

ТЭХНАЛОГІІ ЗАХАВАННЯ КУЛЬТУРНАЙ СПАДЧЫНЫ Ў ІНФАРМАЦЫЙНАЙ ПРАСТОРЫ

П. В. Гляков,

*канд. фізіка-матэматычных, доц.,
зав. каф. інфармацыйных тэхналогій
в культуры БГУКИ*

КУЛЬТУРОМИКА – НОВЫЙ НАУЧНЫ ПІДХОД В ІССЛЕДОВАНИИ КУЛЬТУРНЫХ ФЕНОМЕНОВ

Ученые Гарвардского университета совместно с сотрудниками энциклопедии «Британика» и проекта «Гугл.книги» предложили новую область знаний, основанную на данных статистического анализа оцифрованных книг. Ее назвали культуромикой. Культуромика занимается изучением с помощью математических методов истории языка и словесности, а также проявлений того или иного культурного феномена в истории [1].

Компьютерный статистический анализ позволяет проводить быстрый поиск слов и символов по базе данных электронных книг, благодаря чему можно быстро отслеживать проявления того или иного культурного феномена в ходе истории, который нашел отражение в книгах.

Например, такой анализ позволил ученым выяснить, что пик популярности актеров приходится на 30 лет – именно в этом возрасте они наиболее часто упоминаются в печати. Писатели восходят на вершину славы на 10 лет позже, имена ученых же часто упоминают лишь в исключительных случаях и, главным образом, уже по достижению ими преклонного возраста. При этом физикам и биологам отводится куда больше внимания, чем математикам [2].

С помощью этого статистического подхода, как показали ученые, можно проследить влияние внешних факторов на проявление в культуре таких понятий, как феминизм, бог, эволю-

ция, здоровый образ жизни и других, в той или иной степени интересующих каждого члена общества.

Культуромика расширяет границы возможного до настоящего времени количественного анализа большого количества явлений в социальной сфере и гуманитарных науках. «Изучение культурного отпечатка конкретного понятия в истории интересно многим людям, однако мы надеемся, что наш подход будет особенно активно использоваться учеными, изучающими гуманитарные науки и социальные дисциплины», по словам одного из разработчиков культуромики, профессора Гарвардского университета Эреза Либермана Эйдена [1].

Построенная база данных из оцифрованных текстов, содержит метаданные около 4 % от всех книг, когда-либо напечатанных. Ее анализ позволяет исследовать культурные тенденции количественно, обеспечить более глубокое представление о таких различных областях, как лексикография, эволюция грамматики, коллективная память, стремление к славе, цензура и историческая эпидемиология [3].

База данных оцифрованных книг разбита на корпуса в соответствии со следующими языковыми группами: английской, китайской, французской, русской, немецкой, итальянской и иврит. Английская языковая группа составляет 72 % от всех книг и содержит 5 подгрупп:

- американский вариант английского языка (книги на английском языке, опубликованные в США);
- британский английский язык (книги на английском языке, которые были опубликованы в Великобритании);
- английский язык (книги на английском языке, опубликованные в любой стране);
- английскую фантастику (книги на английском языке, которые библиотека или издатель идентифицировали как фантастика);

– один английский миллион. Эта подгруппа включает 1 млн книг на английском языке, изданных с 1500 по 2008 гг. Для каждого года выбрано не более 6000 книг. Выбор книг осуществлялся так, что книги, выпущенные в ранний период, представлены полностью, а изданные позже, отобраны случайным образом с учетом того, что, например, книг, изданных в 2000 г., будет представлено больше, чем изданных в 1980 г.

Каждый корпус оцифрованных книг, за исключением последней подгруппы английского языка, представлен двумя

версиями: 2009 и 2012 гг. Версия 2012 г. содержит больше книг, чем версия 2009 г.; в ней улучшена структура метаданных и при оцифровке книг использован улучшенный метод оптического распознавания символов.

Для проведения научных экспериментов с базой данных оцифрованных книг Лаборатория Гугл разработала онлайн-инструмент Ngram Viewer [4]. Эта программа отображает график, показывающий, как введенные фразы (N-граммы) распределяются в выбранном корпусе книг по годам. Фразы для построения графика надо вводить на языке, соответствующем выбранному корпусу. При наборе фраз следует различать строчные и прописные буквы.

В качестве примера выполним исследование, показывающее историю развития культуры и науки в странах, в которых издание книг осуществляется на английском и русском языках. Из графика, построенного программой Ngram Viewer, мы видим (см. рис. 1), что в странах, в которых книги издаются на русском языке, науке всегда уделялось больше внимания, чем культуре. Исключение, пожалуй, составляют 20-е гг. 20 в.

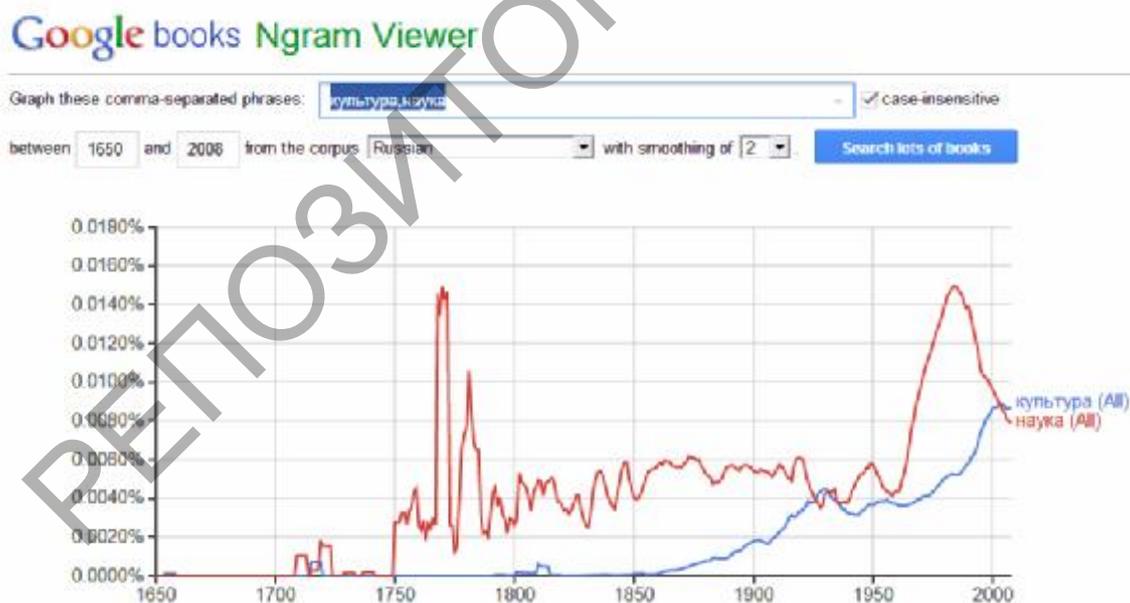


Рис. 1. График частот употребления слов в русскоязычном корпусе

Стремительное падение науки началось в 1984 г. и продолжается, к сожалению, до сих пор. Культура же, начиная с 60-х гг. прошлого столетия, непрерывно развивалась вплоть до 2003 г. В 2003 г. она превзошла науку и остановилась в развитии.

График, приведенный на рис. 2 для англоязычных стран, также показывает, что в них наука всегда доминировала над культурой и лишь в конце 20 столетия культура обогнала науку. С 1998 г. началось резкое падение в развитии культуры. В этот же период стало замедляться и развитие науки.

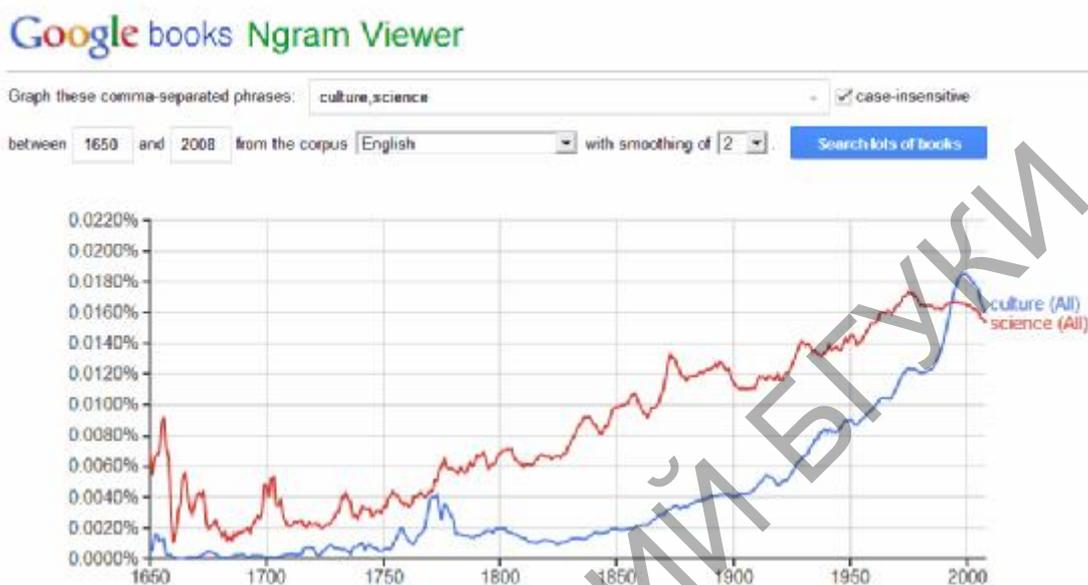


Рис. 2. График частот употребления слов в англоязычном корпусе

Сравнивая между собой графики, приведенные на рис. 1 и рис. 2, можно отметить, что в русскоязычных странах развитие науки происходило резкими скачками. Такие скачки в развитии науки для англоязычных стран были характерны лишь для периода с середины 17 столетия до начала 18 столетия.

Дальнейшее развитие культуромика получила в проекте, который возглавил ученый из университета Иллинойса Калев Лиитару [5]. Проведенные им эксперименты позволили успешно спрогнозировать в ретроспективе революции в арабских странах. Свои исследования ученый выполнял на суперкомпьютере Nautilus, принадлежащем Национальному институту вычислительных систем при университете Теннесси. Машина проанализировала сто миллионов новостных статей, вышедших в свет за период с 1979 по 2011 гг. Лиитару продемонстрировал, что культуромика действительно может делать прогнозы некоторых культурных, политических и экономических событий. В своих исследованиях он использует средства семантического анализа текстов и географического позиционирования упомянутых в них сущностей, а также визуализации результатов.

1. *Michel, J. B.* Quantitative Analysis of Culture Using Millions of Digitized Books. Science / J. B. Michel [and other]. – 2010. – [Электронный ресурс]. – Режим доступа: <https://books.google.com/ngrams/info>. – Дата доступа: 25.05.2013.

2. *Горный, Е.* Проблемы сохранения культурного наследия в эпоху цифрового текста / Е. Горный. – 2012. – [Электронный ресурс]: Режим доступа: www.netslova.ru/gorny/digttext.html?2012. – Дата доступа: 24.05.2013.

3. Оцифровка книг позволит по-новому изучать историю и культуру [Электронный ресурс]. – Режим доступа: <http://ria.ru/science/20101217/-309883432.html#13702723185043>. – Дата доступа: 10.05.2013.

4. Google Books Ngram Viewer: What does the Ngram Viewer do? [Электронный ресурс]. – Режим доступа: <https://books.google.com/ngrams/info>. – Дата доступа: 20.05.2013.

5. *Leetaru, K. H.* Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space / К. Н. Leetaru. – [Электронный ресурс]. – Режим доступа: <https://books.google.com/ngrams/info>. – Дата доступа: 18.10.2013.

А. Г. Зезюля,

доц. каф. информационных технологий в культуре БГУКИ

ОБЕСПЕЧЕНИЕ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ В УЧРЕЖДЕНИЯХ КУЛЬТУРЫ И ИСКУССТВ

Необходимость защиты информации осознавалась на протяжении всей истории развития человечества.

Особенно актуальной эта проблема стала в последние годы, которые характеризуются массовым использованием автоматизированных информационных систем в режиме интенсивного сетевого взаимодействия. Достижения современных компьютерных технологий в сфере обработки информации (сбора, хранения, поиска, преобразования, представления) предоставили новые возможности и существенные удобства. Результатом этого прогресса явилось то, что значительные объемы информации переведены и переводятся с традиционных носителей в виртуальную среду и процесс имеет ярко выраженную тенденцию. Значительная часть информации уже непосредственно функционирует только в виртуальном пространстве, а