# **On-The-Fly Translator Assistant** (Readability and Terminology Handling)

Svetlana Sheremetyeva National Research South Ural State University / pr.Lenina 74, 454080 Chelyabinsk, Russia LanA Consulting ApS/ Moellekrog 4, Vejby, 3210, Copenhagen, Denmark lanaconsult@mail.dk

#### Abstract

This paper describes a new methodology for developing CAT tools that assist translators of technical and scientific texts by (i) on-the-fly highlight of nominal and verbal terminology in a source language (SL) document that lifts possible syntactic ambiguity and thus essentially raises the document readability and (ii) simultaneous translation of all SL document one- and multicomponent lexical units. The methodology is based on a language-independent hybrid extraction technique used for document analysis, and language-dependent shallow linguistic knowledge. It is targeted at intelligent output and computationally attractive properties. The approach is illustrated by its implementation into a CAT tool for the Russian-English language pair. Such tools can also be integrated into full MT systems.

#### 1 Introduction

Exploding volume of professional publications demand operative international exchange of scientific and technical information and thus put in focus operativeness and quality of translation services. In spite of the great progress of MT that saves translation time, required translation quality so far cannot be achieved without human judgment (Koehn, 2009). Therefore in great demand are CAT tools designed to support and facilitate human translation.

CAT tools are developed to automate postediting and often involve controlled language. The most popular tools are translation memory (TM) tools whose function is to save the translation units in a database so that they can be re-used through special "fuzzy search" features. The efficiency of TM (as well as translation quality as such) is directly related to the problem of the comprehensiveness of multilingual lexicons. A translator who, as a rule, does not possess enough of expert knowledge in a scientific or technological domain spends about 75% of time for translating terminology, which do not guarantee the correctness of translation equivalents she/he uses. The percentage of mistakes in translating professional terminology reaches 40% (Kudashev, 2007). It is therefore essential to develop methodologies that could help human translators solve this problem, the huge resource being the Internet, if properly used. In this paper we suggest one of the possible ways to do so.

We would like to address the importance of text readability in the human translation performance. Readability relates to (though does not coincide with) the notion of translatability in MT research. Readability in human translation is associated with the level of clarity of a SL text for human understanding. Every translator knows how difficult it can be to understand professional texts, not only because of the abundance of terminology but also due to complex syntax and syntactic ambiguity. The ultimate example of a low readability text is the patent claim (Shinmori et al., 2003) that is written in the form of one nominal sentence with "inhuman" extremely complex syntactic structure that can run for a page or more. Low readability is often the case with scientific and technical papers as well.

In this paper we describe our effort to develop a portable between domains and languages CAT tool that can on-the-fly improve the readability of professional texts and provide for reliable terminology translation.

We paid special attention to multiword noun terminology, the most frequent and important terminological unit in special texts that can rarely be found in full in existing lexicons. When translated properly, multicomponent NPs do not only provide for the correct understanding of the corresponding target language (TL) term but in many cases lift syntactic ambiguity.

The tool can find a broad application, e.g., it can be useful for any non-SL speaker for a quick document digest. The settings of the tool allow the extraction of keyword translation pairs in case it is needed, e.g., for search purposes. It can also be integrated into a full MT system.

We implemented our methodology into a fully functional tool for the Russian-English

language pair and conducted experiments for other domains and language pairs. In selecting Russian as a first SL we were motivated by two major considerations. Firstly, Russia has a huge pool of scientific and technical papers which are unavailable for non-Russian speakers without turning to expensive translation services. Secondly, our scientific challenge was to develop a hybrid methodology applicable to inflecting languages. Popular SMT and hybrid techniques working well configurational on and morphologically poor languages, such as English, fail on non-configurational languages with rich morphology (Sharoff, 2004). Russian is an ultimate example of such a language. It has a free word order; a typical Russian word has from 9 (for nouns) up to 50 forms (for verbs). In what follows we first present the tool and then describe the underlying methodology.

	Файл Помощь			
Ваше сторь 2 свидляенных теренности теренности теренности в ваш терении  Можать  ЧИСЛЕННОС РЕШЕНИЕ ЗАДАЧИ ОПТИМАЛЬНОГО УПРАВЛЕНИЯ ВИРОИДЕННОИ УЛИТЕРА СИДОРОВА Продложены апторита ликейной сисстемой начальных условнями шохолитера Сидорова нарожиени коздропния начальных условнями Шоуолтера Сидорова продложены апторитального управления начальных условнями Системой начальных условнями Коши Ключеные система начальных условнями Коши начальных				
Image:	🔲 🛩 📔 Русский 📉 Английски 🖳 Разден Математиче	eckoe M	юделир 📉	
Image:				
Image:				
Иколь           ЧИСЛЕННОЕ РЕШЕНИЕ ЗАДАЧИ           ЧИСЛЕННОЕ РЕШЕНИЕ ЗАДАЧИ           ОПТИМАЛЬНОГО УПРАВЛЕНИЯ           ВИРОЖДЕННОЙ ЛИНЕЯНОЙ СИСТЕМОЙ УРАВНЕНИЙ           С           НАЧАЛЬНЫМИ УСЛОВИЯМИ ШОУОЛТЕРА СИДОРОВА           Проелложены           алгоритм           и           Варожденной системой           оптимального управления           варожденной планейкой системой           обыковенных дикреренциальных уравнений           С           Постоянными коэффициентами           Коловенных дикреренциальных уравнений           С           Постоянными коэффициентами           Коновенных дикреренциальных уравнений           С           Постоянными коэффициентами           Коновенных дикреренциальных уравнений           Коновенных дикреренциальных уравнений           Коновенных дикреренциальных уравнений           Коновенный состемой           Проеловен           Постоянными коэффициентами           Коновенсого применения           Мапонькозмального управления           Постоянными коэффициентами           Коновенсого применения           Мололи           На           проекоре           Проекор	Ваша статья с выделенными терминами		Перевести Очистить окно	Ваштермин
численное Решение задачи            Оптимального управления задачи            инстенное порта            Инчальный колописации задачи            инеет вы            варальные условия Шкурлтера-Сидорова            Ала            произованы            прасширики систра практического применения            прасши	Искать			
ОПТИМАЛЬНОГО УПРАВЛЕНИЯ ВЫРОЖДЕННОЙ ЛИНЕЙНОЙ СИСТЕМОЙ УРАВНЕНИЙ С 4 ЧАЛЬНЫМИ УСЛОВИЯМИ ШОУОЛТЕРА СИДОРОВА Продложены алгориты HAЧАЛЬНЫМИ УСЛОВИЯМИ ШОУОЛТЕРА СИДОРОВА Продложены алгориты начального управления адагориты начального управления адагориты начального управления состемой обысковенных ликроеренциальных уранении состемой обысковенных порторенсии начальных условия Шоуолтера-Сидорова алери расширения спектра практического применения начальные условия Шоуолтера-Сидорова алери начальные условия Шоуолтера-Сидорова состемо состем сос				
BLIPOX_CENTRACKS SAURCENCY SPABHEHUM C H4 ADAIDATEMAN VC/DOBUSMM WOYD/TEPA CULOOPDA TpoIntoxets C Ploces and the back C Ploce and the ba				
Dot Ordital HIM VICINOBUSH MULTION CONCERSION / KOINE HIM         H4-YAJAIDHIBMU VICINOBUSH MULTION CONCERSION / KOINE HIM         H4-YAJAIDHIBMU VICINOBUSH MULTION CONCERSION / KOINE HIM         HAYAAIDHIBMU VICINOBUSH MULTION CONCERSION / KOINE HIM         HAYAAIDHIBMU VICINOBUSH MULTION CONCERSION / KOINE HIM         HAYAAIDHIBMU VICINOBUSH MULTION CONCERSION / KOINE HIM         HIM         Impune vacanetwork outling with the first outling of the form of the service outling operators' theory in the form of the service outling operators' theory in the se	ОПТИМАЛЬНОГО УПРАВЛЕНИЯ		1	
HAYADAbHbMM YCDOBURAMU WUOYOJTEPA CULQOPOBA Tpozinowetei anroputr w anroputr m puepe vacneteioro peweteis agaze anroputr au anroputr m puepe vacneteioro peweteis agaze anroputr au anropu	ВЫРОЖДЕННОЙ ЛИНЕЙНОЙ СИСТЕМОЙ УРАВНЕНИЙ			
Предложены anroptrm         Prodoku         Prodoku         Prodoku           anroptrm         dies of the sensigoup opendors' theory.         dies of the sensigoup opendors' theory.           i         mpukep численного управления         bis         metrical science         welf science           popendorsy transmission         is of the sensigoup opendors' theory.         dies of the sensigoup opendors' theory.           popendorsy transmission         is of the model         dies of the sensigoup opendors' theory.           popendorsy transmission         is of the model         dies of the sensigoup opendors' theory.           of ontrivands.horo ynpasnehus ypasnehus yp	- C		Все из загруженного текста Пользовательский сло	варь
Предложены anroptrm         Prodoku         Prodoku         Prodoku           anroptrm         dies of the sensigoup opendors' theory.         dies of the sensigoup opendors' theory.           i         mpukep численного управления         bis         metrical science         welf science           popendorsy transmission         is of the sensigoup opendors' theory.         dies of the sensigoup opendors' theory.           popendorsy transmission         is of the model         dies of the sensigoup opendors' theory.           popendorsy transmission         is of the model         dies of the sensigoup opendors' theory.           of ontrivands.horo ynpasnehus ypasnehus yp	НАЧАЛЬНЫМИ УСЛОВИЯМИ ШОУОЛТЕРА СИЛОРОВА			
anroputra des regular on parsona des regular on parsona des offer des anrigeous persons "heory on the semigrou persons" theory on the semigrou persons "heory on the semigrou persons" theory on the semigrou persons "heory on the semigrou persons" theory on the semigrou persons "heory on the semigrou persons "heory on the semigrou persons "heory on the semigrou persons" theory on the semigrou persons "heory on theory on the semigrou persons "heory heory on the semigrou persons "heory on the s			Русский	Английский
P     Heartransition     Welfkomm system     L     Sectransition     Welfkomm system     L     Sectransition       Intrumanishoro ynpabnetura     is of the form of     is of the form of     Is of the form of       Bup oxtpathion in welfkom system     is of the form of     Is of the form of       Contransition of ynpabnetura     is of the form of     Is of the form of       Contransition of ynpabnetura     is of the form of     Is of the form of       Contransition of ynpabnetura     is of the form of     Is of the form of       Contransition of ynpabnetura     is of the form of     Is of the form of       Contransition of ynpabnetura     is of the form of     Is of the form of       Micronsition of the algorithm     is of the form of     Is of the form of       Micronsition of the algorithm     is of the form of     Is of the form of       Micronsition of the algorithm     is of the form of     Is of the form of       Ana     Is of the form of     Is of the form of     Is of the form of       Ana     Is of the form of     Is of the form of     Is of the form of       Ana     Is of the form of     Is of the form of     Is of the form of       In pose pace     Is of the form of     Is of the form of     Is of the form of       In pose pace     Is of the form of     Is of the form of     Is of	<ul> <li>A state of the sta</li></ul>			
пример численного решения задачи     нете тиа     has     is of the form of	алгоритм			
пример численного решения задачи оттимального управления и честоя	чИ			
ontrusianshoro ynpabnehus     beline umu       bipokapeihok ynpabnehus     beline umu       bipokapeihok ynegeehujuanshus ypabnehus     beline umu       c     beline umu       c     beline umu       c     beline umu       c     beline umu       diama     beline umu       diamu     beline umu       d	пример численного решения задачи			
вырожденной линейной системой     еспользованы     ерисок       обыковенных дифференциальных уравнений     использованы     вегизе       с     вегизе     диала       постояниями коэффициентами     какчевое слово     каумона       использованы     каумона     вегизе       изимона     вегизе     вегизе	оптимального управления			
обыкновенных дифференциальных уравнений         использовык         weied           c         incresses         quadratic matrix           c         incresses         quadratic matrix           noronssosanus         quadratic matrix         quadratic matrix           Havanshus yonogeanus         incresses         Quadratic matrix         quadratic matrix           Ana         konvesses conce         Bieffhorerical justification of the algorithm         incresses           Ana         konvesses         Bieffhorerical justification of the algorithm         incresses           nooros         nooros         Incresses         Incresses         incresses           nooros	· ·			
Observation         Interpation-with         Interpation-with           C         exaptor Hark Hongwith         exaptor Hark Hongwith         exaptor Hark Hongwith           Micro has been Hark         Kongwith Hongwith         Kongwith Hongwith         exaptor Hark Hongwith         exaptor Hark Hongwith           Micro has been Hark         Kongwith Hongwith         Kongwith Hongwith         Kongwith Hongwith         Kongwith Hongwith           Ana         Kongwith Hongwith Hongwith         Kongwith Hongwith Hongwith         Kongwith Hongwith Hongwith         Kongwith Hongwith Ho				
C     kadapat nagina     quadratio matrix       Cortonshute kosquputuentramu     kadapat nagina     kayword     kayword       Maronbaserahu     Kouw     Couv/     couv/       Havanbute yonesus llayon repa-Cupopea     Kouw     Bieffhorerica justification of the algorithm       Ana     Kortone respertive/coxe docidocheseke anorphitma     Bieffhorerica justification of the algorithm       Mongonu     Havanbute yonesus llayon repa-Cupopea     Kata       Mongonu     Havanbute yonesus layon repa-Cupopea     Beorterica yonesus layon       Mongonu     Havanbute yonesus layon repa-Cupopea     Havanbute yonesus layon repa-Cupopea       Mongonu     Havanbute yonesus layon repa-Cupopea     Havanbute yonesus layon repa-Cupopea       Ha     Havanbute yonesus layon repa-Cupopea     Havanbute yonesus layon repa-Cupopea       Ha     Havanbute yonesus layon repa-Cupopea     Havanbute yonesus layon repa-Cupopea       Ha     Havanbute yonesus layon repa-Cupopea     Havanbute yonesus layon repa-Cupopea       C     Havanbute yonesus layon repa-Cupopea     Havanbute yonesus layon repa-Cupopea       G     Havanbute yonesus layon repa-Cupopea     Havanbute yonesus layon repa-Cupopea       G     Havanbute yonesus layon repa-Cupopea     Havanbute yonesus layon repa-Cupopea       G     Havanbute yonesus layon repa-Cupopea     Havanbute yonesus layon repa-Cupopea       Havanbut				
Vcronbasesses         vou         cou/         cou/           Havanbuse ycnosus lloyonrepa-Curgopesa         berides         indef horeitagistication of the algorithm           Ana         horto e reopertivecce docubesses anorphirita         berides         indef horeitagistication of the algorithm           pacultypeinus criektipa npaktruveckoro npumeelenus         horto e reopertivecce docubesses anorphirita         berides         indef horeitagistication of the algorithm           moorphil         neortisee         leontef ype         leontef ype         leontef ype           Ha         neortisee         leontef ype         leontef ype         leontef ype           nposependo         linear system         model         linear system         linear system           rudonehus         nagenb         model         linear system         linear system           rudonehus         nagenb         model         linear system         linear system           rudonehus         nagenb         model         linear system         linear system           rudonehus         nagenbe         euler method         model         linear system           rudonehus         nagenbe         euler method         model         linear system           rudonehus         marguespe         linear system	- C			quadratic matrix
начальные условия Шоуолтера-Сидорова         ретисе тораетическое обоснование алгоритна         breinferentical justification of the algorithm           для         besides           для         kata           расширения спектра практического применения         kata           mode/nu         leonite/type           Ha         leonite/type           nopespeleon         methas acurcreva         leonite/type           nopespeleon         methas acurcreva         methad of phase space           nopespeleon         methad acurcreva         methad of phase space           nopespeleon         methad acurcreva         methad of phase space           nopespeleon         macanhoe gronewe koaw         methad of phase space           c         macanhoe gronewe koaw         methad of phase space           c         macanhoe gronewe koaw         methad of phase space           c         macanhoe gronewe koaw         methad icondition           c         macanhoe gronewe koayonrepo-cuopopea	постоянными коэффициентами		ключевое слово	keyword
Havanshibe ycnosus uoyon tepa-cxuzopoola     powe     besides     in       Azia     yrta     kata       pacuuperus cnektpa npaktruveckoro npumelerus     noothtescuki tru     leonfelyc       Molgrinu     noothtescuki tru     leonfelyc       Ha     noothtescuki tru     leonfelyce       npumepe Jeonttesea     metod sphase space     incertification       npumepe Jeonttesea     metod sphase space     incertification       npumepe Jeonttesea     metod sphase space     incertification       nposeparence     metod sphase space     incertification       nposeparence     metod sphase space     incertification       c     mozample d     indextere     determination       c     mozample d     indextere     indextere       determination     metod sphase space     indextere       determination	- Использованы		коши	
And     Np0net     Desides       And     Np1net     Np1net       pacumpentras cnextrpa npaxtruveckoro nprumehetura     Acontaeccusi trai     leontel ype       Monanni     Acontaeccusi trai     leontel ype       Ha     Acontaeccusi trai     leontel ype       npuepe Joentres     method apsacero nprumehetura     incer system       npuepe Joentres     method apsacero nprumehetura     method apsacero nprumehed       npuepe Joentres     method apsacero nprumehetura     method apsacero nprumehetura       npoepateuro     Ha     on       rpoepateuro     Ha     on <th>начальные условия Шохолтера-Силорова</th> <th></th> <th></th> <th></th>	начальные условия Шохолтера-Силорова			
pacumpenus cnektpa npaktruveckoro npumenenus         nearth securit nn         leonier/yee           Molariu         Inearth securit nn         Ieonier/yee           Ha         neeth securit nn         Inear system           npumepe Леонть сва         method of phase system         method           npogeneethene         method sonce phase         method of phase system         method           c         moderne         moderne         moderne         moderne           c         moderne         on         method of phase system         method           repartering         userspice         method of phase system         method         method           repartering         method system         method of phase system         method         method           repartering         method system         moderne         moderne         moderne         method         method         method         method system         method system condition         method system condition         method system condition         method system condition         method s				
Procession         Provide of the second truin         Ieonite flype           Mageinut         Ieonite flype         Ieonite flype           Ha         Ieonite flype         Ieonite flype           Inpole pleon to be a         method of phase space         Ieonite flype           Inpole pleon to be a         method of phase space         Ieonite flype           Inpole pleon to be a         method of phase space         Ieonite flype           Inpole pleon to be a         method of phase space         Ieonite flype           Inpole pleon to be a         method of phase space         Ieonite flype           Inpole pleon to be a         method of phase space         Ieonite flype           Inpole pleon to be a         method of phase space         Ieonite flype           Inpole pleon to be a         machoe grane te usgo repare         Infel condition           Inter condition to be a         machoe grane te usgo repare         Infel condition           Inter condition to be a         machoe grane te usgo repare         Infel condition           Inter condition to be a         machoe usgo repare         Infel condition           Inter condition to be a         machoe usgo repare         Infel condition           Inter condition to be a         machoe usgo repare         Infel condition           Inte				
Molgania         Inservise         Inservise           Ha         Inservise         method sphase space           rpuwepe Jeontbeega         evento dascoro ropoctpanctes         evento dascoro ropoctpanctes         evento dascoro           rpuwepe Jeontbeega         evento dascoro ropoctpanctes         evento dascoro ropoctpanctes         evento dascoro           rposepatienue         evento dascoro ropoctpanctes         evento dascoro ropoctpanctes         evento dascoro           c         model         on         model         evento dascoro           c         wacante         onte evento         evento dascoro         evento dascoro           reno dascoro ropoctpanctes         intel condition         evento         evento           reno dascoro ropoctpanctes         intel condition         evento         evento           reno dascoro ropoctpanctes         intel condition         evento         evento           reno dascoro ropoctpanctes         scordy midel condition         evento         evento dascore         evento           reno dascore         userserve condition         evento         evento         evento         evento           reno dascore         odocene-ve concese usgonreps concore         scorela vento         evento         evento         evento         evento	расширения спектра практического применения			
Ha purspen Disort resa method of phose space index ind	модели			
проведено         нетод алера         euler method           проведено         mode/h         mode/h           проведено         mode/h         mode/h           c         mode/h         mode/h           c         mode/h         mode/h           ruc         mode/h         mode/h           ruc         mode/h         mode/h           ruc         mode/h         determindon           ruc         mode/h         determindon           ruc         mode/h         mode/h           ruc         mode/h         mode/h<	- Ha			
inpose zerio         index model         index model           cpaseleerio         in	примере Леонтьева			
Image: constant set in the set i			модель	model
c     Nal pumpe     Outline statupe of determination       vuccentum production     Advantage of determination     Advantage of determination       vuccentum production     Intel condition     Advantage of determination       vuccentum production     Advantage of determination     Advantage of determination       neontrum anabero synpapine run anontrum anabero synpapine run determination     Intel showed of determination     Advantage of determination       neontrue determination     Hermination     Intel showed of metanabero synpapine run devantage of conset upgonteps cuapose     Intel showed of metanabero synpapine run devantage of conset upgonteps cuapose       c     Hermination     Hermination     Intel showed of metanabero synpapine run devantage of conset upgonteps cuapose       c     Hermination     Hermination     Intel showed of metanabero synpapine run doctage upgonteps cuapose       c     Hermination     Hermination     Intel showed of metanabero synpapine run doctage upgonteps cuapose       c     Hermination     Hermination     Intel showed of metanabero synpapine run doctage upgonteps cuapose       c     Hermination     Hermination     Intel showed of metanabero synpapine run doctage upgonteps cuapose       c     Hermination     Hermination     Intel showed of metanabero synpapine run doctage upgonteps cuapose       c     Hermination     Hermination     Intel showed of metanabero synpapine run doctage upgonteps cuapose <th></th> <th></th> <th></th> <th></th>				
численным решением задачи         нечельное усление одстовина         infel condition           оптимального управления системой         нечельное усление кошуоптере сидорове         infel showler sidorov condition           песнтъевского типа         нечельное усление кошуоптере сидорове         infel showler sidorov condition           с         нечельное усление кошуоптере сидорове         showlere sidorov condition           к         обощение усление кошуоптере сидорове         showlere sidorov condition           численное решение         обощение усление кошуоптере сидорове         showlere sidorov condition           оптимальное управление         обощение усление кошуоптере сидорове         showlere sidorov condition           оптимальное управление         обощение усление кошуоптере сидорове         showlere sidorov condition           оптимальное управление         ofosmethere ycanose         ordenoit         ofosmethere ycanose           оптимальное управление         ofosmethere ycanose         ofosmethere ycanose				
Autonation contrata protection of the autonation of the autonatis autonation of the autonation of the autonation of the a				
Ontmanship         Heranshoe         Heranshoe         Initel showler sidoro: condition           Incentre         Heranshoe         Initel showler sidoro: condition         Initel showler sidoro: condition           Incentre         Heranshoe         Initel showler sidoro: initel condition         Initel showler sidoro: initel condition           Incentre         Heranshoe         Incentre         Initel showler sidoro: initel condition           Incentre         Genereures         Incentre         Incentre           Haranshoe         Genereures         Incentre         Incentre           Haranshoe         Genereures         Incentre         Incentre           Haranshoe         Googeneree         Incentre         Incentre           Haranshoe         Googeneree         Incentre         Incentre           Intel showler         Incentre         Incentre         Incentr	численным решением задачи			
Indemtse         Havanshee uponteparvacpope         showeller evidorio initial condition           C         Havanshee uponteparvacpope         showeller evidorio initial condition           Havanse uponteparvacpope         showeller evidorio initial condition         initial control initial condition           Kinovesite conda         odostatente uponteparvacpope         showeller evidorio initial condition         initial control initial condition           Vucnethoe petitette         odostatente uponteparvacpope         showeller evidorio initial condition         initial control initial condition           Vucnethoe petitette         odostatente uponteparvacpopea         generalized showeller sidorio condition         initial control initianantee upopenense concreteroli initianantee upopenense control initianatee upopenense control initianatee upopenense control initianantee upopenense control initianatee upopenen	оптимального управления системой			
с herucrose existencia in onempty convex subset in provides herucrose existencia in operatorse consistencia in operatorse donation showeller generalized condition oбоещение gunose e exopartepa conseleratione oбоещение gunose e exopartepa conseleratione oбоещение gunose e exopartepa conseleratione oбоещение gunose e exopartepa conseleratione offer existencia in the state offer existencia in the state order existencia in the state on retreactione gunose e exopartepa conseleratione offer existencia in the state order existencia in the state on retreactione gunose existencia in the state on one parter on retreactione gunose existencia in the state on the	леонтьевского типа			
начальным условием Коши         Обеспечиеат         provides           Ключевые слова         Обеспечиеа условие видоптера         showeller generalized condition           численное решение         Обобщенное условие видоптера         showeller generalized condition           численное решение         Обобщенное условие видоптера         generalized showeller sidorov condition           отитимальное управление         Обобщенное условие видоптера сидорова         generalized showeller sidorov condition           система леонтьевского типа         Обобщенное управление         ode из а         some of           вередение         Опинальное управление         oneportop         operator           опинальное управление         опинальное управление         opfinal system control           Вередение         Основена управление состеой         opfinal system control           Вередение         Основена управление состеой         opfinal system control           В [1]         Основена управление состеой         main	- C			
Hardination y Criticities No.000     doctager-log y consete augorances augorances     showalter generalized condition       Kinoreesee znosa     oodcager-log y consete augorances     generalized showalter generalized condition       Yucchetkoe petuelke     ordisay-terus y consete augorances     generalized showalter generalized condition       ontrukanabeloe yingas petue     ordisay-terus y consete augorances     generalized showalter generalized condition       ontrukanabeloe yingas petue     ordisay-terus y consete augorances     generalized showalter sidors y condition       ontrukanabeloe yingas petue     ordisay-terus y consete augorances     ordisay-terus y consete       ontrukanabeloe yingas petue     ordisay-terus y condition     ordisay-terus y consete       we to appaso to npo cripater tsa     ordisay-terus y consete     optimal system control       beginetive     ordisay-terus y consete     ordisay-terus y control       ordisay-terus y consete     ordisay-terus y consete     ordisay-terus y consete       ordisay-terus y consete     ordisay-terus y consete     ordisay-terus y consete       ordisay-terus y consete     ordisay-terus y consete     ordisay-terus y consete       ordisay-terus y consete     ordisay-terus y consete     ordisay-terus y consete       ordisay-terus y consete     ordisay-terus y consete     ordisay-terus y consete       ordisay-terus y consete     ordisay-terus y consete     ordisay-terus y cons				
Kino deshe Croba         odscientification         generalized showeller sidory condition           Yuncheho puliehke         odseuty-oe up_onsere aug_ontepa chapposa         generalized showeller sidory condition           Yuncheho puliehke         odseuty-oe up_onsere aug_ontepa chapposa         generalized showeller sidory condition           Yuncheho puliehke         odseuty-oe up_onsere aug_ontepa chapposa         generalized showeller sidory condition           Ontrivenan-body tripasinehke         odseuty-oe up_onsere-oe up_operative up_onsere up_ontepa chapposa         generalized showeller sidory condition           CHCTexa neohrtsebsckoro trina         oneportop         operativ         optimal system control           Beogenewe         ontevana-bed ynposnewe chrevé         optimal system control         ontevana-bed ynposnewe chrevé           Beogenewe         occeseat yne up_onteve durposa         tis based         occeseat yne up_onteve durposa           B [1]         occeseat yne up_onteve durposa         main         main				
оптимальное управление         Одни из         some of           система леонтъевского типа         оператор         operator           метод фазового пространства         оптимальное управление         optimal control           Ведение         осковен         optimal control         optimal system control           Ведение         осковен управление системой         optimal system control         optimal system control           В [1]         осковен управление системой         main purpose of this antice         occosent ymain         main			обобщенное условие шоуолтера сидорова	
оптичальное управление ореног система леонтъевското типа метод фазового пространства Ведение В[] система леоне управление ореганото оп оптичальное управление ореганото оп основана и вла данооб статьи май от система денооб статьи основана и вла данооб статьи май оптичальное управление ореганото оп ваното оптичальное управление ореганото оп основана и вла данооб статьи май опространства востана и опространото оп основана и вла данооб статьи май опространото оп основана и вла данооб статьи основана и вла данооб статьи ос	численное решение			
СИСТЕМА ЛЕОНТЪЕВСКОГО ТИПА МЕТОД ФАЗОВОГО ПРОСТРАНСТВА Ведоние В [1] ОПТИЧальное управление систеной ОСновен и ОСНОВИИ ОПТИЧАЛЬНОЕ И ОСНОВИИ ОПТИЧНИИ ОПТИЧНИИ ОПТИЧНИИ ОПТИЧНИИ ОПТИЧНИИ ОПТИЧНИИ ОПТИЧНИИ ОСНОВИ ОПТИЧНИИ ОПТИЧНИИ И ОСНОВИИ ОПТИЧНИИ О	оптимальное управление			
метод фазового пространства         оптимальное упревление         ортипаl силта           Ведение         основена цань данной статьи         is based           В [1]         основена цань данной статьи         main	система леонтьевского типа			
Введение         основен цельданной статьи         is based           В [1]         основена цельданной статьи         main purpose of this article				
В [1] основная цаний статьи main purpose of his and e main				
основной тей				
ПСЕЛПОЖЕН				
	предложен	-	от	from

Figure 1. A screenshot of the Russian-to-English CAT tool user interface at the bookmark "show all". The left pane displays a SL interactive text of a scientific paper in mathematical modelling with explicitly marked (bold faced) nominal terminology and verbs (in blue). The left pane contains the alphabetically ordered list of all 1-4 component Russian terms with their English equivalents. On the top of the right pane there is a type-in area which permits searching for the translations of terms longer than 4 words in the tool knowledge base. The second bookmark on the top of the Ru-En equivalent area allows opening a user dictionary for the user to collect terms she/he might need in the future.

# 2 The Tool

The tool takes a SL text an as input and on the fly produces output at two levels:

- a marked-up interactive SL text with highlighted multi-component nominal and verbal terminology (NPs and VPs);
- a list of all single- and multi-component SL-TL units found in the input text.

Text mark-up improves input readability and helps translator quicker and better understand the syntactic structure of the input. This feature combined with on-the-fly translation of *all* 1-4 component SL text lexical units reduces translation time and effort and raises translation quality. The tool can be used as an e-dictionary where terms are searched through a type-in area in the user interface.

Translation equivalents are normalized as follows. SL NPs are outputted in nominative singular, while VPs are presented in a finite form keeping the SL voice, tense and number features. For example, in the Russian-to-English tool the Russian VP wordform "смонтированные"\_past participle, perfective, plural (literally "done") will be outputted as "смонтированы"\_ finite, past, plural = "were mounted". The tool user interface has a lot of effort-saving functionalities. A click on a unit in the marked up input text in the left pane highlights its TL equivalent in the alphabetically sorted list of translations on the right pane. It is possible to create user dictionaries accumulating terminology from different texts, saving these dictionaries and projects, etc. A screenshot of the user interface in shown in Figure 1.

### **3** Methodology and Development Issues

#### 3.1 Architecture

The overall architecture of the tool is shown in Figure 2. The tool engine consists of a shallow analyzer including three fully automatic modules, - a SL hybrid NP extractor, shallow parser and imbedded machine translation module meant to translate terminology. The knowledge base contains shallow linguistic knowledge, - lexicons and rules.

The NP extractor is a hybrid stand-alone tool pipelined to the system. We built it following the methodology of NP extraction for the English language as described in (Sheremetyeva, 2009) and ported it to the Russian language.



Figure 2. The architecture of the CAT tool.

The extraction methodology combines statistical techniques, heuristics and very shallow linguistic knowledge. The knowledge base consists of a number of unilingual lexicons, - sort of extended lists of stop words forbidden in particular (first, middle or last) positions in a typed lexical unit (Russian NP in our case).

NP extraction procedure starts with n-gram calculation and then removes n-grams, which cannot be NPs by successive matching components of calculated n-grams against the stop lexicons. The extraction itself thus neither requires such demanding NLP procedures, as tagging, morphological normalization, POS pattern match, etc., nor does it rely on statistical counts (statistical counts are only used to sort out keywords). The latter makes this extraction methodology suitable for inflecting languages (Russian in our case) where frequencies of ngrams are low.

Porting the NP extractor from English to Russian consisted in substituting English stop lexicons of the tool with the Russian equivalents. We did this by translating each of the English stop lists into Russian using a free online system PROMT (http://www.translate.ru) followed by manual brush-up.

The NP extractor does not rely on a preconstructed corpus, works on small texts, does not miss low frequency units and can reliably extract *all* NPs from an input text. We excluded a lemmatizer from the original extraction algorithm and kept all extracted Russian NPs in their textual forms. The noun phrases thus extracted are of 1 to 4 components due to the limitations of the extractor that uses a 4-gram model. The extractor was also used for lexicon acquisition.

The shallow parser consists of an NP chunker, VP chunker and tagger. The first users the knowledge dynamically produced by the NP extractor (lists of all NPs of an input text in their text form). The VP chunker and tagger turn to the Russian entries of the tool bilingual lexicon. The tagger is actually a supertagger as it assigns supertags coding all morphological features, such as part-of-speech, number, gender, tense, etc.

The machine translation module translates text chunks into English using simple transfer and generation rules working over the space of supertags as found in the CAT tool bilingual lexicon.

# 3.2 Bilingual lexicon

To ensure correct terminology translation the bilingual lexicon of the tool should necessarily be tuned to a specific domain for which it is to be used. The lexicon is organized as a set of shallow cross-referenced monolingual entries of lexical units listed with their part-of-speech class and paradigms of domain-relevant explicit wordforms. This is the type of resource that, once build for some other purpose, can be simply fed into the system. Acquisition of this type of knowledge for every new pair of languages is what existing SMT tools can provide either in advance or on the fly, as reported in (2012 et al.,). In our work striving for correctness we combined automatic techniques with manual check and manual acquisition.

The Russian vocabulary was created in two steps. First, an initial corpus of Russian scientific papers on mathematical modelling of approximately 80 000 wordforms was acquired on Internet. We then ported the NP extractor described above to other Russian parts-of-speech and automatically extracted domain specific typed lexical units (NPs, VPs, ADJs, etc) consisting of 1 up to 4 components from the corpus. These automatically extracted lists of lexemes were further checked by human acquirers and 14 000 of them were used as a seed Russian vocabulary.

The seed vocabulary was then used to acquire longer Russian lexemes both from the initial corpus, and the Internet, which is in fact an unlimited corpus. The following methodology was applied. The seed lexical units were used as keywords in the Internet search engines. New Russian terminological units including seed terms highlighted in the two first pages of the search results were included in the lexicon. For seed example, for the (key) term «псевдообращение» the following multicomponent terms popped-up on the Internet: «псевдообращение сопряженной системы», «псевдообращение матриц с вырожденными весами», «псевдообращение Мура-Пенроуза», etc. As a result, the seed Russian vocabulary was extended to 60 000 single- and multi-component units up to seven-eight words long.

Lexical acquisition of English equivalents was done based on existing domain lexicons, parallel/comparable corpora and raw Internet resources. The last needs to be explained. In case neither existing lexicons, nor parallel/comparable corpora could provide for a reliable English equivalent, which was mostly the case with long terms, translation hypotheses were made based on different combinations of translation variants of component words. Every translation hypothesis was then checked in the Internet search engine. If an engine (we used Google) showed a translation version in the search results, the hypothesis was considered confirmed and the English equivalent was included in the tool lexicon. For example, the Russian term *«poesoe npedcmasnenue чacmuцы»* could not be found in any of existing lexicons, the following English equivalents of the Russian term components were found:

рой – swarm; представление - conception, expression, representation, performance, configuration; частица – bit, fraction, particle, shard, corpuscle.

If you create a translation hypothesis by using the first translation variant for every component of the Russian term you will get: «swarm conception of a bit» or «bit swarm conception». Used as key words in Google, the search results do not contain these words combined in a term. This translation hypothesis was rejected. Another hypothesis «particle swarm representation» used as key words in Google gives the English term «Particle Swarm Optimization and Priority *Representation*» from the paper on mathematical modelling by Philip Brooks, a native English speaker. «Particle swarm representation» is accepted as a correct English translation of the term представление Russian «роевое частицы». Though tedious, this methodology allowed careful detection of the up-to-date highly reliable translation that could hardly be achieved otherwise.

# 3.3 Workflow

The raw SL document first goes to the automatic NP extractor, which produces a list of one- to four component noun phrases. The dynamically created NP list is then used as knowledge for the NP chunker, which by matching the extracted list against the input text chunks (brackets) noun phrases in the document. The morphological tagger completes morphological analysis of these chunks by looking them up in the NP entries of the tool lexicon. The text strings between chunked NPs is then supplied to the VP chunker that matches this input against verb wordforms, as listed in the morphological zones of verb entries. In case of a match the text string is chunked as VP and a corresponding supertag from the lexicon is assigned. The text strings which were left between NP and VP chunks are then looked up in the rest of the entries of the lexicon and tagged. The fact that in every chunking/tagging pass only the type-relevant lexicon entries are searched practically lifts the ambiguity problem in morphological analysis.

Finally, based on classified chunk borders, the document is turned into an interactive ("clickable") text with NP and VP phrases highlighted in different colours.

The output of the shallow analysis stage (fully (super) tagged lexical units) is passed to the machine translation module that following simple rules generates SL-TL lexical pairs for all the lexica of the text (See Figure 1).

# 4 Status and Conclusions

The viability of the methodology we have described was proved by its implementation in a Russian-English CAT tool for the domain of scientific papers on mathematical modelling. The tool is fully developed. The domain bilingual static knowledge sources have been carefully crafted based on corpora analysis and internet resources. The programming shell of the tool is language independent and provides for knowledge administration in all the tool modules to improve their performance.

The extractor of Russian nominal terminology currently performs with 98, 4 % of recall and 96, 1% precision. The shallow clunker based on the extraction results and lexicon shows even higher accuracy. This is explained, on the one hand, by the high performance of the NP extractor, and, on the other hand, by the nature of inflecting languages. Rich morphology turns out to be an advantage in our approach. Great variety of morphological forms lowers ambiguity between NP components and verb paradigms.

We could not yet find any publications describing research meant for similar output. This leaves the comparison between other methodologies/tools and ours as a future work. In general user evaluation results show a reasonably small number of failures that are being improved by brushing up the bilingual lexicon.

We intend to a) improve the quality of the tool by updating the tool knowledge based on the user feedback; b) integrate the tool into a full MT system and c) develop a search facility on the basis of the our extraction strategy.

## References

- Enache Ramona, Cristina Espana-Bonet, Aarne Ranta, Lluis Marquez. 2012. A Hybrid System for Patent Translation. *Proceedings of the EAMT Conference*. Trento..Italy, May
- Koehn Philipp. 2009. A process study of computeraided translation, Philipp Koehn, *Machine Translation Journal*, 2009, volume 23, number 4, pages 241-263
- Kudashev Igor S. 2007. Desining Translation Dictionaris of Special Lexica /I.S.Kudashev. – Helsinki University Print, – 445 p.
- Sharoff, Serge . 2004. What is at stake: a case study of Russian expressions starting with a preposition. *Proceedings of the ACL Workshop on*

Multiword Expressions: Integrating Processing, July.

- Sheremetyeva, Svetlana. 2009. On Extracting Multiword NP Terminology for MT. *Proceedings of the EAMT Conference*. Barcelona, Spain, May.
- Shinmori A., Okumura M., Marukawa Y. Iwayama M. 2003. Patent Claim Processing for Readability -Structure Analysis and Term Explanation, Workshop on Patent Corpus Processing. conjunction with ACL 2003, Sapporo. Japan, July.