НМЦ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ: ИТОГИ И ПЕРСПЕКТИВЫ МЕЖДИСЦИПЛИНАРНЫХ ИССЛЕДОВАНИЙ

А.А. Кретов

Воронежский государственный университет

Работа НМЦ КомпЛи разделяется на три части: научно-исследовательскую, учебно-методическую и издательскую, разделить которые можно чисто условно.

Научно-исследовательская работа НМЦ КомпЛи ведётся по четырём направлениям.

1-е — анализ и синтез лингвистических объектов. Итоги этой работы подведены на трёх конференциях «Проблемы компьютерной лингвистики», а также в монографии И.Е.Ворониной «Компьютерное моделирование лингвистических объектов» (Воронеж, ВГУ, 2007. — 177 с.). По результатам конференций опубликовано два научных сборника, третий находится в работе.

2-е — Лингвистическая прогностика. Итоги этой работы доложены на 4 конференциях «Проблемы лингвистической прогностики», организованных с участием НМЦ КомпЛи. По результатам этих конференций опубликовано 4 научных сборника и защищено 4 кандидатских диссертации (Л.В. Молчановой, О.Л. Гамовой, И.В. Домбровской и М.А. Глушко). По решению Ученого совета факультета это научное направление становится известным научной общественности благодаря изданию серии научных монографий «Библиотека лингвистической прогностики», в которой на сегодняшний день насчитывается три тома: монографии А.А. Кретова, Л.В. Молчановой и О.Л. Гамовой.

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» (Воронеж, ВГУ, 2002) и «Частная квантитативная лексикология романских языков» (Воронеж, ВГУ, 2004). По этой проблематике защищена кандидатская диссертация О.В.Богдановой «Полипараметрическое исследование ядра лексической системы французского языка» (Воронеж, 2003), завершена и скоро будет представлена на кафедру диссертация Т.А.Казаковой «Параметрический анализ немецкой лексики». Параметрическому анализу каталанской лексики посвящено диссертационное исследование Е.В.Долбиловой, английской лексики — соискателя кафедры А.В. Ходыкиной, карачаево-балкарской лексики — выпускницы ОТиПЛ и соискателя кафедры ТиПл И.Д. Семёновой. Эта проблематика активно разрабатывается в дипломных работах студентов ОТиПЛ: параметрический анализ малагасийского языка осуществлён в дипломной работе К.П. Буравлёвой (2005), карачаево-балкарской лексики — в дипломной И.Д.Семёновой (2005), турецкой лексики — в дипломной работе В.П. Бугаёва (2006), македонской лексики — дипломной И.С. Лысенко (2007), украинской лексики — в дипломной А.А. Кобцева (2007), чешской лексики — в дипломной В.В. Тарховой (2008), в этом году параметрическому анализу лексики посвящены дипломные работы И.А.Терентьевой (польский язык), И.В. Воржевой (тофаларский язык), Ю.В. Макаровой (старославянский и литовский) преддипломные работы по параметрическому анализу пишут О. Марасанова (белорусский язык), В. Терехова (словацкий язык), Н. Хайткулова (узбекский язык). Параметрическому анализу лексики посвящены курсовые А.В. Булавиной (японский язык) и С. Колядко (латышский).

Ведется работа над докторской диссертацией И.А.Меркуловой, посвящённая «Параметрическому анализу лексики славянских языков». В этой работе анализируются результаты параметрического анализа 12 славянских языков: русского, украинского, белорусского, польского, чешского, словацкого, полабского, словенского, сербо-хорватского, македонского, болгарского и старославянского языков. В работе материалы ещё по трём языкам: нижне- и верхнелужицкому и кашубскому. Предварительные итоги докладывались в 2007 году на IV-й Международной научно-практической конференции «Проблемы изучения живого русского слова на рубеже тысячелетий» (Воронеж, ВГПУ). Обобщающий доклад на эту тему публикуется в материалах 14-го Международного съезда славистов и будет доложен в сентябре 2008 г. в Охриде (Македония).

На сегодняшний день НМЦ КомпЛи располагает также результатами параметрического анализа лексики новогреческого и финского языков, а также элект-

ронными версиями некоторых германских, иранских, финно-угорских, тюркских, малайско-полинезийских, палеоазиатских и банту-русских словарей.

Тем самым заложена хорошая основа для исследований в области лексико-семантической типологии языков мира.

4-е и едва ли не наиболее интересное для большинства сотрудников и студентов нашего факультета направление научной работы НМЦ КомпЛи — это создание и использование параллельных корпусов текстов.

Эта работа ведётся НМЦ КомпЛи последние 5 лет — с самого начала в тесном сотрудничестве с РАН и её Институтом русского языка, которым создан и продолжает совершенствоваться Национальный корпус русского языка (www.ruscorpora.ru). В прошлом году это сотрудничество было официально оформлено договором между ИРЯ РАН и ВГУ. В соответствии с этим договором студенты 2 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 и немецко-русский.

Об этих корпусах я и расскажу подробнее.

Англо-русский корпус содержит оригинальные тексты на английском языке и их переводы на русский.

№	Название произведения	Слов
1	Arthur Conan Doyle «A Scandal in Bohemia» Артур Конан-Дойль Скандал в Богемии	15800
2	Arthur Hailey. The Final Diagnosis; Артур Хейли Окончательный Диагноз	156000
3	Bram Stoker «Dracula»	286400
4	Charles Dickens THE PICKWICK PAPERS Чарльз Диккенс «Записки Пиквикского клуба»	526170
5	Charles Dickens «Oliver Twist» (1838)	293150
6	Charles Dickens «The Tale of two Cities»	312000
7	Erle Stanley Gardner «The Case of the Blonde Bonanza» Эрл Стэнли Гарднер Белокурая удача	142700
8	Erle Stanley Gardner «The Case Of The Daring Divorcee»; Эрл Стэнли ГАРДНЕР ДЕЛО СМЕЛОЙ РАЗВЕДЁНКИ	102000
9	Ernest Miller Hemingway «A farewell to arms» Э. Хемингуэй, «Прощай, оружие!»	162900
10	Ernest Miller Hemingway «FOR WHOM THE BELL TOLLS»	332000
11	Francis Scott Key Fitzgerald «The Great Getsby» Ф.С. Фицджеральд «Великий Гэтсби»	100000
12	Gilbert Keith Chesterton «THE INNOCENCE OF FATHER BROWN»	59600
13	Graham Greene «THE POWER AND THE GLORY»; Грэм Грин. Сила и слава.	83000
14	HELEN FIELDING «Bridget Jones's Diary»	142000
15	Herman Melville «Moby Dick»; Герман Мелвилл Моби Дик, или Белый кит	398030
16	Jack London «The BURNING DAYLIGHT»	225680
17	Jack London «The Call of the Wild»; Джек Лондон ЗОВ ПРЕДКОВ	64450
18	Jack London «Martin Eden»	104030
19	Jane Austen. Persuasion; ДЖЕЙН ОСТИН Доводы рассудка	149000
20	Jerom Klapka Jerom «Three men in a boat (to say nothing of the dog)»	130500
21	Joanne Kathleen Rowling «Harry Potter and the Sorcer's Stone»	148000
22	Kenneth Grahame THE WIND IN THE WILLOWS	97600
23	KURT VONNEGUT «HOCUS POCUS»	154800
24	Lyman Frank Baum Баум The Marvelous Land of Oz Чудесная страна Оз	79890
25	MARK TWAIN «THE ADVENTURES OF TOM SAWYER»	144580
26	MARK TWAIN «THE ADVENTURES OF HUCKLEBERRY FINN (Tom Sawyer's Comrade)» 1884	218400
27	MARK TWAIN «TOM SAWYER, DETECTIVE»; Марк Твен Том Сойер — сыщик	46300
28	MARY MAPES DODGE «HANS BRINKER OR THE SILVER SKATES»	164000
29	Ray Bradbury «Fahrenheit 451»	95000
30	Robert Louis Stevenson «THE STRANGE CASE OF DR. JEKYLL AND MR.HYDE»	49520
31	Robert Louis Stevenson «TREASURE ISLAND»	130050
32	Theodor Dreiser «Sister Carrie» Т. Драйзер Сестра Керри	227880
33	Thomas Hardy FAR FROM THE MADDING CROWD 1874; Томас Гарди «Вдали от обезумевшей толпы»	262900

34	Thomas Harris «The Silence of the Lambs».	146200
35	Ursula Le Guin «The tombs of Atuan»	90000
36	Walter Scott «IVANHOE»	339670
37	William Golding «Lord Of The Flies»	120000
	ВСЕГО:	6.300.200

В настоящее время на сайте выставлены не все тексты, а лишь немногим более 3,6 млн словоупотреблений.

Русско-английская часть корпуса содержит около 1 млн словоупотреблений (точнее — более 760 000 словоупотреблений), но будет увеличена, так как сейчас идёт работа над созданием параллельных текстов «Анны Карениной» и «Воскресения» Л.Н.Толстого.

№	Автор	Произведение	Слов	Накоплено
1	А.П. Чехов	Скучная история	42 754	42 754
2	А.П. Чехов	Мужики	22 609	65 363
3	А.П. Чехов	Попрыгунья	17 767	83 130
4	А.П. Чехов	Дама с собачкой	11 879	95 009
5	А.П. Чехов	Человек в футляре	9526	104 535
6	А.П. Чехов	Крыжовник	7585	112 120
7	А.П. Чехов	О любви	6919	119 039
8	А.П. Чехов	Счастье	6889	125 928
9	А.П. Чехов	Мечты	5998	131 926
10	А.П. Чехов	Свирель	5754	137 680
11	А.П. Чехов	Красавицы	5709	143 389
12	А.П. Чехов	Пари	5010	148 399
13	А.П. Чехов	Сапожник и нечистая сила	4732	153 131
14	А.П. Чехов	В сарае	3985	157 116
15	А.П. Чехов	Тоска	3608	160 724
16	А.П. Чехов	На святках	3199	163 923
17	А.П. Чехов	Егерь	3065	166 988
18	А.П. Чехов	Злоумышленник	2594	169 582
19	М. Ю. Лермонтов	Герой нашего времени	101 070	270 652
20	Н.В. Гоголь	Тарас Бульба	89 210	359 862
21	Н.В. Гоголь	Мертвые души	164 800	524 662
22	А.С. Пушкин	Капитанская дочка	60 840	585 502
23	И.С. Тургенев	Новь	176 000	761 502

Немецко-русский подкорпус насчитывает в совокупности более 1,1 млн словоупотреблений и также будет увеличен до конца года.

№	Название произведения	Кол-во слов	Накоплено
1	Erich Maria Remarque «Der schwarze Obelisk»	248000	248000
2	Ernst Theodor Amadei Hoffmann «Der goldne Topf»	54300	302300
3	Ernst Theodor Amadei Hoffmann «Klein Zaches genannt Zinnober.Ein Märchen»	66600	368900
4	Franz Kafka «Der Prozess»	135100	504000
5	Heinrich Boell «Ansichten eines Clowns»	147100	651100
6	Heinrich von Kleist «Michäl Kohlhaas«	59800	710900
7	Herman Hesse «Siddhartha»	66300	777200
8	Michael Andreas Hellmuth Ende «Momo oder Die Geschichte von den Zeit-Dieb»	118600	895800
9	Sueskind, Patrick «Das Parfuem: Die Geschichte eines Moerders».	140800	1036600
10	Thomas Mann «Der Zauberberg» (1-4 главы)	104300	1.140.900

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

Например, из англо-русского корпуса:

FAR FROM THE MADDING CROWD

Вдали от обезумевшей толпы

by Thomas Hardy, 1874

Томас Гарди 1874

From the Penguin edition, 1978

Роман Перевод М. Богословской (вступление, части I-II) и Н. Высоцкой (части III-V) М., Художественная литература, 1970

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.

His Christian name was Gabriel, and on working days he was a young man of sound judgment, easy motions, proper dress, and general good character.

Звали его Габриэль, и в будние дни это был рассудительный молодой человек, одетый как полагается, державшийся спокойно и просто, словом, во всех отношениях вполне положительная личность.

Фрагмент русско-английскогокорпуса:

ВОСКРЕСЕНИЕ

THE AWAKENING (The Resurrection)

{A}Л. Н.ТОЛСТОЙ

{A}By COUNT LEO TOLSTOI Translated by William E.Smith

{A}I

{A}CHAPTER I.

{A}Как ни старались люди, собравшись в одно небольшое место несколько сот тысяч, изуродовать ту землю, на которой они жались, как ни забивали камнями землю, чтобы ничего не росло на ней, как ни счищали всякую пробивающуюся травку, как ни дымили каменным углем и нефтью, как ни обрезывали деревья и ни выгоняли всех животных и птиц, — весна была весною даже и в городе.

{A}All the efforts of several hundred thousand people, crowded in a small space, to disfigure the land on which they lived; all the stone they covered it with to keep it barren; how so diligently every sprouting blade of grass was removed; all the smoke of coal and naphtha; all the cutting down of trees and driving off of cattle could not shut out the spring, even from the city.

Фрагмент немецко-русского корпуса:

Der schwarze Obelisk.

Черный обелиск.

Erich Maria Remarque.

Эрих Мария Ремарк.

Geschichte einer verspäteten Jugend.

{A}Die Sonne scheint in das Büro der Grabdenkmalsfirma Heinrich Kroll und Söhne.

{A}Солнце заливает светом контору фирмы по установке надгробий «Генрих Кроль и сыновья».

Es ist April 1923, und das Geschäft geht gut.

Сейчас апрель 1923 года, и дела идут хорошо.

Das Frühjahr hat uns nicht im Stich gelassen, wir verkaufen glänzend und werden arm dadurch, aber was können wir machen — der Tod ist unerbittlich und nicht abzuweisen, und menschliche Trauer verlangt nun einmal nach Monumenten in Sandstein, Marmor und, wenn das Schuldgefühl oder die Erbschaft beträchtlich sind, sogar nach dem kostbaren, schwarzen, schwedischen Granit, allseitig poliert.

Весна не подкачала, мы торгуем блестяще, распродаем себе в убыток, но что поделаешь — смерть немилосердна, от нее не ускользнешь, однако человеческое горе никак не может обойтись без памятников из песчаника или мрамора, а при повышенном чувстве долга или соответствующем наследстве — даже из отполированного со всех сторон черного шведского гранита.

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.

С помощью параллельных текстов написали свои кандидатские диссертации И.В.Чарычанская и Ю.П.Плешкова, свои дипломные работы — Е.Е. Гурина «Формальная синтаксическая и лексическая асимметрия в параллельных текстах» (2005), и О.Катова (2008).

На Отделении ТИПЛ использование параллельных корпусов или НКРЯ стало обычной практикой. Хотелось бы, чтобы эта практика распространилась и на другие отделения факультета РГФ.

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 сможет обойтись без сотрудничества с другими подразделениями факультета.

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 воспользоваться вкладом других.

Ну, и в заключение скажу, что корпуса текстов (одноязычные и параллельные) предоставляют лингвисту-исследователю возможности, которые можно сравнить лишь со сказочным ковром-самолётом.

Временные рамки не позволяют подробно раскрыть учебно-методическую и издательскую часть работы НМЦ КомпЛи, но о ней уже немало было сказано по ходу описания работы научно-исследовательской составляющей.

Воронежский государственный университет
А.А. Кретов, доктор филологических наук, профессор, зав. кафедрой теоретической и прикладной лингвистики
a_a_kretov@rambler.ru
тел. 20-41-49

Voronezh State University
A.A. Kretov, Doctor of Philological Science, Professor, the Head of the Department of Theoretical and Applied Linguistics
a_a_kretov@rambler.ru
tel. 20-41-49