МИХАИЛ МИХАЙЛОВ

Параллельные корпуса художественных текстов

принципы составления и возможности применения в лингвистических и переводоведческих исследованиях

на примере русско-финского параллельного корпуса художественных текстов

■

ACADEMIC DISSERTATION

To be presented, with the permission of
the Faculty of Humanities of the University of Tampere,
for public discussion in the Pinni auditorium B 1097
of the University, Kanslerinrinne 1, Tampere,
on September 27th, 2003, at 12 o'clock.

Acta Universitatis Tamperensis 956
University of Tampere
Tampere 2003

ACADEMIC DISSERTATION
University of Tampere, The School of Languages and Translation Studies
Finland

Distribution

University of Tampere
Bookshop TAJU
P.O. Box 617
33014 University of Tampere
Finland

Tel. +358 3 215 6055
Fax +358 3 215 7685
taju@uta.fi
http://granum.uta.fi

Cover design by
Juha Siro

Printed dissertation
Acta Universitatis Tamperensis 956
ISBN 951-44-5753-6
ISSN 1455-1616

Electronic dissertation
Acta Electronica Universitatis Tamperensis 280
ISBN 951-44-5754-4
ISSN 1456-954X
http://acta.uta.fi

Tampereen yliopistopaino Oy Juvenes Print
Tampere 2003

ОГЛАВЛЕНИЕ

English abstract

Parallel corpora of literary texts: principles of compilation and possible uses in linguistics and translation studies (as applied to a Russian-Finnish parallel corpus of literary texts)

The aim of the dissertation was to compile a Russian-Finnish parallel corpus of literary texts and to study possible applications of its data in linguistics and translation studies. The *ParRus* corpus includes classical as well as modern Russian literary texts and their translations into Finnish. The size of the Corpus is about 2.2 million running words in each language. Full texts rather than samples were collected, which is important for the use of the Corpus in translation studies and humanities. Different authors, translators, genres, and chronological periods are represented. To some extent, the standard principles of linguistic text corpora compilation were revised.

A software package *COCOS-P* was developed to study the texts of the Corpus. It includes a paragraph-level aligner, Russian and Finnish lemmatizers, a parallel concordancer, a collocation list builder, etc. Furthermore, a special tool for searching possible translation equivalents was designed. The Corpus is stored in a Microsoft Access database and the software is written in Microsoft Visual Basic programming language.

A number of case studies of *ParRus* texts was carried out. It was checked whether translations really tend to be longer than source texts, as suggested by Eugene Nida, applying methods of the theory of information. The research shows that such a tendency really does exist.

Another study topic was whether the language of translations differs from that of texts originally written in the same language. The language of Finnish translations from Russian was compared to the language of the texts of the subcorpus of original literary Finnish texts from *The Savonlinna Corpus of Translated Finnish*. Differences in lexicon, use of certain grammatical constructions and punctuation were studied, using statistical methods. In most cases, the differences found can be explained by the influence of the source language, which implies that the language of translations from different source languages is different from that of original texts.

Finally, Finnish translation equivalents for some Russian words were studied. The data from both parallel texts and bilingual dictionaries was considered. Equivalents used by translators were often different from those suggested by dictionaries. However, certain influence of bilingual dictionaries on translators

can also be traced. Sometimes translators use old-fashioned and even erroneous equivalents; e.g. *ylioppilas* ‘high school graduate’ is used as an equivalent for *student* ‘university student’ more often than *opiskelija* ‘student’, which would serve the purpose much better.

Parallel corpora of literary texts seem to be a valuable source of data for studying culturally-bound words. Some Russian words, which are believed to be important for the Russian culture and Russian way of thinking, were studied on the Corpus material. It was discovered that, although such words sometimes are difficult to translate, it is still the text and not a single word that is translated, and, in most cases, the translator can find a solution on the context level. The core of the problem is whether the translator understands the text correctly.

The results of the research presented in the dissertation are by no means final. The work on *COCOS-P* software will continue. There are certain search and query functions, which would be useful to develop; e.g. part-of-speech and grammar-form tagging would be very important for various research tasks, and sentence-level aligning of the texts would have produced more accurate results. Development of the software will help to carry out a more consistent analysis of the Corpus. Nevertheless, most of the results achieved in this first stage of research would have been impossible or very difficult to obtain using traditional methods.

Предисловие

Настоящая диссертация подводит итог моей четырехлетней исследовательской работы на Отделении переводоведения Тамперского университета. Написание большой работы в относительно сжатые сроки было бы невозможным без наличия определенного исследовательского опыта. Важную роль в его приобретении сыграло обучение в аспирантуре Московского лингвистического университета, написание и защита кандидатской диссертации под руководством проф. Б.Ю. Городецкого, а затем — работа на кафедре лингвистической семантики МГЛУ. Там я написал свои первые компьютерные программы. Очень полезным для меня в то время было общение с Дэрилом Гиббом, исследователем из США, работавшим по контракту на нашем отделении. Именно Дэрил познакомил меня с принципами работы с электронными текстовыми массивами.

Важным этапом в моей исследовательской биографии стала работа в Отделе экспериментальной лексикографии Института русского языка им. В.В. Виноградова РАН. Участие в словарных проектах требовало работы с большими текстовыми массивами, создания баз данных и оболочек для их обработки. Атмосфера творческого поиска и хорошее техническое оснащение Отдела уже тогда позволяли получать интересные результаты. Много дало мне сотрудничество с проф. А.Н. Барановым и проф. Д.О. Добровольским, которое продолжается и по сей день.

Однако по-настоящему заняться работой с корпусами текстов и параллельными текстами удалось только в Тамперском университете. Многое из того, что ранее существовало лишь в виде «бесцветных, прозрачных идей», удавалось реализовывать в невероятно короткие сроки. И это оказалось возможным в первую очередь благодаря моему научному руководителю проф. Ханну Томмола, который не только проявлял интерес к моей работе и создал все условия для ее выполнения, но и фактически сделал корпусную лингвистику одним из приоритетных направлений исследовательской деятельности кафедры. Без ценных советов, доброжелательной критики, дружеского участия и поддержки Ханну эта работа вряд ли была бы когда-нибудь написана.

Я благодарен Тамперскому университету за предоставленную мне исследовательскую стипендию (июль — декабрь 2000 г.), что позволило начать работу над диссертацией. Я благодарен моим рецензентам, проф. Ларсу Борину (Гетеборгский университет, Швеция) и проф. Игорю Богу-

славскому (Институт проблем передачи информации РАН), за внимательное чтение моей рукописи и ценные замечания.

Отдельная благодарность проф. Анне Мауранен (Тамперский университет, Школа современных языков и переводоведения), давшей разрешение на использование в моей работе Савонлиннского корпуса текстов (*The Savonlinna Corpus of Translated Finnish*. Savonlinna School of Translation Studies, University of Joensuu, 2001) и проф. Пекке Песонену и Бену Хелмману (Хельсинкский университет, кафедра русского языка и литературы) за библиографию переводов с русского языка на финский.

Благодарю за поддержку моих коллег с кафедры русского перевода: Арто Лехмускаллио, Катриину Муурайнен, Хейкки Энквиста, Светлану Пробирскую-Турунен и Пяйви Кууси.

Большое значение для моей работе имело общение с соратниками по прикладной лингвистике и автоматизированной обработке текстов: Григорием Сидоровым (Национальный политехнический институт, Мехико, Мексика), Михаилом Копотевым и Игорем Кудашевым (Хельсинкский университет, Финляндия), Ханно Бибером, Эвелин Брайтенедер и Карлхайнцем Мёртом (Австрийская академия наук), Галиной Денисовой (Пизанский университет, Италия) и Натальей Кузиной (Смоленский педагогический университет, Россия). Спасибо вам всем и успехов вам в нашем тяжелом деле!

Спасибо моим студентам за интерес к работе с параллельными текстами, трудные вопросы (ответил почти на все!) и за готовность часами выслушивать консультации.

Спасибо моим родным — жене Наталье, детям Алексею и Александре, и конечно же, родителям Николаю Николаевичу и Марине Михайловне Михайловым — за веру в меня, терпение и поддержку в трудную минуту.

Тампере, август 2003.

Михаил Михайлов

Введение

Одной из интенсивно развивающихся областей современной корпусной лингвистики является развитие многоязычных ресурсов. О важности этого направления говорит, в частности, проведение целого ряда международных симпозиумов и конференций, специально посвященных многоязычным корпусам текстов, например, симпозиум в Уппсальском университете (Швеция) в 1999 г. и 6-й семинар ассоциации TELRI в Банско (Болгария) в 2001 г. На большинстве конференций и симпозиумов, посвященных корпусной лингвистике, в той или иной мере затрагивались проблемы работы с многоязычными и параллельными корпусами текстов. Причина состоит в том, что, несмотря на активную работу по созданию многоязычных корпусов текстов, их доля по сравнению с одноязычными корпусами текстов остается весьма малой (см., например, McEnery & Wilson 2001). Отставание многоязычных ресурсов тормозит развитие многоязычной лексикографии, исследований по типологии, контрастивной лингвистике, теории перевода и т.п. Довольно большую ценность представляют **параллельные корпуса текстов (ПКТ)**, то есть текстовые массивы, включающие в себя тексты на одном языке и их переводы на другой язык (другие языки).

Использование параллельных корпусов текстов, в частности, дает реальные возможности изучения использующихся в переводческой практике эквивалентов, что может поднять на новый уровень переводные словари и обучение переводу[1]. Другое важное применение параллельных корпусов текстов — это сравнение исходных текстов и переводов. Например, появляется возможность исследования стратегий, которыми пользуется переводчик для разрешения различных грамматических и стилистических несоответствий языка оригинала и языка перевода. ПКТ представляют определенный интерес и как материал для исследования языка переводных текстов и их отличий от языка текстов, изначально написанных на данном языке. Не следует забывать и о важности параллельных корпусов текстов для компьютерной лингвистики: они являются своего рода «испытательным полигоном» для «обкатки» различных программ обработки естественного языка, в первую очередь — систем автоматизированного перевода. Наконец, параллельные корпуса художественных текстов могут ис-

[1] Следует отметить, однако, что параллельные корпуса текстов вовсе не отменяют, а лишь дополняют традиционные источники данных. Сведения, полученные из корпусов текстов, зачастую нуждаются в основательной проверке и уточнении.

пользоваться и в других гуманитарных областях, например, в литературоведении, культурологии и др.

Предлагаемая диссертация представляет собой промежуточный итог исследований, которые велись на Отделении переводоведения Тамперского университета с 1999 года. **Цель работы** — составление русско-финского параллельного корпуса художественных текстов и исследование возможностей его применения в лингвистических и переводоведческих исследованиях. Название корпуса — «ПарРус». Тексты, ставшие объектом коллекционирования, — произведения русской художественной литературы XIX-XX вв. и их переводы на финский язык. Для реализации поставленной цели было необходимо решить следующие задачи:

- изучить существующие подходы к составлению параллельных корпусов текстов и сформулировать принципы составления данного ПКТ;
- оценить объем существующих переводов с русского языка на финский и выработать стратегию отбора текстов для «ПарРус»;
- разработать для корпуса текстов программное обеспечение.

После выполнения данных прикладных задач была сделана попытка найти пути решения для следующих теоретических проблем:

- Отличается ли язык оригинальных текстов от языка текстов, являющихся переводами на этот язык с другого языка? Если да, то в чем заключаются отличия?
- Влияют ли наборы межъязыковых соответствий, предлагаемые словарями, на выбор эквивалента при переводе?
- Являются ли «ключевые слова» культуры (Вежбицкая 2001) непереводимыми?

Диссертация состоит из введения, четырех глав, заключения, списка литературы и приложений. В **первой главе** проводится критический анализ существующего опыта в области составления параллельных корпусов текстов и формулируются основные принципы составления параллельных корпусов художественных текстов, применяющиеся при работе над данным проектом. Во **второй главе** работы содержится краткий экскурс в историю художественного перевода с русского на финский, описывается исходный массив текстов и методика, использовавшаяся для классификации текстов. В этой же главе дается описание ПКТ «ПарРус». В **третьей главе** диссертации рассматриваются основные проблемы разработки программного обеспечения для корпусов текстов и описывается пакет программ, разработанный для «ПарРус». Рассматриваются перспективы развития программного обеспечения для параллельных корпусов текстов. В **четвертой главе** представлены первые результаты использования материалов корпуса для исследования различных аспектов перевода: влияние ИЯ на ПЯ, переводные эквиваленты в словаре и в переводах, ключевые слова культуры в

ИЯ и их передача в ПЯ. Итоги исследования подводятся в **заключении**. Для удобства чтения были составлены **краткий глоссарий терминов** и **список сокращений**, которые помещены в конце работы. В **списке литературы** указаны работы, на которые в диссертации имеются ссылки. В **приложения** включены различные материалы, полученные в процессе выполнения исследования, в их числе — списки текстов корпуса, статистические данные по текстам, коды компьютерных программ для обслуживания корпуса текста, некоторые из словников, полученных в ходе работы, и т.п.

Глава 1. Параллельные корпуса текстов и корпусная лингвистика

1.1. Лингвистические корпуса текстов и проблемы современной лингвистики

Понятие корпуса текстов, прочно вошедшее в последнее время в научный обиход, нельзя считать принципиально новым. Выполнение лингвистических исследований всегда требовало больших массивов эмпирических данных. Вообще, до XIX века наука о языке была по преимуществу прикладной дисциплиной, одной из главных задач языковедов было составление словарей и грамматик. Уильям Фрэнсис, один из пионеров современной корпусной лингвистики, определяет корпус текстов как «собрание текстов, считающееся репрезентативным по отношению к данному языку, диалекту или иной части языка и предназначенное для использования в лингвистических исследованиях»[2] (Francis 1992: 17).

В настоящее время лингвистическая теория по-прежнему остается на первом плане, но словари и грамматики и сейчас являются важнейшим практическим результатом, которого ждут от лингвистов. Хороший словарь вряд ли можно получить, записывая все слова, которые приходят в голову. Поэтому с незапамятных времен лексикографы собирали колоссальные картотеки примеров. Так, для Оксфордского толкового словаря английского языка за период с 1858 по 1933 гг. была собрана картотека, насчитывавшая 11 миллионов примеров (Francis 1992). Использование корпусов примеров — одна из излюбленных методик дескриптивистов (McEnery & Wilson 2001: 2–3).

Тем не менее, лишь с появлением компьютеров появилась возможность быстро собирать и обрабатывать большие массивы данных. Эра корпусной лингвистики началась в 1960-е годы, когда был составлен первый лингвистический электронный корпус текстов — Брауновский корпус (A Standard Corpus of Present-Day Edited American English, for use with Digital Computers). Он состоял из 500 текстов, каждый длиной примерно 2000 слов, т.е. объем этого корпуса — чуть больше одного миллиона словоупотреблений (Francis and Kučera 1964).

[2] Здесь и далее, кроме особо оговоренных случаев, перевод цитат мой. — М.М.

В конце 1960-х годов была завершена работа над другим проектом, начатым еще в 1949 году, — корпусом текстов св. Фомы Аквинского и других средневековых философов, собранным Роберто Буса (Roberto Busa). Объем корпуса — 10 600 000 словоупотреблений. С этого корпуса начинается другая традиция электронных корпусов текстов — корпусов текстов для исследований в области гуманитарных наук (McEnery & Wilson 2001: 20–21).

В настоящее время уже существует множество самых разных корпусов текстов, некоторые из которых по объему превышают 100 миллионов словоупотреблений, например, Британский национальный корпус (British National Corpus), Банк английского языка (Bank of English). Рубеж в 100 миллионов словоупотреблений пройден, и многие исследователи уже говорят о реальной возможности создания корпусов текстов объемом в миллиард словоупотреблений (Atkins et al 1994, Rundell 1996). Английский язык в настоящее время занимает «командные высоты» в корпусной лингвистике, причем корпуса англоязычных текстов собирают не только в Великобритании, Соединенных Штатах и других странах, где этот язык является государственным, но и за пределами англоговорящего мира. Вместе с этим, последнее десятилетие ознаменовалось активной работой по составлению корпусов текстов и для других языков — немецкого, французского, испанского, итальянского, португальского, шведского и др.

МакЭнери и Уилсон отмечают, что мнение о том, что корпуса текстов появились в 1960-х годах и получили широкое распространение в 1980-х является заблуждением. До появления порождающей грамматики в лингвистике как раз доминировало именно изучение массивов эмпирических данных, то есть корпусов. Другое дело, что анализ выполнялся вручную, вследствие чего объем данных был ограничен. Однако еще в конце прошлого века некоторые исследователи и без помощи компьютера обрабатывали огромные массивы данных. Например, Кэдинг в 1897 г. исследовал частотность букв и буквосочетаний в немецком языке на материале корпуса объемом порядка 11 млн. словоупотреблений (!!!) (McEnery & Wilson 2001: 2–3).

Расцвет порождающей грамматики и хомскианства на некоторое время прервал эту традицию. Хомский подверг корпуса текстов («доэлектронные») резкой и во многом справедливой критике:

> В любом корпусе естественного языка существуют искажения. Некоторые предложения не встретятся, потому что они очевидны, другие — потому что они ложны, третьи — потому что они невежливы. Таким образом, естественноязыковой корпус даст настолько сильно искаженную картину, что основанное на нем описание окажется простым списком. (цит. по McEnery & Wilson 2001: 10).

Хомский, как последователь рационализма, считал, что исследование эмпирических данных — абсолютно бессмысленное занятие, поскольку

суть лингвистики — в изучении именно **языковой способности** (language competence), а не отражения последней — **языковой деятельности** (language performance) (McEnery & Wilson 2001: 5–8)).

Многие лингвисты-теоретики того времени вообще относились к корпусам текстов очень критично, заявляя, что любой тренированный лингвист может легко придумать сотню хороших примеров, не обращаясь ни к каким текстам (см., напр., Фрэнсис 1988). Тем не менее, корпусная лингвистика постепенно укрепляла свои позиции и завоевывала все новых сторонников.

Критика Хомского была учтена: при создании электронных корпусов текстов искажение реальной картины старались минимизировать за счет упорядочивания выборки. К сожалению, даже тщательная работа над репрезентативностью выборки все равно оставляет вопрос об искажении (skewedness) открытым. Несмотря на это, возможность быстро получать большое количество иллюстративного материала, а также количественную информацию делала корпуса текстов все более привлекательными для исследователей. Мощность компьютеров увеличивалась, они становились все более эргономичными и удобными в обращении, все более доступными массовому пользователю. Все бóльшие массивы текстовых данных переводилось в электронную форму, программное обеспечение позволяло выполнять все более изощренные операции. Количество проектов с использованием корпусов текстов неуклонно росло (Svartvik 1992), а скептиков становилось все меньше (Fillmore 1992).

Рост популярности корпусов текстов можно проиллюстрировать, например, следующей цитатой из статьи Чарльза Филлмора:

> Я могу сделать два замечания. Первое заключается в том, что я не думаю, что могут существовать корпуса текстов, в которых содержалась бы информация по всем сферам английского лексикона и грамматики, которые я хотел бы исследовать, как бы велики эти корпуса ни были. Второе замечание состоит в том, что каждый корпус, который мне довелось изучать, как бы мал он ни был, сообщал мне такие факты, которые я никогда не мог бы выяснить никаким другим способом. (Fillmore 1992: 35)

Недавняя дискуссия в журнале *Applied Linguistics* (Widdowson 2000, Stubbs 2001, Beaugrande 2001, Widdowson 2001) наглядно демонстрирует, что хотя корпусная лингвистика и стала модным направлением, она попрежнему нередко становится объектом критики.

Главной причиной скептицизма является то, что даже современные гигантские корпуса текстов не могут удовлетворять все нужды исследователей. Дело даже не в размерах корпуса, хотя чем бóльшие текстовые массивы оказываются доступными, тем больше информации можно из них получить. Часто оказывается невозможным получить требуемые данные из-за примитивного программного обеспечения или из-за того, что не удается достаточно четко определить объект поиска (например, исследователь

ищет примеры не на конкретные слова или обороты, а на слова, относящиеся к определенной семантической группе, или на употребление определенной грамматической формы, синтаксической конструкции и т.п.). Таким образом, помимо собственно составления корпусов текстов стоит и вопрос об аннотировании (tagging) текстов в ручном или автоматическом режиме. В противном случае корпус вряд ли можно использовать за пределами лексикографической работы. Программное обеспечение также может «подрезать крылья» исследователю, если возможности пакета сводятся только к составлению словников и конкордансов.

Бурное развитие корпусной лингвистики, несомненно, связано с расцветом информационных технологий, которые требуют колоссальных массивов данных для разработки и тестирования различных лингвистических утилит, то есть сервисных программ, входящих в состав больших пакетов (например, орфографический корректор, поставляющийся вместе с текстовым процессором или системой оптического распознавания символов). Кроме того, в современной лингвистике период умозрительного конструирования вновь сменяется периодом сбора и обработки эмпирического материала. Выполняется множество прикладных исследований, так что трудно оставаться в рамках чистой лингвистики. Предмет лингвистики становится все более размытым (о различных приложениях корпусов текстов в лингвистических и гуманитарных исследованиях см. напр. McEnery & Wilson 2001: гл. 4–5).

Первые корпуса текстов представляли собой коллекции образцов относительно небольших размеров. Основным объектом такого коллекционирования был нормативный письменный вариант языка. Корпуса текстов составлялись по жесткой продуманной схеме, исключительно большое внимание обращалось на репрезентативность массива. В настоящее время сфера интересов составителей корпусов текстов заметно расширились. Все чаще внимание лингвистов привлекают весьма специфические варианты языка и подъязыки, например, детская устная и письменная речь (корпус CHILDES), речь неносителей языка (International Corpus of Learner English). Кроме корпусов текстов, отражающих письменную речь, появляются корпуса устной речи[3] (IBM-Lancaster Spoken English Corpus, Corpus of Spoken American English), некоторые большие корпуса текстов (например, BNC, COBUILD/Birmingham Corpus) включают в себя, кроме письменной речи, также и образцы устной речи. Еще одна характерная тенденция — появление так называемых исторических (или диахронических) корпусов текстов (historical corpora), фиксирующих язык на различных этапах его развития (ARCHER Corpus, Helsinki Corpus, etc.) (подробнее о развитии корпусной лингвистики в разных странах см., напр., Biber 1998; Баранов,

[3] Таким образом, в корпусной лингвистике термин «текст» понимается в максимально широком значении, которое включает в себя и расшифровку устной речи — диалогической или монологической.

Добровольский 1998, Баранов 2001: 121–128, McEnery & Wilson 2001: 20–24).

Объектом постоянных дискуссий является размер образца, а также сам вопрос о целесообразности составления корпусов текстов из образцов. В настоящее время корпуса текстов достаточно часто составляют из целых текстов (см., напр., Biber 1998: 281–284). Многие исследователи констатируют, что работа с фрагментами заданной длины накладывает определенные ограничения при исследовании текста как единого целого, связности текста и т.п. Кроме того, на основе полнотекстового корпуса достаточно легко генерируется традиционный корпус образцов, в то время как обратная операция невозможна (Baker 1995: 240).

1.2. Многоязычные корпуса текстов

Наиболее важными для теории и практики перевода типами корпусов текстов являются **многоязычные** корпуса текстов (multilingual corpora) и **параллельные** корпуса текстов (parallel corpora, далее ПКТ), а также **сравнительные** корпуса текстов (comparable corpora) (Baker 1995, Kujamäki & Jääskeläinen 2001).

Многоязычный корпус текстов представляет собой несколько аналогичных по структуре одноязычных корпусов текстов. Сравнительный корпус — это одноязычный корпус текстов, включающий в себя в качестве субкорпусов как оригинальные тексты, так и переводы на этот язык. Пример сравнительного корпуса — Савонлиннский корпус текстов, собранный в Институте переводчиков г. Савонлинна, Университет г. Йоэнсуу (Финляндия). Корпус включает в себя оригинальные финские тексты и переводы на финский язык, сделанные с английского, русского, немецкого и др. языков (Mauranen 2000, Jantunen 2001).

Параллельные корпуса текстов составляются из оригинальных текстов на языке A и переводов этих текстов на язык B. Для параллельных корпусов текстов выделяется ряд подтипов: 1) Тексты на языке A и их переводы на язык B; 2) Тексты на языках A и B и их переводы соответственно на языки B и A; 3) Только переводные тексты на языках A, B, C, ..., X, оригинальные тексты были написаны на языке D (Teubert 1996: 245). Кроме того, к ПКТ можно отнести и диахронические параллельные корпуса текстов, которые составляются из текстов на более раннем варианте языка и их переводов на современный язык (например, таковым был бы корпус древнерусских текстов и их переводов на современный русский язык), транскрипционные корпуса текстов, включающие тексты на литературном языке, прочитанные носителями разных диалектов. Кроме того, можно выделить в качестве подтипов «шумные» ПКТ (noisy parallel corpora) (то есть с пропусками в переводе, без точного соответствия между

оригиналом и переводом), «зеркальные» ПКТ (reciprocal corpora), состоящие из текстов на языках А и В и переводов этих текстов соответственно на языки В и А (см. Borin 2002: 2–5).

Важность ПКТ в теории перевода обусловлена тем, что «прескриптивный пафос исследований постепенно сменяется дескриптивным. Они <параллельные корпуса текстов> позволяют нам объективно установить, как переводчики на практике преодолевают трудности, и использовать эти данные для разработки соответствующих реальности моделей для начинающих переводчиков. Они также играют важную роль в исследовании переводческой нормы в специфических социокультурных и исторических контекстах» (Baker 1995: 231)[4].

В то же время, не следует забывать, что параллельный корпус не является истиной в последней инстанции. Корпус текстов может содержать ошибки самого разного рода: неправильные или неточные эквиваленты (в случае, если переводчик недостаточно хорошо владеет языком оригинала), орфографические, пунктуационные и стилистические ошибки (если язык, на который выполнялся перевод, не являлся для переводчика родным). Ошибки могут встречаться даже в текстах, выполненных опытными переводчиками. По этой причине «профессиональные переводчики не слишком доверяют переведенным текстам как источнику информации» и, как правило, «предпочитают использовать документы, изначально написанные на ПЯ» (Bowker 2000: 20). Все это делает многоязычные корпуса текстов в некоторых отношениях более привлекательными для переводчиков-практиков.

Тем не менее, ПКТ остаются незаменимым источником данных как для проведения исследований в области прикладной лингвистики («обкатка» систем автоматизированного перевода, заполнение систем «переводческой памяти» (translation memory), разработка систем автоматического поиска переводных эквивалентов и т.п.), так и для переводоведческих исследований (сравнение структуры исходного текста и перевода, определение степени информационных потерь при переводе, изучение различных переводческих стратегий и т.п.).

---

[4] В настоящей работе за основу взята классификация корпусов текстов, предложенная Моной Бэйкер (Baker 1995). В то же время необходимо отметить, что в настоящее время эти термины употребляются разными исследователями в разных значениях. Так, нередко термином “сравнительный корпус” обозначают то, что в нашей работе называется “многоязычным корпусом” (см., напр., Teubert 1996). В теории перевода под “параллельными текстами” нередко понимается оригинальный текст на другом языке, близкий по тематике, структуре и жанру к переводимому тексту (см. напр. Алексеева 2001: 258–259). Ларс Борин использует термин параллельный корпус (parallel corpus) по отношению к любым корпусам текстов, включающих в себя тексты на разных языках или разных вариантах одного и того же языка, подтипами которого являются многоязычные корпуса (multilingual corpora), корпуса переводов (translation corpora), сравнительные корпуса (comparable corpora) и др. (Borin 2002: 2–5).

1.3. Обзор исследовательских проектов по параллельным корпусам текстов

Параллельные корпуса текстов являются относительно новым типом языковых ресурсов. Первые ПКТ — собранные в Швейцарии сообщения о снежных лавинах на немецком, французском и итальянском языках, прогноз погоды в канадских СМИ на английском и французском — были ориентированы на специальные подъязыки с очень жестким синтаксисом и, как правило, конечной целью являлось создание системы машинного перевода (Teubert 1996: 264). Первые ресурсы данного типа появились в конце 1980-х — начале 1990-х гг. За последнее десятилетие был дан старт целому ряду проектов, так или иначе связанных с параллельными корпусами текстов. Назовем некоторые из них.

Англо-французский параллельный корпус дебатов в Канадском парламенте (Canadian Hansards English-French parallel corpus). Работа над проектом начата в 1986 и продолжается в настоящее время. Это самый первый крупный проект по ПКТ. Корпус использовался при разработке различного лингвистического программного обеспечения: программы-стыковщика параллельных текстов (aligner) (Gale, Church 1993), системы автоматической обработки текстов для машинного перевода, программы для снятия смысловой неоднозначности на уровне слова (word sense disambiguation) (Oakes 1998). В Интернете по адресу http://www.tsrali.com размещен поисковый интерфейс (McEnery & Wilson 2001: 207).

Проект **INTERSECT** (International Sample of English Contrastive Texts) в Брайтонском университете. Работы начаты весной 1994 г. Цель проекта — сбор и изучение французско-английского и немецко-английского ПКТ. На более поздних этапах работы возможно включение материалов других языков. В настоящее время объем французско-английского корпуса — около 1,5 млн. словоупотреблений на каждом из языков, немецко-английского — около 800 тыс. словоупотреблений на каждом из языков. В корпуса включаются тексты разных жанров: публицистика, документы, юридические тексты, речи политиков, учебники, техническая документация, художественная литература и др. (Salkie 2002, http://www.bton.ac.uk/edusport/languages/html/intersect.html).

Корпус **ET10-63** — англо-французский ПКТ, включающий официальные документы ЕС по телекоммуникациям. Корпус аннотирован по частям речи (POS tagging). Объем корпуса: примерно 1 250 000 словоупотреблений для каждого из языков (см.: http://www.comp.lancs.ac.uk/computing/research/ucrel/corpora.html).

Корпус **CRATER** (International Telecommunications Union (ITU) or CRATER Corpus) трехъязычный французско-испанско-английский ПКТ объемом 1 млн. словоупотреблений, состыкованный на уровне предложений. В корпус включены тексты из области телекоммуникаций. Для всех

трех языков выполнена разметка по частям речи. Обеспечен доступ к корпусу в режиме «он-лайн» (см. http://www.comp.lancs.ac.uk/computing/research/ucrel/corpora.html, McEnery & Wilson 2001: 207).

Проф. Шмид (Schmied) из Технического университета Хемниц-Цвикау работает над параллельным англо-немецким корпусом, который состоит из материалов ЕС, учебников, художественной литературы и туристических брошюр. Общий объем корпуса: примерно 500 000 словоупотреблений. Одна из проблем, особенно интересующих участников проекта — влияние языка оригинала на язык перевода (http://www.ruf.rice.edu/~barlow/para.html).

Англо-норвежский параллельный корпус, Университет г. Осло (Норвегия), 1994–1997, глава проекта Стиг Йоханссон (Stig Johansson). Корпус составлен из оригинальных художественных английских и норвежских текстов и их переводов, соответственно, на норвежский и английский. Тексты корпуса представляют собой образцы длиной по 10 000 — 15 000 словоупотреблений. Объем корпуса составляет примерно 2,6 млн. словоупотреблений. Тексты корпуса состыкованы с помощью программы, написанной Кнутом Хофландом (Knut Hofland). (Johansson 2002, http://www.hf.uio.no/iba/prosjekt/index.html).

Еще в стадии работы над двуязычным корпусом существовали планы его расширения до многоязычного ПКТ с английским языком в качестве исходного и норвежским, шведским, немецким, нидерландским, финским и португальским в качестве ПЯ (Johansson 2002: 48–49, http://www.hf.uio.no/iba/prosjekt/ index.html). В настоящее время работы по расширению корпуса уже ведутся, новый корпус получил название **Oslo Multilingual Corpus (OMC)**. Исходный англо-норвежский корпус дополняется немецкими и французскими текстами. (http://www.hf.uio.no/german/sprik/english/corpus.shtml).

В Швеции с 1993 по 2001 в университетах гг. Лунда и Гётеборга составлялся параллельный англо-шведский корпус текстов (English-Swedish Parallel Corpus — ESPC). Объем корпуса — около 2,8 млн. словоупотреблений, принципы составления во многом похожи на те, которые применялись в англо-норвежском ПКТ (см. выше). В корпус вошли фрагменты художественных текстов, а также мемуаристика, научно-популярные тексты, научные тексты, лекции нобелевских лауреатов и др. При составлении корпуса работы по отбору английских текстов проводились совместно с двумя другими скандинавскими проектами по ПКТ — вышеупомянутым англо-норвежским корпусом и англо-финским корпусом (Finnish-English Contrastive Corpus, Kari Sajavaara, Jyväskylä). Таким образом, в перспективе можно получить «вторичные» параллельные тексты для этих трех скандинавских языков, два из которых являются близкородственными, а третий (финский) относится к другой языковой семье (Altenberg & Aijmer. 2000, http://www.englund.lu.se/research/corpus/corpus/espc.html).

С проектом ESPC тесно связаны еще два крупных шведских проекта: ETAP и PLUG. Работа над проектом **ETAP** (Etablering och annotering av parallellcorpus för fastställande av översättningsekvivalenter: Создание и аннотирование параллельного корпуса для поиска переводных эквивалентов) ведется в Стокгольмском и Уппсальском университетах. Корпус состоит из нескольких субкорпусов, представляющих разные регистры языка (техническая документация, правительственные документы, пресса, художественные тексты и др.), и включает в себя тексты на шведском языке и семи языках национальных меньшинств Швеции: арабском, английском, фарси, финском, польском, сербском, испанском. В корпус включаются также немецко-шведские параллельные тексты. Важной задачей проекта, что отражено и в его названии, является аннотирование текстов и разработка программного обеспечения для автоматического поиска переводных эквивалентов (Borin 2002: 12–14).

Цель проекта **PLUG** (Parallel Corpora in Linköping, Uppsala, and Göteborg) — составление серии ПКТ со шведским языком в качестве ИЯ и разработка программного обеспечения для ПКТ. Было собрано три корпуса текстов: шведско-английский, шведско-немецкий и шведско-итальянский. Все три корпуса состыкованы на уровне предложений. В корпусах представлены следующие типы текстов: техническая документация, политические и административные документы, художественная литература. Общий объем корпуса: около 2 млн. словоупотреблений (шведско-английский субкорпус: около 1 млн., шведско-немецкий и шведско-итальянский — каждый примерно по 500 000 словоупотреблений). Важной задачей проекта является работа над программным пакетом *PWA* (the PLUG Word Aligner), программы для стыковки параллельных текстов на уровне слов. В пакете будут объединены две системы, одна из которых разрабатывается в Университете Линчепинга (Linköping Word Aligner — *LWA*), другая — В Университете Уппсалы (Uppsala Word Aligner — *UWA*) (Ahrenberg et al 2000, Hein 2002, Merkel et al 2002, Tiedemann 2002, http://numerus.ling.uu.se/~corpora/plug/).

Корпус COMPARA. Близок к вышеупомянутым проект по корпусу текстов португальского языка (Computational processing of Portuguese), начатый в 1998 году по инициативе Министерства науки и технологий Португалии. Работы по проекту проводит норвежская компания SINTEF Telecom and Informatics. В числе ресурсов и программных пакетов, разработанный проектной группой ПКТ **COMPARA**, в который входят художественные тексты на португальском языке и их переводы на английский язык. Для обработки текстов используется программный пакет *DISPARA* (DIStributing PARAlell corpora on the Web) (Borin 2002: 11, http://www.portugues. mct.pt/).

В Университете Ювяскюля (Финляндия) и Университете Вюрцбурга (Германия) в 1995–1997 составлялся параллельный немецко-финский корпус текстов, в который включались как художественные тексты, так и

публицистика, документы и т.п. В корпусе имеются как немецко-финские, так и финско-немецкие параллельные тексты. Корпус составлен из целых текстов. Тексты состыкованы на уровне предложений и размечены по стандарту SGML. Для работы с данными используется программное обеспечение, разработанное в Университете Тюбингена — *TUSTEP* (Tübingen System of Text Processing Programs) (см. Stahl 2002).

В Восточной Европе также ведутся работы по созданию параллельных корпусов текстов.

В Загребском университете составляется **хорватско-английский** параллельный корпус текстов (Tadić 2001). Источником данных является газета *Croatia Weekly*, которая выходила на хорватском и английском языках с начала 1998 по май 2000 г. В корпус вошли все номера газеты, кроме первых пяти, которые оказались недоступны в электронной форме.

В Карловом университете в Праге (Чехия) собран **англо-чешский** ПКТ, который состоит из двух частей: 1) тексты по вычислительной технике, а именно — сообщения операционной системы IBM AIX и руководств по операционным системам IBM AS/400 и VARP 4; 2) статьи из журнала Reader's Digest Výběr (от 30 до 60 % публикаций журнала — переводы с английского) и их английские оригиналы (Čmejrek, Cuřín 2001).

Другой англо-чешский ПКТ — **KACENKA** (Korpus anglicko-cesky - elektronicky nastroj Katedry anglistiky) собран на кафедре английского языка Университета им. Масарика (Брно, Чешская республика) (Department of English, Faculty of Arts, Masaryk University). Работа начата в 1997 году, в настоящее время объем корпуса более 3 млн. словоупотреблений. Бóльшую часть корпуса составляет английская литература (В. Шекспир, Ч. Диккенс, Р. Киплинг, Д.Х. Лоуренс, Т. Харди, А. Азимов и др.) и их переводы на чешский язык. Имеется и нелитературная составляющая: биржевые сводки и help-файлы (см. http://www.phil.muni.cz/angl/kacenka/kacha.html).

В Любляне составлен **словенско-английский-англо-словенский** параллельный корпус текстов IJS-ELAN. Объем корпуса небольшой — 1 миллион словоупотреблений, корпус размечен по частям речи и состыкован на уровне предложений (http://nl.ijs.si/elan, Erjavec 1999).

Ассоциация **ELRA** (European Language Resources Association), основанная в 1995 г., финансирует целый ряд проектов по ПКТ. Следует упомянуть по крайней мере следующие два ресурса:

1. Многоязычный параллельный корпус **MLCC** включает в себя тексты на девяти европейских языках: датском, нидерландском, английском, французском, немецком, греческом, итальянском, португальском и испанском. Параллельные тексты, предоставленные Европейской Комиссией, представляют собой материалы из «Официального журнала Европейского сообщества» (Official Journal of the European Communities). Корпус состоит из двух субкорпусов:

а) запросы депутатов Европейского Парламента Европейской Комиссии (1993 г.) и девять параллельных версий ответов. Объем: примерно 10,2 млн. словоупотреблений, по 1,1 млн. на каждый язык.

б) дебаты в Европейском Парламенте за 1992–1994 гг. Материалы представляют собой стенограммы заседаний на девяти языках ЕС. Примерный объем корпуса — от 5 до 8 млн. словоупотреблений на каждом из языков.

2. Корпус **MULTEXT JOC**. Представляет собой часть корпуса **MULTEXT** (см. ниже), работы финансировались Европейской Комиссией (LRE 62-050). Эта часть корпуса состоит из письменных вопросов и ответов из «Официального журнала Европейского Сообщества» (Official Journal of the European Communities). Корпус состоит из параллельных текстов на английском, французском, немецком, итальянском и испанском языках, общий объем — около 5 млн. словоупотреблений, по 1 млн. на каждый язык. Для части корпуса объемом около 800 000 словоупотреблений была сделана грамматическая разметка (английский, французский, итальянский и испанский). Тексты этой же части корпуса были состыкованы на уровне предложений с английским языком в качестве исходного (http://www.icp.grenet.fr/ELRA/home.html).

Кроме ELRA финансированием проектов по лингвистическим ресурсам занимается также ассоциация **TELRI** (Trans-European Language Resources Infrastructure). Ассоциация создала в 2000 году исследовательский архив **TRACTOR** (TELRI Research Archive of Computational Tools and Resources). Среди ресурсов архива имеются следующие параллельные корпуса текстов.

1. CD-ROM **MULTEXT-EAST**. Корпуса текстов, утилиты, тексты на различных языках Европы.

2. CD-ROM **East meet West**. Параллельные тексты на многих европейских языках.

3. Параллельные тексты на болгарском, английском и французском языках, собранные в Лаборатории лингвистического моделирования Академии наук Болгарии (София). Тексты хранятся в формате MS Word, строки исходного текста и перевода чередуются (http://www.telri.de/, http://www.tractor.de/).

Кроме ПКТ, в которых представлены только европейские языки, начинают появляться и корпуса текстов с такими парами языков, как английский и китайский, английский и панджаби (McEnery & Wilson 2001: 70).

Наш небольшой обзор наглядно показывает рост интереса к ПКТ. Однако, несмотря на то, что определенная работа в этом направлении ведется, до решения даже насущных проблем еще довольно далеко. Параллельных корпусов текстов пока довольно мало, размеры их довольно скромны по сравнению с одноязычными корпусами текстов с объемом в десятки и сотни миллионов словоупотреблений. Многие ПКТ содержат только «сырые» тексты, которые даже не состыкованы. Пока представлено

довольно мало пар языков, причем, как правило, одним из языков пары является английский язык.

То, что почти во всех ПКТ (за редкими исключениями) в качестве входного или выходного языка представлен английский язык, представляется вполне логичным. Английский язык фактически является в настоящий момент lingua franca в очень многих областях, и получить параллельные тексты с английским в качестве ИЯ и ПЯ проще, а корпуса с английским языком легче находят практическое применение. Однако представляется, что такое увлечение одним языком несколько сужает возможности использования ПКТ в исследованиях по типологии и в многоязычной лексикографии.

1.4. Проблемы составления параллельных корпусов текстов

При составлении параллельных корпусов текстов, в отличие от одноязычных и сравнительных корпусов текстов, обязательно следует учитывать фактор межкультурных связей. Множество текстов исходного языка (ИЯ) составляют лишь те тексты, которые были переведены на второй язык (ПЯ). Таким образом, если межкультурные связи полностью отсутствуют, получение ПКТ невозможно. Чем слабее связи, чем меньше связаны культуры, тем меньше переводов выполняется и тем более проблематично составление полноценного ПКТ.

Например, наличие политических и культурных связей между двумя странами вызывает потребность в переводе с одного языка на другой различных документов, инструкций, руководств, брошюр и т.п. Развитие «народной дипломатии», то есть поток туристов, малый бизнес, рабочие контакты, браки и т.п. являются важным фактором в укреплении связей между странами, который даже в большей степени, чем развитие «официальных» отношений, способствует усилению интереса к другой культуре. Интересно, что в современном мире такой фактор, как географическая близость, не играет столь важной роли, как этого можно было бы ожидать. Хотя англоязычные страны не являются соседями России (если не принимать во внимание границу с США по Берингову проливу), количество текстов, переводимых с английского языка на русский (а также текстов, переводимых в России с русского на английский), значительно превышает количество переводов с любого другого языка. Польша, Чехия и Словакия ближе к России, чем Германия и Франция, кроме того, эти страны — бывшие партнеры России по Варшавскому договору и СЭВ, однако совершенно очевидно, что переводов на русский с немецкого и французского языков выполняется больше, чем с польского или чешского.

Если языки сосуществуют на одной территории, то появление большого количества параллельных текстов неизбежно: официальные тексты — документы, инструкции и т.п., рекламные тексты, учебники, переводы художественной литературы и т.д. Это могут быть «большие» языки[5] в качестве языков национальных меньшинств (например, испанский в США, украинский в России, финский в Швеции, русский в Финляндии) или второго государственного языка (шведский в Финляндии, французский в Канаде) либо «малые» языки (например, язык коми в России или саамский в Швеции или Финляндии). В случае территориального пересечения возникает специальный местный вариант языка, приспособленный к функционированию в новой культурной среде, которой может отличаться от языка «метрополии» (напр. финский шведский или шведский финский) (Trosterud 2002).

С другой стороны, в некоторых случаях может ставиться и задача создания ПКТ для пар «больших» языков, которые не пересекаются территориально или по крайней мере не образуют больших диаспор, например английский и немецкий, французский и шведский, русский и испанский. В этих случаях количество параллельных текстов существенно меньше, и вообще в этом случае параллельными зачастую оказываются тексты совсем других типов и жанров. Очень трудно найти переводы муниципальных документов, зато есть тексты международных соглашений и переводы речей политиков. Важную роль начинает играть художественный перевод.

Следует отметить, что для пар «больших» языков в случае территориального пересечения могут быть получены параллельные тексты обоих указанных выше типов. Такие пары оказываются в наиболее выигрышном положении при составлении ПКТ. С другой стороны, для многих пар типа «большой язык» — «малый язык» и «малый язык» — «малый язык» при отсутствии территориального пересечения ПКТ могут быть получены лишь через третий язык (например, для языков национальных меньшинств России могут быть получены параллельные тексты через русский язык), а в некоторых случаях даже такие «псевдопараллельные тексты» отсутствуют (можно ли получить ПКТ саами–мари?) (о составлении ПКТ для языков национальных меньшинств см. подробнее Trosterud 2002).

Параллельный корпус, таким образом, является как бы точкой пересечения двух языковых культур. ПКТ состоит из двух (иногда — более, чем двух) субкорпусов — текстов на ИЯ (далее — субкорпус ИЯ) и их переводы на один или несколько ПЯ (далее — субкорпус(ы) ПЯ). Тексты на ИЯ, хотя и являются первичными, отбираются с учетом ПЯ. И вообще,

[5] Будем так называть языки, имеющие в какой-либо стране статус основного государственного языка, хотя термин по своей сути противоречив: эстонский язык попадает в разряд «больших», поскольку является государственным языком Эстонии, а татарский будет «малым», поскольку является лишь государственным языком субъекта Российской Федерации, хотя количество говорящих на татарском языке существенно больше, чем носителей эстонского языка.

структура субкорпуса ИЯ определяется наличием или отсутствием переводов на ПЯ, а также тем, какого рода тексты переводятся.

В целом, при составлении ПКТ в распоряжении исследователя могут быть следующие языковые ресурсы:

- специальные тексты;
- тексты СМИ;
- научные тексты;
- художественные тексты.

1.4.1. Специальные тексты

Документы. Это личные документы (свидетельства о рождении, браке, документы об образовании); деловые письма, контракты, коммерческие предложения, бизнес-планы, лицензии; тексты международных договоров, материалы дипломатических переговоров и т.п. Параллельных текстов такого рода особенно много, если в стране два официальных языка, например финский и шведский языки в Финляндии, английский и французский в Канаде и т.п. Существование такого рода параллельных текстов зависит также от наличия деловых, дипломатических и политических контактов между двумя странами. Процессы интеграции в странах ЕС также ведут к появлению большого количества документов, составляющихся на нескольких языках.

Главная проблема при составлении корпуса из текстов данного типа — конфиденциальный характер многих документов. Эта проблема решается путем исключения из текстов имен, названий организаций, географических названий, дат и т.п. Получение текстовых массивов такого рода осложняется и тем, что многие из документов являются «эфемеридами», то есть перевод выполняется один раз для одного клиента, после сдачи работы текст уничтожается. Еще одна трудность заключается в том, что исходный текст чаще всего существует только в «бумажной» форме. Другая проблема — низкое качество перевода многих личных документов и деловой переписки (при переводе на родной язык переводчика возможны фактические ошибки, неточные эквиваленты, при переводе же на неродной язык — грамматические и стилистические ошибки, неправильные переводные эквиваленты). Следует, впрочем, отметить, что качество перевода важно лишь в том случае, если ресурс будет использоваться для составления переводных словарей или в качестве справочного источника для переводчиков. Если же составители корпуса планируют исследовать межъязыковую интерференцию, типичные ошибки при переводе и т.п., то именно плохие переводы могут быть весьма ценным источником данных.

Параллельных корпусов документов очень мало; нам известно о существовании корпуса документов ЕС (см. выше). В Университете

Мангейма собран корпус текстов НАТО (http://www.tractor.de), к сожалению, тексты этого корпуса, насколько нам известно, не состыкованы.

Инструкции и пособия. Тексты данного типа чрезвычайно широко распространены и довольно разнообразны как по форме, так и по содержанию — от текста на упаковках пищевых продуктов до туристической брошюры. Для некоторых пар языков и для некоторых сфер экономики тексты этого типа доступны в огромных количествах. Например, финские туристические брошюры всегда переводятся на шведский и английский, нередко — также и на немецкий и русский. Инструкции по эксплуатации бытовой электроники всегда переводятся на несколько языков, один из которых, как правило, английский.

Этот тип текстов является очень привлекательным не только для исследовательских целей, но и для разработки различных практических приложений. Техническая документация является важной составляющей многих ПКТ (см. обзор в разделе 1.3. настоящей работы).

Однако для некоторых пар языков найти параллельные тексты этого жанра оказывается не так просто, как это может показаться. Россия почти не экспортирует электронику, поэтому трудно рассчитывать на большое количество переводов инструкций по эксплуатации мобильных телефонов с русского языка. Финляндия мобильные телефоны экспортирует, но инструкции по их эксплуатации переводятся в самой Финляндии на английский (причем, в некоторых случаях инструкции могут быть составлены именно на английском, а потом переведены на финский) и шведский. При экспорте финских мобильных телефонов документация, скорее всего, переводится не с финского, а с английского языка. Таким образом, параллельные корпуса, полученные из таких текстов, корректнее называть псевдопараллельными. Аналогичная ситуация, по-видимому, и с продукцией японских и корейских фирм и переводами их документации.

1.4.2. Язык СМИ

Перевод материалов прессы на разные языки происходит довольно часто, но, как правило, на основе разовых заказов. Однако бывают и исклюения из правил. Уникальным в своем роде является советский опыт. Задачей Агентства печати «Новости» (АПН), созданного в 1961 г. и просуществовавшего до 1991 г. (когда оно было преобразовано в агентство РИА «Новости»), было пропагандировать советский образ жизни и достижения социализма. Этой работой занимались бюро АПН в 120 странах мира, одной из задач которых было «продвигать» материалы советской прессы в местные издания. В каждом представительстве АПН работало несколько штатных переводчиков и редакторов, которые переводили статьи из советской прессы. Часть переводов АПН публиковались в газете «Московские новости», которая в период с 1955 по 1980-е гг. выходила на анг-

лийском, французском, немецком, итальянском, испанском, арабском, венгерском, финском, эстонском, греческом, эсперанто и некоторых других языках, причем газета эта до 1980-го года на русском языке не выходила (см. интернет-страницу газеты: http://www.mn.ru/ publishing-house/history. html). Кроме того АПН издавало дайджест советской прессы «Спутник», также на нескольких языках (подробнее об АПН см. на сайте РИА: http://history.rian.ru).

Однако советский опыт перевода материалов прессы, по-видимому, является единственным в своем роде. В настоящее время газета «Московские новости» («Moscow news») выходит только на русском и английском языках. Иногда на приграничных территориях существует двуязычная пресса, например, в Выборге издается на финском языке дайджест региональной российской газеты «Выборгские ведомости» («Viipurin Sanomat»). Впрочем, языковые ресурсы такого типа встречаются довольно редко: даже в случае официального многоязычия предпочитают издавать разные газеты на разных языках.

Другим, гораздо более характерным видом многоязычной прессы являются международные журналы, такие как *National Geographic*, *PC World*, *Burda* и др. Эти журналы издаются не только на языке оригинала, но и распространяются во многих странах в переводах на местный язык. Однако ресурсы подобного рода оказываются доступны лишь с одним из мировых языков (в большинстве случаев — английским) в качестве ИЯ.

Несмотря на всю привлекательность языковых ресурсов данного типа, проектов по составлению параллельных корпусов языка СМИ пока немного. Тексты СМИ входят в шведские многоязычные корпуса текстов (см., например, Borin 2002: 13). Целиком состоит из текстов СМИ хорватско-английский корпус (Tadić 2001), журнальные статьи входят в англо-чешский ПКТ (Čmejrek, Cuřín 2001).

Многоязычная пресса, как правило, является источником переводов хорошего и высокого качества. Однако в некоторых случаях мы имеем дело с переводами на неродной язык, как, например, в Загребском проекте, когда газетные статьи переводились на английский язык с хорватского, по-видимому, в основном неносителями английского языка, хотя, возможно, тексты редактировались носителем языка. Аналогичная ситуация обстоит с переводами, публикующимися в газете «Московские новости».

1.4.3. Научные тексты

Научные тексты нередко становятся объектом перевода, однако здесь следует сделать ряд уточнений. Многие ученые — основные адресаты научных текстов — владеют иностранными языками. Нередко сам ученый пишет на языке, знакомом большей части его аудитории (латинский — в Средневековье, французский или немецкий — в XIX веке, английский — в

настоящее время). Поэтому только классические научные труды (Дарвин, Маркс, Карлайл, Соссюр и др.) переводились на многие языки. Таким образом, только так называемые «языки науки» могут дать достаточное количество текстового материала для получения параллельного корпуса текстов, а для многих языковых пар параллельные научные тексты могут быть получены только через третий язык.

1.4.4. Язык художественной литературы

Ни один из вышеупомянутых типов текстов не может быть стабильным и универсальным источником для ПКТ. Как уже отмечалось, существуют пары языков, для которых параллельные тексты некоторых типов вообще могут отсутствовать. В то же время, переводы художественной литературы выполняются даже в том случае, если для данной пары языков отсутствуют переводы брошюр, технической документации или научных текстов. Разумеется, во многих случаях и такие ресурсы могут быть довольно ограниченными. Тем не менее, нам представляется, что перевод художественной литературы играет более важную роль в развитии отношений между культурами, чем перевод деловых писем или туристических проспектов.

Чрезвычайно интересный языковой ресурс представляют собой поэтические переводы, но сфера практического применения параллельных корпусов поэзии более ограничена по сравнению с прозой, которая, в частности, может быть использована в качестве лексикографического источника. Прозаические тексты содержат монологи и диалоги, повествования и описания, нормативный язык и сленг, диалект.

Нельзя утверждать, конечно, что в художественной прозе содержится вся лексика и все языковое богатство. В художественных текстах, как правило, относительно мало терминологии, кроме тех терминов, которые стали общеупотребительными (тем не менее, в составленный нами корпус текстов включен роман В. Дудинцева «Белые одежды», в котором довольно много научной терминологии, поскольку герои романа — ученые-генетики). Клише и неологизмы более типичны для СМИ. Сленговые и диалектные вкрапления в художественном тексте также являются стилизацией. Тем не менее, художественные тексты играют очень заметную роль в развитии и формировании любого естественного письменного языка.

Появление своей художественной литературы нередко является важным этапом формирования национального самосознания. Интересно, что в этот момент во многих случаях не последнюю роль играет перевод художественной литературы с других языков (подробнее об этом см. в гл. 2 настоящей работы). Таким образом, художественные тексты, несомненно, оказываются очень важным ресурсом для ПКТ.

В последующих разделах будут рассматриваться проблемы, которые возникают при составлении параллельных корпусов художественных текстов.

1.5. Проблема репрезентативности параллельных корпусов текстов

Классическая лингвистическая традиция не считает переводные тексты объектом, достойным изучения, однако в последнее время к языку переводных текстов как к одному из вариантов языка стали относиться более уважительно (см., в частности, Baker 1996: 175, 1999: 282, Eskola 2002: 2). Язык переводных текстов отличается от языка оригинальных текстов, как с точки зрения грамматики, так и своей лексикой. Достаточно долгое время этот вариант языка считался «неправильным», в английской традиции даже появился термин «translationese», которым обозначался язык плохих переводов с сильным влиянием ИЯ (см. например, Newmark 1991). По этим причинам, как правило, в исследованиях по грамматике переводные тексты в качестве источника данных не использовались. МакЭнери и Уилсон отмечают в качестве одного из серьезных недостатков параллельных корпусов текстов — грамматические конструкции, употребление которых вызвано влиянием ИЯ, лексические кальки с ИЯ, а также неправильные и неточные эквиваленты (McEnery & Wilson 2001: 72).

Первые электронные корпуса текстов составлялись исключительно из оригинальных текстов. Разработчики Брауновского корпуса даже по возможности проверяли, являлся ли автор текста носителем американского варианта английского языка (Francis, Kučera 1964). Таким образом, ясно, что язык текстов субкорпуса ПЯ не является нормативным. Одним из способов снятия этого перекоса — сделать ПКТ «двунаправленным», т.е. включить в него оригинальные и переводные тексты на обоих языках, и таким образом сбалансировать корпус. В результате архитектура корпуса примет следующий вид (рис. 1). Так устроен, например, Англо-норвежский параллельный корпус (см. Johansson 2002, http://www.hf.uio.no/iba/prosjekt/ENPCmanual.html).

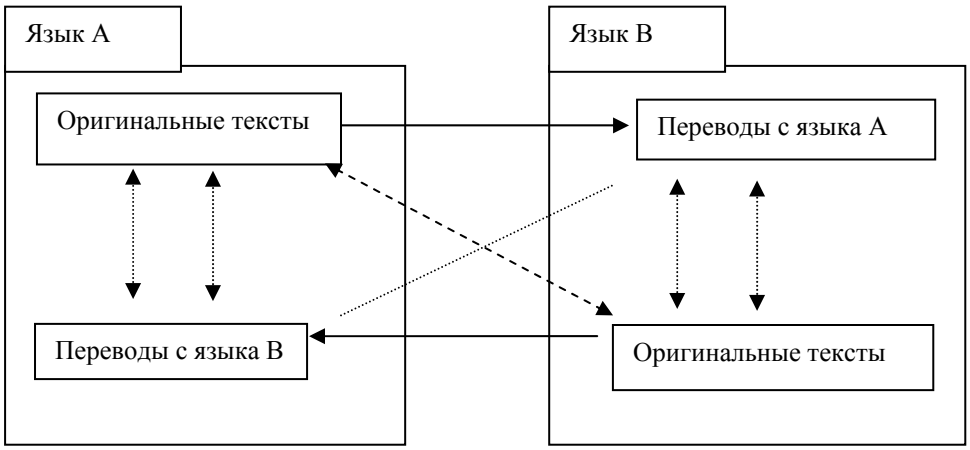

Рис. 1. *Сбалансированный параллельный корпус текстов*

Следует, тем не менее, отметить, что и такая модификация структуры корпуса текстов в итоге не устраняет имеющийся перекос полностью. Главная причина в том, что задача коллекционирования оригинальных текстов и их переводов ограничивает выбор оригинальных текстов (ведь в корпус в итоге включается только то, что реально переводилось). Поэтому субкорпусы ИЯ нельзя рассматривать как обычные одноязычные корпуса текстов. О трудностях, возникающих при составлении подобного рода корпусов текстов пишет Стиг Йоханссон (Johansson 2002: 58). Причем трудности, о которых говорит норвежский ученый, возникают уже при составлении корпуса текстов на европейских языках — английском, норвежском, шведском, немецком. Эти проблемы становятся еще более серьезными при составлении ПКТ с языками, у которых не было столь тесных культурных контактов, как в вышеуказанном случае. Структура ПКТ на рис. 1 нередко может оказаться чисто умозрительным конструктом, который теоретически возможен только для близких и тесно связанных культур с давними традициями перевода с одного языка на другой (напр. английский и французский). В большинстве же случаев мы получим на каждом языке два совершенно разных корпуса текстов с довольно слабыми возможностями сравнения оригинальных и переводных текстов.

Рис. 2. *Сбалансированный ПКТ. Усовершенствованный вариант структуры.*

Например, если начать составлять сбалансированный русско-финско-финско-русский ПКТ, то скоро станет ясно, что текстов, переводимых с русского языка на финский намного больше, чем текстов, переводимых с финского на русский. Кроме того, не всегда окажется возможным найти тексты, аналогичные по тематикам и жанрам. Таким образом, получить «зеркальные» субкорпуса оригинальных и переводных текстов представляется проблематичным. Русско-англо-русский ПКТ также вряд ли реализуем, но по другим причинам. Хотя в обе стороны переводится довольно много текстов, различия в культурных традициях приведут к такому сильному дисбалансу в структуре субкорпусов, что вряд ли удастся удержать их «под одной крышей».

Структура корпуса текстов, предлагаемая на рис. 1 может быть улучшена путем введения одноязычных корпусов текстов в качестве «контрольных» массивов (рис. 2).

Следует впрочем отметить, что таким образом мы получим скорее собрание корпусов текстов, чем один ПКТ.

Такова была главная причина отказа (по крайней мере — временного) от идеи разработки русско-финско-финско-русского параллельного корпуса текстов. Было принято решение сосредоточиться именно на русско-финс-

ком корпусе текстов, поскольку с русского на финский переводится больше, чем с финского на русский.

Полученный корпус, как уже было показано выше, будет давать в некоторой степени искаженную картину. Поэтому, для получения более объективных данных сведения, полученные из субкорпуса ИЯ, будут сравниваться с «ТамРус» — одноязычным корпусом русских художественных текстов, составляемом одновременно с «ПарРус» в Тамперском университете; субкорпус ПЯ будет сравниваться с данными корпуса финских текстов, собранного в Школе переводчиков Университета г. Йоэнсуу (г. Савонлинна, Финляндия) (субкорпус оригинальных финских художественных текстов)[6].

1.6. Традиционные корпуса текстов, корпуса художественных текстов, параллельные корпуса художественных текстов

Параллельный корпус художественных текстов нельзя назвать чисто **лингвистическим** корпусом текстов. Лингвистика оказывается лишь одной из областей, в которых могут использоваться данные, полученные из ПКТ. С одной стороны, ресурсы такого типа чрезвычайно полезны для лингвистических исследований в области типологии, сравнительного языкознания, контрастивной лингвистики. ПКТ — прекрасный источник данных для многоязычных словарей. С другой стороны, из ПКТ можно получать сведения для культурологических и литературоведческих исследований.

Переводоведение — еще одна (возможно, даже основная) сфера применения ресурсов такого рода. Пользователями ПКТ могут быть также и переводчики-практики, поскольку корпус может использоваться как справочная система, в которой зафиксированы прецеденты перевода, а поскольку художественные переводы, как правило, выполняются на достаточно высоком уровне, то полученной информации в целом можно доверять (хотя ошибки и в художественных переводах — не такая уж редкость).

Наконец, большой ПКТ может рассматриваться как справочная система для поиска цитат и их переводов.

Все это следует принимать во внимание при определении структуры корпуса, его объема, критериев отбора текстов, а также при разработке программного обеспечения.

---

[6] Автор выражает признательность проф. Анне Мауранен (Тамперский университет, Школа современных языков и переводоведения), давшей разрешение на использование Савонлиннского корпуса текстов (*The Savonlinna Corpus of Translated Finnish*. Savonlinna School of Translation Studies, University of Joensuu, 2001).

1.6.1. Традиционный корпус текстов. Репрезентативность и выборка

Проблема репрезентативности (representativeness) — одна из важнейших проблем, обсуждение которой идет с самого момента появления электронных корпусов текстов. Хотя на различие языка разных стилей и жанров и обращалось внимание задолго до появления корпусов текстов, только появление больших массивов данных в электронной форме позволило наглядно продемонстрировать всю непохожесть разных подъязыков одного и того же естественного языка.

Основным средством для достижения баланса между разными подъязыками и идиолектами в массиве данных ограниченного объема было составление корпусов текстов из фрагментов (sampling). Использование образцов одинаковой длины предотвращает доминирование языка и стиля более длинных текстов над языком коротких текстов. Здесь возможны два подхода — пропорциональная и стратифицированная выборка (proportional vs. stratified sampling). В пропорциональной выборке учитывается значимость и вес каждого типа текстов в бытовании языка. Стратифицированная выборка предполагает равную представленность текстов разных типов в корпусе (Biber, Conrad, Reppen 1998: 247–248).

Многие исследователи, в частности Мартин Геллерстам (Martin Gellerstam), отмечают в качестве слабого места пропорциональной выборки тот факт, что при такой стратегии в корпусе оказываются представленными, в основном, язык СМИ и язык документов. Такие важные для развития и функционирования языка тексты, как художественная проза, поэзия и драматургия, научная и учебная литература либо оказываются представленными по минимуму, либо вообще не попадают в корпус текстов (Gellerstam 1992: 154). Хотя пропорциональная выборка в целом ряде случаев может оказаться полезной для дальнейших исследований (например, чтобы выяснить, насколько частотна в языке та или иная грамматическая форма или лексема), стратифицированная выборка представляется более перспективной стратегией для решения большинства задач лингвистических исследований. Этот принцип используется в Брауновском корпусе (500 фрагментов по 2000+ слов), в Ланкастерско-Осло-Бергенском корпусе английского языка (LOB Corpus) и многих других корпусах текстов.

С увеличением размеров корпусов текстов и с расширением сферы их использования становилась все более очевидной нечеткость понятия «репрезентативность». Множество всех текстов языка плохо поддается описанию и каталогизации, даже в том случае, если, как при разработке Брауновского корпуса, заранее четко определены временные рамки, вариант языка и т.п., а также тексты, остающиеся за пределами корпуса (см., например, Engwall 1994).

Во многих случаях выделение или невыделение той или иной категории оказывается на совести исследователя. Чем больше размеры корпуса, тем сложнее становится классификация текстов. Объем данных растет, но корпус не становится более представительным. Разработчики Лонгман-Ланкастерского корпуса (Longman/Lancaster Corpus) осторожно определяют репрезентативность как «то, что мы считаем типичными и центральными аспектами языка, что дает достаточное количество употреблений слов и выражений для лексикографов и изучающих иностранные языки и что позволяет получать из корпуса достаточное количество данных, дающее реалистическую картину функционирования лексики» (Summers 1993: 189–190).

Диахронические корпуса текстов в еще большей степени создают проблемы с репрезентативностью из-за появления дополнительного временного фактора. Тексты разных периодов трудно сбалансировать из-за того, что количество публикуемых в разное время текстов может сильно различаться, некоторые стили и жанры могут вообще отсутствовать. Многие из больших корпусов текстов фактически являются диахроническими, поскольку «синхронический корпус представляет собой как бы «фотографию» языка, которая быстро устаревает, что делает такой принцип малопригодным для общеязыкового корпуса текстов» (Summers 1993: 193).

Структура корпуса текстов — своего рода модель языка, которая может в большей или меньшей степени отражать описываемый объект. Представляется, что ясная и четкая структура реальна только для небольших синхронических корпусов текстов, в идеале описывающих очень ограниченный подъязык.

Поэтому в течение последнего десятилетия все большее распространение получает идея **мониторного** корпуса текстов, то есть такого корпуса, который не создается раз и навсегда подобно классическим корпусам текстов, а постоянно дополняется новыми текстами. Подобной стратегии формирования корпуса текстов придерживаются при создании больших корпусов текстов, ориентированных в первую очередь на лексикографическую работу. В качестве примера мониторного корпуса текстов можно назвать Bank of English, самый большой на сегодняшний день корпус текстов английского языка, объем которого составляет около 400 млн. словоупотреблений (McEnery & Wilson 2001: 30–31). Такой корпус текстов можно также назвать **динамическим**, поскольку из него в зависимости от потребностей исследователя можно получать субкорпусы, являющиеся репрезентативными для определенного жанра, автора, периода и т.п. (см., например, Баранов, Михайлов, Сидоров 1998; Баранов 2001: 116).

1.6.2. Корпус художественных текстов и классическое понятие корпуса текстов

Репрезентативность в классическом понимании предполагает составление корпуса образцов ограниченной длины. Главная проблема этой стратегии заключается в том, что при получении корпусов из фрагментов зачастую изначально закладываются значительные ограничения на использование корпуса. Полезность выборок вполне очевидна при составлении корпуса нормативного языка или языка какой-либо ограниченной сферы. Однако использование того же принципа при составлении корпуса художественных текстов весьма проблематично. Ведь важной сферой применения ресурсов такого рода являются исследования языка и стиля конкретных писателей, что представляется не менее важным, чем изучение языка художественной литературы в целом.

Другая проблема, возникающая при составлении корпусов художественных текстов — определение принципов выбора авторов и отбора текстов. Классические корпуса текстов составляются для изучения нормативного языка, т.е. «правильного» языка, на котором пишет и говорит средний образованный носитель языка. Таким образом, выдающиеся лица языкового сообщества (знаменитые писатели, популярные журналисты, крупные политические деятели, известные ученые) в расчет не принимаются. Лингвистический корпус текстов должен составляться из обыкновенных текстов: обычных информационных сообщений, ничем не примечательных газетных статей, второсортных романов, речей никому не известных политиков и т.п. (см. например, Sinclair 1991). В итоге подобные обезличенные массивы данных дают возможность получить сведения о том, какова общая практика употребления тех или иных слов, выражений или грамматических форм.

При составлении корпуса художественных текстов идея «средневзвешенности» начинает противоречить идее литературного произведения как такового. Ведь суть литературного стиля состоит в том, чтобы удивить читателя необычным языком, нестандартными идиомами, непредсказуемыми коллокациями, запоминающимися метафорами, авторскими неологизмами. Таким образом, язык художественной литературы нельзя назвать стандартизованным, писатели не являются обычными носителями языка: они скорее создают язык, чем являются обычными его «пользователями». По этим причинам представляется, что составителям корпуса художественных текстов не следует ограничиваться второразрядными писателями, полезнее было бы сосредоточиться именно на известных авторах.

Выдающееся литературное произведение — это всегда событие. Даже если за один год было издано одно произведение, ставшее классическим, можно сказать, что для культуры этот год не прошел зря. Таким образом, в большинстве случаев для синхронического корпуса художественной лите-

ратуры материала может просто не хватить. Итак, если ориентироваться именно на крупных писателей и значимые для культуры произведения, то полученный корпус текстов будет скорее диахроническим.

Таким образом, принимая во внимание все сказанное выше, можно предложить следующие принципы составления корпуса художественных текстов:

- корпус должен по возможности составляться из полных текстов, а не из образцов;
- в состав корпуса должны в первую очередь включаться произведения известных писателей, а также тексты, сыгравшие важную роль в исследуемом языковом сообществе;
- произведения могут быть созданы в течение относительно значительного промежутка времени (20 лет, 50 лет, 100 лет).

Тем самым отвергаются все основные принципы составления лингвистических корпусов текстов: фрагменты равной длины, «второразрядность текстов», синхроничность. Да и само использование термина «корпус текстов» для коллекции художественных текстов представляется проблематичным. Такого рода собрание текстов можно было бы назвать, например, **электронной антологией**.[7] Примерно таким образом составлялся упомянутый выше корпус текстов Фомы Аквинского, составленный Роберто Буса (McEnery & Wilson 2001: 20–21). Предлагаемое нами понятие приближается и к концепции Машинного фонда русского языка, работа над которым идет в Российской Академии Наук с начала 1980-х гг. Основной идеей проекта было перевести в электронную форму все тексты, значимые для развития русского языка и русской культуры, начиная с Древней Руси и кончая современностью (Андрющенко 1989, Караулов 1986). Близких взглядов придерживаются в проекте «Корпус текстов Австрийской Академии Наук» (AAC = Austrian Academy Corpus) (Biber et al 2002).

Репрезентативность электронной антологии следует отличать от репрезентативности корпуса текстов. Дело в том, что количество «обычных», «стандартных» текстов для любого языка стремится к бесконечности (даже для мертвого языка, поскольку, во-первых, археологи и архивариусы периодически находят все новые, ранее неизвестные тексты на латинском, древнегреческом или древнерусском языках, а во-вторых, даже после «смерти» язык может в какой-то степени продолжать функционировать, как это случилось, например, с латинским языком). А вот количество классических и прецедентных текстов для любой культуры не так велико. Таким образом, репрезентативность для электронной антологии — это представленность в ней по возможности всех текстов, значимых для данной культуры и языка.

[7] Тем не менее, наряду с этим новым термином, мы будем для удобства продолжать пользоваться и термином «корпус текстов».

Следует отметить, что и возможность составления корпуса художественных текстов по классическим принципам тоже нельзя отвергать (корпус художественных текстов может быть составлен из отрывков равной длины, взятых из никому не известных произведений написанных никому не известными авторами и изданных в течение одного года). По такой методике формировался, например, субкорпус художественных текстов в Брауновском корпусе текстов. Такой корпус позволит получить общую картину языка художественной литературы за такой-то год. Однако использование такого корпуса не будет выходить за пределы грамматики и, может быть, отчасти стилистики.

Следует отметить, что «законодателями мод» в литературе являются все-таки известные писатели и оставление их за пределами корпуса представляется в определенной степени искусственным (а при составлении параллельного корпуса художественных текстов, как будет показано ниже, применять принцип «второразрядности» вообще проблематично, поскольку переводятся как правило лишь произведения в тех или иных отношениях выдающиеся, по крайней мере — бестселлеры).

1.6.3. Параллельный корпус художественных текстов: дополнительные проблемы

Итак, корпус художественных текстов можно составлять как по традиционной методике, так и как электронную антологию. Результирующие массивы будут сильно отличаться, сфера применения полученных корпусов также будет разной, но оба варианта вполне выполнимы. Однако параллельный корпус художественных текстов вряд ли может быть составлен по классической методике, это оказывается возможным лишь в случае тесного взаимодействия языковых культур и наличия большого количества переводов.

Первая проблема состоит в том, что заурядные, второразрядные художественные произведения редко становятся объектом перевода. Разумеется, в числе произведений, переводящихся на другие языки, есть и серьезная литература и беллетристика. Например, в последние годы с русского на финский язык переводились детективы Александры Марининой и исторические детективы Бориса Акунина — несомненно литература легкого жанра, произведения Виктора Пелевина, которые одни критики считают новым словом в художественной литературе, а другие — дурным вкусом, и рассказы Татьяны Толстой, представляющие серьезную литературу.

Но чтобы стать объектом перевода, произведение должно пользоватся достаточно широкой известностью, по крайней мере на родине писателя, хотя бы быть бестселлером. Таким образом, с языка на язык переводится классика и модные авторы. (Конечно, бывают и исключения. Например, в Советском Союзе часто переводили иностранных писателей-коммунистов,

в результате Джеймс Олдридж, Джанни Родари и Матти Ларни гораздо больше известны в России, чем у себя на родине).

Временной фактор вносит дополнительные трудности даже по сравнению с одноязычными корпусами художественных текстов. Дело в том, что объектом перевода могут оказываться как современные, так и классические произведения художественной литературы. Нередко книги одного и того же писателя выходят с довольно большими временными интервалами. Допустим, что автор писал свои произведения в течение пятидесяти лет, последняя книга вышла сто лет назад. Предположим, что эти произведения переводил один переводчик и работа заняла двадцать лет. Первый перевод вышел тридцать лет назад, последний — десять лет назад. В синхронический ПКТ не попадет ни один из этих переводов, поскольку автор и переводчик не являются современниками. Ведь если создатели корпуса ориентируются на синхронический корпус, то и тексты на ИЯ, и тексты на ПЯ должны быть изданы если не в один год, то по крайней мере с интервалом менее пяти лет. Представляется, что концентрирование внимания только на такого рода «злободневных» переводах обедняет объект исследования.

Одна и та же книга может быть переведена несколько раз. Нужно ли ограничиться лишь одним переводом? Для теории перевода как раз такого рода многократные переводы представляют большой интерес, да и для сравнительной типологии и контрастивной лингвистики возможность изучать разные варианты перевода одного и того же текста может быть весьма продуктивной.

Таким образом, применение традиционных методик составления корпусов текстов при работе над ПКТ художественных текстов представляется весьма затруднительным. Такого рода собрание текстов скорее всего осуществимо в виде параллельной антологии.

Нам известно о трех проектах, целью которых было создание параллельных корпусов художественных текстов: англо-норвежский параллельный корпус, COMPARA и PLUG (см. раздел 1.3). Ни в одном из полученных ПКТ не удалось до конца выдержать классические принципы составления корпусов текстов. Так, корпус COMPARA составлялся из целых текстов, создатели англо-норвежского корпуса брали фрагменты текстов по единственной причине: для того, чтобы сделать корпус небольшого объема, включающий большое количество текстов разных жанров, написанных разными авторами. «Синхроничности» в классическом смысле (т.е. тексты должны быть изданы в течение одного-двух лет) также, насколько нам известно, добиться не удалось.

Глава 2. «ПарРус» — русско-финский параллельный корпус художественных текстов

«ПарРус» — параллельный корпус русских художественных текстов и их переводов на финский язык. Работа над корпусом была начата в 1999 году на Отделении переводоведения Тамперского университета. Корпус пополнялся усилиями преподавателей, сотрудников и студентов кафедры русского языка. В настоящее время объем русского субкорпуса составляет около 2 270 000 словоупотреблений, объем финского — около 2 240 000 словоупотреблений. Для обслуживания корпуса был разработан пакет программ «КОКОС-П».

В этой главе будут кратко сформулированы общие принципы формирования данного корпуса текстов и обрисовано состояние работ по «ПарРус» на текущий момент. Отдельно будет описано программное обеспечение и основные операции, которые с его помощью можно выполнять.

2.1. Перевод художественной литературы с русского на финский: краткий исторический экскурс

Прежде чем начинать исследование, важно иметь четкое представление об объекте исследования. К счастью, переводы художественной литературы с одного языка на другой представляют собой более или менее обозримые множества текстов и в идеале для каких-то пар языков вполне возможно собрать все существующие на данный момент времени переводы с языка А на язык Б. Тем не менее, в любом случае на предварительном этапе исследования полезно выяснить, какие тексты переводились часто, какие — редко, какие никогда не переводились. Важным этапом представляется выделение основных периодов развития литературного перевода для данных языков. Все это может помочь выработать критерии отбора текстов для корпуса. Итак, попытаемся выяснить, каковы были основные тенденции в переводе с русского языка на финский за сто тридцать с небольшим лет,

прошедших с момента опубликования первого перевода с русского языка на финский.

В настоящем исследовании использовалась база данных по переводам с русского на финский язык, основой для которой послужила библиография, любезно предоставленная проф. П. Песоненом и Б. Хеллманом (Хельсинкский университет, кафедра русского языка и литературы). Основной массив данных был собран Анной Хейнямаа. Исходный список был импортирован нами в базу данных Microsoft Access, а затем дополнен из Интернет-каталогов библиотек Финляндии. Полученная библиография была выверена, добавлены некоторые дополнительные поля, уточнены названия оригиналов. Работа над библиографией еще не завершена, однако имеющиеся данные позволили выяснить основные тенденции и выработать стратегии работы над корпусом.

Важным критерием для оценки продуктивности какого-либо исторического периода в плане переводческой деятельности является количество переведенных текстов и их объем. Единственной мерой количества, зафиксированной в нашем каталоге, было количество страниц в книге. Собственно страницу нельзя считать стабильной единицей измерения, поскольку количество знаков на странице зависит от формата, размера шрифта и т.п. Таким образом, в нашем случае использование точных чисел ни в коей мере не помогает работе. Кроме того, для многих произведений информация о количестве страниц в библиографии отсутствовала. Поэтому было решено присваивать текстам различные веса в зависимости от объема произведения (см. табл. 1). Использовавшиеся при этом критерии были чисто произвольными, за основу было взято общее представление о том, сколько страниц в коротком, длинном или очень длинном тексте.

Таблица 1. Весовая мера для текстов корпуса

Количество страниц в тексте	**Эмпирические представления об объеме текста**	**Вес**
1–49	малый	1
50–149	средний	50
150–500	большой	150
500+	очень большой	500

После того, как всем текстам библиографии были присвоены веса, стало достаточно легко сравнивать общие объемы текстов, например, для того, чтобы определить долю переводов книг разных писателей, вклад разных переводчиков, продуктивность разных исторических периодов. В работе будет использоваться термин **текстовая масса**, под которым понимается сумма весов всех текстов, относящихся к какой-либо группе. В некоторых случаях будут использоваться и данные по количеству наименований опубликованных книг.

2.1.1. Общие тенденции

Сколько книг переводится? Книги каких авторов переводят чаще? Какие темы и жанры пользуются наибольшей популярностью? Это зависит в первую очередь от отношения к другому народу. Можно ли у них чему-нибудь научиться? Интересна ли их культура и их проблемы?

Роль художественных переводов огромна: переводчики выступают в роли «почтовых лошадей просвещения» (Peuranen 1985), заполняя лакуны принимающей культуры, принося в нее новую лексику и идиоматику. Достаточно вспомнить, что славянская культура начиналась с переводов текстов Священного Писания, выполненных Кириллом и Мефодием. Не менее важную роль сыграли переводы и в развитии финского литературного языка и языка художественной литературы (см. Paloposki 2001).

Вполне очевидно, что в развитии перевода художественной литературы — как в целом, так и в нашем конкретном случае — этот «политический фактор» играет весьма значительную роль. Достаточно сравнить отношение к русскому языку и культуре в конце XIX века и во время советско-финской войны 1939 года. Наличие интереса к другой стране в определенной степени оказывают влияние и на людей, которые пришли в книжный магазин или в библиотеку, и на издателя, решающего вопрос о том, перевод какой книги он будет заказывать.

Однако еще более важно наличие в литературе другой страны книг, перевод которых может обогатить принимающую культуру. Поэтому книги продолжают переводиться даже когда отношения между странами очень плохие. Достаточно вспомнить, что в Советском Союзе в разгар холодной войны выходило много переводов с английского языка, хотя издательства и должны были каждый раз объяснять, что тот или иной писатель — «прогрессивный».

По мнению Э. Пеуранена, перевод русской художественной литературы на финский язык сыграл в развитии финской культуры особую роль, особенно в конце XIX века, когда шел процесс становления финского литературного языка. Дело в том, что хотя первые письменные тексты на финском языке появились еще в XVI веке (перевод Библии на финский язык, сделанный Микаэлом Агриколой), финский язык существовал преимущественно в качестве языка устного общения. Художественная литература на финском языке появляется только в XIX веке.

В качестве официального языка, языка культуры и образования и в период шведского владычества, и во времена Великого княжества Финляндского использовался шведский язык (русский язык так и не получил в Финляндии распространения) (Saari 1997). Только со второй половины XIX века финский язык становится вторым официальным языком Финляндии. В это же время начинается развитие национальной литературы, и одновременно публикуются первые переводы на финский язык. И это совпадение не случайно. Переводы с русского стали как бы подготовительным

этапом, предшествовавшим появлению собственной литературы. Нередко переводом художественной литературы на финский язык занимались именно писатели (например, Илмари Каламниус-Кианто (Ilmari Calamnius-Kianto) и Арвид Ярнефельт (Arvid Järnefelt)). Молодому языку нужны были образцы, по которым можно было бы построить собственную литературу. Таким образцом и стала русская литература (Peuranen 1985).

По мнению Э. Пеуранена и Т. Тихменевой, русская культура воспринималась в Финляндии как «другая молодая культура», в том смысле, что литература в европейском смысле этого слова (то есть светская литература) появилась в России в конце XVIII — начале XIX вв. (а этому, в свою очередь, предшествовал период переводов немецкой, французской и английской литературы на русский язык). В то же время русская литература уже успела получить мировое признание, что также способствовало ее успеху в Финляндии (Tihmeneva 1985). Интересно, что в основном переводилась проза, доля поэзии и драматургии в переводах с русского на финский язык чрезвычайно мала. Великие русские поэты Пушкин и Лермонтов больше известны в Финляндии как прозаики, их поэтические произведения были переведены на финский язык намного позже и далеко не все. Интересно, что в соседней Швеции поэзия Пушкина переводилась одновременно с прозой. «Перекос» в сторону прозы сохраняется в Финляндии и в настоящее время (Peuranen 1985).

За периодом увлечения русской литературой, переводами произведений Пушкина, Тургенева, Гоголя, Толстого, Достоевского и Горького последовал длительный спад: после того, как Финляндия получила независимость в 1917 году, интерес к России и Советскому Союзу сохранился только у «левых». Затем началась советско-финская война. Во Второй мировой войне Советский Союз и Финляндия опять оказались «по разные стороны баррикад».

Только после заключения мира в 1944 году количество издаваемых переводов с русского языка начало постепенно увеличиваться. Определенную роль здесь сыграло и Общество Дружбы Финляндия — СССР, которое было основано в начале 40-х годов. В 50-х годах в Финляндии опять наблюдается заметный спад в издании переводов с русского языка. Это в какой-то степени связано с началом холодной войны. Зато период с 1960-х по конец 1980-х, несомненно, является пиком переводов с русского языка. В то время ежегодно переводилось в среднем по пятнадцать наименований произведений русской и советской литературы. В 1976 году финские издательства начали выпускать серию «Советская литература» (Neuvostokirjallisuus). Эта серия издавалась до 1987 года, в ней вышло 81 наименование произведений советской литературы (Pitkänen 1999: 22).

В начале 1990-х годов издание переводов с русского языка катастрофически упало. Например, в 1995 году в Финляндии были изданы лишь четыре новых перевода с русского языка (Pitkänen 1999: 23). Причин

этому много: это и распад Советского Союза, и последовавший за этим политический и экономический кризис в России, и экономический кризис в Финляндии, и «переориентация» Финляндии на ЕС. В настоящее время количество переводов с русского языка несколько возросло, но ситуация нестабильная, интерес к русской литературе упал, и вообще переводы с английского вытесняют переводы с других языков (подробнее см. Saari 1989, Pitkänen 1999).

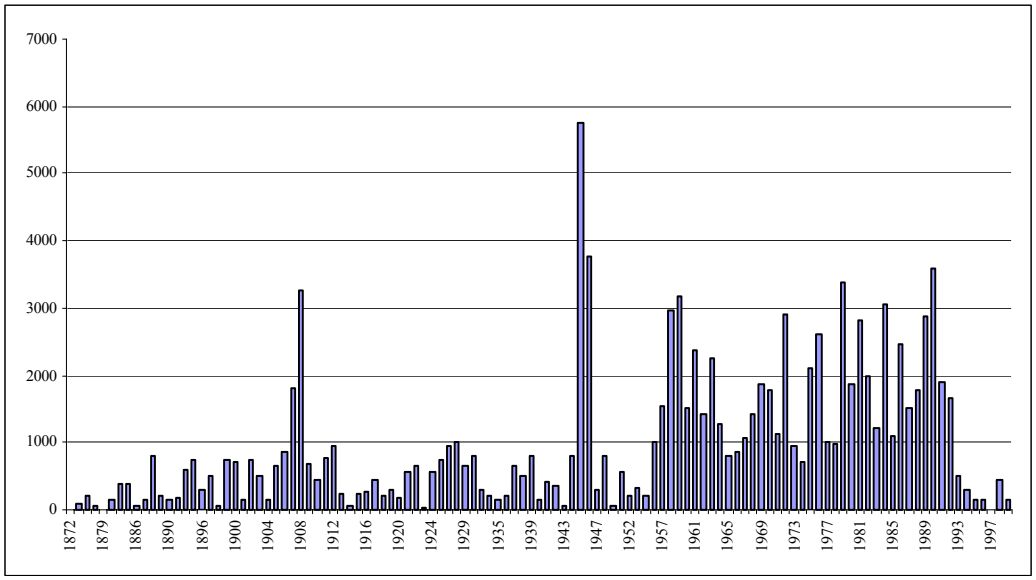

Рис. 3. *Динамика публикации переводов художественной прозы с русского языка на финский с 1870-х по настоящее время (текстовая масса)*

Вышеизложенное наглядно иллюстрируется на рис. 3 и 4. На рис. 3 показана текстовая масса всех изданных за каждый год произведений. На рис. 4 также показана динамика издания переводов русской прозы, но по количеству наименований (отдельное издание / сборник). Переиздания одного и того же перевода в обеих диаграммах в расчет не принимаются.

Обе диаграммы показывают аналогичные тенденции и подтверждают высказанные выше соображения: отражается и становление перевода с русского языка в период автономии с пиком в 1910-е гг, спад 20-х — 30-х гг., неожиданный пик издания русских переводов в 1945–46 гг., и резкий спад 50-х, и «золотое время» переводов с русского языка 1955–1990.

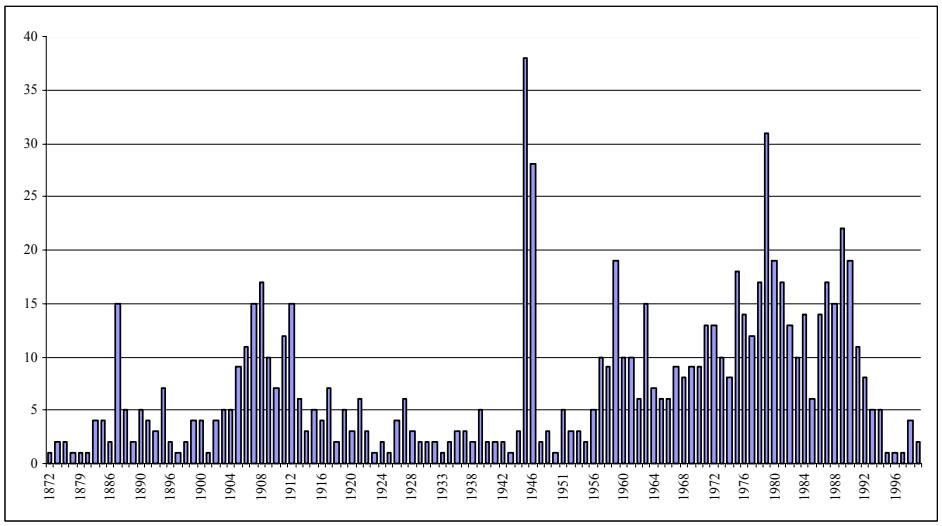

Рис. 4. *Динамика публикации переводов художественной прозы с русского языка на финский с 1870-х по настоящее время (количество опубликованных книг)*

Таким образом, суммируя вышесказанное, можно условно выделить в истории художественного перевода с русского языка на финский следующие периоды:

a. 1870-е — 1912. Зарождение традиций художественного перевода с русского языка. Период открывается публикацией первых переводов художественной литературы с русского языка в начале 1870-х гг. и заканчивается в начале 1910-х годов, когда количество переводов с русского языка резко сокращается.
b. 1913 — 1943. Интерес к России и русской литературе после получения независимости невысок. Черная полоса в российско-финских отношениях. Две войны.
c. 1944 — 1955. Кратковременное оживление в издании переводов с русского языка в конце 1945–46 гг., за которым следует новый спад.
d. 1956 — 1992. Пик в издании переводов с русского языка.
e. После 1993. Очередное уменьшение количества издаваемых переводов с русского языка.

В последующих разделах попытаемся кратко охарактеризовать каждый из вышеназванных периодов с точки зрения публикации переводов художественной литературы с русского языка.

2.1.2. Первый период (1870-е — 1912)

В 1809 г. Швеция проиграла войну России, и Финляндия вошла в состав Российской империи в качестве автономного великого княжества. Политика России на новоприобретенных территориях была на удивление разумной: Финляндия получила особый статус, и во второй половине XIX столетия у Великого княжества Финляндского были многие права, о которых остальные российские провинции не могли даже и мечтать, например, своя валюта и финский язык в качестве одного из официальных языков. Русский язык в Финляндии пропагандировался, однако целенаправленную политику руссификации (заключавшуюся в обязательном изучении русского языка в школах, ограничении использования шведского и финского языков и т.п.) начали проводить только при Николае II, на рубеже XIX и XX веков, что, возможно, и привело к активизации борьбы за независимость Финляндии.

Первые переводы произведений русской художественной литературы на финский язык появились только в начале 1870-х гг., поскольку в первой половине XIX века в Финляндии доминировал шведский язык. Поэтому первыми переводами русской литературы, изданными в Финляндии, были переводы на шведский язык (Peuranen 1985, Tihmeneva 1985, Нумминен 1997). Финский язык начал укреплять свои позиции только в середине века. Первым художественным произведением большого объема, переведенным с русского языка на финский, стала повесть А.С. Пушкина «Капитанская дочка» (1876, переводчик — Самули Суомалайнен (Samuli Suomalainen))[8].

В течение этого периода с русского языка переводились произведения примерно сорока авторов. Наиболее популярными были Л.Н. Толстой, Ф.М. Достоевский, И.С. Тургенев, Л. Андреев, М. Горький, И.А. Гончаров, И. Потапенко, Н.В. Гоголь, В. Крестовский, А.С. Пушкин. Среди признанных классиков русской (а иногда — даже мировой) литературы стоят особняком два писателя. Это Всеволод Крестовский, о котором в России вспомнили только в конце XX века, и Игнатий Потапенко, довольно популярный в свое время писатель, в настоящее время совершенно забытый.

Переводы произведений трех авторов, возглавляющих список — Толстого, Достоевского и Тургенева — составляют 49% от всей текстовой массы этого периода. Интересно, что Толстой занимает первое место в списке не только потому, что он писал длинные романы и все они уже тогда были переведены, а еще и потому, что на финский язык переводилась и практически все статьи и трактаты Толстого на политические, философские и религиозные темы (Нумминен 1997).

---

[8] Интересно, что роман Алексиса Киви «Семеро братьев» (Aleksis Kivi, Seitsemän veljestä), первое крупное произведение финской художественной литературы, увидел свет почти в то же самое время, в 1870 г.

Довольно большое количество переводов с русского языка на финский объясняется целым рядом факторов, важнейшими из которых были вхождение Финляндии в Российскую империю, связи русской и финской интеллигенции (например, дружба Льва Толстого и Арвида Ярнефельда), использование русской литературы в качестве образца для национальной литературы.

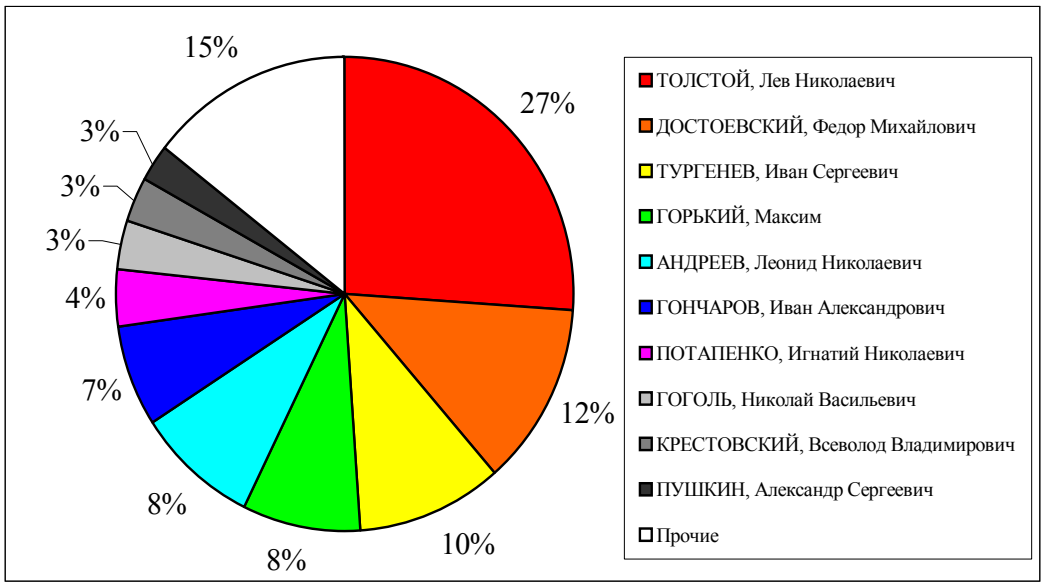

Рис. 5. *Русские писатели, наиболее популярные в Финляндии в период Великого княжества Финляндского*

Около семидесяти переводчиков занимались в этот период художественным переводом с русского языка. Из них наиболее известны Арвид Ярнефельт (Arvid Järnefelt), Самули Суомалайнен (Samuli Suomalainen, Samuli S.), Антон Хелве (Anton Helve), Мартти Вуори (Martti Wuori), Илмари Каламниус-Кианто (Ilmari Calamnius-Kianto), Эйно Калима (Eino Kalima) и Ялмари Аалберг (Jalmari Aalberg).

2.1.3. Второй период (1913 — 1943)

После 1912 года интерес к русской литературе в Финляндии заметно падает, возможно, это связано и с кампанией руссификации, и с ростом освободительного движения, и с началом Первой мировой войны. После провозглашения Финляндией независимости в 1917 г. российско-финские отношения постепенно ухудшались, хотя левые и симпатизировали СССР — первому в мире государству рабочих и крестьян. К концу 30-х годов отношения между странами обострились, что кончилось Советско-финской

войной 1939 года (Talvisota) и участием Финляндии во Второй мировой войне на стороне Германии (Jatkosota).

Весь период характеризуется спадом переводческой деятельности, это хорошо видно на рис. 3 и 4. Все же нельзя утверждать, что переводов с русского языка не делалось вовсе. Ведь идеи социал-демократии и в то время были в Финляндии достаточно популярны, а Советский Союз, первое в мире социалистическое государство, был для многих финнов привлекателен. За этот период на финский язык было переведено более четырехсот произведений русской литературы.

Собственно количество авторов, произведения которых переводились, по сравнению с предыдущим периодом не выросло. Список авторов, вошедших в первую десятку, мало отличается от предыдущего периода. По-прежнему много переводились только два классика — Ф.М. Достоевский и Л.Н. Толстой. Однако начинают переводить и новых авторов — М.А. Шолохова, А.Н. Толстого, Ф. Гладкова, Д. Мережковского (поистине удивительно появление в одном списке столь разных писателей).

В течение этого периода, как и следовало ожидать, интерес издателей и переводчиков сосредоточился на классической русской литературе. Доля литературы XIX века и рубежа веков составила 71% текстовой массы.

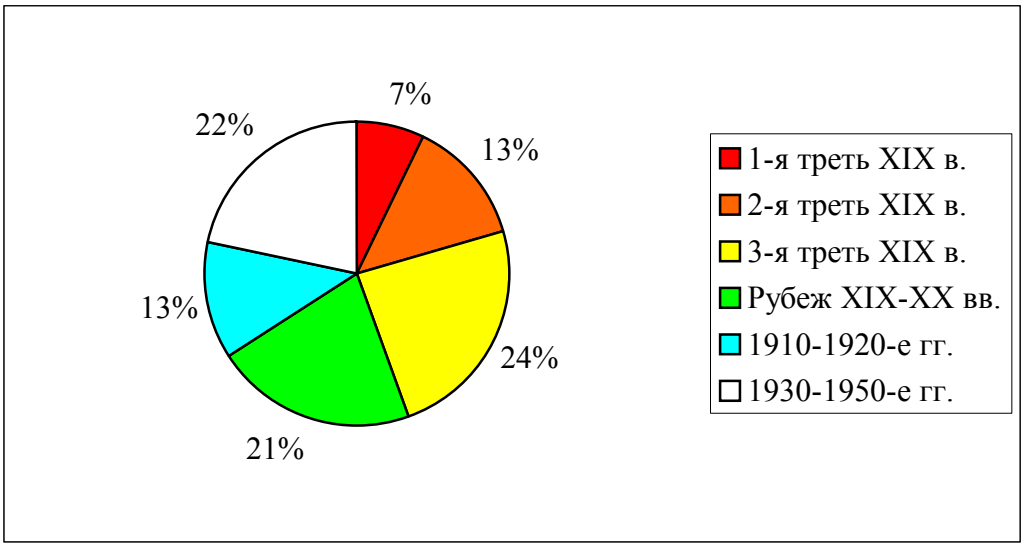

Рис. 6. *Диаграмма 2.5. Русская литература, переводившаяся на финский язык в период 1913–1943 гг.*

Среди переводчиков этого времени наиболее активно работали Виктор Траст (Viktor Kustaa Trast), Юхо Холло (J. A. Hollo), Вернер Анттила (Werner Anttila) и Юхани Конкка (Juhani Konkka).

2.1.4. Третий период (1944 — 1955)

В 1944 году между СССР и Финляндией был заключен мир, а четыре года спустя, в 1948 году подписан «Договор о дружбе и сотрудничестве» (YYA sopimus), давший старт новому этапу в развитии отношений между двумя странами.

1940–50-е годы — переходный этап в переводах художественной литературы с русского языка на финский. На рис. 3 и 4 можно заметить резкий скачок вверх, приходящийся на 1944–1945 гг., который затем сменяется новым спадом. Количество русских авторов, произведения которых переводятся на финский язык, продолжает оставаться небольшим.

«Перекос» в сторону литературы более ранних периодов, характерный для предыдущего периода, полностью ликвидирован: современная для того периода литература составляет более 60% от общего объема переводимой с русского языка художественной прозы. В это время переводились произведения пятидесяти авторов. Обращает на себя внимание тот факт, что в это время нет «любимых авторов», единственное исключение — Максим Горький (переведено 27 произведений, в числе которых — романы «Жизнь Клима Самгина» и «Дело Артамоновых»).

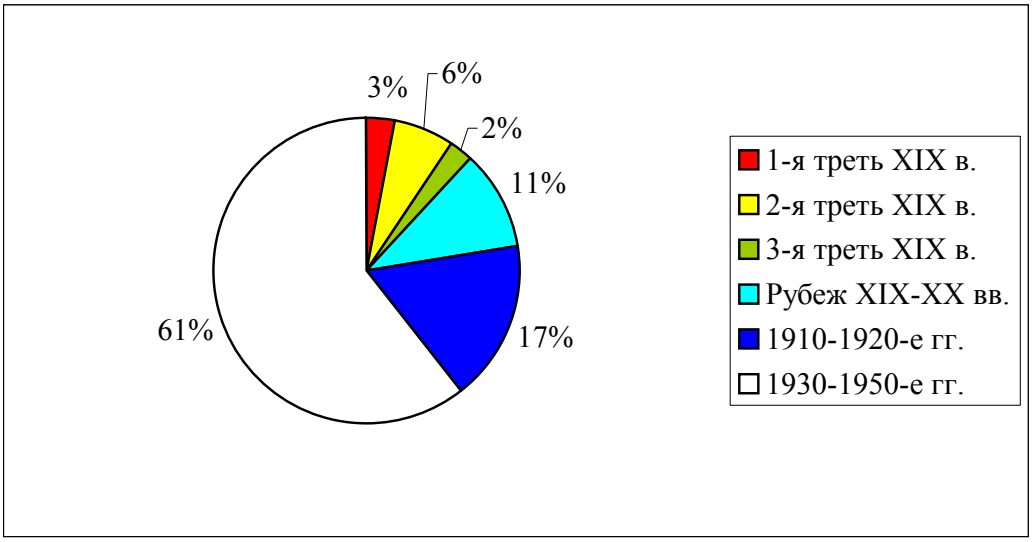

Рис. 7. *Русская литература, переводившаяся на финский язык в период 1944–1955 гг.*

«Застой» в переводческой деятельности второго и третьего периодов связан не только с «похолоданием» в российско-финских отношениях, но и с тем, что в течение первого и в начале второго периода были переведены практически все крупные произведения русской классической литературы

и переводы еще не успели «устареть». Поэтому в переводах третьего периода классическая литература составляет лишь около 11%.

Среди переводчиков этого периода выделяются Юхани Конкка (Juhani Konkka) и Матти Лехмонен (Matti Lehmonen).

2.1.5. Четвертый период (1956 — 1992)

В конце пятидесятых годов начинается неуклонный рост количества произведений русской художественной литературы, переводившейся на финский язык. Начинался период расцвета перевода с русского языка на финский. Только в начале 90-х годов обозначился новый спад.

В это время были переведены произведения более 200 разных авторов. Наиболее популярными были Ф.М. Достоевский, Л.Н. Толстой, А.И. Солженицын, М. Горький, Н.В. Гоголь, И.С. Тургенев, Ч.Т. Айтматов, А.С. Пушкин.

Переводчиков художественной литературы с русского языка в этот период было более ста, из них четверо — Юхани Конкка (Juhani Konkka), Эса Адриан (Esa Adrian), Улла-Лийса Хейно (Ulla-Liisa Heino) и Леа Пююккё (Lea Pyykkö) перевели около 56% всей текстовой массы этого периода (Конкка — 20%, Адриан — 17%, Хейно — 14%, Пююккё — 5%, см. рис. 8).

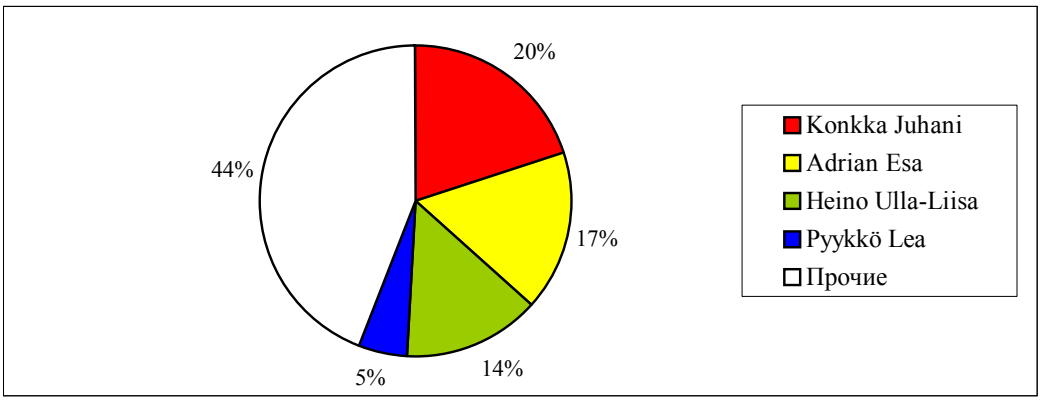

Рис. 8. *Переводчики с русского языка в период 1956–1992*

Интересная особенность этого периода заключается в том, что некоторые произведения русской классики переводились по нескольку раз. «Любимыми» стали «Преступление и наказание» и «Идиот» Достоевского, «Анна Каренина» Толстого, «Шинель» Гоголя. При этом другие произведения нередко оказывались незаслуженно забытыми. Приведем только

один пример: роман Гончарова «Обломов» переводился четыре раза, «Обыкновенная история» — один раз, «Обрыв» — ни разу.

Другая характерная особенность периода состоит в том, что финские издатели (а вслед за ними, вероятно — и финские читатели) начинают переключаться с классической литературы XIX века на советскую послевоенную литературу, доля которой приближается к половине текстовой массы всех переводов с русского языка этого периода. Однако и классическая литература по-прежнему занимает важное место: это около 26% всей текстовой массы (см. рис. 9). Среди наиболее переводимых авторов по-прежнему большинство — писатели XIX века.

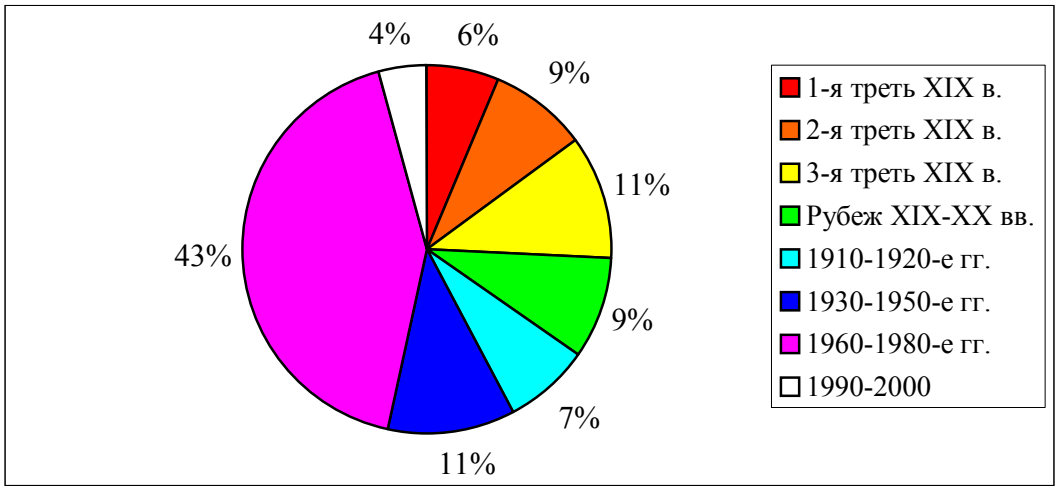

Рис. 9. *Русская литература, переводившаяся на финский язык в период 1956–1992*

2.1.6. Пятый период (1993 — настоящее время)

После распада Советского Союза роль России в экономике Финляндии резко пошла на убыль. Договор о дружбе и сотрудничестве не был продлен, многие из российско-финских контрактов прекратили свое существование. Политический и экономический кризис в России и странах СНГ тоже не способствовал упрочению отношений. В 1995 году Финляндия вступила в Европейский Союз. С 1980 по 1999 согласно данным Статистического центра Финляндии неуклонно росла доля переводов с английского языка с одновременным уменьшением доли переводов с русского и других европейских языков: доля английского языка выросла с 46,6% до 69,4%, доля русского языка упала с 4,9% до 0,9%, шведского — с 17,4% до 8,6%, немецкого — с 8,8% до 6,7%, датского — с 2,2% до 1,0%, доля французского языка осталась неизменной — 5% (см. http://statfin.stat.fi/statweb/). В этой статистике учитывались все публиковавшиеся переводы, то есть и

художественная литература, и публицистика, и научная литература, доля определялась количеством наименований, однако тенденция является достаточно показательной.

Уменьшение количества издаваемых в Финляндии переводов с русского языка не может быть сведено только к политическим и экономическим факторам. Достаточно важными представляются следующие три причины:

1. За длительным ростом обязательно следует спад, который в нашем случае был абсолютно закономерен после сорокалетнего периода, в течение которого было издано огромное количество переводов с русского языка.

2. В течение предыдущего четвертого периода были сделаны новые переводы всех крупных произведений русской классической литературы (Гоголь, Толстой, Достоевский, Пушкин и др.). Вообще, многие произведения русской классики переводились несколько раз. Например, «Идиот» Достоевского был переведен трижды, причем первый перевод был издан в 1929 году (переводчик — Виктор Траст (V.K. Trast)), а два других вышли с интервалом в десять лет: второй в 1968 году (переводчик — Юхани Конкка (Juhani Konkka)), а третий — в 1979 году (переводчик — Леа Пююккё (Lea Pyykkö)). Аналогичная ситуация и с романом Толстого «Анна Каренина»: первый перевод вышел в 1911 году (переводчик — Эйно Калима (Eino Kalima)), второй — в 1969 (переводчик — Улла-Лийса Хейно (Ulla-Liisa Heino)), третий — в 1979 (переводчик — Леа Пююккё (Lea Pyykkö)). В результате «рынок» оказался заполненным на много лет вперед, до того момента, пока изданные в 1970–80 гг. переводы не устареют. Срок жизни перевода обычно составляет порядка 50 лет, после чего, как правило, выполняется новый перевод (см., например, Landers 2001: 10).

3. В самой русской литературе начала 1990-х годов был кризис: «живые классики» публиковались мало, а новых имен почти не появилось. Вообще, современная русская литература — уже другая литература, совсем не похожая на советскую литературу 1960–1980-х гг. Поэтому финские переводчики и издатели какое-то время, по-видимому, просто не могли решить, что может заинтересовать финского читателя. В результате, в основном переводились произведения тех русских писателей, которых переводили на другие европейские языки, а именно произведения А. Битова, Т. Толстой, Л. Петрушевской, В. Пелевина и др.

В конце 1990-х — начале 2000-х годов отношения с Россией опять начинают меняться в лучшую сторону, что связано с постепенной политической и экономической стабилизацией в РФ, а также наличием общих интересов в сфере политики, экономики и культуры. Об этом можно судить, например, по частотности упоминания России в финском Интернете. На финском сайте поискового сервера «Altavista»

(http://fi.altavista.com)[9] 10.07.2003 был выполнен поиск с целью выяснить количество упоминаний разных стран в ФИНЕТ. Кроме России были проверены другие ближайшие соседи Финляндии и Соединенные Штаты. При поиске задавались шаблоны, позволявшие искать как название страны, так и прилагательные, от него образованные (например, для поиска упоминаний Швеции применялись шаблоны *ruotsi** и *ruotsalai**). Результаты поиска представлены в таблице 2. Поиск показал, что Россия в финском Интернете упоминается довольно часто, частота упоминаний приближается к США и Германии, хотя и отстает от другого ближайшего соседа — Швеции. (Впрочем, такое большое количество страниц с с упоминанием Швеции, скорее всего, объясняется тем, что шведский язык является вторым государственным языком Финляндии, в связи с чем очень многие интернет-страницы выполняются на двух языках, и в результате слова *ruotsi* и *ruotsiksi* могут использоваться просто в качестве линка к шведскому варианту страницы или указанием на язык документа).

Таблица 2. Упоминание разных стран в ФИНЕТ

Страна	**Шаблоны поиска**	**Количество найденных страниц**
США	usa, yhdysval*, amerikkalai*	192 919
Россия	venäjä*, venäläi*	166 099
Швеция	ruotsi*, ruotsalai*	223 535
Дания	tanska*	60 880
Норвегия	norja*	67 094
Германия	saksa*	184 178
Эстония	viro*	74 656

Другим фактором, способствующим развитию российско-финских отношений, является наличие в Финляндии русскоязычного национального меньшинства, численность которого стала быстро расти в начале 1990-х годов, в связи с программой репатриации ингерманландцев, «русских финнов». Например, по статистике, предоставленной Управлением по делам иностранцев Финляндии (Ulkomaalaisvirasto), из 4730 человек, получивших финское гражданство в 1999 году, 800 были гражданами России, 379 — гражданами Эстонии и 135 — гражданами бывшего СССР или стран СНГ; в 2000 году финское гражданство получили 2977 человек, из которых 666

[9] Сервер *Altavista* был выбран по той причине, что он позволяет выполнять поиск по шаблону. Серверы *Google* и *Ihmemaa* такой возможносити не предоставляют, а это сильно затруднило бы нашу задачу.

были гражданами России, 353 —гражданами Эстонии и 48 — граждане бывшего СССР или стран СНГ (http://www.uvi.fi/ajankoht.html, 27.11.2002).

Из классики издавались Гоголь (вышел новый перевод «Вечеров на хуторе близ Диканьки», переводчик — Юкка Маллинен (Jukka Mallinen)), Пушкин (только поэтические произведения — в 1999 г., к 200-летию со дня рождения поэта, впервые переведены сказки и многие поэмы, всего 13 произведений[10]), вышли новые переводы некоторых произведений Льва Толстого («Хаджи-Мурат», «Смерть Ивана Ильича») и Тургенева («Первая любовь»).

В переводившейся в это время с русского языка литературе заметное место занимает детская литература: издается очень много произведений Эдуарда Успенского. Переводы книжек для детей про крокодила Гену и Чебурашку и, особенно, про дядю Федора имели огромный успех. Повесть «Дядя Федор, пес и кот» была переведена на финский язык еще в 1975 году («Fedja-setä, kissa ja koira», пер. Мартти Анхава (Martti Anhava)) и выдержала 9 изданий. Перевод книги «Крокодил Гена и его друзья» вышел в 1977 году («Krokotiili Gena ja hänen ystävänsä», пер. Мартти Анхава) и выдержал 5 изданий. В 90-е годы были изданы четыре новых перевода произведений Успенского.

В конце 90-х годов увеличивается доля переводов беллетристики. Это связано также с тем, что литературы такого рода в советский период было очень мало. Переведен на финский язык роман А. Марининой «Стечение обстоятельств» («Irina tietää likaa», переводчик О. Куукасъярви (O. Kuukasjärvi)), два романа Б. Акунина — «Азазель» и «Турецкий гамбит» («Asaselin salaliitto», «Turkkilainen gambiitti», переводчик А. Никкиля (A. Nikkilä)).

В настоящее время в Финляндии работает около пятнадцати переводчиков художественной литературы с русского языка. В отличие от четвертого периода, нельзя определить переводчика-«лидера».

2.2. Структура корпуса «ПарРус»

2.2.1. Параметры классификации

Как уже отмечалось выше, составление параллельного корпуса художественных текстов, репрезентативного в традиционном смысле, представ-

[10] Как уже говорилось выше, русская поэзия вообще переводилась на финский язык довольно мало. Так, в нашей библиографии зафиксировано всего 38 переводов поэтических произведений Пушкина (включая стихотворения), все они были выполнены после 1936 г.

ляется проблематичным. С другой стороны, случайная выборка текстов также не может быть решением проблемы. Поэтому мы попытаемся все-таки определить более или менее приемлемые критерии, которые можно было бы использовать при отборе текстов. Корпус, составленный таким образом, не будет полностью аморфным, в нем будет заложена некая идея, у него будет структура. Такой корпус, конечно, нельзя назвать репрезентативным в классическом понимании этого термина, поскольку он будет состоять из целых текстов. Тем не менее, для полнотекстового корпуса, как уже говорилось выше, всегда можно разработать программное обеспечение, которое будет в случае необходимости генерировать корпус образцов из полнотекстового корпуса.

Параметры классификации текстов делятся на **внутренние** и **внешние** (internal vs. external). Внутренние параметры связаны с языком, содержанием и композицией текстов, например, диалогичность / монологичность, нарративность / аргументативность, жанр, тематика. Внешние параметры имеют отношение к экстралингвистическим аспектам текстов — время, автор, страна, пол автора, возраст автора и т.п. (Atkins et al. 1992: 5).

В нашем случае внутренние критерии использовать довольно сложно, ведь корпус составляется из целых текстов, зачастую — довольно больших по объему. Многие из внутренних параметров, например диалогичность / монологичность, нарративность / аргументативность и т.п., можно применять только к текстовым фрагментам, поскольку в полном тексте может быть все: и диалог, и монолог, и повествование, и аргументативный текст. Во многих случаев использование внутренних критериев классификации становится возможным только при наличии текстов в электронном виде. Однако, как уже говорилось выше, решение о включении текста в корпус принимается еще на той стадии, когда электронная версия отсутствует.

По-видимому, один из немногих внутренних параметров, которые могут использоваться при составлении полнотекстовых корпусов текстов — это жанр. В то же время этот параметр зачастую оказывается трудноприменимым, поскольку до сих пор даже в литературоведении не существует общепринятых определений жанров художественной литературы. Другой внутренний параметр, применимый к целым текстам — это длина текста, однако он представляется еще менее продуктивным, чем жанр, поскольку длина текста зачастую слабо связана с его содержанием. Одновременно оба параметра применять нецелесообразно, поскольку они пересекаются: часть жанров относится к «большим», часть — «к малым». По этой причине длина текста будет использоваться лишь для определения доли группы текстов в генеральном множестве всех переводов художественной литературы с русского языка на финский.

На первый взгляд представляется полезным использовать в качестве параметра тематику текста. Этот параметр непосредственно связан с используемой в произведении лексикой. Однако описать содержание текста (то есть сказать, о чем текст), используя один или несколько дескрипторов,

нередко оказывается невозможным, особенно в случае, если мы имеем дело с эссе отвлеченного содержания. Даже в тех случаях, когда подобное описание возможно, это мало помогает в плане определения пластов лексики, задействованной в тексте. Особенно малопродуктивно подобное для больших романов: описание содержания романа «Анна Каренина» через дескрипторы «семья, высший свет, любовь», а «Тихого Дона» — через дескрипторы «война, революция, любовь, семья», — мало что говорят даже о содержании этих произведений, не говоря уже об их лексическом наполнении. Более подробные описания содержания делают каждый текст уникальным и не помогают классификации. Поэтому было решено отказаться от тематики в качестве параметра классификации.

Таким образом, мы используем по большей части внешние параметры, поскольку последние легче определяются и при их присваивании возникает меньше противоречий.

Итак, для корпуса «ПарРус» используется следующий набор параметров:

1. Жанр
2. Время публикации (написания) оригинала
3. Время публикации перевода
4. Рейтинг автора
5. Рейтинг переводчика

Жанр. При составлении корпуса используются традиционно принятые в литературоведении жанры, такие как «эссе», «рассказ», «повесть», «роман» и т.п. Некоторые жанры с течением времени претерпевают заметные изменения. Например, в русской литературе жанр повести первой трети XIX века (например, «Бедная Лиза» Карамзина, «Повести Белкина» Пушкина, «Петербургские повести» Гоголя) довольно сильно отличается от этого жанра в русской литературе середины XX века (например, «Белый Бим, черное ухо» Троепольского, «Дом на набережной» Трифонова, «Батальоны просят огня» Бондарева). Современная повесть становится заметно больше и по объему, и по количеству действующих лиц и скорее похожа на короткий роман, в то время как повесть начала XIX века ближе к рассказу. К счастью, это не создает дополнительных проблем для исследования, поскольку тексты относятся к разным историческим периодам и все равно попадают в разные группы.

Время публикации (написания). В некоторых случаях даже этот внешне очень простой вопрос может оказаться совсем не простым. Например, повесть Булгакова «Собачье сердце» была написана в 1925 году, однако в России впервые опубликована только в 1987 году. Роман Пастернака «Доктор Живаго» написан в 1955 году, опубликован за рубежом в 1957 году, но в России — только в 1988 году. Список можно продолжать. Для подобных случаев этот параметр будем понимать как время написания произведения, поскольку оно предназначалось автором

для современников и было написано в литературных традициях своего времени, в соответствии с тогдашним состоянием языка (что особенно важно для лингвистических исследований), а публикация через полстолетия — лишь трагическая случайность.

Использовать в качестве этого параметра точный год написания (опубликования) нецелесообразно, поскольку это приведет к появлению очень большого количества значений. Поэтому было решено разработать «рабочую» периодизацию как для оригиналов, так и для переводов.

Вопрос о периодизации русской литературы, как и все, что связано с любой периодизацией в искусстве, является чрезвычайно сложным. В настоящее время в российском литературоведении критикуются традиционно выделявшиеся периоды развития русской литературы, связанные с ленинскими «периодами освободительного движения». Особенно трудно поддается периодизации литература XX века, хотя бы потому, что наряду с литературой социалистического реализма была и литература, не вписывающаяся в это направление (Грин, Ильф и Петров, Булгаков, Паустовский, Пастернак и др.), и «андерграунд», и эмигрантская литература (подробнее см., например, Баевский 1999, Белая 1990, Чудакова 2001).

При разработке периодизации художественной литературы могут учитываться разные факторы: политическая ситуация в стране, политика в области культуры, господствующий художественный метод, литературные школы и кружки, творчество наиболее известных писателей. По-видимому, ни один из подходов не может дать до конца последовательной периодизации. Дополнительную проблему в нашем исследовании создает обязательность определения точных временных рамок: расплывчатые рамки типа «начало 1840-х гг.» непродуктивны при выполнении анализа с помощью компьютера.

Поскольку корпус текстов будет использоваться в первую очередь для выполнения лингвистических и переводоведческих исследований, принципиально важно, чтобы периодизация отражала именно изменения в языке. Поэтому, разрабатывая предлагаемую ниже периодизацию, мы в первую очередь учитывали не столько «политические вехи», сколько смену поколений писателей. Используя хронологию публикации произведений[11] писателей с наиболее высоким рейтингом (о рейтингах см. ниже), мы выделили следующие периоды в истории русской литературы XIX-XX веков[12]:

1. До 1842 г. (Пушкин, Лермонтов, Гоголь)
2. 1843 — 1863 (Гончаров, Тургенев)

[11] Речь идет, разумеется, о первых изданиях на языке оригинала.

[12] Периодизацию начинаем с XIX века, поскольку произведения более ранних периодов на финский язык переводились очень мало (переведены «Слово о полку Игореве», «Повесть временных лет», «Недоросль» Фонвизина, некоторые оды Державина и Ломоносова) и в любом случае, материал для ПКТ дать вряд ли смогут.

3. 1864 — 1883 (Толстой, Достоевский)
4. 1884 — 1912 (Чехов, поздний Толстой, Андреев, Горький)
5. 1913 — 1928 (Горький, Булгаков)
6. 1929 — 1955 (Шолохов, Алексей Толстой, Леонов)
7. 1956 — 1986 (Паустовский, Симонов, Распутин, Солженицын)
8. 1987 — настоящее время. (Битов, Искандер, Евгений Попов)

Фамилии писателей, указанные в скобках, не означают, что данный период сводится только к этим писателям, а только то, что временные рамки периода определены главным образом по хронологии публикаций произведений этих писателей. Выбор писателей ни в коей мере не отражает наши вкусы или симпатии; он обусловлен лишь тем, что произведения именно этих писателей больше переводились на финский язык.

При определении временных рамок мы стремились к тому, чтобы писатели с наиболее высоким «рейтингом» не выходили за рамки «своего» периода. В некоторых случаях это удалось, например, 1-й период кончается в 1842 году публикацией «Мертвых душ» Гоголя. Творчество некоторых писателей принадлежат сразу нескольким периодам (Тургенев, Лев Толстой, Горький и др.), в этих случаях мы старались при выделении периодов учитывать годы публикации центральных произведений этих авторов.

Вопрос о периодизации финского литературного перевода обсуждался в предыдущем разделе (см. стр. 44 настоящей работы). В итоге были выделены следующие пять периодов:

1. 1870-е — 1912
2. 1913 — 1943
3. 1944 — 1955
4. 1956 — 1992
5. 1993 —

Корпус текстов, в котором были бы представлены все периоды перевода с русского на финский, был бы интересен во многих отношениях. Однако выполнение такой масштабной работы в данный момент не представляется возможным. Поэтому было решено ограничиться переводами, изданными начиная с третьего периода.

Рейтинги авторов и переводчиков. Для того чтобы оценить значимость того или иного писателя или переводчика для принимающей культуры, применялась весовая мера, то есть вычислялись рейтинги по текстовой массе всех переводов, выполненных с произведений данного автора или сделанных данным переводчиком. Использовалась следующая формула:

$$R = Int(\lg \sum m)$$

где R — рейтинг, m — «масса» текста. Логарифм использовался для того, чтобы избежать сильных перепадов коэффициента. Результат округлялся до целых. В итоге рейтинги всех авторов и переводчиков, фигурирующих в библиографии, варьируются от 0 до 4.

Более высокий или более низкий «рейтинг» автора или переводчика ничего не говорит о литературных достоинствах их произведений (или переводов), а является лишь показателем того, насколько часто данный автор переводился на финский язык или насколько много издавалось переводов данного переводчика. Например, рейтинг Лермонтова оказался равным 2, поскольку основную часть его наследия составляет поэзия, которая переводилась довольно мало, а из прозы переведен только «Герой нашего времени», правда, роман переведен на финский язык трижды полностью и два раза частично («Бэла» и «Максим Максимыч»). Аналогично, рейтинги Пушкина и Чехова, несмотря на огромную роль, которую они сыграли в развитии русской литературы, равны 3, поскольку ни Пушкин, ни Чехов не писали длинных романов, и текстовая масса произведений, переводившихся на финский язык, оказывается небольшой по сравнению с произведениями Толстого или Достоевского.

2.2.2. Классификация текстов

Все все художественные прозаические тексты базы данных были сгруппированы по указанным выше пяти параметрам (см. стр. 55). В итоге получилось 436 групп; из них переводы, вышедшие после войны, образуют 282 группы. Внутри каждой из выделяемых групп все параметры текстов совпадают. Значимость каждой группы оценивается по ее текстовой массе, т.е. по сумме масс текстов, в нее входящих (количество наименований представляется менее надежным критерием, поскольку в этом случае группы, в которые входят большие романы, могут быть весьма малочисленны, но из этого не следует, что их значимость мала). Таким образом,

$$M = \sum m,$$

где M — текстовая масса группы, m — масса отдельного текста.

Как и следовало ожидать, многие из выделенных групп представлены единичными текстами малого объема. Это позволяет отбросить малые группы и ориентироваться при составлении корпуса по возможности на группы с большой текстовой массой. Из выделенных 282 групп текстов 64 группы с текстовой массой более 400 составляют 70% общей текстовой массы всех переводов. Таким образом, корпус текстов, составленный из текстов этих групп, должен отражать основные тенденции языка оригинальных текстов и переводов.

Полный список 64 наиболее значимых групп с их описаниями дан в приложении 1 (таблица 1). Приведем здесь только описания первых трех групп.

Группа 1. Текстовая масса — 3350, в нее вошло 27 повестей, относящихся к 7-му периоду (1956 — 1986), написанных авторами с рейтингом 2 (Виктор Астафьев, Сергей Залыгин, Фазиль Искандер и др.), переведенных на финский язык в 4-й период (1956 — 1993) переводчиками с рейтингом 3 (Улла-Лийса Хейно, Марья Коскинен и др.)

Группа 2. Текстовая масса — 3250, в нее вошло 17 романов, относящихся к тому же 7-му периоду, написанных авторами с рейтингом 2, переведенных на финский язык в 4-й период переводчиками с рейтингом 3. Таким образом, единственное отличие от группы 1 — в жанре.

Группа 3. Текстовая масса — 3000, в нее вошло 6 больших романов, написанных в 3-й период (1864 — 1883) авторами с рейтингом 4 (в группу попали только два писателя — Ф.М. Достоевский и Л.Н. Толстой), переведенных на финский язык в 4-й период переводчиками с рейтингом 3.

Полные списки текстов первых трех групп см. в приложении 1.2.

Большие группы включают в себя 10 и более текстов. Если в корпус будет включаться примерно десятая часть каждой группы (т.е. 5–10 текстов из больших групп, 1–2 текста из малых групп, из групп, в которые входит менее пяти текстов — один текст или ни одного текста), то мы получим достаточно большой по объему корпус, в котором пропорционально представлена вся популяция текстов.

Однако, составляя корпус, мы отдавали себе отчет, что предложенные критерии могут быть неточными, а данные, которые использовались, могут быть неполными. Поэтому, хотя мы и ориентировались в основном на тексты из больших групп, в некоторых случаях брали и тексты из групп с массой менее 400, особенно если речь шла о переводах одного и того же текста, выполненных разными переводчиками. Такого рода материал может быть чрезвычайно интересным и полезным, особенно в области переводоведения, поэтому представлялось неразумным не включать в корпус второй перевод только потому, что он попал в малую группу.

В целом, «ПарРус» является мониторным корпусом, то есть у него нет заранее жестко заданной структуры, он пополняется постоянно и исследования проводятся параллельно с работами по составлению корпуса.

2.2.3. Описание корпуса

В настоящее время параллельный русско-финский корпус художественных текстов «ПарРус» состоит из 128 произведений русской художественной литературы различных объемов, жанров, написанных в различные периоды разными авторами, и 137 текстов на финском языке, являющихся переводами последних (для некоторых русских текстов имеется более одного

финского перевода). Объем русского субкорпуса составляет 2 270 263 словоупотребления, объем финского субкорпуса — 2 241 020 словоупотреблений (полный список текстов корпуса см. в Приложении 2).

При создании корпусов текстов необходимо решать сложный и деликатный вопрос интеллектуальной собственности. Актуален этот вопрос и для «ПарРус». Часть текстов русского субкорпуса была создана в XIX — начале XX вв. и таким образом не подпадает под законодательство об авторских правах. Однако значительная часть текстов русского субкорпуса и все тексты финского субкорпуса вышли в свет во второй половине XX века. Поэтому для полномасштабного использования «ПарРус» необходимо получение разрешения держателей копирайта русских текстов и финских переводов. Работа по получению таких разрешений начата, в настоящее же время корпус используется исключительно в рамках исследовательской работы кафедры русского языка Отделения переводоведения Тамперского университета.

При составлении корпуса мы в целом руководствовались принципами, сформулированными в предшествующих разделах. Корпус является полнотекстовым, основным критерием отбора текстов является классификация русско-финских художественных переводов, описанная выше. Представить в корпусе все выделенные в классификации группы не удалось: даже если бы мы пытались представить 64 группы с наибольшей текстовой массой, ориентировочно потребовалось бы включить в состав массива 32 романа, 7 исторических романов и 42 повести, не считая рассказов. Если принять средний объем романа за 100 000 словоупотреблений, а повести — за 40 000, то объем корпуса составит около 6 X 2 миллионов словоупотреблений. Подготовка массива такого объема малыми силами — задача малореалистичная. Поэтому при составлении корпуса мы пытались по возможности отразить наиболее значительные группы. Русские тексты брались из текстовых архивов «Рунета», большая часть была получена из электронной библиотеки Максима Мошкова (http://www.lib.ru), некоторые — из публичной электронной библиотеки Евгения Пескина (http://public-library.narod.ru/), архива «Общий текст» (http://text.net.ru/index.html) и др. Отдельные небольшие по объему тексты, отсутствовавшие в электронных библиотеках, были отсканированы самостоятельно. Финские тексты сканировались силами студентов и практикантов кафедры русского языка Отделения переводоведения. Корпус текстов поддерживается с помощью пакета программ «КОКОС-П», описанного в следующей главе.

В корпусе представлены произведения 31 автора. Поскольку корпус полнотекстовый, доли разных авторов различаются довольно сильно. Доля различных жанров, представленных в корпусе, также неодинакова. Романы составляют 27,77% от объема корпуса, повести — 55,78% (то есть более половины), рассказы — всего 16,45%, несмотря на то, что из 128 текстов корпуса 90 — рассказы. Наиболее значительна в «ПарРус» доля произведений следующих авторов: Л.Н. Толстого, Ф.М. Достоевского, В. Дудин-

цева, И. Ильфа и Е. Петрова, М. Булгакова, Б. Пастернака и А.П. Чехова (см. табл. 3). В целом это соответствует рейтингу писателей по библиографии переводов. Полнотекстовость корпуса делает неизбежными и некоторые перекосы. Например, третье место В. Дудинцева представляется слишком высоким, однако это оказалось неизбежным (во всяком случае — на данном этапе работы с корпусом), из-за большого объема романа «Белые одежды», включенного в корпус.

Вообще, включение в полнотекстовый корпус большого романа вызывает резкое увеличение объема текстовой массы по данному автору и периоду, в результате чего возникает потребность включать в корпус большие романы других авторов, писавших в другой период. Например, романы Л.Н. Толстого и Ф.М. Достоевского можно «уравновесить» «Тихим Доном» М.А. Шолохова и романами Ф. Абрамова. Таким образом, настоящего баланса периодов и авторов можно достичь, только включив в корпус все значительные по объему произведения, которые переводились на финский язык, а это — очень большой объем работы.

Таблица 3. Представленность разных авторов в «ПарРус»

Автор	**Количество словоупотреблений**	**Доля (%)**
Толстой Л.Н.	295 190	13,00%
Достоевский Ф.М.	204 505	9,01%
Дудинцев В.	190 655	8,40%
Ильф И., Петров Е.	166 140	7,32%
Булгаков М.А.	153 778	6,77%
Пастернак Б.Л.	148 351	6,53%
Чехов А.П.	109 746	4,83%
Семенов Ю.	77 860	3,43%
Приставкин А.	66 351	2,92%
Трифонов Ю.	65 378	2,88%
Распутин В.	65 132	2,87%
Шукшин В.М.	61 040	2,69%
Пушкин А.С.	57 355	2,53%
Стругацкие А. и Б.	56 964	2,51%
Троепольский Г.	52 040	2,29%
Бакланов Г.	48 298	2,13%
Тургенев И.С.	47 133	2,08%
Аксенов В.	45 905	2,02%
Белов В.	44 904	1,98%
Фадеев А.	44 406	1,96%
Гроссман В.	42 607	1,88%
Лермонтов М.Ю.	41 499	1,83%
Лесков Н.	41 379	1,82%
Ерофеев В.	33 562	1,48%
Солженицын А.И.	32 390	1,43%
Олеша Ю.	29 507	1,30%
Горький М.	20 232	0,89%

Автор	Количество словоупотреблений	Доля (%)
Толстая Т.	13 284	0,59%
Гоголь Н.В.	10 013	0,44%
Бабель И.	2 731	0,12%
Зощенко М.	1 928	0,08%

Бóльшую часть корпуса составляют произведения советской литературы 1950-х — середины 1980-х гг. (46,02%), наименьшая доля — у произведений второй трети XIX века (3,2%), что связано с тем, что этот период представлен в настоящее время лишь тремя произведениями: романом И.С. Тургенева «Дворянское гнездо» и повестями Л.Н. Толстого «Метель» и «Два гусара». Значительность доли литературы середины XIX века связана с тем, что из произведений этого периода в корпус были, в частности, включены роман Л.Н. Толстого «Анна Каренина» — самое большое по объему произведение из вошедших в «ПарРус» (270 тыс. словоупотреблений) и роман Ф.М. Достоевского «Преступление и наказание» (170 тыс. словоупотреблений).

Как видно из таблицы 4, соотношение произведений различных периодов в корпусе и в отображаемом массиве текстов различно, в некоторых случаях различия оказываются довольно существенными (например, по 3-му периоду). Однако в наситоящий момент представляется более важным отразить в корпусе значимые произведения, чем пропорционально представить все периоды.

Таблица 4. Распределение текстов «ПарРус» по времени написания оригинала

Период	Количество словоупотреблений	Доля (%)	Доля (%) по библиографии
1	108 867	4,80%	5,7%
2	72 634	3,20%	7,9%
3	515 573	22,71%	9,2%
4	129 978	5,73%	9,4%
5	156 239	6,88%	8,7%
6	242 251	10,67%	18,0%
7	1 044 721	46,02%	34,9%

В корпус вошли переводы на финский язык, выполненные двадцатью разными переводчиками (см. табл. 5). Наиболее значительна доля переводов Э. Адриана, У.-Л. Хейно, Ю. Конкка и Л. Пююккё. В целом это соответствует данным, полученным из библиографии, однако в библиографии доли этих переводчиков различаются намного больше, чем в нашем корпусе, где они оказались представлены почти одинаковыми объемами текста. Это опять же связано с тем, что корпус является полнотекстовым. Среди включенных в «ПарРус» переводов Хейно — два больших романа («Мастер и Маргарита» Булгакова и «Белые одежды» Дудинцева). В кор-

пусе только два перевода Леа Пююккё — «Очарованный странник» Лескова и «Анна Каренина» Толстого, однако огромный объем романа Толстого выводит эту переводчицу на четвертое место. Хотя место Пююккё в корпусе соответствует ее месту в библиографии, доля ее переводов в корпусе значительно больше (ср. рис. 8).

Таблица 5. Доля разных переводчиков в «ПарРус»

Переводчик	**Количество словоупотреблений**	**Доля (%)**
Adrian E.	485 048	21,64%
Konkka J.	466 452	20,81%
Heino U.-L.	456 520	20,37%
Pyykkö L.	293 157	13,08%
Aarto A.	81 767	3,65%
Pienimäki N.	74 019	3,30%
Koskinen M.	55 242	2,47%
Hollo J.A.	53 472	2,39%
Iranto L.	50 752	2,26%
Orlov V.	49 154	2,19%
Laaksonen H.	42 106	1,88%
Прочие:	133 331	5,95%

В корпусе представлены следующие группы, выделенные по нашей классификации:

Таблица 6. Доля разных групп текстов в «ПарРус»

Группа	**Количество словоупотреблений**	**Доля (%)**
Group01	63 947	2,85%
Group02	181 174	8,08%
Group03	255 801	11,41%
Group04	138 450	6,18%
Group05	142 012	6,34%
Group06	70 224	3,13%
Group08	53 472	2,39%
Group09	96 979	4,33%
Group10	36 865	1,65%
Group12	163 430	7,29%
Group14	62 322	2,78%
Group15	14 598	0,65%
Group18	81 767	3,65%
Group19	171 366	7,65%
Group27	67 608	3,02%
Group28	18 314	0,82%
Group29	105 250	4,70%
Group30	46 601	2,08%
Group40	16 027	0,72%

Группа	Количество словоупотреблений	Доля (%)
Group43	23 882	1,07%
Group44	53 182	2,37%
Group56	20 543	0,92%
Group60	56 059	2,50%
Прочие:	301 147	13,44%

Таким образом, в корпусе «ПарРус» представлены тексты различных групп, относящиеся к разным периодам, написаные разными авторами и переведенные разными переводчиками. Однако изучение статистики по корпусу показывает, что разные типы текстов до конца не сбалансированы. Сбалансированность различных компонентов корпуса можно улучшить, используя продуманные стратегии отбора новых текстов. Но полный баланс в полнотекстовом корпусе недостижим. Тем не менее, представляется, что в своем нынешнем виде корпус «ПарРус» можно использовать для получения эмпирических данных и исследования общих тенденций, которые наблюдаются при переводе художественных произведений.

2.2.4. Составление параллельного корпуса: проблема исходного текста

Проблема критериев отбора текстов и представительности корпуса не являются единственными трудностями, возникающими при составлении параллельного корпуса художественных текстов.

Очень серьезной проблемой является поиск исходного текста. В идеале в корпус должно быть включено именно то издание, с которого выполнялся перевод. Однако найти это издание иногда бывает непросто. Довольно часто переводы выполнялись с журнальных изданий. Перевод романа И. Ильфа и Е. Петрова «Двенадцать стульев» был, по всей видимости, сделан с первого, довоенного издания романа, которое текстуально очень сильно отличается от издания 1957 года (хотя проверить это довольно сложно, поскольку издание это уже стало библиографической редкостью). Перевод романа Б. Пастернака «Доктор Живаго» был сделан с зарубежного издания 1957 года (поскольку других изданий романа в то время, когда был издан финский перевод, в 1958 году, не существовало), чем объясняются расхождения с российским изданием 1988 года. Даже в переводах произведений литературы XIX века обнаруживаются разночтения с академическими изданиями, показывающие, что в качестве исходного текста выступали другие, по-видимому, дореволюционные, издания. Приведем только один пример. В переводе на финский язык повести А.С. Пушкина «Капитанская дочка» (перевод Й.А. Холло (J.A. Hollo), 1962) обнаружилось следующее разночтение с оригинальным текстом, воспроизводящимся по

полному собранию сочинений А.С. Пушкина в 10-ти томах (текстуальные расхождения подчеркнуты):

(1)

<...> Мы проходили через селения, разоренные бунтовщиками, и поневоле отбирали у бедных жителей то, что успели они спасти. Правление было повсюду прекращено: помещики укрывались по лесам. Шайки разбойников злодействовали повсюду; начальники отдельных отрядов самовластно наказывали и миловали; состояние всего обширного края, где свирепствовал пожар, было ужасно... Не приведи бог видеть русский бунт, бессмысленный и беспощадный! Пугачев бежал, преследуемый Иваном Ивановичем Михельсоном. Вскоре узнали мы о совершенном его разбитии.	<...> Me marssimme kapinallisten perinpohjin hävittämien asuttujen paikkojen kautta, ja meidän oli pakko ottaa asukasraukoilta sekin, mitä he olivat pelastaneet rosvojen kynsistä. Missään ei ollut enää järjestystä; tilanomistajat piileksivät metsissä. Rosvojoukot tekivät tuhoa kaikkialla. Eri joukko-osastojen komentajat, joiden olisi pitänyt ajaa takaa jo Astrakaniin päin pakenevaa Pugatšovia, rankaisivat ja armahtivat omavaltaisesti syyllisiä ja syyttömiä. Laaja alue, jolla sodan palo raivosi, oli kauheassa tilassa... Varjelkoon Jumala meitä Venäjällä mielettömästä ja säälimättömästä kapinasta! Ne, jotka suunnittelevat meillä mahdottomia vallankumouksia, ovat joko nuoria tai eivät tunne kansaamme, tai kovasydämisiä ihmisiä, joille oma henki on kopeekan arvoinen ja vieras vielä halvempi. Pugatšov pakeni Ivan Ivanovitš Michelsonin takaa ajamana. Saimme pian kuulla että hän oli joutunut perin pohjin häviölle.

В академическом издании Пушкина цитируемый фрагмент повторяется дважды: в основном тексте и в тексте «Пропущенной главы», которая не была включена Пушкиным в окончательный текст и сохранилась в виде чернового автографа, в академическом издании эта глава публикуется в виде приложения. В отрывке о русском бунте в «Пропущенной главе» имеются некоторые текстуальные расхождения с основным текстом повести (подчеркнуты):

(2)

> <...> Мы проходили через селения, разоренные Пугачевым, и поневоле отбирали у бедных жителей то, что оставлено было им разбойниками.
>
> Они не знали, кому повиноваться. Правление было всюду прекращено. Помещики укрывались по лесам. - Шайки разбойников злодействовали повсюду. Начальники отдельных отрядов, посланных в погоню за Пугачевым, тогда уже бегущим к Астрахани, самовластно наказывали виноватых и безвинных. — Состояние всего края, где свирепствовал пожар, было ужасно. Не приведи бог видеть русский бунт — бес<с>мысленный и беспощадный. Те, которые замышляют у нас невозможные перевороты, или молоды и не знают нашего народа, или уж люди жестокосердые, коим чужая головушка полушка, да и своя шейка копейка.
>
> Пугачев бежал преследуемый Ив. Ив. Михельсоном. Вскоре узнали мы о совершенном его разбитии.

Таким образом, в переводе на финский язык в основном тексте воспроизводится именно отрывок из «Пропущенной главы», расхождений с которым в финском переводе почти нет, в переводе же самой «Пропущенной главы» этот фрагмент отсутствует.

Еще одно расхождение между академическим изданием «Капитанской дочки» Пушкина и переводом Холло — имена персонажей повести в

«Пропущенной главе», которые приведены в соответствие с основным текстом (в пушкинской рукописи Гринев — Буланин, а Зурин — Гринев).

Все издания Пушкина советских времен выполнены по академическому собранию сочинений, поэтому разночтения практически отсутствуют. Таким образом, либо финский перевод выполнялся по какому-либо дореволюционному изданию повести, в котором имеются расхождения с современными изданиями, либо переводчик осознанно отклонился от текста оригинала, чтобы перевод был более понятен и доступен адресатам. Первая версия представляется более правдоподобной.

Нам удалось найти два дореволюционных издания «Капитанской дочки», 1899 и 1901 года. В издании 1989 года (Издание А.И. Мамонтова) в основном тексте отрывок (1) полностью соответствует современным изданиям Пушкина, но «Пропущенная глава» отсутствует. Более интересным с этой точки зрения оказалось издание типографии Товарищества «Общественная польза» (С.-Петербург, 1901). В этом издании текст «Пропущенной главы» вставлен в соответствующее место текста, имена персонажей приведены в соответствие с основным текстом. Отрывок о русском бунте текстуально соответствует фрагменту (2).

Таким образом, хотя издания «Капитанской дочки», полностью соответствующего тексту перевода, найти не удалось, вполне можно предположить, что такое издание существовало и что с этого издания был выполнен первый финский перевод повести (пер. С. Суомалайнена, 1876), а в последующих переводах использовался этот перевод, сделанный со старого издания, причем по каким-то причинам перевод не был приведен в соответствие с более поздними изданиями Пушкина (которые, в данном случае, лучше соотвествуют замыслу писателя, не включившего главу в окончательную редакцию повести).

Поскольку в «ПарРус» специальной работы по поиску именно тех изданий, с которых выполнялись переводы, не выполнялось, необходимо особо оговорить, что далеко не все случаи пропусков в переводе, текстуальных несовпадений или неточных эквивалентов являются результатом работы переводчика.

Глава 3. Программное обеспечение параллельного корпуса текстов

3.1. О существующем программном обеспечении для корпусов текстов

3.1.1. Вопрос об общепризнанной «корпусной» утилите

Поскольку обработка больших массивов текстов — задача специфическая и лежащая несколько в стороне от «массовой» компьютеризации, ни один из крупных производителей программного обеспечения этой проблемой всерьез не занимается. Стандартные утилиты DOS и UNIX, обеспечивающие поиск некой последовательности символов в тексте, разумеется, не удовлетворяют всем нуждам исследователей, работающих с корпусами текстов. Это стало причиной появления в начале 1980-х гг. специализированных пакетов программ для обработки корпусов текстов (более подробные обзоры см. в Oakes 1998, Burnard 1992).

COCOA (Count and Concordance on Atlas) позволяет получать конкордансы, как для отдельных слов, так и для словосочетаний, а также частотные словники. Кроме обычной сортировки программа выполняет и реверсивную сортировку (по концам слов). *OCP* (Oxford Concordance Program) строит словники, сортирует их по алфавиту и по частотам. С помощью программы можно получать конкордансы, которые сортируются по правому или левому контексту. Поиск может выполняться и по целым словам, и по шаблонам для поиска с использованием знаков «*» (пропущена одна или несколько букв) и «@» (пропущена буква). Предусмотрено получение элементарной статистики по текстам (количество знаков, количество слов и т.п.) (Oakes 1998).

Американская компания Electronic Text Corporation совместно с Brigham Young University разработала программы *ETC* и *WordCruncher* (WordCruncher 1989), которые поставлялись вместе с массивами текстов. Пакет состоит из двух программ — *IndexETC* и *ViewETC*. С помощью первой программы тексты индексируются, с помощью второй можно просматривать контексты для слов и словосочетаний и строить конкордансы.

Главные проблемы, возникающие при создании программ обработки текстов — это обеспечение работы с разными алфавитами и поддержка ра-

боты с конкретными языками (например, обеспечение лемматизации). Первая проблема решается относительно легко; все вышеупомянутые программы поддерживают работу с разными алфавитами, в них задается алфавитный порядок для разных языков, а в программе *WordCruncher* можно даже задать свой порядок сортировки. (Подробнее см. Oakes 1998: 156–158). Вторая же проблема оказалась значительно более серьезной: даже современные программы, как правило, поддерживают только один язык, а декларируемая поддержка нескольких языков (например, в *Word Smith Tools*) на деле означает опять же поддержку алфавита, порядка сортировки и т.п. Даже стандартные способы оформления текста, принятые в данном языке, как правило, остаются без внимания, и для того, чтобы программа работала корректно, следует переоформить текст по неким универсальным стандартам (например, использовать в качестве разделителя абзаца именно пустую строку, а не отступ).

Многие из старых программ для обработки текстов плохо подходили для работы с кириллицей, а также не учитывали многие из потребностей исследователей. Поэтому в 1980–90-х гг. в России разрабатывалось свое программное обеспечение для обслуживания корпусов текстов. В отделе Машинного фонда Института русского языка РАН был разработан пакет программ «УНИЛЕКС» (Аношкина 1992), который, помимо стандартных операций, предусмотренных для программ подобного рода, выполняет лемматизацию русских текстов. В отделе экспериментальной лексикографии того же института создана программа «ДИАЛЕКС» (Исаев 1996). Существует большое количество экспериментальных разработок. В свое время, каждая организация, занимавшаяся анализом или обработкой текстов (для выполнения лингвистических исследований, производства словарей, конкордансов, контент-анализа и т.п.), в конечном итоге, создавала свою программу обслуживания корпусов текстов (Михайлов 1998).

В 1990-е годы бурное развитие вычислительной техники и программного обеспечения создало условия для создания более сложных пакетов программ для обработки текстов. Появился целый ряд программ, с помощью которых можно быстро обрабатывать очень большие по объему массивы текстов и получать ранее малодоступные статистические данные (Lager 1995: 8–11, Oakes 1998).

В начале 1990-х для обслуживания архивов текстов ICAME (International Computer Archive of Modern/Mediaeval English) был создан пакет программ *Lexa*. Программа позволяет не только получать конкордансы и статистику по текстам, она выполняет грамматическую разметку текстов (tagging), лемматизацию (Oakes 1998: 194–195). К сожалению, эти функции программы работают только с английским языком.

В 1996 году Майкл Скотт создал пакет программ *WordSmith*, с помощью которого можно получать частотные словники для отдельных текстов или групп текстов, где кроме абсолютной частоты указывается также и относительная частота. По каждому тексту можно получить следующую ста-

тистику: количество словоупотреблений, количество словоформ, отношение количества словоформ к количеству словоупотреблений (type-token ratio), средняя длина слова в знаках, средняя длина предложения (количество слов), количество предложений. Утилита *Concord* позволяет получать конкордансы на заданные слова и выражения и обнаруживает часто повторяющиеся обороты с использованием этих слов. Утилита порождает списки коллокаций, а также графически показывает распределение слова по разным текстам массива. Утилита *Keywords* ищет в текстах «ключевые» слова, т.е. слова, частота которых в данном тексте существенно отличается от их частоты во всем массиве (Oakes 1998: 193–194). В качестве альтернативной программы выступает пакет программ *MonoConc*, который в целом позволяет выполнять те же операции (подробнее см. http://devoted.to/corpora).

В целом, компьютерные программы, обслуживающие корпус текстов, в той или иной степени решают две основные проблемы — получение словников и построение конкордансов. Степень автоматизации этих процедур и гибкость работы пакета (то есть степень трудоемкости подготовки массива текстов к обработке, возможность импорта результатов анализа в разные форматы, а также эргономичность программы) и определяют его популярность. К сожалению, существующие программные продукты пока весьма громоздки и неудобны в работе, вследствие чего ни одна не завоевала по-настоящему мирового признания. Последнее время все более популярной становится программа Майкла Скотта *WordSmith Tools*, однако и ее нельзя пока назвать общепризнанной, поскольку она работает только с ANSI-файлами, а для работы с текстами, размеченными по стандартам XML или SGML, не приспособлена.

Главная проблема программного обеспечения для корпусов текстов состоит в том, что универсальными могут быть только операции обработки строк (разбивка на слова, предложения, абзацы и т.п.). Такие операции, как лемматизация и грамматическая разметка, оказываются тесно привязанными к языку, вследствие чего действительно мощные пакеты программ должны разрабатываться под конкретный язык.

3.1.2. Разметка текстов

Практически ни одна из существующих программ обработки текстов не может работать с текстовыми файлами, если они не были заранее специальным образом подготовлены. Файлы, как минимум, должны быть сохранены в формате ANSI. Старые MS-DOS-программы работали с текстовыми файлами, разбитыми на строки. В конце каждой строки ставилось два знака — конца строки и перехода на новую строку. Такие же знаки ставились и в конце абзаца. Таким образом, возникала омонимия и программа не могла отличать строки от абзацев. Современные программы, работающие в среде

Windows, как правило, требуют файлы в формате ANSI с длинными строками. Если же на вход подаются файлы ANSI с «короткими» строками, программа будет интерпретировать каждую строку как отдельный абзац.

Как правило, для реализации всех возможностей программы требуется специальная предварительная структурная «разметка» текстов, при которой в файле специальными знаками должны быть отмечены интересующие исследователя структурные единицы, например, главы, страницы, абзацы, а иногда — даже предложения. Вообще говоря, это неизбежно: если исследователю нужно получать «документированные» примеры употребления со ссылками на главы и страницы, текстовые файлы должны быть разбиты на разделы.

Теоретически такая разбивка вполне может быть автоматизирована: достаточно сообщить программе, что является маркером новой главы (слово «глава», цифра с новой строки, «три звездочки») или страницы (знак жесткого переноса страницы, пустая строка и т.п.). Однако в реальности все обстоит по-другому: разметка выполняется исследователем, который сам пишет для этого программу, если он умеет программировать, либо вынужден делать разметку вручную[13].

Маркеры, применяющиеся для разметки, как правило, малоинформативны, это один или несколько символов, обозначающие уровень разметки, и порядковый номер, например “|P1, |P3, |L100” в WordCruncher. Этого достаточно для ссылки на страницы или главы в большом романе, однако, если в одном файле хранится сборник рассказов, это гораздо менее удобно. Исследователь должен помнить, например, что в сборнике рассказов Чехова текст №30 — это “Лошадиная фамилия”, а №45 — “Мальчики”. Следует также отметить, что большинство программ обработки текста накладывают определенные ограничения на количество разделов в файле, а также на длину раздела (кстати, ограничение на размер модуля в УНИЛЕКСе — 60 тыс. символов — сильно снижает гибкость этой программы). Многие из новых программ (например, WordSmith) никакой разметки не требуют, что повышает дружественность интерфейса, но одновременно делают программу малопривлекательной для ряда пользователей, например, в том случае, если при выполнении работы требуется указание номера страницы цитируемого сочинения.

Еще одно неудобство известных нам программ, работающих с корпусами текстов — все они работают только со стандартными ANSI-файлами. В тех случаях, когда исследователя интересует шрифтовая разметка оригинала, а также используемое в изданиях графическое

[13] Справедливости ради следует заметить, что в пакете УНИЛЕКС разметка строк и абзацев происходит автоматически (правда абзацы должны быть разделены пустыми строками, что для русскоязычного текста нехарактерно), однако разбивка на “модули” — единицы макроуровня — программой не выполняется.

оформление, этот формат делает исследование в лучшем случае трудно выполнимым.

По этим причинам за последнее десятилетие все большее внимание уделяется разработке универсальной разметке текстов (markup), которая позволяла бы, используя стандартную ANSI-кодировку, сохранять по возможности все структурные особенности текста, а также эксплицировать различные лингвистически релевантные признаки (например, частеречную принадлежность слова, его синтаксическую функцию, семантику и т.п.). Для этого в настоящее время используется SGML (Standard Generalised Markup Language). Этот язык представляет собой систему конвенций, позволяющую с помощью специальных меток вида <X> ... </X> обозначать различные структурные и понятийные элементы текста (предложения, абзацы, главы, а также авторские ремарки, сноски, шрифтовые выделения и т.п.), а с помощью знаков вида &X записывать информацию, которая не может быть передана стандартной латиницей (например, запись *&auml* обозначает букву *ä*). Кроме SGML нередко используется его упрощенный вариант — XML (Extended Markup Language).

С 1987 года идет разработка TEI (Text Encoding Initiative) — система разметки текстов для выполнения лингвистических исследований. TEI является вариантом SGML, предназначенным для обозначения лингвистически релевантных характеристики текста и его элементов. В разработке TEI участвуют три крупнейших международных ассоциации, связанные с использованием вычислительной техники в гуманитарных науках: Association for Computational Linguistics (ACL), Association for Literary and Linguistic Computing (ALLC) и Association for Computers and the Humanities (ACH). В TEI разработана система конвенций для описания текста (header), грамматические пометы, система записи букв разных алфавитов, уточняется система разметки структурных единиц текста. Наряду с TEI используется и его «облегченный» вариант, включающий только наиболее часто употребляющиеся конвенции — TEI-LITE. В настоящее время разметка по TEI уже применяется в ряде крупных проектов по корпусам текстов, например в BNC (British National Corpus) (McEnery & Wilson 2001: 36–37). (Подробнее о разметке текстов см. Johansson 1994, McEnery & Wilson 2001: 32–69, Hockey 2000: 24–48).

Текст, размеченный по TEI, особенно если применялась детальная разметка, человек может читать лишь с большим трудом. Поэтому для работы с такими текстами требуется специальное программное обеспечение, которое «понимало» бы разметку, и, кроме того, «переводило» бы TEI-тексты в привычный для человека формат, «пряча» метки (tags) и визуализируя различные типографские элементы (абзацы, изменения шрифта, жирный шрифт, курсив, сноски и т.п.). Однако в настоящее время стандартное программное обеспечение существует только для работы с другим подмножеством SGML — с HTML (HyperText Markup Language), применяющимся главным образом для создания web-страниц в Интернете. Постепенно

становится стандартом и XML, который, возможно, в обозримом будущем будет использоваться в Интернете наряду с HTML (Hockey 2000: 46). Однако вопрос о стандартных программных пакетах для обработки текстов, размеченных по SGML / XML / TEI, остается открытым. Насколько нам известно, такие инструменты разрабатываются самими исследовательскими группами, в каждом случае под конкретный корпус текстов и под конкретные исследовательские задачи. Приведем в качестве примера «Британский национальный корпус» (British National Corpus, BNC), для которого специально разрабатывалось свое программное обеспечение — *CLAWS* и *SARA* (см. http://thetis.bl.uk/).

В то же время идею разметки текстов нельзя считать общепризнанной. Например, Т. Лагер в своей диссертации подвергает эту идею жесткой критике, называя разметку «нотацией без формализации, стряпней для безмозглых программ рутинной обработки текста» (Lager 1995: 2). В этой критике есть доля правды, поскольку на сегодняшний день единого стандарта разметки действительно не существует, TEI, несмотря на огромный объем описания этого стандарта и сложность его применения, все равно не учитывает всех потребностей, которые могут возникнуть при выполнении исследований[14], что может привести впоследствии к возникновению «диалектов». Еще одна проблема — возможность наложения друг на друга разметок разных уровней, что также может усложнять обработку текстов.

3.1.3. «Индексирование» текстов

После того, как тексты подготовлены к обработке, для большинства существующих программ необходима еще одна операция, требующая времени, — **индексирование**. Программа должна последовательно обработать весь текст и составить по нему индекс — список адресов всех словоформ файла. Это делается для того, чтобы избежать последовательной обработки на этапах выполнения пользовательских запросов. На этом же этапе программа создает словник текста. Так, например, работает *Word Cruncher*.

Альтернативой к этому подходу является подход, использованный в программе *Dialex*, в которой индексирование не предусмотрено, зато при выполнении запросов всякий раз происходит последовательная обработка текста. В этом главный недостаток данного подхода. Поиск в индексированном тексте происходит во много раз быстрее, однако индекс может занимать очень много места на диске.

Таким образом, программа, работающая с неиндексированными текстами, хорошо справляется с небольшими запросами на текстах небольшого

[14]А учесть все потребности не представляется возможным: лингвистическая разметка на несколько порядков сложнее, чем, скажем, типографская разметка. Кроме того, она должна быть применима к разным языкам.

объема и становится плохо управляемой при работе с текстами большого объема. «Большие многопользовательские системы обработки текстов, работающие в режиме прямого доступа, работают эффективно лишь в том случае, если они работают с предварительно построенными индексами» (Hockey 200: 65).

Недостаток индексирования заключается в том, что его требуется выполнять даже в том случае, если исследователю нужно сделать лишь два-три запроса к текстам.

3.1.4. Словники

Первый продукт, ожидаемый от программы обработки текста, — **словник**. Большинство словников, получаемых такими программами, — словники по словоформам. Следует также отметить, что понятие «словоформа» понимается здесь несколько упрощенно, как «графическое слово», то есть последовательность знаков, ограниченная определенными символами (.!?,:-« и др.). Некоторые слова состоят из нескольких графических слов, например *на ощупь, в результате, в течение* и т.п., и эти слова при составлении словника разбиваются на части. К счастью, большая часть словоформ совпадает с графическими словами.

Словник по словоформам есть лишь промежуточный продукт, для получения словника по лексемам требуется операция **лемматизации** — приведения словоформ к словарной форме (например, *бежал* → *бежать*, *его* → *он*). Любой лексикограф знает, насколько трудоемка эта операция даже при работе со сравнительно небольшими массивами текстов. Словник корпуса русских художественных текстов «ТамРус», объем которого составляет около 10 млн. словоупотреблений, превышает 400 тыс. единиц. Поэтому лемматизатор оказывается для лексикографа чрезвычайно важной программой.

Нам известно о целом ряде программ, выполняющих лемматизацию русскоязычных текстов в автоматическом режиме.

Так, лемматизацию русских текстов выполняет упомянутый выше пакет программ «УНИЛЕКС». Первые варианты программы основывались на правилах с использованиями пятибуквенных финалей, например «-овском → -овский». Позднее от этой идеи отказались и стали использовать списки словоформ с заданными леммами (см., например, Мошкович 1989, 1990).

Другие принципы используются в лемматизаторе, разработанном Г.О. Сидоровым (Сидоров 1995, Сидоров 1996). На этапе индексации программа делает «глобальный» морфоанализ с использованием словаря основ, полученного из «Грамматического словаря» А.А.Зализняка (Зализняк 1980). Результаты анализа сохраняются в формате СУБД *Paradox*. В случае получения нескольких лемм для одной словоформы (например, *банку* → *банка*, *банк*), фиксируются все варианты, при этом в специальном поле

фиксируется наличие омонимии. Словоформы, лемма для которых не была найдена, записываются в отдельную базу данных. После завершения лемматизации исследователь может вручную снять омонимию, просматривая контексты употребления словоформ с неоднозначными леммами, а также указать леммы для словоформ, для которых те не были найдены (в основном — имена собственные, неологизмы и просторечные формы). Недостатком программы является то, что она не позволяет «разводить» омонимичные словоформы, имеющие разные леммы (например, если в тексте имеется десять словоформ *банка*, пять из которых — родительный падеж от *банк*, а пять — именительный от *банка*).

Кроме обычного словника исследователю нередко требуется **частотный** словник, т.е. словник с указанием абсолютной и/или относительной частоты употребления единицы. Все известные нам программы выполняют этот вид обработки текста.

На определенном этапе развития программ этого класса большое внимание уделялось **сортировкам**. Например, программа «Диалекс» делает как «стандартную» (по начальным буквам), так и «обратную» (по конечным буквам слова, например, *а — баба — арба — ага — мама — папа*) сортировку, причем и восходящую, и нисходящую. Кроме того, выполняется частотная сортировка словника. В настоящее время распространение баз данных делает проблему сортировок менее актуальной: современные СУБД легко выполняют все указанные виды сортировок, кроме, пожалуй, только «обратной» сортировки, для выполнения которой средствами базы данных требуются некоторые дополнительные действия (например, создание дополнительного поля, в котором хранятся «перевернутые» записи основного входного поля, например, в основном поле — *книга*, в реверсивном — *агинк*).

3.1.5. Конкордансы

Еще один важный продукт работы программы — конкорданс. Это набор контекстов употребления данной единицы. На заре конкордансов длина контекста, как правило, исчислялась в знаках. Поэтому получаемые контексты были довольно короткими и нередко обрывались на середине предложения и даже слова. Так, «бумажный» конкорданс к «Преступлению и наказанию» Ф.М. Достоевского, выполненный японскими исследователями (см. Atsushi et al 1994, Шайкевич 1995), дает контексты длиной в 50 знаков. Современные программы обработки текстов позволяют задавать длину контекста как в символах, так и в строках, предложениях и абзацах.

Построение конкорданса — особенно в тех случаях, когда список слов и размеры текстового массива очень велики — процесс долгий. Скорость построения конкорданса зависит от способа обработки текста — последовательного или с обращением к предварительно построенному индексу.

Совершенно очевидно, что после получения конкорданса исследовательская работа только начинается. Нередко примеров оказывается слишком много. Конкордансы на высокочастотные слова могут занимать сотни страниц текста, причем бóльшая часть — однотипные и малоинтересные для исследователя примеры. В этой ситуации становится ясно, что работа с конкордансом в виде текстового файла — процесс крайне неэффективный. Конкорданс в формате базы данных гораздо предпочтительнее — пользователь имеет возможность хоть в какой-то степени автоматизировать работу с примерами: быстро искать примеры из требующихся источников, «фильтровать» примеры, использовать базу данных конкорданса для пополнения словарной базы данных. Попытка получения конкордансов в формате *Paradox* сделана в версии программы «Диавин» — версии «Диалекс» для *Windows*. Но, разумеется, идеалом была бы некая «смычка» программы обработки текста с СУБД, в которой были бы реализованы все необходимые рутинные процедуры.

3.1.6. Коллокации

Нередко не меньший интерес, чем конкордансы, для пользователя представляют коллокации, то есть информация о ближайших соседях лексемы и словоформы. Это дает информацию о сочетаемости слов, об ассоциациях, связанных с данным словом или группой слов. Списки коллокаций позволяют выяснять, в каких значениях употребляется слово в корпусе, и даже «разводить» омонимы и разные значения многозначных слов. Однако на первоначальном этапе построение конкордансов вызывало гораздо больший интерес, чем получение списков коллокаций. Большинство исследователей видели в корпусах текстов в первую очередь источник примеров употребления. Поэтому многие из первых программ обработки текстов вообще выполняли только две операции: построение словника и получение конкордансов. При необходимости исследователи получали списки коллокаций из конкордансов вручную. В современных пакетах программ набор операций многообразнее, многие программы позволяют получать и списки коллокаций. Особого упоминания заслуживает *WordSmith Tools*, в котором специально предусмотрен поиск наиболее частотных для заданной словоформы коллокаций.

3.1.7. Запросы

Следует сказать несколько слов и о пользовательских запросах. В большинстве случаев пользователю требуется получить информацию по определенному списку слов или словоформ. Большинство программ позволяет либо непосредственно ввести одно слово в окошко запроса, либо

дать имя файла со списком слов. Некоторые программы, например *WordCruncher*, позволяют выбирать слова непосредственно из словника. В большинстве программ кроме поиска на полное совпадение строк можно выполнять поиск на частичное совпадение с использованием шаблонов. Это позволяет искать разные словоформы одного и того же слова (например, *wom*n* позволит найти примеры на *woman* и *women*, по запросу *домашн** будут найдены все формы слова *домашний*).

Очень часто пользователю требуется искать не слова, а словосочетания, нередко — разрывные сочетания типа двойных союзов *не только ..., но и; чем ..., тем* и т.п. Кроме того, при проведении некоторых исследований требуется проверить возможность появления / непоявления тех или иных единиц в общем контексте. Такие функции имеются в программах *WordCruncher* и «Диалекс». В первой из них возможности такого поиска довольно ограничены: пользователь может исследовать совместную встречаемость только двух слов, причем эта функция реализована только в режиме диалога, то есть исследователь лишен возможности задать список всех интересующих его единиц и спокойно ждать конца работы программы. В «Диалексе» данный режим работы решен значительно более удачно: количество слов в сочетании не ограничено, пользователь делает запрос списком, правда должен сохранить список в виде отдельного текстового файла и задать его в качестве входного списка для конкорданса.

3.1.8. Работа с большим количеством текстов

Корпус текстов — это всегда большое количество файлов. Далеко не во всех случаях пользователю нужны данные из всех текстов корпуса. Вообще, хорошо организованный корпус текстов должен иметь сложную структуру и разбиваться на подмножества текстов. Далеко не всегда пользователя интересует весь корпус: у него могут возникать самые разные потребности. Поэтому программа, обслуживающая корпус текстов в идеале должна давать исследователю возможность получать данные как из всего корпуса, так и из любого его подмножества, которое может быть задано авторами корпуса или даже самим исследователем.

Эта задача, к сожалению, пока решается неудовлетворительно. Для того, чтобы обрабатывать группу файлов, пользователь должен подготовить список файлов для обработки. Для этого он должен: а) быть хорошо знакомым с вычислительной техникой и б) не только хорошо знать структуру корпуса, какие тексты в него входят, как они группируются и т.п., но и имена файлов, которые желательно если не знать наизусть, то хотя бы всегда иметь под рукой распечатку с их списком. Это сильно усложняет работу даже с относительно небольшим корпусом текстов.

Представляется, что идеальной была бы система, которая выдавала бы список заранее заданных подмножеств корпуса с краткой аннотацией, а

также список всех текстов корпуса со всей необходимой сопроводительной информацией и возможностью создавать свои собственные группы, не выходя из системы. Следует также отметить, что такие возможности не только вполне реальны при нынешнем развитии программного обеспечения, но и жизненно необходимы при работе с большими корпусами текстов.

3.1.9. Выводы

Подводя итоги, перечислим основные качества, существенные для компьютерной программы, обслуживающей корпус текстов:

- минимальная предобработка текстов корпуса, желательно в автоматическом режиме, наличие утилиты, помогающей объединять тексты в корпус;
- возможность работы с большим количеством текстов, гибкое выделение подмножеств корпуса;
- индексирование текстов;
- диалоговый режим работы;
- лемматизация;
- конкордансы с возможностью задания длины контекста в символах, строках, словах, предложениях, абзацах;
- результаты работы программы как в текстовом формате, так и в формате СУБД;
- простой и доступный интерфейс с возможностью выполнения всех операций (составление списка текстов для обработки, составление запроса и т.п.) и не требующий от пользователя специальных навыков работы с компьютером.

3.1.10. Программное обеспечение для параллельных корпусов текстов

Параллельные корпуса текстов пока не получили столь широкого распространения, как одноязычные корпуса текстов. Поэтому и программное обеспечение для ПКТ находится в стадии разработки.

Программы, предназначенные для работы с одноязычными корпусами текстов, можно использовать для выполнения операций, общих для одноязычных и многоязычных корпусов текстов. Например, составление словника или получение коллокаций выполняется лишь на одном из субкорпусов ПКТ и таким образом, не отличается принципиально от аналогичных операций на одноязычном корпусе текстов. С другой стороны, такие процедуры, как получение параллельных конкордансов, стыковка текстов и поиск лексических соответствий являются специфическими для ПКТ.

Программ для работы с ПКТ немного. *ParaConc* Майкла Барлоу (Barlow, 1995) предусматривает лишь получение параллельных конкордансов, причем до начала работы необходимо вручную отметить параллельные места в текстах. Больше возможностей дает пакет *MultiConcord*, работа над которым идет в Бирмингемском университете (подробнее см. http://devoted.to/corpora). Эта программа не только составляет параллельные конкордансы, но и позволяет в автоматическом режиме выполнять стыковку параллельных текстов на уровне предложений (правда, перед этим пользователь должен выполнить стыковку текстов на уровне абзацев). Программа поддерживает работу с десятью языками: английским, греческим, датским, испанским, итальянским, немецким, португальским, финским, французским и шведским.

Разработку программного обеспечения для ПКТ сильно осложняет тот факт, что оно еще сильнее, чем программное обеспечение для одноязычных корпусов текстов, ориентировано на конкретные языковые пары. В основе работы программ могут лежать некоторые универсальные принципы, но все же пакет должен модифицироваться для каждой новой пары языков.

3.2. Пакет программ «КОКОС-П»

Для обслуживания корпуса «ПарРус» была создана специальная программная оболочка «КОКОС-П» (Конкордансы, Коллокации, Словники из Параллельных текстов. Описание системы и всех режимов ее работы см. в Интернете на странице http://www.uta.fi/~lomimih/Tutkimus/cocos-p.htm). Данный пакет программ разрабатывался на основе системы «КОКОС», предназначенной для поддержки одноязычных массивов текстов и также созданной автором настоящей работы.

Корпус текстов хранится в виде двух баз данных *Microsoft Access*. В одной базе данных хранятся тексты корпуса и индексы (далее — база данных с текстами, БДТ), во второй — каталог текстов корпуса, словники, а также формы, отчеты и программные модули для работы с корпусом (далее — база данных с программами, БДП). Все программы написаны автором настоящей работы на языке *Microsoft Access Basic*. Часть кодов программ приводится в приложении 4.

Пакет программ представляет собой гибкую, быстро развивающуюся систему, к которой легко добавляются все новые модули и становятся доступными все новые режимы работы. Поскольку система организована в виде базы данных *Microsoft Access*, у пользователя есть возможность использовать также стандартные возможности этой СУБД.

3.2.1. Хранение данных в «КОКОС-П»

Как уже было сказано выше, весь ПКТ хранится в двух базах данных. БДП отвечает в основном за обработку данных и хранение результатов, в БДТ хранятся собственно тексты и индексы.

База данных с программами (БДП).

Ядром всего корпуса текстов являются каталоги текстов. В каталоге исходных текстов (КИТ) хранятся сведения об оригинальных текстах: автор, название текста, переводчики, дополнительная информация о тексте (в нашем варианте корпуса: жанр, время издания, рейтинг автора, — однако эта часть таблицы легко модифицируется и, в случае необходимости, могут быть добавлены новые поля), статистика по тексту: количество знаков, слов, предложений, абзацев. Аналогичную структуру имеет каталог переводных текстов (КПТ), только здесь указывается финское название произведения. Эти две таблицы связаны между собой (в КПТ есть специальное поле, в котором указывается идентификационный номер оригинального текста). У каждого оригинального и переводного текста есть свой уникальный идентификационный номер (ИН).

Вторая важнейшая часть БДП — словники. В настоящем варианте пакета программ таковых четыре: нелемматизированные словники по русским и финским текстам и лемматизированные словники. В нелемматизированном словнике указывается словоформа, ее абсолютная и относительная частота. У каждой словоформы есть ИН. В лемматизированном словнике для каждой лексемы также указывается абсолютная и относительная частота, каждая запись в таблице так же, как и в нелемматизированном словнике, имеет идентификационный номер. Лемматизированный и нелемматизированный словники связаны друг с другом через отсылку к ИН леммы в нелемматизированном словнике.

ID	Word	Count	RelCount	LinkToLemm
53843	балет	1	0,001	661
73186	балета	2	0,002	661
134650	балете	1	0,001	661
127745	балетным	1	0,001	36338
128778	балетных	1	0,001	36338
105767	балка	1	0,001	662

Рис. 10. *Фрагмент русского нелемматизированного словника*

База данных с текстами (БДТ)

Тексты хранятся в виде таблицы в БДТ. В каждой записи таблицы хранится один фрагмент исходного русского текста (часть абзаца, один абзац или несколько абзацев), в другом поле — соответствующий ему фрагмент финского перевода. В остальных полях записываются ИН ори-

гинального текста и перевода (из каталогов текстов в БДП). У каждой состыкованной пары фрагментов каждого из текстов имеется, таким образом, свой ИН.

ID	RussianText	FinnishText	Original	Translation
8628	Витька дал ей прикурить от своей папироски, а сам с интересом разглядывал лицо девушки — молодая, припухла, пальцы трясутся.	Vitka antoi tytölle tulen tupakastaan ja katseli samalla tytön kasvoja kiinnostuneena - nuori, pöhöttynyt, sormet vapisevat.	15	16
8629	— С похмелья? — прямо спросил Витька. — Ну, — тоже просто и прямо ответила девушка, с наслаждением затягиваясь «беломориной».	— Krapula vai? Vitka kysyi suoraan. — Joo, tyttö vastasi, mutkattomasti ja suoraan hänkin ja veteli nautinnollisesti «belomorkaansa».	15	16
8630	— А похмелиться не на что, — стал дальше развивать мысль Витька, довольный, что умеет понимать людей, когда им худо.	— Eikä löydy krapularyyppyä, Vitka kehitteli ajatusta eteenpäin, tyytyväisenä siitä, että hän pysyi ymmärtämään ihmisiä, kun näillä oli kurjaa.	15	16

Рис. 11. *Фрагмент таблицы с текстами (ТТ)*

Вторая часть БДТ — индекс. Это самая большая таблица системы, количество записей в ней может доходить до нескольких миллионов (в зависимости от размеров корпуса текстов). В индексе два поля: номер слова (идентификационный код слова из словника в БДП) и код предложения (по таблице текстов БДТ), в котором зафиксировано данное слово.

В «КОКОС-П» — два индекса: русский и финский.

WordNo	PhraseNo
1	34216
1	36558
1	40236
2	1
2	2
2	8

Рис. 12. *Фрагмент индекса*

Связи между различными компонентами системы наглядно показаны на рис. 13.

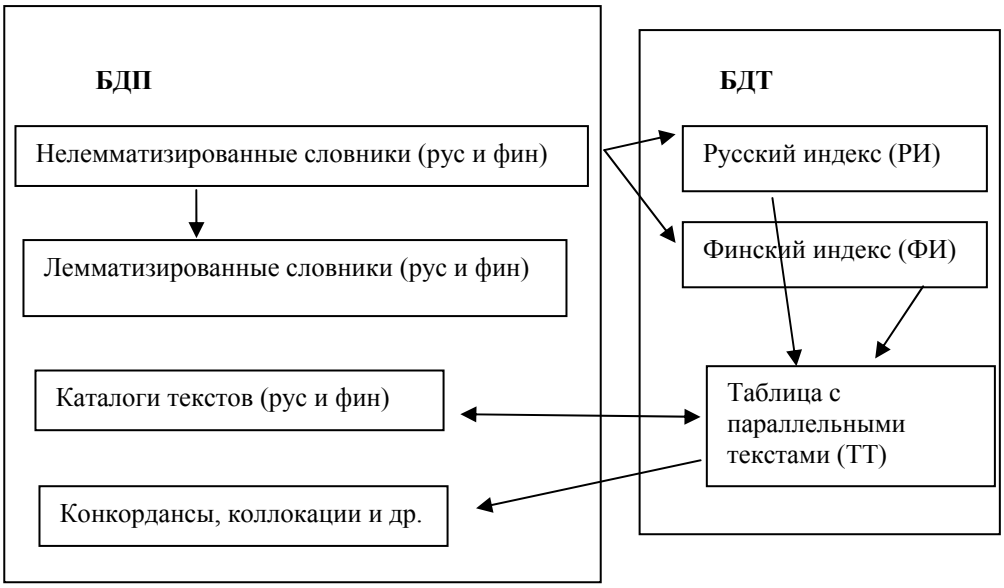

Рис. 13. *Связи между компонентами системы «КОКОС-П»*

3.2.2. Обработка запросов в «КОКОС-П»

В БДП организован интерфейс, с помощью которого пользователь может делать запросы по корпусу текстов. Программные модули по цепочке «словоформа → индекс → предложение текста → контекст» находят требуемые фрагменты текстов оригинала или перевода. Поскольку последние хранятся в соседних столбцах одной и той же таблицы, то можно легко получать, в частности, параллельные конкордансы. Последовательного поиска по текстам удается избежать благодаря использованию индексов. Это значительно ускоряет выполнение запросов (медленно обрабатываются только запросы на очень высокочастотные слова, поскольку многократная считка данных и запись результатов на диск занимает много времени). При проверке быстродействия программы конкорданс объемом в 1000 контекстов был получен за 13,6 секунды (использовался компьютер с процессором Pentium-III 900 Mhz, оперативная память 265 Mbt). Все же о скорости работы программы говорить трудно, поскольку на нее влияют как технические характеристики компьютера, так и характер выполняемого запроса: чем выше частотность леммы или словоформы, тем больше обращений к диску, и, соответственно, тем ниже скорость работы программы.

Алгоритм построения конкорданса выглядит следующим образом:

1. поиск заданной словоформы по таблице словоформ. Если словоформа не найдена, программа выдает сообщение, что этого слова в корпусе нет;
2. сохранение ИН словоформы, обращение к БДТ, к таблице индексов;
3. поиск в индексе записей на искомую словоформу;
4. до тех пор, пока не будет проверена последняя запись по ИН словоформы, выполняются следующие операции:
5. обращение к таблице текстов, поиск контекстов с требуемым ИН;
6. восстанавление данных о тексте и переводе с помощью каталогов текстов из БДП по ИН текстов;
7. если в запросе пользователя определена выборка текстов, проверяется, входит ли текст, в котором был обнаружен пример, в выборку; если нет, то возврат к 3;
8. запись полученного контекста в таблицу с конкордансом;
9. переход к 3.

Текст программы см. в приложении 4.2.

3.2.3. Подготовка текстов в «КОКОС-П»

«КОКОС-П» рассчитан на предварительную обработку текстов: новые тексты регистрируются в БДП, затем они проходят предварительную обработку, в ходе которой программа разбивает исходный текст и перевод(ы) на предложения и записывает их в виде таблиц в БДТ с сохранением информации об абзацах. После этого тексты обрабатываются с помощью программы-стыковщика (о работе этого модуля подробнее см. в следующих разделах), результат сохраняется в виде временной таблицы с состыкованными параллельными текстами. Если для одного и того же текста есть несколько переводов, то текст на ИЯ стыкуется с каждым переводом по отдельности; в итоге оригинал в корпусе текстов оказывается представленным несколько раз, каждый раз — в паре с разными переводами.

С одной стороны, такая стратегия придает системе некоторую гибкость: если один исходный текст одновременно стыкуется с несколькими переводами, структурные единицы оригинала могут в одном тексте укрупняться, а в другом — наоборот дробиться, в результате чего оказывается сложным найти фрагмент исходного текста, к которому можно «привязать» структурно целостные фрагменты всех переводов. С другой стороны, один и тот же исходный текст оказывается представленным в ТТ несколько раз, что нарушает статистику по частотности слов в корпусе[15].

[15] К счастью, в настоящее время в «ПарРус» довольно мало текстов, у которых более одного перевода, и среди них нет текстов большого объема, таким образом, серьезных «перекосов» в данных по частотности не происходит.

Затем программа-индексатор обрабатывает таблицу с параллельными текстами. Обработка происходит следующим образом. Записи из временной таблицы с параллельными текстами добавляются в конец главной таблицы ТТ в БДТ. Одновременно русский и финский фрагменты разбиваются на слова, новые слова добавляются соответственно в русский и финский словники. Для слов, которые уже есть в словниках, обновляется информация по их частотам.

При разбивке на слова программа ориентируется на символы, разграничивающие слова в письменном тексте: пробелы, кавычки, знаки препинания. Некоторые знаки имеют двойственную функцию. Так, знак «-» во многих текстах может быть и дефисом и тире, знак апострофа в финском языке может использоваться и в функции кавычек, и для обозначения «зияния» (напр. *rei'itin*, *rei'illä* и т.п.). Поэтому для обработки этих знаков программа выполняет дополнительную проверку. Создаваемые программой списки можно назвать списками словоформ лишь условно, поскольку среди полученных графических слов есть строки символов, совпадающие с несколькими словоформами (*города* — род. ед. ч. или им. мн.ч. от слова *город*) или могущие быть формами разных слов (*печь* — глагол или существительное). Кроме того, нередко словоформа может состоять из нескольких графических слов, например *на ощупь*, *в результате*, *кое с кем*. По этой причине получаемые таким образом цепочки символов в англоязычной литературе по компьютерной лингвистике называют не *words*, а *tokens*. Таким образом, даже построение словника не является столь простой задачей, как может показаться на первый взгляд.

Параллельно с обновлением словников происходит также обновление индексов в БДТ. Для каждого слова в индексе дописывается ИН словоформы и ИН записи в ТТ.

После того, как индексация нового текста завершена, с корпусом текстов уже можно работать: доступно построение конкордансов на словоформы или псевдоосновы, получение коллокаций и словников по одному или нескольким выбранным пользователем текстам. Однако возможности программы можно расширить, выполнив лемматизацию словника. Для финского и русского субкорпусов были специально написаны модули лемматизации, которые пока работают исключительно со словниками без привлечения контекстов, то есть являются контекстно-свободными.

Лемматизаторы «ЛемКС-Р» и «ЛемКС-Ф» хранятся во внешних к «КОКОС-П» базах данных. Это было сделано в целях экономии места, поскольку и финский, и русский лемматизаторы используют довольно большие по объему словарные массивы. «ЛемКС-Р» и «ЛемКС-Ф» обрабатывают словники, записывая леммы в отдельные таблицы — лемматизированные словники, связывая таблицы через ИН леммы. Словоформы, для которых существует несколько лемм, и словоформы, для которых леммы найдены не были, записываются в таблицы неоднозначных словоформ и в таблицы неопознанных словоформ. После выполнения лемматизации пользо-

ватель может вручную указать леммы для неопознанных словоформ и выбрать правильную лемму для омонимичных форм (подробнее о работе лемматизаторов — в последующих разделах).

3.2.4. Работа со словниками

Как уже говорилось выше, генеральные нелемматизированные словники составляются в ходе индексации текстов, а лемматизированные словники — в ходе лемматизации. Поскольку словники хранятся в таблицах базы данных, пользователь может пользоваться стандартными возможностями СУБД *Microsoft Access*: алфавитная сортировка в восходящем и нисходящем порядке, сортировка по частотам в восходящем и нисходящем порядке, запросы к таблице, фильтры и т.п.

Кроме того, для тех случаев, когда пользователя интересует только часть текстов корпуса, в БДП имеется утилита для построения словников по одному тексту или группе текстов. Полученные таким образом списки словоформ также представляют собой таблицы, с которыми можно выполнять все вышеуказанные операции.

3.2.5. Конкордансы

Центральной операцией системы является получение конкордансов. Программа дает возможность выполнять поиск как по русскому, так и по финскому субкорпусу и получать параллельные конкордансы, т.е. такие конкордансы, в которых даются **битексты** — фрагменты исходных текстов и соответствующие им фрагмент переводов. Выполняя поиск через русский субкорпус, можно выяснить, какие финские слова, выражения, грамматические конструкции используются на практике при переводе тех или иных слов, выражений, грамматических конструкций русского языка. При поиске через финский субкорпус решается обратная задача: какие слова, выражения, грамматические конструкции русского языка соответствуют в параллельных текстах данным финским словам, выражениям, грамматическим конструкциям.

Для построения поисковых запросов и просмотра результатов работы программы организовано диалоговое окно, которое можно видеть на рис. 14.

У пользователя есть возможность составить себе экспериментальный массив из текстов корпуса или выбрать все тексты (как это сделано на рис. 14). При составлении своего субкорпуса можно выбрать вручную один или несколько текстов из списка текстов корпуса, либо сделать автоматическую выборку по одному или нескольким параметрам (например, все тексты одного и того же автора (группы авторов), все тексты одного и того же жанра (группы жанров) и т.п.). Список выбранных текстов появляется в правом окне программы; указывается автор, переводчики, название произведения (в режиме работы с финскими переводчиками имя автора указывается в транслитерации, название произведения — в финском переводе). В окошко под списком текстов выборки выводится ее объем в словоупотреблениях.

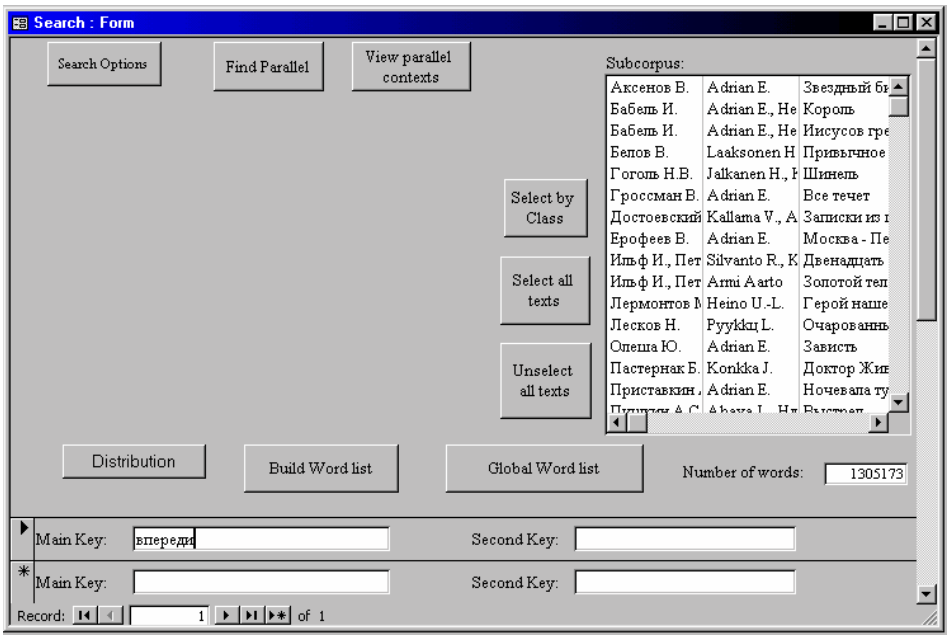

Рис. 14. *КОКОС-П. Диалоговое окно для построения конкордансов*

Запрос может теоретически включать сколько угодно единиц для поиска. В качестве списка для запроса может быть задан даже весь генеральный словник (в этом случае будет получен конкорданс по всем словоформам, зафиксированным в корпусе). Поскольку запрос хранится в виде отдельной таблицы, пользователь может скопировать в нее подготовленный заранее список любой длины. Наиболее часто, впрочем, используется ручной ввод одного или нескольких слов.

Программа позволяет выполнять поиск как на одиночные слова, так и на словосочетания. Для этого в поисковом запросе имеется второе поле, в котором можно указать второй ключ для поиска. Как показывает опыт отладки программы, в большинстве случаев словосочетание можно задать двумя поисковыми ключами (по аналогии с эвклидовым постулатом о том,

что через две точки проходит только одна прямая). «Шум» оказывается минимальным в тех случаях, когда ведется поиск на контактное словосочетание на полное совпадение слов. В остальных случаях качество результатов зависит от того, насколько частотными являются составляющие словосочетания. Например, для поиска сочетания *изо дня в день* лучше задать поисковый образец «дня» + «изо», чем «дня» + «день», который скорее всего даст много «шума».

При поиске словосочетаний программа сначала находит все примеры на первый ключ, а потом отбрасывает те контексты, в которых второй ключ не найден. При проверке знаки препинания игнорируются. Если требуется найти примеры на несколько словосочетаний с одинаковым первым ключом и разными вторыми ключами, имеется возможность задать такой поиск одной записью, используя знак «/»; например, сочетания *много народа* и *много народу* можно искать по запросу «много» + «народа/народу». Поскольку при проверке контекстов на второй ключ словник не используется, в поле для второго ключа можно заносить и контактные словосочетания; например, вышеупомянутое сочетание *изо дня в день* можно искать и по поисковому образцу «дня» + «в день».

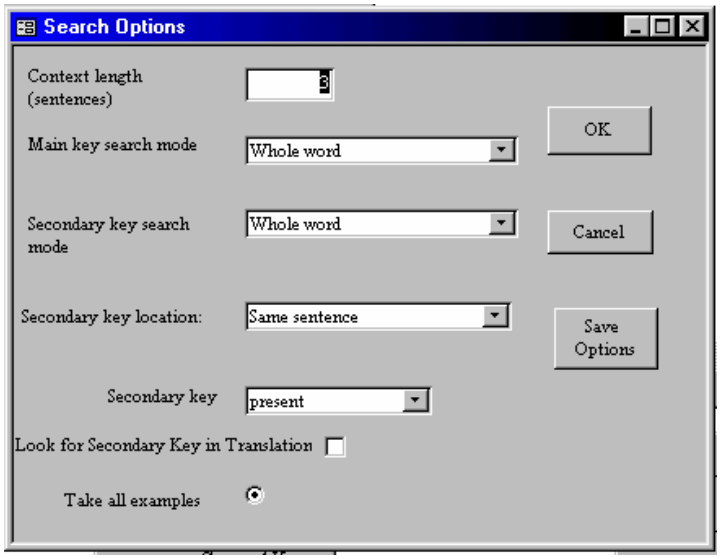

Рис. 15. *КОКОС-П. Диалоговое окно для настройки поиска*

Настройка поиска выполняется с помощью специального диалогового окна (см. рис. 15). Длина контекстов конкорданса (Context length) не ограничена. Программа выполняет поиск по целому слову, по начальной части слова, по концу слова, по любой части слова, а также по лемме (если была выполнена лемматизация словников). Эти режимы поиска могут быть заданы как для первого, так и для второго ключа.

При поиске словосочетаний задается также положение второго поискового ключа относительно первого (Secondary key location): в том же фраг-

менте текста, соседнее слово, соседнее слово справа, соседнее слово слева. Пользователь может искать как контексты, в которых присутствует заданное двумя ключами сочетание (Secondary key: present), так и контексты, в которых присутствует только первый ключ, а второй отсутствует (Secondary key: absent).

Для поиска высокочастотных слов предусмотрена специальная опция (Take all examples / Take every ... example), которую можно настроить на поиск всех примеров или на включение в конкорданс только каждого десятого (сотого, тысячного и т.п.) примера, что позволяет уменьшить объем конкорданса. В противном случае конкордансы на некоторые слова могут образовывать целые корпуса текстов, например, в ПарРус более 100 тысяч вхождений союза *и*.

Конкорданс, полученный в результате работы программы, сохраняется в виде таблицы, которую можно просматривать и редактировать (например, выбрасывать ненужные или сокращать слишком длинные контексты и т.п.), используя специально подготовленную форму. Конкорданс можно распечатать через отчет (стандартный объект Microsoft Access для вывода данных на печать) или экспортировать в Microsoft Word.

В качестве примера работы программы приведем результат поиска сочетания «изо дня в день» (рис. 16, в целях экономии места некоторые контексты даны в сокращенном виде). Для удобства работы пользователя программа выделяет первый ключ поискового запроса знаками «#».

RusExample	**FinExample**	**Original**	**Translation**
Однако почему он обижается на пушку? Какая странная претензия требовать от пушки разнообразия! Отчего вместо пушки лучше не удивится он самому себе, изо #дня# в день стреляющему перечислениями, запятыми и фразами, отчего не прекратит стрельбы журнальным человеколюбием, торопливым, как прыжки блохи?	Kuitenkin, miksi hän moittii tykkiä? Kuinka omituinen pyyde: vaatia tykiltä vaihtelua, erilaisuutta! Sen sijaan että hän vihoittelee tykille, eikö hänen olisi parempi hämmästellä omaa minäänsä, joka päivästä päivään laukoo luetteloita, pisteitä, pilkkuja ja fraaseja; miksi hän ei voisi lopettaa ammuntaa journalistisella ihmisrakkaudella, kiireisellä kuin kirpun hypyt?	Пастернак Б. Доктор Живаго	Konkka J. Tohtori Živago
Нельзя без последствий для здоровья изо #дня# в день проявлять себя противно тому, что чувствуешь; распинаться перед тем, чего не любишь, радоваться тому, что приносит тебе несчастье.	Ilman ikäviä seuraamuksia terveydelle ei voida päivästä päivään tulkita itseään tunteidensa vastaisesti, ei myötäillä sitä, mistä ei pidä, ei iloita siitä, mikä tuottaa onnettomuutta.	Пастернак Б. Доктор Живаго	Konkka J. Tohtori Živago

RusExample	FinExample	Original	Translation
— Как ты думаешь, процесс будут публиковать изо #дня# в день, с прокурорской речью, с допросами, с последним словом подсудимых, или дадут только сообщение о приговоре Военной коллегии?	«Mitä luulet, selostetaanko oikeudenkäyntiä päivittäin syyttäjän puheineen ja kuulusteluineen ja syytettyjen loppupuheenvuoroineen, vai julkaistaanko vain tiedonanto sotilaskollegion tuomiosta?»	Гроссман В. Все течет	Adrian E. Kaikki virtaa
Что это было — ленивое легкомыслие и упование на «кривую, которая вывезет», или же растерянность перед жизнью, что постоянно, изо #дня# в день подсовывает большие и малые распутья?	Kumpaa se oli — laiskaa kevytmielisyyttä ja toivonsa panemista» käyrään joka vie korkealle» vai hämmennystä elämän edessä, sen edessä mikä jatkuvasti, päivästä toiseen suoltaa eteen suuria ja pieniä risteyksiä?	Трифонов Ю. Дом на набережной	Koskinen M. Talo rantakadulla

Рис. 16. *Параллельный конкорданс (изо дня в день)*

В тех случаях, когда примеров оказывается очень много, полезной может оказаться сортировка примеров конкорданса. В «КОКОС-П» используется сортировка «зигзагом» (Oakes 1998: 155–156). Из найденных контекстов программа получает строки для сортировки вида W L_1 R_1 L_2 R_2 L_3 R_3 L_4 R_4 L_5 R_5, где W — первый ключ поискового запроса, L_1, L_2 ... — соседние слова слева, R_1, R_2 ... — соседние слова справа. Конкорданс сортируется по полученным строкам. В результате примеры с поисковым образцом в похожих контекстах оказываются рядом. Так, на рис. 17 видно, что в конкордансе на слово *после* соседними оказываются примеры на сочетания *после болезни* и *после дождя*.

RusExample	FinExample	Original	Translation
Я травил за борт, и меня всего трясло. Меня выворачивало, черт знает как. Потом стало холодно и очень легко, как #после# болезни. Я лежал животом на борту и представлял себе, как усмехнется Баулин, когда я обернусь. А черт с ним, в конце концов.	Minä pumppasin yli laidan ja tärisin kauttaaltaan. Kaikki kääntyi sisälläni, piru ties miten. Sitten tuntui kylmälle ja hyvin keveälle niin kuin sairauden jälkeen. Minä riipuin mahallani partaalla ja kuvittelin, miten Baulin virnistelee, kun minä käännyn. Mitä pirua sillä oikeastaan on väliä.	Аксенов В. Звездный билет	Adrian E. Matkalippu tähtiin
Она шла тихо, еще слабая #после# болезни; сокращала дорогу узкими тропками, глядела на деревню и, улыбаясь, плакала от радости.	Hän kulki hiljakseen, voimat eivät olleet vielä sairauden jäljiltä palautuneet. Hän oiusti pikkupolkuja, hymyili kylää katsellessaan ja itki ilosta.	Белов В. Привычное дело	Laaksonen H. Tuttu tarina

RusExample	FinExample	Original	Translation
#После# дождя город приобрел блеск и стереоскопичность.	Sateen jälkeen kaupunki kävi kimaltavaksi ja stereoskooppiseksi.	Олеша Ю. Зависть	Adrian E. Kateus
И, увидев опять перед собой провисшее и некрасиво заросшее, как замшелое, лицо Андрея, его провалившиеся глаза, острые и измученные страданием, его полусогнутую настороженную фигуру в грязной одежде; попав #после# дождя в сырое темное зимовье с горьким запахом спертого, задушенного воздуха, - увидев и ощутив все это, Настена с новой болью содрогнулась.	Ja nähtyään taas edessään Andrein roikkuvat ja rumasti .karvoittuneet - kuin sammaloituneet - kasvot, hänen vajonneet silmänsä, terävät ja kärsimyksen piinaamat, hänen likaisten vaatteiden verhoaman puoleksi kumaran ja varuillaan olevan hahmonsa; tultuaan sateen jälkeen tähän kosteaan pimeään kämppään, jonka ilma oli tunkkainen ja haisi ummehtuneelta, - nähtyään ja tunnettuaan kaiken tuon Nastena vavahti uudenlaisesta tuskasta.	Распутин В. Живи и помни	Adrian E. Elä ja muista

Рис. 17. *Фрагмент конкорданса, отсортированного «зигзагом».*

Еще одна специальная возможность, предусмотренная в программе — проверка наличия или отсутствия заданного слова в переводе. Для этого надо в настройке поиска отметить опцию «Look for secondary key in translation». В этом случае система будет искать второй ключ в параллельном фрагменте. При этом работает режим сравнения строк («Whole word» / «Start of the Word» / «Any part of the word»). Если установлена опция «Secondary key: absent», то программа не будет включать в конкорданс примеры, в которых второй ключ присутствует. Такой режим поиска позволяет дополнительно уменьшить «шум» при составлении конкорданса и получить только те контексты, где заданное слово переведено с использованием заданного переводного эквивалента.

На рис. 18 в качестве примера приводится конкорданс, полученный в результате поиска на *весьма* в качестве главного ключа и *sangen* в качестве второго ключа (подчеркнуто).

RusExample	FinExample	Original	Translation
— Но к делу, к делу, Маргарита Николаевна. Вы женщина #весьма# умная и, конечно, уже догадались о том, кто наш хозяин.	— Mutta asiaan, Margarita Nikolajevna. Te olette sangen älykäs nainen ja olette tietenkin jo arvannut, kuka on isäntämme.	Булгаков М.А. Мастер и Маргарита	Bulgakov M. Saatana saapuu Moskovaan. Heino U.-L.

RusExample	FinExample	Original	Translation
Состояние у ней было #весьма# хорошее, не столько наследственное, сколько благоприобретенное мужем. Обе дочери жили с нею; сын воспитывался в одном из лучших казенных заведений в Петербурге.	Hän oli sangen hyvissä varoissa; hän oli osaksi perinyt, mutta enimmäkseen saanut mieheltään, joka varat oli hankkinut. Hänen molemmat tyttärensä asuivat kotona, mutta poika oli eräässä Pietarin parhaimmista valtion oppilaitoksista kasvatustaan saamassa.	Тургенев И.С. Дворянское гнездо	Turgenev I. Aateliskoti. Heino U.-L.
Ключи он тотчас же вынул; все, как и тогда, были в одной связке, на одном стальном обручке. Тотчас же он побежал с ними в спальню. Это была очень небольшая комната, с огромным киотом образов. У другой стены стояла большая постель, #весьма# чистая, с шелковым, наборным из лоскутков, ватным одеялом.	Avaimet hän veti esiin heti; ne olivat kuten ennenkin samassa nipussa, kaikki samassa teräsrenkaassa. Ne kädessään hän juoksi makuuhuoneeseen. Se oli hyvin pieni huone, nurkassa valtavan iso ikonikaappi. Toisella seinällä oli kookas, sangen siisti vuode, vanupeitto päällystetty silkkitilkuilla.	Достоевский Ф.М. Преступление и наказание	Dostojevski F.M. Rikos ja rangaistus. Konkka Juhani
— Очень добрый и любознательный человек, — подтвердил арестант, — он высказал величайший интерес к моим мыслям, принял меня #весьма# радушно...	— Erittäin hyvä ja tiedonhaluinen ihminen, vanki vakuutti. — Hän osoitti mitä suurinta kiinnostusta ajatuksiani kohtaan, otti minut vastaan sangen ystävällisesti ...	Булгаков М.А. Мастер и Маргарита	Bulgakov M. Saatana saapuu Moskovaan. Heino U.-L.

Рис. 18. *Конкорданс на словарные эквиваленты*

3.2.6. Получение сведений по частотности слова в разных текстах массива

Довольно часто при работе с корпусом текстов требуется установить, насколько равномерно распределено данное слово или сочетание слов по разным текстам корпуса. Для этого была написана специальная утилита, которая вызывается из того же диалогового окна, что и конкордансы.

Запрос задается таким же способом, как и запрос для построения конкорданса. Нажав кнопку «Distributions», пользователь получает абсолютные и относительные частоты по разным текстам корпуса. На рис. 19 приведены частоты для наречия *быстро* в некоторых из текстов ПарРус, полученные с помощью этой утилиты.

Word	Source	Count	RelFr
быстро	Ильф И., Петров Е. Двенадцать стульев	58	0,74
быстро	Ильф И., Петров Е. Золотой теленок	55	0,63
быстро	Пастернак Б.Л. Доктор Живаго	43	0,29
быстро	Фадеев А. Разгром	29	0,65
быстро	Солженицын А.И. Один день Ивана Денисовича	26	0,80
быстро	Приставкин А. Ночевала тучка золотая	24	0,36
быстро	Семенов Ю. Семнадцать мгновений весны	21	0,27
быстро	Распутин В. Живи и помни	16	0,25
быстро	Трифонов Ю. Дом на набережной	16	0,37
быстро	Стругацкие А. и Б. Попытка к бегству	15	0,53
быстро	Лермонтов М.Ю. Герой нашего времени	15	0,36

Рис. 19. *Абсолютные и относительные частоты слова быстро в некоторых из текстов ПарРус*

3.2.7. Коллокации

Еще одна функция, отсутствовавшая в ранних пакетах программ для работы с корпусами текстов, но довольно популярная в программах современных — поиск коллокаций. Коллокации — это словоформы, встречающиеся в ближайшем окружении заданного слова или словоформы. Такого рода информация помогает найти клише и идиомы с использованием заданного слова, установить устойчивые ассоциации и даже выяснить, в каких значениях употребляется заданное слово в корпусе текстов.

В «КОКОС-П» можно получать коллокации как для русских, так и для финских текстов. Даже в языках со свободным порядком слов, каковыми являются русский и финский, наиболее важную роль для изучения сочетаемости слова являются ближайшие два соседа справа и слева. Однако иногда и более далекие соседи связаны с заданным словом семантически или даже грамматически. «КОКОС-П» получает набор коллокаций для контекста ±5 (т.е. пять соседних слов слева и пять соседних слов справа[16]).

Для получения коллокаций в пакете «КОКОС-П» организовано следующее диалоговое окно (рис. 20).

[16] Программа рассматривает в качестве коллокаций только слова, встретившиеся в пределах одного и того же предложения.

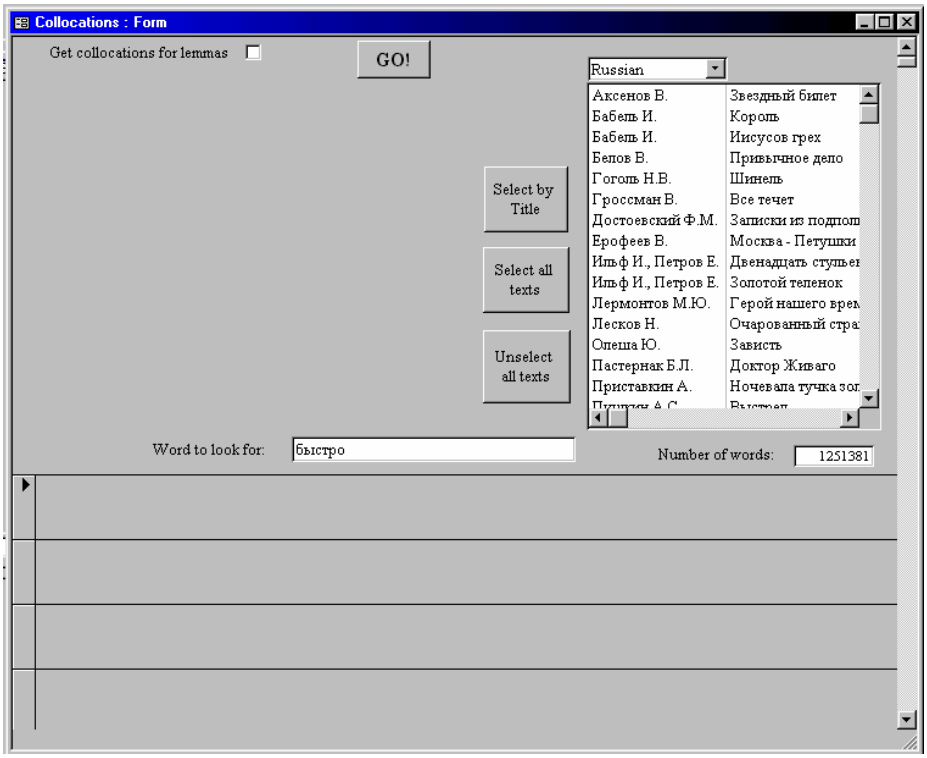

Рис. 20. *КОКОС-П. Диалоговое окно для получения коллокаций*

Так же, как и при получении конкордансов, пользователь может выбрать тексты, по которым будет выполняться поиск, либо выделить все тексты. Одна и та же оболочка используется для работы и с русским, и с финским субкорпусом. Коллокации можно получать как для словоформ, так и для лексем.

После завершения поиска в нижней части окна появляется список найденных коллокаций (рис. 21), которые можно просматривать, сортировать по частотам разных коллокатов или в алфавитном порядке, а также выводить на печать или копировать в текстовый процессор.

Программа (коды см. в приложении 4.3) работает следующим образом: на заданное слово строится конкорданс с длиной контекста равной одному предложению, затем для каждого контекста выделяется пять левых и пять правых коллокаций. Информация обобщается в виде сводной таблицы. Затем для каждого обнаруженного коллоката вычисляется коэффициент значимости коллокации. Использовался коэффициент *z*, предложенный Бэрри-Роггхом (Berry-Rogghe).

Сначала вычисляется вероятность появления слова B в одном контексте с A. Вероятность появления B в тех частях текста, где A не встречается, равна

$p = F_c/(Z-F_n)$, где Z — количество словоупотреблений в корпусе текстов, F_n — частота A, F_c — частота B.

Ожидаемая частота появления A вместе с B в пределах контекста в S слов будет таким образом равна:

$E = pF_nS$

Для того, чтобы оценить разницу между наблюдаемой (K) и ожидаемой (E) частотами, применяем следующую формулу:

$z = \frac{K - E}{\sqrt{Eq}}$, где q = 1-p.

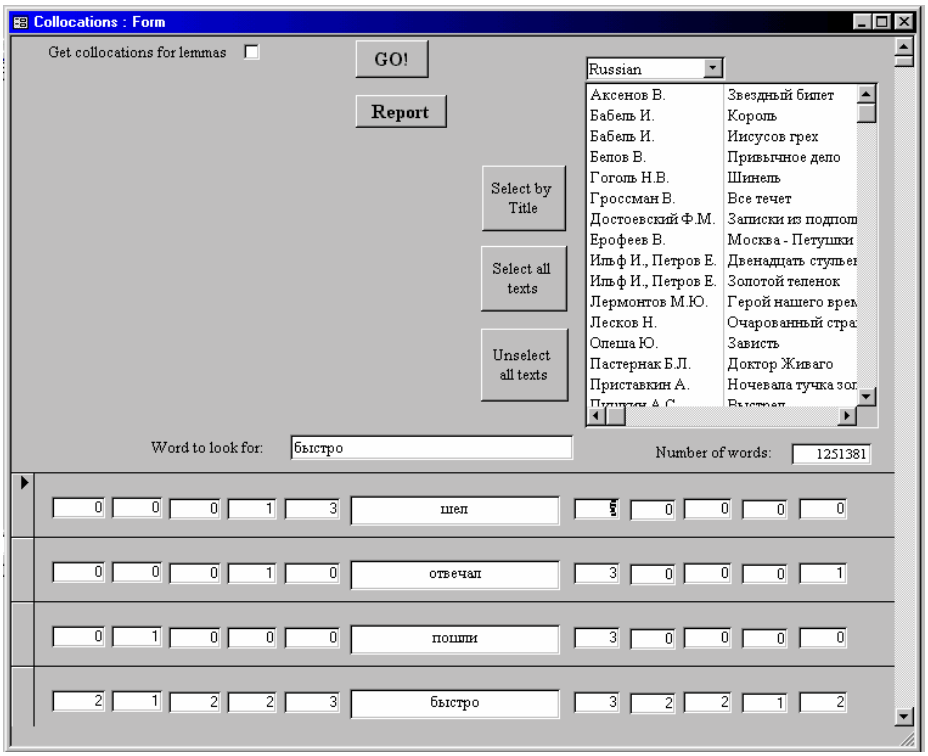

Рис. 21. *КОКОС-П. Диалоговое окно для получения коллокаций со списком коллокаций на слово быстро*

Для статистически значимых коллокаций величина *z* должна составлять по крайней мере 2,576 (подробнее см. Oakes 1998: 163–166).

Если полученный коэффициент оказывается слишком низким или сумма частот всех коллокаций равна единице (то есть появление данного коллоката скорее всего случайно), то запись уничтожается.

Таким образом, в итоговом списке присутствуют только те слова, которые фиксируются в ближайшем окружении заданного слова чаще, чем это прогнозирует теория вероятности.[17] В качестве примера приведем список

[17] В некоторых случаях, в процессе исследования требуются не только статистически значимые коллокаты, а только высокочастотные коллокаты, для таких целей был написан режим без удаления статистически незначимых коллокатов.

первых десяти коллокатов слова *быстро*, отсортированных в нисходящем порядке по частоте первого правого коллоката (табл. 6).

Таблица 7. Первые 10 коллокатов слова *быстро* по русскому субкорпусу «ПарРус», отсортированные по первому правому соседу

Collocation	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5
шел	0	0	0	1	3	5	0	0	0	0
отвечал	0	0	0	1	0	3	0	0	0	1
пошли	0	1	0	0	0	3	0	0	0	0
быстро	2	1	2	2	3	3	2	2	1	2
пробегали	0	0	0	0	0	2	0	0	0	0
повернул	0	0	0	0	0	2	0	0	0	0
зашагал	0	0	0	0	0	2	0	0	0	0
бегает	0	0	0	0	0	2	0	0	0	0
вынырнет	0	0	0	0	0	2	0	0	0	0

В некоторых случаях, например, когда исследуется грамматическая сочетаемость слов, помимо статистически значимых коллокаций, интерес могут представлять все коллокации заданного слова/словоформы. Для проведения таких исследований функцию отбрасывания коллокаций с низким *z* можно отключить.

3.2.8. Перспективы развития «КОКОС-П»

Программа довольно быстро развивается и модифицируется, для нее создаются все новые модули. Использование *Microsoft Access* позволяет использовать как написанные специально для программы процедуры, так и стандартные операции СУБД (сортировки, фильтры, запросы, отчеты и т.п.). Программа хорошо приспособлена к работе с большими массивами текстов, позволяет легко задавать различные выборки из текстов. Хотя программа писалась для поддержки русско-финского параллельного корпуса текста, ее нетрудно адаптировать и для других пар языков.

При разработке программы было решено ориентироваться на уход от работы с текстовыми файлами. Традиционно исследователи и операторы «готовят» тексты, подгоняют их под заданный стандартный формат, создают для них «заголовки» (headers), в полях которых указывается вся релевантная информация о текстах, выполняют вручную или с помощью специальных утилит грамматическую или иную разметку текстов. Программа же представляет собой лишь пустую оболочку, которая обрабатывает цепочки символов. Такой подход создает определенные проблемы. В корпусе, организованном таким образом, очень сложно что-либо

изменять. Например, после изменения «заголовков», как правило, необходимо переиндексировать корпус, а переделать грамматическую разметку корпуса еще труднее. Довольно трудно задать массив текстов. Все эти проблемы решаемы в рамках большого проекта, однако оказываются довольно трудными в случае, если работа выполняется «малыми силами».

«КОКОС-П» и «КОКОС» хранят описание текста отдельно от самого текста. Описание текста хранится в виде таблицы, его легко исправлять и модифицировать. Сами тексты копируются в базу данных, это облегчает работу с текстами (правда объем, который занимает корпус текстов, оказывается довольно большим, больше, чем при работе с традиционными текстовыми файлами). Проблема разметки текста при использовании такого подхода решается следующим образом: из «физической» разметка становится «виртуальной». Разметка будет выполняться в виде дополнительных словников, связанных с основным нелемматизированным словником. Подобно тому, как лемматизированный словник «привязывается» к словоформам, к основному словнику могут «привязываться» списки грамматических форм, имеется возможность составлять списки семантических категорий, связанные со списками лексем и т.д., и т.п. Такого рода решение позволит выполнить на одном и том же корпусе текстов несколько разметок, что при использовании традиционной технологии создает некоторые трудности.

Для работы с параллельными корпусами текстов использование данного подхода также представляется довольно продуктивным. Например, используя параллельные контексты, можно организовать поиск возможных переводных эквивалентов; к основным словникам можно подключать двуязычные словари и т.п.

3.3. Лемматизация в «КОКОС-П»

В процессе работы над программным обеспечением «ПарРус», кроме чисто технических задач, о которых шла речь в предыдущем разделе, пришлось решать и более серьезные проблемы теоретического характера, а именно проблему лемматизации для русских и финских текстов и проблему стыковки параллельных текстов. К разработке своих систем вместо использования уже существующих нас подталкивал мониторный характер создаваемого корпуса и связанную с этим необходимость регулярного пополнения корпуса новыми текстами, их стыковку и лемматизацию обновленного генерального словника. Самостоятельная разработка лемматизаторов и стыковщиков позволяет также лучше учесть характер задач, поставленных перед проектом. В настоящем и следующем разделах будет обсуждаться проблема лемматизации, затем будет рассмотрена проблема стыковки параллельных текстов.

Значение автоматического морфологического анализа и синтеза для прикладных исследований переоценить трудно. Морфологический модуль оказывается необходимым и в орфографических корректорах, и в системах оптического распознавания символов, и в системах машинного перевода, и в системах искусственного интеллекта. Особенно важным морфоанализ оказывается для языков с богатой морфологией, к которым, безусловно, следует отнести и русский язык. Количество различных словоформ в таких языках столь велико, что хранить грамматическую информацию в виде словарей словоформ представляется нерациональным, хотя бы развитие вычислительной техники и подталкивало к подобному расточительству.

Важным компонентом систем, обслуживающих корпуса текстов, является модуль «лемматизации», т.е. получения для каждой словоформы ее начальной формы, «леммы». Выполнение лемматизации не только упрощает построение поисковых запросов, уменьшает «шум» при построении конкордансов, но и позволяет получить для корпуса словник словарного типа. У пользователя появляется возможность породить словарный массив с примерами, т.е. фактически «полуфабрикат» словаря, в котором уже будут обозначены словарные статьи, состоящие из «лексического входа», грамматического описания и корпуса примеров.

Для обработки корпусов текстов, поддерживающихся с помощью оболочек «КОКОС» и «КОКОС-П», были разработаны два лемматизатора: «ЛемКС-Р» (Лемматизатор контекстно свободный), обрабатывающий русские словники, а позднее — «ЛемКС-Ф», обрабатывающий финские словники. Оба лемматизатора — контекстно-свободные и позволяют найти однозначные начальные формы примерно для 70% слов списков. Словники русских и финских текстов были обработаны с помощью лемматизаторов, затем было выполнено постредактирование результатов работы. Были указаны леммы для неопознанных словоформ с частотой более 10 и для омонимичных словоформ с частотой более 20. Пороги частот были выбраны исходя из физических возможностей. Более высокий частотный порог для омонимичных форм связан с большей трудоемкостью работы (практически во всех случаях требуется просматривать контексты). Обработать вручную менее частотные словоформы оказалось невозможным, поскольку в русском субкорпусе осталось около 17 тыс. неопознанных и около 4 тыс. омонимичных словоформ, в финском субкорпусе — около 45 тыс. неопознанных и около 10 тыс. омонимичных словоформ. Большее количество неопознанных словоформ в финском субкорпусе не связано с более низким качеством работы финского лемматизатора, а скорее с особенностями финского словоизменения.

Программы создавались как служебные утилиты, попыток разработки принципиально новых стратегий к морфологическому анализу не делалось. В целом использовались уже существующие подходы (см., например, Сидоров 1995). В данном разделе делается попытка обобщить личный опыт

работ над лемматизаторами для русского и финского языков и выработать общие рекомендации по разработке систем данного класса.

3.3.1. Лемматизация в «КОКОС-П»: русский язык

Общая характеристика системы

Когда говорят о морфологическом анализе, как правило, имеется в виду их применение в первую очередь в орфографических корректорах. Вообще, следует отметить принципиальное отличие лемматизатора от орфографического корректора, «спеллера». Хотя спеллер зачастую выполняет морфологический анализ слова для того, чтобы определить, существует ли данная словоформа, эта программа вовсе не должна «знать» все слова. Программа работает в интерактивном режиме, взаимодействуя с пользователем в тех случаях, когда слово оказалось не зафиксированным в словаре системы. Для лемматизатора такой принцип оказывается малоприемлемым, поскольку анализируемые массивы текстов могут быть очень большими, составлять миллионы и даже десятки миллионов словоупотреблений, в результате чего даже при высоком проценте однозначных разборов «шум» может быть очень значительным — десятки тысяч неоднозначно проанализированных или неопознанных словоформ. Скорость работы для спеллера принципиально важна, поскольку он работает в реальном времени, для лемматизатора же важность этой характеристики зависит от приложения, частью которого он является. В «КОКОС» и «КОКОС-П» скорость работы лемматизатора не очень существенна, поскольку лемматизация не выполняется в реальном времени: словники корпуса лемматизируются, затем проводится постредактирование, после чего начинается его эксплуатация.

Другое отличие лемматизатора от спеллера — отсутствие алгоритма поиска правильного написания слова, «подсказки». Вообще, эти два класса программ строятся на двух полярных принципах: первый основан на презумпции наличия в каждом тексте ошибок и опечаток, второй — на оптимистической идее об отсутствии в тексте ошибок. Объединить эти подходы пока не представляется возможным. Недостатки и спеллеров и лемматизаторов заложены в их природе: оба класса программ пока используют по преимуществу только морфологическую информацию, игнорируя или почти игнорируя синтаксис и семантику.

В качестве исходного словарного массива для лемматизатора «ЛемКС-Р» использовалась база данных объемом около 100 000 записей, любезно предоставленная Г.О. Сидоровым. Эта база данных представляет собой словарь псевдооснов, полученный на основе «Грамматического словаря русского языка» А.А. Зализняка (Зализняк 1980). База данных затем была

дополнена именами собственными, географическими названиями, а также лексикой, появившейся в языке уже после выхода словаря. В то же время ряд архаичных слов, перегружавших базу данных и являвшихся потенциальным источником грамматической омонимии (например, *юр*, *миро*, которые в некоторых формах совпадают с *Юра* и *мир*), пришлось убрать. «Обкатка» системы происходила на довольно больших текстовых массивах: русская проза XIX века (в то время объем корпуса составлял около 3,5 млн. словоупотреблений) и XX века (около 10 млн. словоупотреблений), российская пресса 1999 г. (1 млн. словоупотреблений). В настоящее время в базе данных лемматизатора около 130 тыс. записей[18].

При построении лемматизатора ставилась задача получить быстро и эффективно работающий модуль, делающий минимальное количество ошибок и по возможности решающий проблему грамматической омонимии. Задача построения морфосинтезатора не ставилась, система ориентирована в первую очередь на анализ.

Вся необходимая для работы «ЛемКС-Р» информация хранится в базе данных *Microsoft Access*, процедуры морфоанализа написаны на языке *Visual Basic*. Программа работает в качестве внешнего модуля для систем «КОКОС» и «КОКОС-П», заполняя и обновляя таблицы для лемматизированного словника и связывая создаваемые записи с записями основного словника. Однако программа используется и как отдельная система для определения начальных форм отдельных слов или для лемматизации списков словоформ.

Массивы данных

При работе «ЛемКС-Р» используются два основных массива данных: словарь основ и таблица окончаний.

В словаре основ фиксируются основы (точнее псевдоосновы, поскольку они далеко не всегда совпадают с основами, выделяемыми по правилам грамматической теории), грамматический класс по Зализняку (*мо*, *жо*, *п*, *св*, *нсв* и т.п.) и тип (*1, 2, 3* и т.п., неизменяемым словам и словам с нестандартным словоизменением присваивается тип *0*), схема ударения[19] (*a, b, c* и т.д.) (подробнее см. Зализняк 1980).

---

[18] Точное количество слов в словаре сообщить затруднительно, поскольку, с одной стороны, для многих лексем дано две или более основы или даже готовые формы, но, с другой стороны, обработка префиксов сильно увеличивает количество распознаваемых системой словоформ.

[19] Хотя система не выполняет расстановку ударений, информация о типе ударения все же в некоторых случаях оказывается необходимой, поскольку нередко у слов, относящихся к одному типу словоизменения, но имеющих разные схемы ударения, грамматические

Для проверки получаемых лемм программа должна располагать некоторой дополнительной информацией, например о том, для образования каких форм используется данная основа, а также о возможности образования от данной основы тех или иных грамматических форм (например, кратких форм у прилагательных). Информация такого рода хранится в поле дополнительных помет.

Кроме того, следует отметить, что грамматический класс не всегда совпадает с традиционными грамматическими характеристиками слова. Например, существительные мужского рода *папа* и *дядя* относятся к классу *жо* (существительное женского рода, одушевленное), поскольку их грамматическая парадигма полностью совпадает с парадигмой существительных женского рода *мама* и *тетя*. Тем не менее, синтаксически эти существительные ведут себя по-другому, сочетаясь с прилагательными мужского рода (*занятой папа*). Существительные *столовая* и *рабочий* относятся к грамматическому классу *n* (прилагательные), поскольку они изменяются по адъективному склонению, однако они не изменяются по родам и не образуют кратких форм. Двойственна и природа причастий, которые, являясь глагольной формой, изменяются как прилагательные. По этим причинам к таблице было добавлено поле «часть речи».

Итак, разные основы для одного и того же слова хранятся в разных записях базы данных. Правила образования лемм также оказываются не столь простыми, как может показаться на первый взгляд. Например, для существительных pluralia tantum леммой является именительный падеж множественного числа, а не именительный единственного числа, как для прочих существительных, леммой причастия является инфинитив глагола, а не именительный падеж мужского рода единственного числа, как для прилагательных. Таким образом, порождение леммы может существенно замедлить процесс морфоанализа. По этой причине в словаре основ хранятся, кроме псевдооснов, готовые леммы.

Грамматические таблицы представляют собой списки псевдоокончаний с грамматическими описаниями, организованные в виде двух связанных таблиц. Например, окончание *—и* может быть окончанием краткой формы множественного числа прилагательных классов 2a, 3a, 4a, 6a, 3b, 4b, родительного падежа единственного числа и именительного и винительного падежей множественного числа для существительных женского рода классов 2, 3, 4, 6, 7, 8, именительного и винительного падежей множественного числа существительных мужского рода классов 2, 3, 4, 6, 7, 8, существительных среднего рода классов 7 и 8, местоимений классов 1, 2, 6, а также императива единственного числа. Используемая форма представления оказывается довольно удобной, поскольку позволяет хранить окончания для разных грамматических классов в одном массиве, а также делает возмож-

---

парадигмы могут различаться с точки зрения орфографии, например, *солнце* и *винцо*, *новый* и *живой*).

ным фиксировать кроме основных также вариантные и факультативные окончания, что оказывается довольно сложным, если попытаться использовать в качестве формы представления традиционные грамматические таблицы.

Работа программы

Алгоритм работы программы в общих чертах выглядит следующим образом.

1. На первом этапе программа пытается найти в словаре основ все слово. Если слово найдено, проверяется зона грамматического типа и зона помет. В том случае, если основе присвоен тип «0» (неизменяемое слово) или в зоне грамматических помет обозначено, что это нестандартная готовая форма, приведенная полностью, то лемма, соответствующая данной форме, добавляется к списку найденных лемм. Если же грамматический тип оказался «ненулевым», то программа проверяет, может ли у данной основы быть нулевое окончание. В случае отрицательного ответа гипотеза отвергается. На этом этапе заканчивается анализ таких слов, как *и*, *пюре*, *им*.
2. Далее программа отсекает символы справа и проверяет, не зафиксирована ли правая подстрока в таблице окончаний. Если потенциальное окончание обнаружено, то проверяется, не зарегистрирована ли левая подстрока в таблице основ. Если да, то проверяется, возможно ли сочетание данной основы с данным окончанием. Если в таблице зарегистрировано несколько совпадающих основ и у окончания есть несколько значений, то проверяются все возможные варианты. Не опровергнутые гипотезы считаются подтвердившимися, леммы добавляются к списку найденных лемм.
3. Следующий шаг работы программы — проверка на постфикс ся[20]. Если на конце слова обнаружена последовательность ся или сь, то она отсекается и происходит возврат к шагу 2 с уточнением, что анализируемая строка должна быть глагольной формой. В словаре основ возможность или необходимость использования постфикса *ся* отмечается в поле грамматических помет. В том случае, если анализ подтвердил возможность существования данной формы у данной основы, то к лемме, в зависимости от того, как оканчивается основа, «приставляется» постфикс *ся* или *сь*.
4. Внешний цикл программы — проверка на префиксы. Значительная часть префиксальных образований зафиксирована в словаре, однако некоторые префиксы — например, префикс не — могут присоединяться

[20]Постфикс императива —*те* для упрощения процедуры анализа считается частью псевдоокончания императива и поэтому в расчет не принимается.

к очень большому количеству основ и по этой причине проверяются при анализе. Если в начале слова обнаружен потенциальный префикс, он отсекается, а «остаток» проходит проверку на этапах 1–3. Если анализ прошел успешно и полученная форма может иметь означенный префикс, то к лемме спереди «приставляется» подходящий вариант префикса. Поскольку в русском языке довольно мало слов с двумя или более префиксами (типа монстра *недоперевыполнить*), рекурсия по префиксам не производится.

Анализ проводится до тех пор, пока не будут проверены все потенциальные основы и окончания, и в этом — еще одно отличие от орфографического корректора, который выполняет анализ до тех пор, пока не удается найти первую приемлемую интерпретацию словоформы. Тексты основных модулей программы дан в приложении 4.5.

Лемматизатор в корпусе текстов

«ЛемКС-Р» используется для «вторичной» обработки текстов, им обрабатывается словник, составленный индексатором системы «КОКОС». Такое решение дает возможность обрабатывать текст значительно быстрее, ибо значительную часть словника любого текста составляют союзы, предлоги, частицы и местоимения, которые постоянно повторяются. Приведем в качестве примера статистику, полученную из «ТамРус», корпуса текстов русской художественной прозы XX века[21]. Как уже было сказано выше, объем корпуса — 10 млн. словоупотреблений. Приведем данные по десяти самым частотным словоформам корпуса (табл. 8).

Таблица 8. Десять самых частотных словоформ корпуса «ТамРус»

Слово	**Частотность**	**Доля в корпусе (%%)**
и	387587	4%
в	256871	3%
не	200387	2%
на	177753	2%
что	121671	1%
с	118186	1%
он	114105	1%

[21] В этом разделе, помимо данных «ПарРус», будут привлекаться и данные «ТамРус», поскольку этот корпус текстов значительно больше по объему, в связи с чем эмпирический материал и количественные данные оказываются в иллюстративном плане нагляднее.

Слово	Частотность	Доля в корпусе (%%)
я	106245	1%
а	97102	1%
как	74662	1%
	Итого:	16%

Таким образом, 10 самых частотных слов корпуса составляют 16% от общего количества словоупотреблений в текстах. Этот сам по себе любопытный факт одновременно подтверждает целесообразность выполнения лемматизации на словнике, а не на самих текстах, что привело бы к малоосмысленному многократному анализу все того же союза *и*, который при лемматизации словников «ТамРус» был бы выполнен более 300 000 раз, а при анализе «ПарРус» — более 100 000 раз.

Неповторяющихся словоформ в «ПарРус» зафиксировано 85 820, а всего в тексте словоформ более 179 635. Таким образом, около половины словоформ корпуса повторяются хотя бы два раза. В случае последовательной лемматизации текста на обработку повторяющихся слов будет уходить значительная часть времени работы программы. Обработка же словника позволяет анализировать каждое слово только один раз, что существенно экономит время.

Следует отметить, что обработка словников, а не самих текстов не означает ориентацию исключительно на контекстно-свободный анализ, о недостатках которого будет сказано ниже. Поскольку по корпусу текстов составляется и индекс (подробнее об этом см. в предыдущем разделе), сохраняется возможность проверки ближайшего контекста анализируемых слов и даже анализа целого предложения.

Результат работы программы записывается в три таблицы: 1) таблицу однозначных лемм (*человеку, человека* → ***человек***), 2) таблицу неоднозначных лемм (*агротехники* → *агротехника, агротехник*), 3) таблицу неопознанных словоформ, в которую попадают архаизмы (*скрыпеть, усумниться*), разговорные и просторечные формы (*щас, эфтот*), редкие слова (*потпрукивать, морильщица*), окказионализмы (*неотдание*), имена собственные (*Юсуп*), варваризмы (*адье*) и т.п. В дальнейшем пользователь может вручную редактировать результаты лемматизации. К сожалению, такой ручной работы может оказаться довольно много.

Обработка русских текстов корпуса «ПарРус», объемом около 2,2 млн. словоупотреблений происходила следующим образом. Общее количество словоформ в текстах составило 180 000.[22] Около 4 800 словоформ попали

[22] Здесь и далее даются округленные числа, поскольку точные цифры все равно ничего не дают: нередко компьютер признает в качестве словоформы то, что человек никогда словоформой не назовет. Так аббревиатура *т.е.* при обработке будет опознана как две словоформы — *т* и *е*, хотя более разумно интерпретировать как одну словоформу. Аналогично, дискуссионным является вопрос, считать ли *300*, *+*, *2%* словоформами (см. тж. стр. 83 настоящей работы).

таблицу неоднозначных лемм, 17 000 не были опознаны, остальные 158 200 словоформ были успешно проанализированы, в результате чего было получено 43 600 однозначных лемм. Из 17 000 неопознанных словоформ около 4 700 встретились в текстах более одного раза, 2 000 — более двух раз, 1 300 — более трех раз и лишь около 220 — более 10 раз. Несколько хуже ситуация в таблице неоднозначных лемм: около 750 из 4 800 записей таблицы имеют частоту более 10. Таким образом, КПД системы составляет около 90%. Такое высокое качество связано с тем, что определяется именно начальная форма без учета частеречной принадлежности слова. В противном случае уровень омонимии был бы гораздо выше (например, слово *что* может быть и местоимением, и союзом, *venäläinen* — и существительным, и прилагательным). От определения части речи на данном этапе было решено отказаться, поскольку в этом случае потребовалось бы вводить дополнительные пометы в грамматических словарях, причем значительную часть работы автоматизировать невозможно. Однако в перспективе эта задача решаться будет.

Для получения грубых результатов, не претендующих на статистическую точность, можно пользоваться результатами лемматизации без постредактирования или же выполнить его только для высокочастотных неопознанных или неоднозначно опознанных словоформ. В случае же снятия грамматической неоднозначности и определения лемм для неопознанных словоформ объем работ может оказаться очень большим.

Получаемый после работы программы лемматизированный словник, «привязанный» к словнику словоформ, не только упрощает получение конкордансов. Лемматизация позволяет более «рельефно» представить себе словарный состав анализируемого корпуса текстов, получать статистическую информацию об употреблении не только словоформ, но и лексем.

Все это, несомненно, расширяет возможности использования конкорданса в лингвистических исследованиях.

Грамматические таблицы versus алгоритм

Полученная система морфоанализа ни в коей мере не является идеальной. Она была задумана как сугубо служебный модуль, недостатки работы которого должны по возможности компенсироваться работой других модулей системы.

Грамматическая классификация А.А. Зализняка, несомненно, является наиболее последовательной и структурированной из ныне существующих грамматических описаний русской морфологии. Это делает ее настоящей находкой для программистов. Тем не менее, в процессе создания программы появлялись самые разные технические проблемы.

Первая группа проблем связана с самой классификацией. В некоторых случаях у разных (по Зализняку) грамматических классов оказывались

почти идентичные наборы окончаний. Например, у существительных среднего рода парадигмы для типов 1а и 3а (*болото* и *благо*), 4а и 5а (*жилище* и *солнце*) с точки зрения орфографии идентичны. Прилагательные в нашей системе различаются не столько по типу, выделяемому в словаре Зализняка, сколько по схеме ударения. С другой стороны, принадлежность слов к одному и тому же грамматическому типу не всегда означает, что у них будет идентичный набор окончаний. У слов, относящихся к одному типу, но имеющих разные схемы ударения, зачастую оказываются разные наборы окончаний (ср. *редкий* и *сухой*). Довольно много типов различаются только одной или двумя формами, например, у существительных мужского рода типы 1а (*завод*) и 3а (*чайник*) различаются только формой именительного-винительного падежа множественного числа (*заводы* — *чайники*). Таким образом, компьютерная реализация грамматических таблиц оказывается весьма неуклюжей: для многих грамматических типов указываются абсолютно одинаковый набор псевдоокончаний, имеющих идентичное грамматическое значение. Приведем в качестве примера фрагмент грамматических таблиц для окончания *-ам* (табл. 9).

Таблица 9. Грамматические таблицы «ЛемКС-Р» для окончания *-ам*

Грамматический класс	**Грамматический тип**	**Грамматическая форма**
ж, жо	1	Д. мн.ч.
ж, жо	3	Д. мн.ч.
ж, жо	4	Д. мн.ч.
ж, жо	5	Д. мн.ч.
ж, жо	8	Д. мн.ч.
м, мо	1	Д. мн.ч.
м, мо	3	Д. мн.ч.
м, мо	4	Д. мн.ч.
м, мо	5	Д. мн.ч.
с, со	1	Д. мн.ч.
с, со	3	Д. мн.ч.
с, со	4	Д. мн.ч.
с, со	5	Д. мн.ч.
с, со	8	Д. мн.ч.

Такая форма представления — хотя она и позволяет системе функционировать — представляется весьма неэкономной.

Другая проблема, возникавшая в процессе работы над программой, была связана с грамматической омонимией. Вообще, возможность получения нескольких лемм для одной словоформы (***три*** → *три*, *тереть*; ***весел*** → *весло*, *веселый*; ***потом*** → *пото́м*, *пот*, и т.п.), без сомнения, можно назвать главной проблемой морфологического анализа. Частично эта омонимия все же может устраняться путем отбрасывания проблематичных форм. Например, для слова *Федор* возможны две леммы — *Фёдор* (им. п.,

ед. ч.) и *Федо́ра* (р. п., мн. ч.). Однако, поскольку имена собственные в формах множественного числа употребляются редко, вторую лемму с высокой долей вероятности можно отбросить. К сожалению, возможностей для такой «фильтрации» лемм оказывается не так много. Дело в том, что в словаре Зализняка зачастую допускается образование грамматических форм, которые реально практически не употребляются. Так, считается, что практически любое прилагательное, за редкими исключениями, может образовывать степени сравнения. Вообще говоря, это действительно так, поскольку окказионально степени сравнения образуются не только для качественных прилагательных (например, форма *колбаснее*, которую в одной из своих статей употребил А.И. Солженицын). Однако подобная гибкость в нашем случае не намного повышает распознаваемость форм системой, поскольку такие формы все-таки очень редки.

Допущение образования сравнительных форм не только от качественных прилагательных само себе не создает большой проблемы. Однако оно имеет некоторые неприятные косвенные последствия. Наличие пометы о наличии сравнительных форм является для системы сигналом того, что для данного слова возможно образование кратких форм (хотя в некоторых случаях краткие формы могут образовываться и у прилагательных, не имеющих степеней сравнения, например, *слепой, зрячий*). А это уже приводит к появлению очень большого количества «паразитических» лемм, например, для слова *домов*, кроме *дом*, система предлагает в качестве леммы также *домовый*, для словоформы *классов — класс* и *классовый*. В итоге пришлось пожертвовать «предсказательной» силой словаря и добавить во многие статьи для прилагательных помету «степени сравнения не образуются». Аналогичная ситуация возникла с существительными singularia tantum. В «Грамматическом словаре» отсутствие форм множественного числа также зачастую специально не обозначается, что потенциально может вызвать появление дополнительного «шума», правда в меньшей степени, чем для степеней сравнения прилагательных.

Фонология, морфология и словообразование

При работе над анализатором, как уже было сказано выше, бросается в глаза появление грамматических типов, различающихся только одной-двумя формами. Это совершенно неизбежно, если не выходить за рамки морфологии, а слова рассматривать, как последовательности символов. Если же привлекать информацию фонологического и морфонологического уровня, многие из грамматических типов сольются в один. Например, все подтипы субстантивного склонения с шипящими и *к, г, х* превратятся в частный случай основного типа с твердым согласным. Мягкие варианты склонения также можно превратить в частные случаи основного типа,

введя компонент *j*, который будет располагаться на конце основы в существительных, изменяющихся по мягкому типу.

Интегрировав в систему морфоанализа фонологию, можно уменьшить количество типов и решить проблему морфонологических чередований типа *к-ч, г-ж*, беглых гласных, «вставного *л*» (*любить — люблю*) и т.п. Это позволит в конечном итоге добиться того, что у всех лексем будет по одной основе, что опять же уменьшит объем словаря.

Многие из упомянутых выше проблем решает предложенная финским лингвистом К. Коскенниеми двухуровневая модель (Koskenniemi, 1995). Суть модели заключается в том, что единицы поверхностного представления обобщаются на втором, более абстрактном уровне, на котором нейтрализуются всевозможные виды варьирования, чередований, выражений в виде нуля и т.п. Эта идея перекликается и с порождающей грамматикой и с порождающей семантикой, но для разработки систем автоматического морфоанализа эти идеи, насколько нам известно, до сих пор не применялись. Использование двухуровневого представления позволяет существенно уменьшить количество грамматических типов и сделать алгоритм анализа более изящным. Формализм Коскенниеми был успешно применен для построения системы морфологического анализа/синтеза для финского языка, а в настоящее время делаются попытки использовать его для построения аналогичных систем для других языков с развитой морфологией, в том числе и для русского языка (см. например, Koskenniemi 1983).

Если фонология позволит упростить описание и процедуру анализа, то вышестоящий уровень словообразования таит в себе огромные резервы для повышения предсказательной силы системы и расширения множества анализируемых ей слов.

Грань между морфологией и словообразованием весьма зыбка. Для многих языков, например, для языков с агглютинативным строем, вообще очень трудно сказать, где кончается морфология и начинается словообразование. Традиционно к морфологии относят регулярные изменения, происходящие в словах, а к словообразованию — то, что выходит за рамки системы, что нельзя выразить в виде грамматических таблиц.

> «Граница между морфологией и словообразованием проходит как граница между окончаниями и другими типами морфем, как граница между значениями, появление которых в словоформах обязательно и регулярно, и значениями, которые этими свойствами не обладают.» (Белошапкова и др. 1999: 454).

Однако при построении систем автоматического морфологического анализа становится вполне очевидно, что система, базирующаяся на одной только морфологии, оказывается в некоторой степени ущербной. Так, в русском языке очень богатое глагольное словообразование. Приведем в качестве примера глаголы, производные от глагола *помнить*, зарегистрированные в «ТамРус» (табл. 10). Отметим, что это не самое продук-

тивное гнездо, для глаголов *идти, лететь, сидеть* производных слов гораздо больше.

Другой «нагруженный» участок русского словообразования — уменьшительные, увеличительные, уменьшительно-ласкательные и тому подобные суффиксы (*кот, котик, котишка, котище, котяра, котенька, коток,* и др.).

Таблица 10. Дериваты глагола *помнить*, зафиксированные в корпусе текстов русской литературы XX века

помнить
вспомнить
вспомниться
запомнить
запомниться
напомнить
опомниться
помниться
попомнить
припомнить
припомниться
упомнить

Автоматический морфемный анализ как компонент системы морфологического анализа был бы чрезвычайно полезен. Тем не менее, словообразование чрезвычайно тяжело поддается формализации. В отличие от окончаний и формообразующих суффиксов, словообразовательные морфемы имеют более конкретную семантику, деривационные модели охватывают ограниченные множества лексем. Например, агентивный суффикс может присоединяться только к агентивным глаголам, например, рус. *играть → игрок*; *читать → читатель, чтец*; *учить → учитель*; фин. *lukea → lukija*, *ajatella → ajattelija*.

Обратим также внимание, что в русском языке для передачи близких значений нередко используются несколько деривационных моделей (кроме приведенного примера с агентивными суффиксами вспомним хотя бы великое множество суффиксов со значением ’житель города / страны’: *Москва → москвич*, *Тула → туляк*, *Смоленск → смолянин*, *Новгород → новгородец*; *Англия → англичанин*, *Франция → француз*, *Италия → итальянец*, *Финляндия → финн*).

Таким образом, для корректного образования новых слов, кроме грамматической информации, требуется хотя бы минимальные данные о семантике исходного слова и семантике словообразовательной морфемы.

В процессе работы над лемматизатором мы попытались построить алгоритм обработки префиксальных дериватов глаголов и прилагательных[23]. В целом следует отметить, что обработка префиксов — проблема, ожи-

[23] Идея предложена Г.О. Сидоровым (Сидоров, 1993).

дающая своего решения. Чисто формальное выделение префиксов по начальной подстроке слова (например, *раз*бить, *с*вить) малоперспективно. Например, слово *нему* может быть при таком подходе проанализировано как существительное среднего рода (*не* + *му*). Задача обработки префиксальных дериватов существительных не ставилась по той причине, что префиксация русских существительных не так развита, как префиксация глаголов и прилагательных. Кроме того, именное словообразование гораздо более непредсказуемое. Например, есть слова *подъезд*, *подход*, *подныр*, но существительное **подплыв* возможно только как окказиональное. Префиксация существительных также в большей степени зависит от семантики корня. Достаточно понаблюдать, как образуются существительные с помощью префикса *не*: *недруг, *невраг, неметалл, *нестекло, непогода, *недождь, *недверь, *некнига*[24].

Обработка слов с префиксами организована в системе следующим образом. Часть приставочных дериватов глаголов и прилагательных не хранится непосредственно в словаре основ, а опознается системой путем отсечения гипотетических префиксов. Например, при анализе формы *переселю* система найдет в начале слова префикс *пере*, отсечет его, получит *селю* для которого будет найдено две леммы — *селить* и *сель* — вторая из которых будет отброшена, поскольку анализ префиксальных существительных не предусмотрен, к первой лемме будет приставлен обратно префикс и получена правильная лемма *переселить*.

Программа учитывает, что в русском языке существуют определенные правила присоединения префикса к основе. В процессе работы над алгоритмом были выделены следующие группы префиксов[25]:

1. присоединяющиеся к любым основам без каких-либо формальных ограничений (*не-, на-, псевдо-, у-*);
2. присоединяющиеся к основам, начинающимся на гласную или звонкую согласную (*воз-, под-*);
3. присоединяющиеся к основам, начинающимся на глухую согласную (*вс-, ис-*);
4. присоединяющиеся к основам, начинающимся на две согласные (*обо-, ото-*);
5. присоединяющиеся к основам, начинающимся на *е*, *я*, *ю* (*отъ-, разъ-*);
6. присоединяющиеся к основам, начинающимся на *о* (*со-, во-*).

[24] Тем не менее, нельзя не отметить, что возможности окказионального использования префикса *не* для образования новых существительных все-таки довольно большие, достаточно вспомнить существительное *нелицо,* употребленное в переводе романа Дж. Оруэлла «1984».

[25] Разумеется, здесь речь идет строго говоря не о префиксах, а о псевдопрефиксах: морфы *вос-, вс-, вз-, воз-, взо- взъ-*, которые логично считать алломорфами одной морфемы, рассматриваются нами как разные префиксы.

Некоторые префиксы относятся сразу к нескольким типам: так префиксы *в-* и *пред-* относятся к типам 2 и 3 (*вгонять, вкопать, предупредить, предстать*), другие — только к одному типу: префикс *раз-* относится к типу 2 (*разбить*), *изъ-* — к типу 5 (*изъять, изъесть*). Возможны и исключения, например, префикс *обо-* в большинстве случаев присоединяется к основам, начинающимся на две согласные (*оборвать, оболгать, обойти, обобрать* и др.), но встречаются и формы типа *обошёл,* в которых это правило не выполняется.

В некоторых случаях в разных формах одной и той же лексемы префикс может претерпевать изменения, если основа изменяется, например, *подлить — подолью*. По этой причине, при присоединении префикса обратно к лемме программа проверяет, не происходит ли изменений в основе, и если таковые имели место, подбирает подходящий вариант префикса, чтобы не получались аномальные формы типа **подоливать*.

Проблемы возникают и с определением вида у производных глаголов. Как известно, глаголы, образованные префиксальным способом, в подавляющем большинстве случаев — глаголы совершенного вида, например, *читать — прочитать, вычитать, перечитать*. Однако можно привести довольно много примеров глаголов несовершенного вида с префиксами: *почитывать, перебирать, приводить, совращать, завышать* и т.д. Следует отметить, что эти глаголы нельзя назвать исключениями из правила, поскольку они не образованы префиксальным способом. Эти глаголы образуются от префиксальных глаголов совершенного вида путем добавления специальных суффиксов, имеющих семантику длительности или повторяемости, что и приводит к имперфективации (*читать* → *прочитать* → *прочитывать*). Однако наша система работает только с глагольными (псевдо)основами и не проверяет, есть ли у анализируемых глаголов суффиксы (это сделало бы алгоритм анализа слишком сложным и сильно замедлило бы работу программы). Таким образом, любой глагол, у которого можно найти нечто похожее на префикс, считается глаголом совершенного вида. Чтобы указанным выше глаголам не присваивался совершенный вид, для псевдооснов *—читыва-, -бира- /-бер-, -води-/-вож-/-вод-, -враща-, -выша-/-выш-* и др. в зоне помет специально указывается, что глаголы с этими основами всегда несовершенного вида. Если бы система выполняла морфемный анализ, такой проблемы, скорее всего, не возникло бы.

Поскольку при анализе проверяется лишь формальная сочетаемость префикса и не учитывается семантическая сочетаемость, в некоторых случаях программа ошибается и находит несуществующие леммы. Так для формы *разошлись* программа предлагает две леммы: *разойтись* и *разослаться*, для *закрою — закрыть* и *закроить*. Ранние варианты программы находили очень большое количество «паразитических» лемм. Избавиться от этого шума удалось только введя правило, по которому леммы, полученные с применением правил отсечения префикса, отбрасы-

ваются, если есть варианты лемм, полученные без применения этих правил[26].

Таким образом, в программе заложена потенциальная возможность «принятия» неприемлемых форм, полученных путем отсечения префиксов или *ся*. Однако все же думается, что вероятность ошибки здесь не очень большая, поскольку программа используется для обработки правильных текстов, она не предназначена для поиска ошибок и их исправления[27]. Как уже говорилось выше, программа на своей нынешней стадии разработки не предназначена для синтеза словоформ, поэтому «всеядность» программы в данном случае нам не сильно вредит. С другой стороны, при лемматизации текста всегда возможно появление слов, не зафиксированных в словаре. Чем лучше лемматизатор справляется с такими незарегистрированными словами, тем выше эффективность работы программы. Один из потенциальных путей появления новых слов — словообразование. Поэтому некоторая «предсказательная сила» лемматизатора будет чаще приводить к успешной обработке не зафиксированного в словаре слова, чем к появлению «паразитического» варианта анализа.

3.3.2. Лемматизация в «КОКОС-П»: финский язык

Принципы работы программы «ЛемКС-Ф»

При разработке финского лемматизатора «ЛемКС-Ф» мы в целом следовали тем же принципам, что и при разработке «ЛемКС-Р». Целью было получить эффективно и быстро работающую программу, результаты работы которой можно было бы редактировать. Основной проблемой, приводящей к сильному снижению КПД программы, является грамматическая омонимия, которая проявляется по-разному в разных языках, но присутствует в любом из них. Разрешение омонимии лежит за пределами контекстно-свободного морфоанализа.

Так же как и в «ЛемКС-Р», основным компонентом «ЛемКС-Ф» является грамматический словарь (ГС), основным источником для которого послужил толковый словарь финского языка «Suomen kielen perussanakirja». В базе данных грамматического словаря хранится следующая

[26] Это, разумеется, приводит к тому, что, кроме неправильных лемм, в некоторых случаях могут быть отброшены и правильные.

[27] Чрезмерное увлечение префиксами и префиксальными компонентами, а также композитами при разработке спеллеров нередко приводит к тому, что программа начинает предлагать в качестве альтернатив несуществующие слова. Так, один из ранних вариантов русского орфографического корректора «Пропись» предлагал исправить *бультерьер* на *буйтерьер*.

информация: лемма, «псевдолемма», грамматический класс, информация о чередовании согласных.

Слова с нестандартным словоизменением: местоимения, глагол *olla* («быть») и т.п. — хранятся в отдельной таблице слов с нестандартными парадигмами (ТНС).

Важной частью системы являются две грамматические таблицы: в первой — таблице суффиксов (ТС) — хранятся псевдосуффиксы, а в другой — в таблице порождения лемм (ТЛ) — правила получения псевдолеммы по концу основы.

Финский язык — агглютинативный язык, поэтому алгоритм морфологического анализа для этого языка нельзя строить абсолютно так же, как для русского языка — классического флективного языка. Если в русском слове вся грамматическая информация, как правило, заключена в одной морфеме — в флексии (например, окончание *—е* в словоформе *доме* сообщает и о единственном числе, и о предложном падеже), то грамматическая информация в финском слове распределяется между несколькими суффиксами (например, в словоформе *taloissaan* — ‘в их (своих) домах’, суффикс *—i* сообщает о множественном числе, суффикс *—ssa* — о падежной форме инессива, притяжательный суффикс *—an* — о третьем лице). Лемматизатор русского языка находит окончание, отсекает его и получает основу, от которой образует начальную форму — лемму. Лемматизатор финского языка должен отсекать все формообразующие суффиксы до тех пор, пока не останется основа слова, от которой образуется «псевдолемма».

Главную проблему при выполнении финского морфологического анализа представляют чередования согласных (*pp/p, tt/t, kk/k, t/d, k/v, k/∅* и др.), которые очень трудно учитывать, если отсутствует информация о слогоделении. Кроме того, для некоторых чередований довольно много исключений, например, в слове *outo* (‘странный’) есть чередование *t/d* (*outo* → *oudon*), а в очень похожем на него слове *auto* (‘машина’) этого чередования нет (*auto* → *auton*), поскольку это относительно недавнее заимствование.

В первоначальном варианте программы делались попытки анализа чередований с использованием формальных признаков. Например, если перед окончанием слова находится согласный *k*, а данная грамматическая форма используется со слабой ступенью чередования, то *k* заменяется на *kk*. Однако для того, чтобы алгоритм работал корректно, нужно правильное слогоделение, а также сведения об исключениях. Программа работала медленно, кроме того довольно часто допускала ошибки. Хорошим решением проблемы был бы подход К. Коскенниеми, но в этом случае потребовалась бы достаточно долгая и трудоемкая работа по созданию словарного массива, задача для одного человека непосильная.

Для системы «ЛемКС-Ф» предлагается следующее решение проблемы. В словаре было добавлено поле псевдолемм. У слов без чередований псевдолемма совпадает с леммой. Для слов же с чередованиями были сгенерированы дополнительные словарные записи, где в поле псевдолемм

заносилась форма с другой ступенью чередования. Например, для слова *katto* существует две словарные записи, в одной из которых в поле псевдолемм записано *katto*, в другой — *kato*. В грамматическом словаре фиксируется также, что в данном слове есть чередование, а также ступень чередования для данной псевдолеммы. После порождения дополнительных записей были проверены чередования, в которых имеются исключения.

При выполнении лемматизации программа генерирует не собственно готовые леммы, а псевдолеммы — фиктивные леммы, которые образуются для заданной словоформы без учета чередований согласных, а также других грамматических особенностей (например, отсутствие форм единственного числа). Затем выполняется поиск псевдолеммы по грамматическому словарю, где проверяется грамматическая информация по слову и где хранится готовая «настоящая» лемма. Например, для словоформы *puvussa* (‘в костюме’) программа, используя грамматические таблицы, породит псевдолемму *puvu*. При порождении псевдолеммы чередование *k/v* не учитывается, правильная лемма — *puku* — хранится в грамматическом словаре в зоне леммы. Аналогичным образом была решена и проблема анализа *pluralia tantum*, причастий и степеней сравнения прилагательных (об этом см. ниже).

Еще одно отличие «ЛемКС-Ф» от «ЛемКС-Р» состоит в том, что в грамматическом словаре хранятся не основы, а именно начальные формы слов. Дело в том, что у финских слов чаще, чем в русском языке бывает несколько вариантов основы, поэтому проще хранить готовую лемму, которая образуется по специальным правилам.

Таким образом, алгоритм работы «ЛемКС-Ф» в общих чертах выглядит следующим образом:

1. поиск анализируемой словоформы в таблице нестандартных словоформ;
2. поиск анализируемой словоформы в грамматическом словаре;
3. проверка на наличие модальных частиц (*—kin/—kaan/—kään, —han/—hän, —pa/—pä* и т.п.) на конце слова; если одна из подстрок такого вида обнаружена, она отсекается;
4. проверка на наличие притяжательных суффиксов (*—ni, —si, —nsa/—nsä, —an/—än*); в случае обнаружения суффикс отсекается;
5. проверка на наличие формообразующих суффиксов; при обнаружении отсекаются;
6. из полученной основы по правилам ТЛ порождается псевдолемма;
7. псевдолемма проверяется по грамматическому словарю; в случае соответствия информации грамматического словаря грамматической информации о словоформе, полученной в ходе анализа, словоформе присваивается лемма из таблицы ГС;
8. возврат к 3.

Анализ выполняется до тех пор, пока не будут обнаружены все возможные леммы для заданной словоформы. Тексты основных модулей лемматизатора даны в приложении 4.5.

Так же, как и русский лемматизатор, «ЛемКС-Ф» сохраняет результаты анализа в трех таблицах:

- в лемматизированном словнике, в который записываются однозначно опознанные леммы;
- в таблице омонимичных форм, для которых система нашла более одной леммы;
- в таблице неопознанных словоформ.

Записи из двух последних таблиц можно просматривать и вручную указывать для них леммы, после чего обновляется лемматизированный словник.

Качество работы лемматизатора было проверено на финском субкорпусе «ПарРус» и оценено как удовлетворительное: из 180 000 словоформ словника однозначно были распознаны 118 000, что составляет 65,56% от всего словника. Количество омонимичных форм составило 11 000, и это намного больше, чем при работе русского лемматизатора. Неопознанных словоформ 51 000. Многие из омонимичных словоформ очень высокочастотны, причем предлагаемые леммы оказываются равновероятными. Например, форма *häntä*, которая может быть и номинативом существительного *häntä* ‘хвост’, и партитивом местоимения *hän* ‘он/она’, в нашем корпусе встретилась более 3 000 раз. Это означает, что для разрешения омонимии пользователь должен просмотреть все 3 000 контекста, в которых встретилась эта словоформа.

Грамматический словарь

Грамматический словарь представляет собой список лексем финского языка с указанием грамматического класса по «Perussanakirja» и наличия чередования согласных. Объем базы данных — около 160 000 записей.

Если грамматический класс не указан — слово неизменяемое.

Для слов, у которых наблюдается чередование, в базе данных заведены две словарные статьи: с псевдолеммой для сильной и слабой ступени чередования. Например, для глагола *ajatella* ‘думать’ в базе данных две записи; в первой из них в зоне псевдолемм будет правильная лемма *ajatella*, во второй — фиктивная *ajattella*, кроме того, в специальных зонах указан тип чередования и ступень — сильная (*v*) и слабая (*h*). Дополнительные словарные статьи были сгенерированы в автоматическом режиме, после чего были проверены вручную.

Таблица 11. Фрагмент грамматического словаря «ЛемКс-Ф»

Pseudolemma	**Lemma**	**Grammar**	**PTK**	**Vaihtelu**
itsehillinnä	itsehillintä	9	XX	h
itsehillintä	itsehillintä	9	nt/nn	v
itsehoido	itsehoito	1	XX	h
itsehoito	itsehoito	1	t/d	v
itseihailu	itseihailu	2		
itseilmaisu	itseilmaisu	2		
itseinduktio	itseinduktio	3		
itseinho	itseinho	1		
itseironia	itseironia	12		

По сравнению с русским словарем основ, в финском ГС содержится гораздо меньше информации. Это связано с тем, что в морфологии финского языка заметно меньше исключений и нерегулярных образований, чем в русской морфологии. Указания грамматического класса в подавляющем большинстве случаев достаточно для того, чтобы полностью задать парадигму слова.

Проблема слов *pluralia tantum* в настоящей версии решается путем указания в зоне леммы формы номинатива множественного числа. Специальной пометы в словаре не используется. Здесь мы исходили из того, что система ориентирована на анализ, а не на синтез. Единственная проблема — возможное совпадение несуществующих форм единственного числа с реально существующими формами других слов. Так, для словоформы *häitä* (слово *häät* ‘свадьба’ в партитиве множественного числа) система предложила две леммы: *häät* ‘свадьба’ и *häkä* ‘гарь’, хотя форма множественного числа для второго слова возможна лишь теоретически.

Аналогично в ГС нет специальной пометы для слов singularia tantum. С одной стороны, неуказание на отсутствие форм множественного числа ведет к некоторому риску появления «шума», с другой стороны это повышает гибкость системы, которая сможет анализировать слова singularia tantum, окказионально употребленные в форме множественного числа.

Грамматические таблицы

Часть грамматической информации, например, списки модальных частиц и притяжательных суффиксов, хранится непосредственно в кодах программы, поскольку эти списки короткие. Более сложная информация хранится в таблице суффиксов. В этой таблице хранятся как собственно суффиксы, так и распространенные финали для слов в определенных грамматических формах. Например, сложно сказать, является ли финальные *o* и *e* в глагольных формах 3-го лица ед. ч. *sano<u>o</u>* ‘говорит’, *lähte<u>e</u>* ‘уезжает’, *mene<u>e</u>* ‘идет’ суффиксами 3-го лица. Однако после отсечения этих ко-

нечных гласных легко образовать начальные формы названных слов, поэтому мы пошли на такое упрощение. В тех случаях, когда грамматическое значение никак не выражено, в таблице стоит специальный знак ∅, который означает, что ничего отсекать не нужно. В таблице 12 приведен фрагмент ТС. В поле «Suffix» указаны псевдосуффиксы трех локативных падежей — адессива, аллатива и аблатива. Названия падежей даны в поле «Description». В поле «Stem» указаны типы основ, от которых могут быть образованы данные формы, в данном случае — основа генитива и основа множественного числа. В поле «PartOfSpeech» указана часть речи (помета subst., приведенная во фрагменте из базы данных, означает, что это формы существительных и прилагательных). Поле «Condition» содержит дополниельный формальный критерий для проверки. Помета «V+e» означает, что перед суффиксом обязательно должна быть гласная, в том числе — гласная *e*. Если проверка показывает, что это не так, этот вариант разбора отбрасывается.

Таблица 12. Фрагмент таблицы суффиксов (ТС)

ID	Suffix	Description	Stem	PartOfSpeech	Condition
8	lla	adess.	gen/mon	subst.	V+e
9	llä	adess.	gen/mon	subst.	V+e
12	lle	allat.	gen/mon	subst.	V+e
10	lta	ablat.	gen/mon	subst.	V+e
11	ltä	ablat.	gen/mon	subst.	V+e

Из дополнительных полей ТС поле «Stem» является самым важным: в нем фактически содержится отсылка к таблице образования лемм (ТЛ).

ТЛ состоит из наборов правил образования лемм по финалям основ. Предположим, что программа анализирует словоформу *keväällä* (‘весной’). Используя информацию из ТС, программа найдет на конце словоформы суффикс адессива —*llä,* проверит букву перед суффиксом, установит, что это гласная, после чего перейдет к ТЛ и проверит полученную основу *kevää* по правилам для основ генитива и множественного числа. В числе разных правил программа найдет и правило ää → ät (см. таблицу 13), после применения которого будет получена лемма *kevät*. В дополнительном поле таблицы (GT) указано, что слово должно относиться к грамматическому классу 44. Далее программа обращается к ГС, где находит слово *kevät*, которое действительно относится к классу 44. Лемма найдена.

Таблица 13. Фрагмент таблицы образования лемм (ТЛ)

ID	Ending	LemmaEnding	Description	PartOfSpeech	GT	Comment
46	he	s	gen	subst.	/42/	mies
47	aa	at	gen	subst.	/44/	kevät
48	ää	ät	gen	subst.	/44/	kevät
49	e		gen	subst.	/32/8/49/	kuningata
50	ame	an	gen	subst.	/33/	morsian
51	i		gen	subst.	/5/	rock

В тех случаях, когда программа не находит никаких правил образования леммы, это может означать, что у слов этого класса начальная форма равна основе. Например, при анализе словоформы *pöydälle* ‘на стол’, программа найдет в ТС *—lle*, суффикс аллатива (см. табл. 12), однако перейдя к ТЛ, не найдет никаких правил образования леммы для основы *pöydä*. Тогда программа обратится непосредственно к ГС, где эта псевдолемма будет найдена с указанием, что в этом слове есть чередование согласных «t/d» и псевдолемма получена для слабой ступени чередования. Поскольку для формы аллатива у существительных на согласный в основе наблюдается слабая ступень чередования, то форма является правильной и программа возвращает лемму *pöytä*, которая хранится в ГС в зоне лемм для этой псевдолеммы.

Причастия и степени сравнения прилагательных

Отдельной проблемой оказалась обработка причастий, а также сравнительной и превосходной степеней прилагательных. У этих форм есть свое собственное словоизменение. Таким образом, если считать леммой причастия инфинитив глагола, от которого оно образовано, то поиск леммы для причастия должен происходить в два этапа. Например, для словоформы *sanoneen* (генитив единственного числа действительного причастия прошедшего времени от глагола *sanoa* ‘говорить’) сначала должен быть получен номинатив причастия — *sanonut*, после чего от него образуется инфинитив глагола — *sanoa*.

Аналогичная проблема возникает и со сравнительной и превосходной степенями финских прилагательных. Обе формы имеют в финском языке полные парадигмы.

Первоначально проблема решалась путем выполнения двухступенчатого анализа, с использованием рекурсии. Функция морфологического анализа, породив в качестве псевдолеммы форму причастия, сравнительной или превосходной степени прилагательного, вызывала сама себя с полученной псевдолеммой в качестве аргумента.

Поскольку финали многих форм косвенных падежей причастий, сравнительных и превосходных степеней прилагательных совпадают с формами обычных существительных и прилагательных, подобная рекурсия применялась очень часто, что сильно замедляло работу программы.

Подобная проблема возникала и в русском лемматизаторе при анализе русских причастий и была решена путем добавления новых записей с основами причастий в зоне основ и инфинитивами глаголов в зоне лемм. Аналогичное решение было применено и для «ЛемКС-Ф». В базу данных словаря были в автоматическом режиме добавлены дополнительные записи с причастиями и сравнительными степенями прилагательных в поле псевдолемм. Таким образом, при анализе слова достаточно получить номинатив причастия, сравнительной или превосходной степени, лемму — глагол или прилагательное — программа берет в готовом виде из поля «Лемма».

Автоматическое порождение всех форм причастий для финских глаголов большой проблемы не составило, поскольку правила образования причастий от инфинитива глагола в финском языке очень ясные и четкие. В отличие от русского языка, в котором не все глаголы имеют полный набор причастий, в финском языке действительные и страдательные причастия образуются от любых глаголов (приведем в качестве примера причастия от глагола *olla* ’быть’ — *oleva, oltava, ollut, oltu*). Однако агентивные причастия (*soittama, piirtämä*) образуются только от переходных глаголов. Поскольку информация о переходности глаголов в нашей базе данных отсутствовала, агентивные причастия были образованы для всех глаголов. Хотя в результате в базу данных попадают несуществующие формы, это не создает проблем при работе программы, поскольку она выполняет анализ, а не синтез.

Значительно сложнее было решить проблему степеней сравнения прилагательных. В финской грамматике у прилагательных нет отдельных словоизменительных классов, поэтому с точки зрения морфологии прилагательные отличаются лишь наличием степеней сравнения. По этой причине программа отличает прилагательные от существительных лишь в тех случаях, когда анализируются прилагательные в форме сравнительной или превосходной степени. Таким образом, локальная проблема русского лемматизатора, возникающая с существительными, изменяющимися по адъективному склонению (*рабочий, столовая*), в финском лемматизаторе становится глобальной. Для того, чтобы программа работала корректно и в этой части, потребовалось ввести частеречную помету и отметить существительные и прилагательные (а в будущем планируется отметить и слова неизменяемых частей речи).

Разметка существительных и прилагательных выполнялась вручную, поскольку прилагательные есть практически во всех именных словоизменительных классах. Отдельная помета применялась для прилагательных, имеющих степени сравнения. После завершения разметки для прилагатель-

ных с соответствующей пометой были в автоматическом режиме порождены формы степеней сравнения.

В результате объем базы данных довольно сильно вырос, однако алгоритм программы стал более прозрачным, а скорость работы увеличилась.

Нерешенные проблемы

В числе проблем, возникающих при работе с «ЛемКС-Ф», часть связана с самой идеологией системы. Решение таких проблем лежит за пределами разработки самой системы и требует привлечения внешних средств. Практически все эти проблемы являются общими и для «ЛемКС-Ф», и для «ЛемКС-Р». Обсуждению этих вопросов будет посвящен раздел 3.3.3.

Однако целый ряд проблем связан лишь с нехваткой данных у системы и может быть легко решен путем пополнения системы новыми данными. Назовем наиболее важные из них.

1. «Словоизменение» неизменяемых слов

В финском языке довольно типична «кластеризация» неизменяемых слов, напоминающая словоизменение. Например, у послелога *luo* (’около, у’) есть варианты *luokse* (’по направлению к’) и *luota* (’по направлению от’), существуют и аналогичные группы наречий *loitolla* ’далеко’, *loitolle* ’вдаль’, *loitolta* ’издалека’. Однако, все эти слова хранятся в отдельных записях, хотя нередко все варианты исчислить сложно, поскольку и к послелогам, и к наречиям могут добавляться притяжательные суффиксы.

Некоторые трудности возникают и с обработкой наречий. Большинство финских наречий образуют степени сравнения, которые в большинстве случаев довольно регулярны (*nopeasti, nopeammin, nopeimmin; helposti, helpommin, helpoimmin* и т.п.). Кроме того, сами наречия нередко образуются от прилагательных (*nopea* → *nopeasti*, *helppo* → *helposti*). Таким образом, многие варианты и производные неизменяемых слов тоже можно анализировать. Для решения этой задачи требуется некоторая дополнительная разметка базы данных.

2. Сложные слова

Проблему анализа сложных слов можно игнорировать применительно к русскому морфоанализу, однако для финского языка она становится даже более острой, чем для немецкого языка, известного большим количеством сложных слов. В финском языке сложные слова необычайно широко распространены, многие образуются окказионально, поэтому в словарях регистрируются только наиболее распространенные композиты типа *rautatieasema* (*rauta* ‘железо’ + *tie* ‘дорога’ + *asema* ‘станция’ = ‘железнодорожный вокзал’) или *osakeyhtiö* (*osake* ‘акция’ + *yhtiö* ‘фирма’ = ‘акционерное общество’), а также сложные слова, значение которых плохо мотивируется значением составляющих корней (например, *revontulet — repo*

‘лиса’ + *tuli* ‘огонь’ = ‘северное сияние’). У многих сложных слов склоняется только последняя часть, однако есть слова, у которых изменяются все части, причем количество частей может быть более двух.

Финские сложные количественные и порядковые числительные также представляют собой сложные слова, причем как у количественных, так и у порядковых числительных склоняются все части (напр. *kolmellatuhannellaviidelläsadallakahdeksallakymmennelläseitsemällä* — ’3587’ в форме адессива). Анализ сложных числительных в нашем лемматизаторе был оставлен без внимания, поскольку в письменных текстах запись словами числительных, обозначающих количество бóльшее 20, встречается довольно редко.

Качество морфологического анализа удалось бы существенно повысить, если бы анализатор распознавал составляющие сложного слова: в этом случае процент неопознанных слов удалось бы существенно снизить.

В связи с вопросом лемматизации сложных слов возникает также другая фундаментальная проблема. Что считать леммой сложного слова: само сложное слово в начальной форме, все его составляющие в начальных формах или ключевое слово? Например, для словоформы *juhlakonserttilla* ‘на праздничном концерте’ в качестве лемм можно было бы предложить: а) *juhlakonsertti*, б) две отдельные леммы *juhla* и *konsertti*, в) *konsertti*. Мы склоняемся к тому, что оптимальным решением была бы комбинация вариантов а) и б) и для каждого сложного слова в качестве леммы предлагалась бы начальная форма всего композита и начальные формы каждой из его составляющих. Это сделало бы работу системы достаточно гибкой.

4. Сингармонизм

Правила гармонии гласных в финском языке внешне выглядят очень просто. Если в начальном слоге слова встречается один из гласных ряда *a-o-u* или дифтонг, в который входит один из названных гласных, то в последующих слогах могут быть только гласные этого же ряда, гласные *i* и *e*, а также дифтонги, включающие эти гласные, но не гласные ряда *ä-ö-y* и дифтонги с гласными этого ряда. Например, *katto, puoli,* **kadullä*. Аналогично, если в первом слоге — гласные ряда *ä-ö-y*, то в последующих слогах могут быть только гласные этого же ряда, гласные *i* и *e*, а также дифтонги, включающие эти гласные, но не гласные ряда *a-o-u* и дифтонги с гласными этого ряда. Например, *yöllä, kävellä,* **käyda*. После гласных *i*, *e* и дифтонгов *ie* и *ei* могут появляться гласные обоих рядов, например, *tieto, miehellä* (см., например, White 1993: 12). Этот последний случай, конечно, представляет некоторую проблему при автоматическом анализе, но основную трудность представляет то, что правила сингармонизма распространяются на самом деле не на слово, а на корень и относящиеся к нему суффиксы. Таким образом, в одном сложном слове в разных его корнях могут встречаться гласные обоих рядов. Например, *syysloma, kevätlukukausi*.

Таким образом, для корректной работы с сингармонизмом система должна уметь разбивать сложные слова на составляющие. Как уже говорилось выше, анализ сложного слова представляет отдельную проблему, для решения которой требуется построение специальных алгоритмов.

В «ЛемКС-Ф» решение проблемы сингармонизма удалось обойти. Суффиксы и финали в грамматических таблицах там, где существует два варианта суффикса или финали, даются в двух вариантах (см. табл. 12 и 13). В некоторых случаях может быть получено два варианта псевдолеммы, один из которых будет несуществующим. Таким образом, в целом нерешенный вопрос о сингармонизме не создает в системе больших проблем и не порождает большого количества ошибок. Однако если бы проблема сингармонизма была бы решена, то программа могла бы, например, находить связанные с сингармонизмом опечатки или ошибки, допущенные при распознавании сканированных текстов (например, *а* вместо *ä*).

3.3.3. Общие проблемы морфологического анализа в программном обеспечении для корпусов текстов и перспективы их решения

Грамматическая омонимия и контекстно-свободный морфоанализ

Основная проблема автоматического морфологического анализа на нынешнем этапе состоит в том, что проблема автоматического морфоанализа с привлечением контекста пока решена далеко не для всех языков.

Для русского языка контекстно-зависимый морфологический анализ развивается в рамках системы машинного перевода «ЭТАП-3». Теоретической базой системы является теория «Смысл ⇔ Текст» (подробнее об этой теории см., например, Мельчук 1999). Модули для автоматического анализа финского языка разрабатываются, в частности, фирмами *Connexor* (система *Machinese Syntax*, см. http://www.connexor.com/products.html, 10.07.2003) и *Kielikone* (http://www.kielikone.fi/, 10.07.2003).

Несмотря на несомненное преимущество контекстно-зависимого анализа, чаще используются все же контекстно-свободные системы (например, все спеллеры *Microsoft Office* являются контекстно-свободными). Даже ограниченный синтаксический анализ используется довольно редко, так называемые «грамматические корректоры» (grammar checkers) в основном ориентируются на чисто формальные признаки — повторение слов в одном предложении, длина предложения, частотность определенных грамматических форм (например, причастий). Некоторые спеллеры (например, одна из первых версий корректора «Орфо») проверяют грамматическое согласо-

вание. Однако, в целом, выход на синтаксический уровень — дело будущего.

Между тем, как явствует из сказанного выше, ограничение анализа только морфологическим уровнем влечет за собой неизбежные проблемы, главной из которых является грамматическая омонимия. Даже самый обычный спеллер в результате может пропускать ошибки, если ошибочная форма совпадает с какой-либо правильной формой или если правильные формы употреблены неправильно, например, **хорошая книгу* или **по дорогу*. Кроме того, спеллеры не могут помочь в тех случаях, когда существует несколько вариантов написания, например, в русском языке в эту сферу попадает слитное и раздельное написание *не*, одна и две *н* в суффиксах прилагательных и причастий и т.п. Таким образом, существующие орфографические корректоры ориентированы прежде всего на носителей языка, а не на пользователей, пишущих на неродном языке.

Разрабатывая лемматизаторы для русского и финского языков, мы постоянно сталкивались в той или иной степени с проблемой грамматической омонимии. В русском языке омонимия представлена весьма богато. Можно выделить следующие случаи омонимии:

- полные лексические омонимы (*брак, лук*);
- слова разных частей речи, одно из которых образовано от другого путем конверсии (*военный, рабочий*);
- слова разных частей речи, совпадающие в некоторых формах (*печь* — инфинитив глагола и именительный падеж существительного, *белил* — форма прошедшего времени мужского рода от глагола *белить* и родительный множественного от существительного *белила*);
- различные формы одного и того же слова (*кости* — родительный, дательный, предложный единственного числа, именительный и винительный множественного числа от существительного *кость*).

В финском языке совпадение различных форм одного и того же слова практически не встречается. Лексическая омонимия (*vaara* — ’гора’ и ’опасность’, *keksi* — ’печенье’ и ’багор’) распространена примерно в той же степени, что и в русском языке. Конверсия встречается, но ее последствия для морфологического анализа ощущаются значительно слабее, поскольку, как уже говорилось выше, парадигмы прилагательных и существительных совпадают, единственное отличие — отсутствие у существительных степеней сравнения. Зато омоформы в финском языке встречаются очень часто.

Кроме единичных совпадений форм, например *sinä* — номинатив местоимения *sinä* ‘ты’ и эссив местоимения *se* ‘оно, это’, встречаются совпадения, носящие регулярный характер. Например, в финском языке есть класс глаголов на *—ta*, которые в форме претерита (imperfekti, прошедшее время) получают в конце слова суффикс *—si* (*vastata* ‘отвечать’ →

vastasi). В свою очередь, на конце существительных тоже может стоять *—si* — притяжательный суффикс 3-го лица. В результате некоторые существительные с притяжательным суффиксом на конце могут совпасть с претеритом глагола, например *vastasi* может интерпретироваться как претерит третьего лица единственного числа от глагола *vastata* ‘отвечать’, и как существительное *vasta* ‘банный веник’ с притяжательным суффиксом 3-го лица; *salpasi* — как претерит третьего лица единственного числа от глагола *salvata* ‘запирать на засов’, и как существительное *salpa* ‘засов’ в номинативе с притяжательным суффиксом 3-го лица.

Глаголы на *—taa* в форме претерита единственного числа первого лица оканчиваются на *—in*, такая же финаль у очень продуктивного класса существительных образованных, как правило, от глаголов. Словоформа *nostin* может быть и формой претерита единственного числа первого лица от глагола *nostaa* ‘поднимать’, и номинатив *nostin* ‘подъемник’, *pakastin* — форма глагола *pakastaa* ‘замораживать’ и форма существительного *pakastin* ‘морозильник’. Списки такого рода грамматических омонимов можно продолжать до бесконечности.

Очевидно, что проблема омонимии и, в первую очередь, грамматической омонимии при выполнении морфологического анализа очень существенна. Доля омонимичных форм по результатам работы «ЛемКС-Р» и «ЛемКС-Ф» составляет для русского языка более 3%[28], для финского языка — более 5% от общего количества словоформ в словниках. И эта проблема грамматической омонимии не разрешима средствами контекстно-свободного морфологического анализа.

Только переход от контекстно-свободного анализа к контекстно-зависимому может частично разрешить проблему грамматической омонимии, что существенно повысит качество анализа. В анализировавшихся нами текстах количество словоформ, для которых определить лемму без привлечения контекста невозможно, довольно велико. Анализ же слова в контексте в большинстве случаев дает однозначную интерпретацию.[29]

Пути разрешения грамматической омонимии

Вопрос о разрешении грамматической омонимии оказывается важным не только в плане повышения качества лемматизации. Ведь только в случае однозначного распознания лексемы и грамматической формы становится возможной грамматическая разметка текстов (tagging).

---

[28] На самом деле, для русского языка процент омонимичных форм более 3%, поскольку ЛемКС-Р не фиксирует как омонимичные словоформы, в которых совпадают падежные формы в пределах одной лексемы (*кости, лыжи, дела* и т.п.).

[29] Кроме, разумеется, случаев, небрежного построения сообщения в устной речи или каламбуров типа *Не прячьте ваши денежки по банкам и углам*

Один из часто предлагаемых путей решения проблемы грамматической омонимии — использование частотности тех или иных лексем. Например, для формы *уже* возможны три интерпретации: частица *уже*, сравнительная форма прилагательного *узкий* или наречия *узко*, предложный падеж существительного *уж*. Однако частица *уже́* намного более частотна и поэтому с высокой степенью вероятности можно присвоить форме *уже* лемму *уже́*. Такой путь вполне приемлем при разработке систем автоматического анализа текста или в системах машинного перевода. Не нанесет он большого вреда и в орфографических и грамматических корректорах (там пользователь все равно контролирует анализ и отказывается от неприемлемых предложений программы). Однако при создании лемматизатора для корпуса текстов такой подход абсолютно неприемлем: полученные по такой методике лемматизированные словники не будут давать ответов на главные вопросы, ради которых создаются корпуса текстов.

Другой путь — переход от контекстно-свободных к контекстно-зависимым анализаторам. Простейшим вариантом такой системы будет анализ ближайшего контекста неоднозначной словоформы. Например, при анализе обсуждавшейся выше формы *уже* даже соседние слова могут помочь разрешить омонимию. Например, микроконтексты типа *уже сейчас*, *уже вчера*, *уже пришел* и т.п. позволяют однозначно определить анализируемую форму как частицу; контекст *уже, чем* с высокой степенью вероятности указывает на сравнительную степень прилагательного[30]. Работа с микроконтекстом в целом поможет разрешить часть омонимичных форм, однако требует ручного ввода огромного количества правил, нередко — для конкретных лексем. Потребуется информация о моделях управления глаголов, о сочетаемости существительных с предлогами и т.п. Причем в некоторых случаях разрешить омонимию на уровне микроконтекста все равно не удается, например, в контексте *Мы говорили об уже...* слово *уже* может быть и существительным в предложном падеже или наречием в зависимости от того, какие словоформы идут после *уже* — *об уже, купленном мной на прошлой неделе* или *об уже начавшейся зиме*.

Для того чтобы приблизить разрешение грамматической омонимии к 100%, в конечном итоге потребуется синтаксический анализатор. Однако в языках со свободным порядком слов, каковыми являются и русский и финский, синтаксический анализ может довольно часто давать сбои, особенно если в тексте встречаются стилистически некорректные или разговорные конструкции.

Оптимальным решением проблемы представляется использование вероятностной модели на уровне микроконтекстов. По такому пути в настоящее время идет разработка лемматизаторов и грамматических разметчиков (taggers) для английского языка. Работа целого ряда «тэг-

[30] Хотя в данном случае и возможны контексты, в которых *уже* функционирует как наречие, например *Он пришел уже, чем меня очень обрадовал*.

геров», наиболее известным из которого является CLAWS, основана на использовании марковских цепей или различных вариантов этого формализма. В качестве исходного источника данных использовался размеченный Брауновский корпус текстов: для каждой двучленной комбинации (существительное — прилагательное, артикль — существительное и т.п.) обнаруженной в корпусе текстов были подсчитаны частоты и вероятность их появления в корпусе. В итоге была получена матрица, которая используется для разрешения неоднозначности. При анализе неоднозначных словоформ «тэггер» вычисляет вероятность появления в данном контексте предлагаемых контекстно-свободным анализатором форм. Из всех вариантов выбирается наиболее вероятный (Oakes 1998: 80–85).

Такой же подход, по-видимому, может использоваться и для разметки русских и финских текстов. Насколько продуктивным он окажется, прогнозировать сложно по целому ряду причин. Как для русского, так и для финского языка разметка по грамматическим формам представляется намного более интересной, чем разметка по частям речи. Алгоритм, о котором шла речь выше, был использован именно для разметки по частям речи, которая является намного более грубой. Причем английский язык — язык с фиксированным порядком слов, а русский и финский — языки со свободным порядком слов. С другой стороны, грамматическая омонимия в английском языке, по-видимому, распространена больше, чем в русском или финском. Таким образом, трудно сказать, насколько эффективным окажется данный подход применительно к русскому и финскому языкам.

Другая проблема: наличие исходных данных для «обучение» программы. Для английского языка существовал размеченный вручную Брауновский корпус, который и был использован для обучения CLAWS (там же). В нашем случае, когда источник данных для «обучение» системы отсутствует, подготовку матрицы можно было бы организовать следующим образом. Вначале контекстно-свободным морфоанализатором обрабатывается небольшой массив текстов, после чего все неоднозначные и неопознанные словоформы определяются вручную. На следующем этапе происходит «обкатка» и «обучение» системы на других экспериментальных массивах. На этом этапе программа работает в интерактивном режиме, сообщая пользователю о своей работе и принятых решениях. После завершения этого этапа должна быть получена работающая программа, способная идентифицировать значительную часть омонимичных форм в автоматическом режиме, оставшуюся часть — в диалоге с пользователем. Программа, возможно, окажется способной даже в некоторых случаях определять грамматическую форму для слов, не опознанных морфоанализатором, если у этих слов есть какие-либо формальные признаки, позволяющие прогнозировать их частеречную принадлежность и грамматическую форму.

Обычно считается, что список лемм, хранящихся в лемматизированном словнике, является списком лексем. Строго говоря, это не так. Ведь лемматизатор определяет начальную форму слова и его часть речи и/или грамматический класс, если при лемматизации выполняется морфологический анализ.

Таким образом, для двух разных лексем, леммы которых совпадают и которые относятся к одному и тому же грамматическому классу, в лемматизированном словнике будет заведена одна запись. Так, для омонимов типа *печь* — существительное и глагол — в лемматизированном словнике будет две записи, а лексические омонимы типа *брак* — ‘некачественное изделие’ и ‘брачный союз’ — различаться не будут. Более того, даже лексемы с разными схемами ударения типа *за́мок* и *замо́к*, *мука́* и *му́ка* при анализе письменного текста не различаются.

Близка к проблеме лексической омонимии и проблема многозначности. Только средствами морфологического анализа различить разные значения слова *нос* или *дом* не представляется возможным.

Для снятия семантической неоднозначности (disambiguation) предлагаются различные алгоритмы. Можно, например, использовать дефиниции слов из толковых словарей: если слова из толкования слова А в значении 1 встречаются в ближнем контексте слова с леммой А, то этому слову присваивается соответствующее значение. Недостаток этого подхода в том, что далеко не во всех случаях в ближайшем окружении слова оказываются слова из дефиниции толкового словаря, а конструирование тезауруса, учитывающего синонимы и слова из тех же семантических групп — долгая и кропотливая работа.

Оригинальное решение предлагает Зерник (Zernik 1991, Oakes 1998: 138–139). Разные значения слов предлагается выделять путем анализа конкорданса на заданное слово. На основе контекста ±5 слов для каждого употребления слова рассчитывается по специальной формуле его вес. В формуле учитываются ожидаемая и наблюдаемая частоты появления слов из ближайшего контекста. После проведения кластерного анализа разные значения слова оказывается «разведенными» по разным кластерам. Метод показал себя довольно эффективным с одними словами (*train*, *rate*), но малопродуктивным с другими словами, например, *office*.

Интересен подход, предлагаемый целым рядом исследователей (Dagan, Itai, Shwall 1991, Gale, Church, Yarowski 1992, Oakes 1998: 186–189), считающих, что «два языка более информативны, чем один» и предлагающих использовать для разрешения семантической неоднозначности параллельные тексты. Правда метод может применяться лишь в том случае, если тексты корпуса состыкованы на уровне слов. Недостаток метода состоит в том, что он не работает, если для слова в разных значениях

используется один и тот же переводной эквивалент, т.е. в обоих языках значения переводных эквивалентов совпадают.

Перспективы

Нельзя утверждать, что системы морфологического анализа, «замкнутые» на морфологии, бесперспективны. Несомненно, во многих случаях системы данного класса будут разрабатываться и успешно применяться. Например, орфографический корректор все же достаточно успешно ищет опечатки, не привлекая семантическую и синтаксическую информацию.

Однако при разработке целого ряда прикладных систем с лингвистическими компонентами — например, систем автоматического анализа связного текста — применение контекстно-свободного морфологического анализа оказывается проблематичным. Включение такой системы в качестве одного из модулей (как это нередко делается) может сделать работу всего пакета неэффективной, ненадежной и нерациональной. Представляется, что будущее все-таки за системами, параллельно выполняющими анализ на всех уровнях (см. тж. Patten 1992).

С точки зрения программиста, проще хранить на диске большие массивы данных, чем заниматься поисками путей для уменьшения их объема за счет различных сложных алгоритмов, работающих с более компактными массивами данных. Ведь даже хранящийся в виде словаря словоформ массив данных, равный по объему словника словарю Зализняка, все же занимает на диске меньше места, чем многие из современных текстовых процессоров. Емкость жестких дисков, скорость доступа к данным и быстродействие, обеспечиваемые современной вычислительной техникой, во всех отношениях благоприятствуют применению подобных подходов. Тем не менее, с такой логикой разработки систем автоматического анализа трудно согласиться. Системы, «знающие» грамматику, не только более экономны, но и обладают большей предсказательной силой. Кроме того, когда нынешний этап построения систем морфоанализа будет завершен и начнется промышленная разработка систем с синтаксическим и семантическим анализом, несомненно, возникнет потребность в хранении очень больших массивов данных, во много раз превосходящих по объему морфологические базы данных, а также — в выполнении очень сложной обработки этих массивов. И если использование словаря словоформ для решения проблемы морфологического анализа может показаться оправданным, то вряд ли кто-либо посоветует хранить списки всех правильных предложений языка для выполнения синтаксического анализа.

Для дальнейшего развития морфологических анализаторов и лемматизаторов, несомненно, потребуется постепенный переход от контекстно-свободных систем к программам, обращающимся к контексту. Причем, как нам представляется, вполне реально строить обработку контекста как

дополнительный модуль, работающий с теми словами, которые не удалось обработать путем контекстно-свободного анализа. Наряду с применением чисто лингвистических подходов чрезвычайно продуктивным может оказаться и использование статистическо-вероятностного подхода. Системы, базирующиеся на статистической информации, нередко оказываются более надежными и менее восприимчивыми к ошибкам и опечаткам в тексте, чем исключительно лингвистические алгоритмы.

3.4. Стыковка перевода с оригиналом

3.4.1. Принципы автоматической стыковки параллельных текстов

Одна из важнейших технических проблем, возникающих при составлении параллельных корпусов текстов — это **стыковка текстов** (aligning), т.е. соотнесение фрагментов исходных текстов с соответствующими им фрагментами оригинальных текстов. Уровень стыковки может быть разным: стыковаться могут целые тексты (стыковка 1-го порядка)[31], абзацы (стыковка 2-го порядка), предложения (стыковка 3-го порядка) и даже слова (стыковка 4-го порядка).

Большинство существующих пакетов программ для ПКТ (например, *ParaConc*) работают с текстами, в которых уже отмечены параллельные места (Barlow, 1995). Выполнение такого рода работы вручную, даже для относительно небольшого корпуса текстов, — операция весьма трудоемкая. Разработка программного обеспечения для стыковки параллельных текстов пока еще делает первые шаги, хотя несколько более или менее эффективно работающих программ уже существует (Oakes, 1998: 135–137, 177–179). При разработке большинства программ-стыковщиков в той или иной степени используется структурное сходство перевода с исходным текстом.

Когда художественный текст переводится на другой язык, задача переводчика — получить красивый, хорошо читающийся текст на ПЯ, который в то же самое время передавал бы особенности языка текста на ИЯ. Поэтому, хотя сюжет, по-видимому, изменять нельзя, и новые главы, отсутствующие в оригинале, вряд ли кто одобрит, при переводе могут все же происходить довольно заметные изменения в структуре текста, если этого требуют нормы языка, на который выполняется перевод.

---

[31] Автоматическая стыковка целых текстов имеет смысл в тех случаях, когда корпус состоит из большого количества очень коротких текстов и их переводов на другой язык. Однако проблема выяснения, является ли текст Б переводом текста А, является очень специфической и выходит за рамки данной работы.

Тем не менее, переводчик, как правило, старается по возможности сохранить особенности языка и стиля оригинала. Если в исходном тексте предложения длинные, то вряд ли стоит переводить короткими предложениями. В то же время, если в переводе просто скопирована структура оригинала, то текст может выглядеть неестественно. Переводчик ищет «золотую середину» между двумя крайностями: слепым копированием структуры текста и вольным переводом, в котором от оригинала остался только сюжет.

Например, повесть Н.В. Гоголя «Шинель» замечательна необычайно длинными абзацами: эта небольшая по объему (около 10 000 слов) повесть состоит из 46 абзацев, средняя длина абзаца — 23 предложения или 217 слов. Эта повесть переводилась на финский язык много раз, и каждый раз переводчикам приходилось решать проблему длинных абзацев. Как правило, абзацы в переводе разбивались. В переводе Хуго Ялканена 206 абзацев, засчет того, что реплики в диалогах образуют отдельные абзацы, в переводе Юхани Конкки — 100. Однако третий переводчик этой повести — Эса Адриан — все же нашел возможным сохранить авторское членение на абзацы; его перевод, как и оригинал, состоит из 46 абзацев.

Для решения задачи автоматической или полуавтоматической стыковки текстов возможны два подхода — лингвистический и статистическо-вероятностный.

3.4.2. Лингвистический подход к стыковке параллельных текстов

Этот подход в первую очередь базируется на лексике и исходит из презумпции, что переводчик, как правило, пользуется стандартными словарными эквивалентами. Таким образом, если в фрагменте Б текста перевода встретились словарные эквиваленты многих слов из фрагмента А текста оригинала, то есть основания считать, что Б является переводом А при соблюдении некоторых условий, а именно: 1) А и Б должны находиться на сопоставимом расстоянии от начала текста, 2) А и Б должны быть сопоставимы по длине (см. рис. 22).

Таким образом, теоретически возможна программа-стыковщик, принцип работы которой основан на поиске эквивалентов слов исходного текста в переводе (см., например, Mihailov and Tommola, 2001, Mihailov, 2001).

Между лексемами двух языков возможны четыре типа корреляций:

- Одна лексема языка А ↔ Одна лексема языка Б
- Одна лексема языка А ↔ Несколько лексем языка Б
- Несколько лексем языка А ↔ Одна лексема языка Б
- Несколько лексем языка А ↔ Несколько лексем языка Б

К сожалению, корреляция A, которую можно было бы наиболее успешно применять для стыковки текстов, является, по-видимому, достаточно редкой. Более или менее часто она встречается только в близкородственных языках. Для таких далеких друг от друга языков, как русский и финский, такая корреляция — редкость. Большинство слов русского языка, значения которых полностью или почти полностью совпадают со значениями их финских эквивалентов, являются словами с очень конкретными значениями, например, *город* ↔ *kaupunki*, *самолет* ↔ *lentokone, книга* ↔ *kirja*, *сутки* ↔ *vuorokausi*.

Корреляции других типов являются более распространенными:

B: *палец* → *sormi*, *varvas*

C: *потолок, крыша* → *katto*

D: *корабль, теплоход, судно* → *laiva*, *alus*

Грамматики языков нередко еще более усложняют существующие лексические корреляции. В финском языке нет рода, поэтому для многих финских существительных, обозначающих лиц и живых существ, в русском языке по крайней мере два эквивалента, например *ystävä* → *друг, подруга*; *kissa* → *кошка, кот; hevonen* → *лошадь, конь*. Видовые пары увеличивают количество русских эквивалентов для финских глаголов, например, фин. *lukea* → рус. *читать, прочитать, почитать, почитывать* и т.д. Зато сложные слова финского языка очень часто соответствуют в русском языке словосочетаниям, например, *matkapuhelin* → *мобильный телефон, сотовый телефон*; *työpäivä* → *рабочий день*.

Взяв наугад несколько самых обычных слов русского языка и сравнив их частоты с частотами их наиболее естественных финских словарных эквивалентов, получаем следующую картину (табл. 14).

Из таблицы ясно видно, что переводчики на практике не так уж и часто пользуются стандартными словарными эквивалентами. Только для четырех слов из нашего списка (выделены курсивом) частоты словарных эквивалентов оказались сопоставимыми.

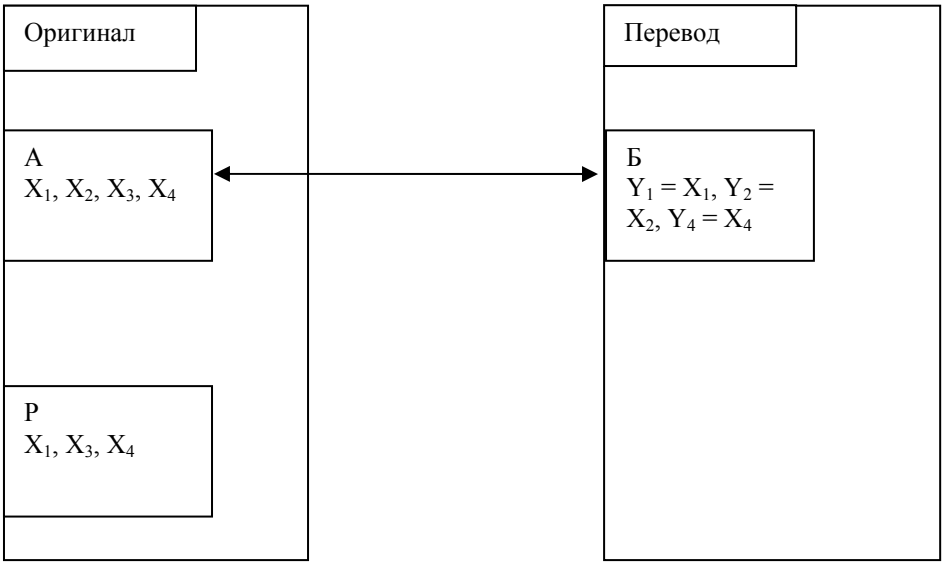

Рис. 22. *Стыковка текстов с использованием словарных эквивалентов*

Таблица 14. Частоты нескольких русских слов в корпусе ИЯ и их финских словарных эквивалентов в корпусе ПЯ

Русское слово	**Частота**	**Финский эквивалент**	**Частота**
человек	958	ihminen	921
женщина	140	nainen	323
мужчина	47	mies	523 (!)
время	448	aika	877
поезд	140	juna	156
ухо	84	korva	56
месяц	*57*	*kuukausi*	*53*
зима	40	talvi	66
хлеб	40	leipä	26
оружие	39	ase	88
политический	*35*	*poliittinen*	*31*
звезда	*33*	*tähti*	*36*
мясо	*18*	*liha*	*19*

Причин этому много. Очень часто ключевое слово оказывается «спрятанным» внутри композита, например русские словосочетания *французская булка* и *ржаной хлеб* переводятся на финский язык соответственно как *ranskanleipä* и *ruisleipä*; *скорый поезд* как *pikajuna*. Другая причина состоит

в том, что для финского языка распространение существительного с помощью генитива существительного является очень частой альтернативой распространению с помощью прилагательного, что приводит к переводу русского прилагательного финским существительным, например рус. *в зимнее время* → фин. *talven aikana*.[32]

Иногда переводчик использует прием генерализации, то есть вводит в качестве эквивалентов слова с более общим значением, чем слова оригинала. Например, для перевода русских слов *оружие, орудие, ружье, револьвер, автомат* используется одно и то же финское слово *ase*. Наконец, иногда лексема может оказаться частью фразеологизма, для перевода которого используется оборот, в котором словарные эквиваленты составляющих отсутствуют. Например, в качестве эквивалента русской идиомы *держать ухо востро с кем-либо* в текстах корпуса используется финская идиома *pitää kieli keskellä suuta*.[33]

Лексическая омонимия и многозначность также сильно изменяют частоты слов, являющихся переводными эквивалентами. Например, в вышеприведенном списке у финского слова *mies* кроме значения ‘мужчина’, есть и значение ‘муж’; у существительного *aika* ‘время’ есть омонимичное наречие *aika* ‘весьма’. Поэтому частоты этих слов существенно выше, чем частоты их русских соответствий.

И наконец, замена существительных на местоимения и опущение слов также приводит к довольно существенным различиям в частотах эквивалентов.

Показанные выше различия в частотах словарных эквивалентов все же не означают, что программу-стыковщик, основанную на лексических эквивалентах, построить нельзя. В основе программы, с помощью которой были состыкованы тексты англо-норвежского параллельного корпуса, лежали именно списки словарных эквивалентов плюс сравнение позиций стыкуемых фрагментов и их длин (Hofland and Johansson, 1998). Несмотря на то, что различия в частотах сильно затрудняет поиск переводных эквивалентов с использованием частотности употребления слов, это не делает такой поиск невозможным. Пересечение слов ИЯ и ПЯ в соответствующих фрагментах исходного текста и перевода позволяет во многих случаях находить переводные эквиваленты в автоматическом режиме (см. раздел 3.5 настоящей работы).

В целом, для лингвистического подхода характерно использование больших массивов данных: чем большей информацией располагает сис-

---

[32] Отметим, что данные такого типа лишний раз показывают проблематичность критикуемого в статье Ларса Борина (Borin 2002a) предположения Меламеда о том, что переводные эквиваленты должны относиться к одной и той же части речи.

[33] На самом деле эта финская идиома является лишь приблизительным эквивалентом для *держать ухо востро* и гораздо ближе по значению к другой русской идиоме *держать язык за зубами*.

тема, тем более эффективно она работает. Так, в статье Штрайтера, Иомдина и Сагаловой описывается программа-стыковщик, работающая с англо-русскими и русско-английскими параллельными текстами, в основе которой лежит система машинного перевода «ЭТАП-3». Программа переводит предложения исходного текста и сравнивает свой «машинный» перевод с переводом, выполненным человеком. При наличии достаточного количества совпадений, предложения или группы предложений стыкуются. Одновременно со стыковкой текстов происходит и «обкатка» системы МП, уточнение словарных эквивалентов и т.п. (Streiter et al 2001).

Крупным недостатком лингвистического подхода к разработке программ-стыковщиков является его слишком большая наукоемкость для такой частной технической задачи. Более интересным представлялось бы получать из параллельных текстов списки возможных переводных эквивалентов, одновременно проверяя эквиваленты, предлагаемые словарями, а не готовить такие списки вручную, чтобы состыковать тексты. Кроме того, такие стыковщики оказываются привязанными к конкретным парам языков, что также снижает их привлекательность.

3.4.3. Статистическо-вероятностный подход к стыковке параллельных текстов

Представляется, что пока наиболее эффективно работают алгоритмы, которые основаны на использовании не лингвистических данных, а данных о структуре текста оригинала и перевода. Эти методы не дают стопроцентных результатов, но в целом работают достаточно эффективно, а главное — это то, что разработка такого рода программ не требует столько времени и сил, как разработка лингвистических алгоритмов. Хорошим примером такого подхода является алгоритм Гэйла и Черча, первоначально разработанный и «обкатанный» на параллельном англо-французском корпусе парламентских дебатов (Canadian Hansards Corpus, см. стр. 247). Алгоритм основан на вполне очевидной закономерности: длинные предложения, как правило, переводятся длинными предложениями, а короткие предложения — короткими. Кроме того, некоторые типы стыковки встречаются чаще, чем другие, например, стыковка «одно предложение — одно предложение» встречается чаще, чем «два предложения — одно предложение» или «одно предложение — три предложения». Программа-стыковщик сравнивает длины предложений (в знаках) и подсчитывает «штраф» (penalty): чем больше наблюдаемая разница отличается от среднего соотношения для этой пары языков (по Гэйлу и Черчу 100 знаков английского текста соответствуют в среднем 120 знакам французского перевода), тем больше штраф. За тип стыковки также приписываются штрафные баллы. За стыковку 1:1 штраф равен нулю, поскольку стыковка такого типа — самая распространенная. За другие типы стыковки (1:2, 2:1,

1:0, 0:1 и т.п.) «начисляется» штраф и его размер зависит от вероятности такой стыковки. В итоге выбирается стыковка, получившая минимальное число штрафных баллов. Алгоритм хорошо работает для стыковок 1:1, для остальных типов процент ошибок довольно высок (Gale and Church 1993; Oakes 1998: 135–137).

Очевидно, оптимальным решением проблемы стыковки параллельных текстов было бы комбинирование двух подходов: за основу можно брать алгоритм Гэйла и Черча, но дополнительно использовать и некоторую лингвистическую информацию: списки ключевых слов (anchor words), пунктуацию (особенно пунктуацию в конце предложения — восклицательные, вопросительные знаки, многоточия), цифры (если и в переводе, и в оригинале числительные записаны цифрами, и эти цифры совпадают, то вероятность соответствия сравниваемых фрагментов становится очень высокой). Программа может также учитывать графически похожие слова, которые могут быть словарными эквивалентами для родственных языков или записями одних и тех же имен собственных для неродственных языков. К сожалению, последнее оказывается проблематичным, если языки используют разные алфавиты, как это имеет место в нашем случае.

3.4.4. Стыковка текстов корпуса «ПарРус»

Поскольку в нашем распоряжении не было финско-русских или русско-финских глоссариев в электронной форме, было решено разрабатывать программу-стыковщик, не использующую лингвистических данных.

Для стыковки текстов «ПарРус» была написана специальная утилита, работающая на уровне абзацев (текст программы см. в приложении 4.1). Решение остановиться на уровне абзаца было связано как с тем, что абзацы чаще стыкуются по принципу 1:1, чем предложения (и поэтому написание программы-стыковщика для абзацев оказывается технически проще), так и с тем, что абзац, как правило, образует более завершенный по смыслу контекст[34]. Недостатком программы оказалась слишком большая длина стыкуемых фрагментов для целого ряда авторов (в первую очередь, для писателей XIX века: Гоголь, Достоевский, Лермонтов), и в результате полученные длинные параллельные фрагменты приходилось впоследствии «резать» вручную. Поэтому стыковщик, работающий на уровне абзацев, является временным решением, стыковка на уровне предложений позволяет получать более корректные статистические данные и более эффективно выполнять поиск переводных эквивалентов (подробнее об этом см. в следующих разделах).

[34] Завершенность контекстов принципиально важна, поскольку главной функцией корпуса в настоящее время является получение параллельных конкордансов с последующим анализом вручную.

Сравнение формальной структуры русских оригинальных текстов и их переводов на финский язык показало те же тенденции, которые были обнаружены Гэйлом и Черчем при сравнении параллельных англо-французских текстов. Наш эксперимент отличался от исследований Гэйла и Черча тем, что они измеряли длины предложений в знаках, а мы измеряли длины текстов, предложений и абзацев в словах.

Данные по количеству слов, предложений и абзацев в текстах корпуса «ПарРус» и в их переводах на финский язык приводятся в приложении 3. Из таблиц ясно видно, что тенденция к сохранению оригинального членения на абзацы существует и что она довольно сильна. Более того, у некоторых переводчиков прослеживается утановка на сохранение и членения на предложения: это видно из того, что количество предложений в некоторых парах текстов совпадает или почти совпадает. Таблица показывает, что отношение количества слов оригинала к количеству слов перевода — более или менее стабильная величина. В дальнейшем будем называть эту величину коэффициентом ИЯ-ПЯ. Значение коэффициента зависит от пары языков. Для переводов с русского языка на финский этот коэффициент составляет примерно 1,06 (интересно, что для сравниваемых нами длинных текстов корпуса коэффициент оказался одинаковым и равнялся 1,07).

Нынешняя версия стыковщика использует обнаруженную тенденцию. Программа поабзацно сравнивает оригинал и перевод. Если отношение количества слов в сравниваемых абзацах оказывается близким к коэффициенту ИЯ-ПЯ для данной пары языков (в нашем случае — 1,06), то абзацы стыкуются и программа переходит к следующей паре абзацев.

Если отношение оказывается значительно больше 1,06, то это скорее всего значит, что переводчик разбил абзац оригинала на несколько абзацев. Программа добавляет абзац из перевода и повторяет вычисление отношения длин стыкуемых фрагментов.

В том случае, если отношение оказалось значительно меньше 1,06, то возможно, что переводчик объединил абзацы оригинала. В этом случае программа добавляет абзац из оригинала и вновь вычисляет отношение длин стыкуемых фрагментов.

Если на предыдущем шаге отношение оказалось слишком низким, а после добавления абзаца перевода — слишком высоким (или наоборот, если сначала отношение было слишком высоким, а после добавления абзаца оригинала стало слишком низким), то программа переходит в интерактивный режим. Пользователь вручную стыкует это сложное место, после чего программа продолжает работу в автоматическом режиме.

В целом программа работает достаточно устойчиво и надежно. Чем ближе к оригиналу перевод, тем быстрее происходит стыковка и тем реже программа обращается к пользователю. Проблемы начинаются в случаях «вольного» обращения с текстом, когда переводчик часто объединяет или разбивает абзацы, а также в случаях пропусков в переводе. Но коэффициент ИЯ-ПЯ оказывается достаточно эффективным барометром:

после двух или трех неправильных стыковок программа в подавляющем большинстве случаев находит «нестыкуемые» фрагменты и переходит в интерактивный режим.

После завершения стыковки текстов корпуса было решено проверить, насколько действенным оказался вычисленный нами коэффициент ИЯ-ПЯ и, в случае необходимости, уточнить его значение. Ведь значение было получено на основе сравнения длин целых текстов. При этом мы исходили из предположения, что отношение количества слов в исходном тексте к количеству слов в переводе показывает общую тенденцию и в соотношениях количества слов фрагментов. Однако некоторые фрагменты перевода могут оказаться заметно многословнее (например, в исходном тексте встретились очень сложные грамматические конструкции, культурные реалии или плохо переводимая игра слов), а другие могут быть, в свою очередь, заметно короче. Наконец, в переводе могут быть пропуски. Таким образом, коэффициент ИЯ-ПЯ, полученный путем вычисления отношений длин целых текстов, является лишь стартовым значением, требующим уточнения.

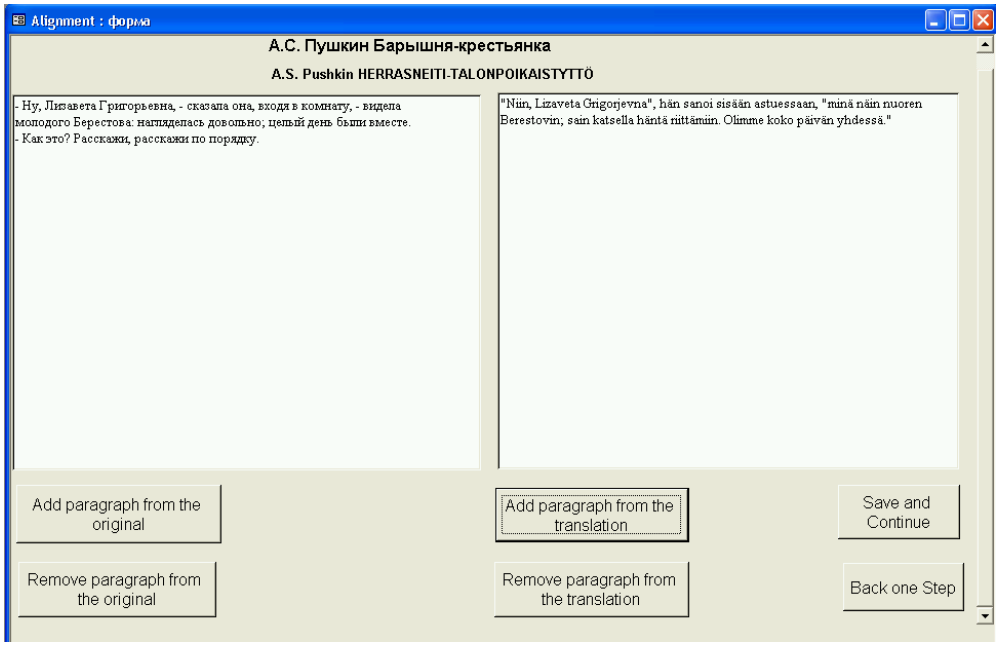

Рис. 23. *Интерфейс для выполнения стыковки параллельных текстов*

Для проведения эксперимента была написана программа, анализировавшая уже состыкованные тексты, и вычислявшая отношение длин (в словах) состыкованных фрагментов. Фрагменты исходного текста, не имеющие соответствий в переводе и фрагменты перевода, не имеющие соответствий в исходном тексте, в расчет не принимались. Затем вычислялось среднее значение коэффициента ИЯ-ПЯ по каждому тексту. В ходе

эксперимента анализировались только тексты, длина которых превышала 4000 слов. В целом, результаты эксперимента не сильно отличались от данных, полученных в ходе анализа целых текстов, а лишь слегка подкорректировали их. Итоговое значение коэффициента оказалось равным 1,08 при стандартном отклонении 0,06 (см. табл. 15).

Таблица 15. Коэффициент ИЯ-ПЯ в текстах «ПарРус»

Автор	**Название**	**Переводчик**	**Название перевода**	**ИЯ-ПЯ**
Аксенов В.	Звездный билет	Adrian E.	Matkalippu tähtiin	1,06
Бакланов Г.	Навеки девятнадцати-летние	Orlov V.	Synnyinmaan puolesta	1,00
Белов В.	Привычное дело	Laaksonen H.	Tuttu tarina	1,10
Булгаков М.А.	Мастер и Маргарита	Heino U.-L.	Saatana saapuu Moskovaan	1,08
Булгаков М.А.	Театральный роман	Adrian E.	Teatteriromaani	1,13
Гоголь Н.В.	Шинель	Konkka J.	Päällysviitta	1,12
Гоголь Н.В.	Шинель	Jalkanen H.	Päällystakki	1,15
Гоголь Н.В.	Шинель	Adrian E.	Päällystakki	1,13
Горький М.	Старуха Изергиль	Mitrošin A.	Isergil-muori	1,09
Горький М.	Челкаш	Mitrošin A.	Tšelkaš	1,05
Гроссман В.	Все течет	Adrian E.	Kaikki virtaa	1,09
Достоевский Ф.М.	Записки из подполья	Kallama V.	Kellariloukko	1,08
Достоевский Ф.М.	Преступление и наказание	Konkka J.	Rikos ja rangaistus	1,09
Достоевский Ф.М.	Записки из подполья	Adrian E.	Kirjoituksia kellarista	1,13
Дудинцев В.	Белые одежды	Heino U.-L.	Valkeat vaatteet	1,07
Ерофеев В.	Москва - Петушки	Adrian E.	Moskova-Petuški	1,15
Ильф И., Петров Е.	Двенадцать стульев	Silvanto R., Konkka J.	Kaksitoista tuolia	1,16
Ильф И., Петров Е.	Золотой теленок	Aarto A.	Kultainen vasikka	1,08
Лермонтов М.Ю.	Герой нашего времени	Heino U.-L.	Aikamme sankari	1,07
Лесков Н.	Очарованный странник	Pyykkö L.	Lumottu vaeltaja	1,13
Олеша Ю.	Зависть	Adrian E.	Kateus	1,09
Пастернак Б.Л.	Доктор Живаго	Konkka J.	Tohtori Živago	1,12
Приставкин А.	Ночевала тучка золотая	Adrian E.	Yöpyi pilvi kultainen	1,08
Пушкин А.С.	Пиковая дама	Pesonen P. Alarik	Patarouva	1,13
Пушкин А.С.	Барышня-крестьянка	Ahava J., Hämeen-Anttila V.	Aatelisneiti talonpoikaistyttönä	1,10
Пушкин А.С.	Пиковая дама	Hollo J.A.	Patarouva	1,06
Пушкин А.С.	Капитанская дочка	Hollo J.A.	Kapteenintytär	1,12

Автор	Название	Переводчик	Название перевода	ИЯ-ПЯ
Пушкин А.С.	Барышня-крестьянка	Hollo J.A.	Herrasneitti-talonpoikalaistyttö	1,05
Распутин В.	Живи и помни	Adrian E.	Elä ja muista	1,02
Семенов Ю.	Семнадцать мгновений весны	Pienimäki N.	Kevään seitsemäntoista hetkeä	1,30
Солженицын А.И.	Один день Ивана Денисовича	Adrian E.	Päivä Stalinin keskitysleirissä	1,01
Солженицын А.И.	Один день Ивана Денисовича	Lahtela M.	Ivan Denisovitšin päivä	1,04
Стругацкие А. и Б.	Парень из преисподней	Adrian E.	Poika helvetistä	1,11
Стругацкие А. и Б.	Попытка к бегству	Adrian E.	Pakoyritys	1,07
Толстой Л.Н.	Метель	Konkka J.	Pyry	1,10
Толстой Л.Н.	Два гусара	Konkka J.	Kaksi husaaria	1,07
Толстой Л.Н.	Анна Каренина	Pyykkö L.	Anna Karenina	1,06
Трифонов Ю.	Предварительные итоги	Anhava M.	Alustava tilinpäätös	1,04
Трифонов Ю.	Дом на набережной	Koskinen M.	Talo rantakadulla	1,03
Троепольский Г.	Белый Бим черное ухо	Iranto L.	Bim mustakorva	1,04
Тургенев И.С.	Дворянское гнездо	Heino U.-L.	Aateliskoti	1,06
Фадеев А.	Разгром	Heino U.-L.	Tuho	1,03
Чехов А.П.	Дама с собачкой	Heino U.-L.	Nainen ja sylikoira	1,04
Чехов А.П.	Дом с мезонином	Heino U.-L.	Taiteilijan tarina	1,09
Чехов А.П.	В овраге	Heino U.-L.	Rotkossa	1,02
Чехов А.П.	Мужики	Heino U.-L.	Talonpoikia	1,00
Чехов А.П.	Тайный советник	Konkka J.	Herra salaneuvos	1,08
Шукшин В.М.	Как зайка летал на воздушных шариках	Adrian E.	Kun pupujussi lensi ilmapalloilla	1,04
Шукшин В.М.	Страдания молодого Ваганова	Adrian E.	Nuoren Vaganovin kärsimykset	1,01
Шукшин В.М.	Алеша Бесконвойный	Adrian E.	Vartijaton Aljoša	1,01
Шукшин В.М.	Охота жить	Rymin R., Parkkinen P.	Halu elää	0,91
			Среднее значение	**1,08**
			Станд. отклон.	**0,06**

Как видно из таблицы 15, разброс значений довольно значительный: от 0,91 (В. Шукшин, «Охота жить», пер. Р. Рюмин и П. Парккинен) до 1,3 (Ю. Семенов, «Семнадцать мгновений весны», пер. Н. Пиенимяки). Насколько можно судить из наших данных, величина коэффициента во многом определяется стилем переводчика: для переводов всех трех наиболее широко представленных в нашем корпусе переводчиков — Э. Адриана, Ю. Конкка и У.-Л. Хейно — получены разные средние значения этого коэффициента (переводы Э. Адриана — 1,08, Ю. Конкка — 1,09, У.-Л. Хейно — 1,05). Кроме того, коэфф. ИЯ-ПЯ для переводов одного и того

же произведения, выполненных разными переводчиками, могут различаться, иногда — довольно сильно, например, для двух разных переводов «Пиковой дамы» он составил 1,13 и 1,06.

Тем не менее, в некоторых случаях соотношение длины исходного текста и перевода может быть довольно разным и в работах одного и того же переводчика, например, коэфф. ИЯ-ПЯ для переводов «Анны Карениной» и «Очарованного странника», выполненных Л. Пююккё, равен соответственно 1,06 и 1,13. Это говорит в пользу того, что язык и стиль оригинала также влияют на величину коэфф. ИЯ-ПЯ. В некоторых случаях его значения для переводов разных произведений одного и того же автора оказываются довольно близкими, например, в переводах произведений Достоевского и Толстого (см. табл. 15).

Но некоторые из текстов корпуса были написаны одним и тем же автором и переведены на финский язык одним и тем же переводчиком, и при этом значения коэффициента ИЯ-ПЯ различаются очень сильно. В качестве примера приведем переводы произведений Чехова «Дама с собачкой», «Дом с мезонином», «В овраге» и «Мужики», сделанные У.-Л. Хейно (см. табл. 3.23). Таким образом, индивидуальные особенности переводимого текста также в некоторых случаях играют немаловажную роль.

Используемый нами алгоритм стыковки был проверен и на другой паре языков — англо-русских переводах. Эксперимент проводился на нескольких произведениях английской и американской художественной прозы и их переводах на русский язык. Результаты оказались примерно такими же, как для русско-финских текстов. Количество абзацев в оригиналах и переводах оказывается довольно близким. Коэффициент ИЯ-ПЯ для англо-русских текстов равняется 1,25. Однако следует отметить, что стыковка англо-русских текстов происходила в целом менее гладко, чем стыковка русско-финских текстов. Вмешательство пользователя требовалось чаще.

В целом, остается впечатление, что переводы с русского на финский в большей степени сохраняют структуру оригинала, чем переводы с английского на русский. Имея в распоряжении довольно ограниченный по объему англо-русский материал, трудно сказать, связано ли это с различием переводческих традиций или с требованиями к художественному стилю в разных языках. Тем не менее, проведенные эксперименты позволяют констатировать, что сохранение целостности абзацев оригинала в переводе типично при выполнении художественного перевода и коэффициент, характеризующий соотношение длин абзацев в оригинале и переводе является достаточно стабильной величиной, которую можно использовать при стыковке оригинала с переводом.

Как уже было сказано, стыковка на уровне абзацев работает вполне эффективно. Тем не менее, нельзя не отметить следующие недостатки нашей программы:

- в некоторых случаях полученные состыкованные фрагменты оказываются слишком длинными; чаще всего это случается в тех случаях, когда абзацы оригинала длинные (как, например, в упомянутой выше «Шинели» Гоголя), либо когда переводчик по каким-либо причинам изменял членение на абзацы;
- чем дальше формальная структура перевода от структуры оригинала, тем хуже работает программа; например, очень плохо стыковался текст романа Р.-Л. Стивенсона «Остров сокровищ» (R.-L. Stevenson, “The Treasure Island”) и перевод этого романа на русский язык, выполненный Н. Чуковским;
- используемый нами алгоритм не является полностью автоматическим, что может означать определенные трудности при стыковке больших массивов текстов; причем получение полностью автоматического стыковщика без использования каких-либо дополнительных данных не представляется возможным, поскольку колебания коэффициента ИЯ-ПЯ для разных параллельных текстов слишком велики.

Многие из указанных недостатков могут быть преодолены в ходе дальнейшего развития программы-стыковщика. Представляется целесообразным в следующей версии программы реализовать алгоритм стыковки на уровне предложений. Для уменьшения количества ошибок, стыковщик на уровне предложений будет организован как дополнительный модуль, работающий с текстами, уже состыкованными на уровне абзацев с помощью программы, описанной в данном разделе. Поскольку программа будет работать с фрагментами текстов небольшой длины, риск ошибки сведется до минимума.

Кроме сравнения длин предложений будут также использоваться списки переводных русско-финских соответствий, полученных в результате работы программы поиска переводных эквивалентов, которая будет описана в разделе 3.5 настоящей работы.

Программа будет выполнять стыковку только в совершенно очевидных случаях, пропуская небольшие по объему абзацы, в которых стыковку предложений в автоматическом режиме выполнить не удается, и обращаться к пользователю только в таких ситуациях, когда не удается состыковать в автоматическом режиме предложения длинных абзацев.

3.5. Автоматический поиск переводных эквивалентов

3.5.1. Можно ли искать переводные эквиваленты в автоматическом режиме?

Одной из важнейших проблем, стоящих перед разработчиками параллельных корпусов текстов, является автоматизация поиска переводных эквивалентов (ПЭ) в ПКТ. Решение этой проблемы во многом облегчит работу лексикографов, а также откроет новые возможности в развитии систем искусственного интеллекта и других высоких технологий. Так, при составлении переводных словарей и по сей день подбор словарных эквивалентов производится практически вручную. Главными инструментами лексикографов до недавнего времени были лишь существующие толковые и переводные словари, недавно в их распоряжение поступили одноязычные корпуса текстов, лишь в последнее время, как уже говорилось в предыдущих разделах данной работы, начинают появляться многоязычные корпуса текстов. Но эти источники данных не позволяют лексикографу получить в автоматическом режиме хотя бы неполный исходный словник с переводными эквивалентами. Самым мощным из вышеуказанных инструментов оказывается ПКТ, с помощью которого исследователь может получить параллельный конкорданс. Однако и в этом случае ПЭ приходится искать вручную.

Использование ПКТ и многоязычных корпусов текстов в многоязычной лексикографии представляется необычайно продуктивным и перспективным делом. Тем не менее, видимо, в связи с тем, что на текущий момент имеется дефицит многоязычных текстовых ресурсов, дву- и многоязычные словари, за редкими исключениями, составляют без применения или с минимальным применением корпусов текстов. В своем довольно обширном обзоре лексикографических корпусных проектов Сьюзен Хоки не упоминает ни одного переводного словаря, составленного с применением корпусов текстов (см. Hockey 2000: 146–164). Работы по автоматическому поиску переводных эквивалентов пока в бóльшей степени ориентированы на разработки в области автоматической обработки текста и искусственного интеллекта.

Поиском путей решения проблемы автоматического поиска переводов слов в параллельных текстах начали заниматься практически одновременно с началом работ над ПКТ. Параллельно с разработкой алгоритмов стыковки параллельных текстов на уровне абзаца и предложения шли работы по стыковке на уровне слова (word alignment). Стыковка текстов на уровне слов фактически означает возможность получения списков переводных эквивалентов. Сама задача поиска ПЭ является, в сущности, более скромной, поскольку не предполагает нахождения в конечном итоге соответствия для каждого слова или словосочетания текста.

Стыковка на уровне предложений и стыковка на уровне слов оказались тесно связанными направлениями, поскольку для стыковки предложений в качестве дополнительного средства нередко используются словарные соответствия, а по состыкованным текстам проще организовать поиск соответствий на уровне слов. Хотя проблему автоматического поиска ПЭ решенной считать нельзя, исследователям удалось добиться на этом направлении впечатляющих результатов (см., напр., Oakes 1998: 174–178).

Поиск переводных эквивалентов, как и многие другие сложные процедуры, нельзя выполнять «с чистого листа». Очень разумным представляется предлагаемый Й. Тидеманном «принцип маленьких шагов» (Tiedemann 1999b), суть которого заключается в том, что сложная проблема разбивается на несколько менее масштабных подпроблем, по мере решения которых решается и основная проблема[35]. Кроме того, до начала поэтапного решения проблемы поиска словарных эквивалентов желательна предварительная подготовка самого параллельного массива. Тексты должны быть состыкованы. Создание программы, которая искала бы ПЭ в несостыкованных текстах, представляется в высшей степени проблематичным. Другая начальная установка состоит в том, что словники субкорпусов должны быть лемматизированными; особенно это важно для языков с богатым словообразованием. В противном случае КПД системы будет очень низким. И, наконец, автоматический поиск ПЭ возможен только в массивах текстов достаточно большого объема. Если объем корпуса маленький, многие из предлагаемых программой соответствий могут оказаться случайными.

Нам ничего не известно о системах, которые использовали бы для автоматического поиска ПЭ только лингвистические алгоритмы. Дело в том, что грамматическая информация мало что дает для поиска лексических эквивалентов. Другое дело, если исходные тексты и их переводы переводятся в некое универсальное семантическое представление (СемП). В этом случае появится возможность сравнить СемП двух текстов, а затем «отследить» поверхностное представление для тех узлов, которым удалось найти соответствия. Разработка систем такого класса, по всей видимости, — дело будущего.

На данном этапе подавляющая часть проектов, связанных с автоматическим построением двуязычных глоссариев, идет по другому пути, по пути механического сравнения текстов оригинала и перевода. Однако к такому подходу нельзя относиться скептически, поскольку, во-первых, он быстро дает результаты, и, во-вторых, таким образом можно накопить большие массивы данных, что позволит впоследствии разрабатывать более «интеллектуальные» системы (снова «принцип маленьких шагов»).

---

[35] Этот принцип был предложен М. Кэйем (M. Kay) применительно к проблеме построения систем автоматизированного перевода (Tiedemann 1999b).

3.5.2. Существующие подходы к автоматизации поиска ПЭ

Компьютерная программа, ищущая ПЭ в параллельных текстах, оказывается в положении Ж.-Ф. Шампольона перед Розеттским камнем, на котором было два текста: один на древнегреческом, другой — на древнеегипетском. Положение французского ученого было даже лучше: он знал один из языков и догадывался о том, к какой языковой группе относится второй. Программа может только механически сравнивать фрагменты текстов, о которых известно, что фрагмент B является переводом фрагмента A. Решение такой задачи со многими неизвестными состоит в привлечении всех имеющихся средств и комбинировании полученных после применения каждого из методов результатов. Так, Тидеманн перечисляет следующие подходы для выполнения стыковки на уровне слов:

- анализ совместной встречаемости слов в битекстах;
- поиск похожих слов;
- стыковка текстовых фрагментов, состоящих из одного слова;
- стыковка низкочастотных элементов;
- использование электронных словарей;
- использование полученных на более ранних этапах анализа пар ПЭ (Tiedemann 1999b).

Ни один из перечисленных методов не позволит найти соответствия для всех слов исходного массива текстов (включая слова исходного текста, для которых в переводе нет соответствия, и наоборот — слова перевода без соответствий в исходном тексте). Однако последовательное применение нескольких или даже всех перечисленных методов и повторная обработка текстов позволит повысить КПД системы.

Анализ совместной встречаемости слов

Этот подход основывается на поиске в оригинальных текстах и их переводах таких слов, которые бы последовательно употреблялись в соответствующих битекстах; иначе говоря, если слово α из языка A встречается в субкорпусе X в фрагментах X_1, X_2, ..., X_n, а слово β из языка B — в параллельном субкорпусу X субкорпусе переводов на язык B в соответствующих фрагментах X_1', X_2', ..., X_n', то велика вероятность того, что слово β является эквивалентом слова α.

Этот подход использовали многие исследователи, работавшие в основном с языками Центральной и Западной Европы (см., например, Gale and Church 1991, Kay and Röscheisen 1993, Tiedemann 1998, Tiedemann 1999a, Tiedemann 1999b). Фактически, при поиске ПЭ на основе совместной встречаемости можно использовать те же принципы, что и при автоматическом

поиске терминологических словосочетаний. Разница заключается только в том, что при поиске терминов используются данные об употреблении нескольких слов в качестве ближних коллокатов, а при поиске ПЭ учитывается употребление слов ИЯ и ПЯ в соответствующих фрагментах параллельных текстов. При выполнении обеих задач учитываются следующие данные о частотах лексем L_i и L_j (мы, однако, дадим формулировки применительно к извлечению ПЭ-пар):

- a — количество параллельных фрагментов, в которых употреблены и L_i, и L_j;
- b — количество параллельных фрагментов, в которых употребляется L_i, а L_j отсутствует;
- c — количество параллельных фрагментов, в которых употребляется L_j, а L_i отсутствует;
- d — количество параллельных фрагментов, в которых не употреблены ни L_i, ни L_j.

Разные исследователи предлагают различные количественные меры, которые можно использовать для того, чтобы оценить степень связи между исследуемыми единицами. Назовем некоторые из них.

Простой коэффициент соответствия (simple matching coefficient, SMC), значения которого варьируются от 0 до 1:

$$SMC = \frac{a+d}{a+b+c+d}.$$

Коэффициент Кульчинского (Kulczinski coefficient, KUC), интервал значений также от 0 до 1:

$$KUC = \frac{a}{2}\left(\frac{1}{a+b}+\frac{1}{a+c}\right).$$

Коэффициент Очиаи (Ochiai coefficient, OCH), варьирует от 0 до 1:

$$OCH = \frac{a}{\sqrt{(a+b)(a+c)}}.$$

Коэффициент Фагера и МакГоэна (Fager and McGowan coefficient, FAG), интервал значений — от минус бесконечности до 1:

$$FAG = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{a+b}}.$$

Коэффициент Йула (Yule coefficient, YUL), интервал значений — от — 1 до +1.

$$YUL = \frac{ad-bc}{ad+bc}.$$

Коэффициент МакКонноти (McConnoughy coefficient, MCC), интервал значений — от -1 до +1.

$$MCC = \frac{a^2 - bc}{(a+b)(a+c)}.$$

Значения коэффициента $\mathbf{\Phi^2}$ составляют от 0 до плюс бесконечности. Этот коэффициент использовался Гейлом и Чёрчем для стыковки параллельных текстов на уровне слов (Gale and Church 1991).

$$\Phi^2 = \frac{(ad - bc)^2}{(a+b)(a+c)(b+c)(b+d)}.$$

Наконец, достаточно часто используются **коэффициенты совместной встречаемости MI** и **MI3**.

$$MI = \log_2 \frac{aN}{(a+b)(a+c)},$$

$$MI3 = \log_2 \frac{a^3 N}{(a+b)(a+c)}, \text{ где N = a + b + c + d.}$$

Коэффициент MI критиковался за слишком высокие значения для слов с низкой частотой, поэтому чаще используется модификация MI3, в которой величина *a* (то есть частота совместной встречаемости) возводится в куб (подробнее обо всех упомянутых выше коэффициентах см. Oakes 1998: 170–172).

Остановимся кратко на применении метода в конкретных проектах.

Французские исследователи Госье, Ланже и Менье (Gaussier, Langé, Meunier 1992) разработали методику извлечения англо-французских эквивалентов из параллельного корпуса текстов. В качестве источника данных использовался Canadian Hansards, параллельный англо-французский корпус дебатов в канадском парламенте (подробнее об этом корпусе см. стр. 18). Тексты корпуса состыкованы на уровне предложений, что делает возможным поиск ПЭ. Для выделения наиболее вероятных ПЭ исследователи ввели коэффициент, характеризующий силу связи между английскими и французскими словами. Этот коэффициент вычисляется по следующей формуле, являющейся вариантом коэффициента MI:

$$I(e,f) = \log_2 \frac{p(e,f)}{p(e)p(f)},$$

где *p(e)* и *p(f)* соответственно — вероятности появления английского слова *e* и французского слова *f* в исследуемом корпусе текстов (вычисляется путем деления абсолютной частоты слова на количество состыкованных фрагментов корпуса), *p(e, f)* — вероятность появления обоих слов в одном и том же фрагменте текста. Экспериментальным путем была установлена пороговая величина коэффициента *I(e, f)*. Программа обнаруживала в текстах те слова, для которых данный коэффициент был

выше пороговой величины, а затем отбрасывались те французские слова, для которых была установлена более сильная связь с другими английскими словами. В результате экспериментов правильные французские ПЭ были найдены примерно для 65% английских слов, для 25% слов эквиваленты найдены не были, для 10% слов были предложены неправильные эквиваленты.

В рамках проекта PLUG (о проекте см. стр. 20) разрабатывались пакеты программ для стыковки параллельных текстов на уровне слов, причем исследование не ограничивалось одной языковой парой: изучались возможности стыковки шведско-немецких и шведско-английских параллельных текстов. В системе UWA (Uppsala Word Aligner) используется несколько подходов к поиску ПЭ (по этой причине эта программа будет упомянута в дальнейшем в связи с другими методиками поиска ПЭ), одним из которых является анализ совместной встречаемости.

В качестве количественного критерия в алгоритме программы используется **коэффициент Дайса**, который изначально применялся в прикладной лингвистике как показатель сходства сравниваемых строк символов (Oakes 1998: 125). Например, этот коэффициент может оказаться весьма полезным при работе орфографического корректора для поиска вариантов правильного написания слова. Другим его применением является поиск похожих слов в параллельном корпусе (см. далее). Коэффициент Дайса вычисляется по следующей формуле:

$$Dice = \frac{2c}{a+b},$$

где c — количество общих контекстов для сравниваемых слов; a, b — частоты этих слов.

В целом, применение этого подхода позволило получать ПЭ для обеих языковых пар, причем с довольно высокой точностью, около 86% для шведско-немецких текстов, для англо-шведских текстов результаты несколько ниже — около 70% (Tiedemann 1998). Интересно, что эти эксперименты лишний раз демонстрируют, различную эффективность одного и того же подхода применительно к разным языковым парам.

Большой интерес представляет методика стыковки текстов, предлагаемая Кэйем и Рёшейзеном (Kay and Röscheisen 1993). Стыковка предложений и слов выполняется одновременно, оба модуля в процессе работы взаимодействуют, что позволяет динамически корректировать результаты работы программы. Программа составляет три индекса, в первом из них соотносятся слова и предложения, во втором дается список полученных лексических соответствий, в третьем — таблица соответствий предложений оригиналов и переводов (то есть номер предложения исходного текста и номер предложения перевода).

Программа начинает строить таблицу соответствий предложений исходя из того, что: 1) первое и последнее предложения перевода стыкуются с

первым и последним предложениями оригинала, 2) предложения оригинала и перевода с близкими порядковыми номерами, в которых используются слова, являющиеся ПЭ, могут быть состыкованы.

После стыковки очередной пары предложений производится проверка лексических пар на совместную встречаемость. Для проверки используется коэффициент Дайса. В том случае, если коэффициент Дайса превышает заданную пороговую величину, в таблицу соответствий записывается очередная ПЭ-пара, после чего обновляется таблица соответствий предложений. Процесс повторяется циклически до тех пор, пока программа не перестанет находит новые лексические соответствия.

Данные о совместной встречаемости слов далеко не всегда позволяют с достаточно высокой точностью определить, являются ли данные слова переводными эквивалентами. Эта методика позволяет находить только те соответствия, которые регулярно повторяются. Поэтому для того, чтобы повысить качество работы системы, целесообразно использовать и другие методики.

Поиск похожих слов

Как известно, в разных языках нередко встречаются слова, похожие фонетически и близкие по графическому оформлению. Это могут быть слова, пришедшие из общего праязыка (англ. *brother* — нем. *Bruder* — рус. *брат*; англ. *sister* — нем. *Schwester* — рус. *сестра*) или заимствования друг у друга или из третьего языка (англ. *sport* — нем. *Sport* — рус. *спорт*, рус. *таракан* — фин. *torakka*). Чем ближе языки друг к другу, чем больше было языковых контактов, тем больше таких похожих слов. Такие слова обычно близки и по значениям и часто используются в качестве переводных эквивалентов. Однако можно назвать довольно много случаев, когда значения слов расходятся очень сильно. Например, значение русского слова *плац* гораздо у́же значения немецкого слова *Platz* ‘место, площадь’, *пенальти* и *гол* также намного конкретнее своих английских прототипов *penalty* ‘наказание’ и *goal* ‘цель, ворота’, рус. *канава* и фин. *kanava* ‘канал’ также довольно далеки друг от друга. Такие слова довольно часто оказываются «ложными друзьями переводчика».

К этой группе примыкают и случайные совпадения, например, англ. *clever* и рус. *клевер*, англ. *hotel* и *хотел* — форма прошедшего времени от русского глагола *хотеть*, финск. *orava* ‘белка’ и рус. *орава*, финск. *porukka* ‘компания, группа людей’ и рус. *порука*[36].

[36] Возможно, что финское *porukka* является заимствованием из русского языка, но значение настолько сильно изменилось, что слово воспринимается как случайное совпадение.

Из вышесказанного следует, что одно лишь формальное сходство слов далеко не всегда означает тождественность их лексических значений, даже в близкородственных языках. Поэтому похожие слова могут быть признаны ПЭ-парой лишь в том случае, если для этого есть дополнительные основания, например, если значение коэффициента совместной встречаемости хотя и является низким, но все же выше некоторого критического значения.

Далее, для сравнения строк можно воспользоваться коэффициентом Дайса:

$$Dice = \frac{2c}{a+b},$$

где c — количество одинаковых букв в сравниваемых словах; a, b — длины этих слов.

Другой часто использующийся коэффициент — отношение длины наибольшей общей подстроки (составляется из всех букв, присутствующих в обоих сравниваемых словах) к длине более длинного слова. Так, для слов *sister* и *Schwester* наибольшей общей подстрокой будет «ster», а коэффициент будет равняться 4/7 = 0,57; наибольшая общая подстрока для пары *sister* и *сестра* (*sister, sestra*) — «sstr», а коэффициент — 4/6 = 0,67.

Тидеманн предлагает усилить эту методику, используя при сравнении строк не только буквы, но и соответствия на уровне буквосочетаний (Tiedemann 1999a).

В целом, сравнение строк может оказаться очень действенным средством для поиска ПЭ в параллельных корпусах текстов родственных языков. Тем не менее, эта методика и в этом случае остается служебным средством, поскольку исследователей в первую очередь интересуют непохожие эквиваленты. Все же применение этого метода облегчает решение и этой задачи, поскольку после установления тривиальных соответствий программа может методом исключения установить и более сложные соответствия.

Вспомогательные методики

Как уже говорилось выше, использование только одного метода не позволит добиться стопроцентного обнаружения ПЭ. Для того, чтобы повысить процент обнаруженных ПЭ и уменьшить количество ошибок, используют различные вспомогательные процедуры, на которых нельзя построить поиск эквивалентов в целом, но которые в некоторых случаях позволяют достаточно точно идентифицировать переводные эквиваленты.

Наиболее очевидной из этих вспомогательных методик является использование существующих двуязычных словарей. Подключение к про-

грамме электронных версий словарей позволяет состыковать те пары эквивалентов, которые зарегистрированы в словарях.

Далее, при анализе параллельных текстов очень важными оказываются короткие текстовые фрагменты, состоящие из нескольких слов или даже из одного слова. Найти соответствия между словами из этих фрагментов не представляет большой трудности.

Тидеманн (Tiedemann 1999b) предлагает еще одну интересную методику получения ПЭ-пар. Задаются две критические частоты t_1 и t_2, $t_2 > t_1$. Программа убирает из текста все слова, частота которых ниже t_1. В некоторых текстовых фрагментах останется по одному слову. Те из полученных пар, частоты которых не превышают t_2, считаются эквивалентами. По словам Тидеманна, такой алгоритм может работать довольно эффективно, если правильно подобрать частоты t_1 и t_2.

Наконец, повторная обработка текстов также повышает эффективность работы программы. Используя информацию об уже найденных ПЭ-парах, программа может найти эквиваленты для тех слов, которые на предыдущих этапах оставались без эквивалентов.

Нерешенные проблемы

Главной проблемой при организации поиска переводных эквивалентов в параллельных текстах является то, что не у всех слов есть переводные эквиваленты. Безэквивалентная лексика нередко пропускается или переводится описательно, и в результате найти эквивалент не удается.

Программы, как правило, находят только однословные соответствия, найти соответствия для словосочетаний намного труднее. В тех случаях, когда в одном языке понятие обозначается сложным словом, а в другом языке — словосочетанием, программа принимает какую-либо из частей словосочетания за однословный эквивалент. С этим связано, например, более низкое качество поиска ПЭ в англо-шведских текстах по сравнению с немецко-шведскими (см. выше): в шведском языке намного больше сложных слов, чем в английском. Но если пару «словосочетание — сложное слово» все же можно в конечном итоге обнаружить на этапах повторной обработки текстов, пара «идиома — идиома» скорее всего будет представлена в виде нескольких ПЭ-пар.

При поиске ПЭ для слов с очень низкой частотой вероятность ошибки очень велика: если слово употреблено в текстах всего один раз, то найти для него ПЭ в большинстве случаев очень сложно.

3.5.3. Поиск переводных эквивалентов в «ПарРус»

Совместная встречаемость или графическое сходство?

Как уже говорилось выше, все известные нам компьютерные программы, выполняющие поиск переводных эквивалентов в параллельных корпусах текстов, работают с родственными языками, причем с языками Западной Европы, в которых имеется довольно большой пласт «общей лексики», которая заимствовалась из одних языков в другие.

В ходе данного исследования была сделана попытка разработки компьютерной программы, которая, используя те же принципы, искала бы переводные эквиваленты в русско-финском параллельном корпусе. Русский и финский языки не являются родственными, в русском языке очень мало заимствований из финского, в финском языке есть заимствования из русского, но таковых относительно немного (*tavara* ‘товар’, *smetana* ‘сметана’, *viesti* ‘весть’, *tuuma* ‘дума’ и т.п.). В финском языке довольно мало интернационализмов, очень многие слова, общие для большинства европейских языков, в финском языке или вообще отсутствуют или являются малоупотребительными (ср. напр. англ. *party* — рус. *партия* — фин. *puolue*; англ. *group* — рус. *группа* — фин. *ryhmä*; англ. *computer* — рус. *компьютер* — фин. *tietokone*). Таким образом, основной акцент при поиске ПЭ приходится делать именно на совместную встречаемость.

Тем не менее, графическое сходство некоторых слов может помочь найти ПЭ, особенно в тех случаях, когда частотность исследуемых лексем низка. Кроме того, разработка модуля, выполняющего поиск эквивалентов на основе графического сходства, может представлять и теоретический интерес.

Поиск ПЭ на основе графического сходства

Эффективность сравнения графического облика слов разных языков может сильно снижаться вследствие различий в графических системах сравниваемых языков, особенно если в них используются разные алфавиты. С этим мы как раз и имеем дело в случае работы с русско-финскими параллельными текстами.

Очевидно, что для сравнения слов таких языков необходимо привести их к единому графическому виду, то есть: 1) если в одном из языков используется другой алфавит, записать слова обоих языков в одном и том же алфавите, 2) заменить все специальные буквы и буквосочетания на сочетания единого стандарта. В итоге слова будут записаны как бы в упрощенной транскрипции. Например, английское *book* и немецкое *Buch* можно записать в виде *buk* и *buh*. Английское *sister*, немецкое *Schwester* и русское

сестра можно записать в виде: *sister, švester, sestra*[37]. Используемая система записи должна учитывать, по крайней мере, особенности фонетики и орфографии сравниваемых языков.

В случае сравнения лексиконов русского и финского языка для повышения качества поиска имеет смысл слегка модифицировать систему транслитерации. Поскольку в финском языке отсутствуют шипящие, то русские шипящие «ш», «щ», «ж» заменяются при транслитерации на латинскую «s». С другой стороны, в русском языке отсутствуют долгие гласные и согласные фонемы, при транслитерации как финских, так и русских слов удвоенные согласные и гласные буквы заменяются на одиночные. Приведем несколько примеров транслитерации:

Русский язык:

чистый → sistij, иллюзия → iljusija, миллион → milion, бар → bar

Финский язык:

siisti → sisti, illuusio → ilusio, miljoona → miljona, baari → bari

Предлагаемые модификации являются наиболее легко осуществимыми, поскольку затрагивают только графику и фонологию. Для повышения качества поиска ПЭ возможно введение дополнительных модификаций, например снятие русских и финских грамматических финалей, например в финском — финали *-nen*, в русском — финали *-ий*. В этом случае русское слово *синий* в «нормализованной» форме приняло бы вид *sin*, а финское слово *sininen — sini*. Проведение такой операции позволило бы еще больше сблизить графическое оформление родственных слов разных языков. Однако в нашем эксперименте было решено ограничиться только буквенной транслитерацией без снятия грамматических показателей.

Как уже отмечалось выше, для проведения теста на графическое сходство слов чаще всего используются коэффициент Дайса и длина максимальной общей подстроки. Оба показателя имеют существенный недостаток: учитывается лишь совпадение символов в определенной позиции. Между тем, по крайней мере для европейских языков, согласные играют существенно более важную роль, чем гласные. Именно согласные составляют «костяк» морфем. Например, при сравнении русского слова *товар* и финского *tavara* решающее значение имеет совпадение согласных *t, v* и *r*, а не гласных *о* и *а*. «Равноправие» гласных с согласными может приводить к появлению шума. Например, для пары *товар — tavara* значение коэффициента Дайса довольно высокое и составляет 0,73. Однако для другой пары никак не связанных друг с другом слов *товар — kova* ’твердый’ значение коэффициента также оказывается высоким и равняется 0,67. Поэтому представляется вполне целесообразным совпадение согласных

[37] Предложенный способ записи, разумеется, не является единственно верным. Для «унификации» графики могут быть выбраны разные системы, не последнюю роль здесь играет и сама пара языков. Главное — добиться того, чтобы из слов исчезли «немые» звуки, аналогичные звуки передавались одной и той же буквой или буквосочетанием.

оценивать выше, чем совпадение гласных. Кроме того, следует различать полное совпадение позиции буквы или нахождение букв в близких позициях. В качестве количественного критерия принимается *p*, который вычисляется по следующей формуле:

$$p = i\left(\frac{l_1}{l_2}\right)$$

где *i* — позиция буквы в слове ИЯ, l_1 — длина слова ИЯ, l_2 — длина слова ПЯ.

Позиция совпавших букв в сравниваемых словах считается близкой при

$$p = j \pm 1,$$

где j — позиция совпавшей буквы в слове ПЯ.

При сравнении русских и финских слов за совпавшие буквы присваиваются следующие веса:

	Согласные	**Гласные**
В той же позиции	1	0,5
В близкой позиции	0,5	0,25

Для оценки степени близости графического облика русских и финских слов используется вариант коэффициента Дайса:

$$Dice = \frac{2c}{a+b},$$

где *c* — суммарный вес совпавших букв в сравниваемых словах; *a, b* — длины этих слов.

Пороговое значение коэффициента было установлено экспериментальным путем и составило 0,6.

Программа поиска ПЭ на основе графического сходства работает в качестве модуля в среде «КОКОС-П». Поиск эквивалентов основан на сравнении русского и финского лемматизированных словников. Для ускорения работы программы в таблицы со словниками были добавлены поля с унифицированной транслитерацией слов. Программа вычисляет коэффициент Дайса для русских и финских слов; в том случае, если коэффициент Дайса для сравниваемой пары больше или равен 0,6, программа считает слова переводными эквивалентами. Совместная встречаемость слов в расчет не принималась.

Программа вполне успешно находит пары эквивалентов, являющихся заимствованиями из западноевропейских языков, например, *агроном — agronomi, баррикада — barrikadi, эскадрон — eskadroona* и т.п.

В некоторых случаях в качестве ПЭ-пар предлагаются неточные соответствия, членами пары оказываются однокоренные слова, относящиеся к разным частям речи, например, *аптекарь — apteekki* ’аптека’,

аристократизм — aristokraattinen ’аристократичный’*, большевизм — bolševikki* ’большевик’и др.

Часть предлагаемых пар оказалась совершенно неправильной, например, *бусы — bussi* ’автобус’*, вельможа — valmis* ’готов’*, заря — sarja* ’серия’*, суслик — šašliikki* ’шашлык’ и т.п.

Как и следовало ожидать, процент графически близких слов, являющихся переводными эквивалентами в исследуемой паре языков, оказался небольшим. Таким образом, можно сделать вывод, что использование графического сходства позволяет лишь несколько повысить качество поиска эквивалентов на основе совместной встречаемости.

Поиск переводных эквивалентов на основе совместной встречаемости

Подготовка «золотого стандарта»

Перед тем, как писать программу, ищущую переводные эквиваленты, мы решили подготовить «золотой стандарт», то есть массив данных, служащих образцом, на который следует ориентироваться при работе над программой (Ahrenberg et al 2000). Для этого была написана программа, которая сделала поиск эквивалентов «начерно».

Сначала была построена матрица, в которую были включены все слова из лемматизированного русского словника «ПарРус» с частотой от 2 до 20 включительно, а также все номера контекстов этих слов в индексе[38]. Затем аналогичная матрица была построена для финского словника. Объем русской матрицы составил 21695 слов, финская матрица оказалась несколько меньше — 14795 слов. На следующем этапе эксперимента программа сравнила полученные две матрицы и те пары слов, для которых количество совпавших контекстов превышало 0,5 от частоты русского слова, были записаны в таблицу возможных эквивалентов. Затем список был проверен вручную, были помечены правильные, отчасти правильные и ошибочные соответствия.

Как и следовало ожидать, такой грубый алгоритм дал довольно низкое качество, правильными оказались чуть более половины предложенных пар (всего было получено 4427 пар, из них правильных — 2166, отчасти правильных — 151, неправильных — 2111). Однако нашей целью не было получение готовой программы, мы хотели проверить на полученном массиве действенность существующих коэффициентов, применяемых для поиска ПЭ (см. предыдущий раздел), а также выяснить, насколько эффективен поиск по совместной встречаемости для русского и финского языков.

[38] Каждый отдельно хранящийся текстовый фрагмент корпуса имеет свой уникальный номер, причем этот идентификационный номер является общим и для фрагмента оригинального текста и для соответствующего ему фрагмента перевода.

Вопрос об эффективности поиска ПЭ на основе совместной встречаемости

Как уже отмечалось выше, чем ближе друг к другу языки, тем результативнее автоматический поиск ПЭ в параллельных текстах с этой парой языков. Тидеманн (Tiedemann 1998) отмечает, что поиск ПЭ в немецко-шведских параллельных текстах проходит более успешно, чем в англо-шведских текстах. Однако английский и шведский неизмеримо ближе друг к другу, чем русский и финский: первые два языка даже относятся к одной и той же германской языковой группе. Единственное фундаментальное различие между английским и шведским языками, сильно усложняющее поиск ПЭ, — это большое количество сложных слов в шведском языке и нетипичность такого словообразования для английского языка.

Как и следовало ожидать, с нашей парой языков результаты оказались намного скромнее. Как уже было указано выше, для 21 000 русских слов было обнаружено чуть более 2000 правильных ПЭ, то есть около 10%. Правда, этот результат получен на низкочастотных словах, для более высоких частот результат оказался лучше (об этом см. ниже). Кроме того, полученную цифру нельзя рассматривать как окончательную, поскольку анализировались не полные словники, а слова в определенном интервале частот, поэтому часть пар эквивалентов не была обнаружена лишь потому, что финское соответствие могло быть более частотным.

Подавляющее большинство обнаруженных пар оказалось существительными или прилагательными (1925 правильных пар), глаголов и неизменяемых слов оказалось несравненно меньше (218 пар глаголов и 23 пары неизменяемых слов). Только для существительных и прилагательных количество правильных пар превышает количество ошибочных пар (см. рис. 24).

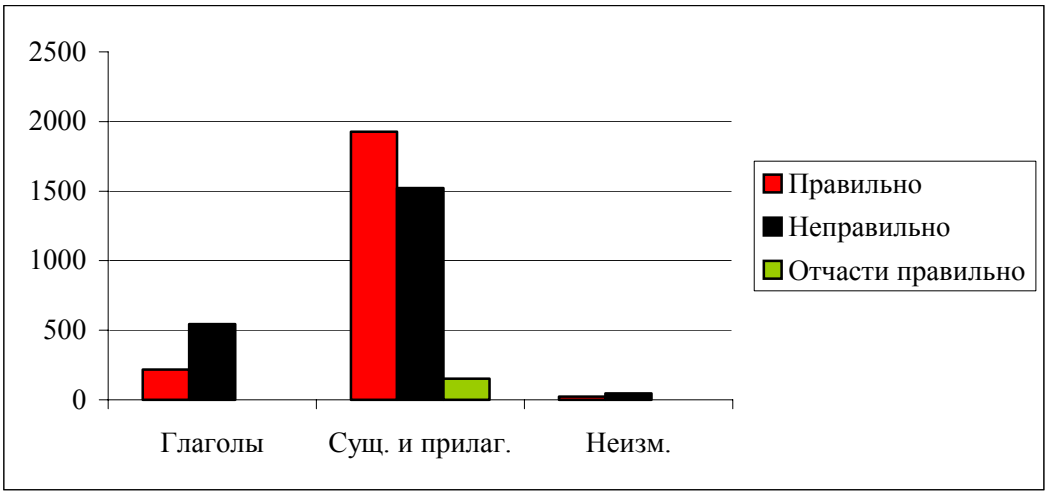

Рис. 24. *Результаты поиска ПЭ: части речи*

Начиная эксперимент, мы считали, что корректный поиск ПЭ возможен только для слов с невысокой частотой, предположительно — от 5 до 10. Выдвигая эту гипотезу, мы исходили из того, что практически все высокочастотные слова многозначны, и поэтому у них должно быть много ПЭ, что должно приводить к значительным различиям между частотами русских слов и их эквивалентов. Интервал от 2 до 20 был выбран для проверки гипотезы.

Эксперимент внес некоторые коррективы в наши представления. На диаграмме 24 показана зависимость между частотой русских слов и количеством правильных и неправильных ПЭ-пар. Действительно, для слов с частотой менее 4 количество ошибок недопустимо высоко: для слов с частотой 2 было найдено всего 9 правильных пар, для частоты 3 количество ошибок более чем в два раза превышает количество правильных пар (1382 против 591). Однако уже при частоте 4 количество ошибок резко снижается и оказывается меньше, чем количество правильных ответов практически для всех исследованных частот.

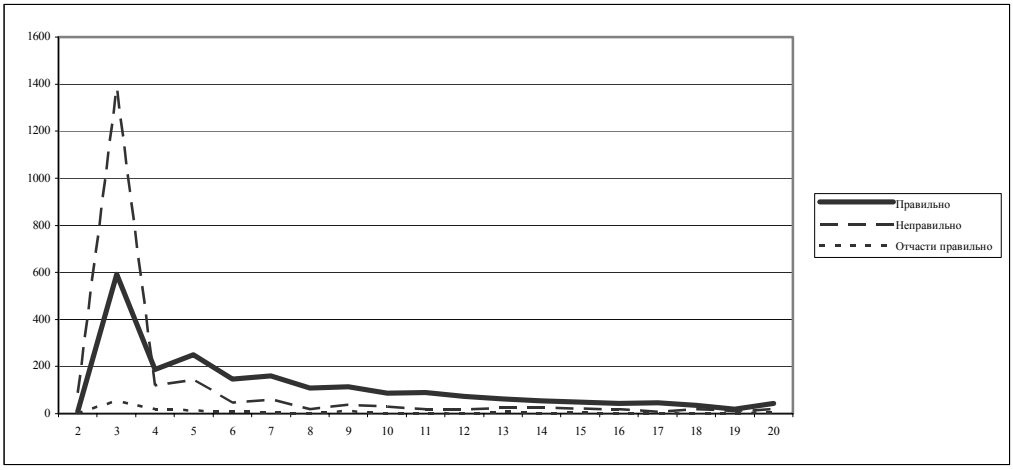

Рис. 25. *Результаты поиска ПЭ: частотность и ошибки*

Проверка действенности различных коэффициентов совместной встречаемости

Эксперимент с низкочастотными словами позволил также проверить действенность различных коэффициентов совместной встречаемости. Для всех полученных для «золотого стандарта» пар эквивалентов вычислялись девять коэффициентов: SMS, Dice, KUC, YUL, FAG, MI3, F2, MCC и OCH (см. стр. 143). Затем определялись критические значения коэффициентов, позволяющие «отсечь» бóльшую часть неправильных эквивалентов и выяснялось, какой процент неправильных пар все же не будет «отсеян». Как и ожидалось, стопроцентного результата не дал ни один из иссле-

довавшихся коэффициентов: порогового значения, которое «отсекало» бы все неправильные ПЭ-пары, обнаружить не удавалось. Таким образом, вопрос заключался в том, какой из коэффициентов позволял бы сохранить значительную часть правильных ПЭ-пар и при этом не «пропускал» бы слишком много ошибок.

Несколько коэффициентов оказались для нашего массива абсолютно неприемлемыми: они практически не снижали количество ошибочных пар. Например, если применять коэффициент Фагера и МакГоэна (FAG, см. рис. 3.26), то при значениях коэффициента от -0,5 до 0 удастся сохранить большую часть правильных ПЭ-пар, однако в нашем эксперименте лишь при значениях коэффициента -0,049 количество правильных ПЭ-пар сильно превышает количество неправильных (204 против 92), для остальных значений количество правильных и неправильных пар оказывается довольно близким. Для некоторых значений в этом интервале количество неправильных пар значительно превышает количество правильных, например, при значении коэффициента -0,429 зафиксировано 20 правильных пар и 84 неправильных. Таким образом, применение этого коэффициента вряд ли повысит качество работы программы. Аналогично обстоит дело с коэффициентами YUL и MCC.

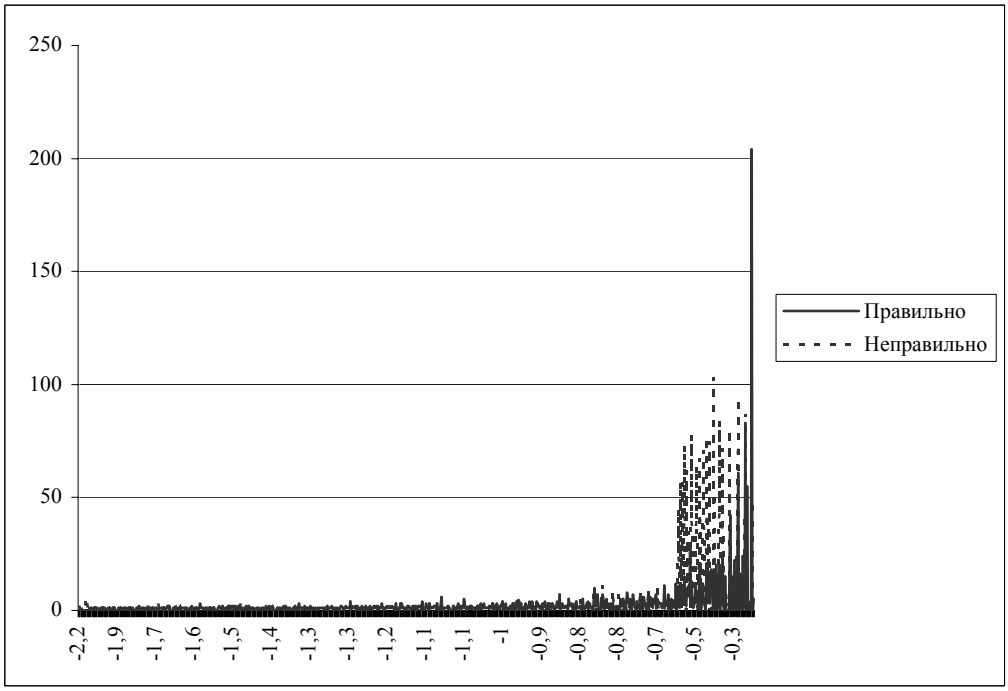

Рис. 26. *Действенность коэффициента FAG при поиске ПЭ-пар*

Другие коэффициенты сводили ошибки к минимуму, но «выплескивали с водой и ребенка»: количество правильных эквивалентов, отброшенных

вместе с неправильными, оказывалось недопустимо велико. Приведем в качестве примера коэффициент F2 (рис. 27). Если определить в качестве порогового значения 10 000, то коэффициент позволить идентифицировать как правильные лишь 548 ПЭ-пар из имеющихся в экспериментальном массиве 2166 правильных пар, то есть всего 25%. Хотя количество неправильных пар, которые будут признаны «правильными», составит всего 138 из 2111, или 7%, продуктивность системы окажется слишком низкой.

Коэффициенты Дайса и Кульчинского оказались более действенными, по крайней мере, на исследовавшемся интервале частот (см. рис. 28). Мы выбрали для дальнейшей работы коэффициент Кульчинского. Экспериментальным путем для этого коэффициента было установлено критическое значение, которое составило 0,5. Для большинства значений коэффициента в интервале от 0,5 до 1 количество правильных ПЭ-пар превышает количество неправильных. В случае применения этого интервала на нашем экспериментальном массиве удается получить 69% имеющихся правильных пар, ошибочных пар «золотого стандарта» остается 25%. Из ПЭ-пар, для которых значение этого коэффициента больше или равно 0,5, неправильными оказывается порядка 35%.

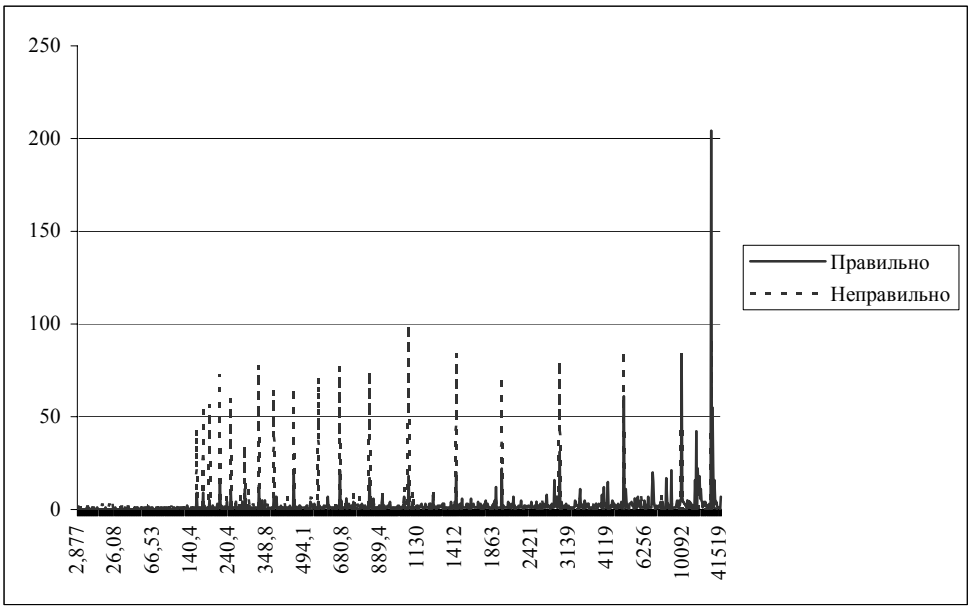

Рис. 27. *Действенность коэффициента F2*

Автоматический поиск ПЭ в системе «КОКОС-П»

Проведенный эксперимент позволил разработать для системы «КОКОС-П» отдельный модуль, позволяющий искать переводные эквиваленты для

части слов русского субкорпуса[39]. Программа позволяет искать эквиваленты как для отдельных слов, так и для списков слов, вплоть до обработки всего генерального словника.

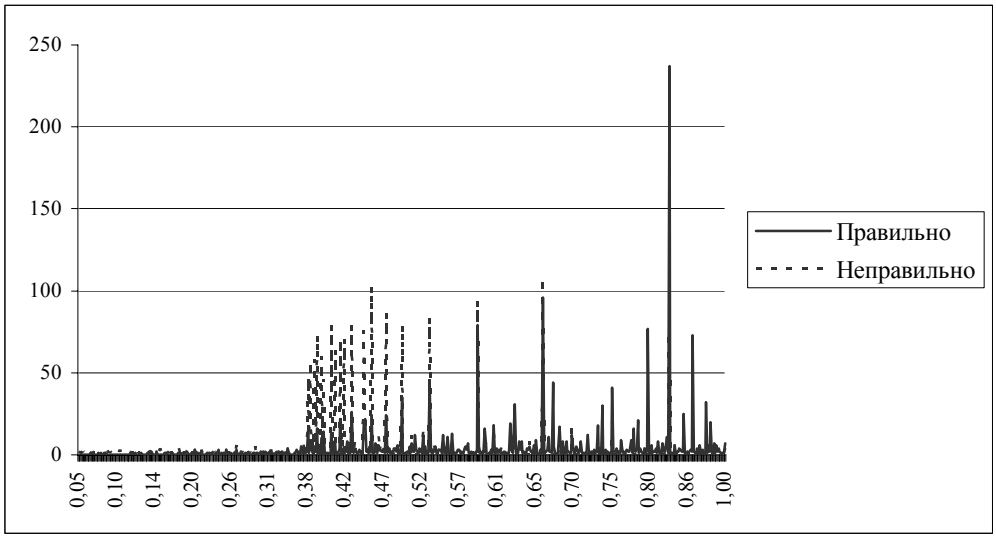

Рис. 28. *Действенность коэффициента Кульчинского*

Алгоритм работы программы в общих чертах строится следующим образом:

1. найти слово в лемматизированном словнике;
2. получить список всех финских словоформ, использовавшихся в параллельных контекстах;
3. найти для финских словоформ начальные формы по лемматизированному финскому словнику;
4. вычислить частоту финских слов, употребленных в параллельных контекстах, а также количество контекстов, в которых были употреблены и русское и финское слово;
5. вычислить для всех найденных пар коэффициент Кульчинского;
6. те пары, для которых KUL ≥ 0,55, считать ПЭ-парами.

[39] Попыток разработки модуля, выполняющего обратную операцию, то есть поиск русских эквивалентов для финских слов не предпринималось — хотя последняя также вполне выполнима — лишь по той причине, что поиск финских эквивалентов для русских слов представляется более естественной операцией для русско-финского корпуса текстов. Другая причина — экспериментальный характер модуля.

В процессе проверки эффективности алгоритма был выявлен один недостаток коэффициента Кульчинского. Напомним, что этот коэффициент вычисляется по формуле:

$$KUC = \frac{a}{2}\left(\frac{1}{a+b} + \frac{1}{a+c}\right),$$

где a — количество параллельных контекстов, в которых употреблено и слово A (ИЯ), и его возможный эквивалент B (ПЯ); b — количество контекстов, в которых употреблено слово A, но не употреблено слово B; c — количество контекстов, в которых не употреблено слово A, но употреблено слово B.

Как и все другие коэффициенты совместной встречаемости, этот коэффициент оказывается неэффективным, если одно из слов встретилось в массиве один или два раза. Например, если оба слова встретились в массиве по одному разу в одном и том же параллельном контексте, то коэффициент оказывается равным единице (максимальное значение этого коэффициента). Аналогично, если частоты сравниваемых слов очень сильно различаются, но количество совпавших контекстов велико, то значение коэффициента также оказывается высоким. В результате программа предлагает большое количество случайно встретившихся в тех же контекстах слов в качестве ПЭ. Например, слово *следственно* встречается в корпусе 34 раза, а *antiteesi* ‘антитеза’ — 4 раза. Однако, поскольку число совпадений равно 4, коэффициент Кульчинского оказывается равным 0,56, поэтому эти слова предлагаются в качестве ПЭ-пары.

Этот недостаток является одновременно и достоинством, поскольку программа оказывается способной находить эквиваленты для многозначных слов и других слов, частоты которых сильно отличаются от частот их ПЭ. Например, слово *мама* встречается в русских текстах 275 раз, а его финский эквивалент *äiti* — 1239 раз (основная причина состоит в том, что это же финское слово чаще всего используется в качестве эквивалента и для слов *мать, мама, мамочка, мамуля* и т.п.). Однако, поскольку количество совпадений составило 275, коэффициент Кульчинского для этой пары равен 0,61, то есть программа опознает эти слова как ПЭ-пару.

Чтобы уменьшить шум, был сделан «стоп-словарь», в который вошли самые высокочастотные слова финского словника, главным образом союзы, местоимения, частицы, предлоги и послелоги. Слова, которые входят в этот список, в качестве эквивалентов не предлагаются. Поиск эквивалентов для этих слов программа также не ведет, но, как уже было сказано выше, поиск для слов с очень высокой частотой в любом случае часто оказывается безрезультатным.

Эффективность работы программы

Для проверки эффективности программы был выполнен поиск переводных эквивалентов для слов, встретившихся в нашем корпусе от 20 до 400 раз (текст программы см. в приложении 4.4). Программа нашла 1708 ПЭ-пар,

из которых 124 оказались неправильными. Таким образом, ошибки составили лишь около семи процентов.

Как и предполагалось в начале эксперимента, процент слов, для которых программа смогла подобрать финский эквивалент, оказался весьма невысоким, в среднем около 20%. Причем, процент этот колеблется довольно сильно (см. рис. 29), вначале постепенно падая от 26% (диапазон 20—40) до 16% (диапазон 160—180), после чего можно наблюдать и резкие пики (32% в диапазоне 280—300), и не менее резкие падения (12% в диапазоне 340—360). Однако, если учесть, что с повышением частоты количество слов уменьшается в геометрической прогрессии (см. рис. 30) и соответственно и уменьшается количество найденных ПЭ-пар, становится ясно, что эффективный поиск переводных эквивалентов можно вести для слов, употребляющихся с частотой в интервале от 20 до 220.

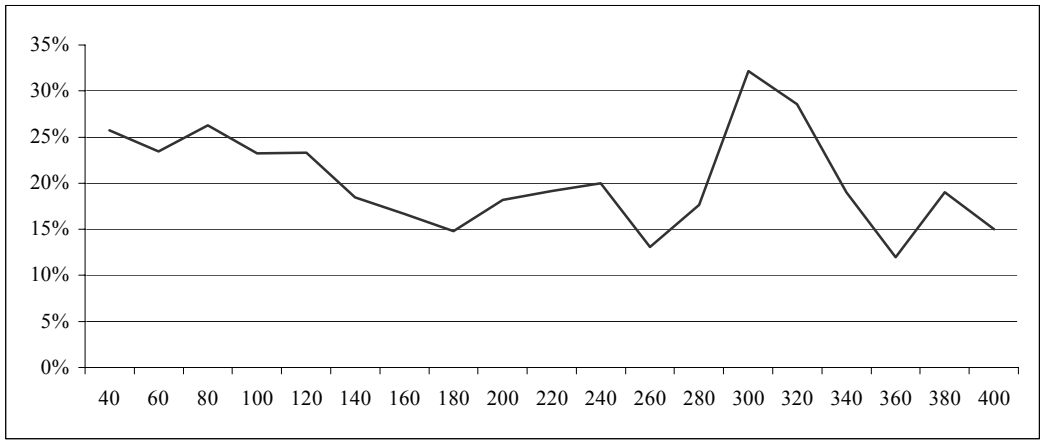

Рис. 29. *Зависимость количества найденных ПЭ-пар от частоты слов (в процентах)*

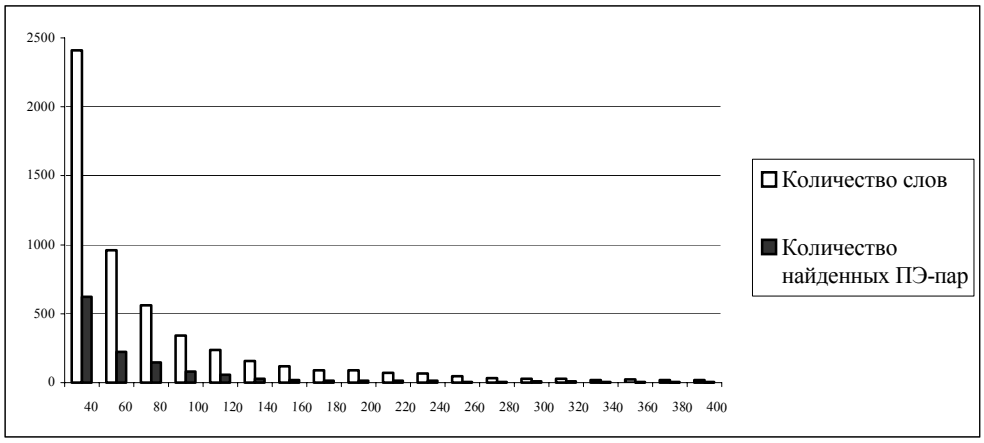

Рис. 30. *Количество слов в заданном диапазоне частот vs. количество предложенных ПЭ-пар*

«Набело» работа программы проверялась на окончательной версии «ПарРус» по состоянию на ноябрь 2002 года. В корпусе была вручную снята грамматическая омонимия для словоформ, частота которых превышала 30 употреблений[40], а также вручную определены леммы для неопознанных словоформ с частотой более 10. Это позволило увеличить объем лемматизированного словника, а также устранить значительную часть высокочастотных словоформ с неустановленными леммами, что также в итоге повышает эффективность работы программы. Поиск производился на интервале частот, установленном в ходе предыдущего эксперимента, то есть от 20 до 220. В русском лемматизированном словнике 7290 слов, имеющих такую частотность. Программа составила глоссарий, включающий 2080 пар (полный список полученных эквивалентов см. в приложении 5, частоты слов и частеречные пометы сняты в целях экономии места). Качество поиска оказалось вполне удовлетворительным: 90% найденных эквивалентов оказались правильными, около 5% были частично правильными (например, эквивалент *rykmentinkommentaja* ’полковой командир’ для слова *командир*) и лишь 5% оказались ошибочными (см. табл. 15).

[40] Снять омонимию для более низких частот не представлялось возможным из-за огромной длины списков, особенно финского, который превышает 10 000 словоформ.

Таблица 16. Результаты эксперимента по поиску переводных эквивалентов на основе совместной встречаемости: качество поиска

	Количество	**%%**
Правильные эквиваленты	1871	90,0 %
Частично правильные эквиваленты	102	4,9 %
Ошибочные эквиваленты	107	5,1 %
Всего найдено	**2080**	**100,0 %**

Как и в ходе предыдущих экспериментов, большая часть полученных ПЭ-пар оказалась существительными или прилагательными. Глаголы составили менее 6% найденных программой пар. Это связано, прежде всего, с глагольным видом в русском языке: одному финскому глаголу, как правило, соответствует два русских глагола — несовершенного и совершенного вида. У некоторых русских глаголов, например, у глаголов движения, может быть довольно много различных дериватов (напр. *идти, пойти, придти, уйти, зайти, подойти, отойти, выйти* и т.д.), и в этом случае межъязыковые соответствия размываются еще сильнее. Не случайно среди полученных глагольных ПЭ-пар нет ни одного глагола движения.

Процент слов, относящихся к другим частям речи, оказался очень невысоким; из 75 обнаруженных правильных пар 50 оказались наречиями и 20 — числительными. Поиск эквивалентов для наречий усложняется тем, что в русском языке наречия, образованные от прилагательных, часто оказываются омонимичными краткой форме прилагательного в среднем роде (например, *быстрый — быстро*). В том случае, если наречие отсутствует в словаре основ лемматизатора, форма опознается однозначно как краткое прилагательное и леммой является прилагательное. В итоге наречие как бы «растворяется» внутри прилагательного, и это уменьшает количество «пересечений» русского наречия с его финским эквивалентом. Соответствие числительных в разных языках, как правило, одно-однозначное, но качество поиска эквивалентов оказывается невысоким по той причине, что числительные нередко записываются цифрами, русские составные числительные являются словосочетаниями, частотность многих числительных оказывается очень высокой и превышает установленную нами границу 220 употреблений (приведем частоты некоторых русских числительных: *два* — 3410, *три* — 1547, *четыре* — 610, *пять* — 914).

Местоимения в полученном списке эквивалентов отсутствуют, поскольку их частоты даже в небольших по объему корпусах текстов почти всегда превышают 220.

Таким образом, поиск переводных эквивалентов в автоматическом режиме позволяет получать в основном списки соответствий для существительных и прилагательных.

Таблица 17. Результаты эксперимента по поиску переводных эквивалентов на основе совместной встречаемости: части речи правильных ПЭ-пар

Часть речи	**Количество**	**%%**
Существительные	1448	77,4%
Прилагательные	243	13,0%
Глаголы	105	5,6%
Прочие	75	4,0%
Всего:	**1871**	**100,0%**

Проверка на графическое сходство по полученному списку позволила в автоматическом режиме установить правильность 177 пар эквивалентов, то есть менее чем для 10% списка. Это лишний раз подтверждает нашу гипотезу о малой продуктивности использования данного критерия для русско-финских корпусов текстов.

Недостатки программы и пути ее совершенствования

Хотя разработанная нами программа позволяет искать в автоматическом режиме переводные эквиваленты для части слов, ее можно считать только первым шагом на пути решения проблемы поиска ПЭ-пар.

Главный недостаток алгоритма состоит в том, что он не позволяет искать переводные эквиваленты для слов с частотами 1 и 2. Эти слова составляют большую часть словника, например, в «ПарРус» их 16 382 или 42,16%. Редкие слова могут быть интересны и с точки зрения перевода, поскольку многие из слов, не относящихся к лексическому ядру, являются труднопереводимыми. Кардинальных путей решения проблемы пока найти не удается, лишь в некоторых случаях ПЭ-пары низкочастотных слов могут быть обнаружены с использованием дополнительных критериев: употребления в коротких контекстах, формального сходства и т.п.

Другая проблема — словосочетания. Кроме ПЭ-пар вида «слово — слово» достаточно распространены пары «слово — словосочетание», «словосочетание — слово» и «словосочетание — словосочетание». Наш алгоритм позволяет искать только соответствия вида «слово — слово». Между тем, для исследуемой пары языков очень характерны ПЭ-пары «словосочетание — слово», поскольку в финском языке чрезвычайно широко распространены сложные слова, которым в русском языке нередко соответствует словосочетание, например, рус. *железная дорога* — фин. *rautatie* (*rauta* ‘железо’ + *tie* ‘дорога’), рус. *многоэтажный дом* — фин. *kerrostalo* (*kerros* ‘этаж’ + *talo* ‘дом’). Наш алгоритм в лучшем случае находит только часть словосочетания, например, предлагает в качестве эквивалента для русского слова *письменный* финское *kirjoituspöytä* ‘письменный стол’ (в данной версии поиска частеречная принадлежность игнорировалась). Вычленение составляющих финского сложного слова и организация поиска русских

словосочетаний с высоким индексом совместной встречаемости по отношению к финскому композиту помогло бы решению данной проблемы.

Еще один недостаток алгоритма состоит в том, что последний рассчитан только на поиск регулярно повторяющихся переводных эквивалентов. Поэтому в большинстве случаев программа предлагает лишь один переводной эквивалент, иногда два; три правильных эквивалента при обработке нашего материала не было найдено ни разу. При поиске ПЭ многозначных слов программа находит переводной эквивалент для более частотного значения. Так, для русского слова *машина* программа нашла только эквивалент *auto* ‘автомобиль’. Таким образом, глоссарий, полученный с помощью программы, содержит по большей части тривиальные ПЭ-пары. С этим недостатком, по-видимому, вряд ли можно бороться, поскольку он вытекает из самого принципа работы алгоритма программы: только те пары слов, которые часто повторяются в параллельных контекстах, могут быть эквивалентами.

Качество работы программы в значительной степени зависит от длины параллельных фрагментов. Во всех проектах, упомянутых в нашем обзоре, программы, выполняющие поиск переводных эквивалентов, работают с текстами, состыкованными на уровне предложений. Тексты же «ПарРус» состыкованы на уровне абзацев. В результате многие параллельные фрагменты оказываются довольно длинными, что вызывает некоторую нестабильность в работе программы: совместная встречаемость оказывается несколько «смазанной». Например, для слова *ариец* программой были предложены следующие финские эквиваленты: *arjalainen, pahennus, luonteenlatu* и др. Правильным является только первый из предложенных эквивалентов (правда, именно у пары *арией — arjalainen* самый высокий индекс Кульчинского — 0,95). Причиной большого количества неправильных эквивалентов явилось то, что слово *арией* в корпусе употреблялось только в романе Ю. Семенова «Семнадцать мгновений весны» в партийных характеристиках, составленных по единому образцу, что вызвало появление одних и тех же слов в большом количестве контекстов. Если бы тексты были состыкованы на уровне предложений, подобная ошибка была бы менее вероятна.

В то же время программа, несмотря на указанные нами недостатки, уже позволяет строить в автоматическом режиме глоссарии, которые можно использовать как для дальнейшего расширения до полноценного словаря, так и для разнообразных электронных словарей и баз данных, а также для развития программного обеспечения самого корпуса текстов. Например, такие двуязычные глоссарии можно использовать в качестве списков пар ключевых слов (anchor words) для повышения качества стыковки параллельных текстов.

Глава 4. Корпус «ПарРус»: Язык русских оригинальных текстов vs. язык финских переводов

4.1. Как сравнивать оригинал и перевод?

Параллельный корпус текстов открывает новые перспективы для исследований в области переводоведения. Только с появлением ПКТ появляется реальная возможность исследовать письменный перевод на большом эмпирическом материале.

Тем не менее, возможности исследователя напрямую зависят от программного обеспечения, обслуживающего корпус текстов. Главная сложность в работе с параллельными текстами состоит в том, что без установления связей между оригинальными текстами и их переводами, параллельный корпус превращается в два одноязычных корпуса с весьма скромными возможностями исследования соответствий между оригиналами и переводами.

Действительно богатый материал может дать только состыкованный параллельный корпус с возможностями поиска лингвистической информации. В корпусе «ПарРус» грамматическая разметка пока не выполнена, однако, поскольку тексты состыкованы и выполнена лемматизация как русских, так и финских текстов, уже есть хорошие возможности для изучения особенностей литературного перевода с русского языка на финский.

В данной главе будут описаны некоторые особенности русско-финского литературного перевода, которые удалось установить на материале «ПарРус». Мы остановимся на вопросах изучения соответствий между исходным текстом и переводом, а также особенностей языка переводных текстов по сравнению с языком текстов того же языка, не являющихся переводами. Задача этой главы — обозначить основные направления, по которым можно проводить полномасштабное исследование, и разработать методику работы.

4.2. Что длиннее — оригинал или перевод?

С точки зрения семиотики перевод с одного языка на другой — это попытка перекодировать текст на одном языке в текст на другом языке. В отличие от других видов перевода, выделяемых Р.О. Якобсоном — пересказа и межсемиотического перевода (Jakobson 1989), — в которых используется та же самая или принципиально другая семиотическая система, в случае межъязыкового перевода применяется семиотическая система того же типа — другой естественный язык. Поскольку языки отличаются друг от друга, информационные потери неизбежны, и хороший переводчик должен стараться как-то их компенсировать. Поэтому Ю. Найда и Ч. Тэйбер утверждают, что хороший перевод всегда несколько длиннее исходного текста, поскольку переводчик должен эксплицировать понятия, отсутствующие в языке перевода (Nida & Taber 1974: 163).

Это утверждение в целом не вызывает никакого протеста.[41] Тем не менее, представляется интересным проверить, насколько оно соответствует реальной практике перевода. Такая проверка оказывается не столь тривиальной процедурой, как это может показаться. Ведь длину текста можно измерять по-разному: в знаках, в словоупотреблениях, в предложениях, в абзацах и т.п. Кроме того, грамматический строй и особенности стиля исходного языка и языка перевода (далее — ИЯ и ПЯ) не могут не влиять на длину перевода. Некоторое сомнение вызывают также утверждения Найды и Тэйбера о том, что во всех языках избыточность (redundancy) составляет примерно 50% и что «плотность» информации для разных языков одинакова (там же).

Для проверки этих утверждений была получена статистика по длинам текстов «ПарРус» в словоупотреблениях, предложениях, абзацах и знаках. Подробная статистика приведена в приложении 3 настоящей работы. По количеству знаков финские тексты, как правило, действительно оказывались длиннее русских оригиналов, в среднем отношение количества знаков оригинала к количеству знаков перевода по исследовавшимся текстам составило 0,87. Зато по количеству словоупотреблений русские оригинальные тексты длиннее финских переводов (в разделе о стыковке параллельных текстов уже упоминался коэффициент 1,06). По количеству предложений и абзацев оригиналы короче переводов, но совсем не намного: отношение количества предложений в оригинальных текстах к количеству предложений в переводах в среднем равно 0,93, аналогичный показатель по абзацам составил 0,94.

---

[41] Следует тем не менее отметить, что единого представления о том, что такое «хороший перевод», не существует: кто-то может считать, что главное в переводе — верность оригиналу, кто-то — что важнее хороший стиль, причем требования это могут меняться от эпохи к эпохи и зависят от типа переводимого текста (см., например, Chesterman 1997).

Теперь попробуем понять, что дает нам эта статистика.

Начнем с более крупных единиц. Использование предложения или абзаца в качестве единицы измерения длины текста представляется малопродуктивным по целому ряду причин. Во-первых, длины предложений и абзацев очень сильно варьируется. Например, в романе Булгакова «Мастер и Маргарита» длина предложения колеблется от одного до ста девяти словоупотреблений. Большая часть предложений романа имеет длину от одного до двадцати пяти словоупотреблений. Во-вторых, как уже было показано выше, переводчик в большинстве случаев — осознанно или неосознанно — переводит одно предложение одним предложением и один абзац — одним абзацем. Поэтому количество предложений и абзацев в переводе всегда оказывается близким к количеству предложений и абзацев в оригинале. То, что в финских переводах русских художественных произведений количество предложений и абзацев все-таки несколько больше, скорее всего, связано с различиями в стилистических традициях.

Использование словоупотребления в качестве единицы измерения длины текста также не даст надежных результатов. Слово — достаточно трудноопределимое понятие, границы между словом и словосочетанием весьма размыты. Поэтому даже такая простая вещь, как подсчет словоупотреблений в тексте на самом деле не так проста. Например, русское выражение *тридцать три* — это одно слово или два слова? А его финский аналог *kolmekymmentäkolme* — это одно слово или два слова? А *33* — это слово или нет?

Разумеется, подсчет словоупотреблений, при котором используются «граничные символы» (пробелы, запятые, точки и т.п.) [42], в целом достигает своей цели, давая слегка огрубленную картину. Другое дело, что сравнивать количество словоупотреблений в текстах, написанных на двух разных языках с очень сильно различающимся грамматическим строем, — весьма проблематичное занятие. Например, если в языке есть артикли и они пишутся отдельно, как в английском, немецком или французском, то количество словоупотреблений в сообщении на этом языке будет, как правило, больше, чем количество словоупотреблений в переводе этого сообщения на безартиклевый язык. Аналитический язык также в целом более «многословен», чем синтетический. С другой стороны, вряд ли можно утверждать, что сообщение на английском языке всегда будет многословнее, чем аналогичное сообщение на русском языке, особенно в ситуации перевода с английского на русский.

И финский и русский — безартиклевые языки. В грамматике преобладают синтетические формы. Однако сходство на этом кончается. В русском языке существительные очень часто употребляются с предлогами. В финском языке конструкции с предлогами и послелогами употребляются

[42] В этом случае словоупотребление понимается как цепочка символов, ограниченная специальными символами, которые задаются списком (.,:;!?- и др.).

заметно менее активно, например, в русском языке локативные значения выражаются сочетанием существительного с предлогом, в то время как в финском языке аналогичные значения выражаются формой существительного без предлога в одном из локативных падежей. В финском языке очень много сложных слов, которым в русском языке нередко соответствуют словосочетания. Финская причастная конструкция (partisiippirakenne) в большинстве случаев переводится на индоевропейские языки придаточным предложением. Притяжательные суффиксы во многих случаях позволяют обходиться без употребления притяжательных местоимений. Финские частицы также функционируют как постфиксы, то есть пишутся слитно со знаменательным словом, после которого они употребляются. Все это делает текст на финском языке более компактным, по крайней мере, с точки зрения количества словоупотреблений.

Таким образом, сравнение длин текстов, написанных на разных языках, с использованием словоупотреблений в качестве меры длины, также мало о чем говорит.

Итак, остается только одно: измерять длины текстов в символах (под символом мы будем понимать букву, цифру, знак пунктуации или пробел). Однако и здесь остается открытым тот же вопрос. В разных языках для передачи одинаковых по смыслу сообщений используются разные слова, длина которых может быть как близкой, так и различной. Слова эти могут употребляться в различных грамматических формах, для чего используются морфемы или служебные слова, длины которых также могут различаться. Таким образом, для передачи одной и той же информации в разных языках могут порождаться сообщения разной длины. Утверждение о том, что текст на языке А длиннее текста на языке Б, потому что в первом тексте больше символов, представляется нам не совсем корректным.

Кроме корпуса «ПарРус», для проверки данных, полученных при анализе финско-русских текстов, использовались относительно небольшие по объему англо-русский и русско-английский массивы художественных текстов. В англо-русский корпус включены произведения английской литературы XIX-XX вв. (Диккенс, Мэри Шэлли, Стивенсон, Конан Дойл, Честертон, Хэммингуэй и др.) и их переводы на русский язык. Объем массива составил порядка 800 тыс. словоупотреблений на каждом из языков. Несколько больше по объему русско-английский корпус, в который были включены произведения русской литературы XIX-XX вв. (Гоголь, Лермонтов, Достоевский, Толстой, Булгаков и др.) и переводы этих произведений на английский язык, выполненные носителями английского языка. Объем этого массива составляет около 1,5 млн. словоупотреблений на каждом из языков.

Сравнение длин исходных текстов и их переводов в символах показала, что соотношение длин в значительной степени определяется парой языков. При переводе с русского на финский и с русского на английский длина перевода в символах оказывается больше, однако при переводе с англий-

ского на русский текст перевода оказывается заметно короче (результаты по некоторым текстам см. в табл. 18 — 20, полные данные по корпусу «ПарРус» см. в приложении 3 настоящей работы).

Таблица 18. Длина текстов в символах: русско-финские тексты

	Символы		
	Оригинал	Перевод	Отношение
Аксенов В., «Звездный билет», пер. Э. Адриан (Esa Adrian)	293833	329404	0,90
Бакланов Г., «Навеки девятнадцатилетние», пер. В. Орлов (Vappu Orlov)	312285	381108	0,82
Белов В., «Привычное дело», пер. Х. Лааксонен (Hugo Laaksonen)	287562	317852	0,91
Горький М., «Старуха Изергиль», пер. А. Митрошин (Anita Mitroshin)	40318	48885	0,83
Ерофеев В., «Москва — Петушки», пер. Э. Адриан (Esa Adrian)	205638	228171	0,90
Лермонтов М.Ю., «Герой нашего времени», пер. У.-Л. Хейно (Ulla-Liisa Heino)	263582	315782	0,84
Пастернак Б.Л., «Доктор Живаго», пер. Ю. Конкка (Juhani Konkka)	989604	1146650	0,86

Таблица 19. Длина текстов в символах: русско-английские тексты

	Символы		
	Оригинал	Перевод	Отношение
Булгаков М.А., «Мастер и Маргарита», пер. Р. Пивир и Л. Волконски (R. Pevear & L. Volokhonsky)	751334	832438	0,90
Гоголь Н.В., «Шинель», пер. И. Хэпгуд (Isabel F. Hapgood)	65095	67862	0,96
Достоевский Ф.М., «Преступление и наказание», пер. К. Гарнетт (Constance Garnett)	1073531	1118797	0,96
Ильф И., Петров Е., «Двенадцать стульев», пер. Дж. Ричардсон (John Richardson)	548410	591748	0,93
Лермонтов М.Ю., «Герой нашего времени», пер. Х. Уиздом и М. Муррэй (H. Wisdom & M. Murray)	263306	323782	0,81
Толстой Л.Н., Анна Каренина, пер. К. Гарнетт (Constance Garnett)	1697480	1938807	0,88

Таблица 20. Длина текстов в символах: англо-русские тексты

	Символы		
	Оригинал	Перевод	Отношение
G. K. Chesterton, «The man who was Thursday» («Человек, который был четвергом»), пер. Н.Трауберг	313400	247205	1,27
A. Conan Doyle, «The Hound of the Baskervilles» («Собака Баскервилей»), пер. Н. Волжина	312332	296092	1,05
Charles Dickens, «The posthuomous papers of the Pickwick club» («Посмертные записки Пиквикского Клуба»), пер. А.В. Кривцова и Е. Ланн	1717788	1641772	1,05
E. Hemingway, «Green hills of Africa» («Зеленые холмы Африки»), пер. Н.Волжина и В.Хинкис	349181	350835	1,00
Jerome K. Jerome, «Three Men in a Boat» («Трое в одной лодке»), пер. М. Донской и Э. Линецкая	353306	359177	0,98
M. Shelley, «Frankenstein, or the Modern Prometheus» («Франкенштейн, или современный Прометей»), пер. З. Александрова	418491	359255	1,16
R. L. Stevenson, «Treasure Island» («Остров сокровищ»), пер. Н. Чуковский	355737	333390	1,07

Таким образом, полученная нами статистика по параллельным текстам не говорит ни за, ни против гипотезы Найды. Для решения этого вопроса необходимо подойти к этому вопросу с другой стороны. Обратимся к математической теории информации. В математике понятие «информация» несколько отличается от бытового и связывается с количеством альтернатив при выборе очередного элемента сообщения. В качестве меры количества информации используется **энтропия**, степень свободы выбора на каждом шаге порождения сообщения. Таким образом, энтропия тесно связана с вероятностью. Для сообщения с алфавитом из *n* букв энтропия (H) вычисляется по формуле:

$$H = -\left[p_1 \log_2(p_1) + p_2 \log_2(p_2) + ... + p_n \log_2(p_n)\right],$$

где p_1, p_2, ... p_n — вероятности появления в сообщении различных символов (Oakes 1998: 58–60).

Для разных языков значение энтропии различается, поскольку даже если используется один и тот же алфавит (хотя полное совпадение алфавитов наблюдается довольно редко, например, почти во всех языках, использующих латиницу, либо есть дополнительные буквы, либо отсутствуют какие-либо из стандартных букв), частота букв и иных знаков (например, знаков препинания), а также их сочетаемость все же различается. В наших массивах были получены следующие значения энтропии: русский язык — 4,56 бит/символ (стандартное отклонение — 0,03), английский

язык — 4,27 бит/символ (стандартное отклонение — 0,04), финский язык — 4,20 бит/символ (стандартное отклонение — 0,03). Таким образом, энтропия всех трех языков отличается.

В наших экспериментах учитывались данные, полученные на достаточно длинных (более 4 000 словоупотреблений) текстах. Тем не менее, значения энтропии, полученные даже для очень коротких текстов корпуса, оказались удивительно стабильными. Например, энтропия рассказов Чехова, длина которых составляла менее тысячи словоупотреблений (самый короткий текст — рассказ «Неудача» — 468 словоупотреблений) не отклонялась от полученного нами среднего значения более чем на 0,05. Таким образом, по-видимому, значение энтропии не зависит от длины текста. Связи между энтропией оригинала и энтропией перевода также не прослеживается. В графике на рис. 31 по оси x отложены значения энтропии русских оригинальных текстов, по оси y — значения энтропии финских переводов. Как видно из полученного графика, высоким значениям энтропии исходного текста могут соответствовать низкие значения энтропии в переводах и наоборот, то есть никакой определенной тенденции не прослеживается.

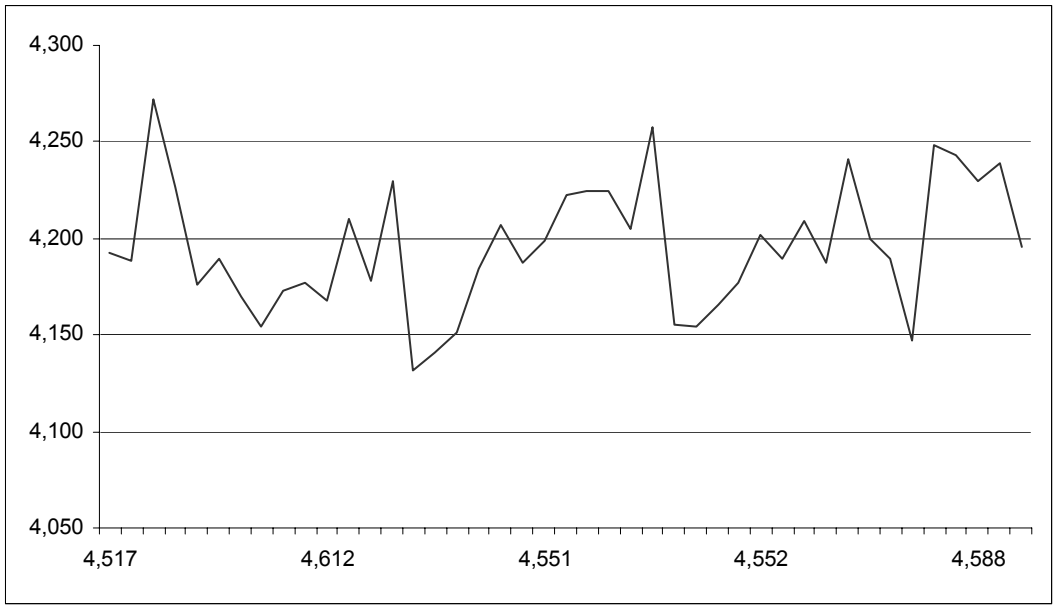

Рис. 31. *Энтропия оригинала vs. энтропия перевода: русско-финские тексты*

Представляется, что энтропия как мера «плотности» информации может помочь нам решить вопрос о соотношении длины оригинального текста и его перевода: ведь количество информации — как бы мостик между разными языками. Предположим, что энтропия влияет на длину сообщения в символах, порождаемого на данном языке. Тогда отношение длин текстов

на разных языках, содержащих одинаковый объем информации (ведь перевод в идеале должен содержать ту же информацию, что и оригинал), будет равно отношению энтропий языков, на которых написаны тексты:

$$k = \frac{n_1}{n_2} = \frac{H_1}{H_2}$$

Коэффициент *k*, равный отношению энтропий ИЯ и ПЯ, можно использовать в качестве поправки при сравнении длин оригинала и перевода:

$$p = k\frac{n_{sl}}{n_{tl}},$$

где *k* — отношение энтропий ИЯ и ПЯ, n_{sl} и n_{tl} — длины исходного текста и перевода (в символах).

Если $p = 1$, то оригинал и перевод «одинаковой длины», то есть различие в длине текста связано лишь со спецификой кодирования сообщения на данном естественном языке. Если $p > 1$, то исходный текст «длиннее», а в случае, если $p < 1$, то «длиннее» перевод.

По данным «ПарРус» коэффициент *k* для русско-финских текстов равен 1,087. Сравнение длин текстов корпуса показало, что переводы действительно «длиннее» оригиналов в том смысле, что длины текстов оказываются больше, чем они должны были бы быть, если бы в переводе заключалось столько же информации, как и в оригинале: значение коэффициента *p* почти для всех текстов оказывается меньше 1. Единственное исключение в нашем корпусе — роман И. Ильфа и Е. Петрова «12 стульев» и его перевод на финский язык, сделанный Р. Сильванто и Ю. Конкка ($p = 1{,}136$). Причина заключается в том, что перевод был сделан с первого издания романа, текст которого очень сильно отличается от издания 1956 г., представленного в «ПарРус». Ниже представлена часть полученных нами экспериментальных данных (см. табл. 21; в колонке n_{sl}/n_{tl} указывается отношение длины исходного текста к длине перевода, в столбце p — значение отношения с учетом поправки).

Таблица 21. Соотношение длин оригинальных текстов и их переводов: русско-финские тексты

	nsl/ntl	**p**
Аксенов В., «Звездный билет», пер. Э. Адриан (Esa Adrian)	0,90	0,97
Бакланов Г., «Навеки девятнадцатилетние», пер. В. Орлов (Vappu Orlov)	0,82	0,89
Белов В., «Привычное дело», пер. Х. Лааксонен (Hugo Laaksonen)	0,91	0,98
Горький М., «Старуха Изергиль», пер. А. Митрошин (Anita Mitroshin)	0,83	0,90
Ерофеев В., «Москва — Петушки», пер. Э. Адриан (Esa Adrian)	0,90	0,98
Лермонтов М.Ю., «Герой нашего времени», пер. У.-Л. Хейно (Ulla-Liisa Heino)	0,84	0,91
Пастернак Б.Л., «Доктор Живаго», пер. Ю. Конкка (Juhani Konkka)	0,86	0,93

Используя лишь данные по переводу с русского на финский, сложно оценить оправданность применения поправочного коэффициента и правильность полученных результатов, поскольку длины финских переводов и так больше длин исходных русских текстов. Полезность этого коэффициента становится очевидной при сравнении англо-русских текстов. Как уже говорилось выше, при переводе с английского на русский язык текст перевода, как правило, оказывается короче исходного текста. Применение поправочного коэффициента показывает, что переводы и в этом случае, как правило, оказываются «длиннее» исходных текстов (см. табл. 22).

Для русско-английских и англо-русских параллельных текстов коэффициент k равен соответственно 1,056 и 0,934. Результаты, полученные в ходе наших экспериментов, представлены в таблицах 22 и 23.

Таблица 22. Соотношение длин оригинальных текстов и их переводов: русско-английские тексты

	nsl/ntl	**p**
Булгаков М.А., «Мастер и Маргарита», пер. М. Глэнни (Michael Glenny)	0,96	1,01
Булгаков М.А., «Мастер и Маргарита», пер. Р. Пивир и Л. Волконски (R. Pevear & L. Volokhonsky)	0,90	0,95
Булгаков М.А., «Собачье сердце», пер. М. Глэнни (Michael Glenny)	0,88	0,93
Гоголь Н.В., «Шинель», пер. И. Хэпгуд (Isabel F. Hapgood)	0,96	1,01
Достоевский Ф.М., «Идиот», пер. Э. Мартин (Eva Martin)	0,98	1,03
Достоевский Ф.М., «Преступление и наказание», пер. К. Гарнетт (Constance Garnett)	0,96	1,01
Ильф И., Петров Е., «Двенадцать стульев», пер. Дж. Ричардсон (John Richardson)	0,93	0,98
Лермонтов М.Ю., «Герой нашего времени», пер. Х. Уиздом и М. Муррэй (H. Wisdom & M. Murray)	0,81	0,86
Толстой Л.Н., Анна Каренина, пер. К. Гарнетт (Constance Garnett)	0,88	0,92

Таблица 23. Соотношение длин оригинальных текстов и их переводов: англо-русские тексты

	nsl/ntl	**p**
G. K. Chesterton, «The man who was Thursday» («Человек, который был четвергом»), пер. Н.Трауберг	1,27	1,18
A. Conan Doyle, «The Hound of the Baskervilles» («Собака Баскервилей»), пер. Н. Волжина	1,05	0,99
Charles Dickens, «The posthuomous papers of the Pickwick club» («Посмертные записки Пиквикского Клуба»), пер. А.В. Кривцова и Е. Ланн	1,05	0,98
E. Hemingway, «Green hills of Africa» («Зеленые холмы Африки»), пер. Н.Волжина и В.Хинкис	1,00	0,93
Jerome K. Jerome, «Three Men in a Boat» («Трое в одной лодке»), пер. М. Донской и Э. Линецкая	0,98	0,92
M. Shelley, «Frankenstein, or the Modern Prometheus» («Франкенштейн, или современный Прометей»), пер. З. Александрова	1,16	1,09
R. L. Stevenson, «Treasure Island» («Остров сокровищ»), пер. Н. Чуковский	1,07	1,00

Как видно из таблиц 22 и 23, для русско-английских и англо-русских параллельных текстов результаты оказались не столь однозначными, как для переводов с русского языка на финский: в ряде случаев *k* несколько больше единицы. Скорее всего, в этих переводах допускались сокращения и незначительные пропуски в переводе для получения более естественного текста на ПЯ. Нередко более важным для переводчика оказывается сделать текст интересным и легко читаемым на языке перевода. В этих случаях переводчики могут опускать некоторые специфические культурные реалии и/или отказаться от передачи труднопередаваемых на ПЯ языковых особенностей исходного текста, что может вызвать в итоге сокращение текста перевода по сравнению с оригиналом.

Такая стратегия нередко использовалась в XIX веке. Например, в переводе на английский язык повести Пушкина «Капитанская дочка» («Marie, or Russian Love», Мари де Зиелинска (Marie H. de Zielinska), 1876) отсутствуют пушкинские эпиграфы к главам, а также довольно много пропусков по сравнению с исходным текстом.

Даже в настоящее время существует традиция сокращенного перевода, например, перевод классической литературы для детей школьного возраста. В качестве примеров можно привести переводы романов Дэфо и Свифта на русский язык. Перевод роман Гарриэтт Бичер-Стоу «Хижина дяди Тома» на русский язык на треть короче оригинального английского текста, несмотря на большое количество довольно подробных примечаний переводчика.

Итак, нам удалось получить косвенные подтверждения гипотезы Ю. Найды о том, что перевод длиннее оригинала. Общая тенденция, по на-

шим данным, действительно состоит в том, что перевод оказывается несколько более громоздким, чем оригинал. Тем не менее, следует сделать уточнение, что перевод оказывается длиннее оригинала лишь в том случае, если переводчик стремится, насколько это возможно, передать в переводе все особенности языка оригинала, сохранить все культурные реалии из исходного текста.

Сравнение длин исходного текста и перевода с учетом энтропий ИЯ и ПЯ, позволяет судить, насколько следует перевод тексту оригинала. Чем меньшее значение коэффициента получено, тем более бережно относился переводчик к исходному тексту. Однако результат такого поверхностного сравнения показывает только общую тенденцию в данном переводе: низкое значение коэффициента вовсе не означает, что в исследуемом переводе отсутствуют пропуски и приблизительные соответствия. Вопрос об определении количества потерь при переводе и возможности их минимизации остается открытым.

4.3. Влияние языка оригинала на язык перевода

Проблема влияния языка оригинала на язык перевода довольно часто рассматривалась в работах по теории перевода. Многие исследователи отмечают, что язык переводных произведений более эксплицитный, текст перевода, как правило, несколько длиннее текста оригинала (Nida & Taber 1974: 163; Chesterman 1997). Коллер (Koller 1992) обращает внимание на то, что язык переводов значительно уступает языку оригиналов в образности и метафоричности. Внимание исследователей привлекает также буквализм в переводе, с которым одни ученые призывают бороться всеми средствами (Nida & Taber 1974: 100, Бархударов 1975: 186), другие же считают, что буквальный перевод приемлем, если смысл передается корректно и стиль не воспринимается как неестественный (Newmark 1991). Переводы типа *keep off the grass — держитесь подальше от травы*, *regular ass — регулярный осёл* (примеры Бархударова (Бархударов 1975: 186), разумеется, остаются неправильными и неприемлемыми при любом подходе. Ньюмарк отмечает, что интерференция присуща любому переводу (Newmark 1991: 78).

В своей статье В.Н. Комиссаров подробно останавливается на проблеме естественности переведенного текста (Комиссаров 1992). Он констатирует неизбежность интерференции и считает, что язык переводов — «особая разновидность литературного языка», которая вовсе не должна совпадать с языком оригинальных литературных произведений.

В статье В.Н. Комиссарова упоминается эксперимент, в ходе которого информантам предъявлялись оригинальные и переводные тексты. Большинство респондентов легко идентифицировали переводы (Комиссаров

1992). В обзоре С. Эскола также говориться о целом ряде экспериментов такого типа, проводившихся на материале разных языков и с аналогичным результатом: участники экспериментов легко отличают переводной текст от оригинального текста (Eskola 2002: 6). Трудно предположить, что интуиция носителя языка не базируется ни на каких отличительных особенностях, характерных для языка переводов. Итак, существуют ли формальные признаки, по которым можно отличить переводной текст от текста оригинального? Попытаемся дать ответ на этот вопрос.

Использование русско-финских параллельных текстов представляется особенно интересным, поскольку финскими исследователями было установлено, что влияние русского языка в качестве ИЯ при переводе на финский язык достаточно сильно и более ярко выражено, чем, например, влияние английского языка в переводах с английского на финский (см. Eskola 2002: 235–239)[43].

Для проведения всех экспериментов, описывающихся в этом разделе, были взяты финские художественные тексты из Савонлиннского корпуса текстов (*The Savonlinna Corpus of Translated Finnish*. Savonlinna School of Translation Studies, University of Joensuu 2001, далее СХТ) и финский субкорпус «ПарРус» (далее ПФ). Оба массива являются близкими по объему (1,8 млн. словоупотреблений и 2,2 млн. словоупотреблений), в обоих корпусах представлены художественные тексты разные по объему и разных жанров. Единственный неблагоприятный для нашего сравнения момент — временной фактор. Все тексты СХТ опубликованы в 1990-х годах, в то время как переводы с русского на финский, включенные в ПФ, были сделаны в период с 1960-х по 1990-е гг. Поэтому в ходе нашего анализа нужно будет делать некоторую поправку на временной фактор, и полученные данные можно будет считать не доказанными фактами, а лишь тенденциями. Впрочем, как уже отмечалось ранее, данным из полнотекстовых корпусов текста никогда нельзя доверять на сто процентов.

Уже проверка таких количественных параметров, как средняя длина предложения и отношение количества словоупотреблений к количеству словоформ (type/token ratio, далее СС) показывают некоторую разницу между оригинальными финскими текстами и переводами с русского на финский.

В качестве одного из критериев богатства языка нередко используется отношение длины текста в словоупотреблениях к количеству словоформ (type/token ratio, СС). Можно использовать и отношение длины текста к количеству лексем, но в этом случае данные будут менее надежными, поскольку часть словоформ исследуемых корпусов текстов осталась нелемматизированными, особенно велико их количество для финских текстов (см. раздел о лемматизации в данной работе). По этой причине было ре-

[43] Причины этого до конца не установлены, по мнению Эскола, это скорее всего связано с традицией перевода с русского на финский.

шено остановиться на отношении количества словоупотреблений к количеству словоформ. Этот коэффициент достаточно связан и с лексическим репертуаром автора, и с разнообразием грамматических форм, которые он применяет в своих текстах. Значение коэффициента в какой-то степени зависит от объема текстов, но в случае исследования больших текстовых массивов зависимость эта оказывается не такой значительной, тем более, что исследуемые нами массивы СХТ и ПФ близки по объему. СС текстов СХТ оказался несколько выше, чем СС текстов ПФ (0,124 и 0,098). Это соответствовало нашей гипотезе, что лексический и грамматический репертуар автора, пишущего на родном языке, несколько больше, чем репертуар переводчика, скованного языком и стилем переводимого им текста.

Другой количественный показатель — средняя длина предложения (СДП). Этот параметр в какой-то степени можно назвать стилистическим, поскольку длина предложения косвенно связана с его синтаксисом: чем длиннее предложение, тем больше в нем различных распространителей, оборотов, придаточных предложений и других компонентов, «утяжеляющих» язык. Вычисляя средние длины предложений по разным текстам и среднее значение по сравниваемым массивам, мы ожидали, что СДП переводных текстов будет больше, поскольку русских писателей часто обвиняют в тяжелом стиле и длинных предложениях. Тем не менее, СДП по оригинальным финским текстам составил 10,89 слов в предложении, в то время, как в переводах с русского языка это значение было несколько меньше: 9,92 слова в предложении. Причем СДП русских текстов «ПарРус» оказалась ближе к текстам СХТ, чем к их же переводам, составив 10,99 слов в предложении. Объяснения этого явления, вероятнее всего, следует искать в функционировании грамматических форм и синтаксических конструкций в переводном тексте; вернемся к этому в соответствующем разделе (4.4.3).

4.3.1. Пунктуация в оригинале и переводе

Как уже говорилось в предыдущих разделах, переводчики, как правило, сохраняют в переводе структуру оригинального текста. В подавляющем большинстве случаев членение на абзацы остается без изменений, существует тенденция сохранять и членение на предложения, хотя с предложениями и обращаются более вольно. Эта тенденция используется для разработки алгоритмов стыковки параллельных текстов. Так, работая над программой-стыковщиком русских и финских текстов, мы использовали именно стремление сохранять в переводе членение на абзацы (см. раздел 3.4).

Интересным представляется проверить, насколько далеко заходит эта тенденция к буквализму в переводе. Членение текста тесно связано с пунктуацией, поэтому естественно предположить, что пунктуация ориги-

нала также накладывает свой отпечаток на перевод. Вполне возможно, что даже употребление запятых, точек с запятой и двоеточий в какой-то степени зависит от пунктуации исходного текста: хотя правила расстановки знаков препинания в каждом языке свои, переводчик может по инерции оставлять в своем переводе пунктуацию оригинала, если она допускается грамматикой и требованиями стиля в ПЯ.

В таблице 24 показаны различия в частотности употребления знаков препинания в русском и финском языках. Для каждого текста вычислялась относительная частота знака препинания на тысячу знаков. Из таблицы ясно, что разница в частотах является существенной, *t*-тест показывает, что результаты подсчетов являются значимыми на уровне 0,005 (для признания результатов теста значимыми достаточным было бы значение коэффициента 2,576).

Таблица 24. Пунктуация в «ТамРус» и СХТ

	Русские тексты		Финские тексты		*t-тест*
Знак препинания	Ср. арифм.	Станд. отклонение	Ср. арифм.	Станд. отклонение	
Запятая	20	3	10,05	2,18	29,17
Точка с запятой	0,23	0,32	0,06	0,19	5,30
Двоеточие	1,27	0,74	0,4	0,29	14,27
Восклицательный знак	1,87	1,33	0,64	0,65	10,22
Вопросительный знак	1,76	1,08	0,97	0,52	8,14

Сравним частоты знаков препинания в оригинальных и переводных финских текстах. Для наглядности возьмем оригинальные тексты и переводы, выполненные одним и тем же человеком. В ходе сравнения частотности знаков препинания в романе известного финского переводчика и писателя Эсы Адриана «Heinäsirkka kulkee kankeasti» и в переводах с русского языка на финский, выполненных им же, были выявлены следующие особенности. Двоеточия и вопросительные знаки в переводах Адриана встречаются в два раза чаще, чем в его оригинальном тексте. Двоеточие в оригинальном тексте употреблено лишь дважды, в то время как в переводах это достаточно распространенный знак препинания. Частотность точки с запятой в 21 раз, а двоеточия — в 257 раз выше в переводных текстах (табл. 25).

Таблица 25. Относительные частоты знаков препинания в романе Эсы Адриана и его переводах с русского языка на финский

	Оригинальный текст	**Переводы**
Запятая	4,629	10,947
Двоеточие	0,005	1,285
Точка с запятой	0,005	0,109
Восклицательный знак	0,258	1,623
Вопросительный знак	0,879	1,856

Продолжим наше сравнение, взяв для этого переводные финские тексты ПФ (31 текст) и оригинальные тексты СХТ (61 текст). Брались только тексты большого объема, превышающие 4000 словоупотреблений. Результаты обобщены в таблице 26. В целом, в переводных текстах знаков препинания больше. Проверка по *t*-тесту показывает, что результаты являются значимыми на уровне 0,005. Лишь статистика по запятым является значимой на уровне чуть более 0,05 (значение *t* для того, чтобы признать результаты значимыми на уровне 0,05, должно составлять 1,645).

Таблица 26. Пунктуация в СХТ и ПФ

	Оригинальные тексты		**Переводные тексты**		***t*-тест**
Знак препинания	**Ср. арифм.**	**Станд. отклон.**	**Ср. арифм.**	**Станд. отклон.**	
Запятая	10,05	2,18	10,73	1,73	1,628
Точка с запятой	0,06	0,19	0,21	0,27	2,765
Двоеточие	0,40	0,29	1,12	0,56	6,716
Восклицательный знак	0,64	0,65	1,32	0,73	4,379
Вопросительный знак	0,97	0,52	1,52	0,76	3,621

Причины интерференции

Итак, пунктуация оригинала оказывает некоторое влияние на пунктуацию перевода. В каждом языке — свои правила расстановки знаков препинания, и нередко пунктуационные нормы ИЯ конфликтуют с нормами ПЯ. Чем более расплывчаты правила расстановки знаков препинания, тем больше вероятность интерференции.

Так, употребление части знаков препинания, как правило, достаточно регламентировано, например, в большинстве случаев понятно, когда следует ставить запятую или двоеточие. Правила употребления точки с запятой или тире несколько более расплывчатые. А употребление точки, знака вопроса или восклицательного знака предполагается ясным и без правил.

Можно обозначить два типа влияния пунктуации оригинала на пунктуацию перевода — непосредственная и опосредованная интерференция.

Опосредованная интерференция является более распространенной. Она вызвана использованием в переводе синтаксических конструкций, нетипичных для ПЯ. Например, и в русском, и в финском языках запятая используется в качестве разделителя повторяющихся определений. Однако для русского стиля — особенно для художественного стиля — более типично использование цепочек определений, например, *Она была красива, тонкая, высокая, с длинной шеей, с немного большим ртом, с немного приподнятым носом* (А.Н. Толстой). В приведенном примере целых пять (!!) определений. Для финского языка, напротив, золотое правило «yksinkertainen on kaunista» («простота красива») и многочисленные определения не приветствуются. Однако, у переводчика с финского на русский нет выбора; ему приходится в большинстве случаев оставлять в переводе нетипичную для ПЯ конструкцию, как это произошло, например, в (3).

(3)

…Саша отвернулся, чтобы скрыть от гостей свое сердитое, отчаянное лицо, и сказал, придавая голосу радостное, благодушное выражение…	… Saša kääntyi sivuittain salatakseen vierailta vihastuneet, epätoivoiset kasvonsa ja antaen äänelleen iloisen, hyväntahtoisen sävyn lausui…

А.П. Чехов. Дачники. пер. Ю. Конкка

В результате, использование грамматических форм и синтаксических конструкций, нетипичных для ПЯ, приводит к изменениям в частотности знаков препинания, которые можно видеть в таблицах 25 и 26.

Точка с запятой в финском тексте, как это можно видеть из нашей статистики, — довольно редкий знак. Главная причина, по-видимому, состоит в том, что, как уже говорилось, длинные предложения, длинные списки, сложные синтаксические структуры и бессоюзие считаются признаками плохого стиля. Чтобы проиллюстрировать это, был проведен следующий небольшой эксперимент. Из небольшого фрагмента перевода с русского на финский (4) были убраны все знаки препинания. Затем носителю финского языка было предложено расставить в данном фрагменте знаки препинания. Единственное отличие от исходного образца было в том, что наш информант поставил в конце текста точку вместо восклицательного знака. Однако на словах он прокомментировал, что одно из предложений текста, содержащее две точки с запятой, несколько громоздко и сам он разбил бы его на три предложения.

(4)

Нет-с, я не хочу лечиться со злости. Вот этого, наверно, не изволите понимать. Ну-с, а я понимаю. Я, разумеется, не сумею вам объяснить, кому именно я насолю в этом случае моей злостью; я отлично хорошо знаю, что и докторам я никак не смогу "нагадить" тем, что у них не лечусь; я лучше всякого знаю, что всем этим я единственно только себе повредю и никому больше. Но все-таки, если я не лечусь, так это со злости. Печенка болит, так вот пускай же ее еще крепче болит!	Ei, nähkääs, häijyyttäni minä en halua mennä lääkärin hoidettavaksi. Te ette varmaankaan voi tätä ymmärtää, mutta minäpä ymmärrän. Minä en tietenkään kykene teille selittämään, kenelle minä tässä tapauksessa teen kiusaa ilkeydelläni; minä tiedän mainiosti, etten voi suututtaa lääkäreitä sillä, etten käy heidän luonaan; tiedän paremmin kuin kukaan muu, että kaikella tällä vahingoitan yksinomaan itseäni enkä kerrassaan ketään toista. Mutta sittenkin, jollen käy lääkärissä, niin se johtuu häijyydestäni. Maksaan koskee, no - antaapa koskea vielä kovemmin!

Ф.М. Достоевский. "Записки из подполья". Пер. В. Каллама.

Несмотря на то, что точка с запятой в финских текстах употребляется редко, нам встретились случаи, когда точка с запятой использовалась в переводе, но отсутствовала в оригинале. Например, в (5) переводчик заменил конструкцию с придаточным предложением на бессоюзное сложное предложение. Однако, по-видимому, главной причиной, по которой в переводе появилось бессоюзное предложение, являлось стремление переводчика сохранить структуру оригинала в переводе.

(5)

Николай Николаевич вез Воскобойникову корректуру его книжки по земельному вопросу, которую ввиду усилившегося цензурного нажима издательство просило пересмотреть.	Nikolai Nikolajevitš oli viemässä Voskoboinikoville tämän uuden, maakysymystä käsittelevän kirjan korjausvedoksia; sensuurin tiukentuneiden otteiden takia kustannusliike oli kehottanut tarkistamaan sen vielä kerran.

Б.Л. Пастернак. "Доктор Живаго". Пер. Ю. Конкка

В (6) переводчик изменил конструкцию с причастным оборотом на бессоюзное предложение, и вновь точка с запятой заменила запятую оригинала.

(6)

А теперь они молчали и еле дышали, подавленные бессмыслицей случившегося. Надя возмущалась и молча негодовала, а у Ники болело все тело, словно ему перебили палкою ноги и руки и продавили ребра.	Mutta nyt he vaikenivat ja hengittivät läähättäen; tapahtuman mielettömyys masensi heitä. Nadja raivosi ääneti. Nikalta kolotti joka paikkaa ikään kuin hänen jalkojaan, käsiään ja kylkiluitaan olisi hakattu kepillä.

Б.Л. Пастернак. "Доктор Живаго". Пер. Ю. Конкка

Непосредственная интерференция — буквальное копирование пунктуации оригинала в переводе. Это, как правило, случается тогда, когда в ПЯ (а иногда и в ИЯ) нет четкого правила, регламентирующего расстановку знаков препинания для данного случая. Непосредственная интерференция возможна даже в употреблении запятых, двоеточий и точек с запятой. В примере (7) запятые в первом предложении обязательны, они выделяют придаточное предложение. Однако в финском переводе запятые при обороте со вторым инфинитивом не нужны, хотя, по словам носителей

финского языка, с ними предложение воспринимается легче. Многоточия в обоих фрагментах не являются обязательными (первое предложение могло завершаться точкой, второе можно было бы не ставить), переводчик воспроизвел в переводе пунктуацию оригинала.

(7)

И в то время, когда я бодрил себя таким образом, я услышал тихие шаги... Кто-то медленно шел, но... то были не человеческие шаги... для человека они были слишком тихи и мелки...	Juuri sillä hetkellä, yrittäessäni tällä tavoin rohkaista itseäni, kuulin hiljaisia askeleita... Joku lähestyi hitaasti, mutta... ne eivät olleet ihmisen askeleita... ihmisen ottamiksi ne olivat liian hiljaisia ja lyhyitä...

А.П. Чехов. "Ночь на кладбище". Пер. Ю. Конкка

В примере (8) прямая речь в оригинале оформлена нетрадиционным способом: кавычек нет, прямая речь начинается с маленькой буквы. Причина состоит в том, что в данном случае прямая речь не является прямой речью в полном смысле слова, рассказчик лишь пытается угадать, о чем думает персонаж. Эта необычная пунктуация буквально воспроизводится в финском переводе, где она является столь же необычной, как и в русском тексте.

(8)

Извозчик косился на водку, потом на ехидное лицо няньки, и лицо его самого принимало не менее ехидное выражение: нет, мол, не поймаешь, старая ведьма!	Ajuri loi syrjäsilmäyksen vodkaan, sitten njanjan huhuihin kasvoihin, ja hänen omillekin kasvoilleen ilmestyi yhtä luihu ilme: ei, etpäs saakaan minua loukkuun, senkin vanha noita!

А.П. Чехов. "Кухарка женится". Пер. Ю. Конкка

Маркеры конца предложения

Хотя наши эксперименты и показывают, что влияние ИЯ может быть обнаружено в употреблении любого знака препинания, интерференция для маркеров конца предложения (МКП) является наиболее сильной.

Специфика МКП в том, что вообще трудно сформулировать четкие правила их употребления. Считается, что во всех языках, в которых есть пунктуация, МКП употребляются одинаково: точка обозначает утверждение, вопросительный знак — вопрос, восклицательный знак — восклицание. Хотя в целом это, по-видимому, и так, все же наша статистика (см. табл. 26) показывает, что частоты МКП в разных языках также различаются.

Точка не представляется привлекательным объектом исследования, поскольку это стандартный МКП. Если переводчик сохраняет членение на предложения, то он сохраняет и точки. Разумеется, в некоторых случаях МКП может быть изменен, например, переводчик может решить, что вопросительный или восклицательный знак будет в переводе выразительнее, чем бесцветная точка оригинала. В любом случае более интересным представляется исследовать особенности употребления вопросительных и восклицательных знаков.

Для того, чтобы сравнивать МКП параллельного корпуса, использовались три коэффициента:

1. **Относительное отклонение** (D) по частоте показывает общую тенденцию сохранения/замены знака препинания. Вычисляется по формуле:

$$D = \frac{|T - O|}{O},$$

где O — частота исследуемого знака в оригинальном тексте, T — частота знака в переводе.

Чем ближе D к 0, тем ближе использование данного МКП в оригинале и переводе, если пунктуация оригинала и перевода совпадает, D равен 0.

2. **Среднее относительное отклонение** (M) показывает точность воспроизведение МКП в параллельных фрагментах текстов. Вычисляется по формуле:

$$M = \frac{\sum d}{N},$$

где N количество параллельных фрагментов, d относительное отклонение по данному МКП в каждом из параллельных фрагментов.

$$d = \frac{|t - o|}{o},$$

где o частота данного МКП во фрагменте оригинального текста, t частота данного МКП в соответствующем фрагменте перевода.

Так же, как и для коэффициента D, чем последовательнее копируется пунктуация оригинала в переводе, тем ближе M к 0.

3. Другой полезной мерой для сравнения пунктуации может быть **коэффициент совпадения** (C), отношение количества параллельных фрагментов, в которых частота исследуемого МКП одинакова в оригинальном тексте и в переводе, к количеству фрагментов оригинального текста, в которых данный МКП употребляется. Коэффициент вычисляется по формуле:

$$C = \frac{T_c}{T_u},$$

где T_c количество фрагментов, в которых частоты данного МКП совпали, T_u количество фрагментов, в которых данный МКП употреблялся.

Чем больше совпадений, тем ближе C к 1. Если совпадение полное, т.е. T_c равно T_u, C равно 1. Если ни одного совпадения не было зафиксировано, C равно 0.

Статистика, полученная из текстов «ПарРус», наглядно показывает, что в большинстве случаев употребление МКП в переводах довольно близко к их употреблению в исходных текстах (см. табл. 27 и 28).

Таблица 27. Восклицательный знак в оригиналах и переводах

Автор	**Переводчик**	**Название**	**ИТ**	**ПТ**	**D**	**M**	**C**
В. Аксенов	E. Adrian	Звездный билет / Matkalippu tähtiin	641	646	0,008	0,013	0,916
В. Белов	H. Laaksonen	Привычное дело/ Tuttu tarina	368	369	0,003	0,093	0,474
Н.В. Гоголь	E. Adrian	Шинель / Päällystakki	30	29	0,033	0,009	0,905
Н.В. Гоголь	H. Jalkanen	Шинель / Päällystakki	30	55	0,833	0,065	0,476
Н.В. Гоголь	J. Konkka	Шинель / Päällysviitta	30	31	0,033	0,044	0,591
В. Гроссман	E. Adrian	Все течет / Kaikki virtaa	90	92	0,022	0,004	0,923
Ф.М. Достоевский	E. Adrian	Записки из подполья / Kirjoituksia kellarista	345	333	0,035	0,017	0,91
Ф.М. Достоевский	V. Kallama	Записки из подполья / Kellariloukko	345	377	0,093	0,071	0,702
Вен. Ерофеев	E. Adrian	Москва-Петушки / Moskova Petuški)	689	670	0,028	0,062	0,714
М.Ю. Лермонтов	U.-L. Heino	Герой нашего времени / Aikamme sankari	571	398	0,303	0,126	0,524

Таблица 28. Вопросительный знак в оригиналах и переводах

Автор	**Переводчик**	**Название**	**ИТ**	**ПТ**	**D**	**M**	**C**
В. Аксенов	E. Adrian	Звездный билет / Matkalippu tähtiin	856	836	0,023	0,012	0,943
В. Белов	H. Laaksonen	Привычное дело/ Tuttu tarina	575	475	0,174	0,08	0,658
Н.В. Гоголь	E. Adrian	Шинель / Päällystakki	32	34	0,063	0,009	0,909
Н.В. Гоголь	H. Jalkanen	Шинель / Päällystakki	32	37	0,156	0,037	0,636
Н.В. Гоголь	J. Konkka	Шинель / Päällysviitta	32	33	0,031	0,037	0,682
В. Гроссман	E. Adrian	Все течет / Kaikki virtaa	189	198	0,048	0,007	0,91
Ф.М. Достоевский	E. Adrian	Записки из подполья / Kirjoituksia kellarista	299	289	0,033	0,031	0,874
Ф.М. Достоевский	V. Kallama	Записки из подполья / Kellariloukko	299	274	0,084	0,057	0,746
Вен. Ерофеев	E. Adrian	Москва-Петушки / Moskova Petuški)	731	689	0,057	0,048	0,802
М.Ю. Лермонтов	U.-L. Heino	Герой нашего времени / Aikamme sankari	526	477	0,093	0,046	0,807

Приведем один типичный пример отношения переводчика к пунктуации оригинального текста:

(9)

Дальше сорока лет жить неприлично, пошло, безнравственно! Кто живет дольше сорока лет, - отвечайте искренно, честно? Я вам скажу, кто живет: дураки и негодяи живут. Я всем старцам это в глаза скажу, всем этим почтенным старцам, всем этим сребровласым и благоухающим старцам! Всему свету в глаза скажу! Я имею право так говорить, потому что сам до шестидесяти лет доживу. До семидесяти лет проживу! До восьмидесяти лет проживу!.. Постойте! Дайте дух перевести...	Elää kauemmin kuin neljäkymmentä vuotta on sopimatonta, säädytöntä, moraalitonta! Kuka elää yli neljänkymmenen vuoden - vastatkaapas vilpittömän rehellisesti? Minä sanon teille ketkä elävät: tyhmyrit ja lurjukset. Minä sanon sen suoraan vasten naamaa kaikille vanhuksille, kaikille noille kunnioitettaville, hopeahapsisille ja hyvätuoksuisille ukoille! Minä sanon sen koko maailmalle vasten naamaa! Minulla on oikeus näin puhua, koska itse elän kuusikymmenvuotiaaksi! Seitsenkymmenvuotiaaksi minä elän! Kahdeksankymmenvuotiaaksi elän! ... Hetkinen! Antakaa, kun vedän henkeä välillä...

Ф.М. Достоевский. "Записки из подполья". Пер. В. Каллама.

В переводе сохранены не только все вопросительные и восклицательные знаки, но даже все тире и многие запятые.

Итак, в качестве основного принципа перевода можно предложить следующий: «Если данное предложение можно перевести буквально и потери будут минимальными, то надо переводить буквально». Однако использовать этот принцип можно далеко не всегда. Наши данные показывают, что коэффициенты совпадения (С) для переводов на финский язык произведений В. Белова, Н.В. Гоголя и М.Ю. Лермонтова довольно низки.

Когда переводчик работает с классическим текстом, написанным более ста лет назад, копировать стилистические особенности текста становится довольно трудно из-за архаичного стиля, книжных слов и т.п. Трудно сохранить и формальную структуру текста, что неизбежно ведет к изменению пунктуации. Например, в примере (10) переводчик заменил часть восклицательных и вопросительных знаков на точки. Это может быть связано с тем, что финский текст с таким количеством восклицаний и вопросов выглядел бы слишком эмоциональным и был бы слишком необычным.

(10)

Он отвернулся и протянул ей руку на прощание. Она не взяла руки, молчала. Только стоя за дверью, я мог в щель рассмотреть ее лицо: и мне стало жаль - такая смертельная бледность покрыла это милое личико! Не слыша ответа, Печорин сделал несколько шагов к двери; он дрожал - и сказать ли вам? я думаю, он в состоянии был исполнить в самом деле то, о чем говорил шутя. Таков уж был человек, бог его знает! Только едва он коснулся двери, как она вскочила, зарыдала и бросилась ему на шею. Поверите ли? я, стоя за дверью, также заплакал, то есть, знаете, не то чтобы заплакал, а так - глупость!..	Petšorin ojensi kätensä jäähyväisiksi ja kääntyi sitten pois. Tyttö ei tarttunut ojennettuun käteen eikä sanonut sanaakaan. Seisoin oven takana ja saatoin sen raosta nähdä hänen kasvonsa. Minun tuli häntä sääli, kun näin kuolonkalpeuden peittävän nuo kauniit pikku kasvot. Kun Petšorin ei saanut mitään vastausta, hän astui muutaman askelen ovea kohti. Hän vapisi ja uskotteko, luulen, että hän oli sellaisessa mielentilassa, että hän olisi varmasti toteuttanut uhkauksensa. Sellainen mies hän oli, Luoja paratkoon! Tuskin hän kosketti ovenripaa, kun tyttö hypähti ylös, purskahti itkuun ja heittäytyi hänen kaulaansa. Uskotteko? Seisoessani siellä oven takana minäkin aloin itkeä, taikka, tuota, enhän minä nyt sentään itkenyt, muuten vaan - kaikenlaista joutavaa!..

М.Ю. Лермонтов. "Герой нашего времени". Пер. У.-Л. Хейно

Другая проблема для переводчика — разговорная речь и различные эксперименты автора произведения. Разговорные обороты почти невозможно переводить буквально, поэтому текст нередко реструктурируется, а изменения находят отражение в пунктуации. В примере (11) из повести В. Белова, разговорная русская речь переводится разговорной финской речью, что ведет к довольно существенным изменениям в синтаксисе. Из примера видно, что иногда восклицательный знак может «переводиться» вопросительным знаком, а вопросительный знак — восклицательным.

(11)

Я говорю, что Дрынова хто зажмет? Нихто Дрынова не зажмет. Дрынов сам кого хошь зажмет. Куда? Это ты куда, дурак старый, воротишь-то? Ведь ты не на ту дорогу воротишь! Ведь мы с тобой век прожили, а ты, понимаешь, куда воротишь? Это тебе домой дорога-то, что ли? Это тебе дорога не домой, а на вырубку. Я тут сто раз ездил, я тебе... Что? Я тебе полягаюсь, я вот тебе полягаюсь! Ты дорогу лучше меня знаешь? Ты, прохвост, вожжей захотел? Нна! Нна, вот тебе, ежели так! Ступай куда велят, свой прынцып не отстаивай! Чего заоглядывался? Ну? То-то, дурак, иди куда велено!	Sanokaas kuka Drynoville mitä mahtaa, hä? Ei kyllä kukaan. Drynov panee itse koville miehen kuin miehen. Hei, minnekä sinä nyt? Minne sinä käännät, vanha hölmö! Väärälle tielle! Ikä on yhdessä eletty ja sinä, tollo, käännät minne sattuu. Sitäkös tietä muka kotia päästään? Ei se vie, kuule kotiin, se vie hakkauksille. Minä olen ajanut tästä sata kertaa. Mitä? Vai vielä tässä potkimaan! Varropas! Vielä minä sinut potkitan! Vai olet sinä tietävinäsi paremmin? Varro kun saat ohjaksista, ryökäle! Kas noin, noin kun kerran haluat! Painu vain minne käsketään, äläkä tuittuile oman pääsi mukaan! Etkä kääntyile siinä! Noh! Sillä lailla, nyt menet minne käsketään.

В. Белов. "Привычное дело". Пер. Х. Лааксонен

Традиции художественного перевода постоянно изменяются. Главный вопрос: кто переводчик, раб или соперник. В истории перевода были периоды, когда считалось, что переводчик должен переводить как можно ближе к оригиналу (т.н. «принцип слова божьего»). Эти периоды сменялись периодами либерализма, когда приоритетом переводчика становились нужды и вкусы читателей, а не собственно сохранение формы (Chesterman 1997: 24). Большинство переводов, представленных в «ПарРус» были выполнены в 1970–80-е гг., однако есть несколько нескольких более ранних переводов в редакции 1960-х гг. (переводы произведений А.С. Пушкина, выполненные в начале века Й. Ахава (Juho Ahava) и П.А. Песоненом (Pekka Alarik Pesonen)). Анализ этих текстов позволяет сделать вывод о том, что в финском художественном переводе также происходила смена традиций. В старых переводах можно заметить гораздо более вольное обращение со структурой текста, что выражается и в употреблении МКП. Все коэффициенты получают довольно низкие значения (см. табл. 29).

Таблица 29. МКП в старых переводах

Автор	**Переводчик**	**Название**		**ИТ**	**ПТ**	**D**	**M**	**C**
А.С. Пушкин	П.А. Песонен	Пиковая дама / Patarouva	Воскл. знак	91	17	0,813	0,221	0,155
			Вопр. знак	69	54	0,217	0,059	0,638
А.С. Пушкин	Й. Ахава, В. Хямеен-Анттила	Барышня-крестьянка / Aatelisneiti talonpoikaistyttönä	Воскл. знак	40	44	0,1	0,132	0,407
			Вопр. знак	56	40	0,286	0,112	0,37

Из приведенного ниже в качестве примера параллельного фрагмента из "Пиковой дамы" А.С. Пушкина и ее финского перевода ясно видно, что употребление пунктуации в финском тексте гораздо менее эмоционально, чем в русском оригинале.

(12)

— А каков Германн! — сказал один из гостей, указывая на молодого инженера: — отроду не брал он карты в руки, отроду не загнул ни одного пароли, а до пяти часов сидит с нами, и смотрит на нашу игру!	— Entäs tuo Herman, virkkoi eräs vieras osoittaen nuorta insinööriä. Ei ole ikinä ottanut kortteja käteensä, ei elämässään ole kertaakaan kaksinkertaistanut panosta, mutta viiteen asti vain on seurassamme istunut peliämme katsellen.
А.С. Пушкин. "Пиковая дама". Пер. П.А. Песонен	

Если сравнивать переводы с русского языка на финский, выполненные в 1950 — 1970-х гг., то можно заметить, что более поздние переводы оказываются более буквальными. Например, по трем переводам гоголевской «Шинели» на финский язык ясно видно, что ранние переводы менее последовательны в воспроизведении вопросительных и восклицательных знаков, чем поздние (см. табл. 27 и 28). Трудно сказать, связано ли это с разными переводческими традициями или с индивидуальными стилями разных переводчиков.

Частью стиля переводчика является степень близости его переводов к переводимым им текстам. Одной из мер такой близости являются использованные нами коэффициенты, описывающие употребление знаков препинания. Было проведено сравнение переводов на финский язык, сделанных тремя известными переводчиками: Э. Адрианом, Ю. Конкка и У.-Л. Хейно. Для каждого из текстов по данной ниже формуле вычислялся сводный индекс МКП (I).

$$I = 1000\left(\frac{M_q}{C_q} + \frac{M_e}{C_e}\right),$$

где M_q, C_q — среднее относительное отклонение и коэффициент совпадения для вопросительных знаков, а M_e, C_e — среднее относительное отклонение и коэффициент совпадения для восклицательных знаков.

Наши данные показывают, что пунктуация большинства переводов Э. Адриана намного ближе к исходным текстам, чем пунктуация переводов Ю. Конкка или У.-Л. Хейно. Среднее значение I для переводов Э. Адриана составило 42,26, переводы У.-Л. Хейно несколько дальше от оригиналов, среднее значение I равно 79,80. Наибольших значений индекс достигает в переводах Ю. Конкка (152,40). Конечно, следует учитывать, что и язык переводимых текстов может влиять на стиль перевода и, как следствие, на употребление знаков препинания. Большинство переводов Ю. Конкка — литература XIX века: Толстой, Чехов, Гоголь. Единственный включенный в «ПарРус» перевод произведения литературы XX века, выполненный Ю. Конкка, — «Доктор Живаго» Б. Пастернака. Тем не менее, даже для этого текста значение индекса весьма высоко — 98,125. И Конкка, и Адриан перевели «Шинель» Гоголя. Индекс МКП для перевода Адриана равен 19,846, для перевода Конкка значение индекса намного выше — 128,702.

Выводы

Сравнивая пунктуацию оригинала и перевода, можно узнать, насколько формальная структура оригинального текста сохранена в переводе. Статистические данные «ПарРус» показывают, что в пунктуации большинства переводов ощущается влияние пунктуации исходных текстов. Степень, в которой переводчик копирует пунктуацию оригинала, зависит от следующих факторов:

1. временной фактор: формальную структуру старых текстов, написанных более 100 лет назад, сохранить труднее;
2. фактор традиции: финские переводчики первой половины XX века менее последовательно повторяют пунктуацию оригинала;
3. особенности авторского стиля (например, длинные абзацы, сложный синтаксис и т.п.) делают трудным копирование структуры текста;
4. индивидуальный стиль переводчика: например, переводы Э. Адриана ближе к оригиналу, чем переводы У.-Л. Хейно или Ю. Конкка.

Тенденция к сохранению формальной структуры оригинала в переводе может оказаться очень полезной в прикладной и компьютерной лингвистике (например, при разработке систем машинного перевода или стыковщиков параллельных текстов). Степень, в которой сохраняется формальная структура исходного текста, может помочь разделить разные переводческие традиции и даже различать индивидуальный стиль разных переводчиков.

4.3.2. Лексика оригинальных и переводных текстов

Отличается ли лексика, используемая в оригинальных финских текстах, от лексики, используемой в переводах с русского языка? Несомненно. При переводе с иностранного языка вместе с текстом, принадлежащим к другой культуре, в ПЯ неизбежно проникают тесно связанные с ней слова. Некоторые понятия могут просто отсутствовать в ПЯ, некоторые присутствуют, но не являются в ней столь важными. Другие понятия отсутствуют в ИЯ, и это также не может остаться незамеченным в переводах.

Особый вопрос, насколько существенны отличия оригинального языка от языка переводов. Это можно выяснить только путем исследования эмпирического материала. Продолжим сравнение художественных текстов из Савонлиннского корпуса (СХТ) и финского субкорпуса «ПарРус» (ПФ). Рассмотрим качественные и количественные различия лексики этих массивов.

Качественные различия

Для выяснения, чем качественно отличаются друг от друга словники СХТ и ПФ, было произведено «вычитание» словников и получены списки слов СХТ, отсутствовавших в ПФ («СХТ — ПФ»), и слов ПФ, отсутствовавших в СХТ («ПФ — СХТ»). Из списков были убраны имена собственные (фрагменты словников см. в приложениях 6.1 и 6.2). Словник «СХТ-ПФ» оказался значительно больше, чем словник «ПФ-СХТ»: первый составил 13 400 слов, второй — лишь 7 600. Это косвенно подтверждает наше предположение, что точкой отсчета во всех случаях является язык оригинальных произведений.

Рассмотрим наиболее частотные слова двух списков.

«СХТ — ПФ»

Значительную часть списка составляют слова, тесно связанные с бытом и обществом Финляндии: *mämmi* ’традиционное пасхальное блюдо’, *joulupukki* ’дед Мороз’, *kunnanjohtaja* ’председатель kunta — муниципального образования в Финляндии’, *terveyskeskus* ’центр медицинского обслуживания’, *kesämökki* ’летний домик’, *ahkio* ’сани’, *viili* ’финская простокваша’ и др. Появление этих слов в переводах с русского языка маловероятно, поскольку они по большей части обозначают финские реалии. Функциональное сходство некоторых понятий, как, например, *joulupukki* и *дед Мороз*, *kesämökki* и *дача*, может использоваться в некоторых видах перевода, однако это ведет к отождествлению понятий, не являющихся полными аналогами, что не очень желательно в художественном переводе.

Часть слов, не будучи «привязанными» именно к финской культуре, все же обозначают понятия, достаточно типичные для финского образа жизни: *viikonloppu* ’уик-энд’, *pizza* ’пицца’, *pizzeria* ’пиццерия’, *makuupussi*

’спальный мешок’, *huoltoasema* ’станция обслуживания’, *kokko* ’костер’, *muovipussi / muovikassi* ’полиэтиленовый пакет’, *sarjakuva* ’комикс’. Часть этих понятий — таких, как *уик-энд*, *пицца*, *полиэтиленовый пакет* и т.п. — уже получила распространение в русском языке и не попала в тексты рассматриваемых переводов лишь потому, что там очень мало произведений последнего десятилетия (в корпус было включено лишь несколько рассказов Татьяны Толстой).

Очень много жаргонизмов, например, *mutsi* ’мать’, *kännykkä* ’мобильный телефон’, *bileet* ’праздник, вечеринка’, *ryssä* ’русский (презр.)’, *baarimikko* ’бармен’, *leffa* ’кинофильм’, *bändi* ’поп-группа’, *telkkari* ’телевизор’, *duuni* ’работа’. Часть этих слов обозначает новые понятия, появившиеся в течение последнего десятилетия, некоторые имеют довольно ярко выраженный финский колорит (в самом деле, трудно представить себе, чтобы в переводе с русского на финский кто-нибудь назвал московского бармена *baarimikko*). Однако отсутствие некоторых слов может быть объяснено либо случайностью, либо боязнью переводчика сделать язык слишком разговорным и не воспринимающимся как перевод, либо стремлением переводчиков писать нормативным языком, либо работой литературного редактора. Так, в приведенном в качестве примера фрагменте из рассказа В. Шукшина встречаются разговорные и диалектные слова, которые остались непереведенными: особенности речи персонажей передаются другими средствами.

(13)

— Садись. Чайку щас поставим.	— Istu. Laitetaan teevesi tulelle.
— Отогреюсь малость... — Выговор у парня нездешний, расейский. Старика разбирало любопытство, но вековой обычай — не лезть сразу с расспросами — был сильнее любопытства.	— Lämmittelen vähäsen... Nuorukaisen puheenparsi ei ollut täkäläinen. Vanhuksen valtasi uteliaisuus, mutta ikivanha tapa — olla sekaantumatta heti toisen asioihin kyselemällä — oli uteliaisuutta voimakkaampi.

В. Шукшин, «Охота жить», пер. Р. Рюмина и П. Паркинена

«ПФ — СХТ»

В переводах с русского язык на финский, как и ожидалось, много лексики, связанной с Россией; эта лексика функционирует в переводах как экзотизмы, например, *kopeekka* ’копейка’, *samovaari* ’самовар’, *husaari* ’гусар’, *kuvernementti* ’губерния’, *kulakki* ’кулак’, *rajoni* ’район’, *ajomies* ’ямщик’. Значительная часть этих слов является также историзмами.

Однако часть слов отсутствует в СХТ лишь потому, что эти понятия не являются столь важными в финской картине мира, как в русской, например, *junamies* ’проводник’, *aamunkoitto* ’рассвет’, *porrasaskelma* ’ступенька’, *traktorinkuljettaja* ’тракторист’, *porttiholvi* ’арка’, *matruusi* ’матрос’. Отметим, что в текстах «ПарРус» слово *junamies* в большинстве контекстов использовалось в качестве эквивалента к русскому *проводник* (в поезде), однако в ФРС для слова *junamies* предлагается эквивалент *помощник кон-*

дуктора, БФРС дает для проводник финские эквиваленты *vaunumies* и *vaunuhoitaja*, а Perussanakirja толкует *junamies* как «работник железной дороги, выполняющий различные работы на станции». Противоречия между нашими лексикографическими источниками вызваны прежде всего тем, что в финских поездах проводников в русском понимании этого слова не существует (во всяком случае — в современной Финляндии). Другое интересное слово — *porttiholvi*. В финском языке это слово обычно обозначает ворота в виде арки, например *linnan porttiholvi* ’ворота зáмка’. В текстах «ПарРус» оно используется для перевода слова *арка* в значении ’сквозной проход’, где в качестве эквивалентов лучше подходят слова *porttikäytävä*, или *kaarikäytävä*.

Некоторая часть представленной в ПФ лексики отсутствует в СХТ по той причине, что в Савонлиннском корпусе отсутствуют тексты, аналогичные по тематике. Например, в «ПарРус» довольно много произведений о войне, в СХТ таких произведений нет, поэтому в список попало довольно много лексики, связанной с войной и армией, например, *taisteluhauta* ’траншея, окоп’, *rykmentinkomentaja* ’командир полка’, *aliupseeri* ’унтер-офицер’ и др.

Таким образом, различия в словниках двух массивах имеют место, и наши соображения по поводу того, что может отсутствовать в переводных текстах или в оригинальных текстах подтвердились. Однако никаких неожиданных фактов установлено не было. Больший интерес представляет не поиск лакун, а исследование того, что представлено в обоих текстовых массивах, но представлено по-разному.

Количественные различия

Даже простое сравнение частотности слов, наиболее употребительных в анализируемых корпусах текстов, позволяет обнаружить некоторые различия в исследуемых массивах текстов. Сравнение ста самых частотных слов из лемматизированных словников сравниваемых массивов дало довольно интересные результаты, показывающие влияние ИЯ на ПЯ при переводе. Два полученных списка были объединены в один сводный словник; частоты для слов, отсутствовавших в каком-либо из списков, были дополнены из соответствующих полных лемматизированных словников (полный сводный список, отсортированный по частоте слов в СХТ, см. в приложении 6.1.). Слов, которые присутствовали бы в списке самых частотных слов СХТ или ПФ и отсутствовали в полном словнике другого текстового массива, зафиксировано не было. Первые 35 слов из двух списков самых частотных слов присутствовали в обоих списках, далее фиксируются лакуны в ПФ относительно СХТ, и лишь в конце списка появляются лакуны в СХТ относительно ПФ. Таким образом, не СХТ отличается от ПФ, а ПФ отличается от СХТ, то есть оригинальные художест-

венные тексты являются стандартом, от которого в той или иной степени отклоняется язык переводных текстов в целом или язык переводов с какого-либо определенного языка.

Проведем выборочное сравнение для некоторых слов, представленных в полученном нами списке.

Поскольку используемые нами корпуса текстов не являются аннотированными, возможности для сравнения списков ограничены. Наблюдать функционирование слов можно по различным косвенным проявлениям. Например, некоторую информацию дает информация по частотности употребления лексем и/или их различных форм. Другой важный инструмент анализа — коллокации. Ближайшее окружение слова является своего рода барометром, который показывает, в каких значениях употребляется это слово, а также какие синтаксические конструкции для него типичны. Для получения наборов коллокаций использовалась соответствующая утилита пакета «КОКОС-П». Эта утилита позволяет анализировать контекст, равный десяти словам: пять слов влево от анализируемого, пять слов вправо, если предложение заканчивается раньше, следующие после конца предложения слова коллокатами не считаются (см. раздел 3.1.6 наст. работы). На основе данных по частотности и коллокациям можно формулировать гипотезы, которые проверяются уже путем построения параллельных конкордансов.

Местоимения. Личное местоимение *hän* ’он/она’ более частотно в переводных текстах (19,25 в СХТ и 32,83 в ПФ[44]), в то же время местоимение *se* ’оно’ более частотно в оригинальных текстах (11,87 против 7,70). Это наглядно показывает тенденцию к употреблению местоимения *se* для обозначения людей, характерную для современного разговорного финского языка. Такое употребление становится вполне распространенным и в СХТ. Приведем один пример: *Se ajoi niin kovaa, ettei huomannut Hartikaisen tiehaaraa*… (Antti Tuuri, «Suomi elää metsistään»). В ПФ контекстов такого типа найти не удалось.

Аналогично выглядит статистика по местоимениям *he* ’они (люди)’ (3,24 и 4,77) и *ne* ’они (вещи и «нелюди»)’ (5,01 и 3,69). Причина, по-видимому, та же: в оригинальных финских текстах местоимение *ne* может употребляться и в тех случаях, когда говорят о людях, например, *Jos minusta tulee rosvo ja pahantekijä, joudun olemaan Manalassa. Telkkarissa sanottiin, että kuolleet joutuu sinne, jos ne ei ole olleet eläessään kunnolla* (Jorma Ranivaara, «Rouva Korhosen tapaus»). В субкорпусе переводов удалось найти примеры такого употребления в переводе «Записок из подполья» Ф.М. Достоевского, выполненного Э. Адрианом:

[44] Здесь и далее будут приводиться относительные частоты на 1000 словоупотреблений.

(14)

Или они все на коленах, обнимая ноги мои, будут вымаливать моей дружбы, или... или я дам Зверкову пощечину!	Joko ne ryömivät jalkojeni juuressa ja polvillaan rukoilevat ystävyyttäni, tai... tai minä isken Zverkovia korvalle!

Ф.М. Достоевский, «Записки из подполья», пер. Э. Адриан

Ср. другой перевод того же места, сделанный В. Каллама: *Joko he polvillaan ryömien, jalkojani syleillen rukoilevat ystävyyttäni tai... tai minä annan Zverkoville korvapuustin!* Употребление местоимения *ne* вместо *he* в переводе Адриана показывает крайне негативное отношение героя к людям, о которых он говорит. В целом все же контексты такого типа более типичны для оригинальных финских текстов.

Бросаются в глаза и отличия частотности употребления местоимения второго лица *te* 'вы' (1,28 и 4,83). Это финское местоимение, так же как и в русском языке, используется как для вежливого обращения на «вы», так и как местоимение 2-го лица множественного числа. В финском языке обращение на «вы» за последние несколько десятилетий стало использоваться только в очень официальных ситуациях. Это и отразилось в нашей статистике: подавляющее количество употреблений местоимения *te* в текстах СХТ — во множественном числе. В то же время обращение на «вы» весьма характерно для переводных текстов (15).

(15)

Давайте сядем вот здесь на лавку, вам придется писать. Вот вам мой блокнот... Вы меня слушаете? Вы думаете о чем-то другом.	Istutaanko tähän penkille, sillä teidän kannattaa tehdä muistiinpanoja. Tässä on teille minun muistikirjani... Kuunteletteko minua? Te ajattelette nyt jotain muuta.

(В. Дудинцев, «Белые одежды», пер. У.-Л. Хейно)

Существительные. Прослеживаются также некоторые различия и в частотности существительных. Например, частотность слов *mies* 'мужчина, муж' и *nainen* 'женщина' выше в оригинальных текстах (*mies*: 3,37 и 2,34, *nainen*: 2,13 и 1,26), а слово *ihminen* 'человек', наоборот, более типично для переводных текстов (1,88 и 2,09). Это связано с тем, что, в связи с отсутствием в финском языке не только грамматического рода, но и местоимений мужского и женского рода, слова *mies* и *nainen* в целом употребляются в финском языке чаще, чем русские слова *мужчина* и *женщина*, частоты которых в русском субкорпусе «ПарРус» составляют, соответственно, 0,19 и 0,63 на 1000 словоупотреблений (в русском языке слово *женщина* является более частотным, а в финском более частотно слово *mies*, что связано с тем, что у слова есть второе значение — 'муж'). Низкая частотность слов *мужчина* и *женщина* в русских текстах вызывает некоторое снижение частотности слов *mies* и *nainen* по сравнению с оригинальными финскими текстами.

Рассмотрим несколько примеров из «ПарРус».

(16)

С виду я не похож на мужика, это точно, потому мне такая судьба вышла, добрый человек.	Näöltäni en muistuta talonpoikaa, se on totta, sillä minulla on oma kohtaloni, hyvä mies.

А.П. Чехов, «Мечты», пер. Ю. Конкка

В данном примере в качестве эквивалента к слову *человек* используется слово *mies*, если бы использовался прямой эквивалент *ihminen*, было бы неясно, какого пола человек, к которому обращается герой. Кроме того, оборот *hyvä mies* является стандартным обращением.

(17)

И он начинает говорить со мной по-алайски, но уже не хриплым басом, а приятным таким, нормальным голосом — тенором или, я не знаю там, баритоном.	Ja mies alkaa puhua minulle alain kielellä mutta ei enää käheällä bassolla vaan miellyttävällä normaalilla äänellä — tenoriko se on vai baritoni, en tiedä tarkalleen.

А. и Б. Стругацкие, «Парень из преисподней», пер. Э. Адриан

Здесь слово *mies* используется для перевода местоимения *он*. Использование местоимения *hän* ’он/она’ во многих случаях может затруднить понимание текста, особенно если в эпизоде действует несколько лиц. Иностранные имена также не очень помогают финскому читателю. Поэтому достаточно часто практикуется перевод русских местоимений финскими существительными. Приведем еще один аналогичный пример.

(18)

Прибежал учитель, молодой еще человек, уважаемый в деревне.	Opettaja, nuori mies, arvostettu kylässä, tuli juosten.

В.М. Шукшин, «Крепкий мужик», пер. Э. Адриан

В Савонлиннском корпусе не зафиксировано ни одного примера на сочетание *nuori ihminen*. При этом достаточно частотны сочетания *nuori mies* и *nuori nainen*.

Несмотря на продемонстрированные выше специальные приемы перевода, зафиксированная нами более низкая частотность в переводах с русского языка слов *mies* и *nainen* при более высокой частотности слова *ihminen* указывает на более частое использование переводчиками личных местоимений (на это указывает и статистика по местоимению *hän*, которое в переводах встречается почти в два раза чаще) и имен собственных, а также буквального эквивалента слова *человек* — *ihminen*. Отметим, что выявленная особенность может в некоторой степени затруднять понимание переводного текста.

Например, использование слова *ihminen* ’человек’ и личного местоимения *hän* при переводе с русского языка нередко приводит к определенным смысловым потерям. Так, в следующем отрывке из «Мастера и Маргариты», при переводе (во всяком случае, в рассматриваемом контексте) оказывается неясным, идет ли речь о *буфетчике* или *буфетчице*.

(19)

— Это низко! — возмутился Воланд, — Вы человек бедный... Ведь вы — человек бедный? Буфетчик втянул голову в плечи, так что стало видно, что он человек бедный.	— Sehän on alhaista! Woland kiihtyi. — Te — köyhä ihminen... Olettehan te köyhä ihminen? Kahvilanpitäjä veti päätään hartioiden väliin, niin että selvästi huomasi hänen olevan köyhän ihmisen.

М.А. Булгаков, «Мастер и Маргарита», пер. У.-Л. Хейно

Другой интересный факт, полученный при сравнении частотных списков — частотность слова *isä* ’отец’. Частотность этого слова в СХТ (2,32) намного превышает частотность того же слова в ПФ (0,76), где это слово не попадает даже в первую сотню, несмотря на то, что в корпус включен роман Ильфа и Петрова «Двенадцать стульев», один из персонажей которого, как известно, *отец Федор* (*isä Fjodor*). Аналогичная закономерность прослеживается и со словом *äiti* ’мать’ (2,43 против 0,88). Исследование частотности этих слов по разным текстам показали, что они распределяются неравномерно как в оригинальных, так и в переводных текстах. Например, частотность слова *isä* в СХТ варьируется от 12,41 на 1000 словоупотреблений (S. Puskala, «Pikkuruu Mustanmusta») до 0,07 (S. Jokinen, «Koskinen ja taikashow»); в «ПарРус» частотность того же слова колеблется от 17,24 (А.П. Чехов, «Устрицы», пер. Ю. Конкка) до 0 (А.П. Чехов, «Дама с собачкой», пер. Ю. Конкка, В.М. Шукшин «Ноль-ноль целых», пер. Э. Адриан). Необходимо отметить, что примерно треть текстового массива СХТ составляет детская литература (500 тыс. словоупотреблений), в «ПарРус» детской литературы пока вообще нет, а доля произведений о детях невелика: несколько рассказов А.П. Чехова и В.М. Шукшина, повесть А. Приставкина «Ночевала тучка золотая» и некоторые другие. Поэтому тексты с высокой относительной частотностью слов *isä* и *äiti*, в основном, — небольшие по объему. Таким образом, различия в частотности этих слов, скорее всего, не связаны с какими-либо особенностями языка переводов, а объясняются тематикой текстов, включенных в корпус.

Сравнение частот остальных ста самых частотных существительных, входящих в список первых ста самых частотных слов, не выявило никаких существенных отличий, выходящих за пределы случайных колебаний.

Прилагательные. С полученном нами списке зафиксировано всего три прилагательных (местоимения-прилагательные в расчет не принимались): *hyvä* ’хороший’, *pieni* ’маленький’ и *suuri* ’огромный / великий’. Частоты для двух из них — *hyvä* и *pieni* — различаются довольно сильно. Прилагательное *hyvä* более частотно в ПФ, *pieni* — в СХТ.

Прилагательное *hyvä* ’хороший’.

Для того, чтобы выяснить причины большей частотности слова *hyvä* в переводных текстов, вначале попробуем проследить, определением к каким существительным чаще всего является прилагательное *hyvä*. Поскольку главное слово для прилагательного в финском языке почти всегда находится справа от прилагательного, рассмотрим первый и второй правые кол-

локаты этого слова (далее для краткости будем обозначать правые коллокаты латинской буквой R, а левые — буквой L, и цифрой, указывающей позицию, например, R1 — первый правый коллокат).

Вообще, бросается в глаза, что набор коллокатов для *hyvä* в ПФ намного богаче, чем в СХТ: в ПФ зафиксировано 370 коллокатов с общей частотой более 10, а в СХТ — лишь около 200. Среди коллокатов, зафиксированных в ПФ, довольно много существительных. В самом начале списка оказываются стандартные обороты, например, *hyvä jumala* ’бог’ (R1 — 167, R2 — 2, всего — 205), *hyvä herrasväki* ’господа’ (R1 — 77, R2 — 0, всего — 79), *hyvät herrat* ’господа’ (R1 — 73, R2 — 1, всего — 83), *hyvä ihminen* ’человек’ (R1 — 72, R2 — 5, всего — 103), *hyvä luoja* ’Создатель’ (R1 — 63, R2 — 0, всего — 71), *hyvä ystävä* ’друг’ (R1 — 62, R2 — 1, всего — 91), *hyvä herra* ’господин’ (R1 — 60, R2 — 4, всего — 85), *hyvä mies* ’мужчина’ (R1 — 51, R2 — 3, всего — 95).

В СХТ набор коллокатов-существительных к *hyvä* намного беднее, из перечисленных выше частотных в ПФ сочетаний, в СХТ были зафиксированы лишь следующие: *hyvä mies* ’мужчина’ (R1 — 25, R2 — 7, всего — 48), *hyvä ihminen* ’человек’ (R1 — 19, R2 — 1, всего — 28), *hyvä jumala* ’бог’ (R1 — 18, R2 — 0, всего — 27), *hyvä herra* ’господин’ (R1 — 16, R2 — 0, всего — 22), *hyvä ystävä* ’друг’ (R1 — 9, R2 — 1, всего — 23). Обращает на себя внимание тот факт, что среди коллокатов, зафиксированных в СХТ, нет таких, которые были бы действительно частотны, в то время как в ПФ у всех перечисленных коллокатов частота в позиции первого правого коллоката превышает 50.

Совершенно очевидно, что обнаруженные нами выражения являются стандартными эквивалентами к русским клише. Для разного рода божбы — *боже мой, господи, боже правый*, *о боже* и др. — в качестве эквивалентов используются финские обороты *hyvä jumala* и *hyvä luoja*. Уважительные обращения типа *милостивый государь, сударь, батюшка* переводятся как *hyvä herra*; обращения к группе людей *господа* и *уважаемые господа* — *hyvät herrat*, *дорогой друг — hyvä ystävä*.

Примеры:

(20)

— Володечка приехали! — завопила Наталья, вбегая в столовую. — Ах, боже мой!	— Voloditška on tullut! kiljui Natalja juosten ruokasaliin. — Ah, hyvä Jumala!
А.П. Чехов, «Мальчики», пер. Ю. Конкка	

(21)

— Ах, боже мой... — говорил он. — Ах, господи помилуй.	— Aah, hyvä Luoja . . . hän puheli. — Aah, Herra armahtakoon.
	А.П. Чехов, «Крыжовник», пер. У.-Л. Хейно

(22)

— Что вы, милостивый государь, — продолжал он отрывисто, — не знаете порядка?	— Hyvä herra, ettekö te ole selvillä järjestyksestä? hän sanoi nykivällä äänellä.

Н.В. Гоголь, «Шинель», пер. Э. Адриан

Другое типичное использование слова *hyvä* — в разного рода приветствиях и пожеланиях: *Hyvää päivää!* ’добрый день’, *Hyvää yötä!* ’спокойной ночи’, *Hyvää matkaa!* ’счастливого пути’. Интересно, что выражение *hyvää yötä* частотнее в текстах СХТ (СХТ — 41, ПФ — 21), что, вероятно, связано с большим количеством детской литературы в СХТ (17 контекстов встретились именно в детской литературе). В то же время *hyvää päivää* чаще встречается именно в переводных текстах (СХТ — 13, ПФ — 45). Менее формальное приветствие *Päivää!* также более частотно в ПФ. Однако неформальное приветствие-прощание *Hei!* частотнее в СХТ (0,25 на 1000 словоупотреблений в СХТ против 0,1 на 1000 словоупотреблений в ПФ).

Таким образом, изучение коллокаций слова *hyvä* в СХТ и ПФ подтверждает интуитивное предположение, что переводные тексты — более официальные, переводчики пишут языком, более близким к стандартному литературному языку, чем писатели. Кроме того, не следует забывать, что бóльшая часть русских текстов была написана в XIX веке или в советское время, и формы обращения в них иногда кажутся слишком официальными даже современному носителю русского языка. Это тоже в определенной степени влияет на язык переводов.

Прилагательное *pieni* ’маленький’

Частотность этого прилагательного в СХТ заметно больше (СХТ — 1,50; ПФ — 0,79). Однако, частотность финского прилагательного *pieni* в «ПарРус» значительно выше, чем его стандартного русского эквивалента *маленький*, частотность которого составила всего лишь 0,45 на 1000 словоупотреблений. Низкая частотность русского прилагательного (даже учитывая возможность его субстантивации: *маленький* = ’ребенок’) объясняется большим количеством уменьшительных и уменьшительно-ласкательных дериватов (*дерево* → *деревце, дом* → *домик, поле* → *полюшко* и т.п.), передающих это значение. При переводе сема ’маленький’ лексикализуется далеко не всегда. Приведем пример.

(23)

И верно: на пасеке у меня — тольки што не рай: березка, знаишь, липа в цвету, пчелки... в-ж-ж... в-ж-ж...	Minun mehiläistarhani se on melkein kuin paratiisi: koivut lehtii ja lehmukset kukkii ja mehiläiset surisee ...

А. Фадеев, «Разгром», пер. У.-Л. Хейно

В исходном тексте употреблено два существительных с уменьшительно-ласкательными суффиксами: *березка* и *пчелка*. В переводе используются только финские лексические эквиваленты этих слов *koivu* и *mehiläinen*, лишенные какой-либо «уменьшительности».

Таким образом, вероятно, на частотность этого прилагательного в переводных текстах косвенно влияет малая частотность его лексического эквивалента в русских текстах. Хотя детская литература, в изобилии представленная в СХТ, и повышает общую частотность прилагательного *pieni* в корпусе финских оригинальных текстов, средняя частотность этого прилагательного без учета детской литературы составляет 1,22, то есть даже в этом случае частотность *pieni* в СХТ выше, чем в ПФ

Глаголы. Среди самых частотных слов исследуемых текстовых массивов глаголов относительно немного, частоты некоторых из них довольно близки. Однако есть целый ряд глаголов, частоты которых различаются довольно сильно (см. табл. 30); *t*-тест дает для них значение 2,07, таким образом результаты являются значимыми на уровне 0,05 для 8 степеней свободы. Однако, поскольку полной уверенности в репрезентативности наших текстовых массивов нет, отметим только самые частотные глаголы, относительные частоты которых различаются более, чем на 0,5 (выделены в таблице жирным шрифтом).

Таблица 30. Частотность финских глаголов в СХТ и ПФ

СХТ			**ПФ**		
Слово	**Частота**	**Отн. Частота**	**Слово**	**Частота**	**Отн. Частота**
olla	105321	55,83	**olla**	115115	51,37
sanoa	13853	7,34	sanoa	15946	7,12
tulla	9379	4,97	**tulla**	9360	4,18
saada	7481	3,97	**voida**	8844	3,95
voida	6223	3,30	**saada**	7642	3,41
pitää	5432	2,88	mennä	6251	2,79
mennä	5360	2,84	**alkaa**	6158	2,75
tehdä	5071	2,69	tehdä	6074	2,71
alkaa	4176	2,21	pitää	5711	2,55

Все глаголы, перечисленные в таблице, имеют очень широкую семантику и могут употребляться в самых разных синтаксических конструкциях. Проводя наш анализ, будем исходить из того, что на употреблении глаголов в текстах ПФ в некоторой степени сказывается влияние ИЯ (ср. анализ употребления глаголов в переводах с английского на немецкий и норвежский в статье Стига Йоханссона (Johansson 2002)). Влияние ИЯ может проявляться в большей частотности употребления глаголов в некоторых синтаксических конструкциях или в большей типичности определенных значений.

Первым в списке зафискирован глагол *olla* ’быть’. Однако функции этого глагола тесно связаны с употреблением времен и наклонений. Поэтому этот глагол будет рассмотрен ниже, в разделе об употреблении грамматических форм в оригинальных и переводных текстах (4.4.3).

Для исследования функционирования глаголов будем использовать коллокации. Наиболее значимыми для глагола, в отличие от существитель-

ного, являются правые коллокаты, поэтому при анализе основное внимание уделялось первому и второму правым коллокатам.

Рассмотрим для примера только один глагол — глагол *tulla* ’приходить / приезжать / появляться / становиться’. При сравнении списков коллокатов этого слова, полученных из СХТ и ПФ, удалось установить следующие существенные различия.

Оборот *tulla takaisin* ’приходить назад’ почти в два раза чаще встречается в СХТ. В качестве коллоката *tulla* наречие *takaisin* встречается в СХТ 206 раз, из них 109 раз в качестве первого правого и 30 раз — в качестве второго правого коллоката. В ПФ зафискировано 134 употребления *takaisin* в качестве коллоката *tulla*, из них 65 в качестве первого правого и 31 — в качестве второго правого коллоката. Это различие связано с влиянием ИЯ. В русском языке обороты типа *идти/ехать назад* не являются частотными: в полученном из русского субкорпуса «ПарРус» списке коллокаций для наречия *назад* наиболее частотным из коллокатов — глаголов движения была словоформа *пошел*, которая встретилась в качестве соседа слова *назад* всего 11 раз. Гораздо типичнее для русского языка глаголы *возвращаться/возвратиться/ вернуться*. Этим глаголам в финском языке соответствует глагол *palata* и упомянутый выше оборот *tulla takaisin*. Поэтому можно прогнозировать более высокую частотность глагола *palata* в переводах с русского. И действительно в СХТ частота этого глагола — 0,56 на 1000 употреблений, а в ПФ — 0,68. Одновременно происходит и снижение частотности оборота *tulla takaisin* по сравнению с оригинальными финноязычными текстами, на которое мы и обратили внимание.

Приведем несколько примеров переводов с использованием этого оборота.

(24)

Назад в Россию пешком шли...	Tulimme takaisin Venäjälle jalkaisin...

А.П. Чехов, «В овраге», пер. У.-Л. Хейно

(25)

— Она говорила, что вернется вечером.	— Hän sanoi, että palaa vasta illalla.

А.П. Чехов, «Лишние люди», пер. Ю. Конкка

(26)

— Так мы же вернемся, дядя Саша! Через месяц.	— Tullaanhan me takaisin, Saša-setä. Kuukauden päästä.

А. и Б. Стругацкие, «Попытка к бегству», пер. Э. Адриан

Оборот *tulla mieleen* ’приходить на ум’ также значительно частотнее в СХТ (СХТ: R1 — 73, R2 — 38, всего 159; ПФ: R1 — 16, R2 — 5, всего — 40). Здесь также прослеживается влияние русских оригиналов. В русском языке довольно много синонимичных слов и оборотов: *вспомниться, прийти в голову, прийти на ум, напоминать (что-либо)* и др. С точки зрения формы наиболее близкими к финскому *tulla mieleen* кажутся *прийти*

на ум и *прийти в голову*. Однако проверка по параллельным текстам показала, что переводчики предпочитают использовать другие эквиваленты.

Оборот *прийти на ум* оказался в «ПарРус» довольно редким, найдено 6 контекстов. Лишь в двух из них переводчики использовали оборот *tulla mieleen*. Приведем один из них (27).

(27)

Выпалил первое, что пришло на ум.	Kakaisi mitä ensimmäiseksi mieleen tuli.

А.Приставкин, «Ночевала тучка золотая», пер. Э. Адриан

В остальных случаях употреблялись *tulla ajatelleeksi*, *johtua mieleen* и более далекие по смыслу обороты.

Оборот *прийти в голову* оказался более частотным, в «ПарРус» он встретился 35 раз. Лишь в четырех случаях переводчики использовали оборот *tulla mieleen*.

(28)

Пришло мне тоже в взбудораженную мою голову, что роли ведь теперь окончательно переменились <...>	Sekasortoiseen mieleeni tuli myös ajatus, että osat ovat nyt täydellisesti vaihtuneet <...>

Ф.М. Достоевский, «Записки из подполья», пер. В. Каллама

Обратим внимание, что другой переводчик Достоевского, Эса Адриан, при переводе этого места использует другое соответствие — *kohota päähän*: *Sekavaan päähäni kohosi myös ajatus, että roolithan olivat nyt lopullisesti vaihtuneet* <...>. Чаще всего переводчики используют следующие эквиваленты: *juolahtaa päähän* (9), *pälkähtää päähän* (7), *tulla ajatelleeksi* (3).

Примеры.

(29)

Ну, да мало ль мне мыслей тогда пришло в голову, так что я положил все это обдумать потом <...>	Mieleeni juolahti silloin paljonkin ajatuksia, niin että päätin harkita asiaa myöhemmin <...>

Ф.М. Достоевский, «Преступление и наказание», пер. Ю. Конкка

(30)

Когда оказалось, что это не так, ему не пришло в голову, что он не прав, упрощая миропорядок.	Kun kävi ilmi ettei asia ollutkaan niin, hänen päähänsä ei pälkähtänytkään, että väärässä olisi hän eikä maailmanjärjestys.

Б.Л. Пастернак, «Доктор Живаго», пер. Ю. Конкка

Другие русские обороты также оказались довольно приблизительными соответствиями для финского оборота: употребление *tulla mieleen* в качестве их эквивалента оказываются единичными.

Таким образом, несоответствия русского и финского лексико-семантических полей «появления мысли» приводят к различиям в частотности употребления слов и выражений, к ним относящимся. Наряду с более высокой частотностью с СХТ оборота *tulla mieleen* зафиксирована более низкая частотность по сравнению с ПФ глаголов, использующихся при переводе русских слов и оборотов *вспомниться*, *прийти в голову* и *прийти на ум*: *muistua* (0,01 - 0,036), *juolahtaa* (0,016 и 0,026), *pälkähtää* (0,009 и 0,03)

Другое частотное в СХТ и редкое в ПФ выражение — *tulla mukaan* ’идти вместе / участвовать’ (СХТ: R1 — 63, R2 — 15, всего — 101; ПФ: R1 — 24, R2 — 9, всего — 54). Вновь различия в частотности коллокатов объясняются «раздвоением»: в русском языке точного соответствия этому выражению нет. В контекстах «ПарРус», в которых для перевода использовалось *tulla mukaan*, в русской части встречаются и *присоединяться*, и *участвовать*, и *быть вместе*, и *посещать* и др.

Примеры:

(31)

— Пойдем-ка, — говорит. — Я тебе кое-что покажу.	— Tule mukaan, hän sanoo. — Minä näytän sinulle jotakin.

А. и Б. Стругацкие, «Парень из преисподней», пер. Э. Адриан

(32)

— Довольно, — сказал Трудолюбов, вставая. — Если ему так уж очень захотелось, пусть придет.	— Riittää, sanoi Trudoljubov nousten. — jos hänen mielensä niin kovasti tekee, niin tulkoon mukaan.

Ф.М. Достоевский, «Записки из подполья», пер. В. Каллама

В приведенных примерах (как и в большинстве полученных контекстов) наречие *mukaan* появляется как маркер совместного действия, а поскольку семантика глагола *tulla* нередко оказывается довольно расплывчатой, то и сам оборот может означать просто ’делать что-либо вместе с другими’

В (33) вообще не удается найти формальное соответствие к финскому обороту, в русском тексте отсутствует идея движения, которая появляется в переводе.

(33)

— Погодите, и я сяду играть, — говорит он.	— Odottakaa, minäkin tulen mukaan, sanoo hän.

А.П. Чехов, «Детвора», пер. Ю. Конкка

Примеры, полученные из СХТ, ничем принципиально не отличаются от примеров из ПФ. Более низкая частотность оборота в ПФ связана с тем, что у переводчика возникает потребность к его употреблению только при реструктуризации.

В то же время некоторые коллокации оказываются более частотными в ПФ. Приведем в качестве примера сочетание *tulla toimeen* (СХТ: R1 — 38, R2 — 9, всего — 58; ПФ: R1 — 83, R2 — 15, всего — 117). Конкорданс на этот оборот показал, что в большей части случаев он используется для перевода русских оборотов *обойтись / прожить без чего-л. / кого-л* или *хватать / быть достаточным для чего-л*.

Примеры:

(34)

Без нас вам нельзя обойтись.	Te ette tule toimeen ilman meitä.

Ф.М. Достоевский, «Преступление и наказание», пер. Ю. Конкка

(35)

Верьте богу, нельзя без лжи!

Uskokaa Jumalan nimessä, en tule toimeen valehtelematta!

А.П. Чехов, «Нищий», пер. Ю. Конкка

(36)

Маловато, но при скромной жизни хватит.

... vähänlaisesti, mutta kyllä sillä vaatimattomasti eläen tulee toimeen...

М.А. Булгаков, «Мастер и Маргарита», пер. У.-Л. Хейно

Русско-финские словари также предлагают оборот *tulla toimeen* в качестве одного из основных эквивалентов для *обойтись без чего-л.* (см., например, БРФС 1963: 459).

Служебные слова

Употребление служебных слов в исследуемых массивах различается довольно сильно. Покажем это на примере союзов. Приведенные ниже данные по наиболее частотным союзам в СХТ и ПФ (см. табл. 31) наглядно демонстрируют различия в частотности.

Таблица 31. Частотность союзов в СХТ и ПФ

Союз	**СХТ**		**ПФ**	
ja ’и’	76319	40,46	92660	41,35
että ’что’	18553	9,84	21038	9,39
mutta ’но’	13409	7,11	20727	9,25
kun ’когда’	11881	6,30	9004	4,02
jos ’если’	4695	2,49	4602	2,05
tai ’или’	3783	2,01	3341	1,49

Наблюдаемые различия, как и в предыдущих случаях, тесно связаны с влиянием ИЯ. Например, союзы *ja* ’и’ и *mutta* ’но’ имеют в переводах с русского языка другую частотность по той причине, что финским союзам *ja* и *mutta* соответствуют три русских союза: *и, а* и *но.* Поэтому при переводе предложений с союзом *а* используются как *ja*, так и *mutta*, а также другие союзные средства.

Примеры:

(37)

Снег пыхал, искрился, а в легких тенях отливало мякотной синью.

Lumi huokui ja kimalteli ja kevyissä varjoissa häivähti pehmeään siniseen.

В. Распутин, «Живи и помни», пер. Э. Адриан

(38)

У меня только глаза работают, а у вас вон и голова, и руки.

Minulla tekevät työtä vain silmät, mutta teillä sekä pää että kädet.

В. Дудинцев, «Белые одежды», пер. У.-Л. Хейно

(39)

— Да, — отвечает, — тоже такой поединок, только это, — говорит, — не насчет чести, а чтобы не расходоваться.	— Sellainen kaksintaisteluhan se on, mutta tätä ei käydä kunnian tähden, vaan siksi etteivät kuluttaisi rahojaan.

Н.С. Лесков, «Очарованный странник», пер. Л. Пююккё

Русскому союзу *или* в финском языке соответствуют два союза — *tai* (нестрогая дизъюнкция) и *vai* (строгая дизъюнкция). В переводах с русского языка частотность союза *tai* оказывается более низкой, чем в оригинальных финских текстах, в то время, как частотность *vai* увеличивается. В СХТ и ПФ частотность *vai* составила соответственно 0,68 и 0,8 на 1000 словоупотреблений.

Причем сводить всю проблему к переводу именно союза *или* нельзя, поскольку достаточно часто при переводе происходит изменение синтаксической структуры, в результате чего бессоюзная конструкция заменяется союзной, а подчинение превращается в сочинение.

Примеры:

tai:

(40)

Попадись этакая собака в Петербурге или Москве, то знаете, что было бы?	Jos tuollainen koira nähtäisiin Pietarissa tai Moskovassa, niin tiedättekö, mitä seuraisi?

А.П. Чехов, «Хамелеон», пер. Ю. Конкка

(41)

Учтите, Сталину приходится сейчас вести войну не в лесах Брянска и не на полях. Украины.	Ottakaa huomioon, ettei Stalin käy nyt sotaa Brjanskin metsissä tai Ukrainan aroilla.

Ю. Семенов, «Семнадцать мгновений весны», пер. Н. Пиенимяки

vai:

(42)

— Сколько ты, интересно, получаешь в месяц? — Павла взволновал вопрос: ворует этот человек или нет?	— Paljonko sinä mahdat saada kuussa? — Pavelia kiihdytti kysymys: varastaako tämä mies vai ei?

В. Шукшин, «Капроновая елочка», пер. Э. Адриан

(43)

— Но, — осторожно отозвался дед и помолчал, не зная, верить, не верить Настене.	— Jaa, ukko lausahti varovaisesti ja vaikeni tietämättä, uskoako Nastenaa vai ei.

В. Распутин, «Живи и помни», пер. Э. Адриан

Частотность союза *или* в русском субкорпусе «ПарРус» составляет 1,58 на 1000 словоупотреблений, если учитывать также его книжные и разговорные варианты (*иль, аль, али*), то она оказывается немного больше — 1,64. Частотность финских соответствий — *tai* и *vai* — оказывается несколько выше — 2,29 на 1000 словоупотреблений. Однако, это не обязательно означает, что в русском тексте реже встречается дизъюнкция. Просто для ее выражения, кроме союза *или* нередко используется также союз *а*. Например:

(44)

Он тотчас постарался ее объяснить, и объяснение было странное: показалось смутно прокуратору, что он чего-то не договорил с осужденным, а может быть, чего-то не дослушал.	Hän yritti selittää sen heti itselleen, mutta selityksestä tuli kovin omituinen: maaherrasta tuntui hämärästi, että häneltä oli vielä jäänyt jotakin sanomatta tuomitulle tai kenties häneltä oli jäänyt kuulematta jotakin tuomitun puheesta.

М.А. Булгаков, «Мастер и Маргарита», пер. У.-Л. Хейно

Таким образом, существование союза *а*, который функционирует и в соединительной, и в разделительной, и в противительной функциях, объясняет причину низкой по сравнению с финскими соответствиями частотность союза *или*, и, кроме того, проливает некоторый свет на снижение частотности союза *tai* в ПФ по сравнению с СХТ. Переводчик с русского языка под влиянием ИЯ чаще употребляет союзы *ja* и *mutta*, чем *tai*.

Исследование употребления в «ПарРус» русского союза *когда* и его финского аналога *kun* также показывает, что ставить знак равенства между этими союзами нельзя.

Примеры:

(45)

Николай Николаевич стоял у окна, когда показались бегущие.	Nikolai Nikolajevitš seisoi parhaillaan ikkunan ääressä, kun ensimmäiset ihmiset juoksivat esiin.

Б. Пастернак, «Доктор Живаго», пер. Ю. Конкка

(46)

Ну и рукой махнешь, потому что не нашел первоначальной причины.	Niinpä sitten viittaat kintaalla, kun et löytänytkään perimmäistä syytä.

Ф.М. Достоевский, «Записки из подполья», пер. Э. Адриан

(47)

Неизвестные злодеи, пока он любовался красавицей, утащили все, кроме контрабаса и цилиндра.	Sillä aikaa kun hän ihaili kaunotarta, tuntemattomat roistot olivat vieneet kaiken, paitsi bassoviulua ja silinterihattua.

А.П. Чехов, «Роман с контрабасом», пер. Ю. Конкка

(48)

А с любовью и без счастья можно прожить.	Ja kun rakastaa, niin voi elää ilman onneakin.

Ф.М. Достоевский, «Записки из подполья», пер. В. Каллама

(49)

А в частности, может быть выгоднее всех выгод даже и в таком случае, если приносит нам явный вред и противоречит самым здравым заключениям нашего рассудка о выгодах <…>	Ja joskus se voi olla kalleinta kaikista arvoista, jopa silloinkin, kun se tuottaa meille selvää vahinkoa ja on ristiriidassa järkemme terveiden päätelmien kanssa eduistamme <…>

Ф.М. Достоевский, «Записки из подполья», пер. В. Каллама

Сравнение частот союза *когда* и *kun* в текстах «ПарРус» показывает, что финский союз значительно частотнее: частота употребления *когда* в русских текстах составляет 2,45 на 1000 словоупотреблений, в то время, как частота *kun* в переводах этих текстов на финский язык — 4,02 на 1000

словоупотреблений. Происходит это потому, что *kun* может использоваться и при переводе других русских союзов и союзных слов: *потому что, что, как, если,* и т.п.

4.3.3. Функционирование грамматических форм и синтаксических конструкций в оригинальных и переводных текстах

Для полномасштабного исследования функционирования грамматических форм необходим аннотированный корпус текстов. Поэтому «ПарРус», в нынешней версии которого грамматическая разметка отсутствует, полных данных по грамматике дать не может. С другой стороны, наличие состыкованных текстов оригиналов и переводов дает хорошие возможности для получения частных фактов по функционированию конкретных слов в конкретных формах. Изучение частотности словоформ позволяет установить некоторые общие тенденции. В данном разделе будет предпринята попытка рассмотреть функционирование некоторых грамматических форм в ПФ и сравнить полученные данные с СХТ.

Глагол olla*: время и наклонение*

Выше, в разделе 4.4.2, отмечались различия в частотности глагола *olla* ’быть’, который оказался более частотным в текстах СХТ. Глагол этот не является полнозначным и функционирует как связка, различные формы этого глагола используются для образования целого ряда глагольных форм: перфекта, плюсквамперфекта, а также форм прошедшего времени кондиционала (konditionaali) и потенциального (potentiaali) наклонения.

Чтобы выяснить, чем обусловлены выявленные различия в частотах, попробуем проследить, какие грамматические формы этого глагола зафиксированы в корпусах текстов и каковы их частоты. В СХТ выявлены 204 разные формы, в ПФ их оказалось несколько больше — 218[45].

Частоты для целого ряда грамматических форм оказались различными, некоторые формы более типичны для оригинальных текстов, некоторые — для переводных. В ПФ намного чаще встречаются формы потенциального наклонения (potentiaali) — *lienen, lienet, lienee* и др., впрочем, и здесь час-

[45] Такое большое количество форм в финском языке вызвано тем, что кроме основных форм есть еще варианты, получаемые добавлением к основной форме притяжательных суффиксов и различных частиц. В результате теоретически возможное количество форм очень велико, реально употребляются далеко не все, однако все равно количество встретившихся в большом текстовом массиве различных форм высокочастотного слова может достигать сотен.

тотность самой частотной формы (3-го лица единственного числа) составляет всего лишь 0,03 на тысячу словоупотреблений. Более частотны также формы настоящего времени, например частотность формы 1-го лица ед. числа (*olen*) в СХТ — 1,5, а в ПФ — 1,92, 3-го лица ед. ч. (*on*) — 10,71 и 11,99, 3-го лица мн. ч. (*ovat*) — 0,97 и 1,25. Более высокая частотность и у форм императива (*olkoon, olkoot*) и у формы транслатива первого инфинитива (*ollakseen* и др.).

Однако картина меняется на противоположную в частотах различных форм прошедшего времени и кондиционала (konditionaali). В СХТ намного выше частотность форм претерита, например, частотность формы 3-го л. ед. ч. (*oli*) по СХТ составила 20,04 на тысячу словоупотреблений, а по ПФ — всего 15,96. Аналогично, у формы 3-го лица единственного числа в кондиционале (*olisi*) частота в СХТ — 3,17, а в ПФ — 2,98. Форма инфинитива (*olla*) также частотнее в СХТ (1,76 против 1,35). Формы пассива также несколько типичнее для оригинальных текстов (*ollaan, oltaisiin* и др.). Довольно сильно отличается частотность форм действительного причастия прошедшего времени (*ollut, olleet*), например у формы ед. числа (*ollut*) частотность в СХТ составляет 4,16 на тысячу словоупотреблений, а в ПФ — 3,22.

Таким образом, частотность глагола *olla* оказывается выше в оригинальных текстах за счет более частого употребления форм кондиционала и прошедшего времени. Формы причастия прошедшего времени используются для образования различных форм перфекта и плюсквамперфекта, поэтому их более высокая частотность говорит о более высокой частоте употребления перфекта и плюсквамперфекта в оригинальных финских текстах. Можно предпололжить, что изменение частотности связано с влиянием ИЯ: в русском языке нет перфекта, а сослагательное наклонение используется реже, чем финский кондиционал[46]. Что касается форм претерита, это, скорее всего, тоже в какой-то степени связано с употреблением плюсквамперфекта, которая образуется путем присоединения к претериту от глагола *olla* формы причастия прошедшего времени (например, *olin syönyt* для глагола *syödä* ’есть, питаться’).

Отсутствие или низкая частотность в ИЯ какой-либо грамматической формы может влиять на употребление этой формы в переводе: у переводчика возникает потребность использовать ее лишь в том случае, когда буквальный перевод выглядит неестественно или вызывает смысловые потери. Приведем несколько примеров с использованием в переводе сосла-

[46] Это интуитивное соображение о большей частотности кондиционала в финском языке по сравнению с русским сослагательным наклонением полностью подтверждается материалом «ПарРус»: частотность форм сослагательного наклонения в русском субкорпусе составляет 3,33 на 1000 словоупотреблений, в финском субкорпусе формы кондиционала употребляются с частотой 11,6 на 1000 словоупотреблений.

гательного наклонения (konditionaali) при его отсутствии в исходном тексте.

(50)

Приказчик поднимал и опускал курки, дышал на стволы, прицеливался и делал вид, что задыхается от восторга. Глядя на его восхищенное лицо, можно было подумать, что сам он охотно пустил бы себе пулю в лоб, если бы только обладал револьвером такой прекрасной системы, как Смит и Вессон.	Myyjä nosteli ja laski hanoja, puhalsi piippuihin, tähtäili ja näytti vallan läkähtyvän ihastuksesta. Katsellessaan hänen riemastuneisiin kasvoihinsa olisi voinut ajatella, että hän ampuisi halusta luodin omaan otsaansa, mikäli itse omistaisi niin erinomaisen aseen kuin Smith & Wesson merkkisen revolverin.

А.П. Чехов. «Мститель», пер. Ю. Конкка

(51)

Он умер от тоски и чрезмерной склонности к обобщениям. Других причин вроде бы не было, а вскрывать мы его не вскрывали, потому что вскрывать было противно.	Se kuoli ikävästä ja liiallisesta taipumuksesta yleistyksiin. Muita syitä ei tainnut ollakaan, eikä me ruvettu tekemään sille avausta, sillä sen avaaminen olisi ollut vastenmielistä.

Ерофеев В. «Москва — Петушки», пер. Э. Адриан

(52)

— Нет, душа моя, для меня уж нет таких балов, где весело, — сказала Анна, и Кити увидела в ее глазах тот особенный мир, который ей не был открыт.	— Ei, kultaseni, minulla ei enää ole sellaisia tanssiaisia, joissa olisi hauskaa, Anna sanoi ja Kitty huomasi hänen silmissään sen erityisen maailman, joka ei ollut avoinna hänelle.

Л.Н. Толстой. Анна Каренина, пер. Л. Пююккё

Во всех трех приведенных примерах используются формы индикатива, однако присутствует семантика ирреальности ситуации, а при переводе на финский с использованием индикатива это значение исчезло бы и смысл высказываний на ПЯ исказился бы.

Комитатив

Даже при простом просмотре переводов с русского на финский бросается в глаза частое употребление комитатива (komitatiivi), падежа со значением ’вместе с чем-либо’ (Eskola & Tommola 2000, Eskola 2002: 207–230). В современном финском языке этот падеж является малоупотребительным и встречается, как правило, в составе клише, например *presidentti seurueineen* ’президент с сопровождающими его лицами’, *ministeri vaimoineen* ’министр с супругой’, *Mikko perheineen* ’Микко с семьей’ (см., например, White 1993). Более высокая частотность комитатива в финских переводных текстах по сравнению с оригинальными уже привлекала к себе внимание исследователей, причем отмечается, что именно в переводах с русского языка на финский частотность комитативов особенно высока (Eskola & Tommola 2000, Eskola 2002: 207–230).

Получить статистику по употреблению комитатива даже по нашему неаннотированному корпусу не составляет большого труда: набор финалей *-ineni, -inesi, -inemme, -ineenne, -ineen, -ineensa, -ineensä* позволяет выделить

из словника почти все существительные в форме комитатива с минимальным «шумом». После выполнения запроса к базе данных в списке кроме форм комитатива оказываются существительные, основа которых оканчивается на *—ine* (*laine* → *laineen*, *esine* → *esineen* и др.) и действительные причастия прошедшего времени от глаголов, оканчивающихся на *—ida* (таких глаголов очень мало) в форме генитива единственного числа (*imuroida* → *imuroineen*). Устранить эти формы из списка не представляет трудности.

Так, из словников СХТ и ПФ были получены списки всех существительных в форме комитатива. В СХТ зафиксировано 415 разных существительных в форме комитатива, сумма их абсолютных частот составила 692. Таким образом, относительная частота употребления комитатива составила для СХТ 0,37 на 1000 словоупотреблений. В ПФ количество существительных, употребленных в комитативе оказалось значительно больше и составило 779, суммарная частота — 1197, относительная частота употребления комитатива для существительных — 0,53 на 1000 словоупотреблений, разница весьма существенная. На полученных частотных списках был проведен *t*-тест (matched pairs *t*-test), значение *t* составило 11,89, что позволяет считать результаты значимыми на уровне более 0,005 (достаточно было бы значения 2,576).

Употребление прилагательных в форме комитатива — еще более редкое явление. В СХТ зафиксировано всего 126 разных прилагательных в этой форме, их суммарная частота — 173, а относительная частота употребления прилагательных в комитативе — 0,009 на тысячу словоупотреблений. В ПФ это явление намного более частое: 340 прилагательных в комитативе, общая частота — 539, относительная частота — 0,24 на тысячу словоупотреблений.

Таким образом, форма комитатива, встречаясь в переводах с русского языка существенно чаще чем в текстах, изначально написанных на финском, является скорее инструментом переводчика, нежели автора.

То же самое значение ’вместе с’ выражается аналитически сочетанием «генитив + kanssa». На самом деле эта конструкция не может считаться полностью синонимичной комитативу, у которого гораздо больше значений и намного более широкая сфера употребления. Конструкция с комитативом, когда главное слово и существительное в комитативе обозначают лиц, ближе по значению к сочинительной конструкции «X и Y». Именно тогда, когда в русском тексте описывается ситуация с двумя агенсами, при переводе на финский может использоваться комитатив.

Примеры:

(53)

— Это мещанин ведь торгует тут на углу, с бабой, с женой, а?	— Tuossa kulmassahan se torikauppias eukkoineen myyskentelee tavaroitaan, vai kuinka?

Ф.М. Достоевский, «Преступление и наказание», пер. Ю. Конкка

(54)

Ее муж, <...>, приехал с семьей в Советскую Россию, принял советское гражданство.	Hänen miehensä, <...>, muutti perheineen Neuvosto-Venäjälle ja otti Neuvostoliiton kansalaisuuden.

В. Гроссман, «Все течет», пер. Э. Адриан

(55)

— Я доктор из Москвы. Следую с семьей в этом эшелоне.	— Olen moskovalainen lääkäri. Matkustan perheineni tässä junassa.

Б. Пастернак, «Доктор Живаго», пер. Ю. Конкка

В то же время достаточно часто употребляется и сочинительный оборот вида «A ja B», и конструкция «генитив + kanssa»:

(56)

Вчера они с Юлией Михайловной не смогли приехать потому, что вдруг напросились в гости Аструги, Борис Львович с женой.	Edellisenä päivänä heidän tulonsa oli estynyt koska Astrugit, Boris Lvovitš ja hänen vaimonsa olivat yllättäen halunneet tulla käymään.

Ю. Трифонов, «Дом на набережной», пер. М. Коскинен

(57)

Ведь дядя Володя после севера оставил ее, уехал в Ташкент с новой семьей.	Volodja-setähän oli pohjoisen jälkeen jättänyt hänet ja matkustanut Taškentiin uuden perheen kanssa.

Ю. Трифонов, «Дом на набережной», пер. М. Коскинен

Комитатив может использоваться и в таких контекстах, когда компонентами ситуации являются «нелица», как, например, в следующем контексте:

(58)

Я очаровал эту даму, внеся ей деньги за всех трех птенцов Катерины Ивановны, кроме того, и на заведения пожертвовал еще денег; наконец, рассказал ей историю Софьи Семеновны, даже со всеми онерами, ничего не скрывая.	Minä lumosin tuon naisen, viemällä hänelle rahasumman Katerina Ivanovnan kaikkien kolmen lapsen hyväksi, sitäpaitsi lahjoitin rahaa myös orpokodeille. Lopuksi kerroin hänelle Sofia Semjonovnan tarinan, jopa kaikkine yksityiskohtineen, salaamatta mitään.

Ф.М. Достоевский, «Преступление и наказание», пер. Ю. Конкка

В контекстах, где оборот *с + Тв.п.* обозначает коагента или контрагента, употребление комитатива невозможно. Например:

(59)

С этой стороны он был верен дедовским обычаям, и, споря с женой, называл ее мирскою женщиной и немкой.	Tässä suhteessa hän oli uskollinen esi-isien tavoille ja vaimonsa kanssa riidellessään hän nimitteli tätä maailmalliseksi naiseksi ja saksattareksi.

Н.В. Гоголь, «Шинель», пер. Э. Адриан

Некоторые обороты с комитативом превратились в клише, в которых грамматическая мотивация сочетания начинает утрачиваться, например *perheineen* ’вместе с семьей’ (см. тж. Eskola & Tommola 2000, Eskola 2002). В тех случаях, когда в финском языке существует клише с комитативом, эквивалентное переводимому выражению, использование комитатива оказывается более частотным, например, обороты *с семьей, с супругой, с*

детьми и т.п. часто переводятся на финский язык именно как *perheineen, vaimoineen, lapsineen*:

(60)

Как бы то ни было, квартира простояла пустой и запечатанной только неделю, а затем в нее вселились - покойный Берлиоз с супругой и этот самый Степа тоже с супругой.	Asunto oli kuitenkin tyhjillään viikon päivät, ennen kuin Berlioz vaimoineen ja Stjopa vaimoineen muuttivat siihen.

М.А. Булгаков, «Мастер и Маргарита», пер. У.-Л. Хейно

Другое значение комитатива — ’характерный признак объекта’ (например *Tampere mustinemakkaroineen* ’Тампере с его кровяной колбасой’) — в финском языке никогда не передается оборотом с *kanssa*, ему может быть синонимична конструкция с придаточным предложением (например, *Tampere, jossa myydään / syödään mustamakkara* ’Тампере, где продают / едят кровяную колбасу’) (см. тж. Eskola 2002: 207–209). В русском языке это значение передается той же конструкцией «*с + Тв.п.*». В финском языке такой оборот может передаваться комитативом, особенно в случаях употребления в русском языке конструкций с обособленными определениями:

(61)

Профессор Хейфец, бледный, с белыми, как сияние, волосами, в длинной болотного цвета кофте домашней вязки, слегка согнувшись, спешил к сцене - головой вперед.	Kalpea professori Heifetz hohtavanvalkeine hiuksineen, yllään pitkä suonvärinen kotikutoinen villäpättä kiiruhti hiukan kumarassa näyttämölle — pää eteen työnnettynä.

В. Дудинцев, «Белые одежды», пер. У.-Л. Хейно

Наконец, материал корпуса показывает, что в финском языке послелог *kanssa* не употребляется с существительными, обозначающими предметы, то есть обороты, аналогичные русскому *человек с ружьем* — *?mies kivääriinsä kanssa*, для финского языка нетипичны (по сообщениям носителей финского языка, такого типа сочетаемость, в принципе, возможна, но в СХТ подобное не зафиксировано). Поэтому при переводе с русского языка оборотов такого типа переводчику приходится использовать другие конструкции.

(62)

Лишь только шоферы трех машин увидели пассажира, спешашего на стоянку с туго набитым портфелем, как все трое из-под носа у него уехали пустыми, почему-то при этом злобно оглядываясь.	Niin pian kuin kolmen taksiasemalla seisovan vuokra-auton kuljettajat näkivät matkustajan, joka oli tulossa kovaa vauhtia pullea salkku kainalossaan, he käynnistivät autonsa ja pyyhälsivät tyhjinä matkaan aivan hänen nenänsä edestä, vieläpä mulkoilivat vihaisesti taakseen.

М.А. Булгаков, «Мастер и Маргарита», пер. У.-Л. Хейно

Другой способ решения проблемы — использование комитатива, который в таком значении употребляться может (напр. *kaikkine tavaroineen / kimpsuineen ja kampsuineen* ’со всеми своими вещами / со всеми своими пожитками’). В некоторых случаях стиль становится более книжным, чем в исходном тексте, зато проблема получает изящное решение (63, 64).

(63)

Из дверей выскакивает мастер художественного слова Филипп Громкий со своими элегантными чемоданами.	Ulko-ovesta syöksähtää taidelausunnan mestari Filip Gromkij hienoine matkalaukkuineen.

В. Аксенов, «Звездный билет», пер. Э. Адриан)

(64)

И Варенуха с портфелем выбежал из кабинета.	Ja Varenuha juoksi salkkuineen ulos kabinetista.

М.А. Булгаков, «Мастер и Маргарита», пер. У.-Л. Хейно

Зафиксированы даже такие случаи, когда в русском тексте конструкция «*с + Тв.п.*» не употребляется, но в переводе все равно используется комитатив.

(65)

В церкви никого, кроме мужиков и дворников и их баб, не было.	Kirkossa ei ollut ketään muita kuin talonpoikia ja kartanon palvelijoita vaimoineen.

Л.Н. Толстой, «Анна Каренина», пер. Л. Пююккё)

(66)

Весловский, желая видеть стрельбу, заехал в болото и увязил лошадей.	Veslovski oli tahtonut katsella toisten ampumista, ajanut suolle hevosineen ja juuttunut sinne.

Л.Н. Толстой, «Анна Каренина», пер. Л. Пююккё

Таким образом, более высокая частотность комитатива может быть определена как особенность переводных текстов. Одна из причин этого — наличие в ИЯ грамматических конструкций, которые трудно передавать другими средствами финского языка. Еще одно достоинство комитатива — его компактность, которая является большим плюсом при переводе с русского языка, славящегося многословными синтаксическими конструкциями.

Темпоральные конструкции

Как уже говорилось выше (см. стр. 176), средняя длина предложения в СХТ несколько больше, чем в переводных текстах ПФ. Возможно, такая краткость предложений в ПФ связана с использованием книжных синтаксических конструкций финского языка: темпоральных конструкций, оборотов со вторым инфинитивом и т.п.

В подтверждение этой мысли приведем данные по частотности форм второго инфинитива глагола *saada* ’получать’ — *saadessa, saadessasi, saadessaan* и др., а также причастных форм того же глагола *saatua, saatuani* и др., которые используются, главным образом, в темпоральных конструкциях. Наша статистика показывает, что для переводных текстов эти формы более типичны: некоторые формы в СХТ не встретились ни разу, другие более частотны в ПФ, а причастная форма saatuaan намного частотнее в ПФ (см. табл. 32).

Таблица 32. Частотность форм второго инфинитива и страдательных причастий от глагола *saada* в СХТ и ПФ

Словоформа	**СХТ**		**ПФ**	
	Абс. част.	Отн. част.	Абс. част.	Отн. част.
saadessaan	6	0,003	19	0,008
saadessamme	0	0	1	0
saadessani	0	0	10	0,004
saatua	6	0,003	9	0,004
saatuaan	52	0,028	165	0,074
saatuamme	2	0,001	1	0
saatuani	6	0,003	11	0,005
saatuanne	0	0,000	5	0,002
saatuasi	1	0,001	1	0

Второй инфинитив и партитив страдательного причастия используются, главным образом, в финских темпоральных конструкциях, которые хорошо подходят для перевода русских деепричастных оборотов и придаточных времени.

Примеры:

(67)

Как он будет горд и доволен, получив мою записку!

Miten ylpeäksi ja tyytyväiseksi hän tuleekaan saadessaan minun kirjelappuni!

Л.Н. Толстой, «Анна Каренина», пер. Л. Пююккё

(68)

Когда только получил пост, был как Керубино

Oli kuin kerubi virkan saadessaan

В. Дудинцев, «Белые одежды», пер. У.-Л. Хейно)

(69)

Когда обулся, то встал, потоптался для верности: не трет, не жмет ли где

Saatuaan saappaat jalkaan hän nousi, polkaisi varmuuden vuoksi muutaman kerran: ettei hierrä eikä purista mistään

В. Белов, «Привычное дело», пер. Х. Лааксонен

Таким образом, можно предположить, что русские деепричастные обороты и придаточные предложения вынуждают финских переводчиков чаще пользоваться книжными конструкциями, что и приводит, в конечном итоге, к уменьшению СДП. Впрочем, эта гипотеза нуждается в дальнейшей проверке.

4.3.4. Выводы

Проведенный анализ наглядно показывает, что язык переводов отличается от языка оригинальных текстов, написанных на том же языке, причем отличия можно найти на самых разных уровнях. С одной стороны, эти отличия вызваны неизбежным при переводе влиянием ИЯ на ПЯ, которое, по-видимому, преодолеть невозможно, да и не нужно (см., например,

Eskola 2002: 37–62). С другой стороны, сам переводчик нередко может намеренно калькировать в своем переводе некоторые особенности языка оригинала. Джулиан Хауз выделяет два типа переводов — «скрытые» (covert) и «открытые» (overt). «Скрытые» переводы не должны восприниматься как переводы, «открытые» переводы не должны восприниматься как самостоятельные тексты (House 1981: 184–194). По мнению Честермана, художественные переводы как раз являются примером «открытых» переводов. Читатели должны знать, что перед ними перевод, а не самостоятельное произведение (Chesterman 1997, 65).

В последнее время исследования языка переводов проводились на материале разных языков (см., например, Mauranen 2000, Eskola 2002, Johansson 2002, Михайлов (в печати)). Все исследователи сходятся в том, что язык переводов отличается от языка оригинальных произведений. Однако достаточно трудно определить особенности, свойственные языку переводов как таковых. Единственная более или менее универсальная тенденция — меньшая степень идиоматичности языка и меньшая эмоциональность (Михайлов (в печати)). По-видимому, для каждой пары языков можно выделить свои особенности и каждый язык создает свои варианты для перевода с разных языков.

Язык переводов нельзя рассматривать как периферийное явление, не заслуживающее внимания исследователей. Ведь переводные тексты во многих случаях оказывают определенное влияние на нормативный язык: они являются одним из источников новой лексики, новых значений слов и т.п. Вполне возможно, они могут влиять и на другие уровни языка.

Исследование языка переводных текстов может быть довольно полезным и в практическом плане. Знания о характере влияний ИЯ при переводе на данный язык могут помочь переводчику осознанно избегать нежелательных эффектов, а в случае необходимости получать текст с минимальной интерференцией.

4.4. Словарные переводные эквиваленты и данные из корпуса текстов

> Существует один знаменательный факт: мы, на нашем еще неустроенном и молодом языке, можем передавать глубочайшие формы духа и мысли европейских языков: европейские поэты и мыслители все переводимы и передаваемы по-русски, а иные переведены уже в совершенстве. Между тем на европейские языки, преимущественно на французский, чрезвычайно много из русского народного языка и из художественных литературных наших произведений до сих пор совершенно непереводимо и непередаваемо. Я не могу без смеха вспомнить один перевод (теперь очень редкий) Гоголя на французский язык, сделанный в средине 40-х годов, в Петербурге, г-м Виардо, мужем известной певицы, в сообществе с одним русским, теперь по праву знаменитым, но тогда еще лишь начинавшим молодым писателем. Вышла просто какая-то галиматья, вместо Гоголя.
>
> Ф.М. Достоевский. Дневник писателя

Приведенная в качестве эпиграфа цитата из Достоевского иллюстрирует довольно распространенное заблуждение, согласно которому одни языки приспособлены для выполнения переводов с других языков, а другие — нет. Между тем эта «приспособленность» языка к переводу, на самом деле зависит от трех факторов: типологической близости ИЯ и ПЯ, близости культур и наличием переводческой традиции для этой пары языков. Поэтому переводы на русский язык с европейских языков уже во времена Достоевского выполнялись вполне удачно, в первую очередь, благодаря тому, что с этих языков уже переводили довольно много и довольно долго, а переводы с русского не удавались в какой-то степени из-за отсутствия традиции, а в какой-то степени из-за специфики самого языка и русской культуры. Кроме того, оценить в полной мере качество художественного перевода на родной язык трудно даже человеку, блестяще владеющему ПЯ: фактические ошибки, непереданная языковая игра, незамеченные аллюзии, плохо переданные стилистические особенности исходного текста остаются незамеченными. В то же время, неточности перевода с родного языка на неродной заметить намного легче.

Таким образом, художественный перевод труден для любой пары языков, потери разного рода при переводе неизбежны, и вообще художественный перевод существует лишь потому, что адресатами перевода являются носители ПЯ, а не ИЯ. Хорош не самый точный и не самый полный перевод, а тот перевод, который был принят носителями ИЯ.

Тем не менее, Достоевский обращает внимание на ключевую проблему художественного перевода, на проблему переводимости / непереводимости. В настоящем разделе мы попытаемся на материале «ПарРус» выяснить, насколько велик пласт лексики, плохо переводимой с русского языка на финский, и какие лексические группы в него входят. С другой стороны,

интересным представляется выяснить, насколько переводима «переводимая» лексика.

Как отмечал в своей классической статье Р.О. Якобсон, единицами перевода, как правило, являются не слова, а завершенные сообщения (Jakobson 1989: 55). Лексические межъязыковые соответствия являются условностью, поскольку лишь часть лексики имеет более или менее точные соответствия в другом языке, другая часть или имеет приблизительные эквиваленты, или вообще таковых не имеет.

Словари межъязыковых соответствий обычно называют дву- или многоязычными. Не менее распространенное наименование для таких словарей — переводные словари, поскольку предполагается, что в них содержатся именно переводные эквиваленты и что ими активно пользуются переводчики. Однако всякий, кому приходилось переводить с одного языка на другой, обращал внимание на несовершенство двуязычных словарей. Предлагаемых эквивалентов оказывается то слишком мало, то наоборот слишком много, нередко предлагаемые соответствия довольно сильно отличаются по значению как от входной лексемы, так и между собой. В итоге переводчику зачастую приходится использовать при переводе слова, отсутствующие в списках эквивалентов.

Следует отметить, что двуязычные словари могут быть ориентированы на носителей входного языка или на пользователей, для которых входной язык не является родным. Так, «Финско-русский словарь» под редакцией И. Вахроса и А. Щербакова (ФРС) адресован в первую очередь носителям русского языка, а недавно вышедший финско-русский словарь Х. Ниеменсиву и Е. Никкиля (H. Niemeisivu & E. Nikkilä, *Suomi-venäjä– sanakirja*, WSOY 2003) — носителям финского языка. Однако, разница главным образом состоит в наличии грамматических помет и информации о словоизменении / словообразовании для входного языка. В словаре Вахроса и Щербакова имеются грамматические таблицы по финскому языку, в словарных статьях указывается тип словоизменения для входного слова, В словаре Ниеменсиву и Никкиля подобная информация отсутствует. Аналогично, во многих словарных статьях ФРС содержится информация страноведческого характера, отсутствующая во втором словаре. На характер предлагаемых эквивалентов и их количество тип словаря, к сожалению, заметного влияния не оказывает.

Поэтому двуязычные словари часто оказываются объектом критики со стороны переводчиков (см., напр., Holmes 1988: 110).

Таким образом, складывается парадоксальная ситуация: двуязычные словари, которые должны быть первым помощником переводчика в его нелегком труде, на самом деле иногда даже мешают ему. И связано это не только и не столько с частными техническими трудностями, сколько с фундаментальными вопросами межъязыковых соответствий. В настоящее время при составлении двуязычных словарей начинают использовать и параллельные тексты, но этот источник данных все же является вспомо-

гательным, в значительной степени из-за технических трудностей. Двуязычные словари, как правило, воспроизводят с той или иной долей модификаций существующую для данной пары языков традицию.

В целом представляется справедливым утверждение о том, что чем больше переводов выполняется с языка А на язык Б, тем легче переводить с языка А на язык Б, причем опыт, накопленный при переводе с языка А на язык Б никак не помогает развитию перевода в обратном направлении (см. Holmes 1988: 13). Наличие богатого переводческого опыта для определенной пары языков способствует и развитию двуязычной лексикографии, причем именно в определенном направлении.

В данном разделе будет идти речь о русско-финских словарных соответствиях. Мы будем сравнивать данные по переводным эквивалентам, полученные из «Большого русско-финского словаря» (Kuusinen & Ollikainen 1963, далее — БРФС 1963), и списки эквивалентов, которые можно найти в параллельных конкордансах из «ПарРус». Попробуем провести выборочное сравнение данных по словарю и по корпусу текстов. Будем исходить из того, что произведения художественной литературы являются одним из основных источников данных для общих словарей.

Сравнение будет проводиться для самых обычных слов, входящих в лексическое ядро языка и не обозначающих культурных реалий. У таких слов, в большинстве случаев, есть соответствия в других языках. Однако, как известно, семантическое поле слова в одном языке никогда в точности не совпадает с семантическим полем какого-либо слова в другом языке (Holmes 1988: 9).

4.4.1. Параллельный корпус текстов как инструмент лексикографа

Покажем, как можно использовать данные ПКТ для уточнения и дополнения двуязычных словарей. Для примера возьмем четыре абстрактных существительных, относящихся к семантическому полю интерперсональных отношений: *любовь*, *ненависть*, *дружба*, *вражда*.

Любовь

Для слова *любовь* БРФС дает следующие эквиваленты: 1) (чувство расположения) *rakkaus, mielisuosio, suosio*, 2) (сердечная склонность) *rakkaus, lempi*, 3) (интерес) *kiintymys, harrastus, halu*. В корпусе «ПарРус» это слово употреблено более 600 раз. Главный эквивалент, предлагаемый словарем — *rakkaus*, оказывется наиболее частым переводным эквивалентом и в корпусе текстов: более, чем в 500 контекстах слово *любовь* переводится именно словом *rakkaus*. Приведем только один пример.

(70)

Она никогда не испытает свободы любви…	Hän ei koskaan saisi tuntea rakkauden vapautta

Л.Н. Толстой. «Анна Каренина», пер. Л. Пююккё

Другой типичный эквивалент — относящееся к высокому стилю слово *lempi*, которое, впрочем, встретилось в «ПарРус» всего 7 раз.

(71)

В ваши годы девичья любовь так дорого стоит!	Teidän iässänne naisen lempi on kallista!

Ильф И., Петров Е. «Двенадцать стульев», пер. Р. Силванто, Ю. Конкка

Третий распространенный эквивалент, *kiintymys*, который также встретился в корпусе 7 раз, используется в тех случаях, когда русское слово *любовь* синонимично словам *дружба, уважение, привязанность*.

(72)

Глеб же Капустин по-прежнему неизменно удивлял. Изумлял, восхищал даже. Хоть любви, положим, тут не было. Нет, любви не было. Глеб жесток, а жестокость никто, никогда, нигде не любил еще.	Gleb Kapustin puolestaan löi entiseen tapaan vääjäämättä ällikällä. Tyrmistytti. Vieläpä ihastutti. Mutta kiintymystä ei hänen osakseen tainnut tulla. Ei, kiintymystä ei ollut. Gleb on julma, eikä kukaan ole vielä milloinkaan eikä missään kiintynyt julmuuteen.

В.М. Шукшин. «Срезал», пер. Э. Адриан

Все эти эквиваленты в словаре зафиксированы. Из остальных предлагаемых БРФС соответствий встретилось только слово *suosio*, и то только один раз.

(73)

Занимая третий год место начальника одного из присутственных мест в Москве, Степан Аркадьич приобрел, кроме любви, и уважение сослуживцев, подчиненных, начальников и всех, кто имел до него дело.	Hoidettuaan kolmatta vuotta päällikön virkaa eräässä Moskovan virastossa oli Stepan Arkaditš yleisen suosion lisäksi voittanut myös virkatovereittensa, alaistensa, esimiestensä ja kaikkien hänen kanssaan asioissa olevien ihmisten kunnioituksen.

Л.Н. Толстой. «Анна Каренина», пер. Л. Пююккё

Далее нас поджидают неожиданности. В словаре вообще никак не отражено значение слова *любовь* — ‘тот/та, кого любят’. В «ПарРус» есть примеры употребления этого слова в этом значении, переводчики используют в качестве эквивалентов *rakastaja* ‘возлюбленный, любовник’ и *rakastajatar* ‘возлюбленная, любовница’, и субстантивированное прилагательное *rakas* ‘дорогой, милый’.

(74)

Подождите, черти! Моя любовь лежит в больнице.	Odottakaa, pirut! Minun rakkaani makaa sairaalassa.

В. Аксенов. «Звездный билет», пер. Э. Адриана

(75)

Там же, в Ленинграде, она узнала, что Брузжак завел в университете новую любовь — совсем молоденькую студентку.	Leningradissa hän myös sai tietää Bruzžakin hankkineen yliopistolla uuden rakastajattaren — aivan nuoren opiskelijatytön.

В. Дудинцев. «Белые одежды», пер. У.-Л. Хейно

Кроме того, из примеров корпуса хорошо видна условность словарных соответствий. Ведь вовсе не обязательно, чтобы при переводе слова в качестве эквивалента использовалось слово, относящееся к той же части речи. И вообще, как правило, перевод происходит по крайней мере на уровне предложения, то есть переводчик должен передать информацию, содержащуюся в предложении и не обязан переводить каждое слово (ср. Jakobson 1989: 55).

Поскольку в финском языке слова *rakkaus* и *lempi* употребляются только для обозначения очень сильных чувств, переводчики нередко употребляют подходящие по контексту менее эмоционально насыщенные эквиваленты. Приведем несколько примеров.

(76)

Он видел проявления общей любви к нему, но не мог отогнать печали, от которой был сам не свой.	Hän näki itseensä kohdistetut yhteisen myötätunnon osoitukset, mutta ei voinut karkottaa surumielisyyttä, joka vaivasi häntä.

Б. Пастернак. «Доктор Живаго», пер. Ю. Конкка

(*myötätunto* — ‘сочувствие’).

(77)

«Действительно, - думал он, с любовью глядя на воодушевленное лицо героя, - глохнешь тут за работой. Великие вехи забываешь».	«Todellakin», hän ajatteli katsellen mieltymyksellä sankarin henkeviä kasvoja. «Tässähän surkastuu työn keskellä. Suuret tienviitat unohtuvat.»

И. Ильф, Е. Петров. «Золотой теленок», пер. А. Аарто)

(*mieltymys* — ‘симпатия’).

В корпусе было обнаружено довольно много примеров, в которых существительное *любовь* переводится с помощью глагола *rakastaa* ‘любить’, т.е. имеет место сдвиг в переводе (translation shift) (Catford 1965: 12).

(78)

Мало сказать: он служил ревностно, — нет, он служил с любовью.	Hän uurasti hartaasti, ei, enemmänkin - hän rakasti työtään.

Н.В. Гоголь. «Шинель», пер. Э. Адриана)[47].

Употребление глагола вместо существительного связано с тем, что финский глагол *rakastaa*, по-видимому, все-таки не так эмоционально нагружен, как существительные *rakkaus* и *lempi*. Оба эти эквивалента русского слова *любовь* в финском языке — очень сильные слова, и в некоторых контекстах, в которых не идет речь о любви мужчины и женщины, они начинают выглядеть в переводе неестественно. Возможно, что употребление при переводе глагола *rakastaa* — один из способов снять

[47] Интересно, что в другом переводе «Шинели», выполненном Юхани Конкка, слово любовь переведено буквально словом *rakkaus*: *Ei riitä, jos sanoo: hän hoiti virkaansa innolla, ei, hän hoiti sitä rakkaudella*, а в третьем переводе того же произведения, принадлежащем Хуго Ялканену, используется наречие *rakastuneesti*, образованное от глагола *rakastua* ‘влюбляться’: *Liian mietoa on sanoa, että hän hoiti virkaansa hartaasti; ei, hän hoiti sitä rakastuneesti*.

излишнюю эмоциональность текста. Но и этот глагол в финском языке употребляется для обозначения очень высоких чувств, более употребителен глагол *pitää* ‘нравиться’, который также используется в качестве переводного эквивалента к существительному *любовь*.

(79)

Пожалуйста, передайте ей от меня мою любовь.	Olkaa hyvä ja kertokaa hänelle, että pidän hänestä.

Толстой Л.Н. «Анна Каренина», пер. Л. Пююккё

Кроме глаголов, в качестве эквивалентов для слова *любовь* употребляются и наречия. Так, в следующем примере выражение *с кошачьей любовью* переведено оборотом *kissamaisen pehmeästi* ‘по-кошачьи мягко’.

(80)

В Черноморском порту легко поворачивались краны, спускали стальные стропы в глубокие трюмы иностранцев и снова поворачивались, чтобы осторожно, с кошачьей любовью опустить на пристань сосновые ящики с оборудованием Тракторостроя.	Nosturit kääntyilivät kevyesti Tšornomorskin satamassa, ne laskivat teräskouransa ulkomaalaisten laivojen syvään ruumaan ja kääntyivät uudelleen laskeakseen laiturille kissamaisen pehmeästi haapapuisia laatikoita, joissa oli traktoritehtaan tarvikkeita.

И. Ильф, Е. Петров. «Золотой теленок», пер. А. Аарто

Вообще, для сочетания *с любовью* буквальный эквивалент *rakkaudella* отнюдь не является наилучшим (хотя он иногда и встречается в переводах, см. например сноску 47). Наиболее часто в контекстах такого рода употребляется наречие *hellästi* ‘нежно’:

(81)

Присоветуй им встретить меня с детской любовию и послушанием; не то не избежать им лютой казни.	Kehota heitä ottamaan minut vastaan niin hellästi ja nöyrästi kuin lapsi isänsä; jos he eivät sitä tee, heidät tuomitaan kuolemaan.

А.С. Пушкин. «Капитанская дочка», пер. Й. Холло

Ненависть

Для слова *ненависть* БРФС дает только один эквивалент — *viha*. В «ПарРус» зафиксировано 128 контекстов употребления слова *ненависть*, и в 80 случаях в качестве эквивалента использовалось предлагаемое словарем существительное *viha*. Однако таким образом покрывается чуть более двух третей материала. Какие же другие эквиваленты использовали переводчики? В ходе анализа параллельных контекстов обнаружились три вполне приемлемых соответствия, не указанных в словаре: это производное от глагола *vihata* ‘ненавидеть’ существительное *vihaaminen* ‘ненавидение’, а также *vihamielisyys* ‘вражда’ и *kauna* ‘неприязнь’.

(82)

Развитой и порядочный человек не может быть тщеславен без неограниченной требовательности к себе самому и не презирая себя в иные минуты до ненависти.	Kehittynyt ja kunnollinen ihminen ei voi olla turhamainen olematta rajattoman vaativa itseään kohtaan, halveksimatta itseään joinakin hetkinä vihaamiseen asti.

Ф.М. Достоевский. «Записки из подполья», пер. Э. Адриана

(83)

Из того, например, что она не обрадовалась приходу мужа, что ей не понравилось, как он держал себя за обедом, она вдруг заключила, что в ней начинается ненависть к мужу.	Esimerkiksi siitä, ettei hän ilahtunut miehensä kotiintulosta ja ettei häntä miellyttänyt miehensä käyttäytyminen päivällispöydässä, hän äkkiä päätteli, että hänessä oli alullaan vihamielisyys miestään kohtaan.

А.П. Чехов. «Несчастье», пер. Ю. Конкка

(84)

<...> Штирлиц понял, что им овладела иная ненависть к этому государству <...>.	<...> Stirlitz oivalsi, millaista kaunaa hän tuntee tätä valtiota kohtaan.

Ю. Семенов. «Семнадцать мгновений весны», пер. Н. Пиенимяки

Кроме того, в корпусе есть и несколько окказиональных эквивалентов, например, *suuttumus* ‘возмущение’, *katkeruus* ‘горечь’.

(85)

— На! Жри... — крикнул он, дрожа от возбуждения, острой жалости и ненависти к этому жадному рабу.	— Ota ne! Syö ne.., ärjäisi hän vavisten kiihtymyksestä, katkerasta inhosta ja suuttumuksesta tuota ahnasta orjaa kohtaan.

М. Горький. «Челкаш», пер. А. Митрошин

(86)

Алексей Александрович и прежде не любил графа Аничкина и всегда расходился с ним во мнениях, но теперь не мог удерживаться от понятной для служащих ненависти человека, потерпевшего поражение на службе, к человеку, получившему повышение.	Aleksei Aleksandrovitš ei aikaisemminkaan ollut pitänyt kreivi Anitškinista, oli aina ollut eri mieltä tämän kanssa, eikä hän nytkään voinut olla osoittamatta sellaista ymmärrettävää katkeruutta, jota virassaan kolhuja saanut virkamies tuntee toista, ylennyksen saanutta kohtaan.

Л.Н. Толстой. «Анна Каренина», пер. Л. Пююккё

Далее, для оборота *с ненавистью* как правило используются наречия *vihaisesti, kiukkuisesti, kiukkuissaan, vihamielisesti, vihoissaan, vimmoissaan*, а также прилагательные *vihainen* и *kiukkuinen* в форме эссива.

(87)

Тьфу, прости господи!.. - оборвал он с ненавистью, резко рванув рукой наотмашь, словно отрубая.	Hyi helvetti! hän keskeytti vihaisesti ja huitaisi kiivaasti kädellään kuin jotakin hakaten.

А. Фадеев. «Разгром», пер. У.-Л. Хейно

И наконец, существительное *ненависть* довольно часто заменяется при переводе на конструкцию с глаголом *vihata* ‘ненавидеть’.

(88)

Ненависть к нововведениям была отличительная черта его характера.	Hän vihasi kaikkia uudistuksia; juuri se oli hänen luonteensa erikoispiirteenä.

А.С. Пушкин. «Барышня-крестьянка», пер. Й. Холло

Дружба

Для слова *дружба* БРФС дает один эквивалент — *ystävyys*. В «ПарРус» этот эквивалент явно доминирует: в 88 контекстах из 108 русское существительное *дружба* переводится финским *ystävyys*. Однако нам встретились в качестве эквивалента к *дружба* близкое к *ystävyys* существительное *toveruus* ‘товарищество’.

(89)

К родной двери, к той самой, знакомой с первых дней жизни, к двери, за которой доверие, наивная святая правда, жалость, дружба и сочувствие были настолько естественны, до абсолютной простоты, что сами эти понятия определять не имело смысла.	Kotiovelleen. Sille joka oli tullut hänelle kaikkein tutuimmaksi elämänsä ensipäivistä, ovelle jonka takana luottamus, yksinkertainen, pyhä totuus, myötätunto, toveruus ja myötäeläminen olivat niin luonnollisia asioita, niin tyystin yksinkertaisia ettei näiden käsitteiden määrittelemisessä ollut mitään mieltä.

Г. Троепольский. «Белый Бим, черное ухо», пер. Л. Иранто)

При анализе контекстов нам встретились — так же, как и для слов *любовь* и *ненависть* — примеры использования при переводе однокоренных с эквивалентом слов, особенно часто использовалось существительное *ystävä* ‘друг’.

(90)

Я вас давно знаю и люблю, и по дружбе со Стивой и за вашу жену...	Stivan ystävänä olen tuntenut teidät jo kauan ja pitänyt teistä ja myös vaimonne takia...

Л.Н. Толстой. «Анна Каренина», пер. Л. Пююккё

Вражда

Для этого слова словарь предлагает следующие эквиваленты: *viha, vihanpito, vihollisuus*. По сравнению с предыдущими тремя словами, частотность каждого из которых в «ПарРус» превышала 100 употреблений, слово *вражда* оказалось довольно низкочастотным: из корпуса было получено всего девять контекстов. Тем более удивляет, что такой маленький материал дал три разных эквивалента: по четыре раза встретились *viha* и *vihamielisyys*, один раз встретилось *vihollisuus*, то есть опять один из эквивалентов, используемых переводчиками, оказался не зафиксированным в словаре.

(91)

Вражда сама собой разъединила Федора Ивановича и Брузжака и поставила по краям шеренги.	Viha jo sinänsä erotti Fjodorin ja Bruzžakin ja asetti heidät rivistön reunoille.

В. Дудинцев. «Белые одежды», пер. У.-Л. Хейно

(92)

Вдруг по совершенной случайности выяснилось, что эта закоренелая вражда есть форма маскировки молодой любви, прочной, прячущейся и давней.	Yllättävästä sattumasta selvisi sitten äkkiä, että tuo hillitön vihamielisyys olikin salatun rakkauden naamiointia.

Б. Пастернак. «Доктор Живаго», пер. Ю. Конкка

(93)

Таким образом вражда старинная и глубоко укоренившаяся, казалось, готова была прекратиться от пугливости куцой кобылки.	Siten pillastunut, lyhythäntäinen hevonen näytti painaneen vanhan syvän vihollisuuden unohduksiin.

А.С. Пушкин. «Барышня-крестьянка», пер. Й. Ахава, В. Хямеен-Анттила

Подведем итоги. Даже такой ограниченный материал показывает, что набор эквивалентов, предлагаемый двуязычным словарем, и эквиваленты, используемые переводчиками художественной литературы, не всегда совпадают[48]. Действительно, основной эквивалент словаря в большинстве случаев оказывается наиболее частотным в корпусе текстов: он, как правило, оказывается наиболее близким и наиболее естественным. Нам представляется, что немаловажную роль в выборе того или иного слова в качестве ПЭ играет, главным образом, авторитет словаря. С другой стороны, сам словарь отражает определенную традицию, сложившуюся в ходе преподавания языка как иностранного, а также в какой-то степени — в практике переводческой деятельности.

Это, в частности, подтверждает зафиксированная в корпусе «ПарРус» традиция использования слова *ylioppilas* в качестве эквивалента для русского слова *студент*. Именно такой эквивалент предлагает БРФС, более того, и ФРС дает для слова *ylioppilas* эквивалент *студент*. Но необходимо учитывать, что в современном финском языке слово *ylioppilas* является специфическим для финской культуры и обозначает ученика выпускного класса гимназии в период сдачи выпускных экзаменов и получения аттестата о среднем образовании. Следовательно, к студентам (т.е. учащимся вуза) это слово отношения не имеет. Эквивалентом для слова *студент* *ylioppilas* стало на рубеже XIX и XX веков, когда гимназист, сдавший выпускные экзамены, автоматически становился студентом (подробнее см. Lehmuskallio et al. 1991: 160). В отношении к студентам высших учебных заведений это слово в настоящее время употребляется только в деловом стиле (*ylioppilaskunta* — 'студенческий союз'), в языке повседневного общения и в нейтральном стиле бытует другое слово — *opiskelija*. Однако в корпусе текстов «ПарРус» слово *opiskelija* в качестве переводного экви-

[48] Это же явление демонстрируется на другом материале в статье (Mihailov 2002). В ходе выполнения лабораторных работ в рамках курса «Корпусная лингвистика», прочитанным автором данной работы на Отделении переводоведения Тамперского университета, большинство студентов получили аналогичные результаты по самой разной лексике — абстрактной и конкретной, относящейся к разным частям речи.

валента для слова *студент* употребляется существенно реже, чем «неправильный» по сути *ylioppilas*.

Анализ параллельных текстов позволяет понять разницу между словарным эквивалентом и переводным эквивалентом. Словарные эквиваленты — те соответствия, которые относятся к той же части речи, что исходное слово, и минимально зависят от контекста. Как мы сами убедились, далеко не всегда эти соответствия можно использовать при переводе, они скорее призваны помочь понять слово при изучении иностранного языка. Переводной эквивалент — межъязыковое лексическое соответствие, которое может быть использовано при переводе. Нельзя забывать, что переводной эквивалент — понятие условное, перевод на уровне лексемы — довольно редкое явление, обычно перевод идет на уровне предложений или абзацев, а иногда привлекаются и более высокие уровни. Юджин Найда в своих работах отстаивал приоритет «динамической эквивалентности» (dynamic equivalence) над «формальной эквивалентностью» (formal equivalence) (Nida, Taber 1974: 14).

Как видно из приведенных примеров, переводные эквиваленты не обязательно относятся к той же части речи, что и исходное слово, поскольку переводятся не конкретные слова а сообщение, которое при переводе может потребовать другого синтаксического оформления, в результате чего в качестве эквивалента данного слова в данном контексте может использоваться слово, относящееся к другой части речи (Jakobson 1989: 55). Более того, даже если имеются словарные эквиваленты, переводчик может использовать в качестве переводного эквивалента слово оригинала как экзотизм, чтобы сохранить колорит текста. Так, в переводах русской художественной литературы на финский язык встречаются такие русские экзотизмы, как *njanja*, *batjuška*, *matuška* и т.п.[49]

Таким образом, трудно сказать, возможно ли создание именно переводных словарей, помогающих подобрать правильный переводной эквивалент и обобщающих опыт практической переводческой деятельности, предназначенных в первую очередь для переводчиков. Однако совершенно очевидно, что параллельные корпуса текстов являются чрезвычайно полезным источником дополнительных данных для двуязычных словарей. Проблема использования параллельных текстов — в том, что не все типы текстов часто переводятся на другие языки. Например, такие жанры, как анекдот, программа телепередач или прогноз погоды переводятся с русского на финский или с финского на русский достаточно редко; так, в Хельсинкском университете в 2000 году был издан сборник русских

[49] Нередко именно таким путем в язык проникает новая лексика: первоначально это экзотизмы в художественных или публицистических текстах, затем часть их попадает в активный словарь. Иногда таким путем заимствуются и идиомы, достаточно упомянуть такие фразеологизмы, пришедшие в русский язык из английского, как *последний из могикан*, *черная метка* или *джентльмен удачи*.

анекдотов с переводами на финский язык[50] и это пока единственная с своем роде книга. Некоторые типы текстов вообще оказываются недоступными (например, частные документы); не все переводы являются высококачественными. Тем не менее, анализ параллельных текстов позволит существенно повышать качество двуязычных словарей, убирать из них неадекватные эквиваленты, дополнять их переводами устойчивых сочетаний и идиом и уменьшать количество лакун.

4.4.2. Параллельный корпус текстов и вопрос о соответствии картин мира

Проблема «языковой картины мира» обсуждается в лингвистике достаточно давно. Еще Вильгельм фон Гумбольдт говорил, что язык описывает вокруг народа круг, выйти из которого можно, лишь вступив в другой круг, описанный другим языком. В XX веке в работах Сепира, Уорфа и других ученых было показано, насколько велика пропасть между разными языками (см. например, Кронгауз 2001: 104–113, Вежбицкая 2001: 18). Обращалось внимание на то, что в разных языках реальность членится поразному, семантические поля разработаны с различной степенью подробности, и даже слова разных языков, на первый взгляд обозначающие одно и то же и зафиксированные в двуязычных словарях как эквиваленты, на самом деле довольно сильно различаются по значению.

В настоящее время проблема языковой картины мира продолжает вызывать живой интерес исследователей. Довольно много работ посвящено отличиям русской языковой картины мира от «европейского стандарта» (см. например, Вежбицкая 1997, Шмелев 2002).

Анна Вежбицкая в своей статье «Русский язык» указывает на следующие важные черты русской языковой картины мира:

- эмоциональность;
- иррациональность;
- неагентивность;
- любовь к морали.

Эти особенности находят проявление на самых разных уровнях: и в лексике, и в словообразовании, и в грамматике (подробнее см. Вежбицкая 1997: 33–88).

Большой интерес вызывают так называемые «ключевые слова» культуры, то есть слова и понятия, которые являются для данной культуры цен-

[50] Arto Mustajoki (päätoim.), Efim Kurganov, Arto Lehmuskallio, Ekaterina Protasova, Katja Manninen, Pirjo Niskanen, Veikko Suvanto, Linda Söderholm, Olga Zhuk (toim.): *Venäläinen ruletti - Russkaja ruletka. Venäläisiä vitsejä suomeksi ja venäjäksi*. Helsingin yliopiston slavistiikan ja baltologian laitos.: Helsinki 2000.

тральными. Вежбицкая отмечает, что «анализ «ключевых слов» культуры не обязательно должен вестись в духе старомодного атомизма. Напротив того, некоторые слова могут анализироваться как центральные точки, вокруг которых организованы целые области культуры. Тщательно исследуя эти центральные точки, мы, возможно, будем в состоянии продемонстрировать общие организационные принципы, вокруг придающие структуру и связность культурной сфере в целом и часто имеющие объяснительную силу, которая распространяется на целый ряд областей» (Вежбицкая 2001: 37).

Анна Вежбицкая говорит о невозможности определить логическим путем набор концептов, уникальных для данного языка. «Нет никакого конечного множества таких слов в каком-либо языке, и не существует никакой «объективной процедуры открытия», которая позволила бы их выявить. Чтобы продемонстрировать, что то или иное слово имеет особое значение для некоторой отдельно взятой культуры, необходимо рассмотреть доводы в пользу этого. Конечно, каждое подобное утверждение потребуется подкрепить данными, но одно дело данные, а другое — «процедура открытия»» (Вежбицкая 2001: 36).

Единственный возможный путь, по ее мнению, — это сравнение исследуемого языка с другими языками и выявление особенностей, отличающих его от всех других языков (там же).

А.Д. Шмелев выделяет следующие «сквозные мотивы» в русской языковой картине мира: «’в жизни всегда может случиться непредвиденное’ (*если что, в случае чего, вдруг*), ’всего все равно не предусмотришь’ (*авось*), ’чтобы сделать что-то, бывает необходимо мобилизовать внутренние ресурсы, а это не всегда легко’ (*неохота, собираться / собраться, выбраться*), но зато ’человек, которому удалось мобилизовать внутренние ресурсы, может сделать очень многое’ (*заодно*), ’человеку нужно много места, чтобы чувствовать себя спокойно и хорошо’ (*простор, даль, ширь, приволье, раздолье*), но ’необжитое пространство может приводить к душевному дискомфорту’ (*неприкаянный, маяться, не находить себе места*), ’хорошо, когда человек бескорыстен и даже нерасчетлив’ (*мелочность, широта, размах*)» (Шмелев 2002: 17).

Работы по русской языковой картине мира дают много интересной и полезной информации. Однако в большинстве случаев русский язык сравнивается с «европейским стандартом», к которому относят английский, французский и немецкий языки, нередко представляемые как некое единое целое (хотя каждый из этих языков имеет свою картину мира, отличающуюся от двух других). Между тем, увеличение базы для сравнения может внести некоторые коррективы в полученные наборы ключевых концептов.

В этом плане «ПарРус» может оказаться источником чрезвычайно полезной информации. Финский язык дальше от «европейского стандарта», чем русский язык, поскольку он относится к другой языковой семье. Одно-

временно можно сказать, что финский язык ближе к «европейскому стандарту», поскольку Финляндия исторически была теснее связана с Европой: много сотен лет она была частью Швеции и лишь чуть больше ста лет входила в Российскую империю, и это не могло не отразиться на финском языке. Параллельные тексты корпуса позволят получать информацию о том, какая лексика и какие грамматические конструкции ИЯ вызывают трудности при переводе с русского на финский.

Выделенный исследователями набор ключевых концептов русской языковой картины мира представляется хорошей базой для проведения нашего обзора лексических соответствий. Одновременно можно будет проверить, являются ли эти концепты столь же чуждыми для финского языка, как они чужды для английского, французского и немецкого. Другой важный вопрос: вызывают ли «ключевые концепты» больше проблем при переводе, чем прочая лексика.

***Мобилизация внутренних ресурсов**: собираться и добираться*

Собираться/собраться

Видовая пара *собираться / собраться* — многозначная, ср. *собраться на площади, собраться в дорогу, собраться с силами*. Рассмотрим только одно значение, связанное с преодолением препятствий и мобилизацией внутренних ресурсов: ’готовиться, намереваться’. А.Д. Шмелев справедливо отмечает, что важным элементом семантики этих глаголов является процесс подготовки к совершению каких-либо действий, причем не рациональное планирование, а «раскачка», преодоление лени или нежелания перейти к активным действиям; этот процесс можно проиллюстрировать фразой типа *целый час лежу и собираюсь встать* (Шмелев 2002: 146). Интересно, что толковые словари русского языка этого значения специально не фиксируют: например, Малый академический словарь выделяет лишь значения ’подготовиться, приготовиться к поездке куда-либо’ (*собраться в поход*) и ’напрячь и возбудить к действию весь имеющийся запас чего-л.’ (*собраться с силами*) (МАС, т. IV: 172), другие словари более или менее повторяют набор значений МАСа.

Несомненно, что в контекстах типа *Я как раз собирался тебе позвонить*; *Он собрался на пенсию*; *Он собирается жениться на старости лет* глагол *собираться / собраться* имеет именно семантику «раскачки», выделяемую Шмелевым. Однако, представляется, что нередко *собираться / собраться* вполне синонимичен глаголу *намереваться*. Например, в контексте типа *Он собирается поступать в этом году в университет*, по-видимому, речь идет о вполне определенных и рациональных намерениях, вопрос лишь в том, выражается ли это намерение лишь в желании или в каких-то усилиях в направлении подготовки к вступительным экзаменам.

И наконец, существуют контексты, в которых довольно трудно решить, о каких «сборах» — «рациональных» или «иррациональных» — идет речь, например, *Он собрался писать очередную статью*.

БРФС дает следующие финские эквиваленты для *собираться / собраться*: (*приготовиться*) *hankkiutua, sonnustautua, varustautua, laittautua, suoriutua, suoria*; (*решиться*) *olla hankkeissa, aikoa* (БРФС 1963: 804). Интересно, что полученный из «ПарРус» параллельный конкорданс на эти глаголы дает совершенно другую картину: часть из предлагаемых эквивалентов не встретилась ни разу (*sonnustautua*, *suoriutua, suoria, olla hankkeissa*). Названный последним глагол *aikoa* на практике оказался самым распространенным соответствием. В то же время корпус дал довольно много очень эффективных и достаточно часто используемых способов перевода оборотов с *собираться/собраться*, не зафиксированных в словаре.

При переводе с русского на финский в качестве эквивалента чаще всего используется глагол *aikoa* ’намереваться’, который связан прежде всего с рациональной подготовкой к совершению какого-либо действия, по крайней мере с принятием решения, имеет ли смысл делать это.

(94)

А француженки, — объяснил Герасим Николаевич, — это двое молодых местных парижских начинающих врачей, которые собирались о нем писать статью	Ja ranskattaret puolestaan — Gerasim Nikolajevitš selitti — olivat kaksi nuorta paikallista aloittelevaa pariisilaislääkäriä, jotka aikoivat kirjoittaa hänestä artikkelin

М.А. Булгаков, «Театральный роман», пер. Э. Адриан)

(95)

Так Левка Шулепников из человека, которого собирались на весь свет опозорить, превратился в героя	Näin Levka Šulepnikov muuttui sankariksi tyypistä joka oli aiottu häpäistä koko maailman silmissä

Ю. Трифонов, «Дом на набережной», пер. М. Коскинен)

(96)

Они собирались обновить эти наряды двадцать седьмого, на традиционной ежегодной елке у Свентицких	Ne aikoivat ottaa ne ensimmäisen kerran ylleen kahdentenakymmenentenä seitsemäntenä päivänä, Sventitskien perinteelliseen kuusijuhlaan

Б. Пастернак, «Доктор Живаго», пер. Ю. Конкка

В приведенных примерах речь идет именно о намерениях что-либо сделать, а не о борьбе с ленью и иных иррациональных процессах. В контекстах с отрицанием глагол *собираться/собраться* также практически всегда имеет семантику ’отсутствия намерения’ и финский глагол *aikoa* вполне ему соответствует:

(97)

Я просто хочу кое-что напомнить вам, а совсем не собираюсь поучать вас	Haluan vain palauttaa mieleenne yhtä ja toista, en lainkaan aio ruveta opettamaan teitä

Б. Пастернак, «Доктор Живаго», пер. Ю. Конкка

Зато следующий контекст является хорошим примером «иррациональной» подготовки, и финский «рациональный» эквивалент *aikoa* к нему не

подходит. Переводчик, чувствуя эту странность, эксплицировал, что «намерение так намерением и осталось».

(98)

Нарубленные еще пять лет тому назад, они так и лежали у большой дороги напротив дома: Иван Африканович все собирался делать дому большой ремонт	Puut oli kaadettu jo viisi vuotta sitten; ne lojuivat maantien laidassa talon kohdalla. Drynov aikoi tehdä talossa suuren remontin mutta aikeeksi se oli jäänyt.

В. Белов, «Привычное дело», пер. Х. Лааксонен

Другой распространенный эквивалент — *hankkiutua* ’готовиться’ — используется в тех случаях, когда в русском тексте идет речь о процессах сборов или подготовки к чему-либо:

(99)

Дома в Москве уже все было по-зимнему, топили печи, и по утрам, когда дети собирались в гимназию и пили чай, было темно, и няня ненадолго зажигала огонь	Kotona Moskovassa oli jo täysi talvi, uuneja lämmitettiin aamuisinkin, kun lapset hankkiutuivat koulumatkalle ja juotiin teetä, oli pimeä, ja kesti hetken, ennen kuin lastenhoitaja sai valkean viritetyksi

А.П. Чехов, «Дама с собачкой», пер. У.-Л. Хейно

(100)

Работники, очевидно, замешкались и теперь наскоро свертывали свою бумагу и собирались домой	Työmiehet olivat ilmeisesti viipyneet, käärivät nyt paperirullia kokoon ja hankkiutuivat lähtemään kotiin

Ф.М. Достоевский, «Преступление и наказание», пер. Ю. Конкка

А вот в следующем примере из романа Достоевского, как нам представляется, выбор *hankkiutua* в качестве эквивалента для *собираться* несколько меняет смысл слов Порфирия Петровича, который в русском тексте ’наконец-то преодолел свою лень, мешавшую посетить Раскольникова’, а в финском тексте ’давно готовился нанести свой визит’.

(101)

— Не ждали гостя, Родион Романыч, — вскричал, смеясь, Порфирий Петрович. — Давно завернуть собирался, прохожу, думаю — почему не зайти минут на пять проведать.	— Ette tietysti odottanut vierasta, Rodion Romanytš, huudahti Porfiri Petrovitš nauraen. — Olen jo kauan hankkiutunut luoksenne ja nyt ohikulkiessani ajattelin, miksen voisi poiketa viideksi minuutiksi

Ф.М. Достоевский, «Преступление и наказание», пер. Ю. Конкка

Иногда в качестве эквивалента к *собираться/собраться* используется глагол *ryhtyä*, который ближе всего по значению к русскому глаголу *приниматься за что-л.* и может обозначать момент перехода от бездействия к действиям:

(102)

Он сидел неподвижно, откинув голову на деревянную спинку дивана и прикрыв глаза, как человек, которого собираются брить.	Hän istui liikkumattomana nojaten päätään sohvan puuselustaan ja silmät ummessa kuin miehellä, jolta ryhdytään ajamaan partaa.

И. Ильф, Е. Петров, «Золотой теленок», пер. А. Аарто

Другой способ передачи значения *собираться/собраться* в контекстах, где речь идет о дороге — обороты *olla menossa*, *olla lähdössä*, *olla tulossa* ’быть в процессе отбытия/прибытия’. Эти обороты уже никак не связаны с

«рациональным» планированием, правда, ничего «иррационального» в них тоже нет; они просто описывают момент приезда/отъезда или приближение такого момента.

(103)

Мы тут на дачу собрались...	Oltiin lähdössä mökille ...

В.М. Шукшин, «Как зайка летал на воздушных шариках», пер. Э. Адриан

(104)

К вам гости собираются	Teille on tulossa vieraita

Б. Пастернак, «Доктор Живаго», пер. Ю. Конкка

(105)

Я вышел из ванны свеж и бодр, как будто собирался на бал	Nousin ammeesta raikkaana ja reippaana, aivan kuin olisin ollut tanssiaisiin menossa

М.Ю. Лермонтов, «Герой нашего времени», пер. У.-Л. Хейно

Близок к вышеназванным оборот *olla aikeissa*, который может относится к любым переходам от действия к бездействию, и, по существу, очень близок к русскому *собираться/собраться*:

(106)

Некоторое время Антон смотрел на него, словно собираясь спросить о чем-то	Anton katsoi häneen jonkin aikaa, ikään kuin aikeissa kysyä jotakin

А. и Б. Стругацкие, «Попытка к бегству», пер. Э. Адриан

Зафиксирован также ряд случаев, когда глагол *собираться/собраться* остался непереведенным. Например,

(107)

Мне Мадсен говорил, ты утром на лыжах собираешься?	Madsen sanoi, että lähdet aamulla hiihtämään? (дословно: «Мадсен сказал, что ты утром пойдешь на лыжах»)

В. Дудинцев, «Белые одежды», пер. У.-Л. Хейно

Добираться / добраться

Другой глагол, связанный с преодолением препятствий — *добираться / добраться*. Отличие его от других глаголов, обозначающих перемещение в пространстве и достижение пункта назначения — *доехать, приехать, прибыть* и др. — в том, что употребление этого слова подразумевает преодоление определенных трудностей (далеко / плохая дорога / плохой транспорт / неблагоприятные погодные условия и т.п.), например *Наконец я добрался до дома* (Левонтина, Шмелев 1999). МАС дает следующее толкование для этого значения: «С трудом или нескоро дойти, доехать и т.п. до какого-л. места, предмета» (МАС 1984, т. I: 408).

БРФС дает для этого слова только один финский эквивалент — *päästä*, и в качестве варианта — *päästä perille* (БРФС 1963: 169). В текстах «ПарРус» используется более тридцати разных финских эквивалентов к этому глаголу: *päästä, tulla perille, jaksaa, kitkutella, valua, saapua* и др. Однако бóльшая часть этих эквивалентов являются чисто контекстуальными и не подходит на роль словарных эквивалентов, например, *valua, tunkeutua* и др.; некоторые являются неточными, например, *saapua, tulla, tavoittaa*.

Главным эквивалентом, безусловно, остается *päästä* ’достигать цели’, функционирование которого довольно близко к русскому *добираться / добраться*, правда значение его намного шире. Трудно сказать, насколько значим для этого слова компонент ’преодоление трудностей’, в толковом словаре финского языка на это указаний нет (CD-Perussanakirja 1997).

Примеры:

(108)

Наконец, какая там власть, и какая она там будет, пока мы туда доберемся?	Ja lopuksi, millainen hallitusvalta siellä on voimassa tai millainen tulee olemaan silloin, kun pääsemme sinne?

Б. Пастернак, «Доктор Живаго», пер. Ю. Конкка)

(109)

Не помнил, как добрался он до Сунжи	Hän ei muistanut, miten pääsi Sunžan rannalle saakka

А. Приставкин, «Ночевала тучка золотая», пер. Э. Адриан

Иногда переводчики используют варианты *päästä / tulla perille, olla perillä* ’оказаться в пункте назначения’. Послелог *perille* лишь указывает на появление агенса там, где он планировал быть, возможно в результате преодоления каких-либо затруднений, но совсем необязательно. Приведем пример из СХТ: *Hän saa heittää teidät perille BMW:llä* ’Он подбросит вас на БМВ’ (Reijo Mäki, «Tatuoitu taivas»).

Примеры:

(110)

Вот и добрались, шибздики, до Кавказа, можете вылезти да пощупать, с чем его едят!	Perillä ollaan, poitsut, Kaukasiassa, voitte painua ulos katsomaan, mitä se on syönyt

А. Приставкин, «Ночевала тучка золотая», пер. Э. Адриан)

(111)

Теперь, считай, добралась, теперь недалеко	Nyt voit katsoa olevasi perillä, enää ei ole pitkä matka

В. Распутин, «Живи и помни», пер. Э. Адриан

Частоты в СХТ и ПФ глагола *päästä* (1,36 и 1,12) и послелога *perillä/perille* (0,036/0,05 и 0,034/0,09) отличаются незначительно, что также указывает на то, что эти слова не являются каким-либо специализированным средством для перевода с русского языка, а вполне органичны для обоих вариантов финского — оригинального и переводного.

Пространство: ***простор, раздолье, ширь, приволье***

МАС определяет слово *простор* как «1) свободное обширное пространство, 2) свобода, раздолье» (МАС 1984, т. 3.: 527). Однако «эмоциональная составляющая занимает в семантике *простора* еще более важное место. *Простор* — это когда легко дышится, когда можно пойти куда угодно, когда *есть разгуляться где на воле*» (Шмелев 2002: 76). Поэтому выде-

ляемые в МАС два значения слова на самом деле очень тесно связаны и между ними иногда бывает трудно провести границу. *Простор речной волны* из известной песни — это обширное пространство, где человек чувствует себя свободным.

БРФС, видимо ориентируясь на два значения, выделяемые в русских словарях, дает к значению ’обширное пространство’ эквиваленты *lakeus, aukea, aava, avaruus*, а к значению ’свобода’ — *vapaus* (БРФС 1963: 678).

Параллельный корпус показывает, что самым распространенным эквивалентом является безликое *tila* ’место, пространство’, в котором отсутствует как сема ’обширности’ так и сема ’свободы’.

(112)

Внутренние комнаты Свентицких были загромождены лишними вещами, вынесенными из гостиной и зала для большего простора.	Sventitskien sisähuoneet olivat täpötäynnä huonekaluja ja esineitä, jotka oli kannettu vierashuoneesta ja isosta salista jotta niihin tulisi tarpeeksi tilaa.

Б. Пастернак, «Доктор Живаго», пер. Ю. Конкка

Предлагаемые БРФС эквиваленты обозначают ’обширное пространство’. Интересно, что только к слову *aava* ФРС предлагает в качестве русского эквивалента слово *простор*, причем примеров использования этого слова в «ПарРус» зафиксировано не было, скорее всего потому, что ни в одном из употреблений слова *простор* в нашем массиве не идет речь о *водном просторе*, а финское слово *aava* связано именно с морскими или озерными просторами (CD-Perussanakirja). Зато было зафиксировано синонимичное ему существительное *aavikko*, связанное со *степными просторами* и не предлагаемое БРФС в качестве эквивалента к слову *простор*.

(113)

Он был географ, и ему были известны такие просторы, о которых обыкновенные, занятые скучными делами люди даже и не подозревают.	Hän oli maantieteilijä ja tunsi aavikot, joista tavallisilla, ikäviä asioita hoitavilla ihmisillä ei ollut aavistustakaan.

И. Ильф, Е. Петров, «Золотой теленок», пер. А. Аарто)

Однако, представляется, что, хотя ФРС и дает *простор* в качестве эквивалента к этим словам, это связано скорее с устоявшейся в русском языке коллокацией *морской простор*, и вряд ли доказывает, что в финских словах содержится коннотация ’свобода, воля’.

Для остальных финских слов, которые даются БРФС в качестве эквивалентов к *простор*, ФРС предлагает эквиваленты типа *открытое место, открытое поле* или *пространство*. Частотность употребления этих эквивалентов довольно близкая.

Примеры:

(114)

Девятьсот лет просторы России, порождавшие в поверхностном восприятии ощущение душевного размаха, удали и воли, были немой ретортой рабства.

Yhdeksänsataa vuotta Venäjän lakeudet, jotka pinnallisessa tarkastelussa synnyttivät vaikutelman sielun laveudesta, hurjuudesta ja tahdonvoimasta, olivat mykkä orjuuden retortti.

В. Гроссман, «Все течет», пер. Э. Адриан)

(115)

Они вышли на простор, как в громадный, тихо и ровно гудящий цех.

He tulivat aukealle kuin valtavan suureen hiljaa ja tasaisesti humisevaan tehtaaseen.

В. Дудинцев, «Белые одежды», пер. У.-Л. Хейно)

(116)

В лугах и на поле, там все ясно: простор, трава, хлеба, хозяина всегда видно, ходи челноком в широком поиске, ищи, найди, делай стойку и жди приказа.

Niityllä ja pellolla kaikki on selkeätä: avaruutta, ruohoa, viljaa, isäntä aina näkyvissä, sen kun kulkee sukkulana laajalla alueella etsien.

Г. Троепольский, «Белый Бим, черное ухо», пер. Л. Иранто

Кроме существительного *avaruus* нередко употребляется и прилагательное *avara* ’обширный’:

(117)

Солнышко поднялось еще выше, далеко в синем просторе выплыло первое кудлатое облако - предвестник ясного дня.

Aurinko oli jo ehtinyt kavuta ylemmäs, ensimmäinen pörheä poutapilvi, hyvän sään airut, souteli avaraan sineen.

В. Белов, «Привычное дело», пер. Х. Лааксонен

Довольно распространенными эквивалентами оказались не зафиксированные в БРФС прилагательное *väljä* и производное от него существительное *väljyys*, для которого ФРС дает эквиваленты *ширина, разреженность, расплывчатость* (ФРС 1975: 765).

(118)

Эта рыба простор любит.

Se kala pitää väljistä vesistä...

А.П. Чехов, «Злоумышленник», пер. Ю. Конкка)

(119)

Городок был невелик. С любого места в нем тут же за поворотом открывалась хмурая степь, темное небо, просторы войны, просторы революции.

Kaupunki oli pieni. Sen joka paikasta levittäytyivät näkyviin synkkä aro, tumma taivas, sodan väljyydet, vallankumouksen väljyydet.

Б. Пастернак, «Доктор Живаго», пер. Ю. Конкка

Существительное *vapaus* ’свобода’ и прилагательное *vapaa* ’свободный’ используются в качестве эквивалентов к слову простор в обоих значениях.

(120)

В Европе одаренному человеку большой простор для достижения славы.

Euroopassa on lahjakkaalla ihmisellä laaja vapaus saavuttaa mainetta.

Ю. Олеша, «Зависть», пер. Э. Адриан)

(121)

Мне надо туда, на простор. Пусти меня, пусти!

Minä haluan tuonne, vapauteen. Päästä minut, päästä!

Г. Троепольский, «Белый Бим, черное ухо», пер. Л. Иранто)

Это, скорее всего, связано с тем, что переводчикам не удается найти финского слова, которое совмещало бы в себе пространство и свободу, и в результате приходится делать выбор, какой частью семантики слова *простор* пожертвовать. Однако в некоторых случаях переводчики пытаются передать оба компонента. Например, в следующем примере слову *простор* соответствует словосочетание *vapaa maisema* ’свободный пейзаж’.

(122)

Остановились у чайной; гляжу, Митька в магазин с ходу; воротился с бутылкой. "Слезай, — говорит, - вся слобода теперь наша, на простор выехали".	No, pysähdyttiin siinä teetuvalle. Ja eikös mitä: salamana käy Mitja hakemassa sieltä pullon. "Kömmihän alas sieltä. Nyt on maailma auki, päästiin vapaisiin maisemiin."

В. Белов, «Привычное дело», пер. Х. Лааксонен

Слово *раздолье* близко по значению к слову *простор*, однако там несколько более сильна коннотация ’свобода’, и оно больше связано с эмоциональным состоянием человека, чем с пространством, «раздолье ориентировано на активное осуществление любых своих желаний» (Шмелев 2002: 77). Семантические отличия от слова *простор* достаточно тонкие, и при поиске иноязычных соответствий они легко теряются. Набор эквивалентов, предлагаемых БРФС, практически не отличается от эквивалентов к слову *простор*, причем первое значение содержит пояснение (*простор*), а второе значение слова *простор* — (*свобода, раздолье*). Второму значению слова *раздолье* (= *свобода*) уделяется больше внимания: предлагаются эквиваленты *täysi vapaus, vapaa elämä* (БРФС 1963: 704), которые, как нам представляется, не полностью передают значение русского слова. Данные, полученные из «ПарРус», показывают, что переводчики не хотят пользоваться эквивалентами, предлагаемыми словарями, и предпочитают приблизительные эквиваленты *vapaus* ’свобода’ и *tila* ’место, пространство’, однако, как правило, эти слова усиливаются путем каких-либо интенсификаторов. В приведенном ниже примере из Пастернака переводчик усилил слово *vapaus* уже упоминавшимся выше прилагательным *väljä* ’обширный’. В примере из Бакланова, слово *tila* усилено глаголом *temmeltää* ’возиться, шалить’.

(123)

Здесь ведь не Бог весть какое раздолье	Eihän teillä täällä ole, luoja ties, kovinkaan väljää vapautta

Б. Пастернак, «Доктор Живаго», пер. Ю. Конкка

(124)

— Вот вшам раздолье, — сказал Китенев	— Siinä oli täillä tilaa temmeltää, huomautti Kitenev

Г. Бакланов, «Навеки девятнадцатилетние», пер. В. Орлов

Практически не отличаются по подбору финских эквивалентов другие синонимы анализировавшихся выше слов — *приволье* и *ширь*. БРФС дает все тот же набор эквивалентов, в данных, полученных из «ПарРус», в качестве эквивалента используется *väljä*.

(125)

Только там приволья больше и люди богаче живут	Siellä on vain väljyyttä enemmän ja ihmiset elävät varakkaammin

А.П. Чехов, «Мечты», пер. Ю. Конкка)

(126)

Докладчик, объезжавший Сибирь с Военною инструкцией Центрального комитета, витал мыслями в ширях пространств, которые ему еще предстояло охватить	Alustaja, joka kierteli Siperiassa keskuskomitean sotilaallisena valtuutettuna, karkasi ajatuksissaan väljillä alueilla, jotka hänen piti muka valloittaa

Б. Пастернак, «Доктор Живаго», пер. Ю. Конкка

***Русский характер**: гостеприимность / гостеприимство, радушие, хлебосольство и размах*

Гостеприимность / гостеприимство, радушие, хлебосольство

Все эти три слова обозначают хорошее отношение к гостям; уже это показывает, насколько это важно в русской культуре — принять гостей так, чтобы они остались довольны. Слова эти полными синонимами не являются. *Гостеприимность* является наиболее широким по своему значению и распространяется не только на людей, ср. *гостеприимный дом, гостеприимный народ, гостеприимный город* и т.п. *Радушие* предполагает *приветливость, доброжелательность* по отношению к гостям. *Хлебосольство* — это обильное угощение, уют и комфорт. БРФС не делает никакого различия между *гостеприимством* и *хлебосольством*: для обоих слов предлагается один и тот же эквивалент — *vieraanvaraisuus* (БРФС 1963: 142, 932), слово является производным от *vieraanvarat* ’запасы на случай прибытия неожиданных гостей’ (ФРС). Для слова *радушие* словарь дает эквивалент *sydämellisyys* ’сердечность’, *lämminsydämisyys* ’теплая сердечность’ (БРФС 1963: 695), таким образом, в данном случае эквивалентом являются слова с более широкой семантикой.

В корпусе текстов для слов *гостеприимный / гостеприимно / гостеприимство / гостеприимность* и *хлебосольный / хлебосол / хлебосольство* в качестве ПЭ используется только *vieraanvarainen / vieraanvaraisesti / vieraanvaraisuus*.

Гостеприимность:

(127)

Своей рукой раскладывала всем куски, особенно же старалась подложить побольше иностранцу. Тот с интересом наблюдал это тяжеловесное гостеприимство	Omakätisesti hän jakoi kaikille makupaloja yrittäen erityisesti tarjota enemmän ulkomaalaiselle. Tämä tarkkaili kiinnostuneena tätä raskaan sarjan vieraanvaraisuutta

В. Дудинцев, «Белые одежды», пер. У.-Л. Хейно

(128)

В журнале "Будни морзиста" Ляписа встретили гостеприимно	"Sähköttäjän arkipäivä"-nimisessä aikakauslehdessä Ljapis otettiin vastaan vieraanvaraisesti.

И. Ильф, Е. Петров, «Двенадцать стульев», пер. Р. Силванто и Ю. Конкка

(129)

Бим нашел лужицу, каких в любом гостеприимном лесу сколько угодно, и утолил жажду	Bim löysi lätäkön, jollaisia on vaikka millä mitalla jokaisessa vieraanvaraisessa metsässä, ja sammutti janonsa

Г. Троепольский, «Белый Бим, черное ухо», пер. Л. Иранто

Хлебосольство:

(130)

Сын Андрея, Петр, Федоров дед, не походил на своего отца; это был простой степной барин, довольно взбалмошный, крикун и копотун, грубый, но не злой, хлебосол и псовый охотник	Andrein poika ja Fedorin isoisä Petr ei ollut tullut lainkaan isäänsä. Hän oli ollut yksinkertainen maalaisherra, melko omituinen, riitaisa ja hidasälyinen, karkea, vaan ei ilkeä, vieraanvarainen ja suuri koirien ystävä

Дворянское гнездо», пер. У.-Л. Хейн

Для слов *радушный / радушно / радушие* кроме предлагаемого словарем *sydämellinen* используется и *vieraanvarainen*, и кроме того нередко встречается и *ystävällinen* ’дружелюбный’.

(131)

Марья Ивановна принята была моими родителями с тем искренним радушием, которое отличало людей старого века	Vanhempani olivat ottaneet Maria Ivanovnan vastaan todella sydämellisesti, vanhan polven ihmisten tavalla

А.С. Пушкин, «Капитанская дочка», пер. Й. Холло)

(132)

Она ни во что не вмешивалась, радушно принимала гостей и охотно сама выезжала, хотя пудриться, по ее словам, было для нее смертью	Anna Pavlovna ei ollut sekaantunut asioihin, olipahan vain ystävällisesti ottanut vastaan vieraita ja käynyt itsekin mielellään vieraisilla, vaikka puuteri oli hänen omien sanojensa mukaan ollut hänelle kuolemaakin kauhistavampi keksintö

И.С. Тургенев, «Дворянское гнездо», пер. У.-Л. Хейно)

(133)

Он славился во всей округе гостеприимством и радушием	Hän oli kaikkialla lähiseuduilla tunnettu vieraanvaraisuudestaan ja hyväsydämisyydestään

А.С. Пушкин, «Метель», пер. Й. Холло)

(134)

Семейство, могу вас уверить... отец, мать и прочие... люди превосходные, радушные такие, религиозные...	... kaikki koko perhe, voin vakuuttaa teille, isä, äiti ynnä muut ... kaikki ovat erinomaisia, vieraanvaraisia, uskovaisia ihmisiä...

А.П. Чехов, «Свадьба с генералом», пер. Ю. Конкк

Размах

Одной из важных характеристик русского характера является *размах*, который «предполагает отсутствие мелочности и внутренних ограничений, связанных со страхом, скупостью или недостатком фантазии» (Шмелев 2002: 96).

БРФС вообще специально не выделяет у слова *размах* этого значения, самым близким оказывается значение с пояснением *(ширь)*, для которого даются в качестве эквивалентов *mittavuus* 'объемность' и *laajuus* 'ширина' (БРФС 1963: 706), использование которых при переводе выражений типа *свадьба с размахом* весьма проблематично. Переводчики используют совсем другие эквиваленты, например, *vauhti* 'скорость, быстрота', *valtavuus* 'масштабность', *tarmo* 'энергия', *laveus* 'обширность, простор' и некоторые другие.

Примеры:

(135)

— Нет, Фагот, — возражал кот, — бал имеет свою прелесть и размах	— Ehei, Fagot, kissa väitti vastaan. — Kyllä tanssiaisillakin on omat hyvät puolensa, on upeutta ja vauhtia!

М.А. Булгаков, «Мастер и Маргарита», пер. У.-Л. Хейно)

(136)

А эта пустоголовая юность, идущая нам на смену, словно бы и не замечает тайн бытия. Ей недостает размаха и инициативы, и я вообще сомневаюсь, есть ли у них всех чего-нибудь в мозгах	Mutta tämä tyhjäpäinen nuoriso, joka on astumassa tilallemme, ei tunnu huomaavankaan elämän salaisuuksia. Siltä puuttuu tarmoa ja aloitteellisuutta, ja yleensäkin epäilen, onko heillä kaikilla mitään aivoissaan

В. Ерофеев, «Москва-Петушки», пер. Э. Адриан)

(137)

Здесь все, что только может понадобиться элегантному гражданину моих лет и моего размаха	Tässä on kaikkea, mikä on tarpeen minun ikäiselleni ja minunlaiselleni rivakkaotteiselle hienolle kansalaiselle

И. Ильф, Е. Петров, «Золотой теленок», пер. А. Аарто

Соответствие картин мира и словарь

В данном и предыдущем разделах были показаны расхождения между русским и финским лексиконами. Возникает вопрос: во всех ли случаях отношения между словами разных языков столь сложны и противоречивы? Существуют ли однозначные ПЭ-пары? В разделе 3.5 описывался поиск ПЭ-пар в автоматическом режиме. Сама возможность реализации поиска эквивалентов на основе совместной встречаемости говорит в пользу существования стандартных ПЭ-пар. Однако, как показали наши эксперименты, количество их относительно невелико, хотя следует делать поправку на соответствия вида «словосочетание — слово», «слово — слово-

сочетание» и «словосочетание — словосочетание», которые наш алгоритм не выявляет. Программа нашла в «ПарРус» лишь порядка 1 800 ПЭ-пар вида «слово — слово» (полный список см. в приложении 5 настоящей работы). Однако не все слова этого списка можно назвать однозначными эквивалентами: у многих из них есть другие соответствия, однако программа обнаружила лишь один или несколько наиболее часто встречающихся. Так, для слова *бульвар* найдено два финских соответствия *bulevardi* и *puistokatu*, для слова *классический — klassinen* и *klassillinen*. Во многих случаях многие из слов, использовавшихся в качестве ПЭ, остаются ненайденными. Например, для существительного *опушка* найдено финское соответствие *aho*, которое употреблено в качестве ПЭ лишь в 5 случаях из 37.

Наибольший интерес в данном случае представляют русские слова, имеющие в финских переводах только один эквивалент. Для того чтобы получить список таких слов, был сделан запрос к полученному из «ПарРус» списку ПЭ-пар, с помощью которого были выделены те пары, в которых количество совпадений с финским эквивалентом составило 90% и более. Допускалось 10% несовпадений, поскольку во многих случаях возможна местоименная замена, использование имен собственных, пропуск слова при переводе и т.п. В результате было получено 218 ПЭ-пар. Список был проверен по БРФС на предмет совпадения нашего списка с эквивалентами, предлагаемыми словарем. Для 168 русских слов списка словарь давал только один эквивалент, совпадающий с зафиксированным в «ПарРус», некоторые из этих пар приводятся в таблице 33.

Таблица 33. Некоторые ПЭ-пары «ПарРус» с единственным финским словарным эквивалентом для русского слова

Русское слово	Финский эквивалент	Частота русск.	Частота финск.	Пересечение
волк	susi	114	148	110
воспитать	kasvattaa	20	139	20
выиграть	voittaa	75	333	69
высочество	korkeus	24	141	24
герцог	herttua	20	24	20
гитара	kitara	27	29	26
грабли	harava	22	20	20
гусар	husaari	64	71	62
датчанин	tanskalainen	40	48	40
дворник	talonmies	193	208	176
двухэтажный	kaksikerroksinen	25	25	23
детдом	lastenkoti	30	66	28
дикобраз	piikkisika	23	23	23
дипломат	diplomaatti	39	40	36
договор	sopimus	46	156	44
доллар	dollari	23	22	21

Русское слово	Финский эквивалент	Частота русск.	Частота финск.	Пересечение
донос	ilmianto	44	61	41
дуэль	kaksintaistelu	79	109	73
замужем	naimisissa	24	105	22
зевнуть	haukotella	59	105	56
землемер	maanmittari	28	28	28
икона	ikoni	50	71	46
инженер	insinööri	153	156	140

Для 45 ПЭ-пар словарь предлагал несколько эквивалентов, лишь один из которых использовался в качестве переводного эквивалента при переводе с русского на финский. Фрагмент этого списка приведен в табл. 34.

Таблица 34. Некоторые ПЭ-пары «ПарРус» с несколькими финскими словарными эквивалентами для русского слова

Русское слово	Эквивалент из «ПарРус»	Частота русск.	Частота финск.	Пересечение	Другие эквиваленты
автомобиль	auto	145	643	139	automobiilli, voimavaunu
барон	paroni	33	39	33	vapaaherra
больница	sairaala	194	378	190	sairashuone
виселица	hirsipuu	24	35	22	hirttopuu
гибрид	hybridi	46	44	44	sekamuoto, sekasinnös, ristisiitos, risteytys
голубь	kyyhkynen	22	44	20	kyyhky
губернатор	kuvernööri	50	55	48	maaherra
десятилетие	vuosikymmen	25	30	24	kymmenvuotiskausi
доброволец	vapaaehtoinen	37	64	37	tarjokas
еврей	juutalainen	78	147	72	heprealainen
жертвовать	uhrata	29	107	27	panna alttiiksi
кивнуть	nyökätä	204	254	186	nyökäyttää
колония	siirtola	78	114	74	siirtokunta, siirtoasutus, kolonia
комплимент	kohteliaisuus	24	71	24	komplimangi, komplimentti
купец	kauppias	103	131	98	kauppamies
лекарство	lääke	58	66	54	rohto
лисица	kettu	22	42	20	repo
матрас	patja	32	96	32	matrassi

Эквиваленты, предлагаемые БФРС, но не использованные в текстах «ПарРус», были проверены по СХТ. Из 25 слов и выражений, приведенных в табл. 34, в корпусе были представлены 9: *panna alttiiksi* (2 примера), *kauppamies* (6), *kyyhky* (11), *maaherra* (1), *nyökäyttää* (8), *risteytys* (1), *ristisiitos* (1), *rohto* (10), *tarjokas* (3). Отсутствие других, в связи с неболь-

шими размерами корпуса, не говорит о том, что эти слова вообще не используются в современном финском языке, однако этот факт позволяет предположить, что часть эквивалентов, предлагаемых словарями, по тем или иным причинам оказывается мало пригодными для перевода, прежде всего, в связи с тем, что являются малоупотребительными (например, *voimavaunu* или *komplimangi*).

Шесть слов списка (*батька, жид, зэк, лохматка, первосвященник, пошехонец*) отсутствовали в БФРС. Для двух слов — *государыня* и *котлован* — переводчики использовали эквиваленты, отсутствовавшие в БФРС. Для слова *государыня* БФРС предлагает *hallitsijatar*, а в «ПарРус» используется *keisarinna*; для слова *котлован* БФРС дает *peruskuoppa*, а в «ПарРус» — отстутствующий в БФРС эквивалент *monttu*.

Итак, полученные нами данные подтверждают распространенное на бытовом уровне представление о том, что переводчик при переводе использует в качестве эквивалентов именно соответствия, предлагаемые словарями. В некоторых случаях переводчики отдают предпочтение одному из нескольких эквивалентов, предлагаемых словарем, являющемуся либо более «стандартным» (*больница — sairaala*, эквивалент *sairashuone* не зафиксирован) либо более похожим на слово ИЯ (*барон — paroni*, эквивалент *vapaaherra* не встретился). Тем не менее, ориентация переводчика на словари на практике оказывается далеко не столь сильной, как можно было бы ожидать. По-видимому, чем дальше оказываются ИЯ и ПЯ друг от друга, тем меньше в переводах «стандартных» соответствий и тем чаще ПЭ-парой становится пара слов, уникальная для данного контекста.

Выводы

Рассмотренный в данном разделе языковой материал, хотя и является достаточно фрагментарным, все же наглядно демонстрирует, что параллельный корпус художественных текстов вполне может использоваться для сравнения языковых картин мира. В тех случаях, когда затрагивается область, «плохо проработанная» в ПЯ, возникают переводческие трудности, а при переводе возможны информационные потери либо потери в экспрессивности. В тех случаях, когда сравниваемые фрагменты картины мира ИЯ и ПЯ оказываются одинаково «подробными», то трудности минимальны.

В некоторых случаях переводчики идут на поводу у словарей и используют стандартные эквиваленты, иногда несколько огрубляя и упрощая переводимое сообщение. Однако, даже небольшой материал, который был проанализирован, показывает, что гораздо чаще находятся более тонкие и более точные способы передачи сообщения ИЯ. В то же время нередко фиксируются случаи использования в качестве эквивалентов слов и выражений, значительно более далеких от ИЯ, чем предлагаемые словарями

эквиваленты, что, по-видимому, чаще всего бывает связано со стилистической неприемлемостью или искусственностью предлагаемых словарем эквивалентов.

Таким образом, параллельные тексты, этот «мостик» между двумя языками, могут, наряду со словарями, предоставлять данные о совместимости картин мира двух языков и об имеющихся лакунах. Причем параллельные тексты и словари взаимодополняют друг друга. С одной стороны, при составлении словарей в какой-то мере используется переводческий опыт, с другой — отдается дань существующей традиции, и, кроме того, имеет место и «конструирование» переводных эквивалентов. Переводчики в свою очередь в каких-то случаях следуют рекомендациям словарей, а в каких-то находят свои собственные решения. Таким образом, используя оба источника данных, можно получить более объективную картину связей между лексическими единицами двух языков.

Кроме применявшегося нами изучения ключевых слов культуры через перевод художественных текстов на другие языки достаточно интересным представляется исследование функционирование этих слов в переводах с других языков. Например, одним из ключевых понятий финской языковой картины мира по праву считается *sisu* ’сильный характер’, которому в русском языке нет однозначного соответствия, например, ФРС дает в списке эквивалентов такие довольно сильно отличающиеся по семантике слова, как *нрав, характер, упорство, стойкость* и *упрямство*. В СХТ зафиксирован 21 случай употребления этого слова, а в бóльшем по объему ПФ — всего 9. Интересно, однако, что слово это все же употребляется в переводных текстах. При переводе с русского языка слово *sisu* довольно часто используется при переводе выражения *хватить духу*:

(138)

Кавалерову не хватало духу проникнуть за триумфальное кольцо	Kavalerovilla ei ollut sisua tunkeutua riemuitševan renkaan sisään

Ю. Олеша, «Зависть», пер. Э. Адриан

Исследование ключевых слов культуры не только позволяет получать новые данные об особенностях языков, но и в конечном итоге позволит совершенствовать словари — как толковые, так и двуязычные, — а также повышать качество художественных переводов.

Заключение

Подведем итоги нашему исследованию.

Главным практическим результатом работы было составление параллельного русско-финского корпуса художественных текстов. В корпус вошли произведения русской художественной литературы XIX-XX вв. и их переводы на финский язык, выполненные в 1960–1990 гг. Объем корпуса — 2,2 млн. словоупотреблений в каждом из субкорпусов — является достаточным для исследований в области переводоведения, а также сопоставительного изучения языков и культуры. Для лексикографической работы, как правило, требуются корпуса текстов большего объема, тем не менее «ПарРус» может использоваться и как источник для лексикографической работы. Полученный материал представляет большой интерес в том плане, что в корпусе представлена пара неродственных языков с долгой историей межъязыковых контактов. С переводоведческой точки зрения тексты интересны тем, что в художественном переводе с русского на финский существует достаточно долгая, более чем столетняя традиция.

При составлении ПКТ было решено включать в корпус целые тексты, а не их фрагменты, как это делалось во многих корпусных проектах последнего десятилетия (см. раздел 1.3 настоящей работы). Полнотекстовость корпуса расширяет сферу его применения, например, появляется больше возможностей для исследования переводов произведений конкретных писателей или переводов, выполненных определенными переводчиками; только полнотекстовый ПКТ может дать материал для исследований в области интертекста, нарратологии и многих других пограничных с литературоведением областях. Полнотекстовый параллельный корпус художественных текстов является одновременно поисковой системой, с помощью которой можно получать переводы цитат.

В начале работы ставился вопрос о включении в корпус и финско-русских параллельных текстов. От этой идеи пришлось отказаться, так как с финского на русский переводилось значительно меньше текстов, и получение сравнимого по объему, жанрам и тематике финско-русского массива представляется проблематичным. Поэтому работы по сравнению языка финских переводов с языком финских оригинальных текстов проводились с привлечением одноязычного корпуса финских текстов в качестве контрольного массива.

При определении структуры «ПарРус» и решении вопросов о критериях отбора текстов использовалась библиография художественных переводов с

русского на финский, составленная на кафедре русского языка Хельсинкского университета, работа над которой была продолжена в рамках данного исследования. Использование библиографии позволило получить классификацию текстов и упорядочить отбор текстов в корпус. Тем не менее, полученный массив трудно назвать репрезентативным в классическом понимании этого термина. С другой стороны, вопрос о репрезентативности любого ПКТ, в том числе — составленного из текстовых фрагментов, является гораздо более серьезной проблемой, чем репрезентативность одноязычного корпуса (см., например, Johansson 2002). В целом, данные «ПарРус» позволяют судить об общих тенденциях, к полученным из корпуса статистическим данным следует относиться осторожно. Количественные данные принимались во внимание лишь в случае наличия явной закономерности, проявляющейся в разных текстах массива. Следует отметить, что «ПарРус», несомненно, является прекрасным источником языковых примеров.

В начале работы было поставлено довольно много технических задач, требовавших решения: наряду с собственно составлением корпуса текстов разрабатывалось и программное обеспечение. Был разработан довольно большой пакет программ, позволяющий выполнять как стандартные операции (построение словников и конкордансов, получение списков коллокаций и т.п.), так и менее тривиальные операции, как лемматизация русских и финских словников, стыковка параллельных текстов на уровне абзацев, поиск переводных эквивалентов.

Автор отдает себе отчет, что многие из задач, поставленных в диссертации, не могли быть решены до конца и вполне могли бы стать темами отдельных диссертаций. Однако в случае сужения темы не удалось бы получить целостной системы, работающей с довольно большим массивом реальных данных. Нередко приходилось идти по пути поиска временного решения крупной проблемы: лемматизаторы системы контекстно-свободные, стыковщик текстов работает на уровне абзацев, поиск переводных эквивалентов ищет только пары вида «слово — слово». Тем не менее, разработанный в ходе выполнения работы пакет программ уже является достаточно надежным и стабильно работающим исследовательским инструментом, с помощью которого можно выполнять разного рода рутинную обработку текстов корпуса и получать из них интересные данные. Каждый из компонентов может совершенствоваться как в качестве отдельного компонента, так и в составе «ПарРус».

В процессе выполнения работы было решено не ограничиваться решением задач, связанных с составлением корпуса текстов и его обслуживанием, но попытаться показать возможные сферы применения данных ПКТ, то есть выполнить выборочный анализ некоторых аспектов полученного материала.

Материал корпуса текстов позволил искать подходы к решению вопроса о сохранении информации оригинала в переводе с точки зрения теории

информации. Проведенный анализ подтверждает мысль Ю. Найды о том, что перевод, как правило, оказывается длиннее исходного текста.

ПКТ является хорошим источником данных для исследования особенностей языка переводных текстов. Субкорпус финских переводов «ПарРус» сравнивался с оригинальными финскими художественными текстами Савонлиннского корпуса текстов, гипотезы о влиянии ИЯ проверялись на русских исходных текстах. В рамках данной работы исследовались различия в пунктуации в переводах с русского языка по сравнению с исходными текстами и с оригинальными финскими текстами. Гипотеза о влиянии пунктуации исходного текста на пунктуацию перевода наглядно подтвердилась, кроме того еще раз подтвердилась тенденция к сохранению структуры исходного текста в переводе.

Исследование лексики и некоторых грамматических явлений в оригинальных и переводных финских текстах также показало, что язык финских переводов с русского несомненно отличается от языка оригинальных финских художественных текстов. Установленные отличия в основном объясняется влиянием ИЯ, устранить которое не удается даже в том случае, когда переводчик допускает отклонения от исходного текста во имя красоты стиля. Для переводов с разных языков формируются разные варианты ПЯ со своими особенностями: в одном варианте может наблюдаться избыток форм сослагательного наклонения, в другом — слишком высокая частотность причастных конструкций и т.п. Таким образом, общих особенностей переводов на данный язык со всех языков, скорее всего, немного и все они настолько абстрактны, что практическая ценность таких универсалий минимальна. В то же время, исследования влияния данного ИЯ на ПЯ может иметь большую практическую ценность как для преподавания перевода, так и для работы переводчика художественной литературы.

Завершает работу выборочный анализ лексических переводных соответствий в корпусе. Данные корпуса сравнивались с данными русско-финских и финско-русских словарей. Анализ, с одной стороны, показывает, что ПКТ является довольно полезным дополнением к двуязычным словарям как в качестве справочной системы для переводчика, так и в качестве источника данных для лексикографа. Данные «ПарРус» далеко не всегда совпадали с данными двуязычных словарей, переводчики нередко используют свои — иногда более, иногда менее удачные — переводные эквиваленты. Однако, с другой стороны, в процессе анализа было обнаружено и влияние словарей на переводчиков, что иногда приводит к распространению неудачных переводных эквивалентов, например, рассматривавшееся в работе соответствие *студент — ylioppilas*. Наряду с чисто лексикографической и лексикологической работой ПКТ может использоваться и для изучения языковой картины мира ИЯ и ее «отражения» в ПЯ.

Главной задачей данной диссертации было создание русско-финского ПКТ и разработка средств для получения из него информации. Эта задача

выполнена. Но тема этим не исчерпана. Остается актуальной задача создания финско-русского корпуса художественных текстов, для решения которой можно воспользоваться разработанной в данной работе методикой, а также программным обеспечением «ПарРус».

Другой важный вопрос — дальнейшее развитие программного обеспечения корпуса. Здесь актуальна задача совершенствования работы лемматизаторов, в первую очередь — снятие грамматической омонимии в автоматическом режиме и лемматизация слов, не зафиксированных в словаре основ. Для финского морфоанализа отдельно стоит вопрос об анализе сложных слов. Актуален также вопрос о грамматической разметке текстов в автоматическом режиме, причем для данной пары языков важным представляется получить не только разметку по частям речи (POS tagging), но и разметку по грамматическим формам.

Программа-стыковщик работает на уровне абзацев, однако уже в настоящее время мы достаточно близки к получению стыковки на уровне предложений. Такой стыковщик второго порядка может использовать результаты стыковки на уровне абзацев и русско-финский глоссарий, полученный в автоматическом режиме.

Автоматическое получение глоссариев на базе ПКТ — другая интереснейшая исследовательская задача. Актуальными остаются вопросы поиска эквивалентов типа «словосочетание — слово», «слово — словосочетание» и «словосочетание — словосочетание». Неясно пока, как организовать поиск эквивалентов для низкочастотных слов.

Другое направление для дальнейших исследований — изучение самих текстов «ПарРус». Даже нынешнее программное обеспечение дает богатые возможности исследования переводных эквивалентов в корпусе, получать данные по сочетаемости слов, их частотности и т.п. Грамматическая разметка корпуса откроет новые возможности в исследовании грамматических особенностей текстов.

Параллельные корпуса текстов представляются важными источниками данных, которые в итоге позволят поднять на качественно новый уровень как теоретические исследования в области переводоведения, сравнительного языкознания и прикладной лингвистики, так и практическую деятельность, связанную с переводом, например, двуязычную лексикографию, преподавание перевода и работу переводчиков-практиков.

Список акронимов

БДТ = база данных с текстами. Компонент систем «КОКОС» и «КОКОС-П».

БДП = база данных с программами. Компонент систем «КОКОС» и «КОКОС-П».

БРФС = Большой русско-финский словарь. Куусинен М., Олликайнен В. *Большой русско-финский словарь*. Porvoo-Helsinki-Juva: WSOY, 1963.

ГС = грамматический словарь. Компонент системы «ЛемКС-Ф».

ИН — идентификационный номер в таблицах словоформ и лемм в пакетах «КОКОС» и «КОКОС-П».

ИЯ — исходный язык; язык, с которого выполняется перевод.

КИТ — каталог исходных текстов, компонент системы «КОКОС-П».

КПТ — каталог переводных текстов, компонент системы «КОКОС-П».

«КОКОС» — пакет программ для обработки одноязычных корпусов текстов.

«КОКОС-П» — версия «КОКОС», поддерживающая параллельные корпуса текстов.

«ЛемКС-Р» — контекстно-свободный лемматизатор русского языка. Является компонентом систем КОКОС и КОКОС-П.

«ЛемКС-Ф» — контекстно-свободный лемматизатор финского языка. Является компонентом систем «КОКОС» и «КОКОС-П».

МАС = Малый академический словарь. *Толковый словарь русского языка: в 4-х тт*. АН СССР, Институт русского языка. Москва: Русский язык.

МКП = маркер конца предложения (например, точка).

«ПарРус» — параллельный русско-финский корпус художественных текстов. Составляется на кафедре русского языка Отделения переводоведения Тамперского университета.

ПКТ = параллельный корпус текстов.

ПЭ = переводной эквивалент.

ПЭ-пара — слово ИЯ и его переводной эквивалент в ПЯ.

ПФ — субкорпус корпуса «ПарРус», включающий переводы с русского языка на финский.

ПЯ — «переводящий язык»; язык, на который выполняется перевод.

СДП — средняя длина предложения в словоупотреблений; отношение длины текста в словоупотреблениях к количеству предложений в тексте

СС — type/token ratio: отношение количества словоупотреблений к количеству словоформ.

СУБД — система управления базами данных.

СХТ (Савонлиннские художественные тексты) — субкорпус Савонлиннского корпуса текстов (*The Savonlinna Corpus of Translated Finnish*. Savonlinna School of Translation Studies, University of Joensuu, 2001), включающий оригинальные художественные тексты на финском языке, опубликованные в 1990-х гг.

Используется в настоящем исследовании в качестве контрольного массива для сравнения его данных с данными ПФ, финского субкорпуса ПКТ «ПарРус».

«ТамРус» — корпус русских художественных текстов. Составляется на кафедре русского языка Отделения переводоведения Тамперского университета.

ТНС — таблица слов с нестандартным словоизменением, компонент системы «ЛемКС-Р».

ТЛ — таблица, в которой хранятся правила получения псевдолемм из основ в «ЛемКС-Ф».

ТС — таблица суффиксов в «ЛемКС-Ф».

ТТ — таблица с текстами, компонент систем «КОКОС» и «КОКОС-П».

ФРС = Финско-русский словарь. Вахрос И., Щербаков А. *Финско-русский словарь*. Москва: Русский язык, 1977.

Глоссарий терминов

Аболютная частота — количество вхождений слова (словоформы) в данный текст/субкорпус/корпус.

Аннотированный/размеченный корпус текстов (tagged corpus) — корпус текстов, в котором содержатся специальные метки, позволяющие получать из корпуса данные (статистика, языковые примеры и др.) по каким-либо лингвистическим явлениям (часть речи, грамматическая форма, синтаксическая функция и т.п.).

Битекст — фрагмент исходного текста и соответствующий ему фрагмент перевода.

Вес текста — условная весовая мера, присваиваемая текстам исследуемого массива на основе их объема.

Грамматический разметчик, «тэггер» (tagger) — программа, выполняющая в автоматическом режиме грамматическую разметку текстов корпуса.

Диахронический корпус текстов — корпус текстов, в который включаются тексты, созданные в разные исторические периоды развития языка.

Индексирование корпуса текстов — составление в автоматическом режиме списков адресов каждого слова текста, индекса.

Коллокация / коллокат — слово или словоформа, встречающаяся в качестве ближнего соседа данного слова (словоформы).

Конкорданс — получаемый в автоматическом режиме набор контекстов для заданного явления (слово / словосочетание / грамматическая форма и др.).

Коэффициент Дайса (Dice coefficient) — коэффициент совместной встречаемости, используется как для поиска ПЭ-пар, так и для поиска похожих слов; в «КОКОС-П» для поиска ПЭ-пар на основе графического сходства.

Коэффициент ИЯ-ПЯ — среднее отношение длины сообщения на ИЯ к длине его перевода на ПЯ (в словоупотреблениях). Используется в «КОКОС-П».

Коэффициент Кульчинского (Kulczinski coefficient) — коэффициент совместной встречаемости слов исходного текста и перевода, используется в «КОКОС-П» для поиска ПЭ-пар на основе совместной встречаемости.

Лемма — начальная (словарная) форма для данной словоформы, например, лемма для формы *работает — работать*.

Лемматизация — процесс поиска начальных форм для словоформ.

Многоязычный корпус текстов (multilingual corpus) — корпус текстов, включающий в себя текстовые массивы на разных языках.

Мониторный корпус текстов (monitor corpus) — постоянно пополняемый и обновляемый корпус текстов, создаваемый в целях мониторинга представляемого корпусом подъязыка (sublanguage) или языка в целом.

Относительная частота — отношение абсолютной частоты слова (словоформы) к объему корпуса (в словоупотреблениях).

Парсер — компьютерная программа, выполняющая автоматическую обработку текста на синтаксическом или семантическом уровне.

Полнотекстовый корпус текстов — корпус текстов, состоящий из целых текстов, а не фрагментов.

Программа-стыковщик, стыковщик (aligner) — программа, выполняющая стыковку параллельных текстов в автоматическом или полуавтоматическом режиме.

Псевдолемма — в лемматизаторах «ЛемКС-Р» и «ЛемКС-Ф» фиктивная лемма, порождаемая по правилам, содержащимся в лемматизаторе; например, для финской формы *kaupassa* порождается псевдолемма *kaupa* (не учтено чередование pp/p), для которой в словаре основ задана правильная лемма —*kauppa*.

Псевдоокончание — в лемматизаторе «ЛемКС-Р» фиктивное окончание, использующееся в грамматических таблицах.

Псевдооснова — в лемматизаторах «ЛемКС-Р» и «ЛемКС-Ф» фиктивная основа, остаток после отделения от слова псевдоокончания (псевдосуффикса).

Псевдосуффикс — в лемматизаторе «ЛемКС-Ф» фиктивный суффикс, использующийся в грамматических таблицах.

ПЭ-пара — слово ИЯ и его переводной эквивалент в ПЯ.

Репрезентативность (представительность) корпуса текстов (representativeness) — степень представленности в корпусе текстов всех типов текстов, существующих в описываемом языке (подъязыке).

Синхронический корпус текстов — корпус текстов, в который включаются только тексты, созданные в течение одного и того же короткого периода времени (например, в течение одного года).

Субкорпус — группа текстов корпуса, объединяемых на основе совпадения какого-либо параметра (язык, жанр и т.п.).

Сравнительный корпус текстов (comparable corpus) — корпус текстов, в состав которого входят оригинальные тексты на каком-либо языке и переводы на этот язык с других языков.

Стыковка текстов (aligning) — установление соответствия фрагментов исходного текста фрагментам перевода, выполняется вручную или автоматически.

Стыковщик (aliner) — компьютерная программа, выполняющая в автоматическом или полуавтоматическом режиме стыковку перевода с исходным текстом.

Текстовая масса — сумма весов всех текстов, относящихся к какой-либо одной группе (например, написанных одним и тем же автором или относящихся к одному и тому же жанру и т.п.).

Утилита — часть пакета программ или отдельная небольшая программа, выполняющая какую-либо несложную операцию.

Электронная антология — полнотекстовое собрание электронных текстов, при составлении предпочтение отдается культурно значимым текстам.

Энтропия — принятая в математической теории информации мера количества информации, показывающая степень свободы выбора на очередном шаге порождения сообщения.

Литература

Altenberg, Bengt, and Aijmer, Karin. 2000: The English-Swedish Parallel Corpus: A resource for contrastive research and translation studies. In *Corpus linguistics and linguistic theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau, 1999*, ed. by Christian Mair and Marianne Hundt, 15–33. Amsterdam/Atlanta: Rodopi.

Ahrenberg, Lars and Merkel, Magnus and Sågvall, Hein Anna and Tiedemann, Jörg 2000: Evaluation of Word Alignment Systems. *Proceedings of LREC 2000.* Athens

Atkins B.T.S. & Levin, Beth & Zampolli A. Computational Approaches to the Lexicon: An Overview. In: B.T.S. Atkins and A. Zampolli (eds.). *Computational Approaches to the Lexicon*. Oxford : Oxford University Press. Pp. 17–48.

Atkins, Sue and Clear, Jeremy and Ostler, Nicholas 1992: Corpus Design Criteria. *Literary and Linguistic Computing.* 7(1), 1–16.

Atsushi, Ando and Yasuo, Urai, and Tetsuo, Mochizuki (eds). 1994: *A Concordance to Dostoevski's "Crime and Punishment"*. Vol. 1–3. — Sapporo, The Slavic Research Center, Hokkaido University.

Baker, Mona 1995: Corpora in Translation Studies: an overview and some suggestions for future research. *Target 7:2.* Amsterdam: John Benjamins, 223–243.

Baker, Mona 1999: The role of Corpora in Investigating the Linguistic Behaviour of Professional Translators. *International Journal of Corpus Linguistics* 4(2), 281—298.

Barlow, Michael 1995. ParaConc: A Concordance for Parallel Texts. *Computers & Texts.* Vol. 10.

Bassnet-McGuire, Susan 1991: *Translation Studies*. London: Routledge.

de Beaugrande, Robert 2001: Interpreting the Discourse of H.G. Widdowson: A Corpus-based Critical Discourse Analysis. *Applied Linguistics* 22/1: 104–121.

Biber, Douglas and Conrad, Susan and Reppen, Randi 1998: *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.

Biber, Hanno & Breiteneder, Evelyn & Moerth, Karlheinz 2002: The Austrian Academy Corpus — Digital Resources and Textual Studies. *New Directions in Humanities Computing. Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities ALLC / ACH 2002.* University of Tübingen 24 – 28 July, 2002.

Borin, Lars 2002: ... and never the twain shall meet? In: Borin, Lars (ed.). *Parallel Corpora, Parallel Worlds.* Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22–23 April, 1999. Amsterdam — New York, NY: Rodopi. Pp. 1–43.

Borin, Lars 2002a: Alignment and tagging. In: Borin, Lars (ed.). *Parallel Corpora, Parallel Worlds.* Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22–23 April, 1999. Amsterdam — New York, NY: Rodopi. Pp. 207–218.

Bowker, Lynne 2000: Towards a Methodology for Exploiting Specialized Target Language Corpora as Translation Resources. *International Journal of Corpus Linguistics.* Amsterdam: John Benjamins. 5(1), 17–52.

Burnard, Lou 1992: Tools and Techniques for Computer-assisted Text Processing. In C.S. Butler (ed.) *Computers and Written Texts.* Cambridge, Massachusetts: Blackwell. Pp. 1–28.

Catford J.C. 1965: *A Linguistic Theory of Translation*. London: Oxford University Press.

CD-PERUSSANAKIRJA 1997 Kotimaisten kielten tutkimuskeskus Kustantaja: Oy Edita Ab

Chesterman, Andrew 1997: *Memes of translation: the spread of ideas in translation theory.* Amsterdam/Philadelphia: John Benjamins.

Čmejrek, Martin and Cuřín, Jan 2001: Automatic Extraction of Terminological Translation Lexicon from Czech-English Parallel Texts. *International Journal of Corpus Linguistics*. Vol. 6 (Special Issue). Amsterdam / Philadalphia: John Benjamins, pp. 1–12.

Erjavec, Tomaž 1999: The ELAN Slovene-English Aligned Corpus. *Proceedings of the Machine Translation Summit VII.* Singapore. Pp. 349–357.

Gale W.A. and Church K.W. 1991: Concordances for Parallel Texts. *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research Using Corpora.* Oxford, pp. 40–62.

Dagan I., Itai A., and Shwall U. 1991: Two languages are more informative than one. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. pp. 130–137.

Engwall, Gunnel 1994: Not Chance but Choice: Criteria in Corpus Creation. In: B.T.S. Atkins and A. Zampolli (eds.). *Computational Approaches to the Lexicon*. Pp. 49–82.

Eskola, Sari & Tommola, Hannu 2000: Komitatiivi ja kääntämisen lainalaisuudet. *Erikoiskielet ja käännösteoria. VAKKI-symposiumi XX. Vaasa 11.–13.02.2000.* Vaasa. S. 96–109.

Eskola, Sari 2002: *Syntetisoivat rakenteet käännössuomessa. Suomennetun kauno-kirjallisuuden ominaispiirteiden tarkastelua korpusmenetelmillä.* Joensuun yliopisto, Joensuu.

Fillmore, Charles J. 1992: "Corpus Linguistics" or "Computer-aided armchair linguistics". In Jan Svartvik (ed.) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991.* — Berlin - New York: Mouton de Gruyter, 35–61.

Francis W. 1992: Language Corpora B.C. In: Jan Svartvik (ed.) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991.* — Berlin – New York: Mouton de Gruyter, 17–35.

Francis W. N. and Kučera H. 1964: BROWN CORPUS MANUAL OF INFORMATION to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Brown University Providence, Rhode Island Department of Linguistics

Gale W.A. and Church K.W. 1993: A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics,* 19(1), 75–102.

Gale W.A., Church K.W. and Yarowski D. 1992: A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities*, 26(5–6), pp. 415–439.

Gaussier E. and Langé J.-M. 1994: Some methods for the Extraction of Bilingual Terminology. in: Jones D. (ed.). *Proceedings of the International Conference on*

New Methods in Language Processing (NewLaP), 14–16 September 1994, UMIST, Manchester, pp. 242–247.

Gellerstam, Martin 1992: Modern Swedish Text Corpora. In J. Svartvik (ed.) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991.* Berlin - New York: Mouton de Gruyter, 149–164.

Hein, Anna Sågvall 2002: The PLUG project: parallel corpora in Linköping, Uppsala, Göteborg: aims and achievements. In: Borin, Lars (ed.). *Parallel Corpora, Parallel Worlds.* Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22–23 April, 1999. Amsterdam — New York, NY: Rodopi. Pp. 61–78.

Hockey, Susan 2000: *Electronic Texts in the Humanities: Principles and Practice.* New York: Oxford University Press.

Hofland, Knut and Johansson, Stig 1998: The Translation Corpus Aligner: A program for automatic alignment of parallel texts. In Johansson, Stig and Signe Oksefjell (eds.). *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies.* Amsterdam: Rodopi. 87–100.

Holmes, James S. 1988: *Translated! Papers on Literary Translation and Translation Studies*. Amsterdam: Rodopi.

House, Juliane 1981: *A Model for Translation Quality Assessment*. Tübingen: Narr.

Itkonen, Tervo 1995: *Kieliopas*. Kirjayhtymä, Helsinki.

Jakobson, Roman 1989: On linguistic aspects of translation. In: A. Chesterman (ed.). *Readings in Translation Theory*. Finn Lectura. Pp. 53–60.

Jantunen, Jarmo 2001: "Tärkeä seikka" ja "keskeinen kysymys": mitä korpus-lingvistinen analyysi paljastaa lähisynonyymeistä? *Virittäjä* 2, 170–192.

Johansson, Stig 1994: Encoding a Corpus in Machine-Readable Form: The Approach of the Text Encoding Initiative. In: B.T.S. Atkins and A. Zampolli (eds.). *Computational Approaches to the Lexicon*. Oxford : Oxford University Press. Pp. 83–102.

Johansson, Stig 2002: Towards a multilingual corpus for contrastive analysis and translation studies. In: Borin, Lars (ed.). *Parallel Corpora, Parallel Worlds.* Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22–23 April, 1999. Amsterdam — New York, NY: Rodopi. Pp. 47–59.

Kay M. and Röscheisen M. 1993: Text-Translation Alignment. *Computational Linguistics*, 19(1), pp. 121–142.

Kilgariff A. 1999: Comparing Corpora. *International Journal of Corpus Linguistics*. – Philadelphia: John Benjamins. – Vol. 4(2).

Koller W. 1992: «A linguistic approach to literary translation: its range and limitation». *Литература и перевод: проблемы теории*. Москва, Прогресс, стр. 85–95.

Koskenniemi K. 1983: *Two-Level Morphology. A General Computational Model for Word-Form Recognition and Production.* Publications of the Department of General Linguistics, University of Helsinki, # 11. Helsinki.

Kujamäki, Pekka 2001: Finnish comet in German skies: Translation, retranslation, and norms. *Target* 13:1, pp. 45–70

Kujamäki, Pekka & Jääskeläinen, Riitta 2001: Korpukset kääntäjän apuna. Oittinen, Riitta & Mäkinen, Pirjo (toim.) *Alussa oli käännös*. Tampereen Yliopistopaino Oy — Juvenes Print, Tampere. S. 239–253.

Kuusinen M. & Ollikainen V 1984: *Venäläis-suomalainen suursanakirja / Большой русско-финский словарь.* Porvoo–Helsniki–Juva: WSOY.

Lager, Torbjörn 1995: *A Logical Approach to Corpus Linguistics.* Gothenburg Monographs in Linguistics 14. Göteborg University.

Landers, Clifford E. 2001: *Literary Translation: A Practical Guide*. Multilingual Matters Ltd. Clevedon–Buffalo–Toronto–Sydney.

Leech, Jeoffrey and Fligelstone, Steven 2002: Computers and Corpus Analysis. In C.S. Butler (ed.) *Computers and Written Texts.* Cambridge, Massachusetts: Blackwell. Pp. 115–140.

Lehmuskallio A. and Podbereznyj V. and Tommola H. 1991: Towards a Finnish-Russian Dictionary of Finnish Culture-Bound Words. Tirkkonen-Condit S. (ed.) *Empirical Research in Translation and Intercultural Studies*. Tübingen: Narr, pp. 157–165.

Mauranen, Anna 2000: Strange things in translated language: a study of corpora. Maeve Olohan (ed.) *Intercultural faultlines. Research models in translation studies. Textual and cognitive aspects.* Manchester, Northhampton: St. Jerome Publishing.

McEnery, Tony and Wilson, Andrew 2001: *Corpus Linguistics: An Introduction.* 2nd edition. Edinbourgh: Edinbourgh University Press.

Merkel, Magnus & Andersson, Mikael & Ahrenberg, Lars 2002: The PLUG Link Annotator — iterative construction of data from parallel corpora. In: Borin, Lars (ed.). *Parallel Corpora, Parallel Worlds.* Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22–23 April, 1999. Amsterdam — New York, NY: Rodopi. Pp. 151–168.

Mihailov, Mihail 2001: Two Approaches to Automated Text Aligning of Parallel Fiction Texts. *Across Languages and Cultures* 2(1), pp. 87–96.

Mihailov, Mihail and Tommola, Hannu 2001: Compiling Parallel Text Corpora: Towards Automation of Routine Procedures. *International Journal of Corpus Linguistics.* Vol. 6 (Special Issue), pp. 67–77.

Mihailov, Mihail 2002: Rinnakaistekstikorpus kääntäjän apuvälineenä. *Kääntäjä* # 4. s. 12.

Nida E.A. & Taber C.R. 1974: *The Theory and Practice of Translation*. Leiden: E.J. Brill.

Newmark P. 1991: «The Virtues of interference and the vices of translationese». Newmark P. *About translation*. Clevedon–Philadelphia–Adelaide, pp. 78–87.

Oakes, Michael 1998: *Statistics for Corpus Linguistics*. Edinburgh University Press.

Paloposki, Outi 2001: Enriching translations, simplified language?: An alternative view-point to lexical simplification. *Target* 13:2, pp. 265–288.

Patten, Terry 1992: Computers and Natural Language Parsing. In C.S. Butler (ed.) *Computers and Written Texts.* Cambridge, Massachusetts: Blackwell. Pp. 29–52.

Peuranen, Erkki 1985: Suomennokset venäjän kirjallisuudesta. *Maailmankirjallisuuden ja sen klassikojen suomentamisesta*. Kääntäjäseminaari Jyväskylässä 5.–6.7.1985. Osa II. Jyväskylä. s. 28 – 44.

Pitkänen, Johanna 1999: *Venäläisen kaunokirjallisuuden suomennokset ja kirjallisuus-kritiikit 1990-luvulla.* Pro-gradu —tutkielma. Tampereen yliopisto.

Rundell, Michael 1996: The corpus of the future, and the future of the corpus. http://www.ruf.rice.edu/~barlow/para.html.

Saari, Elisa 1989: *Käännöskritiikin teoriaa ja käytäntöä — havaintoja venäläisen ja neuvostokirjallisuuden käännösarvosteluista suomalaisissa päivälehdissä*. Pro-gradu —tutkielma. Tampereen yliopisto.

Saari, Mirja 1997: Kieli ja suomalainen identiteetti. *Studia Finlandica kevät 1997. Itsenäinen Suomi 80 vuotta*. Ed. H. Karjalainen & H. Westermarck. Helsinki: University of Helsinki. S. 91– 104.

Salkie, Raphael 2002: How can linguists profit from parallel corpora? In: Borin, Lars (ed.). *Parallel Corpora, Parallel Worlds.* Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22–23 April, 1999. Amsterdam — New York, NY: Rodopi. Pp. 93–109.

Savo Corpus 2001: *The Savonlinna Corpus of Translated Finnish*. Savonlinna School of Translation Studies, University of Joensuu

Sinclair, John 1991: *Corpus, Concordance, Collocation.* London: Oxford University Press.

Sinclair, John 2001: Data-derived Multilingual Lexicons. *International Journal of Corpus Linguistics*. Vol. 6 (Special Issue). John Benjamins Publishing Company, Amsterdam / Philadalphia, pp. 79–94.

Stahl, Peter 2002: Building and processing a multilingual corpus of parallel texts. In: Borin, Lars (ed.). *Parallel Corpora, Parallel Worlds.* Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22–23 April, 1999. Amsterdam — New York, NY: Rodopi. Pp. 169–179.

Streiter, Oliver and Iomdin, Leonid and Sagalova, Irina 2001. Learning Lessons from Bilingual Corpora: Benefits for Machine Translation. *International Journal of Corpus Linguistics.* Amsterdam: John Benjamins. Vol. 5 (2), 199–230.

Stubbs, Michael 2001: Texts, Corpora, and Problems of Interpretation: A Response to Widdowson. *Applied Linguistics*. 22/2, 149–172.

Summers, Della 1993: Longman/Lancaster English Language Corpus — Criteria and Design. *International Journal of Lexicography*. 6(3), 181–208.

Svartvik, Jan 1992: Corpus Linguistics Comes of Age. In Jan Svartvik (ed.) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991.* Berlin – New York: Mouton de Gruyter, 7–17.

Tadić, Marko 2001: Procedures in Building the Croatian-English Parallel Corpus. *International Journal of Corpus Linguistics*. Vol. 6 (Special Issue). John Benjamins Publishing Company, Amsterdam / Philadalphia, pp. 107–124.

Teubert, Wolfgang 1996: Comparable or Parallel Corpora. *International Journal of Lexicography.* Oxford University Press. 9(3), 238–264.

Teubert, Wolfgang 2001: Corpus Linguistics and Lexicography. *International Journal of Corpus Linguistics*. Vol. 6 (Special Issue). John Benjamins Publishing Company, Amsterdam / Philadalphia, pp. 125–153.

Tiedemann, Jörg 1997: *Automatical Lexicon Extraction from Aligned Bilingual Corpora*. M.A. thesis. Department of Linguistics, University of Uppsala / Otto-von-Guericke Universität Magdeburg.

Tiedemann, Jörg 1998: Extraction of Translation Equivalents from Parallel Corpora. In *Proceedings of the 11th Nordic Conference on Computational Linguistics*, Center for Sprogteknologi, Copenhagen.

Tiedemann, Jörg 1999a: Automatic Construction of Weighted String Similarity Measures. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.* University of Maryland, MD, USA, 1999.

Tiedemann, Jörg 1999b: Word Alignment Step by Step. In *Proceedings of the 12th Nordic Conference on Computational Linguistics, 1999*, Technical University of Trondheim. Department of Linguistics.

Tiedemann, Jörg 2002: Uplug — a modular corpus tool for parallel corpora. In: Borin, Lars (ed.). *Parallel Corpora, Parallel Worlds.* Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22–23 April, 1999. Amsterdam — New York, NY: Rodopi. Pp. 181–197.

Tihmeneva, Tatjana 1985: Venäläisen kirjallisuuden teitä Suomeen. *Maailman-kirjallisuuden ja sen klassikojen suomentamisesta*. Kääntäjäseminaari Juväskylässä 5.–6.7.1985. Osa II. Jyväskylä. s. 28–44.

Trosterud, Trond 2002: Parallel corpora as tools for investigating and developing minority languages. In: Borin, Lars (ed.). *Parallel Corpora, Parallel Worlds.* Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22–23 April, 1999. Amsterdam — New York, NY: Rodopi. Pp. 111–122.

Váradi, Tamás and Kiss, Gábor 2001: Equivalence and Non-equivalence in Parallel Corpora. *International Journal of Corpus Linguistics*. Vol. 6 (Special Issue). John Benjamins Publishing Company, Amsterdam / Philadalphia, pp. 167–177.

White, Leila 1993: *Suomen kielioppia ulkomaalaisille*. Finn Lektura.

Widdowson, Henry G. 2000: On the Limitations of Linguistics Applied. *Applied linguistics* 21/1: 3–25.

Widdowson, Henry G. 2001: Scoring Points by Critical Analysis: A Reaction to Beaugrande. *Applied linguistics* 22/1: 266–272.

Wierzbicka, Anna 1997: Understanding Cultures Through Their Key Words. English, Russian, Polish, German, and Japanese. Oxford, OUP.

Wills, Wolfram 1982: *The Science of Translation. Problems and Methods*. Tübingen: Narr.

WordCruncher 1989: *WordCruncher. WC View Text Retrieval Software. WordCruncher. WC Index Text Retrieval Software*. Brigham Young University.

Zernik U. 1991: Train 1 vs train 2: tagging word sense in a corpus. Zernik U. (ed.) *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Алексеева И.С. 2001: *Профессиональный тренинг переводчика*. Санкт-Петербург: Издательство «Союз».

Аношкина Ж.Г. 1992: Лингвистический программно-источниковый пакет UNILEX+. Тексто-ориентированная компонента UNILEX-T. *Бюллетень Машинного фонда русского языка*. — Вып. 2. — С. 3–7. Москва.

Андрющенко В.М. 1989: *Концепция и архитектура машинного фонда русского языка*. Москва: Наука.

Баевский В.С. 1999: *История русской литературы XX века. Компендиум.* Москва: Языки русской культуры.

Баранов А.Н. 1998: Автоматизация лингвистических исследований: корпус текстов как лингвистическая проблема. *Русистика сегодня*. Москва. — №.1–2. — С.179–191.

Баранов А.Н. 2001: *Введение в прикладную лингвистику*. Москва: Эдиториал УРСС.

Баранов А.Н., Добровольский Д.О. 1998: Немецкая корпусная лингвистика. *Вестник МГУ. Сер. Иностранные языки*. №1.

Баранов А.Н., Михайлов М.Н., Сидоров Г.О. 1998: «Динамический корпус текстов» как новая технология прикладной лингвистики. *Труды международного семинара Диалог'98 по компьютерной лингвистике и ее приложениям.* Т. 2. Казань.

Бархударов Л.С. 1975: *Язык и перевод*. Москва: Международные отношения.

Белая Г.А. 1990: Белая Г.А. (ред.) *История советской литературы: Новый взгляд. По материалам всесоюзной научно-творческой конференции 11–12 мая 1989 г.* Москва: Наука.

Белошапкова и др. 1999: Белошапкова В.А. (ред.) *Современный русский язык*. Москва: Азбуковник.

БРФС 1963: Куусинен М., Олликайнен В. *Большой русско-финский словарь*. Porvoo–Helsinki–Juva: WSOY.

Вежбицкая А. 1997: *Язык. Культура. Познание*. Москва: Русские словари.

Вежбицкая А. 2001: *Понимание культур через посредство ключевых слов.* Москва: Языки славянской культуры.

Виноградова В.Б. и др. 2001: Виноградова В.Б., Кукушкина О.В., Поликарпов А.А., Савчук С.О. Компьютерный корпус текстов русских газет конца 20-го века: создание, категоризация, автоматизированный анализ языковых особенностей. *Русский язык: исторические судьбы и современность. Международный конгресс русистов исследователей. Москва, филологический факультет МГУ 13–16 марта 2001. Труды и материалы.* Москва: Издательство МГУ. — С. 398.

Зализняк А.А. 1980: *Грамматический словарь русского языка.* Москва: Русский язык.

Зализняк Анна А. 2000: Преодоление пространства в русской языковой картине мира: глагол *добираться*. *Логический анализ языка. Языки пространств.* Москва.

Зализняк Анна А., Шмелев А.Д. 1997: Время суток и виды деятельности. *Логический анализ языка. Язык и время.* Москва.

Исаев И.А. 1996: Опыт автоматизации лексикографических исследований. Система DiaLex. *Слово Достоевского*. — С. 254–266. Москва.

Караулов Ю.Н. 1986: Караулов Ю.Н. (ред.) *Машинный фонд русского языка: идеи и суждения*. Москва: Наука.

Комиссаров В.Н. 1992: ««Естественность» художественного перевода». *Литература и перевод: проблемы теории*. Москва: Прогресс, стр. 101–110.

Кронгауз М.А. 2001: *Семантика*. Москва: РГГУ.

Левонтина И.Б., Шмелев А.Д. 1999: На своих двоих: лексика пешего перемещения в русском языке. *Логический анализ языка. Языки динамического мира.* Дубна.

Левонтина И.Б., Шмелев А.Д. 2000: Родные просторы. *Логический анализ языка. Языки пространств.* Москва.

Леннгрен Л., Ферм Л. 1991: Уппсальский машинный фонд русского языка. *Труды машинного фонда русского языка*. Т. 1. Москва.

МАС 1984: *Толковый словарь русского языка: в 4-х тт*. АН СССР, Институт русского языка. Москва: Русский язык.

Мельчук, И.А. 1999: *Опыт теории лингвистических моделей "Смысл ⇔ текст": Семантика, синтаксис.* Москва: Языки русской культуры.

Михайлов М.Н. 1998: Компьютерное обеспечение корпуса текстов (взгляд пользователя). *Русистика сегодня*. № 1–2. 1998

Михайлов М.Н. 2002а: Черная кошка в темной комнате, или Можно ли автоматизировать поиск переводных эквивалентов в параллельном корпусе текстов? *Алфавит: филологический сборник*. Смоленск: СГПУ, 2002. — С. 181–188.

Михайлов М.Н. 2002б: Контекстно-свободная лемматизация как временное решение насущных проблем. *Алфавит: филологический сборник*. Смоленск: СГПУ, 2002. — С. 197–210.

Михайлов М.Н. 2002в: От двуязычного словаря к словарю переводному. *Язык. Человек. Культура. Материалы международной научно-практической конференции 22–24 октября 2002 г., Смоленск*. Смоленск: СГПУ.

Михайлов М.Н. 2003а: Чем длиннее, тем лучше? Как сравнить длины исходного текста и перевода? *Математическая морфология. Электронный математический и медико-биологический журнал.* — т. 5. — Выпуск 1. Смоленск. http://www.smolensk.ru/user/sgma/MMORPH/N-9-html/mihailov/mihailov.htm

Михайлов М.Н. 2003б: Что длиннее — оригинал или перевод? *Русская филология. Ученые записки Смоленского государственного педагогического университета.* — Т.6. / Сост. и ред. В.С.Баевский, М.Л.Рогацкина. Смоленск.

Михайлов М.Н. (в печати): Язык переводных текстов как особый вариант литературного языка.

Мошкович, Ж.Г. 1989: *Автоматизированная лексикографическая система УНИЛЕКС–2*. Москва.

Мошкович Ж.Г. 1990: Проблема лемматизации при автоматизированной подготовке словарей и текстов. *Материалы III Всесоюзн. конф. по созданию машинного фонда русского языка.* Москва.

Нумминен, Сейя 1997: *«Моя жизнь в переводе»: Эса Адриан — портрет финского переводчика русской литературы*. Дипломная работа. Институт переводчиков г. Савонлинна, Университет Йоэнсуу.

Рыков В.В. 1996: Корпусная лингвистика (научно-аналитический обзор). *РЖ: Социальные и гуманитарные науки: Зарубежная литература*. — Москва: ИНИОН. — №.4 — С.43–51.

Сидоров Г.О. 1995: *Разработка и реализация лингвистического обеспечения систем с морфологическим анализом/синтезом для русского языка.* Автореф. дисс. . . . канд. филол. наук. — Москва, МГУ.

Сидоров Г.О. 1996: Лемматизация в автоматизированной системе построения словарей языка писателей. *Слово Достоевского*. — С. 266–301. Москва.

ФРС 1977: Вахрос И., Щербаков А. *Финско-русский словарь*. Москва: Русский язык.

Фрэнсис У.Н. 1988: Проблемы формирования и машинного представления большого корпуса текстов. *Новое в зарубежной лингвистике*. Вып. XIII. — С. 334–353. Москва: Прогресс.

Шайкевич А.Я. 1995: Конкорданс к прозаическому тексту (К выходу в свет конкорданса к “Преступлению и наказанию”). *Русистика сегодня*. № 2. С. 5–31.

Шмелев А.Д. 2000: «Широта русской души». *Логический анализ языка. Языки пространств.* Москва.

Шмелев А.Д. 2002: *Русская языковая модель мира. Материалы к словарю.* Москва: Языки славянской культуры.

Чудакова М.О. 2001: *Литература советского прошлого*. Москва: Языки славянской культуры.

Приложения

Приложение 1. Произведения русской художественной литературы и их переводы на финский язык

1.1. Рубрикация русских художественных произведений и их переводов на финский язык

Код группы	Количество текстов	Текстовая масса	Жанр
Group01	27	3350	повесть
Group02	17	3250	роман
Group03	6	3000	роман
Group04	12	2850	роман
Group05	16	2000	повесть
Group06	11	2000	роман
Group07	5	1800	роман
Group08	22	1800	повесть
Group09	13	1650	повесть
Group10	4	1650	роман
Group11	3	1500	роман
Group12	4	1300	роман
Group13	4	1300	роман
Group14	6	1250	роман
Group15	16	1200	повесть
Group16	3	1150	роман
Group17	3	1150	исторический роман
Group18	5	1100	роман
Group19	10	1100	повесть
Group20	5	1100	исторический роман
Group21	5	1100	детектив
Group22	7	950	повесть
Group23	4	950	роман
Group24	4	950	роман
Group25	4	950	роман
Group26	10	900	повесть
Group27	109	893	рассказ
Group28	11	850	повесть
Group29	3	800	роман
Group30	3	800	роман
Group31	16	800	рассказ
Group32	3	800	роман
Group33	6	751	повесть
Group34	5	750	роман
Group35	5	750	повесть

Период (оригинал)	Период (перевод)	Рейтинг авторов	Рейтинг переводчиков
7	4	2	3
7	4	2	3
3	4	4	3
7	4	2	4
7	4	2	2
7	4	2	2
7	4	3	2
1	4	3	3
7	4	3	4
6	4	3	4
7	4	3	4
3	4	4	4
6	3	3	2
5	3	2	2
1	4	3	2
6	3	3	3
7	4	3	4
6	4	2	2
7	4	2	4
6	3	2	2
8	5	2	2
8	4	2	3
6	3	2	4
7	4	3	3
6	4	2	4
7	4	3	3
4	4	3	4
1	4	3	4
6	4	3	3
2	4	3	3
4	3	3	2
2	4	3	4
2	4	4	3
6	3	2	2
6	3	2	2

Код группы	Количество текстов	Текстовая масса	Жанр
Group36	6	700	повесть
Group37	2	650	исторический роман
Group38	5	650	повесть
Group39	4	600	роман
Group40	27	566	рассказ
Group41	5	550	повесть
Group42	11	550	рассказ
Group43	7	550	повесть
Group44	5	550	фантастическая повесть
Group45	7	501	повесть для детей
Group46	1	500	роман
Group47	1	500	исторический роман
Group48	1	500	исторический роман
Group49	1	500	роман
Group50	1	500	роман
Group51	4	500	повесть
Group52	1	500	исторический роман
Group53	31	474	рассказ
Group54	10	451	рассказ
Group55	3	450	роман
Group56	5	450	повесть
Group57	3	450	роман
Group58	3	450	роман
Group59	3	450	повесть
Group60	35	427	рассказ
Group61	32	424	рассказ
Group62	17	411	рассказ
Group63	9	401	рассказ
Group64	4	400	повесть

Период (оригинал)	Период (перевод)	Рейтинг авторов	Рейтинг переводчиков
6	4	2	3
7	4	2	2
4	4	3	4
2	4	3	2
4	4	3	3
5	4	2	3
2	4	4	4
2	4	4	4
7	4	2	4
7	4	2	3
4	4	4	3
3	4	4	4
8	4	2	3
6	4	2	3
3	4	4	2
5	4	3	4
5	4	2	4
7	4	3	4
8	4	2	3
5	4	2	4
4	4	3	3
2	4	4	3
8	4	2	3
5	3	3	2
7	4	2	4
7	4	2	3
5	4	3	4
7	4	2	2
4	4	4	4

1.2. Список текстов группы Group01

Автор	Название	Год
АСТАФЬЕВ В.П.	Царь-рыба	1975
АСТАФЬЕВ В.П.	Пастух и пастушка	1974
АСТАФЬЕВ В.П.	Перевал	1959
ВАСИЛЬЕВ Б.Л.	А зори здесь тихие..	1969
ГРЕКОВА И.	Вдовий пароход	1981
ЕРОФЕЕВ В.В.	Василий Розанов глазами эксцентрика	1982
ЗАЛЫГИН С.П.	Оска — смешной мальчик	1980
ИСКАНДЕР Ф.А.	Защита Чика	1983
КАРЕЛИН Л.В.	Змеелов	1975
КРУПИН В.Н.	Живая вода	1980
МАКАРОВА Е.	На сохранении	1980
ПУЛАТОВ Т.И.	Завсегдатай	1980
ПУЛАТОВ Т.И.	Браслет	1980
ПУЛАТОВ Т.И.	Строжевые башни	1980
СОРОКИН В.Г.	Очередь	1985
СОРОКИН В.Г.	Тридцатая любовь Марины	1984
СОРОКИН В.Г.	Проездом	1984
СОРОКИН В.Г.	Любовь	1984
ТРИФОНОВ Ю.В.	Время и место	1980
ТРИФОНОВ Ю.В.	Предварительные итоги	1980
ТРИФОНОВ Ю.В.	Долгое прощание	1971
ТРИФОНОВ Ю.В.	Опрокинутый дом	1980
ТРИФОНОВ Ю.В.	Дом на набережной	1976
ТРИФОНОВ Ю.В.	Другая жизнь	1975
ТРИФОНОВ Ю.В.	Обмен	1969
ТРИФОНОВ Ю.В.	Старик	1978
ШУКШИН В.М.	Калина красная	1974

Переводчик	Название перевода	Год перевода
Heino Ulla-Liisa	Kuningaskala	1980
Heino Ulla-Liisa	Paimenlaulu	1979
Heino Ulla-Liisa	Selville vesille	1982
Heino Ulla-Liisa	Ja ilta oli rauhaisa	1974
Jaakkola Marja-Leena	Leskien laiva	1982
Mallinen Jukka	Vasili Rozanov eksentrikon silmin	1991
Heino Ulla-Liisa	Oska - yhden miehen maa	1989
Kalliomäki Tuomo-Pekka	Tsikin puolustus	1989
Heino Ulla-Liisa	Käärmeenpyytäjä	1984
Koskinen Marja	Elävää vettä	1982
Jaakkola Marja-Leena	Ehkäisevää hoitoa	1989
Salminen Eila	Kanta-asiakas	1985
Salminen Eila	Rannerengas	1985
Salminen Eila	Vartiotornit	1985
Mallinen Jukka	Jono	1990
Mallinen Jukka	Marinan kolmaskymmenes rakkaus	1992
Mallinen Jukka	Ohimennen	1991
Mallinen Jukka	Rakkaus	1991
Koskinen Marja	Aika ja paikka	1983
Anhava Martti	Alustava tilinpäätös	1989
Koskinen Marja	Pitkät jäähyväiset	1981
Anhava Martti	Talo kumollaan	1989
Koskinen Marja	Talo rantakadulla	1977
Anhava Martti	Toinen elämä	1986
Koskinen Marja	Vaihto	1981
Koskinen Marja	Vanhus	1979
Heino Ulla-Liisa	Punainen heisipuu	1979

1.3. Список текстов группы Group02

Автор	Название	Год
АБРАМОВ Ф.А.	Две зимы и три лета	1968
АБРАМОВ Ф.А.	Дом	1978
АБРАМОВ Ф.А.	Пути-перепутья	1973
БАРУЗДИН С.А.	Повторение пройденного	1964
БОГОМОЛОВ В.О.	В августе сорок четвертого	1973
ГУСАРОВ Д.Я.	За чертой милосердия	1970
ГУСАРОВ Д.Я.	(Название оригинала неизвестно)	1970
ДУДИНЦЕВ В.Д.	Белые одежды	1986
ЕВТУШЕНКО Е.А.	Ягодные места	1983
ЗАЛЫГИН С.П.	Южноамериканский вариант	1973
НИКУЛИН Л.В.	Мертвая зыбь	1960
ОКУДЖАВА Б.Ш.	Свидание с Бонапартом	1983
РУДНЕВ О.	Долгая дорога в дюнах	1980
РЫБАКОВ А.Н.	Тяжелый песок	1978
СВИРИДОВ Г.И.	Стоять до последнего	1970
ТЕНДРЯКОВ В.Ф.	Свидание с Нефертити	1964
ЧАКОВСКИЙ А.Б.	Блокада	1970

1.4. Список текстов группы Group03

Автор	Название	Год
ДОСТОЕВСКИЙ Ф.М.	Идиот	1869
ДОСТОЕВСКИЙ Ф.М.	Братья Карамазовы	1880
ДОСТОЕВСКИЙ Ф.М.	Подросток	1875
ДОСТОЕВСКИЙ Ф.М.	Бесы	1872
ТОЛСТОЙ Л.Н.	Анна Каренина	1877
ТОЛСТОЙ Л.Н.	Анна Каренина	1877

Переводчик	Название перевода	Год перевода
Heino Ulla-Liisa	Kaksi talvea ja kolme kesää	1976
Heino Ulla-Liisa	Koti	1979
Heino Ulla-Liisa	Teitä ja risteyksiä	1978
Heino Ulla-Liisa	Sinun kanssasi, Nataša	1969
Koskinen Marja	Elokuu 1944	1978
Heino Ulla-Liisa	Korpi ei tunne armoa	1980
Heino Ulla-Liisa	Partisaanimusiikkia	1985
Heino Ulla-Liisa	Valkeat vaatteet	1987
Heino Ulla-Liisa	Marjamaat	1982
Heino Ulla-Liisa	Eteläamerikkalainen vaihtoehto	1981
Heino Ulla-Liisa	Operatio Trusti	1966
Heino Ulla-Liisa	Päivälliset Bonapartelle	1985
Jaakkola Marja-Leena	Pitkä matka dyyneille	1986
Jaakkola Marja-Leena	Raskasta hiekkaa	1981
Pyykkö Lea	Viimeiseen mieheen	1977
Heino Ulla-Liisa	Nefertitin hymy	1965
Heino Ulla-Liisa	Piiritys	1972

Переводчик	Название перевода	Год перевода
Pyykkö Lea	Idiootti	1979
Pyykkö Lea	Karamazovin veljekset	1976
Pekari Ida	Keskenkasvuinen	1964
Pyykkö Lea	Riivaajat	1982
Heino Ulla-Liisa	Anna Karenina	1961
Pyykkö Lea	Anna Karenina	1979

Приложение 2. Список текстов корпуса «ПарРус»

№	Автор	Название	Переводчик
1.	Аксенов В.	Звездный билет	Adrian E.
2.	Бабель И.	Иисусов грех	Adrian E.
3.	Бабель И.	Иисусов грех	Heino U.-L.
4.	Бабель И.	Король	Adrian E.
5.	Бабель И.	Король	Heino U.-L.
6.	Бакланов Г.	Навеки девятнадцатилетние	Orlov V.
7.	Белов В.	Привычное дело	Laaksonen H.
8.	Булгаков М.А.	Мастер и Маргарита	Heino U.-L.
9.	Булгаков М.А.	Театральный роман	Adrian E.
10.	Гоголь Н.В.	Шинель	Adrian E.
11.	Гоголь Н.В.	Шинель	Jalkanen H.
12.	Гоголь Н.В.	Шинель	Konkka J.
13.	Горький М.	Макар Чудра	Pienimäki N.
14.	Горький М.	Старуха Изергиль	Mitrošin A.
15.	Горький М.	Челкаш	Mitrošin A.
16.	Гроссман В.	Все течет	Adrian E.
17.	Достоевский Ф.М.	Записки из подполья	Adrian E.
18.	Достоевский Ф.М.	Записки из подполья	Kallama V.
19.	Достоевский Ф.М.	Преступление и наказание	Konkka J.
20.	Дудинцев В.	Белые одежды	Heino U.-L.
21.	Ерофеев В.	Москва - Петушки	Adrian E.
22.	Зощенко М.	Аристократка	Losowitch K.
23.	Зощенко М.	Нервные люди	Losowitch K.
24.	Зощенко М.	Ночное происшествие	Losowitch K.
25.	Ильф И., Петров Е.	Двенадцать стульев	Silvanto R., Konkka J.
26.	Ильф И., Петров Е.	Золотой теленок	Aarto A.
27.	Лермонтов М.Ю.	Герой нашего времени	Heino U.-L.
28.	Лесков Н.	Очарованный странник	Pyykkö L.
29.	Олеша Ю.	Зависть	Adrian E.
30.	Пастернак Б.Л.	Доктор Живаго	Konkka J.
31.	Приставкин А.	Ночевала тучка золотая	Adrian E.
32.	Пушкин А.С.	Барышня-крестьянка	Ahava J., Hämeen-Anttila V.
33.	Пушкин А.С.	Барышня-крестьянка	Hollo J.A.
34.	Пушкин А.С.	Выстрел	Ahava J., Hämeen-Anttila V.
35.	Пушкин А.С.	Выстрел	Hollo J.A.
36.	Пушкин А.С.	Гробовщик	Hollo J.A.
37.	Пушкин А.С.	Капитанская дочка	Hollo J.A.
38.	Пушкин А.С.	Метель	Hollo J.A.
39.	Пушкин А.С.	Пиковая дама	Hollo J.A.
40.	Пушкин А.С.	Пиковая дама	Pesonen P. Alarik
41.	Пушкин А.С.	Станционный смотритель	Hollo J.A.
42.	Распутин В.	Живи и помни	Adrian E.
43.	Семенов Ю.	Семнадцать мгновений весны	Pienimäki N.
44.	Солженицын А.И.	Один день Ивана Денисовича	Adrian E.
45.	Солженицын А.И.	Один день Ивана Денисовича	Lahtela M.

Название перевода	Жанр	Период (ориг.)	Период (пер.)	Рейтинг автора	Рейтинг перев.	Группа
Matkalippu tähtiin	повесть	7	4	2	4	Group19
Jeesuksen synti	рассказ	5	4	2	4	X37
Jeesuksen synti	рассказ	5	4	2	3	X39
Kuningas	рассказ	5	4	2	4	X37
Kuningas	рассказ	5	4	2	3	X39
Synnyinmaan puolesta	повесть	7	4	2	2	Group05
Tuttu tarina	повесть	7	4	2	2	Group05
Saatana saapuu Moskovaan	роман	6	4	3	3	Group29
Teatteriromaani	роман	6	4	3	4	Group10
Päällystakki	повесть	1	4	3	4	Group28
Päällystakki	повесть	1	1	3	1	X120
Päällysviitta	повесть	1	4	3	4	Group28
Makar Tšudra	рассказ	4	4	3	2	X128
Isergil-muori	рассказ	4	4	3	2	X128
Tšelkaš	рассказ	4	4	3	2	X128
Kaikki virtaa	повесть	7	4	2	4	Group19
Kirjoituksia kellarista	повесть	3	4	4	4	X170
Kellariloukko	повесть	3	4	4	2	X169
Rikos ja rangaistus	роман	3	4	4	4	Group12
Valkeat vaatteet	роман	7	4	2	3	Group02
Moskova-Petuški	повесть	7	4	2	4	Group19
Aristokraatti	рассказ	5	4	2	3	X39
Kireähermoista väkeä	рассказ	5	4	2	3	X39
Tapaus yöllä	рассказ	6	4	2	3	X40
Kaksitoista tuolia	роман	5	3	2	2	Group14
Kultainen vasikka	роман	6	4	2	2	Group18
Aikamme sankari	роман	1	4	2	3	X236
Lumottu vaeltaja	повесть	3	4	2	2	X243
Kateus	повесть	5	4	1	4	X265
Tohtori Živago	роман	7	4	2	4	Group04
Yöpyi pilvi kultainen	повесть	7	4	2	4	Group19
Aatelisneiti talonpoikaistyttönä	повесть	1	4	3	2	Group15
Herrasneitti-talonpoikalaistyttö	повесть	1	4	3	3	Group08
Laukaus	повесть	1	4	3	2	Group15
Laukaus	повесть	1	4	3	3	Group08
Ruumisarkuntekijä	повесть	1	4	3	3	Group08
Kapteenintytär	повесть	1	4	3	3	Group08
Lumimyrsky	повесть	1	4	3	3	Group08
Patarouva	повесть	1	4	3	3	Group08
Patarouva	повесть	1	4	3	2	Group15
Asemanhoitaja	повесть	1	4	3	3	Group08
Elä ja muista	повесть	7	4	3	4	Group09
Kevään seitsemäntoista hetkeä	роман	7	4	2	2	Group06
Päivä Stalinin keskitysleirissä	повесть	7	4	3	4	Group09
Ivan Denisovitsin päivä	повесть	7	4	3	2	X12

№	Автор	Название	Переводчик
46.	Стругацкие А. и Б.	Парень из преисподней	Adrian E.
47.	Стругацкие А. и Б.	Попытка к бегству	Adrian E.
48.	Толстая Т.	Любишь — не любишь	Koskinen M.
49.	Толстая Т.	Милая Шура	Koskinen M.
50.	Толстая Т.	На золотом крыльце сидели…	Koskinen M.
51.	Толстая Т.	Река Оккервиль	Koskinen M.
52.	Толстая Т.	Соня	Koskinen M.
53.	Толстой Л.Н.	Анна Каренина	Pyykkö L.
54.	Толстой Л.Н.	Два гусара	Konkka J.
55.	Толстой Л.Н.	Метель	Konkka J.
56.	Трифонов Ю.	Дом на набережной	Koskinen M.
57.	Трифонов. Ю.	Предварительные итоги	Anhava M.
58.	Троепольский Г.	Белый Бим черное ухо	Iranto L.
59.	Тургенев И.С.	Дворянское гнездо	Heino U.-L.
60.	Фадеев А.	Разгром	Heino U.-L.
61.	Чехов А.П.	Агафья	Konkka J.
62.	Чехов А.П.	Актерская гибель	Konkka J.
63.	Чехов А.П.	Беглец	Konkka J.
64.	Чехов А.П.	В овраге	Heino U.-L.
65.	Чехов А.П.	В потемках	Konkka J.
66.	Чехов А.П.	Ванька	Konkka J.
67.	Чехов А.П.	Ведьма	Konkka J.
68.	Чехов А.П.	Гриша	Konkka J.
69.	Чехов А.П.	Дама с собачкой	Heino U.-L.
70.	Чехов А.П.	Дачники	Konkka J.
71.	Чехов А.П.	Детвора	Konkka J.
72.	Чехов А.П.	Добрый немец	Konkka J.
73.	Чехов А.П.	Дом с мезонином	Heino U.-L.
74.	Чехов А.П.	Дорогие уроки	Konkka J.
75.	Чехов А.П.	Дочь Альбиона	Konkka J.
76.	Чехов А.П.	Житейская мелочь	Konkka J.
77.	Чехов А.П.	Злой мальчик	Konkka J.
78.	Чехов А.П.	Злоумышленник	Konkka J.
79.	Чехов А.П.	Знакомый мужчина	Konkka J.
80.	Чехов А.П.	Крыжовник	Heino U.-L.
81.	Чехов А.П.	Кухарка женится	Konkka J.
82.	Чехов А.П.	Лишние люди	Konkka J.
83.	Чехов А.П.	Лошадиная фамилия	Konkka J.
84.	Чехов А.П.	Любовь	Konkka J.
85.	Чехов А.П.	Мальчики	Konkka J.
86.	Чехов А.П.	Мечты	Konkka J.
87.	Чехов А.П.	Мститель	Konkka J.
88.	Чехов А.П.	Мужики	Heino U.-L.
89.	Чехов А.П.	Налим	Konkka J.
90.	Чехов А.П.	Несчастье	Konkka J.
91.	Чехов А.П.	Неудача	Konkka J.
92.	Чехов А.П.	Нищий	Konkka J.
93.	Чехов А.П.	Ночь на кладбище	Konkka J.
94.	Чехов А.П.	Ночь перед судом	Konkka J.
95.	Чехов А.П.	Отец семейства	Konkka J.

Название перевода	Жанр	Период (ориг.)	Период (пер.)	Рейтинг автора	Рейтинг перев.	Группа
Poika helvetistä	фант. повесть	7	4	2	4	Group44
Pakoyritys	фант. повесть	7	4	2	4	Group44
Rakastaa — ei rakasta	рассказ	7	4	1	3	X15
Shura-kulta	рассказ	7	4	1	3	X15
Kultaportailla istuivat	рассказ	7	4	1	3	X15
Joki nimeltä Ockerville	рассказ	7	4	1	3	X15
Sonja	рассказ	7	4	1	3	X15
Anna Karenina	роман	3	4	4	3	Group03
Kaksi husaaria	повесть	2	4	4	4	Group43
Pyry	повесть	2	4	4	4	Group43
Talo rantakadulla	повесть	7	4	2	3	Group01
Alustava tilinpäätös	повесть	7	4	2	3	Group01
Bim mustakorva	повесть	7	4	2	2	Group05
Aateliskoti	роман	2	4	3	3	Group30
Tuho	роман	5	4	2	3	X292
Agafja	рассказ	4	4	3	4	Group27
Näyttelijän lähtö	рассказ	4	4	3	4	Group27
Karkuri	рассказ	4	4	3	4	Group27
Rotkossa	повесть	4	4	3	3	Group56
Pimeässä	рассказ	4	4	3	4	Group27
Kirje isoisälle	рассказ	4	4	3	4	Group27
Noita	рассказ	4	4	3	4	Group27
Grisha	рассказ	4	4	3	4	Group27
Nainen ja sylikoira	рассказ	4	4	3	3	Group40
Huvila-asukkaita	рассказ	4	4	3	4	Group27
Lapsia	рассказ	4	4	3	4	Group27
Kunnon saksalainen	рассказ	4	4	3	4	Group27
Taiteilijan tarina	рассказ	4	4	3	3	Group40
Kalliita kielitunteja	рассказ	4	4	3	4	Group27
Albionin tytär	рассказ	4	4	3	4	Group27
Elämän pikkuseikka	рассказ	4	4	3	4	Group27
Ilkeä poika	рассказ	4	4	3	4	Group27
Pahantekijä	рассказ	4	4	3	4	Group27
Tuttu mies	рассказ	4	4	3	4	Group27
Karviaismarjoja	рассказ	4	4	3	3	Group40
Keittäjätär menee naimisiin	рассказ	4	4	3	4	Group27
Tarpeettomia ihmisiä	рассказ	4	4	3	4	Group27
Hevosenkaltainen sukunimi	рассказ	4	4	3	4	Group27
Rakkaus	рассказ	4	4	3	4	Group27
Poikia	рассказ	4	4	3	4	Group27
Haaveita	рассказ	4	4	3	4	Group27
Kostaja	рассказ	4	4	3	4	Group27
Talonpoikia	повесть	4	4	3	3	Group56
Made	рассказ	4	4	3	4	Group27
Onnettomuus	рассказ	4	4	3	4	Group27
Ei onnistunut!	рассказ	4	4	3	4	Group27
Kerjäläinen	рассказ	4	4	3	4	Group27
Yö hautausmaalla	рассказ	4	4	3	4	Group27
Yö ennen oikeudenkäyntiä	рассказ	4	4	3	4	Group27
Perheen isä	рассказ	4	4	3	4	Group27

№	Автор	Название	Переводчик
96.	Чехов А.П.	Папаша	Konkka J.
97.	Чехов А.П.	Пересолил	Konkka J.
98.	Чехов А.П.	Произведение искусства	Konkka J.
99.	Чехов А.П.	Роман с контрабасом	Konkka J.
100	Чехов А.П.	Свадьба с генералом	Konkka J.
101	Чехов А.П.	Сирена	Konkka J.
102	Чехов А.П.	Смерть чиновника	Konkka J.
103	Чехов А.П.	Событие	Konkka J.
104	Чехов А.П.	Старость	Konkka J.
105	Чехов А.П.	Страшная ночь	Konkka J.
106	Чехов А.П.	Счастливчик	Konkka J.
107	Чехов А.П.	Счастье	Heino U.-L.
108	Чехов А.П.	Тайный советник	Konkka J.
109	Чехов А.П.	Толстый и тонкий	Konkka J.
110	Чехов А.П.	Тоска	Konkka J.
111	Чехов А.П.	Устрицы	Konkka J.
112	Чехов А.П.	Хамелеон	Konkka J.
113	Чехов А.П.	Хирургия	Konkka J.
114	Чехов А.П.	Хористка	Konkka J.
115	Шукшин В.М.	Алеша Бесконвойный	Adrian E.
116	Шукшин В.М.	Беспалый	Adrian E.
117	Шукшин В.М.	Версия	Adrian E.
118	Шукшин В.М.	Верую!	Adrian E.
119	Шукшин В.М.	Выбираю деревню на жительство	Adrian E.
120	Шукшин В.М.	Змеиный яд	Adrian E.
121	Шукшин В.М.	Как зайка летал на воздушных шариках	Adrian E.
122	Шукшин В.М.	Капроновая елочка	Adrian E.
123	Шукшин В.М.	Крепкий мужик	Adrian E.
124	Шукшин В.М.	Мастер	Adrian E.
125	Шукшин В.М.	Материнское сердце	Adrian E.
126	Шукшин В.М.	Миль пардон, мадам!	Adrian E.
127	Шукшин В.М.	Ноль-ноль целых	Adrian E.
128	Шукшин В.М.	Обида	Adrian E.
129	Шукшин В.М.	Ораторский прием	Adrian E.
130	Шукшин В.М.	Осенью	Adrian E.
131	Шукшин В.М.	Охота жить	Rymin R., Parkkinen P.
132	Шукшин В.М.	Свояк Сергей Сергеевич	Adrian E.
133	Шукшин В.М.	Случай в ресторане	Adrian E.
134	Шукшин В.М.	Срезал	Adrian E.
135	Шукшин В.М.	Страдания молодого Ваганова	Adrian E.
136	Шукшин В.М.	Танцующий Шива	Adrian E.
137	Шукшин В.М.	Хахаль	Adrian E.

Название перевода	Жанр	Период (ориг.)	Период (пер.)	Рейтинг автора	Рейтинг перев.	Группа
Isä-kulta	рассказ	3	4	3	4	X380
Liikaa suolaa	рассказ	4	4	3	4	Group27
Taiteen tuote	рассказ	4	4	3	4	Group27
Romaani bassoviulusta	рассказ	4	4	3	4	Group27
Häät kenraalin kera	рассказ	4	4	3	4	Group27
Seireeni	рассказ	4	4	3	4	Group27
Virkamiehen kuolema	рассказ	4	4	3	4	Group27
Surkea tapaus	рассказ	4	4	3	4	Group27
Vanhuus	рассказ	4	4	3	4	Group27
Kauhunyö	рассказ	4	4	3	4	Group27
Onnenpoika	рассказ	4	4	3	4	Group27
Onni	рассказ	4	4	3	3	Group40
Herra salaneuvos	рассказ	4	4	3	4	Group27
Paksukainen ja ohukainen	рассказ	4	4	3	4	Group27
Suru	рассказ	4	4	3	4	Group27
Osterit	рассказ	4	4	3	4	Group27
Kameleontti	рассказ	4	4	3	4	Group27
Hammaskirurgi	рассказ	4	4	3	4	Group27
Kuorotyttö	рассказ	4	4	3	4	Group27
Vartijaton Aljoša	рассказ	7	4	2	4	Group60
Sormeton	рассказ	7	4	2	4	Group60
Versio	рассказ	7	4	2	4	Group60
Minä uskon!	рассказ	7	4	2	4	Group60
Valitsen asuinkylää	рассказ	7	4	2	4	Group60
Käärmeenmyrkky	рассказ	7	4	2	4	Group60
Kun pupujussi lensi ilmapalloilla	рассказ	7	4	2	4	Group60
Kapronkuusi	рассказ	7	4	2	4	Group60
Kova äijä	рассказ	7	4	2	4	Group60
Mestari	рассказ	7	4	2	4	Group60
Äidin sydän	рассказ	7	4	2	4	Group60
Mille pardons, madame!	рассказ	7	4	2	4	Group60
Nolla-nolla kokonaista	рассказ	7	4	2	4	Group60
Mielipaha	рассказ	7	4	2	4	Group60
Puhujan tehokeino	рассказ	7	4	2	4	Group60
Syksyllä	рассказ	7	4	2	4	Group60
Halu elää	рассказ	7	4	2	0	X54
Lankomies Sergei Sergejevitch	рассказ	7	4	2	4	Group60
Tapahtui ravintolassa	рассказ	7	4	2	4	Group60
Teurastus	рассказ	7	4	2	4	Group60
Nuoren Vaganovin kärsimykset	рассказ	7	4	2	4	Group60
Tanssiva Shiva	рассказ	7	4	2	4	Group60
Naistennaurattaja	рассказ	7	4	2	4	Group60

Приложение 3. Статистические данные по текстам корпуса «ПарРус»

3.1. Тексты «ПарРус»: количество слов

Оригинал	**Перевод**	**Количество слов**		
		Оригинал	**Перевод**	**Отношение**
Аксенов В., Звездный билет	Matkalippu tähtiin (Adrian E.)	45905	43923	1,05
Бакланов Г., Навеки девятнадцатилетние	Synnyinmaan puolesta (Orlov V.)	48298	49154	0,98
Белов В., Привычное дело	Tuttu tarina (Laaksonen H.)	44904	42106	1,07
Булгаков М.А., Мастер и Маргарита	Saatana saapuu Moskovaan (Heino U.-L.)	112619	105250	1,07
Булгаков М.А., Театральный роман	Teatteriromaani (Adrian E.)	41159	36865	1,12
Гоголь Н.В., Шинель	Päällystakki (Adrian E.)	10013	9143	1,10
	Päällystakki (Jalkanen H.)	10013	8890	1,13
	Päällysviitta (Konkka J.)	10013	9171	1,09
Горький М., Макар Чудра	Makar Tšudra (Pienimäki N.)	4084	3795	1,08
Горький М., Старуха Изергиль	Isergil-muori Mitrošin A.	6922	6421	1,08
Горький М., Челкаш	Tšelkaš (Mitrošin A.)	9226	8805	1,05
Гроссман В., Все течет	Kaikki virtaa (Adrian E.)	42607	39245	1,09
Достоевский Ф.М., Записки из подполья	Kirjoituksia kellarista (Adrian E.)	34939	31388	1,11
	Kellariloukko (Kallama V.)	34939	32675	1,07
Достоевский Ф.М., Преступление и наказание	Rikos ja rangaistus (Konkka J.)	169566	163430	1,04
Дудинцев В., Белые одежды	Valkeat vaatteet (Heino U.-L.)	190655	181174	1,05
Ерофеев В., Москва - Петушки	Moskova-Petuški (Adrian E.)	33562	29440	1,14
Ильф И., Петров Е., Двенадцать стульев	Kaksitoista tuolia (Silvanto R., Konkka J.)	78315	62322	1,26
Ильф И., Петров Е., Золотой теленок	Kultainen vasikka (Aarto A.)	87825	81767	1,07

Оригинал	Перевод	Количество слов		
		Оригинал	Перевод	Отношение
Лермонтов М.Ю., Герой нашего времени	Aikamme sankari (Heino U.-L.)	41499	40155	1,03
Лесков Н., Очарованный странник	Lumottu vaeltaja (Pyykkö L.)	41379	37356	1,11
Олеша Ю., Зависть	Kateus (Adrian E.)	29507	27476	1,07
Пастернак Б.Л., Доктор Живаго	Tohtori Živago (Konkka J.)	148351	138450	1,07
Приставкин А., Ночевала тучка золотая	Yöpyi pilvi kultainen (Adrian E.)	66351	58758	1,13
Пушкин А.С., Барышня-крестьянка	Aatelisneiti talonpoikaistyttönä (Ahava J., Hämeen-Anttila V.)	5466	4929	1,11
	Herrasneitti-talonpoikalaistyttö (Hollo J.A.)	5466	5203	1,05
Пушкин А.С., Капитанская дочка	Kapteenintytär (Hollo J.A.)	32547	29750	1,09
Пушкин А.С., Пиковая дама	Patarouva (Hollo J.A.)	7022	6793	1,03
	Patarouva (Pesonen P.)	7022	6271	1,12
Распутин В., Живи и помни	Elä ja muista (Adrian E.)	65132	64503	1,01
Семенов Ю., Семнадцать мгновений весны	Kevään seitsemäntoista hetkeä (Pienimäki N.)	77860	70224	1,11
Солженицын А.И., Один день Ивана Денисовича	Päivä Stalinin keskitysleirissä (Adrian E.)	32390	32476	1,00
	Ivan Denisovitšin päivä (Lahtela M.)	32390	32191	1,01
Стругацкие А. и Б., Парень из преисподней	Poika helvetistä (Adrian E.)	28789	26824	1,07
Стругацкие А. и Б., Попытка к бегству	Pakoyritys (Adrian E.)	28175	26358	1,07
Толстой Л.Н., Анна Каренина	Anna Karenina (Pyykkö L.)	269689	255801	1,05
Толстой Л.Н., Два гусара	Kaksi husaaria (Konkka J.)	17038	16112	1,06
Толстой Л.Н., Метель	Pyry (Konkka J.)	8463	7770	1,09
Трифонов Ю., Дом на набережной	Talo rantakadulla (Koskinen M.)	43619	42639	1,02
Трифонов Ю., Предварительные итоги	Alustava tilinpäätös (Anhava M.)	21759	21308	1,02

Оригинал	Перевод	Количество слов		
		Оригинал	Перевод	Отношение
Троепольский Г., Белый Бим черное ухо	Bim mustakorva (Iranto L.)	52040	50752	1,03
Тургенев И.С., Дворянское гнездо	Aateliskoti (Heino U.-L.)	47133	46601	1,01
Фадеев А., Разгром	Tuho (Heino U.-L.)	44406	44086	1,01
Чехов А.П., В овраге	Rotkossa (Heino U.-L.)	11527	11445	1,01
Чехов А.П., Дама с собачкой	Nainen ja sylikoira (Heino U.-L.)	5117	4973	1,03
Чехов А.П., Дом с мезонином	Taiteilijan tarina (Heino U.-L.)	5609	5227	1,07
Чехов А.П., Мужики	Talonpoikia (Heino U.-L.)	9213	9098	1,01
Чехов А.П., Тайный советник	Herra salaneuvos (Konkka J.)	5217	4819	1,08
Шукшин В.М., Алеша Бесконвойный	Vartijaton Aljoša (Adrian E.)	4190	4159	1,01
Шукшин В.М., Как зайка летал на воздушных шариках	Kun pupujussi lensi ilmapalloilla (Adrian E.)	4237	4336	0,98
Шукшин В.М., Охота жить	Halu elää (Rymin R., Parkkinen P.)	5503	6399	0,86
Шукшин В.М., Страдания молодого Ваганова	Nuoren Vaganovin kärsimykset (Adrian E.)	4433	4461	0,99
			Среднее значение:	**1,06**

3.2. Тексты «ПарРус»: количество предложений

Оригинал	Перевод	Количество предложений		
		Оригинал	Перевод	Отношение
Аксенов В. Звездный билет	Matkalippu tähtiin (Adrian E.)	6217	6726	0,92
Бакланов Г. Навеки девятнадцатилетние	Synnyinmaan puolesta (Orlov V.)	4742	5061	0,94
Белов В. Привычное дело	Tuttu tarina (Laaksonen H.)	4481	4687	0,96
Булгаков М.А. Мастер и Маргарита	Saatana saapuu Moskovaan (Heino U.-L.)	8527	9474	0,90
Булгаков М.А. Театральный роман	Teatteriromaani (Adrian E.)	3939	4188	0,94
Гоголь Н.В. Шинель	Päällystakki (Adrian E.)	373	436	0,86
Гоголь Н.В. Шинель	Päällystakki (Jalkanen H.)	373	515	0,72
Гоголь Н.В. Шинель	Päällysviitta (Konkka J.)	373	385	0,97

Оригинал	Перевод	Количество предложений		
		Оригинал	Перевод	Отношение
Горький М. Макар Чудра	Makar Tšudra (Pienimäki N.)	374	394	0,95
Горький М. Старуха Изергиль	Isergil-muori Mitrošin A.	645	683	0,94
Горький М. Челкаш	Tšelkaš (Mitrošin A.)	975	1050	0,93
Гроссман В. Все течет	Kaikki virtaa (Adrian E.)	2761	2968	0,93
Достоевский Ф.М. Записки из подполья	Kirjoituksia kellarista (Adrian E.)	2323	2702	0,86
Достоевский Ф.М. Записки из подполья	Kellariloukko (Kallama V.)	2323	2429	0,96
Достоевский Ф.М. Преступление и наказание	Rikos ja rangaistus (Konkka J.)	12959	13470	0,96
Дудинцев В. Белые одежды	Valkeat vaatteet (Heino U.-L.)	23436	23112	1,01
Ерофеев В. Москва - Петушки	Moskova-Petuški (Adrian E.)	3356	3440	0,98
Ильф И., Петров Е. Двенадцать стульев	Kaksitoista tuolia (Silvanto R., Konkka J.)	9280	7825	1,19
Ильф И., Петров Е. Золотой теленок	Kultainen vasikka (Aarto A.)	8549	8931	0,96
Лермонтов М.Ю. Герой нашего времени	Aikamme sankari (Heino U.-L.)	2755	4123	0,67
Лесков Н. Очарованный странник	Lumottu vaeltaja (Pyykkö L.)	2204	2939	0,75
Олеша Ю. Зависть	Kateus (Adrian E.)	3676	3753	0,98
Пастернак Б.Л. Доктор Живаго	Tohtori Živago (Konkka J.)	14428	14443	1,00
Приставкин А. Ночевала тучка золотая	Yöpyi pilvi kultainen (Adrian E.)	7920	7641	1,04
Пушкин А.С. Барышня-крестьянка	Aatelisneiti talonpoikaistyttönä (Ahava J., Hämeen-Anttila V.)	331	458	0,72
Пушкин А.С. Барышня-крестьянка	Herrasneitti-talonpoikalaistyttö (Hollo J.A.)	331	422	0,78
Пушкин А.С. Капитанская дочка	Kapteenintytär (Hollo J.A.)	2775	3082	0,90
Пушкин А.С. Пиковая дама	Patarouva (Hollo J.A.)	629	666	0,94

Оригинал	Перевод	Количество предложений		
		Оригинал	Перевод	Отношение
Пушкин А.С. Пиковая дама	Patarouva (Pesonen P.)	629	760	0,83
Распутин В. Живи и помни	Elä ja muista (Adrian E.)	4697	5054	0,93
Семенов Ю. Семнадцать мгновений весны	Kevään seitsemäntoista hetkeä (Pienimäki N.)	7274	7989	0,91
Солженицын А.И. Один день Ивана Денисовича	Päivä Stalinin keskitysleirissä (Adrian E.)	3121	3248	0,96
Солженицын А.И. Один день Ивана Денисовича	Ivan Denisovitšin päivä (Lahtela M.)	3121	3256	0,96
Стругацкие А. и Б. Парень из преисподней	Poika helvetistä (Adrian E.)	3225	3407	0,95
Стругацкие А. и Б. Попытка к бегству	Pakoyritys (Adrian E.)	3606	3784	0,95
Толстой Л.Н. Анна Каренина	Anna Karenina (Pyykkö L.)	18146	19856	0,91
Толстой Л.Н. Два гусара	Kaksi husaaria (Konkka J.)	1195	1237	0,97
Толстой Л.Н. Метель	Pyry (Konkka J.)	521	565	0,92
Трифонов Ю. Дом на набережной	Talo rantakadulla (Koskinen M.)	3816	3794	1,01
Трифонов Ю. Предварительные итоги	Alustava tilinpäätös (Anhava M.)	1700	1850	0,92
Троепольский Г. Белый Бим черное ухо	Bim mustakorva (Iranto L.)	4934	5732	0,86
Тургенев И.С. Дворянское гнездо	Aateliskoti (Heino U.-L.)	2917	3762	0,78
Фадеев А. Разгром	Tuho (Heino U.-L.)	3107	4017	0,77
Чехов А.П. В овраге	Rotkossa (Heino U.-L.)	898	893	1,01
Чехов А.П. Дама с собачкой	Nainen ja sylikoira (Heino U.-L.)	309	312	0,99
Чехов А.П. Дом с мезонином	Taiteilijan tarina (Heino U.-L.)	324	318	1,02
Чехов А.П. Мужики	Talonpoikia (Heino U.-L.)	616	622	0,99
Чехов А.П. Тайный советник	Herra salaneuvos (Konkka J.)	403	400	1,01

Оригинал	Перевод	Количество предложений		
		Оригинал	Перевод	Отношение
Шукшин В.М. Алеша Бесконвойный	Vartijaton Aljoša (Adrian E.)	444	459	0,97
Шукшин В.М. Как зайка летал на воздушных шариках	Kun pupujussi lensi ilmapalloilla (Adrian E.)	541	564	0,96
Шукшин В.М. Охота жить	Halu elää (Rymin R., Parkkinen P.)	868	928	0,94
Шукшин В.М. Страдания молодого Ваганова	Nuoren Vaganovin kärsimykset (Adrian E.)	511	534	0,96
		Среднее значение:		**0,93**

3.3. Тексты «ПарРус»: количество абзацев

Оригинал	Перевод	Количество абзацев		
		Оригинал	Перевод	Отношение
Аксенов В., Звездный билет	Matkalippu tähtiin (Adrian E.)	2320	2299	1,01
Бакланов Г., Навеки девятнадцатилетние	Synnyinmaan puolesta (Orlov V.)	1736	1796	0,97
Белов В., Привычное дело	Tuttu tarina (Laaksonen H.)	1614	1515	1,07
Булгаков М.А., Мастер и Маргарита	Saatana saapuu Moskovaan (Heino U.-L.)	3951	3921	1,01
Булгаков М.А., Театральный роман	Teatteriromaani (Adrian E.)	1809	1811	1,00
Гоголь Н.В., Шинель	Päällystakki (Adrian E.)	46	46	1,00
	Päällystakki (Jalkanen H.)	46	206	0,22
	Päällysviitta (Konkka J.)	46	100	0,46
Горький М., Макар Чудра	Makar Tšudra (Pienimäki N.)	105	109	0,96
Горький М., Старуха Изергиль	Isergil-muori Mitrošin A.	127	144	0,88
Горький М., Челкаш	Tšelkaš (Mitrošin A.)	348	359	0,97
Гроссман В., Все течет	Kaikki virtaa (Adrian E.)	1275	1299	0,98
Достоевский Ф.М., Записки из подполья	Kirjoituksia kellarista (Adrian E.)	511	524	0,98
	Kellariloukko (Kallama V.)	511	522	0,98

Оригинал	Перевод	Количество абзацев		
		Оригинал	Перевод	Отношение
Достоевский Ф.М., Преступление и наказание	Rikos ja rangaistus (Konkka J.)	3810	3913	0,97
Дудинцев В., Белые одежды	Valkeat vaatteet (Heino U.-L.)	6444	6315	1,02
Ерофеев В., Москва — Петушки	Moskova-Petuški (Adrian E.)	1151	1164	0,99
Ильф И., Петров Е., Двенадцать стульев	Kaksitoista tuolia (Silvanto R., Konkka J.)	3712	3488	1,06
Ильф И., Петров Е., Золотой теленок	Kultainen vasikka (Aarto A.)	3103	3289	0,94
Лермонтов М.Ю., Герой нашего времени	Aikamme sankari (Heino U.-L.)	1268	1443	0,88
Лесков Н., Очарованный странник	Lumottu vaeltaja (Pyykkö L.)	1409	1494	0,94
Олеша Ю., Зависть	Kateus (Adrian E.)	1038	1037	1,00
Пастернак Б.Л., Доктор Живаго	Tohtori Živago (Konkka J.)	4029	4036	1,00
Приставкин А., Ночевала тучка золотая	Yöpyi pilvi kultainen (Adrian E.)	3374	3477	0,97
Пушкин А.С., Барышня-крестьянка	Aatelisneiti talonpoikais-tyttönä (Ahava J., Hämeen-Anttila V.)	78	197	0,40
	Herrasneitti-talonpoikalais-tyttö (Hollo J.A.)	78	157	0,50
Пушкин А.С., Капитанская дочка	Kapteenintytär (Hollo J.A.)	791	1077	0,73
Пушкин А.С., Пиковая дама	Patarouva (Hollo J.A.)	240	242	0,99
	Patarouva (Pesonen P.)	240	240	1,00
Распутин В., Живи и помни	Elä ja muista (Adrian E.)	1314	1348	0,97
Семенов Ю., Семнадцать мгновений весны	Kevään seitsemäntoista hetkeä (Pienimäki N.)	3050	3013	1,01
Солженицын А.И., Один день Ивана Денисовича	Päivä Stalinin keskitysleirissä (Adrian E.)	1281	1304	0,98
	Ivan Denisovitšin päivä (Lahtela M.)	1281	1244	1,03
Стругацкие А. и Б., Парень из преисподней	Poika helvetistä (Adrian E.)	923	854	1,08

Оригинал	**Перевод**	**Количество абзацев**		
		Оригинал	**Перевод**	**Отношение**
Стругацкие А. и Б., Попытка к бегству	Pakoyritys (Adrian E.)	1536	1417	1,08
Толстой Л.Н., Анна Каренина	Anna Karenina (Pyykkö L.)	7779	7674	1,01
Толстой Л.Н., Два гусара	Kaksi husaaria (Konkka J.)	562	543	1,03
Толстой Л.Н., Метель	Pyry (Konkka J.)	238	237	1,00
Трифонов Ю., Дом на набережной	Talo rantakadulla (Koskinen M.)	764	736	1,04
Трифонов Ю., Предварительные итоги	Alustava tilinpäätös (Anhava M.)	207	256	0,81
Троепольский Г., Белый Бим, черное ухо	Bim mustakorva (Iranto L.)	1810	1908	0,95
Тургенев И.С., Дворянское гнездо	Aateliskoti (Heino U.-L.)	1125	1131	0,99
Фадеев А., Разгром	Tuho (Heino U.-L.)	1498	1538	0,97
Чехов А.П., В овраге	Rotkossa (Heino U.-L.)	327	327	1,00
Чехов А.П., Дама с собачкой	Nainen ja sylikoira (Heino U.-L.)	129	125	1,03
Чехов А.П., Дом с мезонином	Taiteilijan tarina (Heino U.-L.)	122	117	1,04
Чехов А.П., Мужики	Talonpoikia (Heino U.-L.)	258	263	0,98
Чехов А.П., Тайный советник	Herra salaneuvos (Konkka J.)	119	120	0,99
Шукшин В.М., Алеша Бесконвойный	Vartijaton Aljoša (Adrian E.)	112	123	0,91
Шукшин В.М., Как зайка летал на воздушных шариках	Kun pupujussi lensi ilmapalloilla (Adrian E.)	186	189	0,98
Шукшин В.М., Охота жить	Halu elää (Rymin R., Parkkinen P.)	341	316	1,08
Шукшин В.М., Страдания молодого Ваганова	Nuoren Vaganovin kärsimykset (Adrian E.)	170	170	1,00
			Среднее значение:	**0,94**

3.4. Тексты «ПарРус»: количество символов

Оригинал	Перевод	Количество символов		
		Оригинал	Перевод	Отношение
Аксенов В., Звездный билет	Matkalippu tähtiin (Adrian E.)	304841	340049	0,90
Бакланов Г., Навеки девятнадцатилетние	Synnyinmaan puolesta (Orlov V.)	321482	389462	0,83
Белов В., Привычное дело	Tuttu tarina (Laaksonen H.)	294627	324247	0,91
Булгаков М.А., Мастер и Маргарита	Saatana saapuu Moskovaan (Heino U.-L.)	772267	872025	0,89
Булгаков М.А., Театральный роман	Teatteriromaani (Adrian E.)	280633	309246	0,91
Гоголь Н.В., Шинель	Päällystakki (Adrian E.)	65423	71768	0,91
	Päällystakki (Jalkanen H.)	65423	73893	0,89
	Päällysviitta (Konkka J.)	65423	73219	0,89
Горький М., Макар Чудра	Makar Tšudra (Pienimäki N.)	23612	27213	0,87
Горький М., Старуха Изергиль	Isergil-muori Mitrošin A.	40795	49418	0,83
Горький М., Челкаш	Tšelkaš (Mitrošin A.)	61107	71755	0,85
Гроссман В., Все течет	Kaikki virtaa (Adrian E.)	296131	338559	0,87
Достоевский Ф.М., Записки из подполья	Kirjoituksia kellarista (Adrian E.)	217502	244628	0,89
	Kellariloukko (Kallama V.)	217502	254595	0,85
Достоевский Ф.М., Преступление и наказание	Rikos ja rangaistus (Konkka J.)	1093575	1265605	0,86
Дудинцев В., Белые одежды	Valkeat vaatteet (Heino U.-L.)	1295020	1457945	0,89
Ерофеев В., Москва - Петушки	Moskova-Petuški (Adrian E.)	211202	233823	0,90
Ильф И., Петров Е., Двенадцать стульев	Kaksitoista tuolia (Silvanto R., Konkka J.)	566953	542203	1,05
Ильф И., Петров Е., Золотой теленок	Kultainen vasikka (Aarto A.)	631645	712147	0,89
Лермонтов М.Ю., Герой нашего времени	Aikamme sankari (Heino U.-L.)	270087	322874	0,84
Лесков Н., Очарованный странник	Lumottu vaeltaja (Pyykkö L.)	249789	282031	0,89
Олеша Ю., Зависть	Kateus (Adrian E.)	201462	223818	0,90
Пастернак Б.Л., Доктор Живаго	Tohtori Živago (Konkka J.)	1009774	1166486	0,87

Оригинал	Перевод	Количество символов		
		Оригинал	Перевод	Отношение
Приставкин А., Ночевала тучка золотая	Yöpyi pilvi kultainen (Adrian E.)	440879	478994	0,92
Пушкин А.С., Барышня-крестьянка	Aatelisneiti talonpoikaistyttönä (Ahava J., Hämeen-Anttila V.)	36066	40044	0,90
	Herrasneitti-talonpoikalaistyttö (Hollo J.A.)	36066	42253	0,85
Пушкин А.С., Капитанская дочка	Kapteenintytär (Hollo J.A.)	215128	241702	0,89
Пушкин А.С., Пиковая дама	Patarouva (Hollo J.A.)	47830	55218	0,87
	Patarouva (Pesonen P.)	47830	53165	0,90
Распутин В., Живи и помни	Elä ja muista (Adrian E.)	417497	487444	0,86
Семенов Ю., Семнадцать мгновений весны	Kevään seitsemäntoista hetkeä (Pienimäki N.)	537489	601094	0,89
Солженицын А.И., Один день Ивана Денисовича	Päivä Stalinin keskitysleirissä (Adrian E.)	209944	250808	0,84
	Ivan Denisovitšin päivä (Lahtela M.)	209944	245859	0,85
Стругацкие А. и Б., Парень из преисподней	Poika helvetistä (Adrian E.)	184965	210821	0,88
Стругацкие А. и Б., Попытка к бегству	Pakoyritys (Adrian E.)	195628	215970	0,91
Толстой Л.Н., Анна Каренина	Anna Karenina (Pyykkö L.)	1737384	1991531	0,87
Толстой Л.Н., Два гусара	Kaksi husaaria (Konkka J.)	112672	131518	0,86
Толстой Л.Н., Метель	Pyry (Konkka J.)	55063	63075	0,87
Трифонов Ю., Дом на набережной	Talo rantakadulla (Koskinen M.)	288787	336783	0,86
Трифонов Ю., Предварительные итоги	Alustava tilinpäätös (Anhava M.)	141554	168671	0,84
Троепольский Г., Белый Бим черное ухо	Bim mustakorva (Iranto L.)	335525	392077	0,86
Тургенев И.С., Дворянское гнездо	Aateliskoti (Heino U.-L.)	311677	362074	0,86
Фадеев А., Разгром	Tuho (Heino U.-L.)	300094	353855	0,85
Чехов А.П., В овраге	Rotkossa (Heino U.-L.)	72609	87429	0,83
Чехов А.П., Дама с собачкой	Nainen ja sylikoira (Heino U.-L.)	31552	38277	0,82
Чехов А.П., Дом с мезонином	Taiteilijan tarina (Heino U.-L.)	35140	41194	0,85

Оригинал	**Перевод**	**Количество символов**		
		Оригинал	**Перевод**	**Отношение**
Чехов А.П., Мужики	Talonpoikia (Heino U.-L.)	58427	70437	0,83
Чехов А.П., Тайный советник	Herra salaneuvos (Konkka J.)	33524	39364	0,85
Шукшин В.М., Алеша Бесконвойный	Vartijaton Aljoša (Adrian E.)	26855	31493	0,85
Шукшин В.М., Как зайка летал на воздушных шариках	Kun pupujussi lensi ilmapalloilla (Adrian E.)	27880	33652	0,83
Шукшин В.М., Охота жить	Halu elää (Rymin R., Parkkinen P.)	37502	48258	0,78
Шукшин В.М., Страдания молодого Ваганова	Nuoren Vaganovin kärsimykset (Adrian E.)	28668	34510	0,83
			Среднее значение:	**0,87**

Приложение 4. Коды программ «КОКОС-П»

4.1. Стыковка параллельных текстов

4.1.1. Стартовый модуль стыковщика

```
Private Sub AlignerBut_Click()
Dim db As Database
Dim Lib As Recordset, FinLib As Recordset
’Каталоги русских и финских текстов
Dim Log As Recordset
’Таблица, в которой хранятся номера текстов для стыковки.
’Эти данные заносятся в таблицу автоматически после регистрации в каталоге
’новой пары текстов. Уничтожаются после
’того, как была завершена стыковка данной пары текстов
Dim Resp As Integer
’Команды пользователя
Dim ST As Boolean

Set db = CurrentDb
Set Lib = db.OpenRecordset("Library", dbOpenTable)
      Lib.Index = "PrimaryKey"
Set FinLib = db.OpenRecordset("LibraryFin", dbOpenTable)
     FinLib.Index = "LinkToOrig"

Set Log = db.OpenRecordset("LogAlign", dbOpenTable)

   If Log.RecordCount > 0 Then
’Если таблица LogAlign непустая, то в корпусе есть новые тексты для стыковки.
’В противном случае программа завершается
   Do Until Log.EOF
     Lib.Seek "=", Log!TextNo

     If Lib.NoMatch Then
       MsgBox "Wrong text number in the LogAlign. Exiting program":  Exit Sub
     End If

    FinLib.Index = "PrimaryKey"
     FinLib.Seek "=", Log!TrNo
     If FinLib.NoMatch Then
       MsgBox "Wrong translation number in the LogAlign. Exiting program" : Exit Sub
      Else
       TrNo = Log!TrNo : FinLib.Index = "LinkToOrig" : FinLib.MoveFirst
         Do Until FinLib!ID = TrNo : FinLib.MoveNext : Loop
      End If
    [TextNo] = Lib!ID : [TrNo] = FinLib!ID
’Вывод в поля формы программы-стыковщика номеров стыкуемых текстов
’Эти номера используются при возврате из интерактивного в автоматический
’режим работы
```

```
   TextTitle.Caption = Lib!Author & " " & Lib!Title
   TranslTitle.Caption = FinLib!Author & " " & FinLib!Title
   ’Вывод в заголовок формы названия русского текста и финского перевода
   TrNo = FinLib!ID

    TextTitle.Visible = True : TranslTitle.Visible = True
    Forms!alignment!AnalOver = False
    RusDone = False : FinDone = False
      ST = Aligner(TextNo, TrNo)
’Вызов функции Aligner, выполняющей стыковку текстов. Возвращает True, если стыковка
’завершена

    If ST Then
      GoSub ProcessNext
’Перейти к стыковке следующего текста
    Else
          Exit Sub
’Прекратить выполнение программы, перейти в интерактивный режим с сохранением
’информации о позициях в стыкуемых текстах
      End If   'St
    Log.MoveNext
 Loop
Else 'Log.RecordCount > 0
        'Log is empty.
     MsgBox "Nothing to work with. Exiting program." : Exit Sub
End If  'Log.RecordCount > 0
  MsgBox "All done"
Exit Sub

 '********
ProcessNext:
’Переход к обработке следующего текста
’Для этого нужно уничтожить данные из протокола обработки предыдущего текста
’и запись в таблице LogAlign
DoCmd.SetWarnings False
   DoCmd.RunSQL "Delete * from AlignerSave"
DoCmd.SetWarnings True
     Log.Delete
       Resp = MsgBox("Text successfully aligned. Process the next one?", vbYesNo)
         If Resp = vbNo Then Exit Sub
’Выход в интерактивный режим в случае нежелания пользователя обрабатывать
’следующую пару текстов
  Return
End Sub
```

4.1.2. Функция, выполняющая сравнение длин абзацев и их стыковку

```
Const Lower = 0.95  : Const Upper = 1.15
‘Глобальные константы с пороговыми значениями коэффициента ИЯ-ПЯ
‘(значения определены экспериментальным путем)

Public Function Aligner(RusTextNumber As Long, FinTextNumber As Long)
Dim db As Database, Sdb As Database
Dim RusLib As Recordset, FinLib As Recordset
Dim RussianText As Recordset, FinnishText As Recordset
```

```
Dim Save As Recordset, OutTb As Recordset
Dim RusPar As String
Dim RPN As Integer, FinPar As String, FPN As Integer
Dim RWordCount As Single, RSentenceCount As Single, FWordCount As Single
Dim Sent As String
Dim FSentenceCount As Single
Dim Count As Integer
Dim CompBase As Single, OldCmp As Single
Dim MoreRus As Boolean, MoreFin As Boolean
Dim PCount As Integer, TCount As Integer
Dim RusSave As Long, FinSave As Long, TrNo As Long
Dim ReturnValue As Variant

Set db = CurrentDb
Set RusLib = db.OpenRecordset("Library", dbOpenTable)
       RusLib.Index = "PrimaryKey"
Set FinLib = db.OpenRecordset("LibraryFin", dbOpenTable)
      FinLib.Index = "PrimaryKey"

Set Sdb = DBEngine.Workspaces(0).OpenDatabase(InitDb & DbName)
Set RussianText = Sdb.OpenRecordset("Texts", dbOpenTable)
     RussianText.Index = "Source"
Set FinnishText = Sdb.OpenRecordset("FinTexts", dbOpenTable)
     FinnishText.Index = "Source"
Set OutTb = db.OpenRecordset("AlignedText", dbOpenTable)
     OutTb.Index = "PrimaryKey"
       If OutTb.RecordCount > 0 Then
        OutTb.MoveLast : TrNo = OutTb!ID + 1
       Else
         TrNo = 1
       End If
Set Save = db.OpenRecordset("AlignerSave", dbOpenTable)
 If Save.RecordCount = 0 Then
  Save.AddNew : Save!Russian = 1 : Save!Finnish = 1 : RusSave = 1 : FinSave = 1
 Else
 Save.MoveLast : RusSave = Save!Russian : FinSave = Save!Finnish
 End If

RusPar = "" : FinPar = "" : RWordCount = 0 : RSentenceCount = 0 : FWordCount = 0
FSentenceCount = 0 : OldCmp = 0 : PCount = 0 : TCount = 1

'Поиск в русском и финском текстах записи, на которой была завершена
'работа при последнем переходе в интерактивный режим. Если процедура
'вызвана в первый раз, то выполняется поиск начала финского и русского текстов
RussianText.Seek "=", RusTextNumber, RusSave
 If RussianText.NoMatch Then
 MsgBox "Russian text. Nothing found. Skipping Text" : Aligner = True : Exit Function
 End If

FinnishText.Seek "=", FinTextNumber, FinSave
  If FinnishText.NoMatch Then
   MsgBox "Finnish text. Nothing found. Skipping Text" : Aligner = True : Exit Function
  End If

MoreRus = False : MoreFin = False

 Do While RussianText!Source = RusTextNumber
```

```
'Переменные MoreFin и MoreRus принимают значение True в тех случаях, когда
'значение коэффициента ИЯ-ПЯ выходит за пределы допустимого интервала
'и программа должна добавить абзац из исходного текста или из перевода
  If Not MoreFin Then GoSub CollectRus
  If Not MoreRus Then GoSub CollectFin
'Сравнение длин сравниваемых абзацев. Если значение коэффициента оказывается
'в пределах, заданных константами Upper и Lower, абзацы стыкуются
  CompBase = RWordCount / FWordCount
   If (CompBase >= Lower And CompBase <= Upper) Or _
          (RSentenceCount = FSentenceCount And CompBase * 2 > Lower _
          And CompBase / 2 < Upper) Then
     OutTb.AddNew
      OutTb!ID = TrNo : OutTb!RussianText = RusPar : OutTb!FinnishText = FinPar
      OutTb!Original = RusTextNumber : OutTb!Translation = FinTextNumber
     OutTb.Update
      If Not RussianText.EOF And Not FinnishText.EOF Then
        Save.AddNew
         Save!Russian = RussianText!Paragraph : Save!Finnish = FinnishText!Paragraph
        Save.Update
      Else
        Save.Edit
        Save!Russian = 0 : Forms!alignment!RusDone = True : Save!Finnish = 0
        Forms!alignment!FinDone = True
        Save.Update
      End If
      RusPar = "" : FinPar = "" : RWordCount = 0 : FWordCount = 0 : RSentenceCount = 0
      FSentenceCount = 0 : MoreRus = False : MoreFin = False
     TrNo = TrNo + 1
     Forms!alignment!Parallel = TrNo
  ElseIf (OldCmp > Upper And CompBase < Lower) Or _
          (OldCmp < Lower And CompBase > Upper) Then
'Переход в интерактивный режим
        Save.AddNew
        If Not RussianText.EOF Then
          Save!Russian = RussianText!Paragraph
        Else
          Save!Russian = 0 : Forms!alignment!RusDone = True
        End If

        If Not FinnishText.EOF Then
         Save!Finnish = FinnishText!Paragraph
        Else
         Save!Finnish = 0 : Forms!alignment!FinDone = True
        End If
        Save.Update
        Forms!alignment!Parallel = TrNo : Forms!alignment!RussianText = RusPar
        Forms!alignment!FinnishText = FinPar : Forms!alignment!AnalOver = False
     DoCmd.GoToControl "AddFin"
             Forms!alignment!AlignerBut.Visible = False

       Aligner = False
         OutTb.Close : db.Close
       Exit Function

  ElseIf CompBase < Lower Then
'Если значение коэффициента ИЯ-ПЯ ниже порогового значения, добавляется абзац
'из русского текста
```

```
    MoreRus = True : MoreFin = False
    RusPar = RusPar & Chr$(13) & Chr$(10)
'Добавить знак абзаца, чтобы новый абзац отделялся от старого
     If RussianText.EOF Then Exit Do

  ElseIf CompBase > Upper Then
'Если значение коэффициента ИЯ-ПЯ выше порогового значения, добавляется абзац
'из финского текста
    MoreFin = True : MoreRus = False
    FinPar = FinPar & Chr$(13) & Chr$(10)
     'If Not FinnishText.EOF Then FinnishText.MoveNext
     If FinnishText.EOF Then Exit Do
  End If

  OldCmp = CompBase
If RussianText.EOF Then Exit Do : If FinnishText.EOF Then Exit Do
Loop

Forms!alignment!RusDone = True : Forms!alignment!FinDone = True

'Цикл завершается, когда один из текстов закончился. Однако, вовсе не обязательно,
'что программа одновременно дойдет до конца обоих текстов. Проверить,
'не осталось ли что-то от второго текста.
If RusPar <> "" Or FinPar <> "" Then
       If RussianText.EOF And FinnishText.EOF Then
         Forms!alignment!AddRus.Visible = False
         Forms!alignment!AddFin.Visible = False
       Else
         If Save.RecordCount = 0 Then Save.AddNew Else Save.Edit
           If Not RussianText.EOF Then
            Save!Russian = RussianText!Paragraph
           Else
             Forms!alignment!AddRus.Visible = False
           End If

           If Not FinnishText.EOF Then
            Save!Finnish = FinnishText!Paragraph
           Else
             Forms!alignment!AddFin.Visible = False
           End If
          Save.Update
       End If
                Forms!alignment!Parallel = TrNo
                 Forms!alignment!RussianText = RusPar
                 Forms!alignment!FinnishText = FinPar
        Aligner = False
        OutTb.Close : db.Close
       Exit Function
End If
             StrMsg = " "
             ReturnValue = SysCmd(acSysCmdSetStatus, StrMsg)

  DoCmd.SetWarnings False
    DoCmd.RunSQL "Delete * from AlignerSave"
  DoCmd.SetWarnings True
  'Уничтожить протокол стыковки текущего текста после окончания стыковки
 Aligner = True
```

```
 Exit Function

 '*************************
 'Подпрограммы
 '*************************
CollectRus:
'Сборка русского абзаца
  RPN = RussianText!Paragraph
   Do While RussianText!Paragraph = RPN
      Sent = RussianText!Phrase
        RussianText.Edit : RussianText!Parallel = TrNo : RussianText.Update
         RWordCount = RWordCount + CountWords(Sent)
         RSentenceCount = RSentenceCount + 1
       RusPar = RusPar & Sent & " "
    RussianText.MoveNext
      If RussianText.EOF Then Exit Do
  Loop

  RusPar = Trim(RusPar)
  PCount = PCount + 1 'счетчик абзацев
  Count = Count + 1
 Return

 '***************
CollectFin:
'Сборка финского абзаца
 FPN = FinnishText!Paragraph
   Do While FinnishText!Paragraph = FPN
      Sent = FinnishText!Phrase
       FinnishText.Edit : FinnishText!Parallel = TrNo : FinnishText.Update
        FWordCount = FWordCount + CountWords(Sent)
        FSentenceCount = FSentenceCount + 1
       FinPar = FinPar & Sent & " "
    FinnishText.MoveNext
      If FinnishText.EOF Then Exit Do
  Loop
  FinPar = Trim(FinPar)
Return
End Function
```

4.2. Построение параллельного конкорданса

```
Public Function SearchWordPar(LookFor As Variant, SecondKey As Variant, _
          FormName As String)
Dim db As Database, dbSource As Database
Dim OrigLib As Recordset, TrLib As Recordset, Words As Recordset
Dim LemmaTb As Recordset, Addresses As Recordset, Concordance As Recordset
Dim Texts As Recordset, List As Recordset, ConcList As Recordset
Dim WildCard As String, FileName As String
Dim Letter As String * 1, LetterCode As Integer
Dim Word As String, WordNum As Long, WordAddress As Long
Dim StatusInt As Integer, StartingAddress As Long, FinalAddress As Long
Dim SearchResult As String, SearchPat As String, CountExamples As Long
Dim SMode As Byte, Flag As Boolean
Dim StepInt As Integer, StartPoint As Integer, PosWord As Integer, PosWord1 As Integer
```

```
Dim InBOF As Boolean
Dim TakeOrNot As Integer
Dim CopySecondKey as String
Dim SecKey As String
Dim PSlash As Integer
Dim RT As String, FT As String, ST As String, SNum(2) As Long
Dim RusT As String, FinT As String, ZZ As String
Dim e As Integer
Dim Success As Boolean, WCOK As Boolean
Dim Neighbour As Integer
Dim ReturnValue As Variant
Dim SC As Integer
Dim ParaNum As Integer, CurPar As Integer, SourceNum As Integer
Dim WholeWord As Boolean, WholeWordSec As Boolean
Dim WordsBkm As Variant, r As Boolean

TakeOrNot = Forms(FormName)![EveryX]
DataLoc = GetTxDir
Set dbSource = DBEngine.Workspaces(0).OpenDatabase(DataLoc & DbName, True, True)
Set db = CurrentDb
'Функция вызывается из двух разных форм, предназначенных для работы со стороны
'исходных текстов или переводов. Режим работы определяется по имени формы
If FormName = "Search" Then
Set LemmaTb = db.OpenRecordset("GlobalLemm", dbOpenTable)
  LemmaTb.Index = "Word"
Set Concordance = db.OpenRecordset("ParallelConc", dbOpenTable)
Set ConcList = db.OpenRecordset("PhrNumTemp", dbOpenTable)
 ConcList.Index = "PhraseNo"
If (Forms(FormName)![SearchMode] <> "Any part of the word" _
    And Forms(FormName)![SearchMode] <> "End of the word") Then
 Set Words = db.OpenRecordset("GlobalWords", dbOpenTable)
 If Forms(FormName)![SearchMode] = "Lemma" Then
   Words.Index = "LinkToLemm"
 Else
   Words.Index = "Word"
 End If
Else
ReturnValue = SysCmd(acSysCmdSetStatus, "Filtering Global Word List. Please wait...")
If Forms(FormName)![SearchMode] = "End of the word" Then
 WildCard = "Select * from GlobalWords Where Word Like " & Chr$(34) _
           & "*" & LookFor & Chr$(34)
Else
 WildCard = "Select * from GlobalWords Where Word Like " _
           & Chr$(34) & "*" & LookFor & "*" & Chr$(34)
End If
 Set Words = db.OpenRecordset(WildCard)
 If Words.RecordCount > 0 Then
 Words.MoveLast : Words.MoveFirst
 Else
  Exit Function
 End If
ReturnValue = SysCmd(acSysCmdSetStatus, "Done")
End If
Set Addresses = dbSource.OpenRecordset("Addresses", dbOpenTable)
   Addresses.Index = "Word"
Else
 Set LemmaTb = db.OpenRecordset("GlobalLemFin", dbOpenTable)
```

```
   LemmaTb.Index = "Word"
 Set Concordance = db.OpenRecordset("ParallelConc", dbOpenTable)
 Set ConcList = db.OpenRecordset("PhrNumTemp", dbOpenTable)
  ConcList.Index = "PhraseNo"

If Forms(FormName)![SearchMode] <> "Any part of the word" _
    And Forms(FormName)![SearchMode] <> "End of the word") Then
Set Words = db.OpenRecordset("GlobalFin", dbOpenTable)
 If Forms(FormName)![SearchMode] = "Lemma" Then
   Words.Index = "LinkToLemm"
 Else
   Words.Index = "Word"
 End If
Else
ReturnValue = SysCmd(acSysCmdSetStatus, "Filtering Global Word List. Please wait...")
If Forms(FormName)![SearchMode] = "End of the word" Then
 WildCard = "Select * from GlobalFin Where Word Like " & Chr$(34) & _
            "*" & LookFor & Chr$(34)
Else
 WildCard = "Select * from GlobalFin Where Word Like " & _
            Chr$(34) & "*" & LookFor & "*" & Chr$(34)
End If
 Set Words = db.OpenRecordset(WildCard)
 If Words.RecordCount > 0 Then
 Words.MoveLast : Words.MoveFirst
 Else
  Exit Function
 End If
ReturnValue = SysCmd(acSysCmdSetStatus, "Done")
End If
 Set Addresses = dbSource.OpenRecordset("FinAddresses", dbOpenTable)
   Addresses.Index = "Word"
End If

Set List = db.OpenRecordset("List", dbOpenTable)
List.Index = "CodeText"

Set Texts = dbSource.OpenRecordset("Aligned", dbOpenTable)
   Texts.Index = "PrimaryKey"
Set OrigLib = db.OpenRecordset("Library", dbOpenTable)
  OrigLib.Index = "PrimaryKey"
Set TrLib = db.OpenRecordset("LibraryFin", dbOpenTable)
  TrLib.Index = "PrimaryKey"

If IsNull(LookFor) Then Exit Function
LookFor = LCase(Trim(LookFor))

CountExamples = 0 : Success = True : WCOK = True

Select Case Forms(FormName)![SearchMode]
Case "Whole word"
  SMode = 0
Case "Start of the word"
  SMode = 1
Case "Lemma"
  SMode = 3
Case Else
```

```
  SMode = 2
End Select

CountExamples = 0 : Success = True : WCOK = True

If SMode = 1 Then
   Words.Seek ">=", LookFor
ElseIf SMode = 0 Then
   Words.Seek "=", LookFor
ElseIf SMode = 4 Then
   Words.Seek ">=", LookFor
ElseIf SMode = 3 Then
   LemmaTb.Seek "=", LookFor
    If Not LemmaTb.NoMatch Then
      Words.Seek "=", LemmaTb!ID
    Else
     Exit Function
    End If
Else
  If Words.RecordCount = 0 Then Exit Function Else Success = True
End If

 If (SMode < 2) And Words.NoMatch Then Exit Function
  Do While Success = True
  Word = LCase(Words![Word])
    If SMode <= 1 Then
     If (SMode = 1 And InStr(Word, LookFor) > 0) Or (SMode = 0 And Word = LookFor) Then
      Success = True
     Else
      Success = False
       GoTo EndSearch
     End If
    ElseIf SMode = 3 Then
      If Words!LinkToLemm = LemmaTb!ID Then
       Success = True
      End If
    End If

   If SMode = 3 Then
    If Words!LinkToLemm <> LemmaTb!ID Then
      Success = False
      GoTo EndSearch
     End If
   End If

  WordNum = Words![ID]
  Addresses.Seek "=", WordNum
  If Not Addresses.NoMatch Then

 DoCmd.SetWarnings False
   DoCmd.RunSQL "Delete * from PhrNumTemp"
 DoCmd.SetWarnings True

  Do While Addresses![WordNo] = WordNum
'Копирование адресов во временную таблицу PhrNumTemp.
'Они могут потребоваться в дальнейшем при поиске словосочетаний
```

```
    ConcList.AddNew : ConcList!PhraseNo = Addresses!PhraseNo : ConcList.Update
    Addresses.Move TakeOrNot
      If Addresses.EOF Then Exit Do
  Loop
’Вызов проверки по второму ключу
  If Not IsNull(SecondKey) Then GoSub CheckSecondKey

’ Построение конкорданса
  ConcList.MoveFirst
Do Until ConcList.EOF
If Not IsNull(SecondKey) Then
'Пропуск контекстов, не подходящих по второму поисковому образу
       If Forms(FormName)![SecKeyPresent] = "present" Then
            If ConcList!Found = False Then GoTo EndConcList
       Else
          If ConcList!Found = True Then GoTo EndConcList
      End If
 End If
  Texts.Seek "=", ConcList!PhraseNo
    If Not Texts.NoMatch Then
      If FormName = "Search" Then
            SourceNum = Texts!Original
     Else
            SourceNum = Texts!Translation
    End If
      SearchResult = ""
     If List.RecordCount <> 0 Then
       List.Seek "=", SourceNum
        If List.NoMatch Then GoTo EndConcList:
’Пропуск текстов, не выбранных пользователем
     End If

FT = "" : RT = ""
If Forms(FormName)![NumSen] > 1 Then
    SNum(0) = Texts!ID : SNum(1) = Texts!Original : SNum(2) = Texts!Translation
     Texts.Move (-Forms(FormName)![ContSize])
LoopMove:
       If Texts!Original <> SNum(1) Or Texts!Translation <> SNum(2) Then
         Texts.MoveNext : GoTo LoopMove
        End If

 For e = 1 To Forms(FormName)!NumSen
     RusT = PMarks(Texts!RussianText) : FinT = PMarks(Texts!FinnishText)
      If Texts!ID = SNum(0) Then
        If FormName = "Search" Then
          ZZ = ZigZag(RusT, Word) : RusT = MarkWords(RusT, Word)
        Else
          ZZ = ZigZag(FinT, Word) : FinT = MarkWords(FinT, Word)
        End If
      End If

         RT = RT & RusT : FT = FT & FinT
       Texts.MoveNext
         If Texts.EOF Then Exit For
         If Texts!Original <> SNum(1) Or Texts!Translation <> SNum(2) Then Exit For
 Next e
Else
```

```
 RT = Texts!RussianText  : FT = Texts!FinnishText
  If FormName = "Search" Then
        ZZ = ZigZag(RT, Word) : RT = MarkWords(RT, Word)
      Else
        ZZ = ZigZag(FT, Word) : FT = MarkWords(FT, Word)
      End If
End If
’Добавить новую запись в таблицу с конкордансом
       Concordance.AddNew
        Concordance!Word = Word : Concordance![RusExample] = RT
        Concordance![FinExample] = FT : Concordance!SearchPat = ZZ
      OrigLib.Seek "=", Texts!Original
      If Not OrigLib.NoMatch Then
        If Not IsNull(OrigLib![Author]) Then
            If Right(OrigLib![Author], 1) = "." Then
                Concordance![Original] = OrigLib![Author] & " " & OrigLib![Title]
            Else
                Concordance![Original] = OrigLib![Author] & ". " & OrigLib![Title]
            End If        'Right(Lib![Author], 1) = "."
         Else
             Concordance![Original] = OrigLib![Title]
         End If        'Not IsNull(Lib![Author])
      End If            'Lib.NoMatch
    TrLib.Seek "=", Texts!Translation
      If Not TrLib.NoMatch Then
        If Not IsNull(TrLib![Author]) Then
            If Right(TrLib![Author], 1) = "." Then
                Concordance![Translation] = TrLib![Author] & " " & TrLib![Title] & _
                            ". " & TrLib![Translator]
            Else
                Concordance![Translation] = TrLib![Author] & ". " & TrLib![Title] & _
                    ". " & TrLib![Translator]
            End If        'Right(Lib![Author], 1) = "."
         Else
             Concordance![Translation] = TrLib![Title]
         End If        'Not IsNull(Lib![Author])
      End If            'Lib.NoMatch
     Concordance.Update
End If
  EndConcList:
ConcList.MoveNext
Loop        'ConcList.EOF

ErrSW:
   End If
   Words.MoveNext
   If Words.EOF Then Exit Do
  Loop

EndSearch:
SearchWordPar = CountExamples
Exit Function

'Подпрограммы
'*****************
CheckSecondKey:
’Проверка на второй поисковый образ
```

```
SecondKey = LCase(SecondKey)
If Forms(FormName)![Trans] = "Present" Then
  GoSub CheckTransl
  Return
End If

CopySecondKey = SecondKey
 If Forms(FormName)!SearchModeSec = "Whole word" Then
            WholeWordSec = True
 Else
            WholeWordSec = False
 End If

    PSlash = 1
Do While PSlash > 0
      PSlash = InStr(CopySecondKey, "/")
      If PSlash = 0 Then
        SecKey = CopySecondKey
      Else
        SecKey = Trim(Left(CopySecondKey, PSlash - 1))
        CopySecondKey = Trim(Right(CopySecondKey, Len(CopySecondKey) - PSlash))
      End If

 If Forms(FormName)![WhereSecKey] = "Same sentence" Then
      GoSub SecKeySameSentence
 Else
       GoSub NextWord
 End If
Loop
Return
'*******
SecKeySameSentence:
’Проверка на наличие второго поискового образа в том же предложении
If SMode = 3 Then
 WordsBkm = Words.Bookmark  : Words.Index = "Word"
End If

 If WholeWordSec Then
   Words.Seek "=", SecKey
 Else
   Words.Seek ">=", SecKey
 End If

 If Not Words.NoMatch Then
  Do While InStr(Words!Word, SecKey) > 0
       If WholeWordSec And Words!Word <> SecKey Then Exit Do
    Addresses.Seek "=", Words!ID
     If Not Addresses.NoMatch Then
       Do While Addresses!WordNo = Words!ID
            ConcList.Seek "=", Addresses!PhraseNo
              If Not ConcList.NoMatch Then
                 ConcList.Edit : ConcList!Found = True : ConcList.Update
              End If
           Addresses.MoveNext
       Loop
     End If
```

```
   Words.MoveNext
  Loop
 End If

If SMode = 3 Then
Words.Index = "LinkToLemm"
Words.Bookmark = WordsBkm
End If
Return

NextWord:
'Проверка соседних слов на совпадение со вторым поисковым образом
ConcList.MoveFirst
Do Until ConcList.EOF
  Texts.Seek "=", ConcList!PhraseNo
    If Not Texts.NoMatch Then
       If FormName = "Search" Then SearchPat = Texts!RussianText Else SearchPat =
Texts!FinnishText
            Neighbour = CheckNeighbours(SearchPat, Word, SecKey,
Forms(FormName)!SearchModeSec)
       If Forms(FormName)![WhereSecKey] = "Next word" And Neighbour <> 0 Then
            ConcList.Edit : ConcList!Found = True : ConcList.Update
       ElseIf Forms(FormName)![WhereSecKey] = "Next word to the left" And _
             (Neighbour = 1 Or Neighbour = 3) Then
            ConcList.Edit : ConcList!Found = True : ConcList.Update
       ElseIf (Or Forms(FormName)![WhereSecKey] = "Next word to the right") And _
             (Neighbour = 2 Or Neighbour = 3) Then
            ConcList.Edit : ConcList!Found = True : ConcList.Update
       End If
    End If
  ConcList.MoveNext
Loop
Return
'*************
CheckTransl:
'Проверка на наличие второго поискового образа в параллельном фрагменте
ConcList.MoveFirst

Do Until ConcList.EOF
Texts.Seek "=", ConcList!PhraseNo
    If Not Texts.NoMatch Then
       If FormName = "Search" Then SearchPat = Texts!FinnishText Else _
             SearchPat = Texts!RussianText
       ConcList.Edit
        r = CheckTransl(SearchPat, SecondKey, Forms(FormName)!SearchModeSec)
        If Forms(FormName)![Trans] = "Present" And r Then
            ConcList!Found = True
        ElseIf Forms(FormName)![Trans] = "Absent" And r = False Then
            ConcList!Found = True
        Else
            ConcList!Found = False
        End If
       ConcList.Update
    End If
ConcList.MoveNext
Loop
Return
```

```
End Function
```

4.3. Построение списка коллокаций

```
Public Sub Colloc(LookFor As String, Lemmas As Boolean, lang As String)
’Получение списка коллокаций для строки LookFor,
’которая может являться словоформой или лексемой.
’Если lang = ”Russian”, обрабатывается русский субкорпус,
’Если lang = ”Finnish” — финский
Dim db As Database, Sdb As Database
Dim Lm As Recordset, Words As Recordset
Dim Addresses As Recordset, Texts As Recordset, List As Recordset, Col As Recordset
Dim Nbr As Variant, Phr As String
Dim i As Integer, j As Integer, N As Integer, Cnt As Integer, k As Integer
Dim WordNum(1000) As Long, Wrd(1000) As String, WC As Integer

Set db = CurrentDb
DataLoc = GetTxDir
Set Sdb = DBEngine.Workspaces(0).OpenDatabase(DataLoc & DbName, True, True)
Set Col = db.OpenRecordset("Collocations")
Col.Index = "Collocation"

 If lang = "Russian" Then
   Set Lm = db.OpenRecordset("GlobalLemm", dbOpenTable)
   Set Words = db.OpenRecordset("GlobalWords", dbOpenTable)
   Set Addresses = Sdb.OpenRecordset("Addresses", dbOpenTable)
 Else
   Set Lm = db.OpenRecordset("GlobalLemFin", dbOpenTable)
   Set Words = db.OpenRecordset("GlobalFin", dbOpenTable)
   Set Addresses = Sdb.OpenRecordset("FinAddresses", dbOpenTable)
 End If

GlCount = 0
Lm.Index = "Word"
 If Lemmas = False Then
   Words.Index = "Word"
 Else
   Words.Index = "LinkToLemm"
 End If

Set Texts = Sdb.OpenRecordset("Aligned", dbOpenTable)
 Addresses.Index = "Word" : Texts.Index = "PrimaryKey"
Set List = db.OpenRecordset("List", dbOpenTable)
   List.Index = "CodeText"

 DoCmd.SetWarnings False
   DoCmd.RunSQL "Delete * from Collocations"
 DoCmd.SetWarnings True
’Уничтожается старый список коллокаций

If Lemmas = False Then
’Если поиск идет по словоформам, программа работает со словником по словоформам,
’если поиск идет по леммам, обрабатывается лемматизированный словник
Words.Seek "=", LookFor
 If Words.NoMatch Then Exit Sub
```

```
Else
 Lm.Seek "=", LookFor
   If Not Lm.NoMatch Then
     Words.Seek "=", Lm!ID
   Else
    Exit Sub
   End If
End If

WC = 0
 If Not Lemmas Then
  WordNum(WC) = Words![ID] : Wrd(WC) = Words!Word : WC = WC + 1
 Else
  Do While Words!LinkToLemm = Lm!ID
   WordNum(WC) = Words![ID] : Wrd(WC) = Words!Word : WC = WC + 1
   Words.MoveNext
  Loop
 End If

 For k = 0 To WC
  Addresses.Seek "=", WordNum(k)
 If Not Addresses.NoMatch Then
  Do While Addresses![WordNo] = WordNum(k)
   Texts.Seek "=", Addresses!PhraseNo
   If Not Texts.NoMatch Then
     If List.RecordCount <> 0 Then
     If lang = "Russian" Then List.Seek "=", Texts!Original Else List.Seek "=", Texts!Translation
          If List.NoMatch Then GoTo SkipText ' Skip the texts not selected by the user
     End If

    If lang = "Russian" Then
       Phr = Texts!RussianText
    Else
       Phr = Texts!FinnishText
    End If

       Nbr = ColList(Phr, " " & Wrd(k) & " ")
        ' Добавить новый коллокат в таблицу
        i = 0
        Do Until Nbr(i, 0) = "###"
          Col.Seek "=", Nbr(i, 0)
           If Not Col.NoMatch Then
            Col.Edit
             For j = 2 To 11
              Col.Fields(j).Value = Col.Fields(j).Value + Nbr(i, j - 1)
             Next j
            Col.Update
           Else
            Col.AddNew
             For j = 1 To 11
              Col.Fields(j).Value = Nbr(i, j - 1)
             Next j
            Col.Update
           End If
       i = i + 1
       Loop
   End If
```

```
SkipText:
    Addresses.MoveNext
     If Addresses.EOF Then Exit Do
  Loop
 End If
Next k

Call CalcZ(GlCount, lang)
'Вычисляется коэффициент Z
End Sub


'******************
Public Function ColList(S As String, W As String)
'Список коллокатов для слова W в строке S
Dim RCount As Byte, LCount As Byte, p As Integer
Dim Sl As String, Rl As String, ST As String
Dim CC(1000, 10) As Variant, i As Integer, j As Integer
Dim Word As String, Letter As String * 1, Count As Integer, t As Integer
Dim InWord As Boolean, Start As Boolean, l As Boolean, r As Boolean, Flag As Boolean
Dim N As Integer

S = " " & S & " "
ST = Prepare(S)
'Вызов функции, убирающей из строки знаки пунктуации
p = 1 :  t = 0 : CC(t, 0) = ""
 For j = 1 To 10: CC(t, j) = 0: Next j

Do Until p = 0
p = InStr(p, ST, W)
  If p = 0 Then Exit Do Else GlCount = GlCount + 1
  If p > 1 Then Sl = Trim(Left(S, p - 1)) Else Sl = ""
  If p + Len(W) < Len(S) Then Rl = Trim(Right(S, Len(S) - p - Len(W) + 1)) Else Rl = ""
Start = True : InWord = False : Word = "" : l = True : r = False : Count = 0

For i = Len(Sl) To 1 Step -1
Letter = Mid(Sl, i, 1)
 GoSub WrdProcess
  If Count = 5 Then Exit For
Next i
If Count < 5 And Word <> "" Then GoSub AddWord

Start = True : InWord = False : Word = "" : l = False : r = True : Count = 0
For i = 1 To Len(Rl)
 Letter = Mid(Rl, i, 1)
 GoSub WrdProcess
 If Count = 5 Then Exit For
Next i

If Count < 5 And Word <> "" Then GoSub AddWord
p = p + 1
Loop

CC(t, 0) = "###" : ColList = CC
Exit Function
'***************
WrdProcess:
```

```
  GoSub TestLetter
 If InWord = False And Start = False Then
   Start = True
    GoSub AddWord
    Count = Count + 1 : Word = ""
ElseIf InWord = True Then
 Start = False
 If l Then Word = Letter & Word Else Word = Word & Letter
End If
Return
'********
AddWord:       'Подпрограмма, добавляющая новый коллокат в список
Word = LCase(Word)
If l Then N = 5 - Count Else N = 6 + Count
Flag = False
For j = 0 To t
 If CC(j, 0) = Word Then
   CC(j, N) = CC(j, N) + 1 : Flag = True
 End If
Next j
If Flag = False Then
 CC(t, 0) = Word
 For j = 1 To 10 : CC(t, j) = 0 : Next j
 CC(t, N) = 1
    t = t + 1
End If
Return
'*******
TestLetter:            'Подпрограмма, проверяющая, является ли символ буквой
  Select Case Letter
  Case " ", ",", ";", ":", ".", "!", "?", "(", ")", "=", "+", Chr$(32)
     InWord = False
   Case Else
     InWord = True
  End Select
Return
End Function
```

4.4. Поиск переводных эквивалентов в автоматическом режиме

```
Public Sub BuildGlos()
'Строит список русско-финских соответствий в заданном диапазоне частот
'в данном случае — от 20 до 220
Dim db As Database
Dim GlosTb As Recordset, LemTb As Recordset
Dim EqTb As Recordset, LogTb As Recordset
Dim Total As Long, Processed As Long, Got As Long, Count As Integer
Dim ReturnValue  As Variant
Dim Scs As Boolean

Set db = CurrentDb
Set LemTb = db.OpenRecordset("Select * FROM GlobalLemm WHERE _
                    Count >= 20 and Count <= 220 ORDER BY Count")
 LemTb.MoveLast : LemTb.MoveFirst
 Total = LemTb.RecordCount : Processed = 0 : Got = 0 : Count = 0
```

```
Set GlosTb = db.OpenRecordset("Glossary") : Set LogTb = db.OpenRecordset("LogGlos")
     LogTb.Index = "Freq"

Do Until LemTb.EOF
  Scs = FindEquivalent(LemTb!Word)
'Вызов функции FindEquivalent, которая ищет возможные эквиваленты
'для текущего слова из словника корпуса. Результаты поиска записываются в
'таблицу CoocTab
   LogTb.Seek "=", LemTb!Count
     If LogTb.NoMatch Then
      LogTb.AddNew
        LogTb!Freq = LemTb!Count
      Else
        LogTb.Edit
      End If
        LogTb!NumberOfWords = LogTb!NumberOfWords + 1
     If Scs Then
          LogTb!Found = LogTb!Found + 1
     Set EqTb = db.OpenRecordset("CoocTab")
      Do Until EqTb.EOF
       GlosTb.AddNew
         GlosTb!Rus = LemTb!Word : GlosTb!Freq = LemTb!Count : GlosTb!Fin = EqTb!Word
         GlosTb!FreqFin = EqTb!Tot : GlosTb!CrossCount = EqTb!Freq
         GlosTb!KUC = EqTb!KUC
       GlosTb.Update
            Got = Got + 1
        EqTb.MoveNext
      Loop
     EqTb.Close
    Else
         LogTb!NotFound = LogTb!NotFound + 1
    End If
       LogTb.Update
            Processed = Processed + 1
 LemTb.MoveNext
 Loop
 End Sub
'**************
Public Function FindEquivalent(Word As String)
'Поиск возможных эквивалентов для слова Word
Dim db As Database, Sdb  As Database
Dim WLem As Recordset, WLemF As Recordset, WList As Recordset, WListF As Recordset
Dim Texts As Recordset, Ind As Recordset, TmTb As Recordset, t As Recordset
Dim StopL As Recordset
Dim WArr As Variant, i As Long
Dim FrW As Long, FrC As Integer
Dim a As Long, b As Long, c As Long, k As Single

Set db = CurrentDb
Set Sdb = DBEngine.Workspaces(0).OpenDatabase(InitDb & DbName)

DoCmd.SetWarnings False
 DoCmd.RunSQL "Delete * from CoocTab"
 DoCmd.RunSQL "Delete * from TmpNum"
DoCmd.SetWarnings True
  Set WLem = db.OpenRecordset("GlobalLemm")
    WLem.Index = "Word"
```

```
  Set WList = db.OpenRecordset("GlobalWords")
    WList.Index = "LinkToLemm"
  Set Ind = Sdb.OpenRecordset("Addresses")
    Ind.Index = "Word"

  Set WLemF = db.OpenRecordset("GlobalLemFin")
    WLemF.Index = "PrimaryKey"
  Set WListF = db.OpenRecordset("GlobalFin")
    WListF.Index = "Word"

Set Texts = Sdb.OpenRecordset("Aligned", dbOpenTable)
   Texts.Index = "PrimaryKey"
Set StopL = db.OpenRecordset("StopList")
   StopL.Index = "Word"

Set TmTb = db.OpenRecordset("CoocTab")
    TmTb.Index = "Word"
Set t = db.OpenRecordset("TmpNum")
   t.Index = "PrimaryKey"

WLem.Seek "=", Word
 If WLem.NoMatch Then
          MsgBox "Word is not in the list"
   Exit Function
 End If
 FrW = WLem!Count

  WList.Seek "=", WLem!ID
  If Not WList.NoMatch Then
    Do While WList!LinkToLemm = WLem!ID
      Ind.Seek "=", WList!ID
       If Not Ind.NoMatch Then
         Do While Ind!WordNo = WList!ID
          Texts.Seek "=", Ind!PhraseNo
           If Not Texts.NoMatch Then
            If CountWords(Texts!FinnishText) = 0 Then
              FrW = FrW - 1
              GoTo SkipEmpty
            End If
             t.Seek "=", Texts!ID
              If Not t.NoMatch Then
                GoTo SkipEmpty
              Else
                t.AddNew : t!PhrNum = Texts!ID : t.Update
              End If
             WArr = GetWordList(Texts!FinnishText)
             i = 0

             Do While WArr(i) <> "0"
              StopL.Seek "=", WArr(i)
               If StopL.NoMatch Then
               WListF.Seek "=", WArr(i)
                If Not WListF.NoMatch Then
                 If WListF!LinkToLemm <> 0 Then
                  WLemF.Seek "=", WListF!LinkToLemm
                   If Not WLemF.NoMatch Then
                     TmTb.Seek "=", WLemF!Word
```

```
                    If Not TmTb.NoMatch Then
                      TmTb.Edit : TmTb!Freq = TmTb!Freq + 1 : TmTb.Update
                    Else
                     TmTb.AddNew
                      TmTb!Word = WLemF!Word : TmTb!Freq = 1 : TmTb!Tot = WLemF!Count
                     TmTb.Update
                    End If
                  End If
                 End If
                End If
               End If
              i = i + 1
              Loop
             End If
SkipEmpty:
           Ind.MoveNext
            If Ind.EOF Then Exit Do
          Loop
        End If
      WList.MoveNext
       If WList.EOF Then Exit Do
    Loop
  End If

If TmTb.RecordCount > 0 Then TmTb.MoveFirst
 Do Until TmTb.EOF
    a = TmTb!Freq
     If a > FrW Then a = FrW
    b = FrW – a : c = TmTb!Tot – a : k = a / 2 * (1 / (a + b) + 1 / (a + c))
'Вычисляется коэффициент Кульчинского. Если значение больше или равно 0,55,
'финское слово считается эквивалентом проверяемому русскому
    If k >= 0.55 Then
      TmTb.Edit
       TmTb!Freq = a : TmTb!KUC = k
      TmTb.Update
    Else
      TmTb.Delete
    End If
  TmTb.MoveNext
 Loop
  If TmTb.RecordCount > 0 Then FindEquivalent = True Else FindEquivalent = False
 TmTb.Close
End Function
```

4.5. Модуль лемматизации

4.5.1. Лемматизатор русского словника[51]

```
Public Sub Lemmatization(Ll As Boolean)
```

[51] В целях экономии места в текстах программ лемматизации для русского и финского языков приводятся только главные процедуры и функции.

```
'Ll = true — лемматизируется словник из корпуса текстов
’Ll = false — лемматизируется словник, помещенный в базе данных лемматизатора
Dim db As Database
Dim WListTb As Recordset , LemTb As Recordset
Dim LemAmb As Recordset, UndefTb As Recordset, Lm As Variant
Dim Lemmas(10, 3) As String
Dim Vl As String
Dim i As Integer, PosDef As Integer
Dim StartTime, EndTime As Single
Dim ProcCount As Long, PercDone As Variant, Counter As Integer
Dim Total As Long
Dim CountI As Integer, CountJ As Integer, Cnt As Integer

ProcCount = 0 : Counter = 0 : Cnt = 0

If Ll Then
Set db = DBEngine.Workspaces(0).OpenDatabase(Forms!GetDatabase!Path)
Else
Set db = CurrentDb
End If

Set WListTb = db.OpenRecordset("GlobalWords", dbOpenTable)
 WListTb.Index = "LinkToLemm"
Set LemTb = db.OpenRecordset("GlobalLemm", dbOpenTable)
 LemTb.Index = "Word"
Set LemAmb = db.OpenRecordset("Ambivalent", dbOpenTable)
 LemAmb.Index = "Word"
Set UndefTb = db.OpenRecordset("Undefined", dbOpenTable)
 UndefTb.Index = "Word"

Total = WListTb.RecordCount
ReturnValue = SysCmd(acSysCmdSetStatus, "Working... Please wait.")

Do While WListTb!LinkToLemm = 0
 If CyrTest(WListTb!Word) Then

’Перед тем, как вызывать морфоанализатор, программа проверяет, не встретилась ли
’данная словоформа в таблице неоднозначных форм или таблице неопознанных форм
 LemAmb.Seek "=", WListTb!Word
   If Not LemAmb.NoMatch Then
         GoTo GoOn
   End If

UndefTb.Seek "=", WListTb!Word
   If Not UndefTb.NoMatch Then
         GoTo GoOn
   End If

 PosDef = InStr(WListTb!Word, "-")
  Lm = Analize(WListTb!Word)
’Вызов морфоанализатора, варианты разбора сохраняются в переменной Lm
 If PosDef > 0 And FoundLemmas = 0 Then
  Lm = Dash(WListTb!Word, PosDef)
 End If

’Снятие повторяющихся вариантов разбора
Lm = EraseDuplicates(Lm)
```

```
  Select Case FoundLemmas
   Case 0
    If Len(WListTb!Word) <= 50 Then
     UndefTb.Seek "=", WListTb!Word
     If UndefTb.NoMatch Then
      UndefTb.AddNew
      UndefTb!Word = WListTb!Word
      UndefTb!LinkToGlobal = WListTb!ID
      UndefTb!Count = WListTb!Count
      UndefTb.Update
     Else
      UndefTb.Edit
       UndefTb!Count = WListTb!Count
      UndefTb.Update
     End If
    End If
   Case 1
    LemTb.Seek "=", Lm(0, 0)
    If LemTb.NoMatch Then
     LemTb.AddNew
     LemTb!Word = Lm(0, 0) : LemTb!Count = WListTb!Count
      WListTb.Edit
       WListTb!LinkToLemm = LemTb!ID
      WListTb.Update
     LemTb.Update
    Else
    LemTb.Edit
     LemTb!Count = LemTb!Count + WListTb!Count
    LemTb.Update
     WListTb.Edit
       WListTb!LinkToLemm = LemTb!ID
     WListTb.Update
    End If
   Case Else
   LemAmb.Seek "=", WListTb!Word
   If Not LemAmb.NoMatch Then
    LemAmb.Edit
     LemAmb!Count = WListTb!Count
    LemAmb.Update
   Else
     LemAmb.AddNew
      LemAmb!Word = WListTb!Word : LemAmb!LinkToGlobal = WListTb!ID
       Vl = ""
       For i = 0 To FoundLemmas - 1
        Vl = Vl & Lm(i, 0) & ";" & Lm(i, 1) & ";"
       Next i
       Vl = Left(Vl, Len(Vl) - 1)
       LemAmb!VarLemm = Vl : LemAmb!Count = WListTb!Count
      LemAmb.Update
     End If
   End Select
  End If

GoOn:
   WListTb.MoveNext
     If WListTb.EOF Then Exit Do
```

```
Loop

End Sub

'*****************
Public Function Analize(Wrd As String)
'Функция выполняет морфологический анализ словоформы Wrd
'и возвращает все найденные леммы для этой формы
Dim DicDb As Database
Dim Dict As Recordset, FreqWords As Recordset
Dim GrammarMain As Recordset, GrammarDescr As Recordset
Dim Prefixes As Recordset, NoPrefix As String, PrPref As String
Dim NoPostfix As String, Ending As Variant
Dim Class As String, postfix As String
Static Lemmas(30, 3) As String
Dim i As Integer, j As Integer
Dim NoPostfixes As Boolean, AnyPostfix As Boolean
Dim FirstPref As String, PrPostfix As String
Dim Take As Boolean
Dim SQLStr As String, QueryStr As String
Dim CL As Integer, CFL As Boolean, Sh As Boolean, GrDes As String
Dim TempLemm As String, TempGr As String, TempGrDes As String

Set DicDb = CurrentDb
Set Dict = DicDb.OpenRecordset("Dictionary")
Dict.Index = "Osnova"
Set GrammarMain = DicDb.OpenRecordset("GrammarMain", dbOpenTable)
GrammarMain.Index = "Ending"
Set Prefixes = DicDb.OpenRecordset("Prefixes", dbOpenTable)
Prefixes.Index = "Prefix"

FoundLemmas = 0 : FirstPref = ""
Wrd = DeleteJo(Wrd) ’Заменить букву ё на е
NoPostfixes = False : AnyPostfix = False
'Сначала делаем цикл по префиксам: отрезаем слева сначала ничего, потом одну
'букву, потом две ... до тех пор пока не будет отрезано пять букв
'То есть сначала проверяется слово целиком, потом без префикса1, потом — без
'префикса2...
For i = 0 To 6
If i <> 0 And i = Len(Wrd) - 1 Then Exit For
Ne:
PrPref = Left(Wrd, i)
If PrPref <> "" Then
 Prefixes.Seek "=", PrPref
  If Not Prefixes.NoMatch Then
   NoPrefix = Right(Wrd, Len(Wrd) - Len(PrPref)) : PAttr = Prefixes!Description
 Else
   GoTo SkipPrefix
  End If
Else
 NoPrefix = Wrd
End If
'  Вложенный цикл по постфиксам — аналогично
TestPostfix: 'Сюда возврат, чтобы проверить слово без постфикса

'    Вложенный цикл по окончаниям: сначала без окончания,
'    затем по всем окончаниям: если финаль зарегистрирована, отрезаем ее
```

```
'    от основы, делаем поиск основы в базе и проверяем грамматику по
'    базе GrammarDescriptions, связанной с гл. грам базой по полю
'    LinkToMain. Найденные варианты запоминаем.
For j = 0 To 6
 If j <> 0 And j = Len(NoPrefix) - 1 Then Exit For
  PrEnd = Right(NoPrefix, j) 'Ending
   GrammarMain.Seek "=", PrEnd
    If Not GrammarMain.NoMatch Then 'Ending exists
     PrStem = Left(NoPrefix, Len(NoPrefix) - Len(PrEnd)) 'Stem = Word - Ending

     Dict.Seek "=", PrStem  'Look for stem in stems db
    If Not Dict.NoMatch Then  'Found
     SQLStr = "Select * From GrammarDescription Where _
                    LinkToMain = " & GrammarMain!ID
     Set GrammarDescr = DicDb.OpenRecordset(SQLStr)

     Do While Dict!Osnova = PrStem
      ExactClass = Dict!Klass
      If Dict!Klass =  "нсв" And PrPref <> "" _
                    And PrPref <>  "не" And _
                    PrPref <>  "само" And _
                    PrPref <>  "полу" And _
                    InStr(Marks, "#" = 0 Then ExactClass = "св"
      Marks = Dict!Marks
      If IsNull(Marks) Then Marks = ""
      Class = Decipher(Dict!Klass, Dict!Scheme)
      Gram = Dict!Grammar
      If Dict!Grammar =  "глаг. несов." And PrPref <> "" And _
        PrPref <>  "не" And PrPref <> "само" And _
         PrPref <>  "полу" And InStr(Marks, "#" = 0 _
                    Then Gram =  "глаг. сов."
      If (InStr(Marks, "[" > 0 Or Dict!Typ = 0) And PrEnd = "" Then
         'аномальные формы
         GrDes = ""
         Take = TakeOrNot(Class, Dict!Typ, GrammarDescr!Type, _
                    PrPref, AnyPostfix, GrDes, GrammarDescr!Marks)
         If Take And PrPref <> "" Then
           If TestPrefix(PrPref, Dict!Lemma) = False Then
            Take = False
           Else
            If (FirstPref = "не" Or _
             PrPref = "не") And _
             InStr(Dict!Grammar, "глаг") = 0 Then
               Take = False
            ElseIf FirstPref = "не" Then
               FirstPref = ""
            ElseIf PrPref = "не" Then
               PrPref = ""
            End If
           End If
         End If

         If Take And AnyPostfix Then 'Проверка постфикса
           If Not PostfixTest(NoPrefix, PrPostfix) Then Take = False
         End If
         If Take Then
          If AnyPostfix Then postfix = GetPostfix(Dict!Lemma) Else postfix = ""
```

```
    If PrPref <> "" Then PrPref = GetPrefix(PrPref, Dict!Lemma)

    Lemmas(FoundLemmas, 0) = FirstPref & PrPref & Dict!Lemma & postfix
    Lemmas(FoundLemmas, 1) = Gram : Lemmas(FoundLemmas, 2) = GrDes
     If PrPref <> "" Then
      Lemmas(FoundLemmas, 3) = "*"
     Else
      Lemmas(FoundLemmas, 3) = ""
     End If
       FoundLemmas = FoundLemmas + 1
    End If
     GoTo ContinueAnalysis 'Пропускаем поиск окончаний
ElseIf (InStr(Marks, "[" > 0 Or Dict!Typ = 0) And PrEnd <> "" Then
   GoTo ContinueAnalysis 'Основа получилась случайно
End If
'Расшифровка класса
 If Not IsNull(Dict!Scheme) Then Scheme = Dict!Scheme Else Scheme = ""
 Class = Decipher(Dict!Klass, Dict!Scheme)
 If Class <> "Verbs" Then
  QueryStr = "Class = '" & Class & "' And Type = " & Dict!Typ
 Else
  QueryStr = "Class = 'Verbs'" ' & " And Type = 0"
 End If
GrammarDescr.FindFirst QueryStr
 Do Until GrammarDescr.NoMatch
    PrLemma = Dict!Lemma
    GrDes = GrammarDescr!Descr
     Take = TakeOrNot(Class, Dict!Typ, GrammarDescr!Type, _
                PrPref, AnyPostfix, GrDes, GrammarDescr!Marks)
     If Take And AnyPostfix Then 'Проверка постфикса
      If Not PostfixTest(NoPrefix, PrPostfix) Then Take = False
     End If

     If Take And PrPref <> "" Then 'Проверка префикса
      If TestPrefix(PrPref, Dict!Lemma) = False Then
       Take = False
      Else
       If (FirstPref = "не" Or PrPref = "не") And _
         InStr(Dict!Grammar, "глаг.") = 0 Then
         Take = False
       ElseIf FirstPref = "не" Then
         FirstPref = ""
       ElseIf PrPref = "не" Then
         PrPref = ""
       End If
      End If
     End If

     If Take Then
      If AnyPostfix Then postfix = GetPostfix(Dict!Lemma) Else postfix = ""
      If PrPref <> "" Then PrPref = GetPrefix(PrPref, Dict!Lemma)
      Lemmas(FoundLemmas, 0) = FirstPref & PrPref & Dict!Lemma & postfix
      Lemmas(FoundLemmas, 1) = Gram : Lemmas(FoundLemmas, 2) = GrDes
       If PrPref <> "" Then
        Lemmas(FoundLemmas, 3) = "*"
       Else
        Lemmas(FoundLemmas, 3) = ""
```

```
            End If
            FoundLemmas = FoundLemmas + 1
          End If
        GrammarDescr.FindNext QueryStr
       Loop
ContinueAnalysis:
      Dict.MoveNext
          If Dict.EOF Then Exit Do
    Loop 'Stems
  End If
 End If
Next j 'Конец цикла по окончаниям

If (NoPostfixes = False) And_
            (Right(NoPrefix, 2) = "ся" Or _
            Right(NoPrefix, 2) = "сь") Then
  PrPostfix = Right(NoPrefix, 2)
  NoPrefix = Left(NoPrefix, Len(NoPrefix) - 2)
  NoPostfixes = True : AnyPostfix = True
  GoTo TestPostfix
End If

If (PrPref = "не" Or _
PrPref = "полу" Or PrPref = "само") And _
    Class <> "Masculine" And Class <> "Feminine" And Class <> "Neuter" Then
 FirstPref = PrPref
 PrPref = ""
 Wrd = NoPrefix & PrPostfix
 i = 0
 GoTo Ne  'надо избежать приращения i
End If
SkipPrefix:
  NoPostfixes = False : AnyPostfix = False : PrPostfix = ""
Next i
EndSearch:

For CL = 0 To FoundLemmas - 1
'Помечаем прилагательные как проблематичные, если есть наречия
 If Lemmas(CL, 1) = "н" Or _
InStr(Lemmas(CL, 1), "предик.") > 0 Or _
    InStr(Lemmas(CL, 1), "вводн.") > 0 Or _
            InStr(Lemmas(CL, 1), "предл.") > 0 _
     Or InStr(Lemmas(CL, 1), "част.") > 0 Or _
            InStr(Lemmas(CL, 1), "сравн.") > 0 Then
   TempLemm = Lemmas(CL, 0) : TempGr = Lemmas(CL, 1) : TempGrDes = Lemmas(CL, 2)
   FoundLemmas = 1
    Lemmas(0, 0) = TempLemm : Lemmas(0, 1) = TempGr : Lemmas(0, 2) = TempGrDes
     GoTo GetTime
 End If
Next CL

'Отбрасываем сомнительные леммы, если есть несомненные
If FoundLemmas > 1 Then
 'Есть ли несомненные леммы?
  For CL = 0 To FoundLemmas - 1
   If Lemmas(CL, 3) = "" Then
    CFL = True
```

```
   Exit For
  End If
 Next CL

 If CFL Then
  For CL = 0 To FoundLemmas - 1
   If Lemmas(CL, 3) = "*" Then
    If CL = FoundLemmas - 1 Then
     FoundLemmas = FoundLemmas - 1
     Exit For
    End If
    For i = CL To FoundLemmas - 2
      Lemmas(i, 0) = Lemmas(i + 1, 0) : Lemmas(i, 1) = Lemmas(i + 1, 1)
      Lemmas(i, 2) = Lemmas(i + 1, 2) : Lemmas(i, 3) = Lemmas(i + 1, 3)
      Sh = True
    Next i
    If Sh Then
      FoundLemmas = FoundLemmas - 1
      CL = CL – 1 : Sh = False
    End If
   End If
  Next CL
 End If
End If
GetTime:
Analize = Lemmas
End Function
```

4.5.2. Лемматизатор финского словника

```
Public Function AnalysisFin(Word As String)
’Получение начальных форм для словоформы Word
Dim db As Database
Dim SuffTb As Recordset, StemTb As Recordset, TestList As Recordset
Dim ExceptionsTb As Recordset
’Таблицы с данными
Dim Lemmas(10, 3) As String, addlemmas As Variant
Dim al As Integer, el As Integer
Dim PrEnd As String, PrSuff As String, PrS As String, PrStem As String
Dim PrStemPTK As Variant
Dim PTKCount As Integer
Dim ProposedForm As String, Ending As String, LEnding As Variant
Dim i As Integer, j As Integer, k As Integer
Dim LWord As Byte
Dim Found As Boolean, DV As Boolean
Dim Osn As String, ActOsn As String
Dim CondStr As Variant, ConCount As Integer
Dim FormName As String, PartOfSpeech As String
Dim LemPS As String, GTyp As String, PTK As String
Dim VarStems As Variant, Partik As Variant, OLit As Variant
Dim FOl As String
Dim CountPart As Byte, CountOL As Byte
Dim TE As String, AddN As Boolean
Dim LemCount As Integer, Lemma As String
Dim PrtkFound As Boolean, Comments As Variant
```

```
Partik = Array("kaan", "kään", "han", "hän", _
          "pas", "päs", "pa", "pä", "ko", "kö", "kin", "0")
CountPart = 0
’Список частиц
OLit = Array("ni", "si", "nsa", "nsä", "mme", "nne", "=n", "0")
CountOL = 0
’Список притяжательных суффиксов

LemCount = 0 : AcForm = True

Set db = CurrentDb
Set TestList = db.OpenRecordset("WordList", dbOpenTable)
TestList.Index = "Word"
Set SuffTb = db.OpenRecordset("Suffixes", dbOpenTable)
SuffTb.Index = "Ending"
Set ExceptionsTb = db.OpenRecordset("Exceptions", dbOpenTable)
ExceptionsTb.Index = "Word"
Set StemTb = db.OpenRecordset("Stems", dbOpenTable)
StemTb.Index = "Description"

OL = False : FOl = "" : PrtkFound = True : Prtk = False : LWord = Len(Word) : Found = False
AddN = False : GTyp = ""

' Начало цикла по частицам
Do While PrtkFound
' Начало цикла по притяжательным суффиксам
Do While Word <> ""
' Начало цикла по словоизменению
  Comments = Null
TestList.Seek "=", Word   'Поиск всего слова
 If Not TestList.NoMatch Then
 Do While TestList!Word = Word
 If IsNull(TestList!PTK) Or Not TestList!PTK = "XX" Then
  If Not (TestList!Grammar > 19 And OL) Or IsNull(TestList!Grammar) Then
   ProposedForm = Word : Lemma = TestList!Lemma
   PartOfSpeech = GetPs(TestList!Grammar)
    Select Case PartOfSpeech
     Case "subst."
      FormName = "nom. y."
     Case "verb"
      FormName = "inf."
     Case Else
      FormName = "-"
     End Select
       GoSub AssignLemma
      If PartOfSpeech = "adverb" Then GoTo EndOfAnalysis
  End If
End If
   TestList.MoveNext
   If TestList.EOF Then Exit Do
 Loop
 End If

   ExceptionsTb.Seek "=", Word ' Поиск всего слова в таблице исключений
   If Not ExceptionsTb.NoMatch Then
    If Not OL Or ExceptionsTb!PartOfSpeech = "subst" Then
```

```
      ProposedForm = ExceptionsTb!Lemma
      Lemma = ExceptionsTb!Lemma : FormName = ExceptionsTb!Description
      PartOfSpeech = ExceptionsTb!PartOfSpeech
        GoSub AssignLemma
        GoTo EndOfAnalysis
     End If
    End If
LWord = Len(Word)
For i = LWord To 0 Step -1
 PrSuff = Right(Word, i)
 If PrSuff = "" Then PrSuff = "0" 'Нулевое окончание
 Found = False
  SuffTb.Seek "=", PrSuff  'Поиск в таблице суффиксов
  If Not SuffTb.NoMatch Then
  PrS = Left(Word, LWord - i)
   Do While PrSuff = SuffTb!Ending
    Osn = SuffTb!Osnovy : ActOsn = Osn

   If Not (IsNull(SuffTb!Condition) Or SuffTb!Condition = "" _
            Or SuffTb!Condition = "-") Then
' Проверяем, указано ли для суффикса условие, после проверки ищем основу без
' модификаций
    PrStem = TestCondition(PrS, SuffTb!Condition)
   Else
    PrStem = PrS
   End If

  If PrStem <> "" Then
   If SuffTb!Condition <> "-" Then
      LEnding = ""
      If SuffTb!Osnovy = "part/mon" Then ActOsn = "part" Else ActOsn = Osn
      FormName = SuffTb!Description
      PartOfSpeech = SuffTb!PartOfSpeech
       If PartOfSpeech = "verb" And (IsNull(SuffTb!Condition) Or _
                                SuffTb!Condition = "") Then
        If Front(PrS) Then     'Учет сингармонизма
                    PrStem = PrStem & "ä"
           Else
                    PrStem = PrStem & "a"
           End If
       End If
       GTyp = ""
      If Not (PartOfSpeech = "verb" And OL And InStr(FormName, "partis.") = 0 And _
        InStr(FormName, "inf. 2.") = 0 And InStr(FormName, "inf. 5.") = 0 And _
        FormName <> "inf. 1. transl.") Then
          GoSub TestEnding
      End If
    End If

  If SuffTb!Osnovy <> "-" Then
    CondStr = SetCondition(Osn) : ConCount = 0

   Do Until CondStr(ConCount) = "0"
    StemTb.Seek "=", CondStr(ConCount)
    If Not StemTb.NoMatch Then
     Do While StemTb!Description = CondStr(ConCount)
     ActOsn = CondStr(ConCount)
```

```
        'Пропустить глагольные основы для притяж. суфф.
        If OL And ((ActOsn <> "gen" And ActOsn <> "mon" And InStr(ActOsn, "part") = 0) _
           And InStr(SuffTb!Description, "inf. 2.") = 0 And _
             InStr(SuffTb!Description, "inf. 5.") = 0 _
            And SuffTb!Description <> "inf. 1. transl.") Then GoTo SkipVerbs
         If Right(PrS, Len(StemTb!Ending)) = StemTb!Ending And PrS <> StemTb!Ending Then
          Comments = StemTb!Comment : Ending = StemTb!Ending
           PrStem = Left(PrS, Len(PrS) - Len(Ending))
           LEnding = StemTb!LemmaEnding : FormName = SuffTb!Description
           PartOfSpeech = StemTb!PartOfSpeech : GTyp = StemTb!GT
           GoSub TestEnding
          Comments = Null
         End If
SkipVerbs:
         StemTb.MoveNext
           If StemTb.EOF Then Exit Do
        Loop
       End If
     ConCount = ConCount + 1
     Loop
End If
End If
  If (Right(Word, 2) = "aa" Or Right(Word, 2) = "ää") _
                And SuffTb!Description = "3. pers. y." And Not OL Then
        PrStem = Left(Word, Len(Word) - 1)
         If Front(Word) Then
           LEnding = "tä"
          Else
           LEnding = "ta"
          End If
         FormName = SuffTb!Description : PartOfSpeech = "verb" : GTyp = "/73/74/75/"
           GoSub TestEnding
  End If
     SuffTb.MoveNext
      If SuffTb.EOF Then Exit Do
   Loop
  End If 'SuffTb.NoMatch
Next i

' Конец цикла по словоизменению
CountOL = 0
If OL = False Then
  Do While OLit(CountOL) <> "0"
  If Len(Word) > Len(OLit(CountOL)) Then
   If Right(Word, Len(OLit(CountOL))) = OLit(CountOL) And _
      Vowel(Mid(Word, Len(Word) - Len(OLit(CountOL)), 1)) Then
     Word = Left(Word, Len(Word) - Len(OLit(CountOL)))
     OL = True : FOl = OLit(CountOL) : TE = Right(Word, 3)
     If TE = "lla" Or TE = "llä" Or TE = "lle" Or TE = "lta" Or _
        TE = "ltä" Or TE = "ssa" Or TE = "ssä" Or _
        TE = "sta" Or TE = "stä" Then
       AddN = False
     Else
      AddN = True
     End If
     Exit Do
   End If
```

```
  If OLit(CountOL) = "=n" And TestEn(Word) Then
    FOl = Right(Word, 2) : Word = Left(Word, Len(Word) - 2) : OL = True : AddN = False
    Exit Do
  End If
 End If
   CountOL = CountOL + 1
 Loop
 Else
  If AddN Then
   Word = Word & "n" : AddN = False
  Else
   OL = False
 Exit Do
  End If
 End If

 If OL = False Then Exit Do
Loop
' Конец цикла по притяжательным суффиксам
Word = Word & FOl : FOl = "" : OL = False : Prtk = False
Do While Partik(CountPart) <> "0"
   If Right(Word, Len(Partik(CountPart))) = Partik(CountPart) Then
     Word = Left(Word, Len(Word) - Len(Partik(CountPart)))
     Prtk = True
     Exit Do
   End If
   CountPart = CountPart + 1
Loop
  CountPart = 0

  If Not Prtk Then PrtkFound = False
Loop
' Конец цикла по частицам
EndOfAnalysis:
  Lemmas(LemCount, 0) = "0" 'Отметить конец списка
  LC = LemCount
AnalysisFin = Lemmas
Exit Function

'********** Подпрограммы ***************
AssignLemma:
 DV = False
 If LemCount >= 1 Then
  For k = 0 To LemCount
  If Lemmas(k, 0) = Lemma Then
   DV = True
   Exit For
  End If
 Next k
 End If
  If DV = False Then
    Lemmas(LemCount, 0) = Lemma : Lemmas(LemCount, 1) = FormName
    Lemmas(LemCount, 2) = PartOfSpeech : LemCount = LemCount + 1
   End If
   Found = True

 Return
```

```
'************************
TestEnding:
   PTK = GetPTK(FormName, ActOsn, PrStem & LEnding)

    If Not IsNull(Comments) Then
      If InStr(Comments, "/c") > 0 Then
        If Len(PrStem) < 3 Then Return
        If Vowel(Mid(PrStem, Len(PrStem) - 2, 1)) Then Return
       End If
      If InStr(Comments, "/v") > 0 And PTK <> "v" Then Return
      If InStr(Comments, "/h") > 0 And PTK = "v" Then Return
    End If

   Found = False
    ProposedForm = PrStem & LEnding
    TestList.Seek "=", ProposedForm
     If Not TestList.NoMatch Then
        Do While TestList!Word = ProposedForm
        If Not IsNull(TestList!Lemma) Then
         Lemma = TestList!Lemma
           If IsNull(TestList!Vaihtelu) Or TestList!Vaihtelu = PTK Or TestList!Vaihtelu = "%" _
                    Or PTK = " " Or Definite = False _
                Or InStr(Comments, "/0") > 0 Then
               GoSub LemGram
                  If Found Then GoSub AssignLemma
           End If
       End If
          TestList.MoveNext
          If TestList.EOF Then Exit Do
        Loop
   End If
GTyp = ""
Return
'*************************
LemGram:
' Проверка грамматики: совпадают ли ожидания с грамм. пометой
LemPS = GetPs(TestList!Grammar)
      If TestList!POS = "v/p" And (InStr(PartOfSpeech, "part") <> 0) Then
        Found = True          'Причастия
        Return
      End If
      If (TestList!POS = "a/cm" Or TestList!POS = "a/sl") And _
          InStr(PartOfSpeech, "adj.") <> 0 Then    'Сравнительная или превосходная степень
        Found = True
        Return
      End If

      If PartOfSpeech = "-" Then
       Found = True
       PartOfSpeech = LemPS
      Else
        If GTyp <> "" Then
         If InStr(GTyp, "/" & TestList!Grammar & "/") > 0 Then Found = True
        Else
           If LemPS = PartOfSpeech Then Found = True
        End If
      End If
```

```
   If GTyp = "" And SuffTb!Condition = "v+e" Then
    Select Case TestList!Grammar
      Case 7, 16
        Found = False
      Case 48
        If Right(ProposedForm, 2) = "ee" Then Found = True Else Found = False
      Case Is > 22
        Found = False
       Case Else
        Found = True
    End Select
   End If
Return
End Function
```

Приложение 5. Список ПЭ-пар, полученных из корпуса «ПарРус» в автоматическом режиме

5.1. Правильные ПЭ-пары

Русск. слова	**Финск. слова**	**KUC**
абажур	lampun-varjostin	0,57
август	elokuu	0,66
автобус	bussi	0,82
автомат	automaatti	0,56
автомат	konepistooli	0,75
автомобиль	auto	0,61
автор	tekijä	0,56
авторитет	auktoriteetti	0,79
агент	agentti	0,86
адвокат	asianajaja	0,89
администратор	apulais-johtaja	0,74
адрес	osoite	0,82
адский	helvetillinen	0,75
адъютант	adjutantti	0,91
академия	akatemia	0,83
актер	näyttelijä	0,56
акцент	korostus	0,60
аллея	puisto-käytävä	0,60
альбом	albumi	0,76
амбар	aitta	0,71
американец	amerikkalai-nen	0,72
американ-ский	amerikkalai-nen	0,76
анализ	analyysi	0,70
ангел	enkeli	0,86
английский	englantilai-nen	0,57
англичанин	englantilai-nen	0,65
анекдот	kasku	0,56
антилопа	antilooppi	0,97
антракт	väliaika	0,66
апостол	apostoli	0,90
аппетит	ruokahalu	0,79
апрель	huhtikuu	0,81
аптека	apteekki	0,80
арбуз	arbuusi	0,86

Русск. слова	**Финск. слова**	**KUC**
армия	armeija	0,62
артиллерия	tykistö	0,61
архив	arkisto	0,61
асфальт	asvaltti	0,55
асфальт	asfaltti	0,80
атмосфера	ilmakehä	0,56
аудитория	kuulijakunta	0,63
афиша	juliste	0,57
аэродром	lentokenttä	0,74
бабочка	perhonen	0,77
базар	basaari	0,68
бакенбарда	poskiparta	0,72
бал	tanssiaiset	0,84
балкон	parveke	0,86
бальный	tanssisali	0,57
банк	pankki	0,66
банка	tölkki	0,63
баня	sauna	0,95
барабан	rumpu	0,83
баран	pässi	0,56
барон	paroni	0,92
бархатный	samettinen	0,61
барыня	rouva	0,56
бас	bassoääni	0,58
бас	basso	0,60
бассейн	allas	0,57
бассейн	uima-allas	0,56
батарея	patteri	0,87
батон	polakka	0,65
батон	patonki	0,71
батька	pappa	0,65
бахрома	ripsu	0,61
башня	torni	0,64
бегун	hölkkääjä	0,76
бедность	köyhyys	0,73
беззащит-ный	turvaton	0,55
беззубый	hampaaton	0,90
безнадеж-ный	toivoton	0,61
бекас	taivaanvuohi	0,79

Русск. слова	Финск. слова	KUC
белизна	valkoisuus	0,59
белобрысый	hailakka	0,58
белье	alusvaate	0,60
бензин	bensiini	0,84
береза	koivu	0,80
березка	koivu	0,57
берет	baskeri	0,85
бесконеч-ность	rajattomuus	0,56
бесконеч-ность	loputtomuus	0,69
бесконеч-ный	loputon	0,56
беспечный	huoleton	0,56
бесплатный	maksuton	0,59
бесполезный	hyödytön	0,58
беспомощ-ный	avuton	0,78
бессилие	voimatto-muus	0,77
бесследный	jäljetön	0,57
бессмерт-ный	kuolematon	0,93
бессозна-тельно	tiedottomasti	0,69
бессонница	unettomuus	0,73
бессонный	uneton	0,67
бесчувст-венный	tunteeton	0,56
бешеный	vesikauhui-nen	0,56
библиотека	kirjasto	0,84
библия	Raamattu	0,83
бинокль	kiikari	0,94
биологичес-кий	biologinen	0,58
биология	biologia	0,77
благодар-ность	kiitollisuus	0,62
благодар-ный	kiitollinen	0,67
благодетель	hyväntekijä	0,66
благоразу-мие	järkevyys	0,56
благородие	jalosukuisuus	0,64
благород-ный	jalosukuinen	0,56
благослове–ние	siunaus	0,67
благосло-вить	siunata	0,58
блаженство	autuus	0,62
бледность	kalpeus	0,66
близость	läheisyys	0,60

Русск. слова	Финск. слова	KUC
блин	blini	0,60
блок	blokki	0,68
блок	väkipyörä	0,57
блондинка	vaaleave-rikkö	0,89
богатство	rikkaus	0,76
болото	suo	0,59
больница	sairaala	0,74
большевик	bolševikki	0,70
большин-ство	enemmistö	0,58
бомба	pommi	0,65
бомбежка	pommitus	0,81
бомбить	pommittaa	0,59
бор	havumetsä	0,57
борода	parta	0,56
борьба	taistelu	0,55
ботва	naatti	0,79
бочка	tynnyri	0,65
браво	bravo	0,77
брак	avioliitto	0,66
браслет	rannerengas	0,83
бред	houre	0,56
бритва	partaveitsi	0,76
бровь	kulmakarva	0,75
бродяга	kulkuri	0,78
бронзовый	pronssinväri-nen	0,58
брызги	pärske	0,65
буква	kirjain	0,64
буквально	kirjaimelli-sesti	0,77
букет	kukkavihko	0,57
булавка	hakaneula	0,58
булавка	nuppineula	0,57
бульвар	bulevardi	0,82
бульвар	puistokatu	0,56
бумажник	lompakko	0,93
бунтовать	kapinoida	0,59
буржуазный	porvarillinen	0,84
буркнуть	murahtaa	0,56
бутерброд	voileipä	0,73
буфетчик	kahvilanpitä-jä	0,80
бухгалтер	kirjanpitäjä	0,87
бухгалтерия	kirjanpito	0,59
бушлат	pomppa	0,69
бык	härkä	0,59
бычок	mulli	0,55
бюстик	rintakuva	0,63
ваза	maljakko	0,75
вал	valli	0,63
валенок	huovikas	0,81

Русск. слова	Финск. слова	KUC
вальдшнеп	lehtokurppa	0,89
вальс	valssi	0,88
валюта	valuutta	0,77
валюта	ulkomaan-valuutta	0,61
варенье	hillo	0,70
варьете	varietee-teatteri	0,76
вахта	vartiotupa	0,72
вдова	leski	0,62
ведро	sanko	0,58
веер	viuhka	0,61
вежливый	kohteliaasti	0,64
веко	luomi	0,58
веко	silmäluomi	0,56
вексель	vekseli	0,92
великоду-шие	jalomielisyys	0,78
велосипед	polkupyörä	0,81
веник	vihta	0,77
венок	seppele	0,72
веранда	veranta	0,64
верблюд	kameli	0,91
верность	uskollisuus	0,61
вероятность	todennäköi-syys	0,60
верста	virsta	0,89
вертикаль-ный	pystysuora	0,71
весенний	keväinen	0,62
весло	airo	0,81
весна	kevät	0,76
весной	kevät	0,57
вечность	ikuisuus	0,76
взвешивать	punnita	0,68
взводный	joukkueen-johtaja	0,73
вздох	huokaus	0,58
вздыхать	huokailla	0,69
взорваться	räjähtää	0,57
взрослый	aikuinen	0,74
вилка	haarukka	0,94
вилы	hanko	0,56
вилы	talikko	0,65
виноград	viinirypäle	0,73
винт	ruuvi	0,59
винтовка	kivääri	0,66
виселица	hirsipuu	0,77
висок	ohimo	0,88
витрина	näyteikkuna	0,63
вице-король	varakuningas	0,91
вкус	maku	0,63
вкусный	maukas	0,62

Русск. слова	Финск. слова	KUC
влево	vasemmalle	0,58
влюбиться	rakastua	0,60
влюбленный	rakastua	0,70
внешний	ulkoinen	0,60
внук	pojanpoika	0,63
вовремя	ajoissa	0,57
во-вторых	toiseksi	0,71
водопад	vesiputous	0,79
возражение	vastaväite	0,55
волк	susi	0,85
волна	aalto	0,68
волосатый	karvainen	0,57
волчица	naarassusi	0,78
вообража-ние	mielikuvitus	0,58
вооружен-ный	aseellinen	0,62
во-первых	ensinnä	0,67
вопроситель-но	kysyvästi	0,73
вор	varas	0,73
воробей	varpunen	0,70
воротник	kaulus	0,65
восемнад-цать	kahdeksantoista	0,74
восемьдесят	kahdeksan-kymmentä	0,62
восемьсот	kahdeksan-sataa	0,70
восклицание	huudahdus	0,58
воскресенье	sunnuntai	0,78
воспален-ный	tulehtua	0,62
воспитание	kasvatus	0,57
воспитать	kasvattaa	0,57
восток	itä	0,83
восточный	itämainen	0,59
восточный	itäinen	0,65
вошь	täi	0,78
вполголоса	puoliääneen	0,82
всадник	ratsastaja	0,55
вселенная	maailman–kaikkeus	0,62
вслух	ääneen	0,66
всхлипнуть	nyyhkäistä	0,64
вторник	tiistai	0,81
вторник	tiistaisin	0,58
втроем	kolmisin	0,65
выбор	valinta	0,57
выбрать	valita	0,58
вывеска	nimikilpi	0,61
выгода	etu	0,62
вызов	haaste	0,56

Русск. слова	Финск. слова	KUC
выиграть	voittaa	0,56
выразитель–ный	ilmeikäs	0,73
высочество	korkeus	0,59
выставка	näyttely	0,79
выстрел	laukaus	0,81
выходной	vapaapäivä	0,59
вышка	vartiotorni	0,61
газ	kaasu	0,68
гайка	mutteri	0,90
галерея	galleria	0,60
галстук	solmio	0,63
гармония	harmonia	0,60
гармонь	huuliharppu	0,55
гвоздь	naula	0,56
гениальный	nerokas	0,80
гений	nero	0,59
германия	saksa	0,73
герой	sankari	0,85
герцог	herttua	0,92
гибкий	notkea	0,57
гибрид	hybridi	0,98
гигант	jättiläinen	0,65
гигантский	jättimäinen	0,58
гимназия	lukio	0,56
гимназия	kymnaasi	0,64
гиря	punnus	0,82
гитара	kitara	0,95
гладкий	sileä	0,55
глоток	kulaus	0,74
глупость	tyhmyys	0,71
голенище	saappaan-varsi	0,56
голод	nälkä	0,69
голодный	nälkäinen	0,73
голубь	kyyhkynen	0,68
гордиться	ylpeillä	0,57
гордо	ylpeästi	0,56
гордость	ylpeys	0,84
гордый	ylpeä	0,66
горло	kurkku	0,58
горница	kamari	0,56
горничная	sisäkkö	0,84
горох	herne	0,56
горшок	ruukku	0,70
горючее	polttoaine	0,59
госпиталь	sairaala	0,60
гостиница	hotelli	0,72
государство	valtio	0,70
государыня	keisarinna	0,87
готовность	valmius	0,57
грабли	harava	0,95
градус	aste	0,59

Русск. слова	Финск. слова	KUC
грамм	gramma	0,81
гранитный	graniittinen	0,57
графин	karahvi	0,83
грек	kreikkalainen	0,75
грести	soutaa	0,62
греться	lämmitellä	0,58
грех	synti	0,72
грешный	syntinen	0,64
гриб	sieni	0,59
грива	harja	0,57
гроб	ruumisarkku	0,56
гроб	arkku	0,58
гробовщик	ruumisarkun-tekijä	0,76
гроза	ukonilma	0,67
гроссмейс-тер	suurmestari	0,87
грубость	karkeus	0,61
грузовик	kuorma-auto	0,73
грузчик	lastaaja	0,59
грузчик	kuormaaja	0,57
груша	päärynä	0,57
губернатор	kuvernööri	0,94
губерния	kuverne-mentti	0,66
гумно	puimala	0,55
гусар	husaari	0,94
гусь	hanhi	0,79
давление	verenpaine	0,64
датчанин	tanskalainen	0,92
дача	huvila	0,79
двенадцать	kaksitoista	0,73
двести	kaksisataa	0,68
двойной	kaksinkertai-nen	0,63
дворец	palatsi	0,77
дворник	talonmies	0,90
дворянство	aatelisto	0,66
двухэтаж-ный	kaksikerroksi nen	0,96
девяносто	yhdeksän-kymmentä	0,67
девятнад-цать	yhdeksän-toista	0,81
девять	yhdeksän	0,77
девятьсот	yhdeksänsata a	0,62
дедушка	vaari	0,56
дежурить	päivystää	0,63
дезертир	rintama-karkuri	0,59
действитель ность	todellisuus	0,61

Русск. слова	Финск. слова	KUC
декабрь	joulukuu	0,85
департамент	virasto	0,59
десятилетие	vuosikymmen	0,88
детдом	lastenkoti	0,69
детина	koljatti	0,94
детство	lapsuus	0,71
джентльмен	gentlemanni	0,56
дивизия	divisioona	0,76
дикобраз	piikkisika	1,00
диктовать	sanella	0,61
диплом	diplomi	0,67
дипломат	diplomaatti	0,91
диссертация	väitöskirja	0,78
дневальный	päivystäjä	0,63
дневник	päiväkirja	0,56
дно	pohja	0,58
доброволец	vapaaehtoinen	0,79
добровольный	vapaaehtoisesti	0,62
добродетель	hyve	0,72
доверие	luottamus	0,79
доверчивый	luottavainen	0,58
договор	sopimus	0,62
доказательство	todiste	0,59
доказать	todistaa	0,56
долина	laakso	0,87
доллар	dollari	0,93
домработница	kotiapulainen	0,88
донос	ilmianto	0,80
допрос	kuulustelu	0,76
дощечка	laudanpätkä	0,56
драка	tappelu	0,73
дракон	lohikäärme	0,81
драма	draama	0,67
драться	tapella	0,62
дремать	torkkua	0,69
дружба	ystävyys	0,78
дряхлый	raihnainen	0,57
дубовый	tamminen	0,57
дуэль	kaksintaistelu	0,81
дьячок	lukkari	0,67
евангелие	evankeliumi	0,76
еврей	juutalainen	0,71
европа	eurooppa	0,70
европейский	eurooppalainen	0,71
елка	kuusijuhla	0,57
елка	joulukuusi	0,55

Русск. слова	Финск. слова	KUC
желчь	sappi	0,75
женатый	naimisissa	0,65
жених	sulhanen	0,83
жертва	uhri	0,74
жертвовать	uhrata	0,59
жест	ele	0,59
жестокость	julmuus	0,77
живопись	maalaustaide	0,65
животное	eläin	0,56
жид	juutalainen	0,58
жидкость	neste	0,69
жилет	liivi	0,58
жилистый	suonikas	0,65
жребий	arpa	0,64
жук	turilas	0,62
журнал	aikakauslehti	0,59
журналист	lehtimies	0,71
забор	aita	0,55
завидовать	kadehtia	0,74
зависимость	riippuvaisuus	0,58
зависть	kateus	0,56
завод	tehdas	0,60
завтрак	aamiainen	0,61
загадка	arvoitus	0,71
загадочный	arvoituksellinen	0,75
загар	rusketus	0,58
заговор	salaliitto	0,71
заговорщик	salaliittolainen	0,78
заикаться	änkyttää	0,63
зайка	pupu	0,67
зайка	pupujussi	0,86
заказ	tilaus	0,57
закат	auringonlasku	0,60
законный	laillinen	0,68
закоптить	nokeentua	0,63
замедлить	hidastaa	0,57
заместитель	varapuheenjohtaja	0,62
замечание	huomautus	0,65
замок	lukko	0,73
замужем	naimisissa	0,56
занавес	esirippu	0,73
заначка	jemma	0,72
запад	länsi	0,75
западный	läntinen	0,61
запрячь	valjastaa	0,56
заработок	ansiotyö	0,55
зарядить	ladata	0,84
заседатель	valamies	0,71
затылок	takaraivo	0,69

Русск. слова	Финск. слова	KUC
заяц	jänis	0,76
звездный	tähtikirkas	0,56
зверек	pikkuotus	0,58
зверь	peto	0,59
звучный	soinnikas	0,56
здешний	täkäläinen	0,60
здоровье	terveys	0,73
зевать	haukotella	0,58
зевнуть	haukotella	0,75
зелень	vihreys	0,60
землемер	maanmittari	1,00
землянка	korsu	0,79
земной	maallinen	0,57
земство	kunnanval-tuusto	0,59
зеркало	peili	0,89
зимний	talvinen	0,56
зимой	talvi	0,56
зловещий	pahaenteinen	0,59
змея	käärme	0,59
знакомство	tuttavuus	0,58
знаменитый	kuuluisa	0,75
значение	merkitys	0,65
зной	helle	0,59
зонтик	päivänvarjo	0,75
зонтик	sateenvarjo	0,60
зритель	katsoja	0,71
зэк	vanki	0,65
зять	vävy	0,71
игемон	hegemoni	0,59
иголка	neula	0,58
игрок	peluri	0,63
игрушка	lelu	0,58
игрушка	leikkikalu	0,58
идеал	ihanne	0,75
идеальный	ideaalinen	0,57
идеальный	ihanteelli-sesti	0,57
идеальный	ihanteellinen	0,60
идеологичес кий	ideologinen	0,88
идеология	ideologia	0,75
идиот	idiootti	0,81
извозчик	ajuri	0,71
изголовье	pääpuoli	0,71
изнутри	sisältäpäin	0,57
икать	nikotella	0,80
икона	ikoni	0,78
икра	kaviaari	0,64
имение	maatila	0,58
император	keisari	0,62
инвалид	invalidi	0,83
иней	huurre	0,56

Русск. слова	Финск. слова	KUC
инерция	jähmeys	0,59
инженер	insinööri	0,91
иностранец	ulkomaalai-nen	0,77
инстинкт	vaisto	0,59
институт	instituutti	0,74
инструмент	työkalu	0,61
интеллиген-ция	sivistyneistö	0,61
интимный	intiimisti	0,55
интонация	intonaatio	0,56
информация	informaatio	0,57
иронически	ironisesti	0,56
исключение	poikkeus	0,57
искренность	vilpittömyys	0,65
искусство	taide	0,70
испанский	espanjalainen	0,61
исповедь	synnintunnus tus	0,58
испытание	koettelemus	0,67
истина	totuus	0,61
историк	historioitsija	0,61
историчес-кий	historiallinen	0,82
италия	italia	0,84
итальянский	italialainen	0,66
июль	heinäkuu	0,84
июнь	kesäkuu	0,93
кабак	kapakka	0,68
кабан	villisika	0,79
кавалер	kavaljeeri	0,77
кадриль	katrilli	1,00
казак	kasakka	0,95
казарма	kasarmi	0,85
календарь	kalenteri	0,69
календарь	almanakka	0,57
калоша	kalossi	0,73
кальсоны	alushousut	0,64
каменка	kiuas	0,63
каменщик	muurari	0,88
камера	selli	0,60
камердинер	kamaripalve-lija	0,87
камыш	kaislikko	0,63
камыш	kaisla	0,69
канал	kanava	0,58
кандидат	kandidaatti	0,89
канцелярия	kanslia	0,75
капитал	pääoma	0,73
капитан	kapteeni	0,71
капрал	korpraali	0,96
капуста	kaali	0,76
карандаш	lyijykynä	0,68

Русск. слова	Финск. слова	KUC
каска	kypärä	0,78
касса	kassa	0,60
катастрофа	katastrofi	0,73
категория	kategoria	0,61
каторга	pakkotyö	0,68
катушка	kela	0,59
кафе	kahvila	0,82
кафедра	kateederi	0,55
кафедра	oppituoli	0,58
каша	puuro	0,75
кашель	yskä	0,64
каюта	hytti	0,65
квартальный	katsastaja	0,56
квитанция	kuitti	0,65
кепка	lippalakki	0,56
керосин	petroli	0,69
кивать	nyökkäillä	0,55
кивнуть	nyökätä	0,83
килограмм	kilo	0,63
километр	kilometri	0,93
килька	kilohaili	0,87
кинжал	tikari	0,89
кино	elokuva	0,65
кирка	hakku	0,69
кирпич	tiili	0,56
кисель	kiisseli	0,88
кисет	massi	0,74
кисет	tupakkamassi	0,68
китайский	kiinalainen	0,67
кишка	suoli	0,57
клад	aarre	0,66
кладбище	hautausmaa	0,89
кладка	muuraus	0,57
класс	luokka	0,63
классичес-кий	klassinen	0,75
классичес-кий	klassillinen	0,66
клевета	parjaus	0,63
клетчатый	ruudullinen	0,84
клещ	punkki	0,61
клиника	klinikka	0,86
клоп	lude	0,58
клоп	lutikka	0,66
клуб	klubi	0,59
клубень	mukula	0,74
клясться	vannoa	0,63
кнут	ruoska	0,55
княжна	ruhtinatar	0,66
кобыла	tamma	0,68
ковер	matto	0,70
кодекс	koodeksi	0,59
кодекс	lakikirja	0,57

Русск. слова	Финск. слова	KUC
коза	vuohi	0,71
колбаса	makkara	0,79
колесо	pyörä	0,70
коллега	kollega	0,83
коллектив	kollektiivi	0,68
колода	korttipakka	0,63
колодец	kaivo	0,87
колокольня	kellotorni	0,60
колонист	siirtolalainen	0,64
колония	siirtola	0,80
колос	tähkä	0,57
колхоз	kolhoosi	0,70
команди-ровка	komennus	0,59
комедия	komedia	0,69
комендант	komendantti	0,59
комендант	komentaja	0,60
комендант-ша	komentajan-rouva	0,83
комик	koomikko	0,75
комиссар	komissaari	0,88
комиссия	komissio	0,68
коммуна	kommuuni	0,90
коммунизм	kommunismi	0,85
коммунист	kommunisti	0,93
коммунисти-ческий	kommunis-tinen	0,57
комод	lipasto	0,77
комплимент	kohteliaisuus	0,67
конверт	kirjekuori	0,64
конгресс	kongressi	0,85
кондитер-ская	konditoria	0,58
кондуктор	konduktööri	0,80
конек	luistin	0,57
конкретный	konkreetti-nen	0,74
консервы	säilyke	0,60
консультант	ekspertti	0,56
контора	konttori	0,63
контрабас	bassoviulu	0,72
контрабас	kontrabasso	0,65
конус	kartio	0,83
конферансье	kuuluttaja	0,71
концерт	konsertti	0,81
концессио-нер	liiketoveri	0,64
концлагерь	keskitysleiri	0,87
коньяк	konjakki	0,91
конюшня	talli	0,84
копейка	kopeekka	0,79
копия	kopio	0,68
копыто	kavio	0,76

Русск. слова	Финск. слова	KUC
копье	keihäs	0,74
корабль	alus	0,58
коричневый	ruskea	0,67
кормилица	imettäjä	0,93
корнет	kornetti	0,92
королева	kuningatar	0,77
король	kuningas	0,83
корреспон-дент	kirjeenvaihta ja	0,86
корточки	kyykkysil-leen	0,57
корыто	soikko	0,59
коса	viikate	0,62
косвенный	välillisesti	0,62
косвенный	välillinen	0,62
костер	nuotio	0,84
костлявый	luiseva	0,65
костыль	kainalosauva	0,90
котелок	knalli	0,67
котлета	kotletti	0,75
котлован	monttu	0,93
котомка	reppu	0,55
кофе	kahvi	0,72
кофта	neuletakki	0,55
кочка	mätäs	0,58
кошелек	kukkaro	0,74
кошка	kissa	0,57
кошмар	painajainen	0,69
крайность	äärimmäi-syys	0,56
кран	vesihana	0,56
красавица	kaunotar	0,64
красота	kauneus	0,80
кредит	luotto	0,59
крем	voide	0,55
крепость	linnoitus	0,84
кролик	kaniini	0,64
кружево	pitsi	0,69
крутой	jyrkkä	0,56
крыло	siipi	0,74
крыса	rotta	0,79
крышка	kansi	0,58
ксендз	pappi	0,55
кубло	kopla	0,75
кузнец	seppä	0,68
кукла	nukke	0,58
кукуруза	maissi	0,63
кулиса	kulissi	0,68
культура	kulttuuri	0,58
купе	vaunun osasto	0,70
купец	kauppias	0,85
купол	kupoli	0,74
курица	kana	0,85

Русск. слова	Финск. слова	KUC
курносый	pystynenäi-nen	0,57
курносый	nykerönenä	0,56
курносый	kippurane-näinen	0,55
курс	kurssi	0,62
кухарка	keittäjätär	0,66
кучер	kuski	0,71
лаборатория	laboratorio	0,75
лакей	lakeija	0,85
лампа	lamppu	0,67
лапоть	virsu	0,65
ласка	hyväily	0,57
латыш	latvialainen	0,94
лгать	valehdella	0,62
лебедь	joutsen	0,81
лев	leijona	0,65
легенда	legenda	0,67
легенда	peitetarina	0,59
легкомыс-ленный	kevytmieli-nen	0,74
легкость	keveys	0,62
лейтенант	luutnantti	0,75
лекарство	lääke	0,87
лекция	luento	0,69
лениво	laiskasti	0,64
лето	kesä	0,68
летом	kesä	0,58
летчик	lentäjä	0,65
либераль-ный	liberaalinen	0,67
ливень	kaatosade	0,67
лизнуть	nuolaista	0,71
лисица	kettu	0,69
литература	kirjallisuus	0,89
литерный	erikoisjuna	0,57
лифт	hissi	0,87
лично	henkilökoh-taisesti	0,62
логика	logiikka	0,60
лодка	vene	0,93
ложка	lusikka	0,77
ложь	valhe	0,74
лозунг	iskulause	0,56
локоть	kyynärpää	0,83
лом	kanki	0,68
лопата	lapio	0,76
лопатка	lapaluu	0,78
лопух	takiainen	0,65
лохматка	takku	0,98
луг	niitty	0,66
лужа	lätäkkö	0,57
лужайка	nurmikko	0,62

Русск. слова	Финск. слова	KUC
лук	sipuli	0,72
луна	kuu	0,63
луч	säde	0,61
лыжа	suksi	0,73
лыжник	hiihtäjä	0,80
лыжня	latu	0,81
лысина	kalju	0,61
любоваться	ihailla	0,59
любовник	rakastaja	0,90
любовница	rakastajatar	0,62
любопыт-ство	uteliaisuus	0,71
люлька	kätkyt	0,64
люстра	kattokruunu	0,83
лягушка	sammakko	0,91
маг	maagi	0,64
маг	taikuri	0,58
магистраль	päärata	0,89
мадам	madame	0,75
мазурка	masurkka	0,97
май	toukokuu	0,60
майор	majuri	0,82
майский	toukokuu	0,65
малиновый	vadelman-punainen	0,60
мансарда	ullakkoka-mari	0,64
мансарда	ullakkohuone	0,57
март	maaliskuu	0,89
маска	naamari	0,55
маска	naamio	0,69
материал	aineisto	0,69
материаль-ный	aineellinen	0,61
материнский	äidillinen	0,56
материя	materia	0,57
матрас	patja	0,67
матрац	patja	0,57
матрос	matruusi	0,68
мачта	masto	0,82
машиналь-ный	koneellisesti	0,66
машинист	veturinkuljet-taja	0,77
машинка	kirjoituskone	0,64
мебель	huonekalu	0,64
мед	hunaja	0,82
медаль	mitali	0,80
медведь	karhu	0,63
медицина	lääketiede	0,81
медицин-ский	lääketieteel-linen	0,58
медный	kuparinen	0,55

Русск. слова	Финск. слова	KUC
междуна-родный	kansainväli-nen	0,85
мел	liitu	0,67
мельница	mylly	0,82
мерин	valakka	0,57
местный	paikallinen	0,61
металл	metalli	0,67
метр	metri	0,75
метро	metro	0,80
механизм	mekanismi	0,66
меч	miekka	0,59
мечта	haave	0,59
мечтать	haaveilla	0,71
микроскоп	mikroskoop-pi	0,98
микротом	mikrotomi	0,98
милиционер	miliisi	0,59
милиция	miliisi	0,60
миллион	miljoona	0,90
миллионер	miljoonamies	0,71
мина	miina	0,55
министер-ство	ministeriö	0,74
министр	ministeri	0,84
мираж	kangastus	0,70
мисс	miss	0,83
митинг	kansanko-kous	0,57
могила	hauta	0,63
мода	muoti	0,58
мозг	aivot	0,66
молдаванин	moldavialai-nen	0,78
молитва	rukous	0,76
молиться	rukoilla	0,67
молния	salama	0,68
молодежь	nuoriso	0,71
молодость	nuoruus	0,64
молоко	maito	0,84
молоток	vasara	0,65
монастырь	luostari	0,83
монах	munkki	0,73
монтер	monttööri	0,72
монтер	asentaja	0,65
мороженое	jäätelö	0,82
мост	silta	0,66
мостовая	ajotie	0,63
мостовая	katukiveys	0,61
мотор	moottori	0,73
мотоцикл	moottori-pyörä	0,85
мох	sammal	0,64
мрачный	synkkä	0,58

Русск. слова	Финск. слова	KUC
мстить	kostaa	0,64
мужествен-ный	miehekäs	0,70
мужичок	ukkeli	0,70
музей	museo	0,79
музыка	musiikki	0,83
музыкант	muusikko	0,64
мускул	lihas	0,65
муха	kärpänen	0,80
мыло	saippua	0,68
мыслитель	ajattelija	0,79
мышь	hiiri	0,83
мясо	liha	0,70
мяч	pallo	0,70
набережная	rantakatu	0,94
наблюдатель	tarkkailija	0,60
навес	katos	0,58
навоз	lanta	0,56
нагнуться	kumartua	0,56
надзиратель	valvoja	0,60
наивный	naiivi	0,61
наизусть	ulkoa	0,62
наказание	rangaistus	0,59
налево	vasemmalle	0,56
налим	matikka	0,70
налим	made	0,65
намек	vihjaus	0,57
наоборот	päinvastoin	0,57
направо	oikealle	0,58
нарезать	viipaloida	0,56
нарушитель	rikkoja	0,57
нары	laveri	0,58
насилие	väkivalta	0,88
наслаждение	nautinto	0,62
наследник	perillinen	0,87
наследство	perintö	0,80
настежь	selkoselälään	0,56
насыпь	penkka	0,57
научный	tieteellinen	0,73
националь-ный	kansallinen	0,71
нация	kansakunta	0,63
начальство	päällystö	0,63
начисто	kauhallinen	0,56
начкар	vartiopäällikkö	0,62
невероятный	uskomaton	0,61
невеста	morsian	0,90
невестка	miniä	0,65
невидимый	näkymätön	0,77
невинность	viattomuus	0,60
невинный	viaton	0,64

Русск. слова	Финск. слова	KUC
невольно	tahtomattaan	0,59
невыноси-мый	sietämätön	0,64
негр	neekeri	0,85
недоволь-ный	tyytymätön	0,70
недоволь-ство	tyytymät-tömyys	0,73
недоразуме-ние	väärinkäsitys	0,83
недостой-ный	arvoton	0,58
нежность	hellyys	0,66
независи-мость	riippumatto-muus	0,73
незнакомец	tuntematon	0,57
неизвестный	tuntematon	0,59
неистово	vimmaisesti	0,55
немой	mykkä	0,67
ненужный	tarpeeton	0,56
необходи-мость	välttämättömyys	0,56
необъясни-мый	selittämätön	0,71
неожидан-ный	odottamaton	0,55
непобеди-мый	voittamaton	0,67
неподвиж–ность	liikkumatto-muus	0,71
непонятный	käsittämätön	0,64
неприят-ность	ikävyys	0,61
нерв	hermo	0,74
нерешитель-ность	päättämättö-myys	0,56
неслышно	kuulumatto-masti	0,56
несправедли-вость	epäoikeuden-mukaisuus	0,56
несправедли-вый	epäoikeuden-mukaisesti	0,57
несправедли-вый	epäoikeu-denmukainen	0,61
несчастье	onnettomuus	0,60
нетерпение	kärsimät-tömyys	0,60
неуверен-ный	epävarmasti	0,55
ничтожество	mitättömyys	0,65
ниша	seinäsyven-nys	0,58
нищий	kerjäläinen	0,63
новость	uutinen	0,60

Русск. слова	Финск. слова	KUC
ноготь	kynsi	0,59
нож	veitsi	0,76
ножницы	sakset	0,61
ножовка	sahanterä	0,64
ноздря	sierain	0,83
ноль	nolla	0,71
норма	normi	0,59
нормальный	normaali	0,74
носилки	paarit	0,79
ноябрь	marraskuu	0,93
нырять	sukellella	0,56
обвинение	syytös	0,72
обезьяна	apina	0,72
обещание	lupaus	0,68
обидеться	loukkaantua	0,58
облава	ratsia	0,65
облако	pilvi	0,60
облегчение	helpotus	0,58
облигация	obligaatio	0,80
обман	petos	0,68
обморок	pyörtyä	0,67
обожать	jumaloida	0,66
обоз	kuormasto	0,76
обои	tapetti	0,59
обои	seinäpaperi	0,62
обрадовать-ся	ilahtua	0,56
образование	sivistys	0,57
образо-ваться	lutviutua	0,58
обувь	jalkine	0,59
общежитие	asuntola	0,63
объект	työkohde	0,55
обыск	kotietsintä	0,66
обыск	kotitarkastus	0,61
овальный	soikea	0,78
овес	kaura	0,87
овощ	vihannekset	0,55
овраг	rotko	0,70
овца	lammas	0,78
оглобля	aisa	0,68
ого	oho	0,58
огород	vihannesmaa	0,58
огород	perunamaa	0,56
одеяло	peite	0,64
одиннадцать	yksitoista	0,68
одинокий	yksinäinen	0,75
одиночество	yksinäisyys	0,69
одноглазый	yksisilmäi-nen	0,82
однообраз-ный	yksitoikkoi-nen	0,71
одобрение	hyväksyntä	0,57
одобритель-ный	hyväksyvästi	0,67
ожидание	odotus	0,57
озеро	järvi	0,74
озимый	syysvehnä	0,57
озимый	syysvilja	0,59
океан	valtameri	0,83
окончатель-но	lopullisesti	0,61
окоп	potero	0,64
октябрь	lokakuu	0,83
опасность	vaara	0,69
опасный	vaarallinen	0,68
опера	ooppera	0,59
операция	operaatio	0,67
опоздать	myöhästyä	0,67
опушка	aho	0,57
опыт	kokemus	0,56
оранжевый	oranssi	0,58
оранжевый	oranssinväri-nen	0,57
оранжерея	kasvihuone	0,91
оратор	puhuja	0,73
организация	organisaatio	0,59
организм	organismi	0,67
организм	elimistö	0,71
орден	kunnia-merkki	0,79
орел	kotka	0,75
орех	pähkinä	0,76
ореховый	pähkinäpui-nen	0,57
оркестр	orkesteri	0,78
оружие	ase	0,59
осел	aasi	0,61
осень	syksy	0,75
осенью	syksy	0,67
осколок	sirpale	0,72
осторож-ность	varovaisuus	0,66
осторожный	varovainen	0,58
остров	saari	0,87
ответствен-ность	vastuu	0,68
ответствен-ный	vastuullinen	0,64
отвыкнуть	vieraantua	0,57
отечество	isänmaa	0,70
отказаться	kieltäytyä	0,57
открове–ность	avomielisyys	0,59
открытка	postikortti	0,75

Русск. слова	Финск. слова	KUC
относитель-но	suhteellisesti	0,55
отомстить	kostaa	0,71
отпуск	loma	0,76
отравить	myrkyttää	0,79
отрывок	katkelma	0,65
отряд	osasto	0,55
отсюда	täältä	0,57
отчим	isäpuoli	0,89
оформление	muodostelu	0,57
охапка	sylyksellinen	0,56
охотник	metsästäjä	0,72
ошейник	panta	0,57
ошейник	kaulapanta	0,73
ошибаться	erehtyä	0,66
ошибиться	erehtyä	0,67
ошибка	virhe	0,58
ошибка	erehdys	0,57
пазуха	povi	0,71
палатка	teltta	0,77
палач	pyöveli	0,96
пальма	palmu	0,94
памятник	muisto-merkki	0,60
памятный	muistorikas	0,63
паника	pakokauhu	0,58
паника	paniikki	0,70
папаня	isi	0,85
папироса	paperossi	0,57
папка	mappi	0,64
папка	kansio	0,64
пар	höyry	0,62
парад	paraati	0,81
паразит	parasiitti	0,58
параллель-ный	yhden-suuntainen	0,55
парашютист	laskuvarjo–jääkäri	0,56
парк	puisto	0,65
паркет	parketti	0,71
паровоз	veturi	0,84
паром	lossi	0,72
паром	lautturi	0,57
партер	permanto	0,59
партизан	partisaani	0,87
партия	puolue	0,61
паспорт	passi	0,87
пассажир	matkustaja	0,78
пастор	pastori	0,95
пастух	paimen	0,79
патриарший	patriarkka	0,91
патрон	patruuna	0,67
паук	hämähäkki	0,83

Русск. слова	Финск. слова	KUC
паутина	hämähäkin-verkko	0,55
певуче	laulavasti	0,56
пейзаж	maisema	0,61
пена	vaahto	0,63
пенсия	eläke	0,73
первосвя-щенник	ylipappi	0,89
перебить	keskeyttää	0,56
перевал	laskusuunta	0,55
переводчик	kielenkääntä-jä	0,57
переводчик	tulkki	0,72
переговоры	neuvottelu	0,66
перегородка	väliseinä	0,63
перекресток	kadunristeys	0,56
перемена	muutos	0,56
переписка	kirjeenvaihto	0,82
переулок	kuja	0,60
переулок	sivukatu	0,58
перила	kaide	0,68
перстень	sormus	0,57
перчатка	hansikas	0,74
перчатка	käsine	0,61
пес	koira	0,55
песня	laulu	0,67
песок	hiekka	0,82
пестрый	kirjava	0,65
петух	kukko	0,91
пехота	jalkaväki	0,89
пехотинец	jalkaväen-sotilas	0,58
пешеход	jalankulkija	0,82
пешком	jalkaisin	0,70
пивная	olutkapakka	0,58
пиво	olut	0,85
пилить	sahata	0,58
пилотка	suikka	0,86
пир	pidot	0,58
писарь	kirjuri	0,66
писатель	kirjailija	0,81
пистолет	pistooli	0,87
письменный	kirjoitus-pöytä	0,78
письмоводи-тель	sihteeri	0,55
плакат	plakaatti	0,56
планета	planeetta	0,91
плашмя	lappeellaan	0,66
племя	heimo	0,76
племянник	veljenpoika	0,60
племянник	sisarenpoika	0,68
племянница	veljentytär	0,61

Русск. слова	Финск. слова	KUC
плод	hedelmä	0,56
плоский	litteä	0,56
плотина	pato	0,74
плотник	kirvesmies	0,85
площадка	tasanne	0,59
плюнуть	sylkäistä	0,58
победа	voitto	0,60
победитель	voittaja	0,74
побледнеть	kalveta	0,69
повар	kokki	0,70
поверенный	valantehnyt	0,63
повестка	kutsukirje	0,56
повинный	katuvainen	0,55
поводок	lieka	0,61
погаснуть	sammua	0,62
погладить	silittää	0,58
погулять	riiustella	0,56
подарить	lahjoittaa	0,65
подарок	lahja	0,68
подбородок	leuka	0,71
подвал	kellari	0,58
подвиг	uroteko	0,67
подвиг	urotyö	0,63
по-детски	lapsenomaisesti	0,60
подлость	kataluus	0,55
подножка	astinlauta	0,65
поднос	tarjotin	0,82
подоконник	ikkunalauta	0,77
подписать	allekirjoittaa	0,61
подпись	allekirjoitus	0,56
подробность	yksityiskohta	0,59
подсолнух	auringonkukka	0,77
подтверждение	vahvistus	0,64
подушка	tyyny	0,88
пожар	tulipalo	0,81
пожарный	palokuntalainen	0,58
пожертвовать	uhrata	0,61
пожилой	iäkäs	0,62
поза	asento	0,60
позвонить	soittaa	0,63
поздравить	onnitella	0,60
поздравлять	onnitella	0,72
познакомиться	tutustua	0,63
поить	juottaa	0,57
поклониться	kumartaa	0,69
поклонник	ihailija	0,62
поколение	sukupolvi	0,77

Русск. слова	Финск. слова	KUC
покраснеть	punastua	0,70
покупатель	ostaja	0,73
покупать	ostaa	0,58
покушение	attentaatti	0,61
покушение	murhayritys	0,63
полезный	hyödyllinen	0,57
политика	politiikka	0,68
политический	poliittinen	0,74
полиция	poliisi	0,58
полк	rykmentti	0,81
полковник	eversti	0,95
положительный	positiivinen	0,58
полоз	jalas	0,71
полосатый	raidallinen	0,66
полотенце	pyyhe	0,57
полотенце	pyyheliina	0,67
полтора	puolitoista	0,78
полтораста	puolitoistasataa	0,58
полушубок	puoliturkki	0,82
помещик	tilanomistaja	0,86
помидор	tomaatti	0,88
поминки	muistotilaisuus	0,55
помощник	apulainen	0,67
понедельник	maanantai	0,85
поп	pappi	0,62
попугай	papukaija	0,90
порог	kynnys	0,75
порода	rotu	0,56
поросенок	porsas	0,71
поросенок	possu	0,57
порох	ruuti	0,87
портной	räätäli	0,87
портрет	muotokuva	0,75
портфель	salkku	0,97
портянка	jalkarätti	0,89
порывистый	puuskainen	0,55
поскользнуться	liukastua	0,86
поскотина	laidunmaa	0,56
посланник	lähettiläs	0,74
послезавтра	ylihuomenna	0,87
пословица	sananlasku	0,60
постепенный	vähitellen	0,59
посуда	astia	0,62
посылка	paketti	0,68
пот	hiki	0,55
потный	hikinen	0,74
потолок	katto	0,55

Русск. слова	Финск. слова	KUC
потомок	jälkeläinen	0,75
по-французски	ranska	0,75
похмелье	krapula	0,63
поход	sotaretki	0,57
похороны	hautajaiset	0,83
похудеть	laihtua	0,66
поцеловать	suudella	0,67
поцелуй	suudelma	0,72
почерк	käsiala	0,80
почка	munuainen	0,60
почка	silmu	0,67
почта	posti	0,82
почтальон	postiljooni	0,78
почтовый	postilaatikko	0,57
пошехонец	hölmöläinen	0,84
пошлость	latteus	0,60
пощечина	korvapuusti	0,63
поэзия	runous	0,55
поэма	runoeepos	0,58
поэма	runoelma	0,70
поэт	runoilija	0,95
поэтический	runollinen	0,85
правительство	hallitus	0,73
православный	kreikkalais-katolinen	0,56
православный	oikea-uskoinen	0,78
праздность	joutilaisuus	0,76
практически	käytännöllisesti	0,65
прапорщик	vänrikki	0,86
превосходительство	ylhäisyys	0,69
преданность	uskollisuus	0,55
предательство	petturuus	0,71
предлог	tekosyy	0,59
предрассудок	ennakkoluulo	0,62
представитель	edustaja	0,63
преждевременный	ennenaikainen	0,65
президент	presidentti	0,88
президиум	puhemiehistö	0,93
президиум	presidiumi	0,55
презирать	halveksia	0,82
презрение	halveksunta	0,61
препарат	preparaatti	0,91
преподаватель	opettaja	0,57

Русск. слова	Финск. слова	KUC
препятствие	este	0,59
преступление	rikos	0,68
преступник	rikollinen	0,69
преувеличивать	liioitella	0,59
прививка	varrennus	0,68
прививка	varrennos	0,66
привидение	aave	0,60
привычка	tottumus	0,62
приданое	myötäjäiset	0,89
призвание	kutsumus	0,76
признание	tunnustus	0,59
приключение	seikkailu	0,75
прикосновение	kosketus	0,60
приличие	säädyllisyys	0,70
примус	priimus	0,79
принадлежность	kirjoitusväline	0,58
принц	prinssi	0,94
принцип	periaate	0,79
присутствие	läsnäolo	0,58
присяжный	valantehnyt	0,57
присяжный	valamies	0,76
притворство	teeskentely	0,65
притворяться	teeskennellä	0,60
причесанный	kammata	0,64
прическа	kampaus	0,76
пробирка	koeputki	0,97
пробка	korkki	0,62
пробка	tulppa	0,56
проблема	ongelma	0,56
проводник	junamies	0,64
программа	ohjelma	0,58
прогресс	edistys	0,76
продавщица	myyjätär	0,58
прозрачный	läpikuultava	0,65
проклясть	kirota	0,56
проклятие	kirous	0,63
прокурор	syyttäjä	0,61
пропуск	kulkulupa	0,58
прораб	työnjohtaja	0,64
пророк	profeetta	0,98
просвещение	valistus	0,67
проситель	anoja	0,58
проспект	prospekti	0,56
простота	yksinkertaisuus	0,62

Русск. слова	Финск. слова	KUC
простыня	lakana	0,84
просьба	pyyntö	0,60
протест	protesti	0,62
противник	vastustaja	0,71
противоре-чие	ristiriitaisuus	0,55
протокол	protokolla	0,61
протокол	pöytäkirja	0,62
профессия	ammatti	0,66
профиль	profiili	0,79
профиль	sivukuva	0,60
прохлада	viileys	0,55
прохожий	ohikulkija	0,72
процент	prosentti	0,61
процесс	prosessi	0,61
процессия	kulkue	0,75
прошлогод-ний	menneen-vuotinen	0,56
прошлогод-ний	viimevuoti-nen	0,62
прошлогод-ний	edellisvuoti-nen	0,58
прощение	anteeksianto	0,56
проявление	ilmenemis-muoto	0,56
пруд	lampi	0,73
пряжка	solki	0,60
пряник	piparkakku	0,71
психология	psykologia	0,55
публика	yleisö	0,67
пуд	puuta	0,58
пулемет	konekivääri	0,86
пульс	valtimo	0,80
пуля	luoti	0,59
пустота	tyhjyys	0,74
пустынный	autio	0,61
пустыня	erämaa	0,59
пушка	tykki	0,67
пчела	mehiläinen	0,75
пшеница	vehnä	0,73
пыльный	pölyinen	0,60
пыльца	siitepöly	1,00
пытка	kidutus	0,62
пьеса	näytelmä	0,85
пьяница	juoppo	0,59
пьяница	juopporatti	0,55
пьянство	juoppous	0,56
пятнадцать	viisitoista	0,77
пятница	perjantai	0,83
пятьсот	viisisataa	0,72
раб	orja	0,72
рабство	orjuus	0,71
равнина	tasanko	0,61

Русск. слова	Финск. слова	KUC
равновесие	tasapaino	0,80
равнодушие	välinpitämät-tömyys	0,66
радио	radio	0,90
развалина	raunio	0,58
разведка	tiedustelu-palvelu	0,68
разведчик	tiedustelija	0,84
развитие	kehitys	0,58
развод	avioero	0,66
разврат	irstaus	0,71
разврат	haureus	0,56
развратный	irstas	0,60
раздеваться	riisuutua	0,62
раздеться	riisuutua	0,61
размышле-ние	tutkiskelu	0,57
разноцвет-ный	erivärinen	0,62
рай	paratiisi	0,91
ракета	raketti	0,77
раковина	näkinkenkä	0,60
рана	haava	0,58
раненый	haavoittua	0,66
раскаяние	katumus	0,75
рассеян-ность	hajamieli-syys	0,60
рассеянный	hajamieli-sesti	0,60
рассказ	kertomus	0,61
расспраши-вать	kysellä	0,57
раствор	laasti	0,64
раствор	ruukki	0,62
растворный	konesali	0,57
растение	kasvi	0,61
ребро	kylkiluu	0,73
ревизия	revisio	0,55
ревниво	mustasukkai-sesti	0,56
ревность	mustasukkai-suus	0,91
револьвер	revolveri	0,93
революцион ный	vallankumou ksellinen	0,78
революция	vallanku-mous	0,92
редактор	toimittaja	0,78
редакция	toimitus	0,67
редко	harvoin	0,75
режиссер	ohjaaja	0,56
резина	kumi	0,61

Русск. слова	Финск. слова	KUC
резолюция	päätöslau-selma	0,58
резолюция	loppuponsi	0,56
результат	tulos	0,57
рейх	valtakunta	0,58
ректор	rehtori	0,87
религиоз-ный	uskonnol-linen	0,70
религия	uskonto	0,83
ремонт	remontti	0,64
репетиция	harjoitus	0,67
ресница	ripsi	0,57
ресница	silmäripsi	0,59
республика	tasavalta	0,87
ресторан	ravintola	0,67
рецепт	resepti	0,86
рецепт	valmistus-ohje	0,58
решимость	päättäväisyys	0,59
ржавый	ruosteinen	0,57
рис	riisi	0,87
робкий	arasti	0,57
робость	arkuus	0,56
ровный	tasainen	0,57
рог	sarvi	0,81
родитель	vanhempi	0,65
родственник	sukulainen	0,63
роза	ruusu	0,65
роковой	kohtalokas	0,83
роль	rooli	0,66
ром	rommi	0,96
роман	romaani	0,81
романс	romanssi	0,92
романтичес-кий	romanttinen	0,80
ропот	napina	0,58
ропот	nurina	0,58
роса	kaste	0,82
роскошь	ylellisyys	0,73
рота	komppania	0,69
рояль	flyygeli	0,68
рубаха	paita	0,56
ружье	haulikko	0,56
рукопись	käsikirjoitus	0,74
румяный	punakka	0,65
ручей	puro	0,80
рыба	kala	0,76
рыбак	kalastaja	0,78
рывок	kiskaisu	0,57
рыжеватый	punertava-tukkainen	0,58
рыжий	punapää	0,56
рысью	ravi	0,58

Русск. слова	Финск. слова	KUC
рыцарь	ritari	0,88
рюкзак	rinkka	0,84
рябина	pihlajan-marja	0,57
рябина	pihlaja	0,82
рябой	rokonarpinen	0,58
сабля	sapeli	0,66
сало	ihra	0,60
сало	läski	0,56
сало	silava	0,56
салфетка	ruokaliina	0,60
салфетка	lautasliina	0,61
самовар	samovaari	0,94
самозванец	valekeisari	0,70
самолет	lentokone	0,75
самолюбие	itserakkaus	0,77
самоубий-ство	itsemurha	0,81
сандалия	sandaali	0,75
сани	reki	0,78
санитар	lääkintämies	0,63
санчасть	sairastupa	0,56
сапожник	suutari	0,77
сарай	vaja	0,62
сахар	sokeri	0,69
свекла	rehujuurikas	0,57
свекла	juurikas	0,68
свекор	appi	0,68
сверху	ylhäältä	0,61
свеча	kynttilä	0,79
свечка	kynttilä	0,60
свидетель	todistaja	0,70
свинья	sika	0,69
святой	pyhä	0,55
святыня	pyhäkkö	0,57
священник	pappi	0,65
священный	pyhä	0,56
сдержанный	pidättyvästi	0,56
север	pohjoinen	0,82
седло	satula	0,80
сезон	sesonki	0,56
сезон	näytäntö-kausi	0,58
сейнер	troolari	0,55
сейнер	kalastuslaiva	0,61
секретарь	sihteeri	0,77
секундант	sekundantti	0,91
селедка	silli	0,75
семерка	seitsikko	0,89
семнадцать	seitsemän-toista	0,70
семьдесят	seitsemän-kymmentä	0,62

Русск. слова	Финск. слова	KUC
сено	heinä	0,65
сентябрь	syyskuu	0,91
сержант	kersantti	0,86
серп	sirppi	0,64
серьга	korvarengas	0,64
серьезно	vakavasti	0,61
сессия	tenttikausi	0,56
сеттер	setteri	0,91
сеть	verkko	0,57
сибирь	siperia	0,74
сигара	sikari	0,79
силуэт	siluetti	0,63
сирень	syreeni	0,61
сирота	orpo	0,69
сказка	satu	0,58
сказочный	tarunomainen	0,60
скала	kallio	0,66
скамейка	penkki	0,62
скамья	penkki	0,57
скандал	skandaali	0,81
скандальный	skandaalimainen	0,61
скатерть	pöytäliina	0,74
скелет	luuranko	0,80
складка	poimu	0,65
скользкий	liukas	0,60
скомандовать	komentaa	0,58
скорлупа	munankuori	0,64
скрепка	klemmari	0,76
скрипка	viulu	0,89
скромный	vaatimaton	0,60
скупой	saita	0,72
слабость	heikkous	0,66
славянский	slaavilainen	0,56
слепой	sokea	0,79
сливаться	sulautua	0,59
сливки	kerma	0,70
сложный	monimutkainen	0,56
слон	norsu	0,76
слуга	palvelija	0,62
слух	huhu	0,64
случайно	sattumalta	0,57
случайность	satunnaisuus	0,56
случайный	satunnainen	0,69
слушатель	kuulija	0,60
слушатель	kuuntelija	0,57
смело	rohkeasti	0,70
смелость	rohkeus	0,61
смелый	rohkea	0,57

Русск. слова	Финск. слова	KUC
смертельный	kuolettavasti	0,56
сметана	hapankerma	0,56
сметана	vuolukerma	0,75
смутный	hämärästi	0,57
снаряд	ammus	0,79
снизу	alhaalta	0,70
собеседник	keskustelukumppani	0,70
собрание	kokous	0,64
собственноручно	omakätisesti	0,77
совершенство	täydellisyys	0,74
советчик	neuvonantaja	0,60
современный	nykyaikainen	0,61
согласие	suostumus	0,55
сок	mehu	0,64
сокол	haukka	0,58
сокровище	aarre	0,65
соловей	satakieli	0,85
соль	suola	0,71
соперник	kilpailija	0,63
сорока	harakka	0,88
сорокалетний	nelikymmenvuotias	0,64
сорт	lajike	0,78
соседний	viereinen	0,57
сосна	mänty	0,79
сотый	sadas	0,73
сотый	sadasosa	0,61
соус	kastike	0,91
социализм	sosialismi	0,98
социальный	sosiaalinen	0,59
сочувственно	myötätuntoisesti	0,59
союзник	länsiliittoutuneet	0,56
спальня	makuuhuone	0,84
спасать	pelastaa	0,56
спаситель	pelastaja	0,63
специалист	spesialisti	0,62
спирт	pirtu	0,72
спичка	tulitikku	0,78
спортсмен	urheilija	0,84
способность	kyky	0,58
споткнуться	kompastua	0,58
спотыкаться	kompastella	0,71
справедливость	oikeudenmukaisuus	0,66
ссора	riita	0,70
стареть	vanheta	0,62

Русск. слова	Финск. слова	KUC
староста	kylänvanhin	0,74
старость	vanhuus	0,62
старшина	vääpeli	0,61
статья	artikkeli	0,75
степь	aro	0,90
стих	runo	0,71
стихотворение	runo	0,58
столетие	vuosisata	0,58
столица	pääkaupunki	0,72
столовая	ruokasali	0,67
столоначальник	toimisto-päällikkö	0,56
страдать	kärsiä	0,58
странник	vaeltaja	0,79
строгость	ankaruus	0,73
строитель	rakentaja	0,78
строиться	rakenteilla	0,61
суббота	lauantaisin	0,56
суббота	lauantai	0,86
субъект	subjekti	0,60
сугроб	kinos	0,57
судорога	kouristus	0,56
судорожный	kouristuksenomaisesti	0,64
судья	tuomari	0,69
сукно	verka	0,64
суконный	verkainen	0,55
сумма	summa	0,56
сумочка	käsilaukku	0,59
сутки	vuorokausi	0,83
сухарь	korppu	0,60
сухопарый	kuivakka	0,78
существование	olemassaolo	0,68
сфера	sfääri	0,60
сходство	yhdennäköisyys	0,58
счастливец	onnenpekka	0,56
сырость	kosteus	0,56
таблетка	tabletti	0,82
таз	pesuvati	0,59
таинственность	salaperäisyys	0,64
таинственный	salaperäinen	0,72
тайга	taiga	0,89
такси	taksi	0,80
талант	lahjakkuus	0,58
талантливый	lahjakas	0,66
танк	panssari	0,70
танк	tankki	0,63
танцевать	tanssia	0,58

Русск. слова	Финск. слова	KUC
танцовать	tanssia	0,58
таракан	torakka	0,87
тарелка	lautanen	0,67
татарин	tataari	0,90
текст	teksti	0,61
телевизор	televisio	0,92
телеграмма	sähke	0,59
телеграф	lennätin	0,66
теленок	vasikka	0,81
телефон	puhelin	0,65
телогрейка	toppatakki	0,73
телятина	vasikanliha	0,64
телятина	vasikanpaisti	0,70
температура	lämpötila	0,61
тенор	tenori	0,79
теория	teoria	0,93
термин	termi	0,76
терпеливый	kärsivällinen	0,58
терпеливый	kärsivällisesti	0,67
терпение	kärsivällisyys	0,67
терраса	terassi	0,63
тесть	appi	0,74
тетка	tantti	0,61
технический	teknillinen	0,59
технический	tekninen	0,63
теща	anoppi	0,75
тигр	tiikeri	0,92
типичный	tyypillinen	0,79
тиф	lavantauti	0,79
тиф	pilkkukuume	0,58
ткань	kudos	0,58
тогдашний	silloinen	0,60
толстовка	työpusero	0,58
толь	kattohuopa	0,65
толь	kattopahvi	0,58
тополь	poppeli	0,79
топор	kirves	0,87
топот	töminä	0,56
торжественный	juhlallisesti	0,57
торжество	voitonriemu	0,59
трагедия	murhenäytelmä	0,63
трагедия	tragedia	0,66
трагический	traaginen	0,59
трактор	traktori	0,92
тракторист	traktorinkuljettaja	0,94
трамвай	raitiovaunu	0,73
траншея	taisteluhauta	0,60
тренер	valmentaja	0,90

Русск. слова	Финск. слова	KUC
трибуна	puhujakoroke	0,63
тринадцатый	kolmastoista	0,67
тринадцать	kolmetoista	0,59
триста	kolmesataa	0,67
тройка	kolmivaljakko	0,63
тройка	troikka	0,63
троллейбус	trollikka	0,73
троллейбус	trolikka	0,59
троллейбус	johdinauto	0,57
тропинка	polku	0,61
тротуар	jalkakäytävä	0,85
трус	pelkuri	0,79
трусость	pelkuruus	0,85
тряпка	riepu	0,55
туз	ässä	0,84
туман	sumu	0,70
тупик	umpikuja	0,61
турок	turkkilainen	0,71
туча	pilvi	0,58
тщеславие	turhamaisuus	0,87
тысячелетний	tuhatvuotinen	0,71
тысячелетний	vuosituhantinen	0,77
тьфу	hyi	0,59
тюрьма	vankila	0,63
убеждение	vakaumus	0,83
убийство	murha	0,57
убийца	murhaaja	0,74
уборная	pukuhuone	0,62
уважать	kunnioittaa	0,63
уважение	kunnioitus	0,64
уговор	suostuttelu	0,56
уголь	hiili	0,70
угроза	uhkaus	0,65
удобство	mukavuus	0,67
удовлетворить	tyydyttää	0,60
ужаснуться	kauhistua	0,57
ужин	illallinen	0,63
узкий	kapea	0,62
украсть	varastaa	0,68
улан	ulaani	0,64
улика	todistuskappale	0,55
умываться	peseytyä	0,57
умыться	peseytyä	0,56
университет	yliopisto	0,81
уполномоченный	valtuutettu	0,62

Русск. слова	Финск. слова	KUC
ура	hurraa	0,68
урожай	sato	0,74
ускорить	jouduttaa	0,67
услужливый	avulias	0,56
услужливый	alttiisti	0,55
успех	menestys	0,56
успешный	menestyksellinen	0,58
успокоить	rauhoittaa	0,55
успокоиться	rauhoittua	0,67
усталость	väsymys	0,73
устрица	osteri	0,97
утешать	lohdutella	0,56
утешение	lohdutus	0,67
утка	sorsa	0,74
учебник	oppikirja	0,83
учитель	opettaja	0,86
учить	opettaa	0,59
учхоз	koetila	0,82
ущелье	sola	0,60
факел	soihtu	0,83
факт	fakta	0,59
факт	tosiasia	0,57
факультет	tiedekunta	0,75
фанатик	fanaatikko	0,90
фантастический	mielikuvituksellinen	0,62
фартук	esiliina	0,65
фарфоровый	posliininen	0,64
фашист	fasisti	0,94
февраль	helmikuu	0,88
фельдмаршал	sotamarsalkka	0,94
фельдшер	välskäri	0,88
фельетон	pakina	0,60
физически	fyysisesti	0,83
физический	fyysinen	0,62
филиал	tytäryhtiö	0,55
философ	filosofi	0,93
философия	filosofia	0,74
философский	filosofinen	0,70
фильм	filmi	0,67
финдиректор	talouspäällikkö	0,68
фирма	toiminimi	0,65
флейта	huilu	0,86
флигель	piharakennus	0,61
флигель	sivurakennus	0,61
фон	tausta	0,55
фонарик	taskulamppu	0,55
фонарь	lyhty	0,68
фонтан	suihkukaivo	0,64

Русск. слова	Финск. слова	KUC
фонтан	suihkulähde	0,69
формула	kaava	0,63
фортепьяно	piano	0,62
форточка	tuuletusik-kuna	0,67
форточка	tuuletusluuk-ku	0,56
фотограф	valokuvaaja	0,76
фотография	valokuva	0,66
фраза	lause	0,59
фраза	fraasi	0,56
фрак	frakki	0,71
фрак	hännystakki	0,56
франт	keikari	0,58
француженка	ranskatar	0,71
француз	ranskalainen	0,72
фронт	rintama	0,77
футляр	kotelo	0,58
халат	aamutakki	0,58
хан	kaani	0,73
хаос	kaaos	0,79
херес	sherry	0,98
химический	kemiallinen	0,63
хитрить	viekastella	0,60
хитрость	viekkaus	0,65
хлопья	hiutale	0,58
холм	kukkula	0,66
хор	kuoro	0,71
храм	temppeli	0,76
храпеть	kuorsata	0,68
хриплый	käheä	0,63
христианин	kristitty	0,69
христианский	kristillinen	0,85
христианство	kristillisyys	0,56
христианство	kristinusko	0,64
хромой	nilkku	0,63
хромосома	kromosomi	0,99
художественный	taiteellinen	0,60
художник	taiteilija	0,74
хутор	maatalo	0,61
царица	valtiatar	0,62
царь	tsaari	0,60
целовать	suudella	0,70
цепь	ketju	0,60
церемониться	kursailla	0,56
цех	tehdassali	0,66
цивилизация	sivilisaatio	0,84

Русск. слова	Финск. слова	KUC
цилиндр	silinterihattu	0,62
цирк	sirkus	0,72
цыган	mustalainen	0,91
цыганка	mustalais-nainen	0,64
цыпленок	kananpoika	0,88
цыпочки	varpaillaan	0,59
цыпочки	varpaisillaan	0,69
цыпочки	varpaiset	0,58
чайник	teepannu	0,72
чайник	teekannu	0,60
частушка	rekilaulu	0,56
чахотка	keuhkotauti	0,84
человечество	ihmiskunta	0,94
чемодан	matkalaukku	0,87
чемоданчик	kapsäkki	0,71
чердак	ullakko	0,83
череп	pääkallo	0,57
черноглазый	mustasilmäinen	0,70
чертеж	piirustus	0,78
честность	rehellisyys	0,84
честолюбие	kunnianhimo	0,86
четверг	torstai	0,86
четвереньки	nelinkontin	0,58
четвертак	neljännes-rupla	0,70
четверть	neljännes-tunti	0,67
четыреста	neljäsataa	0,76
четырнадцать	neljätoista	0,65
чиновник	virkamies	0,82
чистка	puhdistus	0,59
чистота	puhtaus	0,77
читатель	lukija	0,85
член	jäsen	0,75
чугунный	valurautainen	0,63
чудо	ihme	0,56
шакал	sakaali	0,82
шампанское	sampanja	0,91
шарик	pallonen	0,56
швейцар	ovenvartija	0,77
швейцария	sveitsi	0,71
швеция	ruotsi	0,80
шелк	silkki	0,61
шестнадцать	kuusitoista	0,61
шифр	koodi	0,77
шкаф	kaappi	0,66
шлакоблок	tiili	0,64
шофер	autonkuljettaja	0,65

Русск. слова	Финск. слова	KUC
шофер	kuljettaja	0,61
шпага	miekka	0,56
шпион	vakoilija	0,79
шпора	kannus	0,84
шрам	arpi	0,71
штаб	esikunta	0,78
штабс-капитан	alikapteeni	0,77
шуба	turkki	0,62
щетина	parransänki	0,57
щетка	lattiaharja	0,63
щи	kaalikeitto	0,72
щиколотка	kehräsluu	0,58
эгоист	egoisti	0,87
эй	hei	0,60
экран	kuvaruutu	0,55
экран	valkokangas	0,63
эксперт	ekspertti	0,57
электричество	sähkö	0,68
энергия	energia	0,65
энергия	tarmo	0,58

Русск. слова	Финск. слова	KUC
эпоха	aikakausi	0,56
эскадрон	eskadroona	1,00
эстонец	eestiläinen	0,84
эхо	kaiku	0,61
юбка	hame	0,62
юг	etelä	0,79
юридический	juridinen	0,79
юридический	lainopillinen	0,56
яблоко	omena	0,74
яблоня	omenapuu	0,90
явление	ilmiö	0,68
яд	myrkky	0,65
ядовитый	myrkyllinen	0,64
яйцо	kananmuna	0,56
ямщик	kyytimies	0,66
январь	tammikuu	0,94
японец	japanilainen	0,79
японский	japanilainen	0,75
ярмарка	markkinat	0,86

5.2. Отчасти правильные ПЭ-пары

Русск. слова	Финск. слова	KUC
бетонный	betonilaatta	0,61
биологический	biologia	0,74
бродячий	kulkukoira	0,63
ватный	toppahousut	0,67
взад	edestakaisin	0,61
виноградный	viinirypäleterttu	0,57
висячий	riippulukko	0,59
висячий	riippusilta	0,59
восковой	vahakynttilä	0,59
выборы	vaalipäivä	0,64
гений	nerous	0,55
городок	kampus	0,57
гражданский	kansalaissota	0,77
граница	ulkomaa	0,62
гребень	mäenharja	0,55
грудной	sylilapsi	0,55
губернский	lääninmarsalkka	0,59
дворницкий	talonmies	0,58
двоюродный	serkku	0,63
дивизион	patteristo	0,77

Русск. слова	Финск. слова	KUC
дивизия	divisioonankomentaja	0,58
заведение	oppilaitos	0,57
звездный	tähtitaivas	0,60
звуковой	äänielokuva	0,66
зеркальный	peilikaappi	0,59
змеиный	käärmeenmyrkky	0,67
извинить	anteeksi	0,58
кавторанг	kapteeni	0,66
каменный	kivimuuri	0,56
капуста	hapankaali	0,56
кирпичный	tiilitehdas	0,56
кирпичный	tiilimuuri	0,56
ковбойка	ruutupaitainen	0,73
командир	rykmentinkomentaja	0,56
комик	esittelijä	0,56
консервы	kalasäilyke	0,57
кружка	peltimuki	0,55
куриный	kanakeitto	0,56

Русск. слова	Финск. слова	KUC
лаборатория	probleemi-laboratorio	0,61
лень	laiskottaa	0,58
липовый	lehmuskuja	0,62
лисий	ketunnahkai-nen	0,58
метро	metroasema	0,63
меховой	turkisliivi	0,56
модный	muotiliike	0,55
морозный	pakkasyö	0,56
мыльный	saippuavesi	0,59
мыльный	saippuakupla	0,66
мышка	kainalo	0,58
мышца	hauislihas	0,55
навоз	hevosenlanta	0,55
навоз	lannanajo	0,55
нарушитель	sopimuksen-rikkoja	0,84
непривычка	oudokseltaan	0,64
областной	aluelehti	0,58
оборона	puolustus-kannalla	0,59
огневой	tuliasema	0,83
ореховый	pähkinän-kuori	0,57
отпечаток	sormenjälki	0,58
палуба	yläkansi	0,60
пассажир-ский	matkustajava unu	0,68
пахучий	tuoksua	0,56
перекрести-ться	ristinmerkki	0,60
печень	maksasyöpä	0,57
пожарный	palokunta	0,60
позиция	tuliasema	0,56
познание	tiedonjano	0,55
полковой	rykmentin-komentaja	0,72
предводи-тель	lääninmar-salkka	0,55
проволока	piikkilanka	0,64
промышлен-ность	tehdas-teollisuus	0,58
птичий	lintu	0,55

Русск. слова	Финск. слова	KUC
пузырек	ilmakupla	0,56
пузырь	saippuakupla	0,61
разочаро-ванный	pettyä	0,59
ревновать	mustasuk-kainen	0,59
речной	jokisatama	0,57
связка	avainnippu	0,65
семейный	perhe-elämä	0,60
скорлупа	pähkinän-kuori	0,57
сниться	uni	0,55
совет	tiedeneu-vosto	0,55
советник	valtioneuvos	0,56
соленый	suolakurkku	0,57
соломенный	olkikatto	0,56
соломенный	olkihattu	0,65
союзник	liittoutua	0,56
спортсмен	urheilla	0,55
статский	valtioneuvos	0,76
строитель-ный	rakennus-materiaali	0,55
товарный	tavaravaunu	0,56
уборка	heinänkorjuu	0,58
указатель-ный	etusormi	0,85
урок	yksityistunti	0,55
фабричный	tehdas-teollisuus	0,56
фокус	korttitemppu	0,56
чиж	varpunen	0,70
штаб	esikunta-päällikkö	0,56
яд	käärmeen-myrkky	0,61
язва	mahahaava	0,56
яровой	kevätvilja	0,58
яровой	kevätvehnä	0,73

5.3. Ошибочные ПЭ-пары

Русск. слова	Финск. слова	KUC
беспощад-ный	pahennus	0,58

Русск. слова	Финск. слова	KUC
биологичес-кий	idealistinen	0,55

Русск. слова	Финск. слова	KUC
бойцовый	baretti	0,56
брякнуть	sankkeri	0,56
брякнуть	pehmyt	0,58
буржуазный	kommentaari	0,55
ввязаться	olympolai-nen	0,55
вейсманизм-морганизм	kommentaari	0,55
втянуть	kauhallinen	0,56
выговор	kuivuri	0,56
выждать	vierre	0,55
выставка	kuivuri	0,56
голенище	baretti	0,56
гостинец	tuliainen	0,74
диванчик	alkovi	0,56
диск	valomainos	0,55
добросовестный	antiteesi	0,60
домработ-ница	kvantti	0,59
дополнитель ный	ladelma	0,55
забвение	viestiä	0,57
задел	nuohooja	0,56
заезжать	alkovi	0,57
зарплата	korkeapaine	0,55
засвистеть	valomainos	0,55
затрепетать	lyöjä	0,55
затягиваться	nootti	0,55
захлебы-ваться	lyöjä	0,55
знамя	käenpoika	0,55
идеологи-ческий	idealistinen	0,55
инспектор	palotarkasta-ja	0,57
информация	tutkiskelu	0,60
исправный	sotilaallisesti	0,55
исследовать	kansatieteel-linen	0,55
киса	kisa	0,84
клониться	haaskaeläin	0,55
коммуни-стический	internatio-naali	0,55
конвенция	sopimuksen-rikkoja	0,70
кондитер-ский	sillisalaatti	0,56
коренастый	vesuri	0,55
корма	pukinpartai-nen	0,55
крохотный	lippi	0,55
крылышко	venonen	0,57

Русск. слова	Финск. слова	KUC
крюк	käenpoika	0,55
кто-либо	valtarakenne	0,55
лисий	konekivääri-pesäke	0,55
лисий	päivämatka	0,55
мастерок	lasta	0,57
молочный	toverukset	0,55
монтер	sähkötyö	0,55
наблюдатель ный	tulenjohtopai kka	0,57
насилие	kerettiläisyys	0,55
обернуть	sorkkatauti	0,55
обшитый	alkovi	0,57
окраска	sopeutumis-kyky	0,55
патриарший	lampi	0,61
пенсне	kakkulat	0,56
перебросить	akustiikka	0,55
пленка	kommentaari	0,55
полутемный	alkovi	0,58
полчаса	tunti	0,60
постарше	kivikasvoi-nen	0,55
предаться	kevätaamu	0,55
предаться	nelikulmio	0,55
пришелец	kauko-ohjain	0,55
протиснуть-ся	epämieluinen	0,55
пряжка	baretti	0,57
раздвинуть	rasvaton	0,55
раздражить	kuivuri	0,56
разрушать	logistiikka	0,58
разыграться	venonen	0,57
расстрел	asevarikko	0,55
резной	puuleikkaus	0,60
рим	käenpoika	0,56
рота	komppanian-päällikkö	0,55
свинство	kiilata	0,55
секция	alkovi	0,57
сени	porstua	0,55
сеттер	sekarotuinen	0,57
славка	teknokraatti	0,57
слесарь	sorvari	0,66
слободка	variksenpesä	0,83
снабдить	asevarikko	0,55
сознать	antiteesi	0,56
стеганый	tikata	0,57
строитель-ный	minareetti	0,55
структура	alkovi	0,57
субъект	sopeutumis-kyky	0,55

Русск. слова	**Финск. слова**	**KUC**
суматоха	baretti	0,56
таинствен-ность	eliksiiri	0,55
телевизор	puimuri	0,55
телевизор	namikka	0,55
телятина	sillisalaatti	0,56
тесемка	korvus	0,55
теснота	keitinpiiras	0,55
тиражный	pukinpartai-nen	0,56
толщина	penaali	0,55
уволить	irtisanomis-aika	0,55
угар	kapiot	0,55
укладывать-ся	laskutikku	0,56
утащить	voitonmalja	0,58
утерпеть	toppi	0,57
флакон	voitonmalja	0,55
франция	pehmyt	0,56
шлаг	pastori	0,63
щиколотка	rakennuslevy	0,55
щиколотка	kavennus	0,55

Приложение 6. Статистические данные по оригинальным финским текстам и переводам на финский язык с русского

6.1. Список слов СХТ, отсутствовавших в ПФ (частота 40 и более)

Слово	**Частота**	**Отн. частота**
mämmi	313	0,17
mutsi	297	0,16
kännykkä	167	0,09
oikeasti	141	0,07
joulupukki	130	0,07
viikonloppu	114	0,06
keiju	104	0,06
imuri	100	0,05
ryssä	99	0,05
pizza	98	0,05
per	87	0,05
bileet	82	0,04
okei	81	0,04
makuupussi	81	0,04
kokko	78	0,04
parkkipaikka	76	0,04
pizzeria	74	0,04
kunnanjohtaja	72	0,04
kataja	71	0,04
pinja	70	0,04
eka	68	0,04
vaaputtaa	68	0,04
baarimikko	67	0,04
konkurssi	66	0,03
noitamummo	66	0,03
seksuaalisuus	60	0,03
muovipussi	58	0,03
semmonen	56	0,03
juotava	54	0,03
pakettiauto	53	0,03
hitti	52	0,03
huoltoasema	50	0,03
kirkkoherra	49	0,03
ornitologi	49	0,03
rööki	47	0,02
etupenkki	47	0,02
hornankattila	46	0,02
muovikassi	45	0,02
merenrantakaupunki	45	0,02
samantien	45	0,02

Слово	Частота	Отн. частота
leffa	44	0,02
terveyskeskus	43	0,02
myymäläpäällikkö	43	0,02
unioni	43	0,02
terapeutti	41	0,02
paapoa	41	0,02
viestinviejä	41	0,02
ahkio	40	0,02
iltapäivälehti	40	0,02

6.2. Список слов ПФ, отсутствовавших в СХТ (частота 40 и более)

Слово	Частота	Отн. частота
kopeekka	239	0,11
huovikas	145	0,07
samovaari	143	0,06
aalto	141	0,06
jottei	135	0,06
tataari	116	0,05
siirtola	114	0,05
tilanomistaja	112	0,05
mainiosti	106	0,05
ajomies	93	0,04
ikäänkuin	93	0,04
halveksivasti	86	0,04
hirvittävä	85	0,04
kiirehtää	75	0,03
liiketoveri	72	0,03
husaari	71	0,03
portaat	71	0,03
hegemoni	69	0,03
kuvernementti	66	0,03
kulakki	62	0,03
kahvilanpitäjä	59	0,03
huomattava	57	0,03
läsnäollessa	55	0,03
tantti	54	0,02
ohjas	54	0,02
puhujakoroke	52	0,02
lukuunottamatta	51	0,02
tiedustelupalvelu	50	0,02
sakaali	46	0,02
tilanhoitaja	44	0,02
hybridi	44	0,02
ehkei	44	0,02
ruukki	44	0,02
kromosomi	44	0,02
joutava	43	0,02
rajoni	42	0,02

Слово	Частота	Отн. частота
apulaisjohtaja	41	0,02
vartiotupa	41	0,02
päärata	40	0,02

6.3. Список самых частотных слов СХТ и ПФ

Слово	СХТ		ПФ	
	Частота	Отн. част.	Частота	Отн. частота
olla	105321	55,83	115115	51,37
ja	76319	40,46	92660	41,35
ei	41633	22,07	51020	22,77
hän	36314	19,25	73567	32,83
minä	25499	13,52	30736	13,72
se	22392	11,87	17261	7,70
että	18553	9,84	21038	9,39
joka	13918	7,38	18054	8,06
sanoa	13853	7,34	15946	7,12
mutta	13409	7,11	20727	9,25
kuin	12114	6,42	14218	6,34
kun	11881	6,30	9004	4,02
niin	10866	5,76	14830	6,62
ne	9447	5,01	8277	3,69
tulla	9379	4,97	9360	4,18
sinä	8596	4,56	11297	5,04
mikä	8139	4,31	14114	6,30
saada	7481	3,97	7642	3,41
mies	6361	3,37	5235	2,34
voida	6223	3,30	8844	3,95
nyt	6186	3,28	8250	3,68
he	6116	3,24	10687	4,77
sitten	5855	3,10	7550	3,37
kaikki	5823	3,09	12936	5,77
sitä	5816	3,08	6621	2,95
me	5718	3,03	6566	2,93
pitää	5432	2,88	5711	2,55
mennä	5360	2,84	6251	2,79
tämä	5299	2,81	9808	4,38
tehdä	5071	2,69	6074	2,71
vain	4953	2,63	7102	3,17
aika	4902	2,60	5681	2,54
jos	4695	2,49	4602	2,05
jo	4607	2,44	7191	3,21
äiti	4579	2,43	**1971**	**0,88**
toinen	4510	2,39	6668	2,98
isä	4381	2,32	**1708**	**0,76**
vielä	4314	2,29	6576	2,93
alkaa	4176	2,21	6158	2,75
katsoa	4155	2,20	4116	1,84
mikään	4104	2,18	5109	2,28
itse	4092	2,17	7182	3,21

Слово	СХТ		ПФ	
	Частота	Отн. част.	Частота	Отн. частота
kanssa	4085	2,17	4294	1,92
lähteä	4053	2,15	5056	2,26
nainen	4020	2,13	2819	1,26
tietää	3878	2,06	5453	2,43
muu	3814	2,02	3669	1,64
kysyä	3812	2,02	3581	1,60
tai	3783	2,01	3341	1,49
nähdä	3770	2,00	5308	2,37
käsi	3693	1,96	4929	2,20
ottaa	3574	1,89	4784	2,14
ihminen	3537	1,88	4678	2,09
jokin	3527	1,87	5337	2,38
vaikka	3466	1,84	2748	1,23
silmä	3453	1,83	4326	1,93
siitä	3384	1,79	4745	2,12
haluta	3195	1,69	3056	1,36
käydä	3160	1,68	3926	1,75
hyvä	2990	1,59	4599	2,05
oma	2964	1,57	3609	1,61
antaa	2923	1,55	4468	1,99
istua	2919	1,55	3155	1,41
miten	2913	1,54	4140	1,85
kertoa	2858	1,52	**1844**	**0,82**
asia	2830	1,50	3269	1,46
pieni	2827	1,50	**1768**	**0,79**
ajatella	2792	1,48	3710	1,66
sillä	2778	1,47	3273	1,46
puhua	2659	1,41	3854	1,72
poika	2647	1,40	2564	1,14
päästä	2575	1,37	2514	1,12
siinä	2527	1,34	3761	1,68
ovi	2513	1,33	2464	1,10
päivä	2504	1,33	3399	1,52
jäädä	2493	1,32	**2039**	**0,91**
näyttää	2454	1,30	**2190**	**0,98**
ennen	2447	1,30	**1881**	**0,84**
te	2406	1,28	10830	4,83
aina	2404	1,27	**2024**	**0,90**
koko	2371	1,26	3856	1,72
enää	2348	1,24	2841	1,27
kyllä	2301	1,22	**2280**	**1,02**
suuri	2269	1,20	**2199**	**0,98**
joku	2262	1,20	**2282**	**1,02**
siellä	2262	1,20	3210	1,43
nousta	2245	1,19	2361	1,05
hetki	2208	1,17	**2324**	**1,04**
tuntea	2203	1,17	3830	1,71
ääni	2169	1,15	2609	1,16
yrittää	2158	1,14	**1754**	**0,78**
tyttö	2147	1,14	**1893**	**0,85**
muistaa	2143	1,14	**2061**	**0,92**
kuulla	2129	1,13	2938	1,31

Слово	СХТ		ПФ	
	Частота	Отн. част.	Частота	Отн. частота
jälkeen	2021	1,07	2369	1,06
vanha	1998	1,06	**1902**	**0,85**
taas	1993	1,06	2815	1,26
juuri	1987	1,05	2675	1,19
vaan	**1921**	**1,02**	3214	1,43
pää	**1909**	**1,01**	**2678**	**1,20**
silloin	**1753**	**0,93**	2938	1,31
elämä	**1722**	**0,91**	2390	1,07
sellainen	**1677**	**0,89**	2915	1,30
hyvin	**1668**	**0,88**	2379	1,06
aivan	**1623**	**0,86**	2920	1,30
täällä	**1623**	**0,86**	2335	1,04
heti	**1586**	**0,84**	2749	1,23
ymmärtää	**1518**	**0,80**	2771	1,24
kuka	**1494**	**0,79**	2650	1,18
kasvot	**1485**	**0,79**	2970	1,33
kaksi	**1470**	**0,78**	2545	1,14
vastata	**1381**	**0,73**	2996	1,34
tuo	**1257**	**0,67**	4975	2,22
sana	**1191**	**0,63**	2358	1,05
no	**938**	**0,50**	3028	1,35