“Искалка” Д. В. Самойлова

© Николай Бузикашвили

Институт системного анализа РАН
buzik@cs.isa.ru

Аннотация

Статья посвящена *Искалке* — двухкомпонентной системе поддержки перевода научной литературы, Одна компонента *Искалки* предназначена для автоматической и полуавтоматической синхронизации параллельных текстов, а вторая обеспечивает синхронизованный многоязыковой поиск в корпусе параллельных текстов, подготовленном с помощью первой. Поисковая компонента *Искалки* является основным элементом рабочего места редактора переводной научной литературы и обеспечивает терминологическое единство переводов. Построенная от начала и до конца Д.В. Самойловым, *Искалка* представляет собой редкий и яркий пример того, как изящная и адекватная формулировка задачи позволяет простыми средствами достичь наивысшей эффективности работы.

1 Введение

Инструмент поиска, которому посвящена статья, создавался для *реального* решения *реальной*, а тем самым, достаточно специфической задачи, учет особенностей которой позволяет найти элегантное и нетрудоемкое решение. Обычно разработчик и [потенциальный] заказчик видят задачу по-разному и имеют разные имплицитные представления о само собой разумеющемся. Как результат, программный продукт часто сочетает избыточность с неполнотой. По счастливому совпадению, в случае с *Искалкой*, такого разрыва не было: один и тот же человек был наделен способностями остроумно и точно поставить задачу, а затем спроектировать и воплотить ее решение. Это Дмитрий Вадимович Самойлов (1958–2005), являвшийся бессменным главным редактором издательства “Практика”. Основное направление деятельности этого издательства — перевод на русский язык актуальной медицинской литературы.

В значительной мере стараниями Самойлова были сформулированы [2] и ежедневно воплощались принципы, сделавшие столь популярными книги этого издательства:

1) перевод — не подстрочник, смысл (если он есть, а уже это критерий отбора, что переводить) должен быть сохранен и доступен врачу-практику;

2) книга должна быть не только ясной, но и переведена на литературном русском;

3) соблюдение терминологического единства — если объект уже был назван по-русски в одной из книг, выпущенных издательством, под этим и никаким другим именем он должен упоминаться и *во всех остальных* книгах.

Если воплощение первых двух принципов связано скорее с божьим даром, которым сполна был наделен Д.В. Самойлов, то для реализации третьего принципа достаточно аккуратности редактора в сочетании с технологической поддержкой его работы. Именно как средство такой поддержки и создавалась система, с самого начала получившая естественное название *Искалка* и нашедшая применение не только в издательстве, но и за его пределами. Этот во всех отношениях успешный проект, выполненный не в подражание и не от противного, ранее упоминался лишь вскользь [3]. Однако, прежде всего как образец правильно и элегантно сформулированной и максимально просто, эффективно и эргономично решенной специальной задачи поиска, он заслуживает более подробного описания.

Задача поддержки терминологически преемственного перевода как не декларативная, а подлежащая автоматизации реальная задача издательства, была сформулирована Самойловым спустя два года после создания издательства, когда корпус переведенных текстов стал достаточно велик, чтобы, с одной стороны, быть уверенным, что новые книги терминологически пересекаются с переведенными ранее, но, с другой стороны, выяснение для каждого [допускающего разные переводы] термина, не переводился ли он уже и как именно, стало слишком трудоемким.

Искалка с самого начала строилась как система *синхронизованного поиска* в корпусе параллельных текстов, при котором одновременно автоматически отображается фрагмент текста, в котором найдено искомое, и соответствующий фрагмент парного ему текста (оригинала или перевода). Поэтому уже первая версия *Искалки* избавила редакторов и переводчиков издательства от трудоемкой обязанности ручного поиска ранее выполненных

Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2007, Переславль-Залесский, Россия, 2007.

вариантов перевода. В результате, *Искалка* сразу же стала незаменимым помощником, резко облегчившим работу над переводом.

Последние десять лет технологическая цепочка подготовки переводной книги в “Практике” завершается этапом пополнения корпуса параллельных текстов, на котором оригинал и перевод синхронизируются и индексируются *Искалкой*. На сегодня оригиналы и переводы всех книг, изданных “Практикой” (порядка 50 млн. слов перевода), входят в общедоступный в локальной сети синхронизованный и проиндексированный корпус, использование которого, с помощью поискового блока *Искалки*, давно стало обыденным для редакторов и переводчиков издательства.

2 Задача поддержки перевода медицинской литературы

2.1 Содержательная постановка задачи

Первое, что обнаруживает начинающий переводчик медицинской литературы, сам по образованию врач, — это отсутствие в актуальной медицинской литературе, причем, прежде всего англоязычной, стандартной терминологии. Один и тот же объект может по-разному называться не только разными, но и одним и тем же автором, а совершенно несхожие объекты нередко носят одно и то же название. С учетом исповедуемого в издательстве принципа терминологического единства, задача переводчика кажется почти невыполнимой. Как же ему помочь?

Задача: перевести новый текст, соблюдая максимальную терминологическую преемственностью с ранее выполненными [в этом издательстве] *переводами*.

Исходные условия:

(1) отсутствие терминологического стандарта в оригиналах в сочетании с требованием терминологического единства переводов;

(2) знание иностранного языка пользователем (переводчиком, редактором) достаточно, чтобы понять соответствует ли и в какой мере фрагмент перевода фрагменту оригинала.

2.2 Предварительная категоризация задачи

Первое приходящее на ум работника издательства решение — снабдить пользователя качественным многоязыковым словарем (тезаурусом) и тем решить проблему. Но (*а*) для *актуальной* медицинской лексики такого словаря в действительности нет; (*б*) словарь должен содержать в русскоязычных ячейках “диалектизмы”, принятые в *этом* издательстве, и может быть создан только в самом издательстве. При этом кто-то должен подробно описать и классифицировать контексты употребления возможных (в иноязычной литературе) и требуемых (в переводе) словарных значений. Но такой контекстно-погруженный словарь отличается от традиционных бумажных словарей и структурой, и объемом, и манерой использования. Более того, он был бы не слишком реален и удобен в случае его организации как гиперссылочного текста со множеством ссылок на примеры “оригинал—перевод” и полным ссылочным покрытием десятков тысяч словарных терминов.

Построение такого словаря как решаемая “в лоб” *самостоятельная* задача невероятно трудоемка, требует привлечения специалистов в разных областях медицины, а с учетом разнообразия тематики и контекстов, объем словаря окажется едва не того же порядка, что и переводимых книг. Т.е. в действительности прямое решение этой задачи невозможно. Единственный реальный способ создать такой словарь — как “*побочный*” продукт самой переводческой деятельности, т.е. подход *словарь для перехода* становится реализуемым лишь при его инверсии — *переводы для словаря*. Заметим, что, на сегодня, вероятно, самый полный англо-русский словарь медицинской лексики создан (продолжая, разумеется, постоянно наращиваться) именно как побочный продукт деятельности использующих *Искалку* переводчиков.

В свою очередь, у специалиста по информационно-поисковым системам, при беглом ознакомлении с задачей, может возникнуть соблазн проблематизировать поддержку медицинского перевода как задачу многоязыкового поиска (напр., [7, 8]). Упомянутая проблема словарной поддержки, скорее всего, была бы сочтена сугубо технической и вполне разрешимой, учитывая, что в издательстве сосредоточены квалифицированные кадры.

Оба подхода были забракованы Д.В. Самойловым, а в качестве альтернативы было предложено то, что сегодня называется поиском в корпусе параллельных текстов. Действительно, помимо исходных условий (1)-(2), задача поддержки преемственного перевода, включает и дополнительные условия.

Дополнительные условия:

(3) наличие корпуса оригинал-текстов, полученных в виде *doc*-файлов непосредственно от издателя оригинала;

(4) наличие парного оригиналам корпуса ранее выполненных в издательстве переводов, представленных также в виде *doc*-файлов.

Мы вернемся к выбранной постановке задачи в разделе 2.4, а сейчас рассмотрим специфику перевода медицинской литературы.

2.3 Особенности медицинской лексики и перевода

Все независимо опрошенные автором высокопрофессиональные и активно работающие специалисты в биомедицинской науке, сходились в том, что потенциальная неоднозначность переводов биомедицинских текстов выше, чем у текстов из любой другой области знания. Конечно, со стороны, это мнение можно было бы отнести на счет

внутрипрофессионального видения. Однако, того же мнения придерживаются и переводчики, и специалисты из других областей, вовлеченные в биомедицинские исследования. Все они согласны с тем, что (*а*) "пословный" перевод медицинского текста без понимания того, о чем идет речь, обречен на ошибочность, (*б*) даже специалисту для перевода небольшого фрагмента медицинского текста требуется "дополнение", существенно большее, чем для текстов из других областей знания. Вероятно, речь идет о причине, названной ниже эффектом сочетания специализации и системности. Однако перечислим по порядку то, что на наш взгляд, делает перевод медицинской литературы особенно сложным.

1. *Обилие комплексных названий и описаний*. Медицинские названия в оригинале и по-русски очень часто являются составными конструкциями из нескольких слов. Более того, описании одного и того же явления (например, симптоматики) дается в свободной форме, но при этом очевидным требованием к различным описаниям одного и того же является узнаваемость, т.е. оформленные в разных словах описания одного и того же должны распознаваться специалистом как описания одного референта, а разных — не быть спутаны. На уровне отдельных термов и их сочетаний мера близости таких описаний построена быть не может, как, уже в силу "технических", но перерастающих в *принципиальные* ограничений, не может она быть построена и средствами семантического анализа.

2. *Отсутствие стандартов, неоднозначность*.

2.1. *Новая терминология, разные коллективы*. В актуальной биомедицине в порядке вещей "диалекты", когда один автор или группа авторов используют одну терминологию, а другие — другую: *антибактериальная терапия* суть то же, что *антимикробная терапия*, а обе они — то же, что *антибиотикотерапия*.

2.2. *Устоявшиеся неоднозначности*. Помимо "сиюминутных" диалектов, и в давно и прочно устоявшейся медицинской лексике одно и то же название часто используется (в том числе одним и тем же автором в одном и том же тексте) для обозначения разных референтов, а один и тот же референт имеет несколько названий.

3. *Эффект международного языка*. Значительная часть актуальных текстов создается не носителями английского, которые, в частности, следуют словарным вариантам (пример на Рис. 1). Эффект международного языка — это эффект двойного перевода — например, испаноязычный исследователь пишет статью по-английски, русскоязычный переводчик переводит ее на русский. Несовпадение испано-английских и русско-английских соответствий умножает вероятность ошибочного перевода на русский.

4. *Эффект специализации исследователей и системности объекта*. Корпус знаний и терминов, охватываемый медицинской наукой, а тем более, биомедициной в целом, никак не уже, чем в других отраслях знания, специализация выше, но, в силу системности объекта, возможностей для стыков разных подразделов существенно больше и встречаются они чаще. Медик при чтении текста по своей специальности нередко сталкивается с вкраплениями из других разделов медицины, о терминологической точности которых он, без специального уточнения, судить не может. Как результат, при переводе медицинской литературы проблема лексических неоднозначностей гораздо острее, чем в любой другой области, а следствия возможные ошибок очевидно более драматичны.

Хотя обсуждаемая в статье задача направлена на терминологическую унификацию, однако говоря о медицинской лексике в целом, нет никаких оснований ожидать снижения терминологического разнобоя, и прежде всего в международной (= англоязычной) литературе. Для других естественнонаучных отраслей этот, только нарастающий, разнобой гораздо менее характерен.

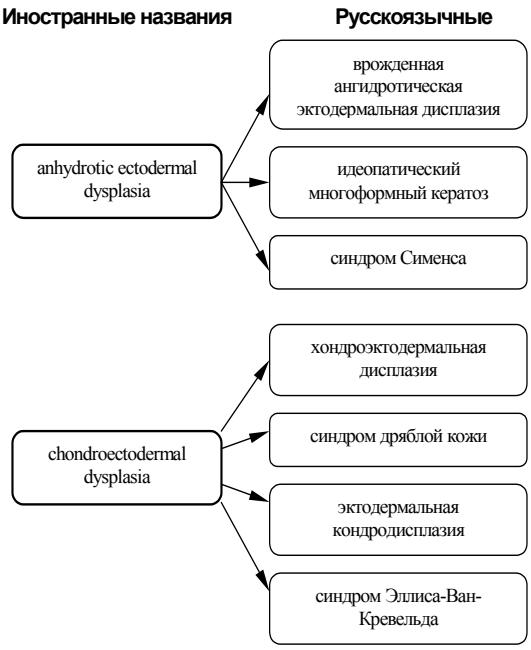

Рис. 1. Типичный граф словарных соответствий

Остановимся чуть подробнее на синонимии и омонимии в биомедицине.

(*а*) биомедицинская *синонимия*. В разных местах переводимого корпуса один и тот же объект или явление (болезнь, симптоматика) имеет разные названия и описывается в разных терминах. Терминологическое разнообразие, естественное для разных изданий и разных школ, на уровне коллективного издания (справочника, руководства) превращается в существенную помеху: не всегда легко догадаться, что разные названия относятся к одному и тому же явлению.

(*б*) биомедицинская *омонимия*. Одно и то же название (причем состоящее вовсе не из одного термина, что было бы объяснимо) используется для обозначения разных объектов или явлений. Пример,

покрывающий целый омонимический класс "*болезь—синдром*": "Nephrotic syndrome — это и нефротический синдром, и болезнь (липоидный нефроз, или болезнь минимальных изменений). Иногда nephrotic syndrome присутствует *в обоих значениях в одном абзаце*" [2]. На Рис. 1 приведены два типичных словарных подслота для термина *displasia*, заимствованные из англо-русского медицинского словаря [1]. Заметим, что значения в правой части — отнюдь не синонимы и имеют разных референтов.

Если изобразить связи название—референт (то, что этим термином обозначается) для двух языков, а затем выкинуть из цепочки *Name—референт—Название* референтов и слить ребра, соединяющие пару *Name—Название*, получим типичный словарь. Например, между названиями с Рис.2 установится взаимно однозначное соответствие, позволяющее переводчику, не вдумываясь, просто заменять иностранное слово на русское. Т.е. когда значения иноязычного термина-омонима совпадают со значениями русскоязычного (Рис.2), никаких проблем с переводом не возникает. Эта ситуация типична, скажем, для математики, где мы найдем порядка 10 значений слова "модуль", но в любом языке все эти значения имеют одно и то же имя.

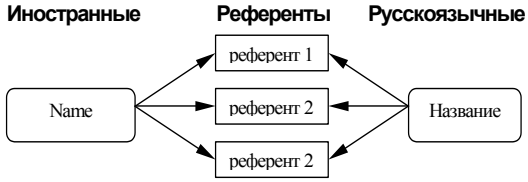

Рис. 2. "Хорошая омонимия"

Синонимия и омонимия в медицине массовы (что само по себе еще не есть отличие от других отраслей знания), однако *классы синонимов и омонимов в одном языке не соотносятся с одноименными классами в другом*.

Может показаться, что (*а*) источником переводческих затруднений могут быть только те англоязычные названия, которым соответствует более одного русскоязычного, (*б*) необходимость референтов (и, соответственно, слов в контексте) не очевидна, а в случае одного *словарного* варианта перевода явно излишня. Но это не так, поскольку русскоязычные названия для каких-то референтов вообще отсутствуют, а использование предлагаемого словарем значения, относящегося к другому референту, способно только ввести читателя в заблуждение.

Приведенные ниже примеры заимствованы из канонической работы по переводу медицинской литературы [2].

Целый класс омонимии соответствует ситуации, когда в одном языке (английский) одно и тот же название используется и для обозначения болезни и как название некоторого, и не обязательно связанного с болезнью, синдрома, тогда как в другом языке (русский) и эти болезнь и синдром имеют разные названия (Рис. 3).

Рис. 3. Примеры омонимии

А вот пример, когда "явно одно и то же" название, обозначает разное (Рис. 4). *Пернициозная анемия* — очевидная и буквальная калька с *pernicious anemia*. При этом *pernicious anemia* и *пернициозная анемия* — реально существующие названия. Казалось бы, и обозначать они должны одно и то же. Однако *пернициозная анемия* — это дефицит витамина B_{12}, тогда как референт *pernicious anemia* по-русски называется аутоимунным (атрофическим) гастритом, который, впрочем, является одной из возможных причин пернициозной анемии.

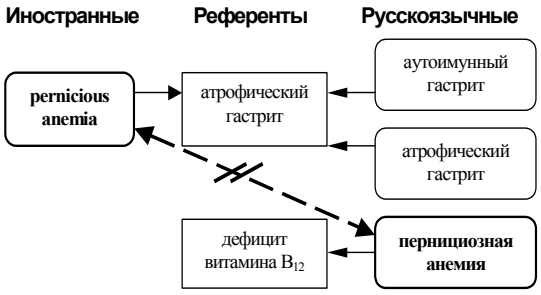

Рис. 4. Ложные эквиваленты

Русскоязычный термин должен быть привязан в первую очередь к *референту*, причем, согласно принципу терминологического единства, из всех допустимых русскоязычных вариантов должен использоваться один (Рис. 5).

В случае с актуальной медицинской литературой имеет место словарная ситуация, изображенная на Рис. 6:

(*а*) несколько оригинальных терминов для одного референта;

(*б*) несколько значений (референтов) для одного оригинального термина, причем в русском отсутствует термин, покрывающий те же значения;

(*в*) отсутствие оригинального термина в словаре, либо, что гораздо хуже, его присутствие, но в сочетании с отсутствием актуального значения (ситуация, типичная для любой быстро и экстенсивно развивающейся отрасли, какой является медицина).

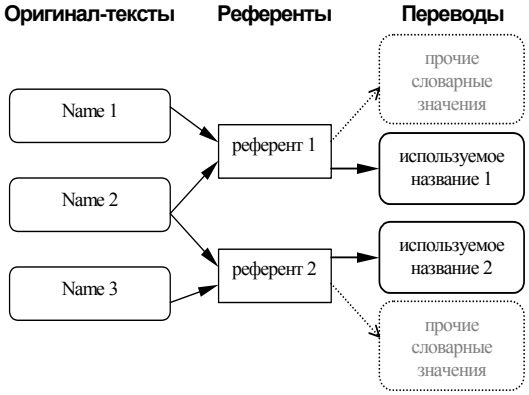

Рис. 5. Граф идеального перевода

Так, на Рис. 6 только для оригинального названия *Name-4* нет русскоязычного медицинского эквивалента, всем остальным оригинал-названиям соответствует хотя бы одно русскоязычное. Однако русскоязычная терминология покрывает не всех референтов и — худший вариант — название *Name -2*, в оригинале использованное в значении *референт-2*, согласно словарю обозначает только *референта-1*, а *Name -3* в значении *референт-3*, при слепом следовании словарю, будет переведен как *Название-2*. Едва ли кого-то прельстит перспектива стать жертвой врачебной ошибки, вызванной отсутствием актуального слота в словаре.

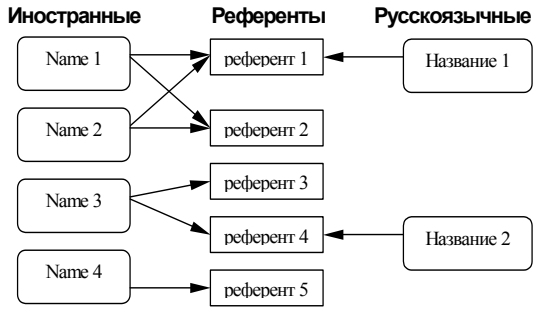

Рис. 6. Словарные соответствия

В свете сказанного, словарь, особенно взятый извне, не может быть *основой* поисковой системы, зато может быть — очень ценным для “повторного использования” в переводческой деятельности — побочным *со-продуктом* ее эксплуатации.

2.4 Выбранная постановка задачи и принципиальные решения

Итак, задачей поддержки перевода медицинской литературы является не многоязыковый поиск, а показ 1) того, как было переведено оригинальное название, употребленное в конкретном окружении, и 2) того, что соответствует в оригинале переведенному названию, употребленному в конкретном окружении перевода. Задача состоит вовсе не в том, чтобы в *однородной* совокупности текстов, найти все вхождения названия и его (весьма многочисленных) значений на иных языках, а в том, чтобы по запросу находить место *в тексте, где слово встретилось, и место в тесте, **парном** этому тексту*. Тем самым, основной задачей оказывается *синхронизация* показа (*а*) отысканного названия и контекста, в котором оно употреблено, и (*б*) образа или прообраза этого контекста соответственно в переводе или оригинале.

Если бы решение было выбрано в рамках многоязыкового поиска, то всего лишь приемлемая работа системы потребовала бы создания очень качественного словаря. При этом система все равно допускала бы пропуски (низкая полнота поиска, как результат неучтенной синонимии) и ее работу отличала бы избыточность (низкая точность поиска, как результат неучтенной омонимии), что не удовлетворяет жестким издательским требованиям к качеству перевода.

Наоборот, выбранное решение *вообще не требует **никакой поддержки** двуязычным словарем* и, более того, *само является инструментом для построения такого словаря*.

Итак, суть задачи определяется не *многоязычием* документов, а их *парностью* (оригинал—перевод). А сама задача формулируется так:

*найти удовлетворяющие запросу единицы текста (не важно, в оригинале или переводе[1]), а найдя, выдать **их и им соответствующие** единицы в парном документе (переводе или оригинале)*.

Возвращаясь к классификации в терминах задач поиска, справедливо назвать *Искалку* “внеязыковой” *системой синхронизованного поиска*. (Собственно словарная поддержка ограничивается в *Искалке* этапом построения индекса, где используются английский и русский стоп-словари для отбраковки слов, не подлежащих индексации.)

Те же принципы лежат в основе технологии *памяти переводов* (translation memory), используемой в таких системах как *Trados*, *MultiTerm* и *EuroLang*. Единицей синхронизации в этих системах является предложение (реже — фрагмент предложения). Дальнейшее измельчение единиц соответствия чревато утратой контекста, т.е. воспроизведением ситуации со словарем, сопоставляющим слово на одном языке словам на другом, но не дающем контекстной подсказки. Другой недостаток

[1] Не составит труда привести примеры, когда, напротив, важно, чтобы поиск производился только среди оригиналов или только среди переводов (синхронно отображаться при этом по-прежнему должны оба документа — и оригинал, и перевод). Однако, на практике, учитывая, что пересечение оригиналов и переводов чрезвычайно незначительно (только латинские названия), эти примеры, при всей их очевидности, совершенно несущественны.

установления соответствия на мелких единицах текста — невозможность выполнить его полностью или существенно автоматически, т.е. необходимость высокозатратной ручной разметки текстов.

3. Техническая реализация

Подчеркнем, что мы описываем реальные возможности и реально существующее устройство реально существующего программного продукта, первая версия которого была создана и начала активно использоваться еще в 1997 г.

3.1 Синхронизация оригинала и перевода

Формулировка задачи поддержки медицинского перевода как задачи *синхронизированного* поиска допускает разные технические решения. Общее у них то, что собственно поиск играет подчиненную роль, а объектами поиска и соответственно индексации являются *не отдельные документы, а именно* ***пары*** *оригинал—перевод*.

В качестве *единицы отображения контекста, или единицы синхронизации* (text aligning, стыковка или выравнивание, парных текстов, см. напр. [4, 5, 6]) был выбран *абзац*. Это гораздо более надежная единица как в плане сохранения контекста (то, что нужно, при переводе по прецедентам), так и в плане возможности установления соответствия автоматически или, по крайней мере, без существенных человеческих затрат. Затраты, требующиеся для более детальной разметки (на уровне предложений, а тем более на уровне составных терминов), не только (*а*) очевидно выше *на много порядков* (возможности автоматизации на этом уровне в случае медицинской терминологии очень ограничены — см. 2.3) и (*б*) *менее эффективны по результатам* (даже при минимальной автоматизации), но, что более важно, (*в*) по крайней мере в рассматриваемой задаче поддержки перевода, *совершенно бессмысленны*, т.к. переводчику все равно потребуется тот же контекстный уровень.

Понятно, что для обеспечения парности достаточно для каждой пары таких файлов создать третий файл — таблицу соответствия единиц контекста (абзацев), на чем основана процедура синхронизации, используемая при построении корпусов параллельных текстов [5, 6]. Однако более простое, прозрачное и в некотором смысле более экстремистское решение состоит в том, чтобы вместо виртуального документа, которому соответствует тройка физических (оригинал, перевод, таблица соответствий) пользоваться одним физическим документом-склейкой, содержащим слитые по абзацному соответствию оригинал и перевод плюс разметочные маркеры. Т.е. производится буквальное слияние парных текстов. Это решение и было применено Д.В. Самойловым в блоке, поддерживающем стыковку парных текстов (блок *проИскалка*).

Для каждого переведенного текста строится файл-склейка, состоящий из *блоков*, в каждом из которых абзацу оригинала сопоставлены соответствующие ему абзацы перевода. Хотя почти всегда абзацу оригинала соответствует абзац перевода, понятно, один абзац оригинала может при переводе превратиться в несколько; несколько абзацев оригинала — слиться в один; а некоторые абзацы вообще выпасть (Рис. 7).

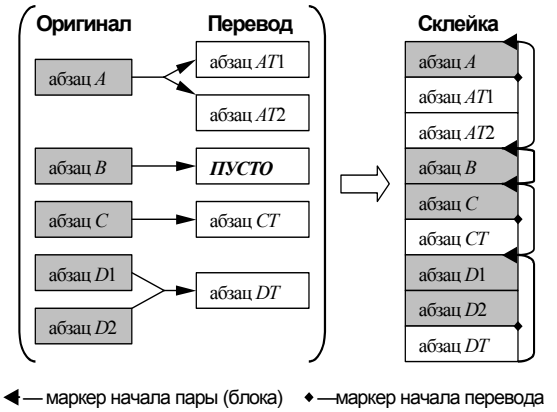

Рис. 7. Слияние оригинала и перевода

Поэтому при создании файла-склейки необходим контроль за правильностью соответствий. Доступны два режима контроля: ручной и автоматический. В обоих случаях при обнаружении расхождений абзацной структуры оригинала и перевода коррекция текста-склейки производится вручную (простым перетаскиванием абзаца из одного блока в другой).

Автоматический контроль можно выполнять на любом уровне — внутри целиком всей книги, внутри отдельной главы или ее раздела. Понятно, чем ниже структурный уровень, на котором выполняется контроль, тем точнее и надежнее локализация межабзацных расхождений. Действительно, зная лишь, что число абзацев оригинала отлично от числа абзацев перевода, нам, чтобы отыскать расхождение, придется вручную просмотреть оригинал и перевод с самого начала. Зная же, что число абзацев различно в некоторой главе, мы и ограничимся сравнением только этой главы.

Структуризация текста не представляет существенной проблемы, поскольку в *doc*-файлах содержится информация об уровне заголовков, а сверх того сами заголовки (в любом файле) пронумерованы. Тем самым, можно автоматически фрагментировать парные файлы, сопоставить фрагменты перевода фрагментам оригинала и проверить совпадение числа абзацев для парных фрагментов. Синхронизация (стыковка) парных текстов выполняется компонентой *проИскалка*, соответственно в режимах автоматической и полуавтоматической стыковки.

Замечание о переносимости *Искалки*. Автору известно о многочисленных случаях “нецелевого

использования” *Искалки* ее создателем, например, для представления параллельных корпусов поэтических текстов. Разумеется, простота использования *Искалки*, прежде всего блока синхронизации, делает ее чрезвычайно привлекательной в плане применения и за границами поддержки научного, прежде всего медицинского, перевода. Однако здесь существует два препятствия, одно юридическое — правообладателем обеих компонент *Искалки* на сегодня является исключительно издательство, в котором работал Д.В. Самойлов. Другое препятствие содержательное. *Искалка* строилась под конкретную потребность — поддержку прецедентного перевода *медицинских* текстов, где избранный уровень синхронизации параллельных текстов — по абзацам — не является компромиссом между затратами и точностью локализации. Он не только практически обнуляет затраты на синхронизацию, но, в данном случае, нисколько не идет в ущерб качеству локализации — более низкий уровень синхронизации парных конструкций при переводе медицинских текстов скорее вреден. Однако, в какой мере синхронизация по абзацам приемлема в прочих возможных применениях *Искалки*, очевидно зависит от области применения.

3.2 Индексация

Индекс строится по файлу-склейке. Собственно словарная обработка при этом вообще не выполняется, а синтаксическая ограничивается снятием разделителей (скобки, кавычки, запятые и т.д.). Выделенные таким образом слова попадают в индекс в той форме, как они встретились в тексте. Приведение к нормальной форме или стемминг не используются. Применяется только (очень небольшой) стоп-словарь, покрывающий служебные слова, например, отбраковывающий артикли.

Это наиболее простое решение, а не результат “многоязыкости” (она минимальна, “иностранный язык” — это английский с вкраплениями латыни, а “русский” — это “биомедицинский русский”) или способ избежать множественности и вариантности нормализаций встреченного слова, а равно и неполноты индексации как следствия неполного учета этой вариантности.

Описатель каждого вхождения индексируемого слова представляет собой пару <*документ-склейка*, *содержащий это слово блок этой склейки*>.

3.3 Поиск

Результатом такого простого решения является то, что при поиске пользователь вынужден активно использовать оператор “звездочка”, покрывающий флективные вариации. Этого достаточно, поскольку биомедицинская терминология не содержит неправильных глаголов и их производных.

На Рис. 8 приведен скриншот *Искалки*. При пользовании ею сразу обнаруживаешь, что она чрезвычайно эргономична: в ней нет лишних кнопок и окон, отображается все нужное и не отображается ничего лишнего. Физически работа производится с одним файлом-склейкой, *Искалка* содержит два окна отображения документа. В одном окне отображается оригинал, а в другом — перевод. Листание в одном окне приведет к синхронному автоматическому отображению соответствующего фрагмента парного текста.

При поиске найденное слово маркируется (значком конца абзаца — ¶) в том окне (оригинала или перевода), к которому относится соответствующая часть блока склейки.

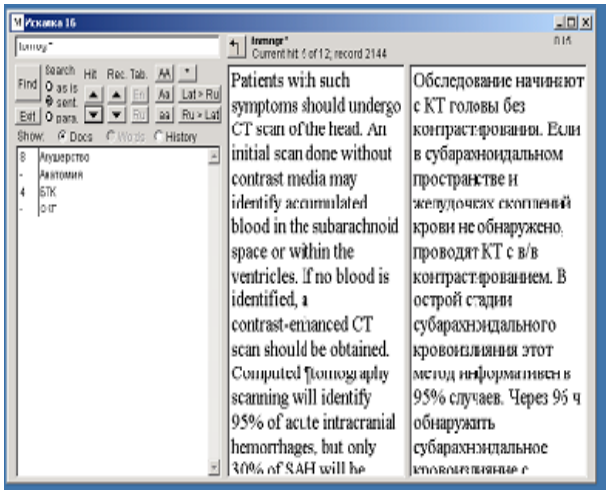

Рис. 8. Скриншот *Искалки*

Искалка — инструмент синхронизованного поиска. В таком качестве она полезна, когда нужно искать в корпусе *параллельных* текстов. Прежде всего, это корпус из оригиналов и переводов. Помимо промышленной эксплуатации *Искалки* в медицинском переводческом издательстве нам известно, например, об *Искалках*, которыми пользуются при работе с переводами поэзии. Но, что более существенно, *Искалка* оказалась нужна и полезна людям, реально занятым медицинской наукой.

4. Заключение

По образованию Д.В. Самойлов — врач. При достаточно скромной практике, он был диагностом от бога. Постановка точного диагноза сродни адекватной постановке задачи. Следствием точной и нетривиальной формулировки задачи, как и нетривиального диагноза, оказывается результативное и часто простое лечение.

Точно также *Искалка*, по эффективности не уступающая промышленным системам, а по простоте и надежности использования (в первую очередь, синхронизации параллельных текстов в *проИскалке*) — превосходящая, являет собой блестящий пример того, как красивая и самостоятельная постановка задачи привела к максимально эффективному и одновременно технически простому решению.

5. Благодарности

Автор считает своим долгом выразить признательность родителям Д.В. Самойлова — Марии Михайловне Каверзневой и Вадиму Ивановичу Самойлову.

Автор благодарен чл.-корр. РАН В.Л. Арлазарову, д.м.н. Н.В. Пономаревой, И.В. Сегаловичу, к.м.н. А.Е. Богорад и А.В. Турусову за ценные замечания, касающиеся разных аспектов статьи.

Литература

[1] Англо-русский медицинский словарь. М. “Русский язык”, 1989. (см. также электронные версии англо-русских медицинских словарей)

[2] Д.В. Самойлов. О переводе медицинского текста. (эта работа имеется практически на всех переводческих сайтах, в частности,. http://practica.ru/Articles/medical.htm)

[3] Н.Е. Бузикашвили, Д.В. Самойлов, Л.И. Бродский. Задача поиска в неструктурированном тексте и лингвистический анализ. *Интеллектуальные технологии ввода и обработки информации*, ИСА РАН, 96-104, 1998

[4] А.Ф. Гельбух, Г.О. Сидоров, А. Вера-Феликс. Словари в задачах автоматической обработки пар переводных текстов. *Диалог-2006*, 110-114. 2006.

[5] М.Н. Михайлов. Черная кошка в темной комнате или Можно ли автоматизировать поиск переводных эквивалентов в параллельном корпусе текстов? *Алфавит: Филологический сборник*. Смоленск, 181-188, 2002.

[6] A. Gelbukh, G. Sidorov. Alignment of Paragraphs in Bilingual Texts using Bilingual Dictionaries and Dynamic Programming. *CIARP-2006*, LNCS 4225, 824-833, Springer 2006

[7] Ch. Fluhr. Multilingual information retrieval. Survey of the State of the Art in Human Language Technology, 291-305, 1995

[8] D. Oard, B. Dorr. A survey of multilingual text retrieval. 31 pages, CS-TR-3615, 1996

Dmitry Samoilov’s *Iskalka*

Nikolai Buzikashvili

The paper describes elegant, highly effective and easy-to-use text aligning and information retrieval solutions of the machine-aided translation completely developed by Dmitry Samoilov (1958-2005).