

Компьютерные ресурсы для гуманитарных исследований

Никипорец-Такигава Г.

Введение

Десять лет назад сбор картотеки для любого гуманитарного исследования, посвящённого России, начинался с трудоёмкого поиска и выписывания сотен примеров из книг и газет. Русский язык был одним из немногих крупнейших языков, не имеющих собственного корпуса¹, находящегося в открытом доступе. Современные достижения компьютерной лингвистики в сочетании с возможностями Интернета позволили гуманитариям существенно оптимизировать процесс сбора и анализа данных. Однако потребовали овладения суммой специальных знаний, что было главной причиной разделения учёных на два лагеря – противников и сторонников использования компьютерных технологий в гуманитарном исследовании.

Компьютерные ресурсы необходимы не всем учёным. Сразу следует исключить тех, кто не занимается современной Россией. А также тех, кто считает, что гипотезу можно строить на нескольких случайно зафиксированных примерах или обходиться без примеров, прозревая мощью собственного интеллекта и силой интуиции процессы, происходящие в современной российской экономике, политике, истории, языке, культуре и обществе. Учёный, не попадающий в две названные категории, стремится изучить все имеющиеся на сегодня компьютерные возможности сбора и анализа данных и найти для себя те, которые наиболее приспособлены для решения интересующих его задач. Для такого учёного может оказаться небесполезной данная статья, цель которой — обзор применяемых в гуманитарных

Г. Никипорец-Такигава

исследованиях компьютерных ресурсов и их сравнительное описание на примере одного конкретного исследования.

1. Корпусы русского языка. Состав, объём и исследовательские задачи, для решения которых они предназначены

В настоящее время в распоряжении учёных имеется несколько корпусов русского языка. В Японии широко известен первый из западных — Тюбингенский корпус русского языка (ТК). Скандинавы — пионеры и энтузиасты корпусной лингвистики, помимо Тюбингенского (в основе которого Уппсальский корпус), работают и над вторым корпусом — Хельсинкским аннотированным корпусом русского языка (Ханко). Первый российский корпус — «Интегрум» — появился раньше западных, но сразу был создан как коммерческий продукт и не выставлен в открытый доступ. Несколько газет из корпуса «Интегрума» было передано в МГУ, где на их основе был создан Корпус Газетных Текстов (КГТ). В 2001 началась работа над Национальным корпусом русского языка (НКРЯ). Среди других российских корпусов следует назвать Полистилевой корпус русского языка (ПКРЯ), первый фрагмент которого вместе со «Словарём 500 самых употребительных слов русского языка» появился в открытом доступе в марте 2006 года.

1.1. ТК — Тюбингенский корпус² был задуман в 1999 году как база примеров для проекта по изучению форм обращений и вежливости в русском языке («Linguistische Datenstrukturen: Theoretische und empirische Grundlagen der Grammatikforschung»). В настоящее время в ТК около 25 миллионов морфологически размеченных словоформ. Материал корпуса разнороден (художественные тексты, СМИ 1985-1988 годов, журнал «Огонёк» 1996-2002), но ограничен тематикой исследования: корпус составлен, главным образом, из текстов, в которых могут быть использованы формы обращения и вежливости. Так, например, средства массовой информации представлены только текстами интервью из некоторых газет и журналов. Второе серьёзное ограничение — количественное: как только закончился проект, корпус пере-

стал пополняться и развиваться.

1.2. **Ханко** — Хельсинкский аннотированный корпус русского языка³ разрабатывается с 2001 года для проекта «Функциональный синтаксис русского языка», предназначен для изучения грамматики русского языка. Задачей создателей корпуса является предоставление максимального количества грамматической информации. В соответствии с задачей, корпус размечают (выполнена подробная синтаксическая и морфологическая разметка, семантическая разметка выполняется в настоящее время), но не пополняют. В настоящее время это самый маленький из названных корпусов: в нём около 100 тысяч словоупотреблений. В качестве источника текстового материала были выбраны несколько номеров журнала «Итоги» за январь 2001 года.

1.3. **НКРЯ** — Национальный корпус русского языка⁴ разрабатывается с 2001 года. Это главное достижение российской корпусной лингвистики, постоянно совершенствующийся и пополняющийся корпус, практически не имеющий недостатков и конкурентов в будущем. Сегодня серьёзным недостатком является его объём. Несмотря на то, что корпус насчитывает 120 миллионов словоупотреблений, превосходит объёмы западных корпусов, он пока не отражает русский язык в совокупности его проявлений, поэтому не вполне соответствует своему названию. Корпус размечен и идеально подходит для поиска грамматических единиц.

1.4. **«Интегрум»**⁵ — первый корпус русского языка, разработка которого началась в конце 70-ых годов. Сегодня в базах «Интегрума» хранится около 350 миллионов русскоязычных документов, что позволяет считать его самым полным корпусом русских текстов, приближающимся по своим количественным показателям к исчерпывающему. Это единственный корпус, который предназначен не только для филологических, но и для любых других гуманитарных исследований.

Г. Никипорец-Такигава

2. Интернет-ресурсы

Помимо корпусов русского языка ещё одним важным компьютерным ресурсом для сбора и анализа данных является Рунет: Yandex (www.yandex.ru), Rambler (www.rambler.ru), Google (www.google.com) и т.д.

Когда гуманитарии познакомились с Интернетом, он представлялся возможностью сделать гуманитарное исследование точным, доказательным, репрезентативным, отказавшись от абстрактной методологии, которая переводит научную работу в жанр субъективной эссеистики. Интернет казался и способом объять необъятное — получить безграничное количество материала, быстро найти любую единицу информации. Однако возможности Интернета скоро оказались иллюзией, как только гуманитарии овладели Интернетом, радость от общения с новым медиумом сменилась осознанием недостатков:

2.1. Первая проблема: Интернет не гарантирует качества и достоверности информации. Это главный аргумент противников использования компьютерных методов в гуманитарном исследовании. Аргумент справедлив — «ручной» сбор картотеки не ставит проблему неаутентичности текста, тогда как электронная копия, найденная в Интернете, нуждается в перепроверке, так как часто серьёзно расходится с оригиналом.

2.2. Вторая проблема: невозможность ограничить пространство поиска. Интернет-поисковые машины (далее ИПМ) вынуждают искать во всём пространстве «Интернета», выбирая необходимые данные из многостраничной выдачи.

2.3. Третья проблема: на страницах выдачи много повторов.

2.4. Четвёртая проблема: отсутствие в Интернете полнотекстовых **архивов** средств массовой информации, затрудняющее сравнительный и исторический анализ.

2.5. Пятая проблема: ИПМ не имеют сложного языка запросов для точного поиска. Исследователя, который хочет на основании примеров, найденных в Интернете при помощи ИПМ, построить систему доказательств, обычно сначала подкупает изве-

стная лёгкость в общении. Но очень скоро перестаёт удовлетворять процент «шума» и невозможность снятия некоторых принципиальных затруднений, например, омонимии. Так, поставив задачу поиска стереотипов, которые наиболее активно разрабатываются в русской языковой картине мира в связи с японским характером, японцами, Японией во всём пространстве Рунета, исследователь сталкивается с тем, что ИПМ не могут исключить из выдачи контексты, в которых японец оказывается машинной, бытовой техникой, электроникой и так далее.

По сравнению с Интернетом корпусы предлагают определённые средства поиска и гарантируют некий баланс текстов, который репрезентативен для современного русского языка. В Интернете баланса не может быть. Документы попадают в Интернет бесконтрольно, хранятся там бессистемно, поэтому исследователь не застрахован от ситуации, в которой сотни найденных ИПМ примеров на самом деле будут найдены в чатах, то есть представлять лишь одну сферу использования языка. Подчеркнём, тем не менее, что все корпусы, кроме «Интегрума», по сравнению с Интернетом имеют существенный недостаток: в них заложены научно целесообразные и тщательно проработанные принципы, но они реализованы пока на малом количестве текстов.

3. «Интегрум» — объединение возможностей Интернета и корпусной лингвистики

Единственный корпус русского языка, который сочетает достоинства Интернета с достижениями корпусной лингвистики — «Интегрум». Так как «Интегрум» первоначально создавался не как корпус, а как информационно-справочная база данных, он стремился сохранить в электронном виде любой русскоязычный документ, следовательно, объединил в себе такой же огромный материал, как и Рунет. С другой стороны, «Интегрум» сразу поставил перед собой задачу преодоления недостатков Интернета:

3.1 Решил первую проблему: проблему качества информации, так как искал не в Интернете, а в собственных базах дан-

Г. Никипорец-Такигава

ных, составленных из специально приобретённых у владельцев, а поэтому наиболее надёжных источников.

3.2 «Интегрум» решил вторую проблему выбора релевантных для анализа источников. Поиск и исследования при помощи «Интегрума» можно осуществлять в ограниченной по желанию исследователя на этапе запроса области: во всех СМИ, в группах СМИ⁶, в одном СМИ; во всей «Библиотеке всемирной литературы», в группе произведений разных авторов, во всех произведениях одного автора, в одном произведении и так далее. Поиск можно ограничивать заданным временным интервалом.

3.3 В выдаче «Интегрума» почти нет повторов документов.

3.4 Как и Интернет, «Интегрум» дал гуманитариям возможность исследовать русскоязычные массивы информации, состоящие из сотен миллионов документов⁷, однако, в отличие от Интернета, **содержал полные архивы тысяч СМИ**, большинство из которых начиналось с середины 90-х годов XX века. Это позволило собирать данные на большом качественном материале и анализировать период продолжительностью десять-пятнадцать лет.

3.5 И, наконец, «Интегрум» придумал многофункциональную информационно-поисковую систему «Артефакт», располагающую разветвлённым языком запросов со сложным синтаксисом (операторами «или» для изучения группы, «не» для снятия омонимии и других ограничений запроса, «*» и «!п» для того, чтобы система искала все слова, начинающиеся с определённой комбинации букв или полную парадигму слова, «/п» и «:» с последующей цифрой для ограничения поиска указанной после буквы «п» группой предложений или указанным после двоеточия расстоянием до следующего слова), осуществляющую поиск с опечатками, поиск дат и людей, сортировку результатов по разным критериям и т.п. Например, запрос, который может обработать «Интегрум» для решения задачи выяснения стереотипов восприятия японцев, выглядит следующим образом: *(настоящ* или типичн* или среднестатистич* или средн* или*

Компьютерные ресурсы для гуманитарных исследований

обычн или обыкновенн*):0 (япон* или самурай или «житель японских островов» или «житель Японии» или «житель страны восходящего солнца») не машина не автомобиль не внедорожник не вездеход не джип не видеомэгнитофон не мэгнитофон не видео не телевизор не камера не видеокамера не фотоаппарат⁸.*

Таким образом, «Интегрум» организовал поиск в не уступающей Интернету по объёму базе данных с учётом всех достижений корпусной лингвистики. Эти достоинства «Интегрума» быстро привлекли к нему учёных, прежде всего, политологов и социологов, по запросу которых «Интегрум» за секунды доставлял на экран компьютера любой российский документ и всю палитру российских СМИ, среди которых были напечатанные в такой российской провинции, куда без «Интегрума» было бы не добраться. Не в последнюю очередь к «Интегруму» обратились и филологи. В то время филологическая мысль была уже достаточно компьютеризирована, поиск примеров в Интернете осознавался вчерашним днём, его несовершенства были очевидны. Филологи начинали активно пользоваться корпусами русского языка, которые идеально подходили для решения многих проблем, кроме тех, для которых важен охват материала. Для таких проблем потенциала корпусной лингвистики не хватало – НКРЯ и другие корпуса были неизбежно количественно ограничены. В этом отношении стремление «Интегрума» сохранить любой русскоязычный документ сулило новые научные перспективы.

Первыми «Интегрумом» заинтересовались создатели корпуса Ханко. Финские коллеги стали успешно применять «Интегрум» для исследования грамматических феноменов, им принадлежит много интересных разработок.

У японских учёных тоже вскоре появилась возможность стать пионерами в освоении новой базы данных — в январе 2002 года «Интегрум» был представлен на Winter International Symposium at Slavic Research Center. С 2002 года автор данной статьи использует этот корпус в своих научных работах, посвящённых нормативному аспекту языка, некодифицированным подсистемам языка, агрессии в языке и обществе, новым лексическим и семантическим процессам, происходящим в современном русском языке и, в частности, в языке

Г. Никипорец-Такигава

средств массовой информации. В качестве материала для дальнейшего описания и сравнения компьютерных ресурсов будет привлекаться материал из одной из последних работ автора⁹.

4. Сравнительный анализ возможностей применения разных современных компьютерных ресурсов и баз данных на примере одного конкретного исследования

Автором было отмечено появление в русском языке таких словоупотреблений: «*К экстравагантным платьям выбираем духи **агрессивные** и тяжёлые*»¹⁰, «*Не волнуйтесь, вы найдёте эти витамины в любой аптеке — они очень **агрессивно** представлены*»¹¹, «***Амбициозность** великолепна!*»¹². «*Когда я вдруг вижу, что человек читает книгу Довлатова, меня как-то так **приятно шокирует***»¹³.

Такое употребление в русском языке до конца XX было невозможно: вне употребления в качестве термина (*шок, шокирующая терапия* в медицине, *агрессивная жидкость, среда* в химии и т.д.) всем этим словам была присуща вполне определённая отрицательная оценочность.

Прежде всего, следовало выяснить, являются ли подобные примеры единичными или частотными настолько, что позволяют говорить о семантическом сдвиге в русском языке. Для этого было необходимо собрать картотеку примеров нового употребления исследуемых слов. Очевидно, что задача вычленения строго определённой группы примеров перед Интернет-ресурсами поставлена быть не может. Они обрабатывают только элементарные запросы, поэтому могут выдать огромное количество любых примеров употребления слова, среди которых многие будут дублировать друг друга, многие будут нерелевантны, не удовлетворяя поставленным условиям (в нашем случае, отражая классическое употребление слов в негативном значении). Среди исследуемых единиц словосочетание «приятно шокировать» само по себе является мини-контекстом употребления слова *шокировать* в положительном значении, поэтому его частотность можно попробовать выяснить с помощью ИПМ, однако ИПМ не работают с дериватами и ищут строго указанные в кавычках словоформы, кроме

этого, не отсекают повторы документов. По этим причинам, количество примеров, найденных ИПМ по запросу «приятно шокировать» (ИПМ не искали контексты «приятно шокирован», «приятно шокировали» и т.д.) – 488 (Yandex) и 591 (Google), не может быть основой серьёзного научного анализа.

Почти все корпуса русского языка для решения такой задачи тоже непригодны, так как содержат малое количество текстов. В частности, в НКРЯ нет ни одного примера словосочетания «приятно шокировать», идиомы «шок – это по-нашему». «Интегрум» же нашёл, соответственно, 389 примеров и 1383.

Помимо репрезентативного количества примеров для обоснования гипотезы о возникновении нового явления в языке следует статистически исследовать явление на временном отрезке, большем, чем пять лет. Известно, что столько существуют проявления языковой моды, которые могут возникать в языке случайно под влиянием экстралингвистических факторов и бесследно исчезать. Серьёзная проблема, стоящая перед современной лингвистикой — способы отделения проявлений языковой моды от устойчивых тенденций в языке. «Интегрум» даёт исследователю надёжный метод различения. Его сервисы «Сравнительной и относительной упоминаемости (статистики)» и «Частотный словарь русского языка» позволяют фиксировать первое появление слова или любой другой единицы (любого другого объекта исследования), проследить его существование в языке на любом заданном отрезке времени, изучая его частотность при помощи вычисления и сравнения:

- о абсолютного количества документов, в которых объект встречается в каждой заданной временной точке;
- о процентного отношения документов, в которых объект встречается, к общему количеству документов, которые содержатся в «Интегруме» в каждой заданной временной точке;
- о процентного отношения документов, в которых объект встречается в одном контексте, к количеству документов, в которых он встречается в другом контексте или в любом контексте;
- о процентного отношения количества документов, в кото-

Г. Никинорец-Такигава

рых упоминается один объект, к количеству документов, в которых упоминается другой объект;

При помощи сервиса «Сравнительная статистика по абсолютной шкале»¹⁴ (см. гистограммы на рисунках 1 и 2) и точного поиска было выяснено, что первое употребление идиомы «шок — это по-нашему» зафиксировано в 1997 году¹⁵; словосочетания «приятный шок»¹⁶ — в 1994 году; прилагательных положительной семантики *агрессивный* — в 1997 году, *амбициозный* — в 1999 году.

Рис. 1: Частотность идиомы «шок – это по-нашему»:



Рис. 2 Частотность словосочетания «приятно шокировать»:



Компьютерные ресурсы для гуманитарных исследований

Непрерывно восходящая на протяжении десяти лет динамика исследуемых единиц языка свидетельствует о том, что они закрепляются и актуализируются, следовательно, из проявлений языковой моды трансформируются в явления языка.

Ещё один способ убедиться, что объект исследования частотен — проверить его индекс в частотном словаре. На сегодняшний день наиболее известны четыре частотных словаря: один печатный — «Частотный словарь русского языка» под редакцией Л.Н. Засориной и три электронных: «Машинный словарь» С.А. Шарова, «Частотный словарь ПКРЯ», «Частотный словарь «Интегрума»». Последний является самым представительным, так как сделан на материале 9 миллиардов слов, тогда как «Частотный словарь русского языка» Л.Н. Засориной сделан на материале миллиона слов, «Машинный словарь» С.А. Шарова - 40 миллионов слов. Частотный словарь ПКРЯ состоит из 500 слов, в то время как частотный словарь «Интегрума» состоит из 10.000. Частотный словарь «Интегрума» обнаружил, в частности, что прилагательное *агрессивный* чаще употребляется в языке СМИ, чем прилагательные *ужасный*, *эмоциональный*, *синий*, *солнечный* (см. таблицу на рис. 3):

Рис. 3 Частотность прилагательного агрессивный:

Группа СМИ	Частотность	Прилагательные с частотностью, равной частотности «агрессивный»	Прилагательные с частотностью ниже, чем у «агрессивный»
Все СМИ	агрессивный (0,69%)	армейский (0,69%) стартовый (0,69%)	дружный (0,67%) кремлёвский (0,67%) подлинный (0,67%) эмоциональный (0,67%) ужасный (0,67%)
Центральные газеты	агрессивный (1,28%)	бухгалтерский (1,28%) книжный (1,28%) кавказский (1,28%)	забытый (1,26%) значимый (1,27%) плотный (1,27%) синий (1,27%) солнечный (1,27%)

Г. Никинорец-Такигава

Для дальнейшего исследования был необходим детальный контекстный анализ, который требует наличия широкой и глубокой картотеки примеров. «Интегрум» является самым удобным полем для сбора картотеки: благодаря возможностям тонкого запроса картотеку возможно составить только из интересующих примеров (в нашем случае, нового позитивного употребления единиц), благодаря объёму баз «Интегрума» удаётся собрать много примеров. «Интегрум», в отличие от других корпусов, показывающих только предложение, в котором используется искомая единица языка, содержит весь документ, в котором она встречается. Благодаря этому были получены полные контексты употребления слова. Ещё одним достоинством является то, что «Интегрум» позволяет собрать картотеку примеров за 15 лет. ИПМ, которые тоже открывают полностью документ, не справляются с такой задачей.

Контекстный анализ употребления исследуемых единиц в русском языке последних 15 лет обнаружил процесс вторичного заимствования¹⁷ из английского языка новых положительно-оценочных слов *агрессивный, амбиции, шок/шокировать*, в результате которого в русском языке XXI века появились пары омонимов с антонимичными значениями. Два первых слова первоначально возникли в языке бизнеса и спорта, где являлись кальками с современного английского языка: *«Может быть, нашим «командирам производства» стоит не только «плакаться» правительству, но и брать уроки агрессивной, умной, продвинутой манеры маркетинга у своих «заморских» конкурентов?»*¹⁸. Калька английского термина *aggressive marketing*, лишённая каких-либо коннотаций, в русском предложении была наделена положительной оценочно-эмотивной модальностью. Из бизнеса и спорта вновь заимствованные положительно-оценочные слова переместились в другие дискурсы.

Что касается нового положительно-оценочного слова *шок*, его тоже можно считать вторичным заимствованием, однако более ранним и не связанным с языком бизнеса и спорта. В «Библиотеке мировой литературы»¹⁹ «Интегрума» было найдено 17 примеров употребления слова *шок* и дериватов²⁰, среди них два принадлежат Л. Толстому, два Н. Лескову, один Н. Гумилёву, три Г. Газданову,

два М. Зощенко, один А. Толстому. В них слова *шок/шокировать* (причастия от этого глагола) употребляются в том значении, в котором были заимствованы в конце XIX века из французского языка, например:

* Ну, а для молодых, невинных — всегда сначала берут невинных, как Лидочка, — для этих первый **шок** ужасен... (Толстой Л. Воскресение)

* — Конечно-с, на что же мне это было? Я не веду моей родословной ни от каких славных гусей: я не граф и не князь, чтобы быть **шокированным** поведением жены. (Лесков Н. На ножах).

* Сурово перешептывались камергеры, молодые камер-юнкера принимали утрированно-солидный вид, и **шокированные** дамы с негодующим удивлением поднимали тонко вырисованные брови. (Гумилёв Н. Последний придворный поэт).

* Он обратился ко мне фамильярно и непринужденно, так, точно мы были знакомы много лет; но, как всегда, во всем, что он говорил или делал, в этом не было ничего **шокирующего**. (Газданов Г. Призрак Александра Вольфа).

* Лида кричала, что ему, старику, следует думать о своей науке, а не о паршивых девицах, что она этим буквально **шокирована**, что она поражена низменными уклонами своего отца. (Зощенко М. Возвращенная молодость).

* Тут еще привезли одного командира, тяжелый случай **шока**... **Шок**, я вам скажу, дает иногда любопытнейшую картину... Доктор оборвал на полуслове, потому что, хотя и в слабой степени, сам получил неожиданный **шок**. (Толстой А. Н. Хождение по мукам).

Шесть примеров принадлежат В. Набокову:

* В Америке мужчины дружатся легче и веселее, чем здесь, — я там привык не стесняться, простите, если я вас **шокирую**... (Камера обскура).

* Позже он объяснил эту вспышку тем, что принял — по

Г. Никипорец-Такигава

вине читальных очков — долгожданного друга за наглеца-торговца, но должен сказать, что я был **шокирован**, крайне **шокирован**, что и позволило мне уже тогда прочесть отвратительный смысл всего, что за этим последовало. (Бледное пламя).

* Тетя Мод была отнюдь не старой девой, а экстравагантный и сардонический склад ее ума должен был порою **шокировать** жантильных нью-вайских дам. (Бледное пламя).

* Это чувство обладало резкостью ретроспективных деталей, составляющей, как уверяют, драматическое достояние утопающих, — особенно на прежнем Русском флоте — феномен удушья, которое бывалый психоаналитик, забыл его имя, объяснял подсознательным возрождением **шока** крещения, вызывающим как бы взрыв воспоминаний, промежуточных между первым погружением и последним. (Пнин).

* Правда, во всем произведении нельзя найти ни одного непристойного выражения; скажу больше: здоровяк-филистер, приученный современной условностью принимать безо всякой брезгливости целую россыпь заборных словечек в самом банальном американском или английском романе, будет весьма **шокирован** отсутствием оных в /"Лолите"/. (Лолита).

* Великое произведение искусства всегда оригинально; оно по самой своей сущности должно потрясать и изумлять, т. е. "**шокировать**". (Лолита).

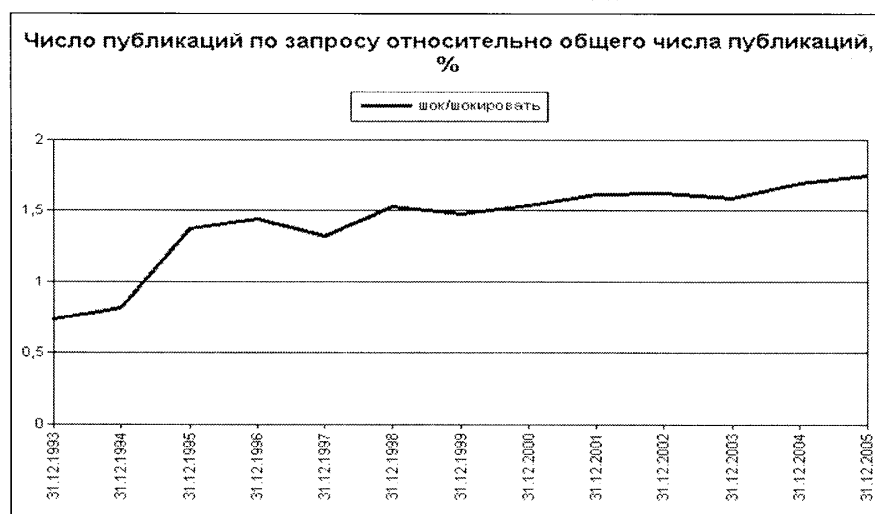
Два последних примера свидетельствуют о вхождении в русский язык вторичного заимствования *шокировать* в нейтрально-положительном значении. На английский источник заимствования указывает авторство и история создания текста. «Приятно шокировать» в современном русском языке также возникло под влиянием английского языка: было калькой английского pleasantly shock(ed). Появление идиомы «шок — это по-нашему» можно считать следствием иностранного влияния, так как идиома и рекламный ролик, в котором она появилась, а также рекламируемый с их помощью шоколадный «Шок» — продукция швейцарского концерна Nestle. Однако идея использования омонимии могла быть отечественной или француз-

Компьютерные ресурсы для гуманитарных исследований

ской, но не английской, в котором такая омонимия невозможна.

В целом, влияние английского языка следует рассматривать лишь как причину возникновения исследуемых слов в положительном значении в русском языке. Причиной их закрепления и дальнейшей актуализации, и, шире, сдвига в семантике русского языка, следует считать воздействие российских СМИ, преимущественно, телевидения на языковую компетенцию среднестатистического носителя русского языка. Подтверждение было найдено в архивах «Интегрума». Статистические измерения исследуемых лексем показали, что их частотность за пятнадцать лет неуклонно повышалась и достигла в настоящее время значительных показателей. Например, слово *агрессивный* употребляется в каждом четвёртом тексте СМИ, слово *шок* во всех значениях употребляется в двух процентах всех текстов СМИ²¹ (см. график на рис 4):

Рис.4 Частотность слова *шок*:



Эти показатели - причина и следствие произошедших в истории этих слов изменений. Слова попали в язык СМИ как кальки с английского, но были немедленно подхвачены и внедрены в активный словарь СМИ. Их многократное повторение в СМИ превратило их в штамп, наделив свойственной штампу вытесняющей способностью. Последнее обстоятельство привело к закреплению нового положи-

Г. Никипорец-Такигава

тельного значения в языке. Примеры свидетельствуют, что в современном русском языке шокировать начинает всё, что может *приятно удивлять, поражать, потрясать, восхищать, радовать*: «*И пусть его упрекают в чрезмерной коммерциализации линии, но, тем не менее, когда он появился на рынке, это был **шок**, эпатаж, это было круто*»²². «*Катя Кобзон **шокирует** и покоряет*»²³. Слово *шок* употребляется в противоположных значениях почти в соседних предложениях: «*Однажды я закурила при своей дочери, и для неё это был **шок**, она очень переживала... Её задатки мы замечали ещё с раннего детства и поначалу не обращали внимания. Но наши знакомые бывали в **шоке** от её дара и стремления играть*»²⁴. Словом *агрессивный* можно определить всё, что в XX веке определяли словами *активный, смелый, энергичный, решительный, яркий, сочный, красочный, эмоциональный, сильный, стремительный, быстрый, ревнивый, экспрессивный и т.д.* Агрессивными могут быть: *талант, красота, эстетика, стиль и одежда, внешность, макияж, причёски, русский балет, цветы, женщины, мужчины, фильмы, машины, манера поведения, характер*²⁵.

Таким образом, контекстный анализ в сочетании с разными видами статистического анализа позволил сделать аргументированный и репрезентативный вывод о том, что в русском языке возникли пары омонимов с антонимичными значениями: *агрессивный, амбиции, шок/шокировать*, а также выводы и предположения о причинах появления и актуализации вторичных заимствований.

Заключение

Обобщить изложенное выше и завершить предпринятый в данной статье обзор современных компьютерных ресурсов необходимо двумя замечаниями. Следует подчеркнуть, что их использование целесообразно при решении многих методических и научных задач. Однако важно помнить, что каждый из ресурсов в современном виде предназначен для решения разных задач.

Интернет-ресурсы полезны изучающим русский язык, особенно студентам, которые привыкли к жизни в киберпространстве и располагают временем для длительных поисков не столько точной, сколь-

ко любопытной информации.

Использование корпусных ресурсов перспективно для разных научных направлений и гуманитарных дисциплин. Для лингвиста это изучение лексики современного русского языка в динамике развития: истории слов конца XX начала XXI веков; лексических изменений, изменений лексической семантики (семантических преобразований, сдвигов и колебаний), инноваций (их первого вхождения в русский язык и этимологии); влияния экстралингвистических факторов на лексику и лексическую семантику; проявлений языковой моды; нормативного аспекта языка. Для таких задач среди корпусов особенно перспективным представляется использование «Интегрума», так как это единственный почти исчерпывающий корпус. Статистический анализ любого слова при помощи инструментов «Интегрума» делает аргументированными рассуждения о его частотности, которые, по существу, остаются абстракцией, если в качестве материала для доказательства вместо всего материала языка служат несколько ярких примеров или безграничное и бесконтрольное пространство Интернета. НКРЯ, Ханко, ПКРЯ могут рассматриваться как полный электронный источник грамматической информации.

При помощи «Интегрума» решаются сложные лексикографические задачи: фиксация вхождения слова в русский язык и умение отличать явления языковой моды от явлений языка. Статистический сервис «Интегрума» вместе с его же Частотным словарём, «Машинным словарём» С.А. Шарова, «Частотным словарём ПКРЯ» позволяют за секунды составить словник реального современного русского языка, хотя в традициях даже развитой западной лексикографии на это уходили годы. Пока Частотный словарь «Интегрума» — лучшая возможность для лексикографов, так как он анализирует самое большое количество слов (9 миллиардов) и имеет важнейшую функцию обновления, позволяющую немедленно отслеживать все изменения, происходящие в языке.

Безусловно, все без исключения корпуса интересны лингвистам. Литературоведам полезны те из них, которые полно и разнообразно представляют художественную литературу. На сегодняшний день это НКРЯ, с 2007 года таковым станет корпус «Интегрума».

Г. Никипорец-Такигава

Что касается нефилологических исследований, то для них самым перспективным и полезным оказывается «Интегрум». Отметим, что «Интегрум» – не только корпус, его базы могут использоваться для сугубо практических информационно-справочных потребностей как огромная электронная библиотека, состоящая из сотен миллионов документов о России, в которой можно быстро и эффективно найти интересующую информацию, быстро отыскать факты, цифры, цитаты, примеры. Так как данные в этой библиотеке непрерывно пополняются и обновляются, это позволяет изучать российские процессы в динамике.

Таким образом, гуманитарии необходимы обширные знания о разных компьютерных ресурсах и умение применять наилучшим образом приспособленный ресурс к конкретному исследованию. Сложная, но увлекательная задача, учитывая, что многие компьютерные ресурсы быстро развиваются и совершенствуются, постоянно расширяя список потенциальных возможностей применения.

Цитируемая литература

Integrum: точные методы и гуманитарные науки / Ред.–сост. Г. Никипорец-Такигава. Вступ. сл. Вяч. Вс. Иванов. Предисл. А.Я. Шайкевич. — М., 2006, 430 с.

Крысин Л.П. Вторичное заимствование и его описание в толковом словаре. // Проблемы русской лексикографии. Тезисы докладов международной конференции Шестые Шмелёвские чтения. 24-26 февраля 2004 года. М., 2004, с. 54-56.

Крысин Л.П. Русское слово, своё и чужое. Исследования по русскому языку и социолингвистике. М., 2004а.

Национальный корпус русского языка: 2003-2005. Результаты и перспективы. М., 2005.

Никипорец-Такигава Г. Агрессивная комедия. История слова «агрессивный» в трёх частях. // Slavic Culture Studies, TUFS, 5, 2005, с. 90-114.

Никипорец-Такигава Г. Агрессия в языке СМИ – опыт статистического анализа. // Язык, сознание, коммуникация, МГУ, 33, 2005а,

Компьютерные ресурсы для гуманитарных исследований

с. 61-76.

Поликарпов А.А., Кукушкина О.В. Полистилевой корпус русского языка. www.philol.msu.ru/humlang/articles/polystylcorp.html.

Шайкевич А.Я., Андリュщенко В.М., Ребецкая Н.А. Статистический словарь языка Достоевского, М., 2005.

Примечания

1. Под корпусом здесь и далее понимается совокупность электронных текстов на данном языке.
2. The Tübingen Russian Corpus (www.sfb441.uni-tuebingen.de/b1/rus/korpora.html).
3. The Helsinki annotated corpus of Russian texts HANCO (www.slav.helsinki.fi/hanco).
4. www.ruscorpora.ru
5. www.integrum.com
6. «Центральные газеты» (ЦГ), «Центральные журналы» (ЦЖ), «Региональная пресса» (РСМИ), «Теле-Радио эфир» (Т-Р), «Интернет-издания», «Компьютерная пресса».
7. На начало сентября 2006 года в «Интегруме» имелось примерно 6000 баз данных, состоящих из 350 миллионов документов. Базы «Интегрума» ежедневно увеличиваются на два источника.
8. Чем длиннее, подробнее будет запрос, тем больше потенциально возможных контекстов придумает исследователь, тем точнее будут соответствовать результаты на выходе ожиданиям. Важно, что «Интегрум» не имеет ограничений на длину запроса, обрабатывает запрос любой сложности.
9. Г. Никопорец-Такигава. Вторичные заимствования в русском языке XXI века. // [Integrum: 87-107].
10. Орловские новости. 6.07.2005.
11. Из разговора с врачом.
12. Смена (Санкт-Петербург). 08.08.2002.
13. Авторское телевидение. Программа «Ночной полёт». 04.11.2002.
14. Статистические исследования имеют разную форму представления результатов, выбираемую исследователем: отчёты в трёх разных форматах, таблицы, графики двух видов, гистограммы двух видов.
15. **Инструмент:** сравнительная статистика/по абсолютной шкале/гистограмма 3 D/ежегодно.
Запрос: «шок — это по-нашему» или «шок это по-нашему» или «шок это понашему» или «шок — это понашему».
Область поиска: все базы «Интегрума».
Временной интервал: без ограничений. **Примечание:** в запросе учтены ошибочные формы написания идиомы, так как они встречаются в российских СМИ.

Г. Никипорец-Такигава

16. **Инструмент:** сравнительная статистика/по абсолютной шкале/график/ежегодно.
Запрос: ((шок* не шокола*);2 приятн*) или (приятн*:2 (шок* не шокола*)).
Область поиска: все базы «Интегрума».
Временной интервал: 1993 – 2006. **Примечание:** сервис «Сравнительной статистики по абсолютной шкале» выясняет абсолютное количество примеров употребления исследуемой единицы языка в базах «Интегрума». Оператор «не» снимает омонимию.
17. «Вторично заимствованным считается иноязычное слово, по форме совпадающее с ранее заимствованным словом, но имеющее иное значение» [Крысин 2004 — с.54].
18. Экономическая газета, Москва. 17.11.1998.
19. «Интегрум» не располагает архивами СМИ раньше 1990 года, поэтому для более глубокого, чем 15 лет, сравнительно-исторического анализа невозможно использовать СМИ.
20. **Инструмент:** точный поиск.
Запрос: шок* не шокола*.
Область поиска: личная коллекция.
Временной интервал: без ограничений.
Примечания: для поиска примеров только из классической литературы удобно выбрать авторов в «Личную коллекцию». Следует сразу исключить слово шоколад (а также шоколат, который встречается, например, у А.С. Пушкина) и его дериваты при помощи оператора не: шок* не шокола*. По такому запросу было получено 26 документов, среди них 9 были нерелевантными: опечатка («шока» вместо «пока»), Шок как имя собственное у Набокова, повторы одного и того же произведения, критические статьи о творчестве писателя вместо его произведения.
21. **Инструмент:** сравнительная статистика/ по относительной шкале/ график/ ежегодно
Запрос: шок* не шокола*
Область поиска: ЦГ, ЦЖ, РСМИ.
Временной интервал: 1993 – 2006.
22. Советская Беларусь (Минск). 03.05.2005.
23. Курортный проспект (Сочи). 02.05.2005.
24. Правда. Ру. 04.05.2005.
25. Примеры приводятся в статьях [ГНТ-2005, с. 102-104, ГНТ-2005a].

ABSTRACT

Databases and computer technologies for research in the humanities

Galina Nikiporets-Takigawa

This article describes the history of the creation of Russian databases, their parameters, operating principles and the goals they were created for. Researchers who are active in Russian studies today can use a large number of tools based on computer technology. Four examples are: the Russian National corpus, the Helsinki annotated corpus of Russian texts HANCO, the Tuebingen Russian Corpus, and the Integrum Database Service. Linguists can also use Internet search engines, such as Yandex, Rambler or Google to find examples of usage or to check the frequency with which words are used.

Herein, the author addresses the fundamental differences between searching in databases and on the Internet, and focuses on the peculiarities of information retrieval systems. The author also gives separate consideration to the specialized options of the Integrum Database Service, developed for research purposes. Compared with other research tools, Integrum has some advantages that make it very effective when performing scientific analysis:

- o Compared with any other special corpus, Integrum is extraordinarily large in size. It consists of more than 350 million documents whereas the Russian National Corpus today includes only slightly more than 120 million running words.
- o Compared with Internet search engines, Integrum provides a much wider range of opportunities thanks to its well-organized structure. It is possible to compare the quantitative results of a search based on several different sources, or to concentrate on a particular type of material.
- o In Integrum, duplicated documents are quite rare, so they do not distort the final results.
- o Integrum material is of a very high quality.
- o It is easy to learn the basic techniques required in the use of Integrum. A 15-minute introduction is generally adequate, after which students are able to initiate small-scale independent analyses.

Г. Никинорец-Такигава

- o Compared with other databases, only Integrum ensures a dynamic approach to research data.

Over the last four years, the author has used a lot of databases and a variety of approaches and tools. In order to describe and to compare different computer technologies used in the social sciences, the author makes use of the results of a recent study. Short linguistic commentaries explain the phenomena researched by the author and the methodology for the use of computer technologies is described as well. The author then describes the research procedure step by step.

The author analyzes semantic changes in the Russian language over the last fifteen years. Modern technology allows us to analyze quantitative aspects of language usage in mass media. For example, using the resources provided by Integrum, the author carried out statistical research with regard to the usage of the words *aggressive*, *ambition* and *shock* by the mass media. Analysis revealed that the frequency of occurrence of these words is increasing. A semantic research and analysis of modern contextual engagement, using Integrum sources, showed that the meaning of these words is changing and that they are acquiring a positive connotation. The words have become more popular, not only owing to extralinguistic factors (the primary linguistic reason is the impact of the English language, where “aggressive” can have both a positive and a negative meaning in the corresponding business discourse, depending on the context). A search was carried out, based on the exact form of the word, and with various forms of fine-tuning of the search domain in the Integrum database. This was then supplemented by a statistical and context analysis of the words агрессивный (aggressive), шок (shock), амбиции (ambitions), using various instruments available on the Integrum Service (“Frequency Dictionary of Russian” and “Relative and comparative statistics”). In this way it is possible to specify the time when these words acquire neutral or positive meanings and the reasons why these meanings have survived and entered common usage. The author uses Integrum to identify the first recorded use of the expressions «приятно шокировать», «шок – это по-нашему» and to establish the frequency with which vocabulary relating to war and military aggression is used by the Russian mass media. The author has undertaken a semantic research and analysis of modern contextual engagement using Integrum sources and found that the

Компьютерные ресурсы для гуманитарных исследований

meaning of such language is shifting, acquiring a positive connotation.

This paper will show that the role of databases and computer technologies varies. In some cases it merely makes the collection of quantitative and qualitative data quicker and easier than it was with previously available methods. In other cases it enables us to carry out research that scholars in the past could only dream of.

Students can use databases and computer technologies to perform special exercises or in the writing of MA or doctoral theses while teachers and researchers can collect material for teaching or research work. In some cases a whole research question has sprung up from the possibility of using Integrum; in others it plays only a complementary role. Some researchers use Integrum merely as a huge electronic library while others take a much more sophisticated approach, e.g. searching for the first occurrence of a word or an expression. We can quite easily study stylistic tendencies in public language use. But maybe the real strength of Integrum is in finding a correlation between the dynamic of language usage and changes in social realities, be they of a cultural, political or economic nature. In conclusion, this paper stresses the need for a better understanding of different computer instruments to ensure that one finds a proper tool with which to achieve various research goals.