УДК 821.161.1 + 81’322.4
ББК Ш141.2 + Ш11
DOI: 10.14529/ling160209

О РАЗРАБОТКЕ ИНСТРУМЕНТАРИЯ ДЛЯ ПОДДЕРЖКИ НАУЧНОГО ПИСЬМА

С.О. Шереметьева
Южно-Уральский государственный университет, г. Челябинск

Подчеркивается необходимость создания электронных инструментов для повышения качества научных текстов. Определяются факторы, препятствующие пониманию и переводу (удобочитаемости) документов человеком. Особое внимание уделяется пригодности научных текстов для машинного перевода (переводимости) в связи со все более широким его использованием, несмотря на возникающие при этом проблемы. Описывается методика решения этих проблем с помощью компьютерного предредактирования. Под компьютерным предредактированием понимается интерактивная процедура, которая оповещает автора о проблематичных фрагментах и контролирует лингвистические параметры текста, повышая его удобочитаемость и/или переводимость. На основании разработанной методики создан компьютерный инструмент для русско-английской пары языков и научных статей по машиностроению. Описанный подход может быть перенесен на другие языки и предметные области.

Ключевые слова: научный текст, компьютерный инструмент редактирования, удобочитаемость, переводимость.

Введение

Научные тексты играют важную роль в распространении и ассимиляции научно-технических знаний. При этом для эффективного выполнения своих функций лингвистические параметры текстов должны обеспечивать для адресной аудитории адекватное и, по возможности, беспроблемное понимание изложенного содержания текста на родном и иностранных языках, т. е. научные тексты должны обладать высокой «удобочитаемостью» и «переводимостью».

Удобочитаемость связана с уровнем ясности текста для понимания их человеком [1, 2]. Последнее определяется формулами удобочитаемости, наиболее популярные из которых предложены в [3]. Алгоритмы, измеряющие текстовую удобочитаемость, чаще всего в качестве параметров (индикаторов удобочитаемости) используют длины слов и предложений [4, 6, 7].

Индикаторы переводимости – это графические, лексические и синтаксические элементы текста, провоцирующие ошибки при машинном переводе. Как таковое понятие «Индикатор переводимости» появилось в области исследований машинного перевода (МП) [9], который в связи с постоянно растущим объемом научно-информационного обмена в международном масштабе все более широко используется как авторами научных статей, так и переводчиками, несмотря на его недостаточно высокое качество существующих систем.

Осознание факта, что определенные элементы текста могут отрицательно влиять на качество его машинного перевода на другие языки, определили два основных направления исследований в этой области.

В рамках первого направления выявление индикаторов переводимости является основой для их количественной оценки и выводов о пригодности или непригодности текста для машинного перевода [9].

В рамках второго подхода вырабатываются рекомендации по ограничению языковых средств (контролируемого языка), которые позволительно использовать в языковой структуре документов, чтобы сделать их более «удобными» для машинного перевода [5, 8, 11].

Реализацию таких рекомендаций на этапе создания текста или его предредактирования, как правило, предлагается осуществлять вручную. Тем не менее, авторов статей практически невозможно заставить писать на контролируемом языке, и ручное редактирование по предложенным правилам также остается проблематичным. Очевидно, что для повышения удобочитаемости и переводимости научных текстов в помощь авторам и переводчикам необходимо разрабатывать электронные инструменты, автоматизирующие решение указанных проблем, а также поддерживающие перевод терминологии и других лексических единиц текстов конкретной предметной области.

В статье предлагается один из возможных подходов к разработке компьютерного инструмента для повышения удобочитаемости и переводимости научных статей. Подход иллюстрируется на материале разработки инструмента для русскоязычных статей по машиностроению.

1. Постановка задачи

Цель нашего исследования – создать методологию для разработки компьютерного инструмента, который поможет адресной аудитории увидеть и исправить проблематичные для понимания и/или перевода фрагменты текста.

Под адресной аудиторией мы имеем в виду

две категории пользователей, активно использующих машинный перевод: ученых, работающих в соответствующих отраслях науки, и переводчиков, которые ответственны за адекватное понимание научного документа на иностранном языке.

Научные тексты часто чрезвычайно трудны для понимания обеими категориями читателей не только из-за изобилия терминологии, но и вследствие сложного, часто неоднозначного синтаксиса. Последнее особенно проблематично для переводчиков, которые, как правило, не обладают знаниями в области научных исследований, что могло бы им помочь в понимании изложенного.

В задачу нашего исследования входила также практическая реализация разработанной методологии в виде электронного ресурса для обработки русскоязычных статей по машиностроению для русско-английского направления перевода.

Поставленные цели были достигнуты через решение следующих задач:

- определение индикаторов удобочитаемости русскоязычных статей по машиностроению;
- определение индикаторов переводимости статей данной предметной области с русского языка на английский;
- разработка контролируемого языка, позволяющего избежать наличия индикаторов переводимости в предназначенных для машинного перевода текстах;
- имплементация электронного инструмента, повышающего удобочитаемость и переводимость научных документов и поддерживающего перевод одно- и многокомпонентных терминов и других лексических единиц.

Индикаторы удобочитаемости коррелируют, но не совпадают с индикаторами переводимости. Удобочитаемость облегчает понимание текста и его перевод для человека-переводчика, но не гарантирует хорошего МП и, как отмечается в [8], набор правил, предназначенных предотвратить в текстах появление индикаторов удобочитаемости, является не более чем подмножеством набора правил, устраняющих индикаторы переводимости.

В литературе отмечается наличие двух типов индикаторов переводимости:

1. *Универсальные*, вызывающие трудности при использовании любых систем машинного перевода для любых языков; к числу таких индикаторов переводимости относят, как правило, длинные предложения, все виды неоднозначности, координацию, телескопические синтаксические структуры, дистантные зависимости;

2. *Специфические,* которые зависят от конкретных языков, направления перевода, типа текста и конкретной системы машинного перевода.

В настоящей работе мы выявляем лингвистические индикаторы переводимости текстов исследуемой предметной области с русского языка на английский, вызывающие ошибки при переводе с помощью широко используемой в настоящее время системы МП, доступной онлайн: Google Translate. При этом мы исключаем из рассмотрения и анализа графические индикаторы, которые легко устранить с помощью стоп-листов формальных символов; ошибки правописания, которые корректируются существующими системами проверки правописания, а также покрываемость словаря системы МП Google Translate и некорректный перевод терминологии в рамках этой системы, поскольку система Google Translate не предназначена для перевода текстов по машиностроению.

Мы акцентируем наше внимание, прежде всего, на тех лингвистических индикаторах переводимости, которые могут быть устранены непосредственно *пользователями* систем МП. Такие индикаторы, как правило, обусловлены сложностью и многозначностью синтаксической структуры предметной области, различиями синтаксического строя русского и английского языков и особенностями архитектуры систем МП Google Translate.

Что касается устранения проблем при переводе терминологии, мы интегрируем в нашу систему лексикографический компонент, содержащий актуальную терминологию по машиностроению.

2. Эксперимент

Чтобы оценить трудности с пониманием (удобочитаемостью) текстов в указанной предметной области, автор этой статьи, обладая профессиональной квалификацией переводчика, проанализировал корпус 120 научных статей, опубликованных в журнале «Вестник ЮУрГУ», серия «Машиностроение» за 2010–2014 гг. объемом 203 729 словоформ. Было выявлено, что основной проблемой, делающей научный документ неудобочитаемым, являлся сплошной («слепой») текст в виде цепочки длинных предложений сложной синтаксической структуры, затрудняющий идентификацию границ лексических групп и определение типов синтаксических зависимостей. В дополнение к указанному научный текст был неудобочитаем, если в нем присутствовали предложения большой длины, дистантные зависимости, телескопические синтаксические структуры, длинные причастные обороты в препозиции к именной группе, неоднозначность аттрибуции предложных групп к именным или глагольным группам, неоднозначность сочинительных структур, грамматические ошибки авторов при согласовании, а также неединообразная терминология.

Консультации с исследователями в области машиностроения показали, что перечисленные выше явления снижают удобочитаемость текстов не только для переводчиков, но и для специалистов в области машиностроения, а визуализация компонентов предложения значительно повышает удобочитаемость текстов для этой категории пользователей.

В нашей работе выявление индикаторов переводимости текстов по машиностроению основано на экспериментальном исследовании, которое проведено следующим образом.

На первом этапе эксперимента был сформирован корпус русских предложений из статей по машиностроению, опубликованных в журнале «Вестник ЮУрГУ» (см. выше), содержащих:

- фрагменты с индикаторами удобочитаемости;
- фрагменты с универсальными индикаторами переводимости;
- фрагменты с «кандидатами» в специфические индикаторы переводимости, выделенными на основе анализа предметной области машиностроения;
- предложения, выбранные случайным образом из указанного корпуса.

На втором этапе отобранные предложения были переведены наиболее широко используемой системой МП Google Translate. Анализ результатов эксперимента показал, что выявленные индикаторы переводимости (вполне ожидаемо) включают практически все индикаторы удобочитаемости, перечисленные выше. Что касается остальных индикаторов, то многие лингвистические явления, традиционно причисляемые к индикаторам переводимости, в нашем случае являются таковыми только с существенными ограничениями. Например, пресловутая длина предложения не всегда отрицательно влияет на качество перевода: сколь угодно длинное предложение хорошо переводится автоматически (при условии полной покрываемости лексикона), если оно не содержит вставленных структур и/или дистантных зависимостей, но состоит из последовательно расположенных сочиненных или подчиненных предложений, которые, в свою очередь, могут содержать последовательно перечисленные однородные члены и/или причастные обороты с причастиями в постпозиции к определяемому существительному. Тем не менее, осознавая, что полную покрываемость лексикона даже для ограниченной предметной области обеспечить невозможно, мы считаем большую длину предложения индикатором переводимости. Ограниченным является и отрицательное влияние на МП координации и эллипсиса. Например, такие предложения как «*решение может быть численным или приближенным аналитическим»* не представляют проблем при МП: «*solution can be numerical or approximate analytical»*. Эллипсис становится индикатором переводимости в сочиненных конструкциях, если опускаются дистантно расположенные повторяющиеся глагол и именная группа в постпозиции к прилагательному: *«Интегральный метод дает хорошие результаты при его применении, а дифференциальный — плохие» -->The integral method gives good results if it is applied, and *the differential – bad.*

В процессе эксперимента выделен достаточно длинный список индикаторов переводимости для указанной предметной области, но перечислить все из них в данной статье не представляется возможным из-за ограничений на ее размер.

На основе выделенных в процессе эксперимента индикаторов удобочитаемости и переводимости разработаны правила контролируемого языка, которые ориентированы на повышение качества русских текстов и их машинного перевода на английский язык.

3. Контролируемый язык

Ниже мы приводим правила контролируемого языка, соблюдение которых авторами и переводчиками при создании или предредактировании научных текстов по машиностроению, повышает удобочитаемость таких текстов и позволяет избежать значительного числа ошибок при АП с русского языка на английский. Эти правила могут применяться как в «ручном» режиме, так и быть инкорпорированы в электронный инструмент, позволяющий автоматизировать решение проблем удобочитаемости и переводимости.

Разработанный контролируемый язык содержит две группы правил. Первая группа предполагает коррекцию всей структуры предложения, вторая группа правил носит более локальный характер и рекомендует коррекцию отдельных элементов предложения. В частности, правила контролируемого языка формулируются следующим образом:

1. Предложения длиной более 20 слов разбивать на несколько более простых предложений
2. В неполных предложениях восстанавливать пропущенное подлежащее и/или сказуемое. Всякое предложение должно быть полным.
3. Порядок членов предложения должен соответствовать порядку членов предложения в английском языке (сказуемое после подлежащего).
4. Предложения, содержащие множественную сочинительную связь (например, несколько союзов «и»), относящуюся к разным частям речи, разбивать на несколько более простых предложений.
5. Предложения с множественной сочинительной связью дистантно расположенных членов предложения разбивать на более простые.
6. Предложения с телескопическими конструкциями (например, вставленными причастными оборотами) преобразовывать в сложноподчиненные предложения с союзами «который», «что» и/или разбивать на несколько более простых предложений.
7. Безличные предложения трансформировать в предложения с явно выраженным подлежащим и сказуемым в личной форме.
8. Восстанавливать эллипсис дистантных членов предложения.

9. Трансформировать предложения с синтаксической омонимией так, чтобы структура предложения обеспечивала только одно понимание.

10. Определение-причастный оборот ставить после определяемого слова.

11. Определение-прилагательное ставить перед определяемым словом.

12. Если в предложении между компонентами составных сказуемых стоят слова или выражения, вынести их за пределы сказуемых.

13. Перед именными группами в творительном падеже использовать по возможности предлоги «с помощью», «посредством».

14. Перед именными группами ставить указательные местоимения или определения, например «этот», «наш» «указанный» и т. д., в том случае, если при их переводе должен использоваться определенный артикль.

15. Вместо местоимений использовать полнозначные слова.

16. Заменять многозначные слова и выражения на их однозначные синонимы.

17. Использовать унифицированную терминологию.

Например, в научных текстах по машиностроению часто можно встретить предложения с такого типа структурой:

«Изучен общий для аэродинамических компоновок с хвостовым коническим стабилизатором механизм возникновения минимума сопротивления».

Это предложение содержит такие индикаторы удобочитаемости и переводимости, как дистантные зависимости, телескопические синтаксические структуры, обратный порядок слов (сказуемое перед подлежащим), что делает предложение неудобочитаемым и плохо переводимым системами МП. В частности, машинный перевод этого предложения системой Google Translate некорректен и выглядит следующим образом:

« Studied for the overall aerodynamic configurations with tapered tail stabilizer mechanism of occurrence of low resistance».*

После применения правил контролируемого языка вышеприведенный русский текст выглядит следующим образом:

«Мы изучили механизм возникновения минимума сопротивления, который является общим для аэродинамических компоновок с хвостовым коническим стабилизатором».

Этот фрагмент удобочитаем и позволяет получить МП хорошего качества:

«We have studied the mechanism of occurrence of minimum resistance, which is common for aerodynamic configurations with a conical tail stabilizer»

4. Инструмент InterWrite

В разделе 3 мы привели пример применения правил контролируемого языка для повышения удобочитаемости и переводимости текста в «ручном» режиме. Ниже мы даем описание компьютерного инструмента InterWrite, автоматизирующего выполнение этих задач. Разработка этого инструмента является продолжением работ в области автоматической обработки текста, проводимой в НОЦ ЛИНТ ЮУрГУ[1].

InterWrite, представляет собой программный продукт, который повторно использует адаптированные для новой задачи алгоритмы автоматической обработки текстов и программную оболочку системы, описанной в [10].

В инструмент инкорпорированы правила контролируемого языка, описанные в разделе 4, а также русско-английская база знаний, ориентированная на предметную область «Машиностроение». База знаний поддерживает анализ научных текстов по машиностроению и предредакцию предложений текста на контролируемый язык для повышения его переводимости на английский язык. Преимуществами вновь созданной базы знаний является то, что она актуальна, разработана на основе новой методологии с применением интернет-ресурсов [11] и содержит

- русско-английские эквиваленты терминов и других лексических единиц длиной до 10 компонентов;
- морфологическую, синтаксическую, семантическую и контрастивную информацию о функционировании лексем в данном типе научных текстов;
- информацию о синтаксических ограничениях контролируемого языка.

Инструмент InterWrite принимает на входе научные тексты по машиностроению и автоматически анализирует его синтаксическую структуру. Результаты анализа визуализируются на левой панели основного окна пользовательского интерфейса в виде интерактивного текста с четкой разметкой: именная терминология выделена жирным черным шрифтом, а глагольная – синим, что позволяет пользователю легко ориентироваться в синтаксической структуре текста и, таким образом, повышает его удобочитаемость. При этом в результате анализа все лексические единицы русского текста связываются с лексической базой инструмента, что позволяет пользователю путем двойного щелчка на предикатные лексемы вызывать из базы знаний инструмента шаблоны для трансформации проблематичных предложений входного текста в предложения на контролируемом языке. Фрагменты интерактивно размеченного текста могут автоматически переноситься в слоты шаблона и, в случае необходимости, редактироваться в соответствии с правилами контролируемого языка. Заполненные в результате интерактивной процедуры взаимодействия пользователя с

[1] Научно-образовательный центр «Лингвоинновационные технологии» Южно-Уральского государственного университета.

компьютером шаблоны подаются на вход автоматического генератора русских предложений на контролируемом языке.

Поле правой панели пользовательского интерфейса горизонтально разделено на две части. В верхней части показываются автоматически сгенерированные инструментом предложения текста повышенной удобочитаемости и переводимости, которые могут быть далее переведены с помощью системы МП Google Translate (или любой другой доступной пользователю системой).

В нижней части правой панели интерфейса показываются в алфавитном порядке автоматически сгенерированные русско-английские эквиваленты всех одно- и, что особенно важно, многокомпонентных (до 10 компонентов) терминов и других лексических единиц научного текста. Именная терминология на русском и английском языках представлена в форме единственного числа именительного падежа. Английские эквиваленты личных форм глаголов (в помощь переводчику) выдаются в соответствующем времени, числе, лице и залоге. Например, *исследованы → are investigated; найдена → is found.* Для причастий выдаются их переводы как в функции определений (*опирающийся → resting*), так и в личной форме (*опирающийся → опирается → rests*), что облегчает использование этих глаголов при переводе.

Пользовательский интерфейс снабжен достаточно большим количеством элементов управления (кнопок, меню, полей поиска и т. д.), что делает его удобным в использовании.

Заключение

Мы представили методологию разработки и инструмент для повышения удобочитаемости и переводимости научных текстов. Эффективность методологии обусловлена эмпирически построенным контролируемым языком, проблемно-ориентированной лексической базой знаний, а также адаптацией и повторным использованием ранее разработанных программных модулей, инкорпорированных в новый интерактивный компьютерный инструмент.

Литература/References

1. Allen D. A Study of the Role of Relative Clauses in the Simplification of News Texts for Learners of English. *System*. 2009. 37 (4), pp. 585–599.

2. Karpov N., Baranova J., Vitugin F. Single-sentence Readability Prediction in Russian in Analysis of Images, Social Networks and Texts. *Communications in Computer and Information Science*. 2014. Vol. 436, pp. 91–100.

3. Kincaid J.P., Fishburne R.P., Rogers R.L. & Chissom B.S. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel, Research Branch Report 8–75, Millington, TN, Naval Technical Training, Memphis, 1975. 75 p.

4. Krioni N., Nikin A., Filippova A. Automated System for Analysis of the Complexity of Educational Texts. *Manag. Soc. Econ. Syst*. 2008, 11, pp. 101–107.

5. Nyberg E., T Mitamura, D. Svoboda, J. Ko, K. Baker, J. Micher. 2003. An Integrated System for Source Language Checking, Analysis and Terminology management. *Proceedings of Machine Translation Summit IX*, 2003. September. New-Orleans.USA, pp. 75–83.

6. Oborneva I. Automatic Assessment of the Complexity of Educational Texts on the Basis of Statistical Parameters. 2006.

7. Pooneh Heydari and A. Mehdi Riazi. Readability of Texts: Human Evaluation Versus Computer Index. *Mediterranean Journal of Social Sciences*, 2012, January, Vol. 3 (1). pp. 104–112.

8. Reuther U. Two in one – Can it work? Readability and Translatability by Means of Controlled Language. *Proceedings of EAMT-CLAW03, Controlled Language Translation*. Dublin, 2003, pp. 124–132.

9. Underwood N.L. and Jongejan B. Translatability Checker: A Tool to Help Decide Whether to Use MT. *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain, 2001, pp. 125–133.

10. Sheremetyeva S. Controlled Authoring In A Hybrid Russian-English Machine Translation System. *Proceedings of the workshop "Third Workshop on Hybrid Approaches to Translation" in conjunction with the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden, 2014, pp. 78–86.

11. Sheremetyeva S. On Getting the Most out of Internet Resources to Raise Translation Quality Of Professional Documentation. *Journal of General and Professional Education*, iss. 4, 2012, pp. 94–102.

Шереметьева Светлана Олеговна, доктор филологических наук, профессор кафедры «Лингвистика и межкультурная коммуникация», Южно-Уральский государственный университет (Челябинск), sheremetevaso@susu.ru

Поступила в редакцию 1 марта 2016 г.

DOI: 10.14529/ling160209

ON DEVELOPING TOOLS FOR SUPPORTING SCIENTIFIC WRITING

***S.O. Sheremetyeva**, sheremetevaso@susu.ru*
South Ural State University, Chelyabinsk, Russian Federation

In the paper a strong need for effective computer support in professional writing is put into focus. Barriers to human understanding (readability) and translation of scientific texts are discussed. Special attention is paid to the suitability of a scientific text for machine translation (translatability), which is now widely used not withstanding its imperfect quality. A computer-based authoring methodology to attend these issues is described. Authoring is viewed as an interactive procedure that makes professionals aware of the typical areas of concern and controls the linguistic parameters of a document thus raising its readability and translatability. The methodology is implemented into a computer tool for research papers on engineering for the Russian-English language pair. The approach described can be applied to other languages and domains.

Keywords: scientific text, computer tool for authoring, readability, translatability.

Svetlana O. Sheremetyeva, Doctor of Philological Science, Professor of the Linguistics and Intercultural Communication Department, South Ural State University, sheremetevaso@susu.ru

Received 1 March 2016

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Шереметьева, С.О. О разработке инструментария для поддержки научного письма / С.О. Шереметьева // Вестник ЮУрГУ. Серия «Лингвистика». – 2016. – Т. 13, № 2. – С. 49–54. DOI: 10.14529/ling160209

FOR CITATION

Sheremetyeva S.O. On Developing Tools for Supporting Scientific Writing. *Bulletin of the South Ural State University. Ser. Linguistics*. 2016, vol. 13, no. 2, pp. 49–54. (in Russ.). DOI: 10.14529/ling160209