки. Различные словоформы одного слова также считаются разными словами, что имеет особое значение для русского языка. Русскоязычная часть корпуса содержит 35 миллиардов слов.

Как сообщают авторы, весь корпус целиком пока не может быть опубликован по соображениям юридического харктера, поэтому опубликован набор данных, который содержит информацию в обобщенном виде: частоты сочетаний из N слов («N-грамм» в авторской терминологии) для N от 1 до 5, а также частоты страниц и книг, содержащих эти N-граммы. Под частотой здесь понимается соотношение количества упоминаний слова или словосочетания (N-граммы) к общему количеству слов (страниц, книг) в корпусе для выбранного года и языка.

Эти данные представлены отдельно для каждого из языков и даны раздельно по годам. Последнее обстоятельство и позволяет рассматривать эти данные как представляющие интерес для анализа в историческом аспекте.

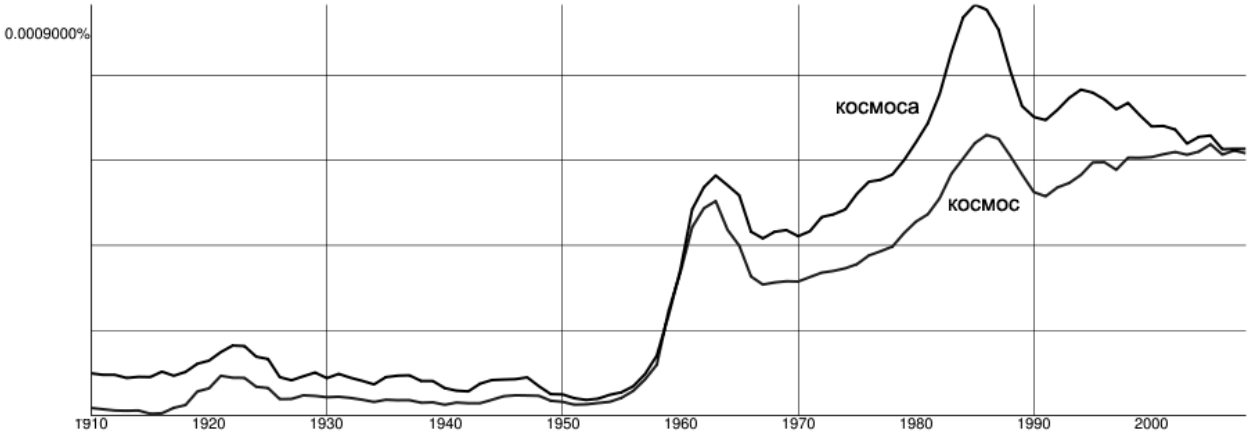
В исходной работе авторы демонстрируют анализ изменений размера английского лексикона со временем, эволюцию грамматики, некоторые свойства «коллективной памяти» и проявления цензуры. Доступен также онлайн-интерфейс для получения простейших зависимостей частоты от времени [3].

Спустя почти семь месяцев, 13 июля 2011 года, проект «Национальный корпус русского языка» [3] в режиме тестирования также запустил аналогичный сервис поиска с распределением по годам [4]. Открыто доступен только интерфейс поиска и ограниченно доступны численные данные по частотам (только по запрашиваемому с помощью интерфейса слову), что можно считать недостатком по сравнению с проектом [2]. С другой стороны, неоспоримое преимущество проекта [3] – это лучшая систематизация и разметка корпуса, возможность работы (поиска) в его подмножествах по типам текстов и по признакам словоформ.

В этой статье использованы данные проекта [2] по причине его лучшей пригодности для автоматизированной компьютерной обработки (доступности наборов обобщенных данных). Однако использование обоих упомянутых и других подобных источников по отдельности либо совместно, очевидно, может представлять особый интерес.

Культурный прогресс человечества, в том числе и прогресс техники как неотъемлемой части культуры, сопровождается изменениями обсуждаемой тематики («повестки дня») как в количественном, так и в лингвистическом аспектах [6]. Эти процессы оставляют наблюдаемый след в виде принятых решений, предпринятых действий, созданных вещей, а также в различных публикациях.

Для обозначения такого подхода к анализу культуры авторы работы [1] ввели термин «культуромика» [7].



Фиг. 1. Частота слова «космос» в именительном и родительном падежах

Например, интенсивность обсуждения темы космоса очевидным образом лавинообразно вырастает в период 1961–1963 и никогда более не опускается до уровня «докосмической эры» (фиг. 1). [1]

Со временем, по мере развития событий, старые темы уходят с повестки дня, замещаясь либо темами новыми, либо собственным их развитием по мере разработки и специализации (фиг. 2).

Подходы аналогичной природы – различные индексы цитирования (citation index) и индексы влияния (impact factor) – давно используются в наукометрии и маркетинге науки, хотя и подвергаются некоторой критике [8] (фиг. 2).

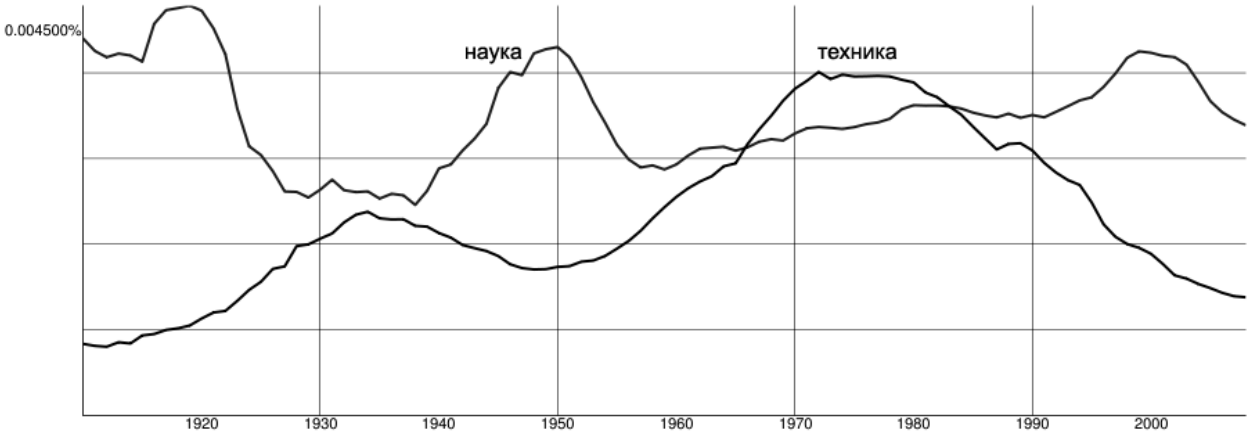
Фиг. 2. Частота терминов «индекс цитирования» и «наукометрия»

Однако в нашем случае речь идет не просто о количестве публикаций некоторой тематики, а о более широком явлении, давно замеченном и описанном в литературе: последние (новые) научные открытия и достижения техники издавна использовались в качестве моделей и метафор для описания и понимания природы и человека. В XVII столетии Вселенная и живые существа часто описывались в терминах часового механизма, а в XIX – паровой машины [9, 10]. Сегодня можно наблюдать, как во вненаучной сфере для описания человеческого мозга и процесса понимания вообще используется метафора компьютера, а термин «социальные сети» вобще стирает разницу между технологическими достижениями и образом жизни. Классический пример частного проявления этой тенденции – слово «бикини», обозначающее вид одежды, получивший свое название в честь одноименного атолла как следствие интереса к происходившим там событиям. Слово «atomic» в середине XX века часто использовалось в качестве названий продуктов и торговых марок (фиг. 3) и т.п. Еще один пример – использование слова «нанотехнологии» в рекламе косметических средств с целью подчеркнуть ис-

[1] Все графики, приведенные в статье, получены автором на основе данных [2]: фиг. 1–23 – с использованием инструментария [3], фиг. 24–27 – с помощью расчетов и инструментов, сделанных самостоятельно.

пользование современных высоких технологий (даже не имеющих отношения к нанотехнологиям вообще) в производстве потребительских и промышленных продуктов.

Фиг. 3. Фотография упаковки безопасных лезвий с использованием в названии популярной в 50-е годы XX века в США атомной тематики (Коллекция Health Physics Historical Instrumentation Museum, Oak Ridge Associated Universities – http://www.orau.org/ptp/collection/brandnames/brandnames.htm)



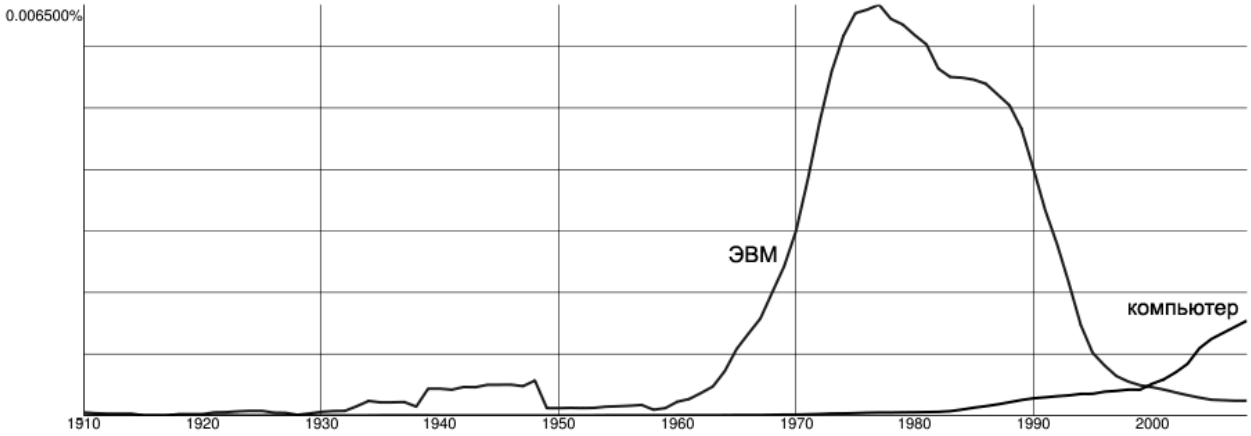
Фиг. 4. Частота слов «наука» и «техника» в русскоязычной части корпуса

Здесь и далее, если не указано особо, частоты выражены в процентах от общего количества за тот же период времени. Также применяется сглаживание (скользящее среднее) с размером окна от 1 до 3 лет. Обратим внимание на максимум частоты «техники» в районе 1971–1975. Далее обратим на это внимание еще один раз.

Появившиеся данные [2] могут использоваться в качестве как вспомогательного, так и – после достаточной их верификации – основного метода

исследования разнообразных, связанных с научно-техническим прогрессом культурных явлений, например, таких, как выявление и локализация во времени различных общественных процессов, анализ характера их развития, исследование влияния различных факторов, как естественных, так и искусственных, изучение тенденций использования терминов и понятий от их рождения до выхода из употребления (фиг. 4, 5, 6, 7) и т.п.

Фиг. 5. Частота слов «самолеты» и «ракеты»



Фиг. 6. Иллюстрация жизненного цикла терминов «ЭВМ» и «компьютер»

Анализируя эти данные, можно также вложить некоторый верифицируемый смысл в такие слабоформализуемые понятия, как «важность», «известность», «признание», «расцвет», «упадок» в применении к тем или иным явлениям, событиям и технологиям.

Появляется возможность быстро (без обращения к специализированным источникам) установить верхнюю границу времени возникновения термина или понятия, а также наблюдать смену значения термина или понятия (фиг. 8, 9).

Фиг. 7. Иллюстрация изменения грамматической нормы слова «эксплуатация»

Фиг. 8. Частота слов «самолетъ» и «аэропланъ» (в дореформенном написании)

3. БЕЗПОЛЕЗНОЕ.

271. ATRACTILIS.

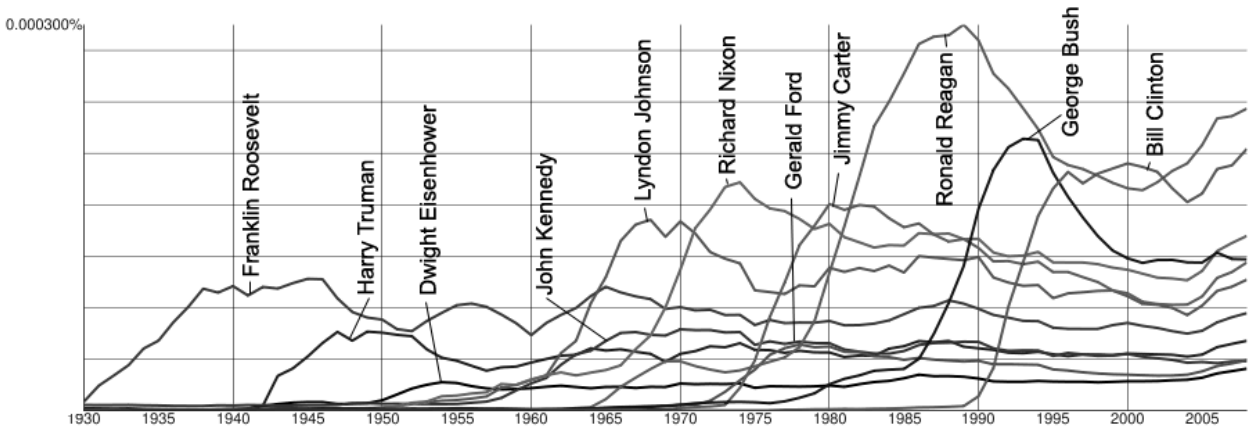
Общій Оцвѣтникъ умноженный. *Вѣнчикъ* съ лучами: въ окружности находящіеся *Женскіе Вѣнчики* пятизубчатые, съмена выкидывающіе. *Ложе* шелушистое. *Самолетъ* перяный.

802. A. *flava.* Стебель и листья пушистые.

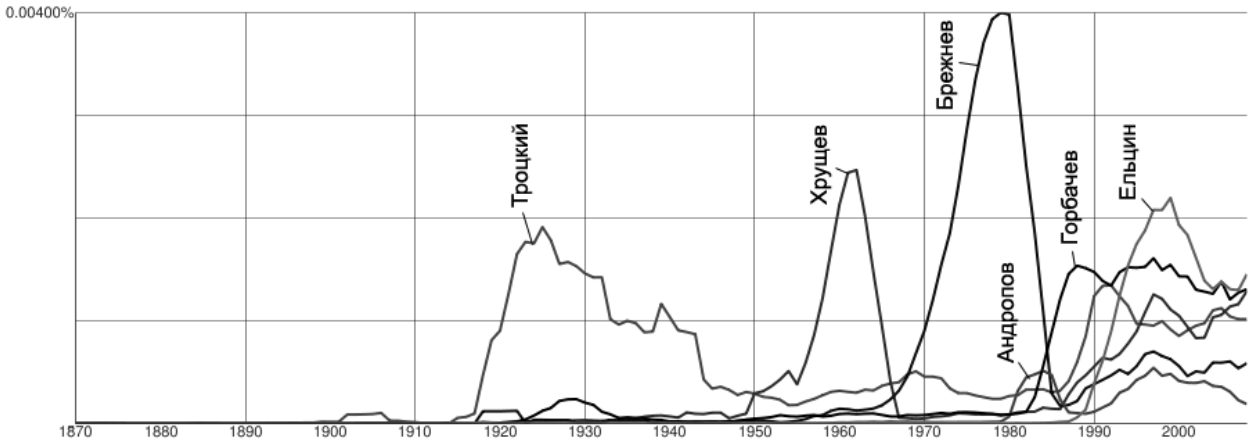
Фиг. 9. Фрагмент текста, содержащего слово «самолетъ» в значении, отличном от современного. Из «Умозрительныя изслъдованія Императорской Санктпетербургской Академіи наукъ» Том. 1, 1808 год, стр. 481

Можно также говорить о возможности выявления факта искусственности существования в ноосфере того или иного понятия. Для этого рассмотрим несколько выходящий за рамки специализации журнала, однако очевидный пример, иллюстрирующий то наблюдение, что при отсутствии искус-

ственного влияния (пропаганды или цензуры), частота упоминания ведет себя во времени схожим образом, спадая по кривой, имеющей характер экспоненты, что свойственно естественным (в широком смысле) процессам. Поэтому можно предположить, что отличающееся поведение может свидетельствовать об искусственном принудительном проталкивании (пропаганда) либо замалчивании (цензура) темы, либо о влиянии неких необычных процессов, приведших к вытеснению одних тем другими. Этот вопрос детально рассмотрен в [1].



Фиг. 10. Частота имен президентов США в англоязычных текстах



Фиг. 11. Частота имен руководителей СССР и России в русскоязычных текстах

Рассматривая частоту упоминания имен американских президентов в англоязычных текстах и руководителей СССР – в русскоязычных (фиг. 10 и 11), это различие легче всего наблюдать на примере Хрущева и Брежнева. После отстранения и смерти соответственно, количество упоминаний этих фамилий в текстах падает практически до нуля. Легко заметить, что после прекращения существования СССР и деполитизации издательского дела характер кривых (фиг. 11, интервал 1990–2008) мгновенно уподобляется характеру кривых, которые можно наблюдать для имен президентов США.

Пример аналогичного поведения можно легко найти и для технических тем (фиг.12,13).

Анализ данных в историческом разрезе позволяет также наблюдать влияние различных исторических событий (фиг.14).

Фиг. 12. Частота темы «Стратегической оборонной инициативы»

Фиг. 13. Частота темы «Стратегической оборонной инициативы» в англоязычном мире

Всплески интереса к истории науки и техники совпадают по времени и, очевидно, связаны с принятым Пленумом ЦК ВКП(б) в 1929 году постановлением о введении в высших технических заведениях страны преподавания курса «марксистской истории техники» [11] и кампанией поиска приоритетов в советской науке в 1948 году (фиг. 15).

Однако при работе с этими данными следует учитывать особенности и ограничения, налагаемые свойствами русского языка. Его синтетическая природа вносит существенные сложности в работе с русскоязычной частью данных. Существование множества флексий и сложного словообразования вынуждает обращаться к использованию морфологических словарей либо использовать специальные приемы, разрабатываемые современной компьютерной лингвистикой.

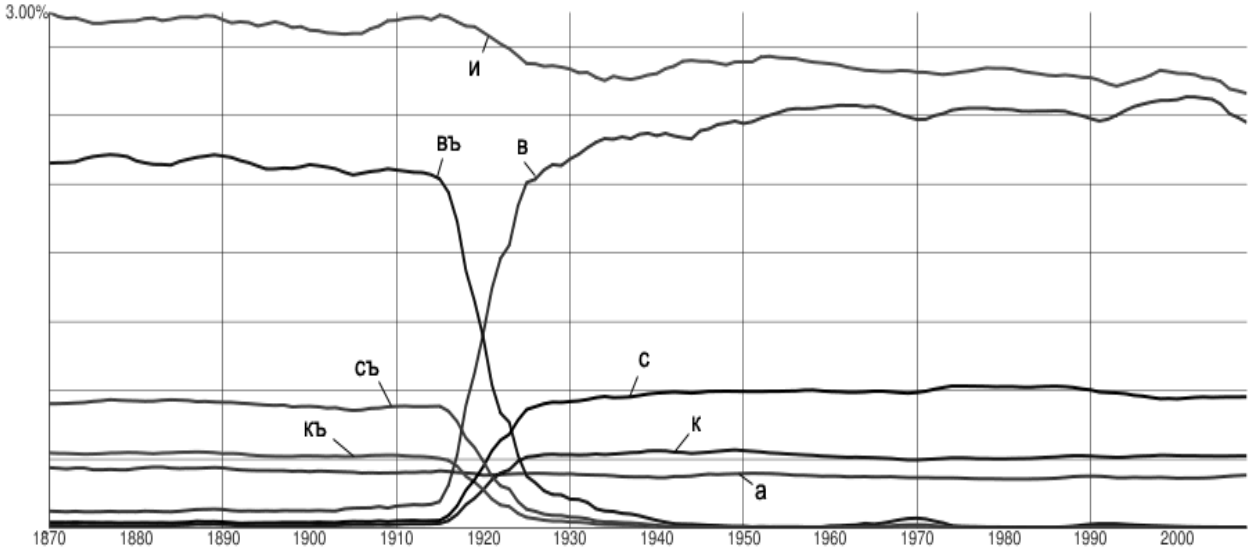
Еще один фактор, который необходимо учитывать – это реформа грамматики русского языка 1917–1918 годов, дополнительно увеличившая множество вариантов слов при переходе анализируемого интервала времени через время реформы.

Фиг. 14. Частота сочетаний «истории науки» и «истории техники». Родительный падеж выбран из тех соображений, что в течение всего исследуемого периода эти словосочетания наиболее часто употреблялись именно в этом падеже, например: «курс истории ...»

Фиг. 15. Порождение кампании по борьбе с «космополитизмом» – лозунг «низкопоклонство перед Западом» [12] совпадает по времени со вторым всплеском интереса к истории науки и техники

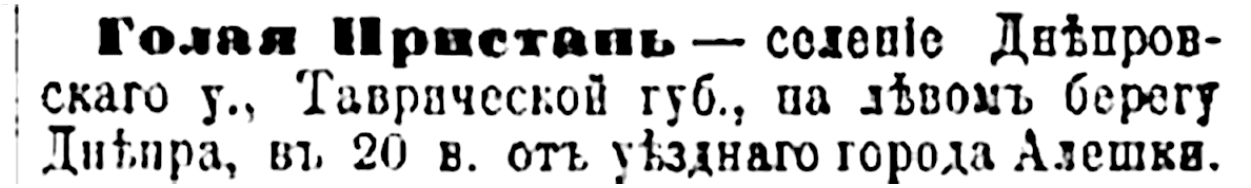
Дополнительная сложность заключается в том, что система автоматического распознавания текста, применявшаяся при оцифровке, не была настроена для распознавания символов дореформенного русского языка (ять, и десятиричное, фита, ижица). В результате в словах они могут быть подменены непредсказуемыми комбинациями других символов (например, «1» или «!» вместо «и десятеричное» («i») или «твердый знак» вместо «ять»). Таким образом, качество данных для периода ранее 1920 года резко отличается в худшую сторону, а количество слов, которые достоверно могут быть исполь-

зованы для анализа (не содержащих «опасные» символы), сильно ограничено по сравнению с периодом 1920–2008. Из графика, полученного путем анализа частоты простых нейтральных общеупотребительных слов (фиг. 16), видно, что реформа практически полностью произошла в интервале между 1917 и 1923 годами.



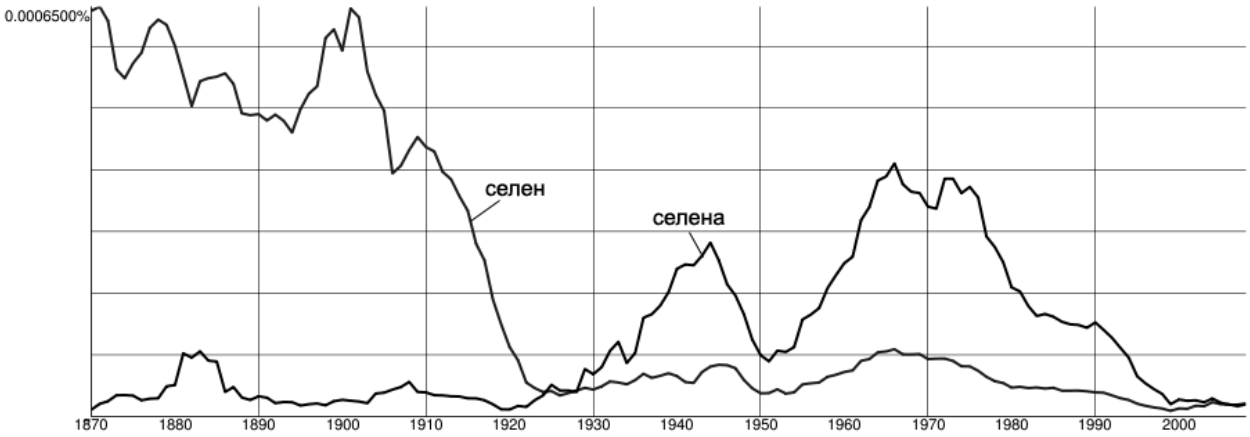
Фиг. 16. Иллюстрация реформы русской грамматики
на примере общеупотребительных нейтральных слов:
предлоги и союзы «и», «а», «к», «в», «с»

Продемонстрируем пример ошибки, возникающей в подсчете частот из-за неверного распознавания текста в дореформенной орфографии, на примере названия химического элемента «селен» в именительном и родительном падежах. На фиг. 17 показан фрагмент текста, в котором слово «селенiе» ошибочно распознано как «селен!е», с последующей естественной интерпретацией восклицательного знака как символа-разделителя. Подобные ошибки в массовом количестве приводят к искажениям результатов, показанным на фиг. 18 в левой части временной шкалы (до 1920 года).


Голая Пристань — селенiе Днѣпровскаго у., Таврической губ., на лѣвомъ берегу Днѣпра, въ 20 в. отъ уѣзднаго города Алешки.

Фиг. 17. Фрагмент текста из Энциклопедического словаря 1803 года

Следующим фактором, требующим особого рассмотрения, является происхождение этих данных. Набор составлялся в англоязычном мире на основе библиотечных книг. Учитывая небольшую роль русского языка в Западном мире, а также политические препятствия («железный занавес», «холод-



Фиг. 18. Иллюстрация ошибочных результатов за счет дефектов распознавания текста на дореформенном русском языке на примере химического элемента «селен»

ная война» и т.п.), строго говоря, нельзя быть уверенными в репрезентативности корпуса текстов на русском языке, использованных при составлении упомянутого набора (это в меньшей степени является проблемой для [4]). Неясно также, какую долю занимает «эмигрантская литература» и насколько существенно ее влияние. Для формирования данных использовались, в основном, доступные исследователям зарубежные библиотеки. Пока трудно сказать, как можно оценить эти факторы, однако в пользу того, что потенциально эта проблема разрешима, можно привести такие доводы:

а) можно предположить, что более важные тексты попадали в зарубежные библиотеки с большей вероятностью, что при оценке важности работает как положительная обратная связь;

б) наибольшие искажения пропорций, вероятно, должны быть в количестве текстов, а соответственно, и данных по терминам политической и пропагандистской тематики, и в меньшей степени должны затрагивать тематику техники и технической культуры;

в) учитывая упомянутые факторы, можно применять методы калибровки по нейтральным либо заведомо известным семантически связанным темам;

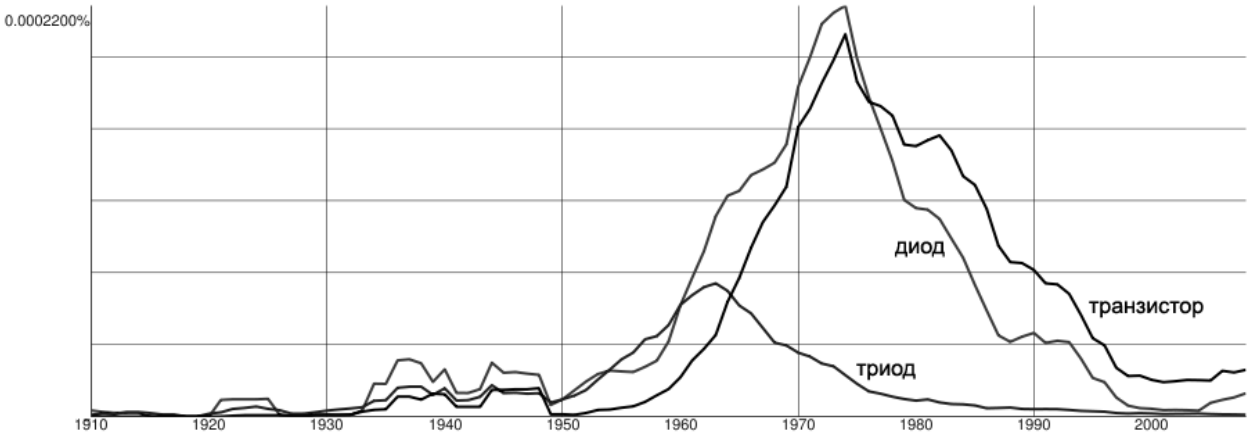
г) количественные измерения в любом случае подвергаются качественной интерпретации;

д) по мере продолжения работ по оцифровке, если таковые будут продолжаться, есть шанс подойти близко к 100% изданных книг, и тогда все искажения этой природы автоматически нивелируются.

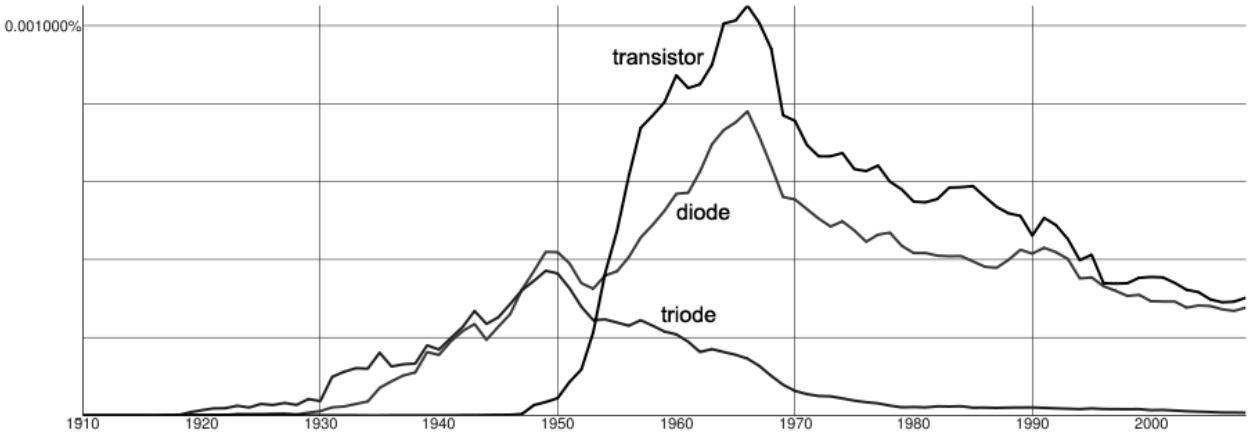
Принимая во внимание перечисленные особенности, эти новые данные можно использовать для анализа характера некоторых культурных явлений в историческом и сравнительно-географическом аспектах.

Ниже в качестве иллюстрации приведены несколько графиков (фиг. 19 – 24), показывающих культурный след различных этапов развития электронной техники в русскоязычном и англоязычном мирах.

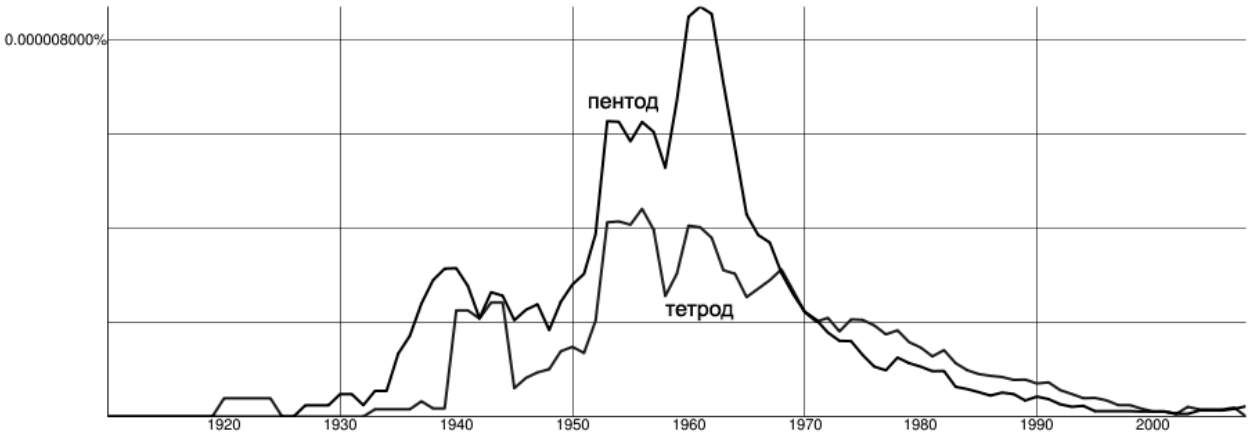
Фиг. 19. Электронные технологии



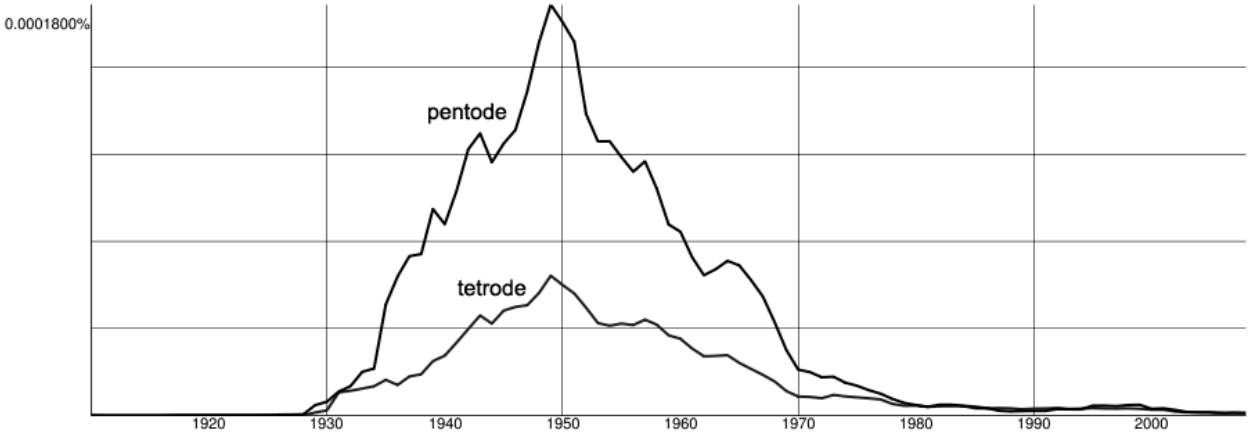
Фиг. 20. Элементная база электронной техники в русскоязычном мире



Фиг. 21. Элементная база электронной техники в англоязычном мире



Фиг. 22. Электронно-вакуумные приборы в русскоязычном мире



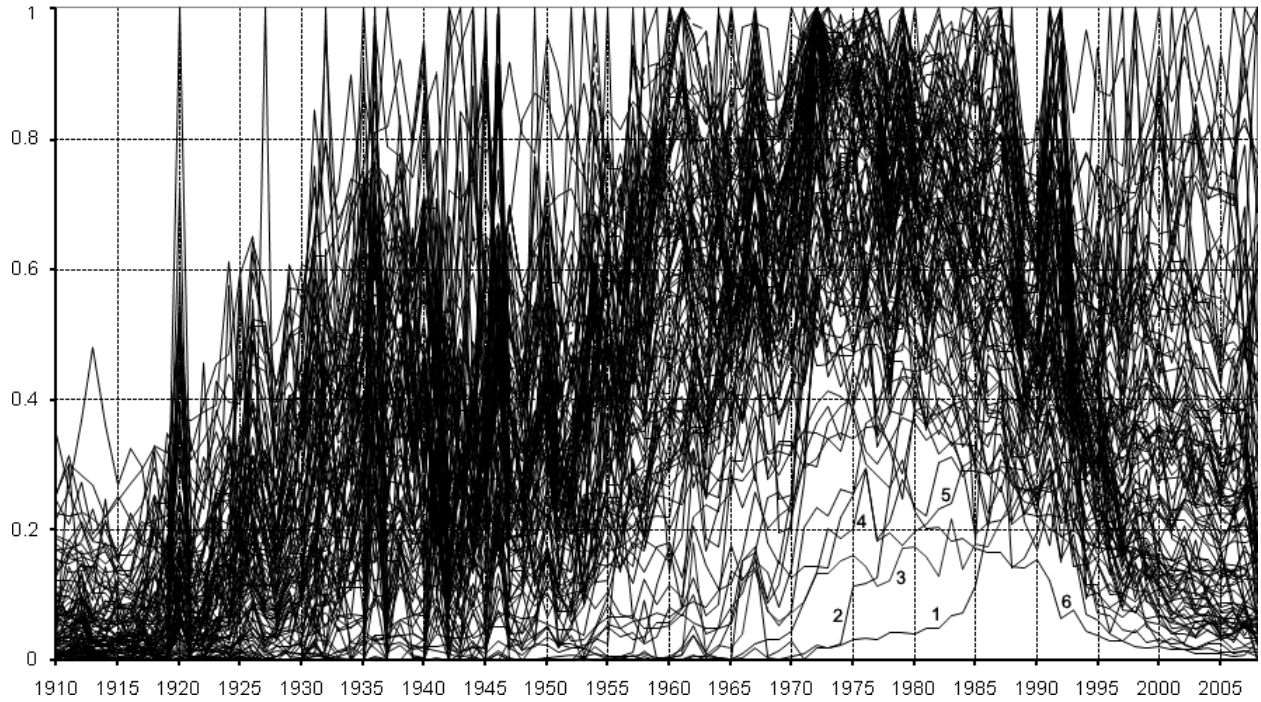
Фиг. 23. Электронно-вакуумные приборы в англоязычном мире

В качестве эксперимента автором была предпринята попытка оценки общественного интереса к технике вообще. Для этого был сформирован список из 112 слов (частей слов), выбранных следующим образом: половину списка составили слова, которые чаще других стояли рядом со словом, содержащим корень «техн» (эта часть списка была получена с помощью автоматизированной компьютерной выборки по указанному условию и вручную очищена от слов и словосочетаний очевидно неподходящей направленности: «техника исполнения», «техника танца» и т.п.). Вторая половина списка была составлена вручную путем выбора из классификатора технических специальностей ВАК (код 05) слов, которые могли бы однозначно отражать направление техники или технологии (Табл. 1).

Табл. 1.

авиа	измер	микропроцессор	сельскохоз
автомат	изоляц	навигац	сельхоз
автомоб	инструмент	нефт	систем
агрегат	информатик	оборуд	скваж

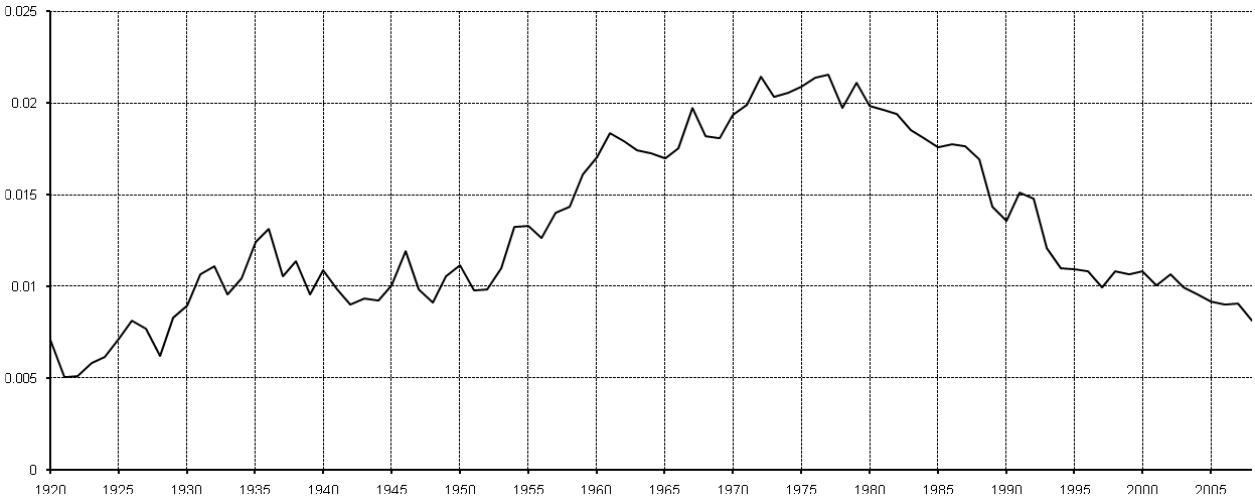
агро	информацион	обработ	строител
акуст	иониз	опти	структур
антенн	кабел	оруж	твердотельн
аппарат	картограф	плазм	текстил
атом	квантов	пластик	телевиз
аэро	керам	пластическ	теплообм
баллисти	комплекс	пневмо	техн
безопасн	компрессор	полиграф	топлив
бытов	компьютер	полупровод	трактор
вакуумн	кондицион	преобразов	транспорт
видео	корабл	прибор	турб
водн	косм	программн	удобрен
воен	крио	проектир	установк
вычислит	лазер	производ	химич
газов	локац	промышл	холодил
геодез	магнит	радио	хроматограф
гидравл	маркшейд	ракет	целлюлознобум
гидро	материал	реактор	шахтн
гиро	машин	регулир	электри
двигател	медицин	ремонт	электроник
железнодор	мелио	рентген	электронн
жизнеобеспеч	металлург	робот	электротех
звукотех	метролог	свар	энергет
излуч	механ	сверхпровод	ядерн



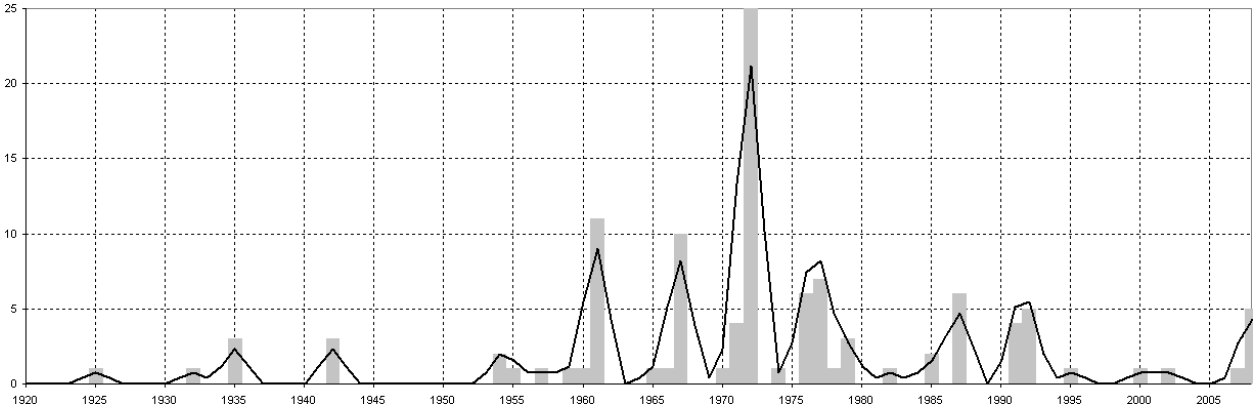
Фиг. 24. Общие траектории частот всех выбранных слов, нормированные к единице (отдельно обозначены: 1 – «компьютер»; 2 – «микропроцессор»; 3 – «жизнеобеспеч»; 4 – «информатик»; 5 – «видео»; 6 – «пневмо»)

Для отобранных слов были рассчитаны индивидуальные частоты, суммарные частоты, и гистограмма годов, на которые приходятся максимумы частот. Результаты приведены на фиг. 24, 25, 26.

Можно видеть, что в общем частота достигает своего максимума в период 1971–1975, а затем начинает спадать. Это видно как по суммарной частоте отобранных технических терминов (фиг. 25), так и по некоторым отдельным темам (фиг. 4,6,19,20), что дает основания предположить либо систематический перекос в исходном наборе данных, либо существование некоторых происходивших в это время явлений, которые изменили место техники в общественном сознании.



Фиг. 25. Суммарная частота всех выбранных для эксперимента слов



Фиг. 26. Гистограмма максимумов частот по годам и скользящее среднее с окном в 3 года

Таким образом, можно сказать, что по мере развития информационной техники, прежде невозможные электронные массивы данных [2, 4] становятся доступными, расширяя возможности квантитативного анализа самых разнообразных явлений и тенденций культурных процессов. В истории техники, как и в других областях науки, исследования, базирующиеся на сборе данных, могут теперь быть дополнены данными за прошедшие периоды време-

ни, а соответствующим образом разработанные и адаптированные методы могут дать возможность более точно наблюдать характер культурных процессов по величине их следа в ноосфере.

Работающие в областях истории науки и техники, социологии, истории и других наук специалисты получили дополнительные возможности для анализа процессов технического прогресса и развития общества.

Особенно ценным является то, что упомянутые и подобные им массивы данных постоянно дополняются и становятся свободно доступными. Ими может воспользоваться любой исследователь как с использованием готового инструментария, так и с помощью специализированных статистических методов и алгоритмов Data Mining применительно к собственным задачам.

Литература и источники

1. Quantitative Analysis of Culture Using Millions of Digitized Books / Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, Erez Lieberman // Science. – 2011. – Т. 331, № 6014. – С. 176-182.
2. Google Books Ngram Viewer : База данных [Электронный ресурс] / Доступ: http://books.google.com/ngrams/datasets
3. Google Books Ngram Viewer [Электронный ресурс] / Доступ: http://books.google.com/ngrams
4. Национальный корпус русского языка [Электронный ресурс] / Доступ: http://ruscorpora.ru/
5. Национальный корпус русского языка. Поиск в корпусе. Распределение по годам. [Электронный ресурс] / Доступ: http://ruscorpora.ru/ngram.html
6. Лексикографические мемуары: о времени, стихах и техническом прогрессе / Максим Кронгауз // Дружба Народов. – 2011. – № 2
7. Culturomics [Электронный ресурс] / Доступ: http://www.culturomics.org/
8. Citation Statistics. / Robert Adler, John Ewing, Peter Taylor // IMU-ICIAM-IMS joint Committee on Quantitative Assessment of Research Report. – 12 июня 2008. – 26 с.
9. Eduard Jan Dijksterhuis. The mechanization of the world picture / Eduard Jan Dijksterhuis. – London: Oxford University Press, 1961. – 539 с.
10. Джеф Раскин. Интерфейс: новые направления в проектировании компьютерных систем / Джеф Раскин. – М.: Символ-плюс, 2004. – 272 с.
11. КПСС в резолюциях и решениях съездов, конференций и пленумов ЦК. 8 изд. – М., 1970. – Т. 4. – С. 338.
12. О советском патриотизме в науке / Зворыкин А. А. // Большевик. – 1948. – № 22. – С. 23-42.