КАФЕДРА

ЭТА РАБОТА ПОХОЖА НА РАБОТУ АРХЕОЛОГОВ В ПОМПЕЯХ

ОБ ИННОВАЦИОННОМ НАПРАВЛЕНИИ В ЛИНГВИСТИКЕ РАССКАЗЫВАЕТ ДИРЕКТОР НАУЧНО-МЕТОДИЧЕСКОГО ЦЕНТРА КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ ВГУ ПРОФЕССОР А.А. КРЕТОВ

Алексей Александрович, в Воронежском университете вы не только заведуете кафедрой теоретической и прикладной лингвистики, но еще и возглавляете научно-методический Центр компьютерной лингвистики. Как это произошло?

– Я действительно являюсь директором НМЦ КомпЛи с 1998 года. Тогда Владимир Тихонович Титов, бывший в то время деканом факультета романо-германской филологии, предложил мне перейти с филфака на РГФ, чтобы создать и возглавить НМЦ КомпЛи. Я дал согласие, а его инициативу поддержал тогдашний ректор ВГУ Иван Иванович Борисов. Так и возник Центр. С самого начала он был результатом сотрудничества гуманитариев и «технарей», представленных сотрудниками факультета прикладной математики и механики и в первую очередь – выпускницей Ленинградского государственного университета Ириной Евгеньевной Ворониной. Сейчас И.Е. Воронина является кандидатом технических наук, доцентом кафедры ПО и администрирования информационных систем, автором двух научных монографий и не планирует останавливаться в своем научном росте. Трудно переоценить, особенно на начальных этапах нашего сотрудничества, моральную и организационную поддержку тогдашнего завкафедрой МО ЭВМ доцента Константина Савича Рыбака. Впрочем, и НМЦ возник не на голом месте: ему предшествовала созданная на филологическом факультете лаборатория компьютерной филологии, возникшая при поддержке проф. Зинаиды Даниловны Поповой, заведовавшей в то время кафедрой общего языкознания и стилистики, и проф. Виктора Михайловича Акаткина, возглавлявшего тогда деканат филологического факультета. В этом столь богатом для ВГУ на юбилеи году наш центр отмечает своё 15-летие, и мы с благодарностью вспоминаем всех, кто способствовал его становлению.

– Чем же занимается НМЦ КомпЛи и каким достижениями встречает свой юбилей?

– Работа НМЦ КомпЛи чисто условно подразделяется на три части: научно-исследовательскую, учебно-методическую и издательскую.

– Полагаю, для наших читателей будет интересна прежде всего научно-исследовательская работа вашего центра, предполагающая максимум инновационности. Не могли бы вы начать с неё?

– С удовольствием. Научно-исследовательская работа НМЦ КомпЛи ведётся по пяти направлениям.

1-ое – анализ и синтез лингвистических объектов. Итоги этой работы подведены на пяти конференциях «Проблемы компьютерной лингвистики», а также в двух монографиях И.Е. Ворониной: «Компьютерное моделирование лингвистических объектов» (Воронеж, ВГУ, 2007. – 177 с.) и «Моделирование и алгоритмизация исследования лингвистической реальности» (LAP LAMBERT Academic Publishing GmbH & Co. KG, Saarbrucken, Germany, 2011 – 263 с.). По результатам конференций центром подготовлено к печати и опубликовано пять выпусков научного продолжающегося издания «Проблемы компьютерной лингвистики». Одним из интересных аспектов этого направления является визуализация языка – представление языка и его подсистем, включая словарь, в виде «картинок».

– И что это даёт?

– Совершенно новые познавательные возможности. До 90% получаемой нами информации поступает по зрительному каналу. Он лучше всего приспособлен к анализу информации. Одно дело иметь формулу множества Б. Мандельброта, и совсем другое дело – любоваться узорами, являющимися её зрительным эквивалентом.

Мы научились представлять слово в виде линии в трехмерном пространстве словообразовательной системы. Научились представлять двуязычный словарь в виде пейзажа и лексико-семантическую систему языка в трёхмерном пространстве (третьим измерением является цвет), наконец, идеографический словарь в виде графа. К слову сказать, это направление называется «компьютерной когнитивной графикой» и в США является одним из приоритетных направлений компьютерных наук.

2-ое – лингвистическая прогностика. Это направление, развиваемое в ВГУ уже 20 лет, обеспечивает нашему университету приоритет в мировой науке. Итоги работы в этом направлении доложены на пяти конференциях «Проблемы лингвистической прогностики», организованных с участием НМЦ КомпЛи, результаты которых отражены в пяти выпусках материалов этих конференций, подготовленных к печати Центром. По лингвистической прогностике защищено пять кандидатских диссертаций. Это научное направление известно и по серии научных монографий «Библиотека лингвистической прогностики» (БЛП), основанной и издаваемой по решению Ученого совета факультета РГФ. На сегодняшний день БЛП насчитывает шесть монографий: А.А. Кретова, Л.В. Молчановой, О.Л. Гамовой, И.В. Домбровской, А.И. Кузнецовой, Ю.А. Силиной (в соавторстве с А.А. Кретовым).

На прошедшей недавно V конференции были подведены некоторые итоги развития лингвистической прогностики, позволившие получить интересный результат – мы научились оценивать и выражать количественно удовлетворённость носителей того или иного языка окружающей их действительностью.

– *Как же это можно измерить?*

– Для этого существует ИнПриМ – Индекс приятия мира. Вычисляется он довольно просто: достаточно определить общее количество обозначений в тексте ситуаций зрения, таких как смотреть, видеть, созерцать/рассматривать, вычислить долю употреблений, приходящихся на ситуации видения и смотрения и вычесть из «доли видения» «долю смотрения». Разность и даст Индекс приятия мира.

– *И как же приятие мира связано с видением и смотрением?*

– Опосредованно, но весьма жестко. Окружающая действительность воздействует на писателей. В зависимости от характера этого воздействия герои произведений чаще видят, чем смотрят или наоборот, что находит отражение в тексте. А это соотношение определяет значение ИнПриМа.

– *А в чем различие видения и смотрения?*

– Видеть толкуется словарями как «воспринимать зрением». Именно это толкование и дало название Индексу приятия мира. Видение всегда направлено от мира к субъекту. Синонимичным глаголу видеть является словосочетание ловить взгляд(ы). А ловят то, что летит к тебе. Субъект не может не видеть по своей воле: он может только не смотреть. Глагол видеть не употребляется в повелительном наклонении. Нельзя сказать «*Видь!», но можно сказать «Смотри!».

– *Но ведь можно сказать «Виждь!». По крайней мере, именно эта форма употреблена в стихотворении А.С. Пушкина «Пророк».*

– Да, эта церковнославянская форма есть у Пушкина. Но кому она принадлежит?

Как труп в пустыне я лежал,
И Бога глас ко мне воззвал:
«Восстань, пророк, и виждь, и внемли...»

Как видим, наделить зрением и слухом может только Бог.

Что касается смотрения, тут всё иначе: это действие всецело подконтрольно человеку. Смотреть или не смотреть – человек решает сам. Смотрение направлено от человека вовне: смотреть куда? Синонимичным глаголу «смотреть» является словосочетание «бросать взгляд». Видение и смотрение противопоставлены как непроизвольное и произвольное, субъектостремительное и отсубъектное действие. Акт видения – акт вос-приятия (принятия в себя) Божьего мира. Акт смотрения – акт проявления своей воли (своеволия) по отношению к Божьему миру. Полученные нами результаты всецело подтверждают предложенную интерпретацию.

В ходе исследований было обнаружено, что в лексико-семантических группах зрительного восприятия (ЛСГ ЗВ) почти всех обследованных языков происходит функциональная рокировка: ситуации видения в какой-то момент начинают встречаться в текстах реже, чем ситуации смотрения.

Если расположить языки по времени этого события, получаем следующую последовательность: французский (вторая половина 18-ого века) [Силина 2011] – русский (первая половина 19-ого века) [Кретов 2006] – немецкий (вторая половина 19-ого века) [Молчанова 2007] – британский вариант английского языка (вторая половина 20-ого века) [Переверзева 2002] (в американском варианте английского языка смотрение преобладает над видением с самого начала XIX века [Домбровская 2009]).

Если воспользоваться предложенным моей аспиранткой Ю.А. Силиной Индексом Приятия Мира (ИнПриМ), то упомянутая выше «функциональная рокировка видения и смотрения» является ничем иным, как изменением знака ИнПриМа с положительного на отрицательный: люди перестают пассивно и благодарно воспри-

Сотрудники НМЦ КомпЛи и участники Первой международной конференции по лингвистической прогностике, организованной НМЦ КомпЛи ВГУ. Воронеж, 2001 год.

нимать окружающий их мир и начинают активно воздействовать на него – в целях его улучшения.

Во французских текстах смена знака ИПМ на отрицательный совпадает с Великой французской революцией (1789–1799). В русских – с Отечественной войной 1812 года, в немецких – с франко-прусской войной (1870–1871), закончившейся объединением Германии в рамках Германской империи, в англо-британских – с последствиями Второй мировой войны: деколонизацией и распадом «Второй империи» (1945–1997).

В Америку же из Англии уезжали те, кого не удовлетворяла окружавшая их действительность. В этом и в притоке богатств из колоний секрет долгого удовлетворения британцев окружавшей их действительностью и изначальная неудовлетворенность ею уехавших – американцев. Жизнь американцев в свете ИнПриМа представляется следующим образом: с самого начала XIX века человек в Америке попадал в недружелюбное окружение индейцев. Индейский вопрос был решён во второй половине XIX века, и именно этот период (несмотря на гражданскую войну) стал самым счастливым в истории Америки (США). Затем удовлетворённость жизнью пошла на убыль: Великая депрессия и Вторая мировая война в первой половине XX века, поражение во вьетнамской войне, нарастание расовой и социальной напряженности, вылившиеся в негритянские и молодёжные протестные движения второй половины XX века, способствовали падению удовлетворённости окружающей действительностью.

Вряд ли все описанные выше совпадения в поведении ИнПриМа с историческими событиями можно признать случайностью. А если это не случайность, то, значит, нами найден способ по лингвистическим данным делать экстралингвистические выводы.

Сопоставление динамики ИПМ в исследованных лингвокультурах дает основание для следующих заключений.

● Духовная жизнь общества находит отражение в создаваемых текстах.

● Между ИПМ и внеязыковой действительностью наблюдается корреляция.

● ИПМ позволяет измерять и сравнивать в сопоставимых величинах удовлетворенность его носителей окружающей действительностью.

● На рубеже XVIII-XIX вв. самым неудовлетворенным окружающей действительностью народом были французы; начиная со второй половины XIX века по настоящее время – русские.

● Трагичным для французов был весь XIX век, для немцев – первая половина XX века.

● В конце XX в. удовлетворенность окружающей действительностью у англичан, французов и немцев практически сравнялась.

● Самым счастливым периодом для французов была 1-я половина XVII века, для англичан – 2-я половина XVII века, для немцев и русских – вторая половина XVIII века, для американцев – вторая половина XIX века.

● На протяжении всего исследованного периода самым благополучным народом были англичане: даже отрицательное значение ИнПриМа у них наименьшее.

● Если меру неудовлетворённости англичан окружающей действительностью во второй половине XX века принять за единицу, то на конец XX века неудовлётворённость окружающей действительностью у немцев и французов в 2 раза больше, у американцев – в 3 раза больше, а у русских – в 5 раз больше, чем у англичан. Попросту говоря, в конце XX века русским их жизнь нравилась в 5 раз меньше, чем англичанам, в 3 раза меньше, чем немцам с французами, и почти в полтора раза меньше, чем американцам. С учётом терпеливости и непритязательности русского народа эти данные весьма красноречивы.

Как видим, даже побочные результаты лингвопрогностических исследований представляют немалый интерес.

3-е направление – «квантитативная лексикология и лексико-семантическая типология языков мира». Это направление представлено докторской диссертацией В.Т. Титова (Тверь 2005) и двумя монографиями, в которых отражено содержание его диссертации: «Общая квантитативная лексикология романских языков»

Сотрудники МНЦ КомпЛи среди участников секции компьютерной лингвистики, сплотившиеся вокруг руководителя секции (сидит) на Девятой международной научно-методической конференции «Информатика: проблемы, методология, технологии», организованной ФКН, ВГУ.
12 февраля 2009 г.

(Воронеж, ВГУ, 2002) и «Частная квантитативная лексикология романских языков» (Воронеж, ВГУ, 2004). По этой проблематике защищены 2 кандидатские диссертации. На подходе ещё три.

Эта проблематика активно разрабатывается также в дипломных работах студентов отделения теоретической и прикладной лингвистики. Близится к завершению работа над докторской диссертацией И.А. Меркуловой, посвящённой «Параметрическому анализу лексики славянских языков». Обобщающий доклад на эту тему опубликован в материалах 14-го Международного съезда славистов и доложен в сентябре 2008 г. в Охриде (Македония). В настоящее время наш научный коллектив с участием завкафедрой романской филологии проф. В.Т. Титова разрабатывает научно-исследовательскую тему «Исследование единства Европы по данным лексики», предполагающую параметрический анализ 33 государственных языков стран Европы.

На сегодняшний день сотрудниками НМЦ КомпЛи заложена хорошая основа для исследований в области лексико-семантической типологии языков мира.

– И чем интересно это направление?

– Оно позволяет «взвешивать» слова и отвечать на вопрос, какое слово самое важное в том или ином языке.

– Верно ли я понимаю, что знание «весов» слов даёт возможность целенаправленно изучать иностранные языки от самых важных слов к менее важным?

– Да, конечно. Знание места слова в системе представляет не только теоретический, но и практический интерес.

– И какие же слова самые важные?

– Для английского – to get «получить (в собственность)», а для русского – дать. Кстати, слова со значением «дать», как установил В.Т. Титов, являются главными также в испанском, португальском, румынском и латинском языках.

– А какое слово главное во французском языке?

– Doux – «сладкий, нежный, мягкий».

– А вам не кажется, что это гармонирует с французским национальным характером, менталитетом и образом жизни?

– Согласен. И не только doux – с французским, но и get – с англо-американским, а дать – с русским.

4-е направление научной работы НМЦ КомпЛи – это создание и использование параллельных корпусов текстов.

Эта работа ведётся НМЦ КомпЛи с 2002 года в тесном сотрудничестве с РАН и её Институтом русского языка, которым создан и продолжает совершенствоваться Национальный корпус русского языка (www.ruscorpora.ru). В 2007 году это сотрудничество было официально оформлено договором между ИРЯ РАН и ВГУ. В соответствии с этим договором студенты 2-го курса ОТИПЛ проходят дистанционную практику по созданию параллельных текстов, руководство и мониторинг которой со стороны ИРЯ РАН осуществляет д.ф.н., проф., известный лексикограф, фразеолог, германист Д.О. Добровольский.

Национальный корпус РЯ содержит параллельные подкорпуса: англо-русский, русско-английский, немецко-русский и русско-немецкий. Об этих корпусах я и расскажу подробнее.

Англо-русский корпус содержит оригинальные тексты на английском языке и их переводы на русский общей длиной в 16 млн. словоупотреблений. Русско-английская часть корпуса содержит 2 млн. словоупотреблений. Немецко-русский и русско-немецкой корпуса насчитывают по 2,5 млн. словоупотреблений.

Параллельные корпуса представляют собой отрезки оригинального текста, сопровождаемые соответствующими им отрезками текста переводного.

Например, из англо-русского подкорпуса:

FAR FROM THE MADDING CROWD
Вдали от обезумевшей толпы
by Thomas Hardy, 1874
Томас Гарди 1874
CHAPTER I. DESCRIPTION OF FARMER OAK – AN INCIDENT
ГЛАВА I. ПОРТРЕТ ФЕРМЕРА ОУКА. ПРОИСШЕСТВИЕ
When Farmer Oak smiled, the corners of his mouth spread till they were within an unimportant

distance of his ears, his eyes were reduced to chinks, and diverging wrinkles appeared round them, extending upon his countenance like the rays in a rudimentary sketch of the rising sun.

Когда фермер Оук улыбался, губы у него так расплывались, что углы рта оказывались где-то возле ушей, а глаза становились узенькими щелками и вокруг них проступали морщинки, которые разбегались во все стороны, словно лучи на детском рисунке, изображающем восход солнца.

Ценность параллельных корпусов для факультета РГФ трудно переоценить. Сбор материала для сопоставительных исследований по лексике, грамматике, переводу становится прост и быстр. То, на что раньше уходили недели, месяцы, а то и годы труда, теперь может быть получено за несколько секунд. На Отделении ТИПЛ использование параллельных подкорпусов или основного корпуса НКРЯ стало обычной практикой.

Настало, видимо, время начинать работу над французско-русским, испанско-русским, а быть может, и над итальянско- и португальско-русскими корпусами. Но тут уже НМЦ КомпЛи не обойтись без сотрудничества с другими подразделениями факультета РГФ.

Достоинства параллельных корпусов в том, что их надо создать лишь один раз, в том что их можно накапливать, в том что они общедоступны и каждый, внеся свой вклад, может пользоваться вкладом других. Корпуса текстов (одноязычные и параллельные) предоставляют лингвисту-исследователю возможности, которые можно сравнить лишь со сказочным ковром-самолётом.

5-е, возникшее в 2007 году и едва ли не наиболее интенсивно развивающееся направление деятельности НМЦ КомпЛи – маркемология.

– Впервые слышу это слово. Что оно значит?

– Этот термин сравнительно недавно введен в науку нами: д.ф.н., проф. завкафедрой русской литературы XIX века ВГУ Андреем Анатольевичем Фаустовым и мной. Он состоит из двух частей: маркемо- и -логия и означает «учение о маркеме или маркемах». Словом «маркема» мы называем самые важные для автора текста слова, обозначающие самые важные понятия в тексте и тем самым маркирующие (отмечающие) эти понятия. Окончание термина -ема указывает на то, что данные слова относятся не к уровню непосредственного наблюдения, а к уровню конструктов.

– Нет ли тут противоречия? Ведь слова в тексте непосредственно наблюдаемы.

– Вы правы: противоречие есть, но это диалектические противоречия: противоречие явления и сущности, формы и содержания. Парадоксальность ситуации состоит в том, что текст в отвлечении от человека – это форма, и компьютер, которым мы пользуемся для его анализа, имеет дело только с формами, только с явлениями, в сущности, – только с комбинациями нулей и единиц. Нас же текст интересует, прежде всего, тем, чем он не является: своим содержанием. Задача состоит в том, чтобы через форму текста проникнуть в его содержание.

– Неужели это возможно?

– Замечательный вопрос! С одной стороны, совершенно невозможно, поскольку содержанием текста является его проекция на личность читателя. Сколько читателей, столько содержаний. Да и мы сами, перечитывая что-нибудь из классики с интервалом в 10-20 лет, читаем как бы другое произведение, извлекая из него новое содержание. И каждая эпоха по-своему прочитывает классические произведения.

– И как же можно такое меняющееся и ускользающее содержание ухватить, да ещё с помощью компьютера?

– Наша работа похожа на работу археологов в Помпеях. Лава, заливавшая людей, охлаждалась их испарявшимися телами и застывала. В ней образовывались пустоты. Археологи заполняют эти пустоты жидким гипсом, а когда тот застывает, удаляют лаву. Гипсовые отливки представляют собой образы людей, погибших при извержении: форма каверн в лаве превращается в полные трагического смысла фигуры людей. Примерно так же и автор текста отливает трагедию своей души в форму текста, которую читатели должны заполнить своими собственными мыслями, чувствами, эмоциями. И если автор со своей задачей справился, читатели воспринимают всё, описанное в книге, как часть собственного жизненного опыта, как то, что пережили они сами.

– Это интересно, но ведь вы ставите задачу проникнуть в содержание текста через его форму. Как же вы это делаете?

– Очень просто. Каждое слово в тексте повторяется какое-то количество раз – имеет частоту употребления. Частота – равнодействующая языка и речи. Для получения речевой составляющей, характеризующей текст, надо вычесть из частоты слова её языковую составляющую.

– А разве это возможно?

– Да, тут помог метод параметрического анализа лексики, предложенный В.Т. Титовым в его монографиях. Им была предложена формула вычисления веса слова в словаре. Мы применили её к вычислению веса слова в тексте. Смысл формулы прост: вес слова тем больше, чем меньше в тексте слов такой же или меньшей длины. Самый большой вес – у самых коротких слов.

– И как вы вычитаете из частоты её языковую составляющую?

– Для каждого слова вычисляются два веса: по частоте и по длине. Вес слова по длине и является языковой составляющей: над длиной слова автор текста не властен, она задана языком.

А вычитание из веса слова по частоте его веса по длине даёт ИнТеМ – индекс текстуальной маркированности словоформы.

– Действительно, просто. А как вы получаете маркемы?

– Маркемами мы называем 50 словоформ с максимальным ИнТеМом, прошедших через целую систему фильтров.

– И что же в таком случае остаётся?

– Остаётся отвлечённая лексика, по большей части, оформленная абстрактными суффиксами: -ость, -ие, -ство. Остаются природные явления: солнце, небо, земля и т.д. Остаются предметы, имеющие символическое значение: например, кинжал в одноименном стихотворении М.Ю. Лермонтова. Остаются обозначения частей человеческого тела в символическом значении: сердце (как обозначение души), голова (как обозначение ума) и т.д.

– А всё-таки, почему маркем именно 50? И что вообще даёт их выделение?

– Конечно, 50 – величина произвольная, определяемая чисто практическими соображениями: меньше – мало, а больше – много. Ответить же на второй вопрос – значит рассказать о важнейших результатах, полученных в рамках маркемологии.

Самыми наглядными результатами являются Карта русской литературы XVIII-начала XX века и Генетическое древо русской литературы того же периода.

Маркемный анализ позволяет измерить содержательную близость авторов, выявить важнейшие (доминантные) понятия как для отдельного автора, так и для литературных эпох.

Так, маркемный анализ позволил установить, что сердцем русской литературы является И.А. Гончаров, а «отцом» русской литературы – Н.М. Карамзин. Литературным «отцом» А.В. Кольцова был И.И. Дмитриев, а И.С. Никитина – Е.А. Баратынский.

Самыми важными понятиями (доминантами) русской классической литературы были: вторая половина XVIII века – добродетель, первая треть XIX века – человек; вторая треть XIX века – обстоятельство (вице-доминанта – действительность); третья треть XIX века – обстоятельство (вице-доминанта – впечатление); начало XX века – противоположность.

Кроме того, с помощью маркемного анализа мы можем ответить на вопрос, какой текст является наиболее «чеховским», «гончаровским» и т.д.

– И как это определить?

– Текст, в котором встречается максимальное количество маркем, характеризующих все тексты данного автора, и является самым характерным для данного автора текстом. Такие скопления маркем в текстах мы с А.А. Фаустовым называем констелляциями. Так вот самым «гончаровским» текстом является «Лучше поздно, чем никогда», а самым «чеховским» – «Скучная история».

– Такое богатство направлений и результатов впечатляет. Велик ли штат НМЦ КомпЛи?

– Три человека: директор (полставки), лаборант и методист. Нам очень нужен программист, но штатным расписанием НМЦ он, к сожалению, не предусмотрен. Надеюсь, традиция поддержки нашего Центра продолжится, и этот вопрос будет решен.

– Откуда такая уверенность?

– Вектор современного развития направлен на всемерную поддержку инновационных направлений. Это позволяет нам смотреть в будущее с оптимизмом.

Беседовала Тамара Дьякова