Н. В. Бунтман, Анна А. Зализняк, И. М. Зацман, М. Г. Кружков, Е. Ю. Лощилова, Д. В. Сичинава, Информационные технологии корпусных исследований: принципы построения кросслингвистических баз данных, *Информ. и её примен.*, 2014, том 8, выпуск 2, 98–110

DOI: 10.14357/19922264140210

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 2. С. 98–110

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ КОРПУСНЫХ ИССЛЕДОВАНИЙ: ПРИНЦИПЫ ПОСТРОЕНИЯ КРОССЛИНГВИСТИЧЕСКИХ БАЗ ДАННЫХ*

Н. В. Бунтман[1], Анна А. Зализняк[2], И. М. Зацман[3], М. Г. Кружков[4], Е. Ю. Лощилова[5], Д. В. Сичинава[6]

Аннотация: Рассматривается информационная технология создания кросслингвистических баз данных текстов на русском языке и их переводов на французский язык, называемых параллельными текстами. Разработанные принципы построения этой базы данных обеспечивают реализацию уникального сочетания трех видов двуязычного поиска: лексического, грамматического и лексико-грамматического. Отличительной чертой рассматриваемой технологии является одновременное формирование русско-французского параллельного подкорпуса Национального корпуса русского языка (НКРЯ) и кросслингвистической базы данных глагольных лексико-грамматических форм русского языка и их функциональных эквивалентов во французских переводах. Подкорпус и база данных обладают разной глубиной выравнивания: в первом случае оно выполняется на уровне предложений, а во втором — на уровне конструкций. Теоретическое значение создания этой базы данных заключается в обеспечении исследований как в области двуязычной контрастивной грамматики, так и в направлении создания грамматики русского языка, опирающейся на современную эмпирическую базу и информационные технологии корпусной лингвистики. Ее основное прикладное назначение заключается в повышении качества машинного перевода.

Ключевые слова: параллельный корпус; информационная технология; кросслингвистические базы данных; двуязычный лексико-грамматический поиск; корпусная лингвистика; контрастивная грамматика
DOI: 10.14357/19922264140210

1 Введение

Возникновение параллельных электронных корпусов ознаменовало начало новой эры контрастивных лингвистических исследований; пионерскими в этой области стали работы 1990-х гг. Стига Йоханссона с англо-норвежским корпусом. Сочетание методов современной компьютерной лингвистики с возможностями сопоставления текстов на двух и более языках, предоставляемыми параллельными корпусами, обеспечило возможность осуществления контрастивного лингвистического анализа на принципиально новом уровне точности (ср. [1]). Благодаря таким корпусам за прошедшие два десятилетия в этой области были достигнуты значительные успехи как в плане разработки методик анализа, так и в плане создания оригинальных лексикографических описаний. О перспективах контрастивных грамматических исследований на базе параллельных корпусов см. работы [1–7].

В работе [8] была описана технология формирования русско-французского параллельного подкорпуса НКРЯ, содержащего литературные произведения на русском языке и их переводы на французский язык.

В настоящее время русско-французский подкорпус НКРЯ содержит тексты произведений совокупным объемом в 2 млн словоупотреблений. При этом часть параллельных текстов представлена в *поливариантном формате*, т. е. одно произведение на русском языке выровнено в корпусе по предложениям с *несколькими вариантами его перевода* на французский язык. Совокупный объем поливариантных текстов — 700 тыс. словоупотреблений, т. е. больше трети всего параллельного русско-французско-русского подкорпуса НКРЯ.

Параллельные корпуса стали включаться в состав НКРЯ с 2005 г. [2, 4, 9, 10]. Сейчас он включает восемь двуязычных параллельных подкорпусов

*Работа выполнена в ИПИ РАН при поддержке фонда «Династия» (грант NG13-036) и РФФИ (грант № 13-06-00403).

[1]Московский государственный университет им. М. В. Ломоносова, факультет иностранных языков и регионоведения, nabunt@hotmail.com

[2]Институт языкознания Российской академии наук; Институт проблем информатики Российской академии наук, anna.zalizniak@gmail.com

[3]Институт проблем информатики Российской академии наук, izatsman@yandex.ru

[4]Институт проблем информатики Российской академии наук, magnit75@yandex.ru

[5]Институт проблем информатики Российской академии наук, lena0911@mail.ru

[6]Институт русского языка Российской академии наук, mitrius@gmail.com

с русским языком оригинала или перевода (английский, немецкий, французский, испанский, итальянский, польский, украинский и белорусский) и один многоязычный параллельный подкорпус. Подкорпус параллельных текстов на русском и французском языках появился в составе НКРЯ в декабре 2012 г. Технологию, используемую для формирования этого подкорпуса и описанную ранее в работе [8], обозначим как Parallel Corpus technology или ParCor-технология.

В 2013 г. ParCor-технология была дополнена новыми операциями: была создана база данных глагольных форм русского языка и вариантов их перевода на французский язык (далее — БД) и сформирован поливариантный подкорпус параллельных текстов на русском и французском языках (далее — подкорпус). Новая технология дала возможность формировать БД одновременно с пополнением подкорпуса и реализовать три вида двуязычного поиска глагольных форм и их переводов: лексического, грамматического и лексико-грамматического. Например, в этой БД можно задать и выполнить запрос на поиск параллельных выровненных текстовых фрагментов, в которых в русском оригинале употреблена глагольная форма прошедшего времени несовершенного вида, а в параллельных французских фрагментах — passé composé. Технологию, ориентированную на одновременное формирование подкорпуса и БД с двуязычным поиском параллельных глагольных форм в оригинальном и переведенных текстах обозначим как Database Parallel Corpus technology или DBParCor-технология.

Цель статьи состоит в том, чтобы описать назначение, задачи и принципы построения БД, формируемой на основе параллельных текстов НКРЯ, а также функции двуязычного поиска, реализованные в этой БД.

2 Назначение базы данных и принципы ее построения

Принципы построения БД во многом диктовались ее назначением. Она создавалась как инструмент описания русской грамматической семантики «в зеркале французского языка», а также с целью уточнения положений русско-французской контрастивной грамматики. При выработке принципов построения БД использовались работы Гака [11, 12], Кузнецовой [13], Гиро-Вебер [14] и др. Эти работы, однако, появились в докорпусную эпоху; теперь, когда созданы и регулярно пополняются русско-французские корпуса, стали доступны параллельные тексты в цифровой форме, их сопоставление и анализ дает возможность уточнить описание русско-французской контрастивной грамматики.

В ходе разработки БД учитывалось то, что объектом анализа являются соответствия глагольных категорий русского и французского языка в параллельных текстах. Было определено несколько новых терминов, которые отражают существо принципов построения БД.

Ключевыми являются понятия «лексико-грамматическая форма», или ЛГФ, и «базовый вид ЛГФ», определения которых даны ниже.

Определение 1. Под *лексико-грамматической формой* (ЛГФ) понимается совокупность элементов конкретного предложения, обладающая набором признаков, задаваемых базовым видом ЛГФ.

Определение 2. Под *базовым видом ЛГФ* понимается определенная комбинация значений следующих параметров:

- категориальная принадлежность языковой единицы (в данной статье рассматриваются только глагольные ЛГФ, т. е. значение этого параметра фиксировано);
- набор значений грамматических категорий, релевантных для выбранного класса единиц;
- (факультативно) определенные элементы структуры предложения, задающие «конструкцию»; например: «PastPF + *если бы*».

Другими словами, базовый вид ЛГФ представляет собой некоторую комбинацию значений глагольных категорий, в совокупности с определенными элементами структуры предложения задающую некоторую «конструкцию»[1].

В процессе формирования БД было выделено 15 базовых видов ЛГФ русского языка; это так называемое множество-источник (табл. 1)[2]. Количество базовых видов ЛГФ французского языка (множество-цель) не фиксировано, поскольку оно возрастает по мере пополнения БД; в текущем варианте БД оно составляет 25 единиц (табл. 2).

Помимо базовых видов ЛГФ для каждого из двух языков сформировано множество дополнительных признаков, которые позволяют специфицировать тип конструкции. А именно: дополнительные признаки характеризуют либо состав глагольной груп-

[1] В значении, которое придается этому понятию в Грамматике конструкций [15–17].

[2] В текущую версию БД включены только те ЛГФ русского языка, которые содержат глагол в финитной форме (т. е. исключались безличные глаголы, слова категории состояния, причастия, деепричастия, а также перифразы с глаголом *быть*). В дальнейшем состав рассматриваемых видов глагольных форм будет расширяться.

Таблица 1 Множество-источник базовых видов глагольных ЛГФ русского языка

	Полное название базового вида ЛГФ	Сокращенное обозначение базового вида ЛГФ
1.	Настоящее	Pres
2.	Прошедшее НСВ	Past-IPF
3.	Прошедшее СВ	Past-PF
4.	Простое будущее	Fut-PF
5.	Сложное будущее	Fut-IPF
6.	Императив СВ	Imperat-PF
7.	Императив НСВ	Imperat-IPF
8.	Форма с *бы* СВ	Past-PF+*бы*
9.	Форма с *бы* НСВ	Past-IPF+*бы*
10.	Форма с *если бы* СВ	Past-PF+*если бы*
11.	Форма с *если бы* НСВ	Past-IPF+*если бы*
12.	Форма с *чтобы* СВ	Past-PF+*чтобы*
13.	Форма с *чтобы* НСВ	Past-IPF+*чтобы*
14.	Форма с *было* СВ	Past-PF+*было*
15.	Форма с *было* НСВ	Past-IPF+*было*

Таблица 2 Множество-цель базовых видов ЛГФ французского языка

	Полное название базового вида ЛГФ	Сокращенное обозначение базового вида ЛГФ
1.	Présent	Pr
2.	Passé composé	PasCom
3.	Passé simple	PasSim
4.	Imparfait	Imparf
5.	Plus-que-parfait	PqParf
6.	Passé antérieur	PasAnt
7.	Passé immédiat	PasIm
8.	Futur simple	Fut
9.	Futur antérieur	FutAnt
10.	Futur immédiat	FutIm
11.	Impératif	Imperat
12.	Subjonctif présent	SubjPres
13.	Subjonctif passé	SubjPas
14.	Subjonctif imparfait	SubjImparf
15.	Subjonctif plus-que-parfait	SubjPqParf
16.	Conditionnel présent	CondPr
17.	Conditionnel passé	CondPas
18.	Participe présent	PartPr
19.	Participe passé	PartPas
20.	Participe passé composé	PartPasComp
21.	Gérondif	en PartPr
22.	Infinitif	Inf
23.	Préposition+infinitif	Prep+Inf
24.	Préposition+infinitif passé	Prep+InfPas
25.	Substantif	Subst

пы (например, наличие при глаголе подчиненного инфинитива, модального детерминанта или отрицания), либо тип предложения, в котором употреблена данная ЛГФ (например, придаточное, вопросительное предложение, диалогическая реплика) (табл. 3 и 4). Каждый признак приложим или ко всем, или к некоторым из базовых видов ЛГФ. На всех рисунках статьи дополнительные признаки указаны в квадратных скобках после базового вида ЛГФ.

Определение 3. Комбинацию базового вида ЛГФ с одним или несколькими из дополнительных признаков назовем *видом ЛГФ*.

Принципы установления соответствия в параллельных выровненных текстах между русскими и французскими ЛГФ состоят в следующем. Сначала из фразы русского оригинала вычленяется фрагмент, включающий ЛГФ, базовый вид которой принадлежит множеству-источнику (см. табл. 1). Далее ищется ее «функционально эквивалентный фрагмент» (ФЭФ)[1] во французском переводе, из которого извлекается ЛГФ, базовый вид которой принадлежит множеству-цели (см. табл. 2).

Лексико-грамматическая форма русского языка и соответствующая ей ЛГФ французского языка образуют *моноэквиваленцию* (см. определение 4 и табл. 5). Если в процессе анализа ФЭФ оказывается, что нужный базовый вид французской ЛГФ в табл. 2 отсутствует, то множество-цель может быть пополнено. Случаи, когда для русской ЛГФ французский эквивалент не найден, отмечаются в БД специальной пометой (Nondetermined), и в процессе обработки данных они пока не учитываются[2].

Поиск ФЭФ и выявление содержащейся в нем ЛГФ французского языка являются первой задачей, решение которой обеспечивается разработанным вариантом БД. Для описания других задач БД, рассмотренных в следующем разделе, определим еще пять терминов: «моноэквиваленция», «тип моноэквиваленции», «полиэквиваленция», «тип полиэквиваленции» и «гиперэквиваленция».

Определение 4. *Моноэквиваленция* (МЭ) — это двухместный кортеж вида $\langle R_n(i); F_m(j)\rangle$, где первую позицию занимает i-е вхождение ЛГФ базового вида R_n русского языка (см. табл. 1) в оригинальном тексте. Вторую позицию занимает j-е вхождение ЛГФ базового вида F_m французского языка (см. табл. 2) в одном из вариантов перевода i-го вхождения русской ЛГФ. Все МЭ, входящие в БД, имеют идентификационный номер.

[1] Термин «функционально эквивалентный фрагмент» введен в работе [2], см. также [9].

[2] Речь идет о таких случаях, когда семантическое содержание, заключенное в выбранной ЛГФ оригинала, передано в переводе столь существенно иными лексическими средствами, что установление соответствия между ЛГФ при помощи того аппарата, который имеется на сегодня, оказывается невозможно. Например: *ты [. . .] так теребишь за носы, что* ***еле держатся*** *— tu tirais tellement sur leur nez [. . .] que* ***tu as failli le leur arracher***.

Таблица 3 Дополнительные признаки для базовых видов ЛГФ русского языка

Полное название дополнительного признака	Сокращенное обозначение дополнительного признака
Подчиненный инфинитив СВ	[SubInf-PF]
Подчиненный инфинитив НСВ	[SubInf-IPF]
Модальный детерминант	[ModDet]
Отрицание	[Neg]
Вопросительное предложение	[Interrog]
Восклицательное предложение	[Exclam]
Глагол, вводящий прямую речь	[VerbDirSp]
Глагол в составе диалогической реплики	[DialRepl]
Глагол в придаточном предложении	[Sub]
Глагол в изъяснительном придаточном	[SubCompl]
Глагол в определительном придаточном	[SubAttr]

Таблица 4 Дополнительные признаки для базовых видов ЛГФ французского языка

Полное название дополнительного признака	Сокращенное обозначение дополнительного признака
Подчиненный инфинитив	[SubInf]
Подчиненный инфинитив прошедшего времени	[SubInfPas]
Добавление подчиняющего предиката	[+SuperPred]
Модальный детерминант	[ModDet]
Отрицание	[Neg]
Вопросительное предложение	[Interrog]
Восклицательное предложение	[Exclam]
Глагол, вводящий прямую речь	[VerbDirSp]
Глагол в составе диалогической реплики	[DialRepl]
Глагол в придаточном предложении	[Sub]
Глагол в изъяснительном придаточном	[SubCompl]
Глагол в определительном придаточном	[SubAttr]
Глагол в условном придаточном	[SubCond]
Accusativus cum infinitivo	[Acc.c.Inf]
Faire + Infinitif	[*faire* + Inf]
Laisser + Infinitif	[*laisser* + Inf]
Sembler + Infinitif	[*sembler* + Inf]
Paraître + Infinitif	[$paraître$ + Inf]

Таблица 5 Моноэквиваленция, зарегистрированная в БД под номером 4711

№ МЭ	ЛГФ русского языка	Вид ЛГФ русского языка	ЛГФ перевода	Вид ЛГФ перевода
4711	потом [. . .] плотно **запер** все двери	Past-PF [ModDet]	après **avoir** bien **fermé** toutes les portes	Prep + InfPas [Sub]

Определение 5. *Типом моноэквиваленции* называется кортеж базовых видов ЛГФ русского и французского языка $\langle R_n; F_m\rangle$, например ⟨Past-PF; Prep + InfPas⟩ (см. 3-й и 5-й столбцы в табл. 5).

Определение 6. *Полиэквиваленция* — это двухместный кортеж вида $\langle R_n(i); \{F_m(j), F_k(r), \ldots\}\rangle$, представляющий собой объединение нескольких моноэквиваленций с идентичной первой позицией ($\langle R_n(i); F_m(j)\rangle$, $\langle R_n(i); F_k(r)\rangle$ и т. д.), отражающих разные варианты перевода одного и того же i-го вхождения ЛГФ базового вида R_n в русском оригинальном тексте: $F_m(j)$ — это ЛГФ французского языка, идентифицированная в первом переводе и соответствующая i-му вхождению русской ЛГФ, $F_k(r)$ — во втором переводе и т. д. (табл. 6).

Определение 7. *Типом полиэквиваленции* называется кортеж базовых видов ЛГФ русского и французского языка $\langle R_n; \{F_m, F_k, \ldots\}\rangle$, например ⟨Pres-IPF; {Pr,Pr}⟩ (см. 2-й и 5-й столбцы в табл. 6).

Таблица 6 Две моноэквиваленции (№№ 596, 5927), составляющие полиэквиваленцию*

ЛГФ русского языка	Вид ЛГФ русского языка	ЛГФ в текстах французских переводов и их виды		
		Номер моноэквиваленции	ЛГФ в текстах французских переводов	Вид ЛГФ французского языка
Я иногда в театр **хожу**	Pres-IPF [ModDet] [DialRepl]	596	**Il m'arrive d'aller** au théâtre,	Pr [SubInf] [+SuperPred] [DialRepl]
		5927	Non, je **vais parfois** au théâtre, et en visite.	Pr [ModDet] [DialRepl]

*Французские ЛГФ, входящие в данную полиэквиваленцию, имеют одинаковый базовый вид, но различаются на уровне дополнительных признаков, указанных в квадратных скобках.

Определение 8. *Гиперэквиваленция* — это двухместный кортеж вида $\langle R_n; \{F\}\rangle$, репрезентирующий соответствие между базовым видом ЛГФ русского языка R_n и множеством базовых видов эквивалентных ЛГФ французского языка, входящих во вторую позицию моноэквиваленций БД с ЛГФ базового вида R_n.

Другими словами, каждая гиперэквиваленция включает один базовый вид ЛГФ русского языка R_n и список базовых видов ЛГФ французского языка — при условии, что хотя бы одна ЛГФ базового вида из этого списка образовала в БД моноэквиваленцию с русской ЛГФ базового вида R_n.

Используя определенные выше термины, перечислим те задачи, для решения которых предназначена спроектированная БД:

- построение моно-, поли- и гиперэквиваленций;
- двуязычный лексический, грамматический и лексико-грамматический поиск моно- и полиэквиваленций;
- вычисление частотности для каждого типа моно- или полиэквиваленций.

Для решения этих задач был разработан веб-интерфейс, который позволяет пользователям-лингвистам взаимодействовать с БД в онлайн-режиме с помощью распространенных веб-браузеров (Internet Explorer, Mozilla Firefox, Google Chrome). Для создания и ведения БД используется СУБД Microsoft SQL Server.

Функции БД можно разделить на две основные группы:

(1) первая группа функций служит для построения и редактирования моноэквиваленций (см. рис. 1 для функции редактирования);

(2) вторая группа функций — для поиска уже построенных моно- и полиэквиваленций (см. рис. 2 с интерфейсом поиска и просмотра полиэквиваленций).

Группа функций построения и редактирования моноэквиваленций в БД позволяет отфильтровывать выровненные фрагменты оригинального и переводных текстов по названию книги, автору перевода и присутствующих в этих фрагментах видам ЛГФ. Используя эти функции, пользователь-лингвист может просматривать выровненные фрагменты параллельных текстов с целью формирования моноэквиваленций.

На начало 2014 г. построено 10 527 моноэквиваленций и на их основе автоматически было сгенерировано 4128 полиэквиваленций (т. е. объединений моноэквиваленций из разных переводов одного оригинального текста с одной и той же ЛГФ русского языка в первой позиции кортежа).

3 Двуязычный поиск

На странице поиска и просмотра полиэквиваленций пользователи БД могут видеть подборки полиэквиваленций (см. рис. 2), которые генерируются в соответствии с поисковым запросом. Пользователи БД могут осуществлять поиск моно- и полиэквиваленций, используя следующие поисковые признаки: название русского произведения, французский перевод, базовые виды и признаки ЛГФ русского и французского языка, лексемы оригинала и переводов, искомые тексты как последовательности знаков, включая знаки препинания (ср. опцию «поиск точных форм» в НКРЯ).

Поисковые признаки можно задавать как по отдельности, так и в сочетании. В результате выполнения поискового запроса можно узнать число найденных полиэквиваленций, удовлетворяющих заданным поисковым признакам, и посмотреть их.

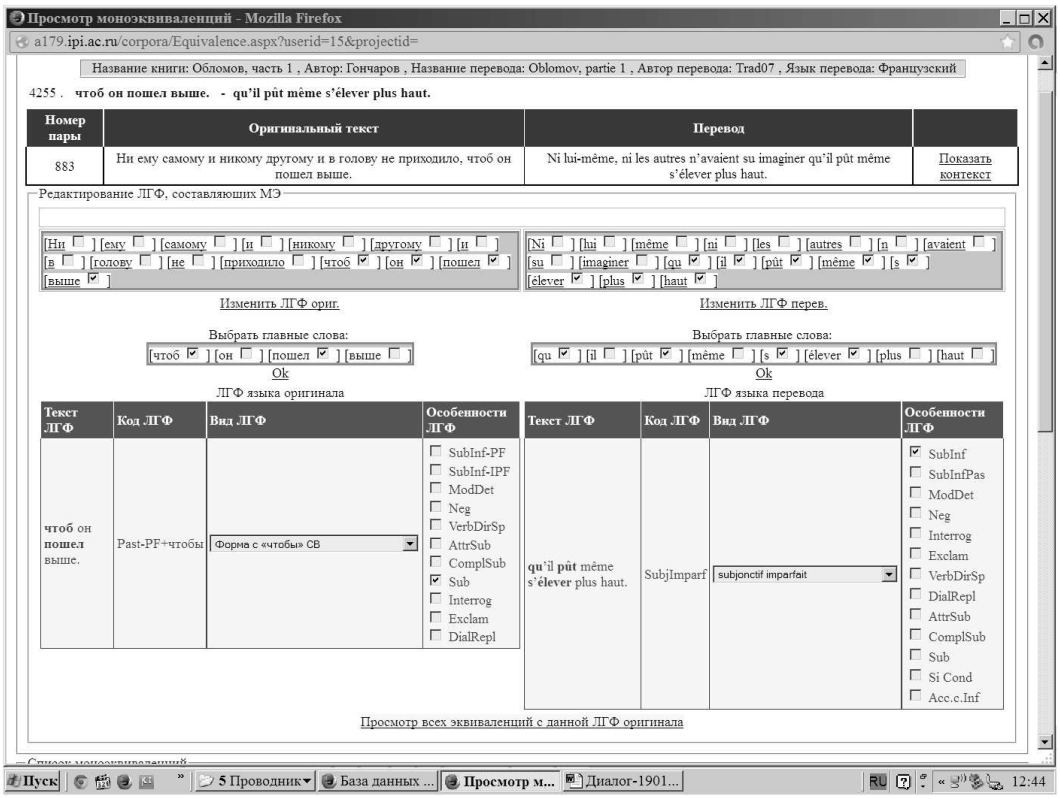

Рис. 1 Интерфейс для редактирования моноэквиваленций

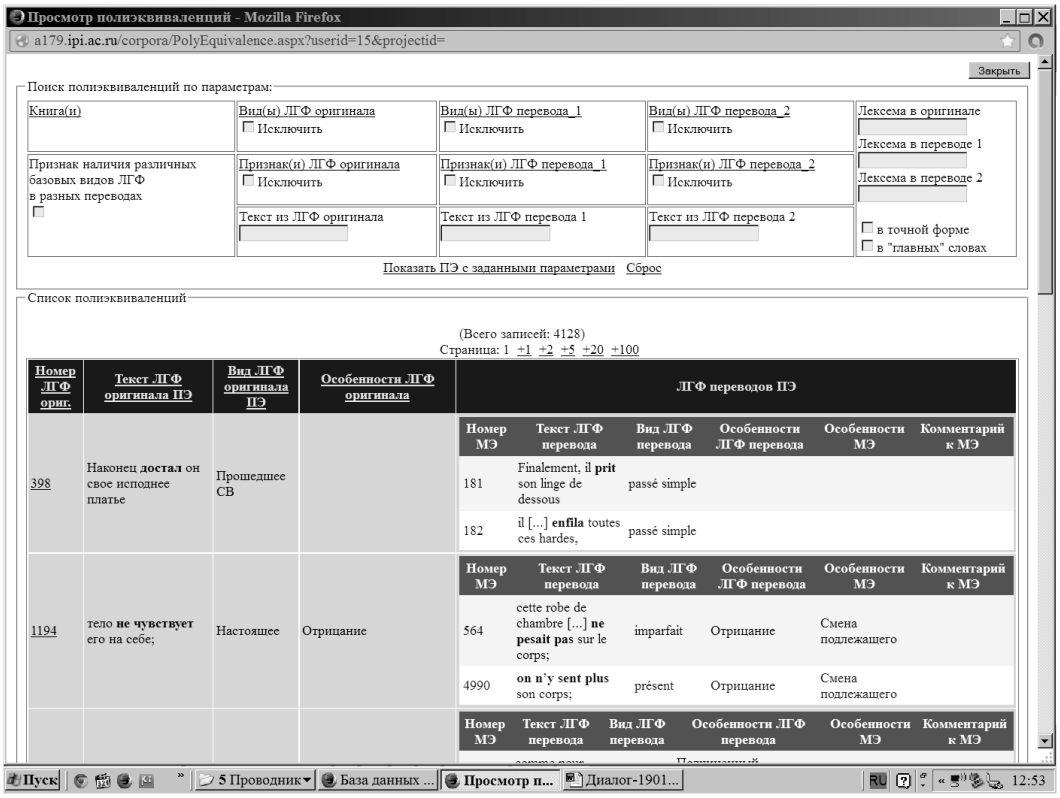

Рис. 2 Интерфейс для поиска и просмотра полиэквиваленций

Таблица 7 Две полиэквиваленции, найденные в БД по поливариантному двуязычному грамматическому запросу

Полиэквиваленция		Номер МЭ	ЛГФ перевода	Базовый вид ЛГФ перевода
он **решил оставить** [. . .] липовые и дубовые деревья	Past-PF [SubInf]	2931	alors qu'**il garderait** les [. . .] tilleuls et chênes,	CondPr
		8011	Il **décida de laisser** tels quels les [. . .] tilleuls et les chênes,	PasSim [SubInf]
он **решил** [. . .] яблони и груши **уничтожить**	Past-PF [SubInf]	2932	il **se débarrasserait** des pommiers et des poiriers	CondPr
		8013	Il **décida** [. . .] **de supprimer** les pommiers et les poitiers	PasSim [SubInf]

Принципиально новой является функция двуязычного грамматического поиска, который применим как к одному, так и одновременно к нескольким переводам (поливариантный двуязычный запрос). Например, если задать один базовый вид ЛГФ русского языка «Past-PF» и соответстующие в двух вариантах перевода два базовых вида ЛГФ французского языка CondPr и PasSim, то в БД будут найдены две полиэквиваленции с заданными поисковыми грамматическими признаками, отраженные в табл. 7.

Кроме базового вида ЛГФ в поисковом запросе могут задаваться также дополнительные признаки для ЛГФ русского языка (из табл. 3) и для ЛГФ французского языка (из табл. 4).

Например, если задать поисковый запрос «Pres [SubInf-PF]» в русской фразе и «CondPr [SubInf]» хотя бы в одном из двух ее переводов, то в БД будут найдены три полиэквиваленции с заданными видами ЛГФ (при этом в найденных полиэквиваленциях могут присутствовать и другие дополнительные признаки) (табл. 8).

Приведенные примеры двуязычного грамматического поиска говорят о том, что разработанная БД является на сегодняшний день уникальным лингвистическим ресурсом, который может быть использован для исследования не только глагольных форм, но и более широкого спектра языковых единиц (см. разд. 4).

4 Исследование лингвоспецифичной лексики с помощью базы данных

В настоящее время DBParCor-технология и БД адаптируются к исследованию лингвоспецифичных единиц (ЛСЕ) русского языка «в зеркале ино-

Таблица 8 Три полиэквиваленции, найденные в БД по видам ЛГФ

Полиэквиваленция		Номер МЭ	ЛГФ переводов	Вид ЛГФ переводов
Не может постараться для барина!	Pres [SubInf-PF] [Neg]	661	**Tu pourrais** tout de même **faire un effort** pour ton maître!	CondPr [SubInf] [Exclam]
		5897	Il **ne peut** même **pas faire** un petit effort pour son maître!	Présent [SubInf] [Exclam]
теперь **можете отдать**	Pres [SubInf-PF]	945	maintenant vous **pouvez** me **rembourser**.	Présent [SubInf]
		7584	alors vous **pourriez** peut-être **me rembourser**?	CondPr [SubInf] [ModDet] [Interrog]
Разве я **могу** все это [. . .] **перенести**?	Pres [SubInf-PF] [Interrog]	8939	Est-ce que je **puis** [. . .] le **supporter?**	Présent [SubInf] [Interrog]
		8940	Je **pourrais** [. . .] **supporter** tout ça?	CondPr [SubInf] [Interrog]

странных языков», включая французский. Для решения задач проекта «Контрастивное корпусное исследование специфических черт семантической системы русского языка», финансируемого по гранту РФФИ, была разработана оригинальная методология контрастивного корпусного анализа, осуществляемого с помощью БД, которая опирается на концептуальный аппарат контрастивного анализа русских лексико-грамматических форм, описанный выше. Данная методология предусматривает:

- статистическое и/или экспертное обоснование гипотез лингвоспецифичности лексических единиц русского языка на основе анализа текстов двуязычных корпусов, БД и других текстовых источников;
- статистическое обоснование с помощью БД гипотез лингвоспецифичности лексических единиц русского языка, сформулированных в ходе предшествующих исследований на основании семантического анализа;
- статистическую и экспертную верификацию гипотез с использованием БД.

Если для статистического обоснования гипотез могут использоваться разные информационные ресурсы (книги, корпуса или БД), то для верификации гипотез используется только БД, так как ключевым этапом верификации является построение моно- и полиэквиваленций и вычисление частотности их типов. Построение моно- и полиэквиваленций позволяет документировать процесс статистической верификации гипотез, а также согласовывать и документировать результаты верификации, выполненной лингвистами-экспертами (экспертная верификация гипотез).

Проведенные эксперименты по формированию, обоснованию, статистической и экспертной верификации гипотез с помощью БД показали, что для их верификации потребуется увеличить ее объем.

Разработанная методика построения статистической и экспертной верификации гипотез основана на использовании количественного статистического и качественного экспертного методов. Суть статистического метода заключается в следующем. Для каждой языковой единицы из списка потенциально ЛСЕ русского языка определяется число ее переводных эквивалентов в тексте переводов, имеющихся в параллельном корпусе, вычисляются частотности переводных эквивалентов и их разброс. Для определения числа переводных эквивалентов могут использоваться книжные источники, корпуса и БД: лингвист-эксперт анализирует причины разброса частотности переводных эквивалентов и отбрасывает те случаи, когда разброс не связан с лингвоспецифичностью (в частности, он может быть обусловлен различием в способе лексикализации, например русскому *плавать* в английском языке соответствует три разных глагола, обозначающих три различных вида плавания: *swim*, *sail*, *float*). Оставшиеся статистически выявленные лексические единицы считаются гипотетически лингвоспецифичными.

Разработанная методика включает стадию уточнения степени лингвоспецифичности языковой единицы. На этой стадии, в частности, проводится анализ условий появления рассматриваемой единицы русского языка в обратных переводах (т. е. множество «стимулов» перевода на русский язык). Чем больше таких стимулов, тем больше вероятность лингвоспецифичности рассматриваемой единицы.

Категория лингвоспецифичных слов находится в отношении пересечения с категорией безэквивалентной лексики, т. е. имеется множество языковых единиц, относящихся к обеим категориям, но есть и непересекающиеся области. Статистические методы применимы только к тем лингвоспецифичным словам, которые относятся также к категории безэквивалентной лексики. Если слово не принадлежит к этой категории, то применяется качественный экспертный метод построения гипотезы, который включает детальный сопоставительный семантический анализ рассматриваемой лексической единицы и его переводного эквивалента и выявление возможных расхождений в составе компонентов, формирующих их семантическую структуру.

Верификация гипотез лингвоспецифичности языковых единиц выполняется лингвистами-экспертами с использованием БД, сформированной на основе параллельных текстов двуязычного корпуса. Каждая построенная в БД моноэквиваленция включает гипотетически лингвоспецифичную лексическую единицу русского языка и один из ее переводных эквивалентов, найденных в параллельных текстах. (На данном этапе ограничим исследуемый материал, с одной стороны, переводами на французский язык и с другой — лингвоспецифичными личными глагольными формами; в дальнейшем исследовательская база будет расширена в обоих направлениях: по числу языков и по спектру конструкций).

Если сопоставительный статистический и семантический анализ гипотетически лингвоспецифичной лексической единицы и ее эквивалентов позволяет лингвисту-эксперту выявить специфический смысловой компонент, присутствующий в русском слове и отсутствующий в переводе, то лингвоспецифичность этой единицы признается им верифицированной. Особую ценность для нужд

семантического анализа представляют полиэквиваленции, которые предоставляют в распоряжение эксперта данные о границах вариативности перевода интересующей его единицы в контексте, зафиксированном в полиэквиваленции.

Так, в двух переводах фразы *Давно собирался к тебе* из БД (рис. 3), оба французских глагола *s'apprêter* и *se préparer* (буквально 'готовиться') не содержат специфического смыслового компонента неконтролируемости, заключенного в русском глаголе *собираться* (см. [18]) и, наоборот, усиливают, по сравнению с оригиналом, семы приготовления и прилагаемых усилий.

База данных предоставляет в распоряжение пользователя практически полный спектр семантических компонентов, составляющих ту сложную концептуальную конфигурацию, которая заключена, например, в русском глаголе *успеть* (ср. [18–20]), что позволяет уточнить проведенный ранее анализ этого лингвоспецифичного слова (рис. 4). Сопоставление с двумя французскими переводами, где использовано выражение со словом 'время', особенно ясно выявляет эти дополнительные семы, определяющие лингвоспецифичный характер данного русского глагола. В русском оригинале речь идет не столько о возможной нехватке времени, сколько о наклонности и способности, с одной стороны, и о случайности и удаче — с другой; кроме того, в русском глаголе имеется отсутствующая во французском переводном эквиваленте оценочная сема (ср. существительное *успех*).

К каждой единице предварительного списка ЛСЕ русского языка применяется процедура анализа, включающая следующие шаги:

- формирование и выполнение поискового запроса по данной лексеме, который позволяет выявить в БД все включающие ее моно- и полиэквиваленции;
- анализ грамматической составляющей полученных моно- и полиэквиваленций, в том числе статистическое распределение реально встречающихся в узусе грамматических форм;
- анализ лексической составляющей полученных моно- и полиэквиваленций, в том числе статистические параметры выбора переводного эквивалента;
- интерпретация полученных результатов с точки зрения семантического анализа исходной единицы русского языка и оценки степени ее лингвоспецифичности.

Разработанная методология контрастивного корпусного анализа специфических черт семантической системы русского языка предполагает использование уже имеющейся БД глагольных форм, а также построение лексических моноэквиваленций, что планируется осуществить в дальнейшем. При этом базовые виды лексических моноэквиваленций будут определяться включенными в них ЛСЕ.

5 Заключение

Сформированная БД позволила уточнить ряд положений русско-французской контрастивной грамматики. В частности, список соответствий, описанных в работах [11, 12] и частично суммированных в работе [13]:

- инвертирован (в работах Гака и Кузнецовой материал рассматривается в направлении от французского к русскому, так как конечной целью там является интерпретация значения и функции форм французского языка);
- существенно расширен, т. е. установлены новые типы переводных соответствий;
- подвергнут статистической оценке.

Особый интерес представляют полученные результаты частотного анализа переводных соответствий. В частности, корреляция между оппозициями «совершенный *vs.* несовершенный вид» в русском языке и «passe composé/passe simple *vs.* imparfait» во французском может быть уточнена на основе количественных показателей: базовому виду русской ЛГФ Past-IPF лишь в 49,4% случаев соответствует базовый вид французской ЛГФ Imparf и в 21% случаев — PasCom/PasSim; особенно значимой представляется последняя цифра, отражающая широту семантического диапазона русского несовершенного вида.

Разработанная методология, DBParCor-технология и созданная БД, сформированная на основе выровненных текстов поливариантного параллельного корпуса, позволили также уточнить семантику русских глагольных форм: варианты перевода на французский язык, обладающий более детализированной сеткой грамматических противопоставлений в области темпорально-модальных значений, выявляют определенные семантические компоненты, заключенные в значении русских глагольных форм.

В заключение отметим, что разработанная DBParCor-технология может быть адаптирована для использования в других кросслингвистических проектах, целью которых является приведение в соответствие знаний о русском языке современному состоянию лингвистической теории и эмпирической базе, представленной современными электронными корпусами, с одной стороны, и, с

Давно **собирался** к тебе, —	Depuis longtemps je **m'apprêtais** à te rendre visite.
	Il y a déjà longtemps que je **me préparais** à venir te voir,

Рис. 3 Лингвоспецифичная единица *собираться*

Когда это он **успел** опять лечь-то	Quand est-ce qu'il **a trouvé le temps** de se recoucher
	Mais, comment **a**-t-il **eu le temps** de se recoucher?

Рис. 4 Глагол *успеть*

другой стороны, потребностям современной системы образования, а также требованиям, предъявляемым новыми информационными технологиями машинного перевода. Необходимость использования кросслингвистических моделей для разработки технологий машинного перевода была обоснована в работах [21–23].

DBParCor-технология может быть использована в проектах, посвященных изучению на базе параллельных выровненных текстов лексико-грамматических форм других категорий без изменения структуры БД или с небольшими ее изменениями. Для адаптации DBParCor-технологии нужно сформировать перечень используемых языков, определить списки базовых видов ЛГФ и их дополнительных признаков для языков оригинала и перевода в соответствии с целями конкретных проектов.

Литература

1. *Aijmer K., Altenberg B.* Advances in corpus-based contrastive linguistics. Studies in honour of Stig Johansson. — Amsterdam: John Benjamins, 2013. 295 p.
2. *Добровольский Д. О., Кретов А. А., Шаров С. А.* Корпус параллельных текстов // Научная и техническая информация. Сер. 2. Информационные процессы и системы, 2005. № 6. С. 16–27.
3. Корпусные исследования по русской грамматике / Под ред. К. Л. Киселевой, Е. В. Рахилиной, В. А. Плунгяна, С. Г. Татевосова. — М.: Пробел-2000, 2009. 516 с.
4. *Добровольский Д. О.* Корпус параллельных текстов в исследовании культурно-специфичной лексики // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. — СПб.: Нестор-История, 2009. С. 383–401.
5. *Сичинава Д. В., Шведова М. А.* Параллельные корпуса в составе Национального корпуса русского языка: технологии и решаемые задачи // Компьютерная лингвистика: научное направление и учебная дисциплина. — Гомель: ГГУ им. Ф. Скорины, 2010. С. 30–34.
6. *Сичинава Д. В.* Комплексное исследование одноязычного и параллельного корпусов в грамматических исследованиях // Корпусная лингвистика-2011: Труды Междунар. конф. — СПб.: СПбГУ, 2011. С. 316–322.
7. *Сичинава Д. В., Архангельский Т. А.* Параллельные белорусско-русский и русско-белорусский корпуса: совместный проект Национального корпуса русского языка // Труды школы-семинара TEL-2012. — Казань: КФУ, 2012. С. 54–60.
8. *Loiseau S., Sitchinava D. V., Zalizniak A. A., Zatsman I. M.* Information technologies for creating the database of equivalent verbal forms in the Russian-French multivariant parallel corpus // Информатика и её применения, 2013. Т. 7. Вып. 2. С. 100–109.
9. *Добровольский Д. О., Кретов А. А., Шаров С. А.* Корпус параллельных текстов: архитектура и возможности использования // Национальный корпус русского языка: 2003–2005. — М.: Индрик, 2005. С. 263–296.
10. *Андреева Е. Г., Касевич В. Б.* Грамматика и лексика (на материале англо-русского корпуса параллельных текстов) // Национальный корпус русского языка: 2003–2005. — М.: Индрик, 2005. С. 297–307.
11. *Гак В. Г.* Русский язык в сопоставлении с французским. — М.: УРСС, 2006. 264 с.
12. *Гак В. Г.* Сравнительная типология французского и русского языков. — М.: УРСС, 2009. 288 с.
13. *Kouznetsova I. N.* Grammaire contrastive du francais et du russe. — M.: Nestor Academic Publs., 2009. 272 p.
14. *Guiraud-Weber M.* Essais de syntaxe russe et contrastive. — Aix: Université de Provence, 2011. 337 p.
15. *Goldberg A.* Constructions: A Construction Grammar approach to argument structure. — Chicago: Univ. of Chicago Press, 1995. 265 p.
16. *Goldberg A.* Constructions at work. The nature of generalization in grammar. — Oxford: Oxford Univ. Press, 2006. 290 p.
17. Лингвистика конструкций / Под ред. Е. В. Рахилиной. — М.: Азбуковник, 2010. 584 с.
18. *Зализняк Анна А., Левонтина И. Б.* Отражение «национального характера» в лексике русского языка (размышления по поводу книги: *Wierzbicka Anna*. Semantics, culture, and cognition. Universal human concepts in culture-specific configurations. — N.Y., Oxford: Oxford Univ. Press, 1992) // Russian Linguistics, 1996. Vol. 20. No. 2/3. P. 237–264.
19. *Виноградов В. В.* История слов. — М.: Толк, 1994. 1138 с.
20. *Плунгян В. А.* Конструкция с *успеть* и *не успеть* в русском языке XIX–XX вв.: корпусное исследование // Русский язык XIX века: Проблемы изучения и лек-

сикографического описания. — СПб.: Наука, 2004. С. 112–115.

21. *Kozerenko E. B.* Cognitive approach to language structure segmentation for machine translation algorithms // MLMTA'03: Conference (International) on Machine Learning; Models, Technologies and Applications Proceedings. — Las Vegas: CSREA Press, 2003. P. 49–55.
22. *Козеренко Е. Б.* Лингвистические фильтры в статистических моделях машинного перевода // Информатика и её применения, 2010. Т. 4. Вып. 2. С. 83–92.
23. *Kozerenko E. B.* Syntactic transformations modelling for hybrid machine translation // ICAI'11, WORLDCOMP'11 Proceedings. — Las Vegas: CSREA Press, 2011. P. 875–881.

Поступила в редакцию 29.03.14

INFORMATION TECHNOLOGIES FOR CORPUS STUDIES: UNDERPINNINGS FOR CROSS-LINGUISTIC DATABASE CREATION

N. V. Buntman[1], Anna A. Zaliznyak[2,3], I. M. Zatsman[3], M. G. Kruzhkov[3], E. Yu. Loshchilova[3], and D. V. Sitchinava[4]

[1]Faculty of Foreign Languages and Area Studies, M. V. Lomonosov Moscow State University, 31-a Lomonosov Str., Moscow 119192, Russian Federation
[2]Institute of Linguistics, Russian Academy of Sciences, 1-1 Bolshyi Kislovskiy pereulok, Moscow 125009, Russian Federation
[3]Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation
[4]Institute of Russian Language, Russian Academy of Sciences, 18/2 Volkhonka Str., Moscow 119019, Russian Federation

Abstract: Information technology for creation of cross-linguistic databases of Russian texts with French translations (also known as parallel texts) is considered. The underlying principles of the developed database provide a unique combination of three types of bilingual search: lexical, grammatical, and lexico-grammatical. A distinctive feature of the considered technology is simultaneous creation of Russian-French parallel subcorpus within the National Russian Corpus and of the cross-linguistic database of Russian verbal lexico-grammatical forms and their French functional equivalents. The subcorpus and the database have different levels of alignment: the former is aligned at the level of sentences, and the later at the level of constructions. The academic relevance of the developed database is due to its support of bilingual contrastive grammar development, as well as to its role in creation of Russian grammar based on the modern empirical base and information technologies of corpus linguistics. The main practical application of the database consists in improvement of quality of machine translation.

Keywords: parallel corpus; information technology; cross-linguistic databases; bilingual lexical grammar search; corpus linguistics; contrastive grammar

DOI: 10.14357/19922264140210

Acknowledgments

The work was performed in the Institute of Informatics Problems of the Russian Academy of Sciences with financial support of Foundation "Dynasty" (grant NG13-036) and Russian Foundation for Basic Research (grant No. 13-06-00403).

References

1. Aijmer, K., and B. Altenberg. 2013. *Advances in corpus-based contrastive linguistics. Studies in honour of Stig Johansson*. Amsterdam: John Benjamins. 295 p.
2. Dobrovolsky, D. O., A. A. Kretov, and S. A. Sharoff. 2005. Korpus parallel'nykh tekstov [Corpus of parallel texts]. *Nauchnaya i Tekhnicheskaya Informatsiya. Ser. 2. Informatsionnye protsessy i sistemy* [Scientific and technical information. Ser. 2: Information processes and systems] 6:16–27.
3. Kiseleva, K. L., E. V. Rahilina, V. A. Plungjan, and S. G. Tatevosov, eds. 2009. *Korpusnye issledovaniya po russkoy grammatike* [Corpus studies on Russian grammar]. Moscow: Probel-2000. 516 p.
4. Dobrovolsky, D. O. 2009. Korpus parallel'nykh tekstov v issledovanii kul'turno-spetsifichnoy leksiki [A corpus

of parallel texts and studying culture-specific lexicon]. *Natsional'nyy korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy* [Russian National Corpus: 2006–2008. New results and prospects]. St. Petersburg: Nestor-Istoriya. 383–401.

5. Sitchinava, D. V., and M. A. Shvedova. 2010. Parallel'nye korpusa v sostave Natsional'nogo korpusa russkogo yazyka: Tekhnologii i reshaemye zadachi [Parallel corpora of the Russian National Corpus: Technologies and problems]. *Komp'yuternaya lingvistika: Nauchnoe napravlenie i uchebnaya distsiplina* [Computational linguistics: Scientific field and academic discipline]. Gomel': Gomel' University. 30–34.
6. Sitchinava, D. V. 2011. Kompleksnoe issledovanie odnoyazychnogo i parallel'nogo korpusov v grammaticheskikh issledovaniyakh [Comprehensive study of monolingual and parallel corpora in grammatical studies]. *Korpusnaya Lingvistika-2011: Trudy Konferentsii* [Corpus-Based Linguistics-2011 Proceedings]. St. Petersburg. 316–322.
7. Sitchinava, D. V., and T. A. Arhangel'skij. 2012. Parallel'nye belorussko-russkiy i russko-belorusskiy korpusa: Sovmestnyy proekt Natsional'nogo korpusa russkogo yazyka [Parallel Belarusian-Russian and Russian-Belarusian corpora: Joint project of the Russian National Corpus]. School-Seminar TEL-2012 Proceedings. Kazan': Kazan' University. 54–60.
8. Loiseau, S., D. V. Sitchinava, A. A. Zalizniak, and I. M. Zatsman. 2013. Information technologies for creating the database of equivalent verbal forms in the Russian-French multivariant parallel corpus. *Informatika i ee Primeneniya — Inform. Appl.* 7(2):100–109.
9. Dobrovolsky, D. O., A. A. Kretov, and S. A. Sharoff. 2005. Korpus parallel'nykh tekstov: Arkhitektura i vozmozhnosti ispol'zovaniya [Corpus of parallel texts: Architecture and usage]. *Natsional'nyy korpus russkogo yazyka: 2003–2005* [Russian National Corpus 2003–2005]. Moscow: Indrik. 263–296.
10. Andreeva, E. G., and V. B. Kasevich. 2005. Grammatika i leksika (na materiale anglo-russkogo korpusa parallel'nykh tekstov) [Grammar and lexicon in the English-Russian corpus of parallel texts]. *Natsional'nyy korpus russkogo yazyka: 2003–2005* [Russian National Corpus 2003–2005]. Moscow: Indrik. 297–307.
11. Gak, V. G. 2006. *Russkiy yazyk v sopostavlenii s frantsuzskim* [Russian language compared to French]. Moscow: URSS. 264 p.
12. Gak, V. G. 2009. *Sravnitel'naya tipologiya frantsuzskogo i russkogo yazykov* [Comparative typology of French and Russian]. Moscow: URSS. 288 p.
13. Kouznetsova, I. N. 2009. *Grammaire contrastive du francais et du russe*. Moscow: Nestor Academic Publs. 272 p.
14. Guiraud-Weber, M. 2011. *Essais de syntaxe russe et contrastive*. Aix: Université de Provence. 337 p.
15. Goldberg, A. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: Univ. of Chicago Press. 265 p.
16. Goldberg, A. 2006. *Constructions at work. The nature of generalization in language*. Oxford: Oxford Univ. Press. 290 p.
17. Rakhilina, E. V., ed. 2010. *Lingvistika konstruktsiy* [Construction linguistics]. Moscow: Azbukovnik. 584 p.
18. Zaliznjak, Anna A., and I. B. Levontina. 1996. Otrazhenie "natsional'nogo kharaktera" v leksike russkogo yazyka (razmyshleniya po povodu knigi: Anna Wierzbicka. 1992. *Semantics, culture, and cognition. Universal human concepts in culture-specific configurations*. — New York, Oxford: Oxford Univ. Press) [Representation of "national character" in the Russian lexicon (reflections on the book: Anna Wierzbicka. 1992. *Semantics, culture, and cognition. Universal human concepts in culture-specific configurations*. New York, Oxford: Oxford Univ. Press]. *Russian Linguistics* 20:237–264.
19. Vinogradov, V. V. 1994. *Istoriya slov* [History of words]. Moscow: Tolk. 1138 p.
20. Plungjan, V. A. 2004. Konstruktsiya s uspet' i ne uspet' v russkom yazyke XIX–XX vv.: Korpusnoe issledovanie [Constructions with "uspet"' and "ne uspet"' in Russian language in XIX–XX centuries: Corpus-based studies]. *Russkiy yazyk XIX veka: Problemy izucheniya i leksikograficheskogo opisaniya* [Russian language in XIX century: Studies and lexicographical description]. St. Petersburg: Nauka. 112–115.
21. Kozerenko, E. B. 2003. Cognitive approach to language structure segmentation for machine translation algorithms. *MLMTA'03: Conference (International) on Machine Learning; Models, Technologies and Applications Proceedings*. Las Vegas. 49–55.
22. Kozerenko, E. B. 2010. Lingvisticheskie fil'try v statisticheskikh modelyakh mashinnogo perevoda [Linguistic filters for statistical machine translation models]. *Informatika i ee Primeneniya — Inform. Appl.*] 4(2):83–92.
23. Kozerenko, E. B. 2011. Syntactic transformations modelling for hybrid machine translation. *ICAI'11, WORLDCOMP'11 Proceedings*. Las Vegas. 875–881.

Received March 29, 2014

Contributors

Buntman Nadezhda V. (b. 1957) — Candidate of Science (PhD) in philology, associated professor, Faculty of Foreign Languages and Area Studies, M. V. Lomonosov Moscow State University, 31-a Lomonosov Str., Moscow 119192, Russian Federation; nabunt@hotmail.com

Zalizniak Anna A. (b. 1959) — Doctor of Science in philology, leading scientist, Institute of Linguistics, Russian Academy of Sciences, 1-1 Bolshoy Kislovskiy pereulok, Moscow 125009, Russian Federation; Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; anna.zalizniak@gmail.com

Zatsman Igor M. (b. 1952) — Doctor of Science in technology, Head of Department, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; izatsman@yandex.ru

Kruzhkov Mikhail G. (b. 1975) — leading programmer, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; magnit75@yandex.ru

Loshchilova Elena J. (b. 1960) — scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; lena0911@mail.ru

Sitchinava Dmitri V. (b. 1980) — Candidate of Science (PhD) in philology, senior scientist, Institute of the Russian Language, Russian Academy of Sciences, 18/2 Volkhonka Str., Moscow 119019, Russian Federation; mitrius@gmail.com