СИСТЕМА МАШИННОГО ПЕРЕВОДА TILDE TRANSLATOR: НОВЫЙ ЭТАП В РАЗВИТИИ АНГЛО / РУССКО-ЛАТЫШСКОГО МАШИННОГО ПЕРЕВОДА

MACHINE TRANSLATION SYSTEM TILDE TRANSLATOR: NEW STAGE OF ENGLISH / RUSSIAN-LATVIAN MACHINE TRANSLATION

Горностай Т. *(tatjana.gornostaja@tilde.lv)*
компания Tilde, Рига (www.tilde.com)

В докладе представлен лингвистический программный продукт – система машинного перевода Tilde Translator. Внимание уделено особенностям разрешения лексической неоднозначности, переводу устойчивых словосочетаний и употреблению семантических фильтров для улучшения качества машинного перевода в рассматриваемой системе.

1. Введение. Машинный перевод – роскошь или необходимость?

Несмотря на общепризнанный факт о невозможности полностью автоматизировать и алгоритмизировать перевод текста с одного естественного языка на другой [1, 3, 5, 7, 9, 10, 11], машинный перевод (МП) представляет собой перспективное научное и практическое направление с почти 200 действующих коммерческих систем [12], в частности систем, производящих перевод текстов на латышский язык: SmartTranslator, Pragma Lite, Tilde Translator. Однако даже сегодня нередко можно услышать вопрос о том, для чего нужны системы МП, и человек в современном информационном пространстве часто оказывается в ситуации, когда ему неизвестны возможности применения языковых технологий для решения различных задач. Задачей как научных исследователей, так и разработчиков систем МП является ознакомление пользователя любой категории с возможностями той или иной системы МП для продуктивного ее использования [8].

Ни одна из систем МП не способна удовлетворить интересы всех пользователей. Следуя современной тенденции в теории МП [14, 15, 16], все системы МП можно разделить на две большие группы: разработки, ориентированные на технологии МП, способные максимально облегчить работу переводчика-профессионала, и на создание программ, обеспечивающих межъязыковое общение для обычного пользователя. Система МП Tilde Translator относится ко второму типу систем. Рассмотрим подробнее особенности данной системы МП.

2. Система машинного перевода Tilde Translator

В рамках традиционной типологии [2, 4, 17], лингвистическая информационная система Tilde Translator может быть классифицирована как многоязычная, полностью автоматизированная система МП. Входной единицей перевода может являться как отдельное слово языка-источника, так и словосочетание или предложение, а выходной единицей – соответствующая структурная единица выходного языка. В первом случае система будет работать как резидентный электронный словарь, извлекая из словарной статьи перевод иностранного слова, а получая на входе единицу, большую, чем отдельное слово, – как собственно система МП.

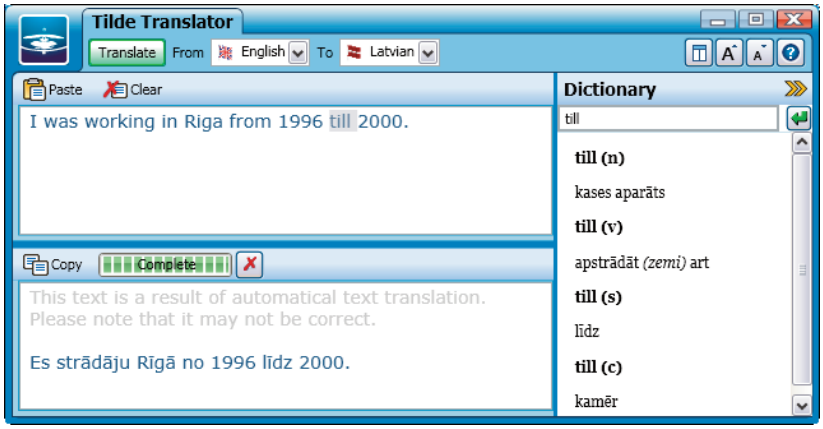

Рис. 1 Электронный словарь в системе МП Tilde Translator

Надо сказать, что использование большинства систем МП в качестве электронного словаря не представляется рациональным, так как система МП, как правило, предложит лишь один вариант перевода заданного слова, а для более полной информации об искомой словарной единице следует обратиться к специально созданным для данных целей электронным словарям. В программе Tilde Translator реализована возможность вызова электронного словаря, интегрированного в систему МП для удобства пользователя в том случае, если система МП предлагает некорректный перевод того или иного слова, или пользователю необходимо просмотреть всю словарную информацию о данном слове (рис. 1).

2.1. Архитектура системы

Сегодня принято считать, что нет единственного метода для достижения МП необходимого качества, и будущее за моделями так называемого ‘гибридного’ МП, объединяющего лучшие технологии, основанные на правилах, статистике и переводческой памяти [15, 16]. Система МП Tilde Translator создавалась на базе современных тенденций в теории и практике МП и является одним из примеров подобных технологий МП, в основу работы отдельных модулей которой заложены статистические методы (модуль разрешения лексической неоднозначности, например) и технология переводческой памяти (перевод устойчивых словосочетаний) наряду со стратегией трансфера, составляющей ядро системы..

Архитектура рассматриваемой системы МП основывается на том же принципе, что и архитектура электронного словаря с элементами МП SmartTranslator [6, 13], – принципе модульной архитектуры. Алгоритмы перевода представляют собой многоуровневый процесс, в самом общем виде состоящий из трех этапов: анализ структуры входной единицы языка-источника; преобразование данной структуры в аналогичную структуру выходного языка (собственно трансфер); синтез единицы выходного языка в соответствии с полученной структурой на втором этапе [18].

2.2. Разрешение лексической неоднозначности

В нашем предыдущем докладе [6] мы рассматривали применение статистических методов в электронном словаре с элементами МП SmartTranslator для разрешения лексической неоднозначности в процессе перевода на латышский и литовский языки, при котором учитывалась вероятностная синтаксическая модель данных языков, построенная на основе одноязычного корпуса. Недостатком упомянутой разработки являлось использование данных выходного языка и игнорирование языка-источника. Была поставлена задача усовершенствования алгоритма разрешения многозначности при МП, и за последний год данный модуль претерпел изменения благодаря новой стратегии, которая на данный момент реализована для англо-латышского МП. При описанном подходе используется двуязычный англо-латышский параллельный корпус, на котором проводилось обучение системы, и инструмент GIZA++ для сопоставления переводных эквивалентов (выравнивания корпуса). Модуль разрешения многозначности подключается на том этапе МП, когда уже построены синтаксические деревья для языка-источника и выходного языка, и приписаны переводные эквиваленты для лексических единиц языка-источника. Итак, мы имеем один или несколько переводных эквивалентов в выходном языке, соотнесенных с одной вершиной дерева языка-источника. Для всех переводных эквивалентов в выходном языке вычисляется их вероятность перевода и выбирается переводной эквивалент с максимальным значением, учитывая локальный контекст данной лексической единицы в тренировочном корпусе. В результате, мы имеем возможность разрешения лексической неоднозначности, выбирая тот переводной эквивалент, который обеспечивает наивысшую вероятность перевода данной лексической единицы языка-источника.

Известные недостатки и неточности данного метода на сегодняшний день: ошибки при анализе корпуса и оценке перевода (морфологическая неоднозначность, например); игнорирование некоторых единиц в процессе анализа корпуса (числительные, например); использование переводных эквивалентов только из автоматического словаря системы (возможно, что в словаре вообще нет перевода для данной лексической единицы, или же в корпусе содержится более подходящий вариант).

2.3. Перевод устойчивых словосочетаний

Система МП, в отличие от систем переводческой памяти, предлагающей перевод ‘по образцу’, позволяет получать синтаксически связный перевод текста, учитывающий морфологию как языка-источника, так и выходного языка [4]. Своего рода модификацией технологии переводческой памяти в рассматриваемой системе МП Tilde Translator является усовершенствованный модуль перевода устойчивых словосочетаний, который произво-

дит лингвистический синтез устойчивых словосочетаний с учетом морфологии выходного языка и используется в англо / русско-латышском МП.

С точки зрения МП, устойчивыми единицами являются как идиоматические выражения, фразеологизмы, терминологические единицы, так и составные союзы и предлоги, которые должны обрабатываться как единое целое. Причем ‘многокомпонентность’ устойчивого словосочетания в рассматриваемой системе является общим понятием для пары языков и не обязательным критерием для каждого языка, то есть, либо в языке-источнике, либо в выходном языке устойчивое словосочетание может быть выражено одним словом, например:

(1) авторское право – autortiesības,
(2) автозавод – automobiļu rūpnīca,
(3) to be late – nokavēties,
(4) fire and fury – liesmainums.

Основное изменение в алгоритме перевода устойчивых словосочетаний – возможность обработки компонентов внутри структуры словосочетания. Напомним, что в предыдущей разработке корректный перевод получали лишь фиксированные, с грамматической точки зрения, устойчивые словосочетания [6].

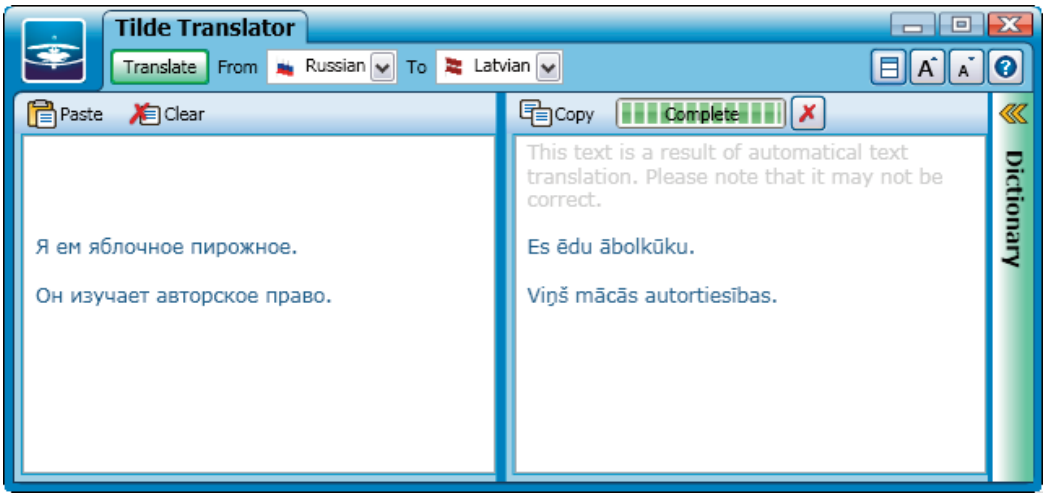

Рис. 2 Пример русско-латышского перевода устойчивых словосочетаний

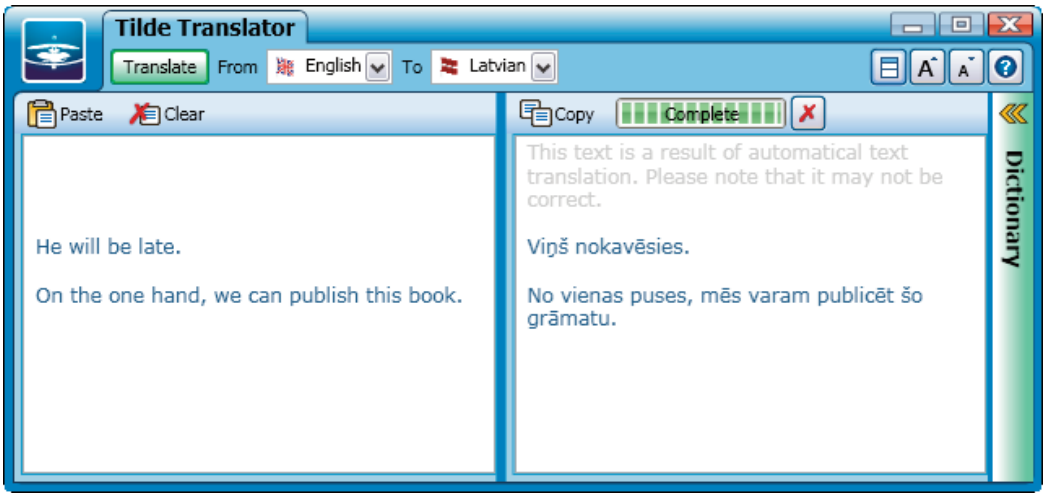

Рис. 3 Пример англо-латышского перевода устойчивых словосочетаний

Модуль обработки устойчивых словосочетаний представляет собой отдельный словарь, в котором хранятся устойчивые словарные единицы[1], а также набор разработанных правил трансфера, которые описывают, каким образом будет изменено синтаксическое дерево языка-источника при синтезе предложения выходного языка в случае несовпадения синтаксической структуры словарной единицы в языке-источнике и выходном языке. Проанализированное и переведенное устойчивое словосочетание интегрируется в синтаксическое дерево всего

[1] Около 25 тыс. словарных единиц для англо-латышского перевода, более 4 тыс. словарных единиц для русско-латышского перевода (в процессе пополнения).

предложения, которое посылается для дальнейшей обработки (трансфер, согласование и другие процессы). На рисунках 2 и 3 приведены примеры русско-латышского и англо-латышского перевода устойчивых словосочетаний.

2.4. Семантические фильтры

Одним из решений в системе МП Tilde Translator является также использование семантических фильтров для лексических единиц языка-источника там, где это необходимо для улучшения качества МП. Лексические единицы могут быть отнесены к той или иной семантической категории. Категории не являются уникальными для данного направления и могут использоваться при переводе на другой язык.

Рассмотрим процесс установления семантического фильтра на примере синонимического ряда *say, tell, talk, speak* в английском языке. Принадлежность к одному синонимическому ряду не означает то, что все данные лексические единицы будут отнесены к одной семантической категории, так как мы используем семантические фильтры для описания структурных типов и конструкций, и не всегда лексические единицы с синонимичным значением будут образовывать однотипные синтаксические конструкции. Так, глаголы *talk* и *speak* отнесены к категории ‘communication’ на основе структурного сходства синтаксических конструкций, в которых они употребляются:
может употребляться без объекта (пример 5),
употребляются с предлогами *about* и *of* (пример 6, 7),
употребляются с предлогом *to* (пример 8).

(5) to be learning to talk / speak,
(6) to talk about / of something,
(7) to speak about / of something,
(8) to talk / speak to somebody.

Глаголы *say* и *tell* объединены в категорию ‘speech_declarative’ на основании следующих особенностей:
употребляются в предложениях с прямой речью (пример 9),
употребляются в предложениях с косвенной речью (пример 10, 11),
употребляются в конструкциях с пассивным залогом (пример 12, 13).

(9) “Hello,” somebody said / told,
(10) to say that / wh- word,
(11) to tell that / wh- word,
(12) something / somebody is said,
(13) something / somebody is told.

Рассмотрим примеры использования семантических фильтров при МП .

2.4.1. Применение семантических фильтров

Семантические фильтры используются на стадиях трансфера и согласования. В синтаксис трансформационных правил и правил согласования введен атрибут *KindOf* для установления принадлежности вершины дерева к семантической категории. На рисунке 4 приведен пример трансформационного правила, которое будет исполнено, при условии принадлежности имени существительного к семантической категории ‘time’.

```
TransferRule(PREP<-pcomp-N)  //pirms gada
{
        Parent.SourceSpelling  == "pirms";
        Child.KindOf  == "time";
        Parent.TargetSpelling  = "назад";
        Parent.POS  = ADV;
        Child.Case  = accusative;
        move_to_left(Child,Parent);
        MakeLink(Child-pcomp->Parent);
}
```

Рис. 4 Пример трансформационного правила с использованием семантического фильтра

При переводе предлогов, например, учитывается информация о принадлежности лексической единицы, употребленной с предлогом, к той или иной семантической категории. Английский предлог *to*, например, употребленный с глаголом с семантическим фильтром ‘motion’ переводится на латышский язык предлогом *uz* (примеры 14, 15), а с семантическим фильтром ‘communication’ – предлогом *ar* (примеры 16, 17):

(14) to go to work – iet uz darbu,
(15) to take to a town – ņemt uz pilsētu,
(16) to talk to John – runāt ar Džonu,
(17) to speak to him – runāt ar viņu.

Этот же предлог, употребленный с глаголом с семантическим фильтром ‘speech_declarative’, не переводится на латышский язык, а косвенное дополнение будет синтезировано в форме дательного падежа, например:

(18) to say to somebody – sacīt kādam.

Для перевода именных предложных конструкций также учитывается семантическая информация о лексических единицах. Так, например, английский предлог *about* с существительным, принадлежащим к семантической категории ‘time’ или ‘period’, переводится на латышский язык предлогом *aptuveni*, в остальных же случаях – предлогом *par* (пример 19, 20):

(19) to work about five hours – strādāt aptuveni piecas stundas,
(20) to speak about something – runāt par kaut ko.

Семантические фильтры используются также для латышско-русского МП. Так, например, латышский предлог *pirms* может переводиться на русский язык предлогом *до* или наречием *назад* (примеры 21, 22), а предлог *pēc* – предлогами *через* и *после* (примеры 23, 24):

(21) pirms lekcijas – до лекции,
(22) pirms desmit gadiem – десять лет назад,
(23) pēc diviem gadiem – через два года,
(24) pēc darba – после работы.

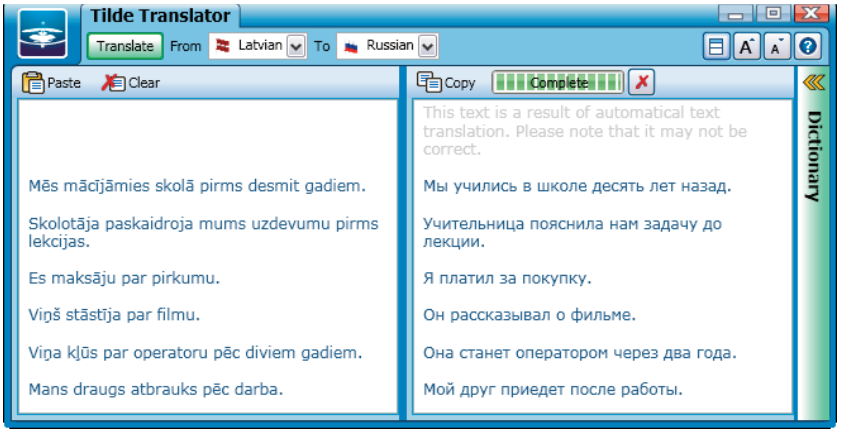

Рис. 5 Пример латышско-русского перевода с использованием семантических фильтров.

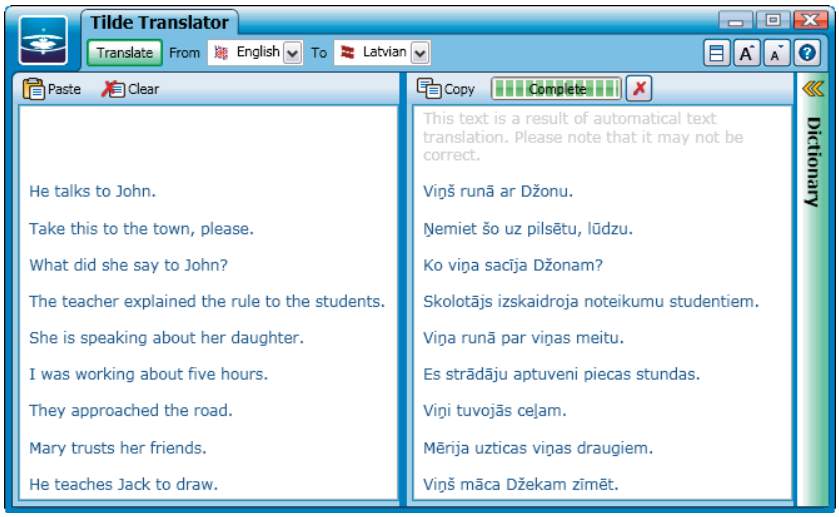

Рис. 6 Пример англо-латышского перевода с использованием семантических фильтров.

Аналогично семантические фильтры используются также и для корректного перевода беспредложных глагольных конструкций, учитывая валентную структуру глагола в выходном языке (примеры 25, 26, 27):

(25) to approach the road – tuvoties ceļam,

(26) to trust somebody – uzticēties kādam,

(27) to teach somebody – to mācīt kādam.

Наряду со статистическим методом разрешения лексической неоднозначности используются и семантические фильтры. Приведем классический пример из латышско-русского перевода с использованием семантических фильтров ‘publication’ и ‘food_plant’ (примеры 28, 29). Рисунки 5 и 6 демонстрируют примеры латышско-русского и англо-латышского МП с использованием семантических фильтров.

(28) lasīt grāmatu – читать книгу,

(29) lasīt ogas – собирать ягоды.

4. Результаты и планы

Для оценки результатов МП системы Tilde Translator с использованием BLEU и NIST метрик автоматического измерения качества МП, полученного после внесения описанных изменений в систему, были составлены тестовые корпусы. Под ‘тестовым корпусом’ мы понимаем параллельный (выровненный по предложениям), не аннотированный (морфологически, синтаксически и лексически не маркированный), представительный и сбалансированный корпус текстов, который состоит из оригинальных предложений (500 единиц), то есть, корпус является естественным и достаточным для оценки качества МП данной системы.

Результаты для латышско-русского и англо-латышского направлений перевода обобщены в таблице 1. За базовый уровень взяты результаты оценки первой версии системы без введения семантических фильтров и нового подхода к разрешению лексической неоднозначности, а также переводу устойчивых словосочетаний.

Языковая пара	ЛАТ-РУС		АНГЛ-ЛАТ	
Метрика автоматического измерения качества МП	BLEU	NIST	BLEU	NIST
Базовый уровень	38,79	6,2762	10,87	3,2449
Последнее измерение (февраль 2008)	43,30	7,1857	21,15	5,3382

Таблица 1. Результаты автоматического измерения качества латышско-русского, англо-латышского МП

Полученные результаты позволяют говорить о положительных изменениях в системе с применением других технологий МП вместе со стратегией трансфера. Для русско-латышского направления автоматическая оценка качества МП проводилась на последней версии системы, полученные показатели: BLUE 17,98, NIST 4,22.

На будущее запланирована разработка модулей разрешения неоднозначности и перевода устойчивых словосочетаний для остальных направлений перевода, а также расширение автоматических словарей.

Автор доклада благодарит коллег Ингуну Скадиня, Дайгу Дексне и Карлиса Гоба за ценные замечания, высказанные в ходе написания статьи.

Список литературы

1. Беляева Л. Н. Проектирование лингвистических информационных баз для систем автоматической переработки текста // Статистика речи и автоматическая переработка текстов : межвуз . сб. науч. тр. / науч. ред. Р. Г. Пиотровский. – Л. : Ленингр. гос. пед. ин-т им. А. И. Герцена, 1988. – С. 8–28.

2. Беляева Л. Н. Автоматический (машинный) перевод / Л. Н. Беляева, М. И. Откупщикова // Прикладное языкознание : учебник / отв. ред. А. С. Герд ; С.-Петерб. гос. ун-т. – СПб. : Изд-во С.-Петерб. ун-та, 1996. – С. 360–388.

3. Беляева Л. Н. Лингвистические автоматы / Л. Н. Беляева, Р. Г. Пиотровский, В. Р. Нымм // Вестн. С.-Петерб. отд-ния Рос. акад. естеств. наук. – 1999. – [Т.] 3, № 1. – С. 73–82.

4. Беляева Л. Н. Лингвистические автоматы в современных информационных технологиях / Л. Н. Беляева ; Рос. гос. пед. ун-т им. А. И. Герцена. – СПб. : Изд-во Рос. гос. пед. ун-та им. А. И. Герцена, 2001. – 130 с.

5. Богуславский И. М., Иомдин Л. Л., Лазурский А. В., Митюшин Л. Г., Бердический А. С. Интерактивное разрешение неоднозначности различных типов в машинном переводе [Электронный ресурс] // Труды Междунар. конф. «Диалог’ 2005»: [сайт] / Ассоц. компьютер. лингвистики и интеллектуал. технологий. – М., 2005. – Режим доступа:
http://www.dialog-21.ru/Archive/2005/Iomdin%20Boguslavski%20Lazurski/Iomdin%20Boguslavski%20Lazurski.htm.

6. Горностай Т., Васильев А., Скадиня И., Скадиньш Р. Опыт латышско↔русского машинного перевода // ред. Л. Л. Иомдин, Н. И. Лауфер, А. С. Нариньяни, В. П.. Селегей. Труды Международной конференции Диалог' 2007 : Компьютерная лингвистика и интеллектуальные технологии. – М. : Издательский центр РГГУ, 2007. – С. 137-146.

7. Зубова И. И. Информационные технологии в лингвистике : Учеб. пособие / И. И. Зубова. – Минск : Изд-во Мин. гос. лингвист. ун-та, 2001. – 211 с.

8. Лингвистические ресурсы автоматизированного рабочего места филолога : Коллективная монография / Л. Н. Беляева, А. А. Виландеберк, Л. А. Деверь [и др.]. – СПб. : Изд-во ИнфоДа, 2004. – 184 с.

9. Пиотровский Р. Г. Инженерная лингвистика и теория языка / Р. Г. Пиотровский ; Акад. наук СССР, Науч. совет по теории сов. языкознания, Ин-т языкознания. – Л. : Наука, Ленингр. отд-ние, 1979. – 112 с.

10. Пиотровский Р. Г. Лингвистический автомат : (в исследовании и непрерыв. обучении) : Учеб. Пособие / Р. Г. Пиотровский ; Рос. гос. пед. ун-т им. А. И. Герцена. – СПб. : Изд-во Рос. гос. пед. ун-та им. А. И. Герцена, 1999. – 256 с.

11. Пиотровский Р. Г. Лингвистический автомат и его речемыслительное обоснование : учеб. Пособие / Р. Г. Пиотровский. – Минск : Изд-во Мин. гос. лингвист. ун-та, 1999. – 196 с.

12. Compendium of Translation Software [Electronic resource] : directory of commercial machine transl. systems and computer-aided transl. support tools / comp. by John Hutchins. – 14th ed. // EAMT : European Association for Machine Translation : [site]. – [Geneve], 2008. – Mode of access: http://www.eamt.org/compendium.html.

13. Deksne D., Skadiņa I., Skadiņš R., Vasiļjevs A. Foreign language reading tool – first step towards English-Latvian commercial machine translation system // Proceeding of the Second Baltic Conference on Human Language Technologies, Tallinn, 2005.

14. Hutchins W. J. Machine Translation and Human Translation : in Competition or in Complementation? // International Journal of Translation vol. 13, No 1-2, 2001, pp. 5-20.

15. Hutchins W. J. Towards a New Vision for MT [Electronic resource] : introductory speech… // Machine Translation Summit VIII, 18–22 Sept. 2001, Santiago de Compostela, Spain : papers and presentations / EAMT. – [Geneve], 2001. – Mode of access: http://www.eamt.org/summitVIII/papers/introduction.html.

16. Hutchins W. J. Current commercial machine translation systems and computer-based translation tools : system types and their uses // International Journal of Translation vol. 17, No 1-2, 2005, pp. 5-38.

17. Hutchins W. J. Machine translation : a concise history // To be published in Computer aided translation : Theory and practice, ed. Chan Sin Wai. Chinese University of Hong Kong, 2007.

18. Skadiņa I., Vasiļjevs A., Deksne D., Skadiņš R., Goldberga L. Comprehension Assistant for Languages of Baltic States // Nivre J., Kaalep H., Muischnek K., Koit M. (eds.) Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007. Tartu : University of Tartu, 2007, pp.167-174.