Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2016»

Москва, 1–4 июня 2016

БАЗА ДАННЫХ МЕЖЪЯЗЫКОВЫХ ЭКВИВАЛЕНЦИЙ КАК ИНСТРУМЕНТ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА[1]

Зализняк Анна А. (anna.zalizniak@gmail.com)

Институт языкознания РАН, Москва;
Институт проблем информатики ФИЦ ИУ РАН, Москва

Ключевые слова: база данных межъязыковых эквиваленций, параллельный корпус, русский язык, французский язык, лингвоспецифичные слова, дискурсивные слова, корпусная лингвистика, контрастивная лингвистика

A DATABASE OF CROSS-LINGUISTIC EQUIVALENCES AS AN INSTRUMENT OF LINGUISTIC ANALYSIS

Zalizniak Anna A. (anna.zalizniak@gmail.com)

Institut of linguistics RAS, Moscow;
Institute of Informatics Problems FRC CSC RAS, Moscow

The paper outlines the principles of creation of a Database of Russian language-specific units and their French equivalents and the possibilities of its use as a tool of linguistic analysis. The entry of the Database is a mono-equivalence (ME), i. e. a dyadic tuple, which consists of a Russian sentence including a language-specific unit and its French translation (automatically extracted from the Russian-French subcorpus of Russian National Corpus), including a functionally equivalent fragment (FEF) of the Russian

[1] Статья написана при финансовой поддержке РФФИ, грант № 16-06-00339.

language-specific unit. Both constituents of the ME are annotated with two-level characteristics, ensuring their faceted classification: “basic type” and “additional feature”. The paper indicates relevant quantitative parameters that can be extracted from such a database and can be accounted for in the analysis of language-specific units; it demonstrates that quantitative methods can be effectively used only in combination with proper methods of semantic analysis. The reliability of statistical data will increase with the extension of the volume of the parallel corpus.

Key words: database of cross-linguistic equivalences, parallel corpora, Russian language, French language, language-specific units, discourse markers, corpus linguistics, contrastive linguistics

В данной статье излагаются принципы создания и возможности использования баз данных межъязыковых эквиваленций, формируемых на основе текстов параллельных корпусов, в качестве инструмента лингвистического анализа — на примере *базы данных лингвоспецифичных единиц русского языка*, созданной в ходе выполнения проекта «Контрастивное корпусное исследование специфических черт семантической системы русского языка», продолжением которого является проект исследования дискурсивных слов русского языка, которое проводится теми же методами и с использованием базы данных аналогичной структуры. Поскольку дискурсивные слова по преимуществу лингвоспецифичны, данный проект является естественным продолжением предыдущего. Основные принципы контрастивного корпусного исследования лингвоспецифичных единиц (ЛСЕ) были изложены в [Зализняк 2015]; в данной статье будут сделаны некоторые уточнения этих принципов, возникшие в ходе формирования базы данных, перечислены дополнительные (по сравнению с параллельными корпусами и другими типами электронных ресурсов) возможности, предоставляемые базой данных межъязыковых эквиваленций, а также продемонстрированы применяемые в наших проектах новые автоматизированные исследовательские технологии.

1. База данных лингвоспецифичных единиц русского языка и их французских функциональных эквивалентов

Как известно, непереводимость, или труднопереводимость — это важнейший отличительный признак лингвоспецифичных единиц. Однако до сих пор утверждения о «труднопереводимости» тех или иных единиц русского языка делались на основании сравнительного семантического анализа (см. в частности, [Зализняк, Левонтина, Шмелев 2005, 2012]), а также отдельных наблюдений над их употреблением, в том числе, частотностью (ср. о сравнительной частотности слов русск. *душа* и англ. *soul* в [Wierbicka 1992]) сопоставляемых единиц разных языков. С появлением корпусов параллельных текстов открылись новые перспективы изучения межъязыковой эквивалентности (ср. [Сичинава 2014, Соколова 2013, Алексеева 2007, Dobrovol’skij 2006, Добровольский 2013:

162–273, Добровольский 2015] и др. — по материалам английского, испанского и немецкого параллельных корпусов).

Однако эффективность контрастивного корпусного исследования существенно возрастает, если сделать еще один шаг, а именно, «надстроить» над параллельным корпусом базу данных, позволяющую фиксировать реально засвидетельствованные соответствия языковых единиц, снабженных характеристикой по множеству признаков. Концепция такой базы данных была создана силами нашего коллектива (см. [Loiseau et al. 2013, Бунтман и др. 2014, Зализняк и др. 2015]), и она была названа «надкорпусной базой данных» (НБД) [Кружков 2015]. Надкорпусная база данных межъязыковых эквиваленций формируется в автоматизированном режиме и предоставляет следующие новые возможности. Во-первых, в НБД для каждого употребления анализируемой языковой единицы в тексте перевода уже выявлен ее «функционально-эквивалентный фрагмент» (ФЭФ)[2]; во-вторых, как анализируемой языковой единице, так и ее ФЭФ приписано значение признаков двухуровневой фасетной классификации, специально разрабатываемой отдельно для каждого типа языковых единиц. Кроме того, в НБД имеется возможность исправления ошибок выравнивания, которая в параллельных корпусах после загрузки текстов отсутствует.

Параллельный корпус, тексты которого используются при формировании надкорпусной базы данных лингвоспецифичных единиц (НБД ЛСЕ), состоит из двух сегментов: русские тексты с их переводом на французский язык (Р-Ф) и французские тексты с их переводом на русский язык (Ф-Р)[3]. Важно подчеркнуть, что в обоих сегментах нас интересуют единицы *русского языка* (список этих единиц составлен на основании Указателя лексем в книге [Зализняк, Левонтина, Шмелев 2012]). Соответственно, в Р-Ф сегменте мы ищем лингвоспецифичные единицы *в текстах оригинала*, а Ф-Р нас интересуют лингвоспецифичные единицы *в текстах перевода*. Перевод с русского языка на французский позволяет выявить те смысловые компоненты интересующей нас русской языковой единицы, которые определили выбор переводчиком того или иного переводного эквивалента. В случае «обратного» перевода в качестве свидетельства о семантике анализируемой единицы русского языка выступают признаки иноязычного текста, послужившие «стимулом» появления интересующего нас русского слова в переводе. При этом часто перевод *на* русский язык оказывается даже более информативным, чем перевод *с* русского языка — поскольку в этом случае лингвоспецифичное слово возникает в переводе как непосредственная реакция на смысловое задание, диктуемое иноязычным текстом (ср. [Шмелев 2015]).

---

[2] Термин введен в [Добровольский и др. 2005]; более распространенный термин «модель перевода» (“translation pattern”), ср. [Hasselgård, Oksefjell (eds.) 1999, Сичинава 2014] имеет несколько более узкое значение.

[3] На момент 15.02.16 сегмент Р-Ф имел общий объем 2 104 900 словоупотреблений, 141 936 предложений; сегмент Ф-Р имел общий объем 470 458 словоупотреблений, 27 055 предложений. Работа по увеличению объема французского параллельного корпуса продолжается.

Возможная лингвоспецифичность единиц французского языка нас в этом проекте не интересует: французский язык в обоих направлениях перевода служит не объектом, а лишь инструментом анализа, т. е. способом выявления скрытых семантических компонентов, содержащихся в значении анализируемых русских языковых единиц. В этом смысле наш метод контрастивного анализа был назван *унидирекциональным*. Это, однако, не означает, что создаваемая согласно этим принципам база данных не может служить различным переводческим и переводоведческим целям, а также целям сопоставительной лексикологии. Надкорпусная база данных лингвоспецифичных единиц русского языка является, тем самым, одновременно результатом и инструментом лингвистического анализа.

Входом НБД ЛСЕ является *моноэквиваленция* (термин был введен в [Loiseau et al. 2103]). Моноэквиваленция (в сегменте Р-Ф) — это двухместный кортеж (упорядоченная пара) вида: фрагмент текста на языке оригинала, содержащий интересующую нас единицу русского языка — ее функционально-эквивалентный фрагмент (ФЭФ) в тексте французского перевода данного фрагмента. Моноэквиваленции автоматически объединяются в *полиэквиваленции* — в тех случаях, когда в БД имеется несколько переводов одного и того же исходного текста; ценность полиэквиваленции как инструмента анализа состоит в том, что она показывает варианты перевода языковой единицы *в одном и том же контексте*. Моноэквиваленции, построенные на основе Ф-Р сегмента корпуса имеют вид: «стимул перевода» (СП)[4], т. е. фрагмент французского текста, «реакцией» на которой служит появление в русском переводе интересующей нас языковой единицы — русский перевод. (Примеры моно- и полиэквиваленций см. ниже.)

В НБД ЛСЕ разработана двухуровневая система аннотирования языковых единиц, входящих в моноэквиваленции: «базовый вид» и «дополнительные признаки». Базовый вид — это классифицирующая категория; каждая анализируемая языковая единица принадлежит одному одноименному «базовому виду» (напр. слово *беда* — базовому виду *беда* и т. д.); базовые виды группируются в кластеры на основании принадлежности к определенной части речи.

Второй уровень — «дополнительные признаки» — обеспечивает фасетность классификации: дополнительные признаки различны по своей природе, и каждой анализируемой языковой единице приписываются значения нескольких признаков. Дополнительные признаки объединены в кластеры: в первом кластере помещены морфологические характеристики анализируемой языковой единицы, во втором — сведения о присутствующих в предложении зависимых элементах. В третьем кластере содержатся сведения о «конструкции» (в понимании Грамматики конструкций, т. е. сюда входит вся сочетаемость и идиоматика в широком смысле); например: *на* [*беду*]; [*беда*] *если*; [*беда*] *в том, что*; *жди* [*беды*] и т. п. В четвертом кластере помещаются характеристики типа предложения, в котором употреблена анализируемая языковая единица (вопросительное, восклицательное, побудительное, отрицательное; диалогическая реплика и др.).

---

[4] Термин «стимул перевода» в том значении, в котором он применяется в наших исследованиях, введен в [Loiseau et al. 2013].

Именно наличие этого второго уровня аннотирования обеспечивает новое качество информации, предоставляемой надкорпусной базой данных, по сравнению с параллельными корпусами, а также, например, такими переводческими ресурсами как Multitran или Linguee: НБД позволяет зафиксировать для исследуемой языковой единицы функционально эквивалентный фрагмент в тексте перевода (в том числе — грамматической природы, в том числе — нулевой) с учетом формально охарактеризованных признаков типа ее употребления.

Несколько иначе устроена система аннотирования единиц французского текста (т. е. ФЭФ). Главное отличие состоит в том, что список «базовых видов» — открытый, он составляется в процессе построения моноэквиваленций. Базовые виды ФЭФ сортируются по категориям: это основные части речи плюс следующие категории: Alia (сюда попадают слова всех остальных частей речи), Compositа (многокомпонентные единицы), Sentential TS (ФЭФ, представляющие собой предложения, напр. *qu'est-ce que cela fait*; *on sait jamais* и т. п.), Grammatical TS (грамматические средства передачи — например, значение русского глагола *успеть* может быть передано французской формой времени Passé antérieur).

Дополнительные признаки фасетной классификации единиц французского текста группируются в два кластера: Collocation и Grammatical. В кластер Collocation попадают словосочетания, содержащие некоторое слово из списка «базовых видов». Например, если в списке «базовых видов» имеется единица *coeur*, то словосочетание *au fond du coeur* попадет в кластер Collocation в виде *au fond du* [coeur]. В кластере Grammatical находятся грамматические характеристики ФЭФ.

Кроме того, некоторые признаки могут быть приписаны самой моноэквиваленции (изменение конструкции предложения по сравнению с оригиналом, необходимость учета контекста более широкого, чем данное предложение, существенное расхождение в способе лексикализации и др.)

Так, например, во фразе

> …во мне был заперт свет, который искал выхода, но только жег свою тюрьму, **не вырвался на волю** и угас (Гончаров. Обломов)

слово *воля* принадлежит базовому виду *воля* (кластер «существительное») и получает следующие значения дополнительных признаков: Sg (кластер 1); *вырваться на* [*волю*] (кластер 3), Neg (кластер 4).

Во французском переводе этой фразы находим ФЭФ для всего словосочетания (выделено жирным):

> …une lumière emprisonnée en moi cherchant une issue ne faisait que consumer sa prison et a fini par s éteindre **sans jamais recouvrer sa liberté.**

Базовый вид ФЭФ — *liberté*, **дополнительные признаки**: *recouvrer sa* [liberté] (**кластер** Collocation); *sans* INF (кластер Grammatical).

Соответствующая моноэквиваленция выглядит следующим образом (в первом столбце указан ее номер, во вором — шифр текста оригинала, в третьем — фрагмент текста, содержащий анализируемую ЛСЕ, в котором выделены сама

ЛСЕ плюс ее релевантный контекст, в четвертом — базовый вид и доп. признаки данной ЛСЕ в данном предложении, в пятом — извлеченный из параллельного корпуса фрагмент перевода, в котором выявлен ФЭФ, в шестом — базовый вид и доп. признаки ФЭФ перевода):

604	ГОБ	во мне был заперт свет, который искал выхода, но только жег свою тюрьму, **не вырвался на волю** и угас.	**воля** < Neg > < Sg > < вырваться на [волю] >	une lumière emprisonnée en moi cherchant une issue ne faisait que consumer sa prison et a fini par s éteindre **sans jamais recouvrer sa liberté**	**liberté** < recouvrer sa [liberté] > < sans INF >

Другие примеры моноэквиваленций из сегмента Р-Ф:

892	ДПН	Нет, то досадно, что врут, да еще собственному **вранью** поклоняются.	**вранье** < Sg >	Non, ce qui est fâcheux, c'est qu'ils se trompent et qu'ils admirent pardessus le marché leurs propres **erreurs**.	**erreur** < Pl >

197	ДПН	**всё** гимнастикой **собираюсь лечиться**;	**собираться** < SubInf-IPF > < Pers1 > < V-IPF > < Pres > < всё >	**je me propose** toujours **de les soigner** par la gymnastique;	**se proposer** < SubInf >

Пример полиэквиваленции:

133	ГОБ	Обломову [...] хотелось бы, чтоб было чисто, [...] он [...] желал, **чтоб это сделалось как-нибудь так**, незаметно, само собой;	**как-нибудь** < так >	Oblomov eût [...] apprécié la propreté, mais à condition qu'elle s'installât d'elle-même, sans qu'il s'en aperçoive.	**ZERO**
171	ГОБ	Обломову [...] хотелось бы, чтоб было чисто, [...] он [...] желал, **чтоб это сделалось как-нибудь так**, незаметно, само собой;	**как-нибудь** < так >	Oblomov aurait [...] voulu que tout devînt propre [...] que la chose se fit insensiblement, et comme allant de soi.	**ZERO**

Примеры моноэквиваленций из сегмента Ф-Р:

2280	B99	pour donner une raison à sa présence à cette PPM (**et par extension** au sein de la société Madone)	**par extension**	чтобы оправдать свое присутствие на этом PPM (**а заодно и** в самой фирме «Манон»)	**заодно** < Adv >
2507	BIG	il lui arriva toujours, […] de se mettre en fureur à cette observation,	**ZERO**	случалось все же, что он **вдруг** свирепел от этого замечания,	**вдруг** < Adv >
2614	MLH	et pleins de cloches qui sonnent **dans l'air bleu** des belles matinées	**air** < dans l'[air] bleu >	и там множество колоколов, которые звонят **в голубом просторе** прекрасного утреннего часа	**простор** < Sg > < в голубом [просторе]>

2. Корпусные методы анализа лингвоспецифичных единиц

В ходе работы с НБД ЛСЕ были выделены следующие пять основных типов отсутствия межъязыкового семантического изоморфизма; каждый тип обозначается реализующим его русским словом:

I. Асимметричное членение концептуальной области:
   a. подтип ЗНАТЬ (русск. *знать* vs. франц. *connaître — savoir*)
   b. подтип ПРАВДА-ИСТИНА (русск. *правда–истина* vs. франц. *verité*);
II. Тип САМОВАР (переводной эквивалент отсутствует; используется заимствованное слово или описательное определение);
III. Тип БАБУШКА (имеется один преимущественный вариант перевода, но он неточный);
IV. Тип РОДНОЙ (переводной эквивалент отсутствует; имеется несколько приблизительно равновероятных вариантов перевода, все неточные);
V. Тип РАЗЛУКА (имеется один преимущественный переводной эквивалент, имеющий более точное соответствие в русском языке): *разлука — séparation* (ср. *расставание*); *беда — malheur* (ср. *несчастье*); *грозный — menaçant* (ср. *угрожающий*); *тоска — angoisse* (*ср. тревога*).

С точки зрения возможностей применения количественных методов наибольший интерес представляют три последние категории, именно о них и будет идти дальше речь. Можно назвать следующий ряд характерных признаков, указывающих на вероятную лингвоспецифичность слова, которые могут быть

установлены при помощи автоматизированных процедур, осуществляемых в НБД, «надстроенной» над параллельным корпусом. Перечислим эти признаки (их список и формулировки существенно уточнены по сравнению с теми, которые были приведены в [Зализняк 2015]).

А именно, на возможную лингвоспецифичность слова указывает:

- при переводе с русского языка:

(1) наличие большого количества ФЭФ, в том числе — имеющих приблизительно равную частотность (ср. ниже);

(2) наличие многокомпонентных ФЭФ (ср.: *обидно — en avoir gros sur le coeur; родной — membre de la famille*), а также ФЭФ, состоящих из двух квазисинонимов; ср. для слова *грозный*:

> *И он, как* ***грозный*** *учитель, глядел на прячущегося ребенка —*
> *Il le regardait comme un maître* ***sévère*** *regarde,* ***menaçant****,*
> *un enfant qui se cache.* (Гончаров. Обломов);

для слова *обидно*:

> должно быть, ей очень ***обидно***… — Sans doute se sentait-elle très ***malheureuse***, très ***déçue*** (Гончаров. Обломов);

(3) слово остается без перевода (модель перевода — ZERO); ср. для слова *душа*:

> *Впрочем, он был* ***в душе*** *добрый человек* (Гоголь. Шинель) —
> *C'était pourtant un brave homme*;

- при переводе на русский язык:

(4) большое количество «стимулов перевода» (СП), в том числе, имеющих приблизительно равную частотность; напр. для слова *беда* из 17 примеров в 15 имеются разные СП;

(5) наличие многокомпонентных СП; напр. для появления слова *душа* в русском переводе имеются следующие СП во французском оригинале: *au fond de l'âme, au fond du coeur, dans la serenité, par un geste naturel, autant qu'ils veulent* и еще 10 различных многокомпонентных СП;

(6) отсутствие какой-либо единицы, которая «стимулирует» появление анализируемой русской ЛСЕ (стимул перевода — ZERO), ср. для слова *родной*:

> *Paris redevenait* […] *ma ville* (P. Modiano. Quartier perdu). *Париж вновь становился* […] ***моим родным городом***.

Посмотрим теперь, как ведет себя слово *тоска* относительно перечисленных выше признаков лингвоспецифичности — на основании данных, полученных из БД ЛСЕ. Напомним, что *тоска* — одно из самых известных лингвоспецифичных русских слов, см. в частности [Wierzbicka 1992, Шмелев 2002][5].

Статистическая таблица, генерируемая базой данных (сегмент Р-Ф), выглядит следующим образом (цифра напротив слова указывает на количество примеров, в первом столбце указан французский переводной эквивалент, во втором — сколько раз он встретился, в третьем — какой процент от общего количества ФЭФ для данного русского слова это составляет):

тоска	111

ФЭФ французского языка	**Кол-во МЭ**	**% в группе**
angoisse	45	40,18%
détresse	19	16,96%
tristesse	8	7,14%
ennui	5	4,46%
angoissé	3	2,68%
nostalgie	3	2,68%
chagrin	3	2,68%
ZERO	2	1,79%
mélancolie	2	1,79%
désespoir	2	1,79%
douleur	1	0,89%
tourment	1	0,89%
triste	1	0,89%
mélancolique	1	0,89%
accablement	1	0,89%
s'ennuyer	1	0,89%
hypocondrie	1	0,89%
stupeur	1	0,89%
inquétude	1	0,89%
idées noires	1	0,89%
déprimé	1	0,89%
déprime	1	0,89%
se sentir triste	1	0,89%
angoisser	1	0,89%
ennui, tristesse	1	0,89%
tristesse et nostalgie	1	0,89%
ennui teinté d'affliction	1	0,89%
ennuyer	1	0,89%
désolé	1	0,89%

[5] В работе [Соколова 2013] обнаружено 24 модели перевода на испанский язык слова тоска в рассказах А. П. Чехова; в работе [Сичинава 2014] обнаружено 22 модели перевода этого слова на английский в англо-русском подкорпусе НКРЯ.

Признак (1) — количество разных ФЭФ — 27: соотношение наиболее частотных моделей перевода: 40%, 17%, 7%, 4,5%; очевидно преобладание одного — *angoisse*.

Признак (2) — наличие многокомпонентных ФЭФ:

> *…не столько от боли,* [...] *сколько от* ***тоски*** *— pas tellement sous l'effet de la douleur physique* [...] *que* ***d'ennui, de tristesse****..;* (Л. Толстой. Смерть Ивана Ильича);
>
> *всё* ***тоска — tristesse et nostalgie*** *sans fin* (Л. Толстой. Смерть Ивана Ильича,);
>
> *Тупая* ***тоска*** *— Cette* ***sorte de stupeur*** *obtuse et* ***chagrine*** (Л. Толстой. Смерть Ивана Ильича);
>
> ***Тоска*** *проглянула в лице Лужина — Une expression* ***d'ennui teinté d'affliction*** *passa sur le visage de Loujine.* (Достоевский. Преступление и наказание).

Признак (3): Модель перевода — ZERO представлена лишь в двух случаях.

В Ф-Р сегменте слово *тоска* встречается всего 11 раз, там преобладание варианта *angoisse* еще более существенное, но общее количество примеров слишком мало для каких-либо статистических выводов.

Следует упомянуть еще один существенный количественный параметр, который может быть извлечен из НБД ЛСЕ: он касается преимущественного переводного французского эквивалента; соответственно:

а) в сегменте Р-Ф: из каких русских «стимулов перевода» возникает выявленный преимущественный эквивалент во французском языке;

б) в сегменте Ф-Р: как переводится этот преимущественный эквивалент на русский язык.

Про преимущественный эквивалент для русского *тоска* — франц. *angoisse* — пока можно сказать, что в Р-Ф сегменте корпуса оно встречается 89 раз (из них 45, т. е. половина, соответствуют в русском слову *тоска*). В Ф-Р сегменте корпуса *angoisse* встречается 17 раз, из них оно переведено на русский язык словом *тоска* 8 раз, кроме того имеются варианты *тоскующий*, *тоскливо*, *тоскливая тревога*.

Из всего этого, как представляется, можно сделать вывод, что русское слово *тоска* и франц. *angoisse* имеют довольно существенную область совпадения. Другими словами, лингвоспецифичность слова *тоска* в паре «русский — французский» достаточно низкая. Эти данные, конечно же, должны быть верифицированы на большем материале.

Лингвоспецифичность существительного *обида*, а также глаголов *обидеться*, *обижаться* и в особенности предикатива *обидно* много обсуждалась

в литературе (см. [Зализняк 2000, Dobrovolskij 2006, Протасова 2006, Апресян 2011]. В [Апресян 2011] приводятся примеры из Набоковской «Лолиты» в английском оригинале и авторском переводе на русский, где русское слово *обида* используется как переводной эквивалент пяти несинонимичных английских слов, а также примеры перевода глаголов *обидеться, обижаться* с русского на английский, выполненный билинвами (8 типичных контекстов употребления; все переводятся по-разному).

В НБД ЛСЕ в сегменте Р-Ф слово *обида* встретилось 56 раз. Представлено 27 вариантов перевода, из них 19 встречаются по одному разу; в 9% случаев оставлено без перевода; наиболее частотный перевод (*offense*) составляет 23%, т. е. меньше четверти случаев.

Если взять нелингвоспецифичное слово *любовь*, то картина будет разительно отличаться, а именно, в 85% случаев употребления данного слова в переводе ему соответствует одно и то же слово — *amour*.

Дискурсивное слово *авось* в 43% случаев переведено как *peut-être* (т. е. с утратой некоторой части смысла); следующий по частотности вариант — ZERO (23%); далее следуют варианты перевода, в которых отражено представление о желательности и малой вероятности обсуждаемого события (*espérer* — 10%, *avec un peu de chance*, букв.: 'если повезет' — 10%) и несколько единичных переводов, по-разному выражающих идею надежды на нечто маловероятное и непредсказуемость будущего: *peut-être bien, au petit bonheur, on sait jamais*. Можно сказать, что в целом множество вариантов перевода достаточно полно отражает набор семантических компонентов этого слова, лингвоспецифичность которого возникает за счет *комбинации* нескольких компонентов: непредсказуемости хода вещей, легкомысленного расчета на осуществление желаемого положительного события и, одновременно, равнодушия к возможному провалу (о слове *авось* см. в частности [Wierzbicka 1992: 433–435]).

Для дискурсивного слова *как-нибудь* показателен, прежде всего, факт преобладания варианта ZERO (37%), а также общее количество разных вариантов перевода, в том числе, встретившихся по одному разу. Это множество вариантов перевода рисует богатую картину смысловых компонентов и возможных вариантов значения русского слова *как-нибудь* — одновременно подтверждая его лингвоспецифичность, определяемую, прежде всего уникальной комбинацией этих компонентов в одном типе употребления.

3. Заключение

Из сказанного могут быть сделаны следующие выводы.

1. Надкорпусные базы данных представляют собой потенциально весьма эффективный инструмент, который может быть использован:
   - для лингвистического анализа тех или иных единиц одного языка;
   - для проведения контрастивных исследований;
   - для совершенствования систем перевода — как, автоматического, так и авторского;

2. Количественные методы установления меры лингвоспецифичности существуют, но эффективно они могут применяться лишь в комбинации с методами собственно семантического анализа.
3. Для того чтобы результаты такого рода исследований были значимыми, нужны корпуса значительно большего объема (т. е. объем параллельного корпуса должен быть увеличен на один-два порядка, при этом он должен быть сбалансированным в отношении времени создания и жанра входящих в него текстов);
4. За те 150–200 лет, которые отделяют нас от эпохи «русской классической литературы» в русском языке произошли очень существенные семантические сдвиги, в особенности для тех языковых единиц, которые сегодня являются лингвоспецифичными: в подавляющем большинстве случаев в XIX веке их значение было иным, и оно не было лингвоспецифичным; так, почти все обсуждавшиеся выше слова претерпели существенную эволюцию в этом направлении: *тоска* (ср. *тоска за жизнь* в «Обломове»), *разлука, совестно, авось* и др. — ср. следующие два примера употребления выражения *без обиды*:

> …да так, что ни с того ни с сего сгреб кирпич и кинул в начальника, *безо всякой обиды* с его стороны (Достоевский. Преступление и наказание).

> Жил папа *без обиды*, он считал, что это время было такое. (С. Алексиевич. Время сэконд-хенд).

Факт увеличения лингвоспецифичности в ходе семантической эволюции уже отмечался исследователями (см., напр., [Шмелев 2001]), однако использование базы данных, надстроенной над параллельным корпусом, может здесь дать результаты более высокого уровня надежности. Таким образом, при должном расширении состава текстов параллельного подкорпуса НКРЯ (в котором на настоящий момент преобладают тексты XIX в.) диахронический аспект изучения лингвоспецифичных единиц также получит дополнительную экспериментальную основу.

Автор благодарит анонимных рецензентов, чьи ценные замечания были по возможности учены в окончательной версии статьи.

Литература

1. *Алексеева М. Л.* (2007), Русские реалии в разновременных немецких переводах романов Ф. М. Достоевского. Словарь-справочник. Екатеринбург.
2. *Апресян В. Ю.* (2011), Опыт кластерного анализа: русские и английские эмоциональные концепты. Часть 1 // ВЯ, №1, 2011.
3. *Бунтман Н. В., Зализняк Анна А., Зацман И. М., Кружков М. Г., Лощилова Е. Ю., Сичинава Д. В.* (2014), Информационные технологии корпусных

исследований: принципы построения кросс-лингвистических баз данных // Информатика и ее применения. Т. 8, вып. 2. С. 98–110.

4. *Добровольский Д. О.* (2013), Беседы о немецком слове. М.: Языки славянской культуры, 2013.
5. *Добровольский Д. О.* (2015), Корпус параллельных текстов и сопоставительная лексикология // Труды института русского языка им. В. В. Виноградова. Вып. 6. М., 2015. С. 411–446
6. *Добровольский Д. О., Кретов А. А., Шаров С. А.* (2005), Корпус параллельных текстов: архитектура и возможности использования. // Национальный копус русского языка: 2003–2005. Результаты и перспективы. М., 2005. С. 263–296.
7. *Зализняк Анна А.* (2000), О семантике щепетильности (*обидно*, *совестно* и *неудобно* на фоне русской языковой картины мира) // Логический анализ языка. Языки этики. М., 2000. С. 101–118.
8. *Зализняк Анна А.* (2015), Лингвоспецифичные единицы русского языка в свете контрастивного корпусного анализа // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2015. М., 2015. С. 651–662.
9. *Зализняк Анна А., Левонтина И. Б., Шмелев А. Д.* (2005), Ключевые идеи русской языковой картины мира. Москва: Языки славянской культуры. М.
10. *Зализняк Анна А., Левонтина И. Б., Шмелев А. Д.* (2012), Константы и переменные русской языковой картины мира. Москва: Языки славянской культуры. М.
11. *Зализняк Анна А., Зацман И. М., Инькова О. Ю., Кружков М. Г.* (2015), Надкорпусные базы данных как лингвистический ресурс // Труды международной конференции» Корпусная лингвистика-2015». 22–26 июня 2015, Санкт-Петербург. СПб, 2015. С. 211–218.
12. *Кружков М. Г.* (2015), Информационные ресурсы контрастивных лингвистических исследований: электронные корпуса текстов // Системы и средства информатики. Т. 25. № 2. С. 140–159.
13. *Протасова Е.* (2006), Какого вкуса *обида*? // Integrum: точные методы и гуманитарные науки. Г. Нечипорец-Такигава (ред.) М.: «Летний сад», 2006. С. 270–288.
14. *Сичинава Д. В.* (2014), Использование параллельного корпуса для количественного изучения лингвоспецифичной лексики // Язык, литература, культура: Актуальные проблемы изучения и преподавания. Вып. 10. М., МАКС ПРЕСС, с. 37–44
15. *Соколова Л. В.* (2013), Способы передачи безэквивалентной лексики в переводах А. Чехова на испанский язык (на материале концепта «тоска») // Rafael Guzmán Tirado, Irina A. Votyakova (ed.). Tipología léxica. Granada, 2013. С. 191–196.
16. *Шмелев А. Д.* (2001), Некоторые тенденции семантического развития русских дискурсивных слов. // Русский язык: пересекая границы. Дубна, 2001. С. 266–279.
17. *Шмелев А. Д.* (2002), Русская языковая модель мира. Опыт словаря. М.: Языки славянской культуры.

18. *Шмелев А. Д.* (2015), Русские лингвоспецифичные лексические единицы в параллельных корпусах: возможности исследования и «подводные камни» // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2015. М., 2015.
19. *Dobrovol'skij D.* (2006), Zur kontrastiven Analyse kulturspezifischer Konzepte // Wörter-Verbindungen. Festschrift Jarmo Korhonen zum 60. Geburtstag. Hrsg. von Ulrich Breuer und Irma Hyvärinen. Frankfurt am Main etc.: Peter Lang, 2006. S. 31–45.
20. *Dobrovol'skij D., Pöppel L.* (2015), Corpus perspectives on Russian discursive units: semantics, pragmatics, and contrastive analysis // Yearbook of Corpus Linguistics and Pragmatics 2015. New York etc.: Springer, 2015, pp. 223–241.
21. *Hasselgård, H., Oksefjell, S.* (eds.) (1999), Out of Corpora: Studies in Honour of Stig Johansson. Amsterdam — Atlanta, GA: Rodopi.
22. *Kruzhkov M., Buntman N. V., Loshchilova E. J. Sitchinava D. V., Zalizniak Anna A., Zatsman I. M.* (2014), The database of Russian verbal forms and their French translation equivalents // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2014. С. 275–287.
23. *Loiseau S., Sitchinava D. V., Zalizniak Anna A., Zatsman I. M.* (2013), Information technologies for creating the database of equivalent verbal forms in the Russian-French multivariant parallel corpus // Информатика и ее применения, 2013. Том 7, вып. 2. С. 100–109.
24. *Wierzbicka A.* (1992), Semantics, Culture, and Cognition. Universal Human Concepts in Culture-Specific Configurations. N. Y.; Oxford: Oxford Univ. Press.