Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Владимирский государственный университет
имени Александра Григорьевича и Николая Григорьевича Столетовых»

О. А. СЕЛИВЕРСТОВА

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ПЕРЕВОДЕ

Учебно-практическое пособие

Владимир 2020

УДК 811.111
ББК 81.2(Англ)
С29

Рецензенты:
Кандидат филологических наук
доцент кафедры философии, истории, права
и межкультурной коммуникации
Владимирского филиала Финансового университета
при Правительстве Российской Федерации
Н. А. Наумова

Кандидат педагогических наук
доцент кафедры русской и зарубежной филологии
Владимирского государственного университета
имени Александра Григорьевича и Николая Григорьевича Столетовых
В. И. Горбатов

Издается по решению редакционно-издательского совета ВлГУ

С29 **Селиверстова, О. А.** Информационные технологии в переводе : учеб.-практ. пособие / О. А. Селиверстова ; Владим. гос. ун-т им. А. Г. и Н. Г. Столетовых. – Владимир : Изд-во ВлГУ, 2020. – 96 с. – ISBN 978-5-9984-1169-4.

Разработано для курса «Перевод с применением современных технологий». Включает в себя как теоретический материал, так и практические задания, дидактический материал для самоконтроля.

Предназначено для студентов вузов 2-го курса направления подготовки 45.03.02 – Лингвистика очной формы обучения.

Рекомендовано для формирования профессиональных компетенций в соответствии с ФГОС ВО.

Ил. 21. Библиогр.: 20 назв.

УДК 811.111
ББК 81.2(Англ)

ISBN 978-5-9984-1169-4

CONTENTS

FOREWORD

Welcome to the exciting word of translation. Some might think translation is a monotonous tiresome work. I can assure you here it is not the case. Today information technology is so closely blended with language that purely manual translation with a good old paper dictionary without technology looks like hunting a mammoth with a stick in the Stone Age. If you do not feel like hunting a mammoth you are very likely to find the tools and resources described in this book really helpful. As soon as you have them at your finger tips you are sure to feel more confident as a translator and more competitive in the labour market.

The route you are going to follow in this book is a sequence of destinations referred to as Units here. Make sure you spend enough time in each of them and get to know the locals (which are software tools) and their way of life. To help you here we have prepared tasks, exercises and tutorials.

When you have accomplished them all and achieved the final destination, we expect you to see the trade of translation in a completely different light and enjoy the process of using IT tools for finding solutions in the most intricately twisted cases, checking your hypotheses and discovering new opportunities.

Enjoy you trip to the world of information technology in translation!

Unit 1

MICROSOFT WORD

Pre-reading tasks

Make sure you know Russian equivalents of the following words and word combinations.

Tab, folder, layout, File button, Home tab, References tab, Mailings tab, Review tab, View tab, clipboard, font, spacing, bullets, lists, justify, line spacing, indents, borders, bold, underline, strikethrough, highlight, page break, cover page, toolbar, text wrapping, print layout, full screen reading, web layout, outline and draft views.

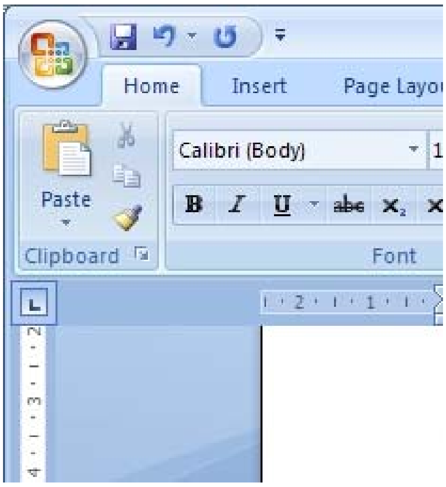
Basics

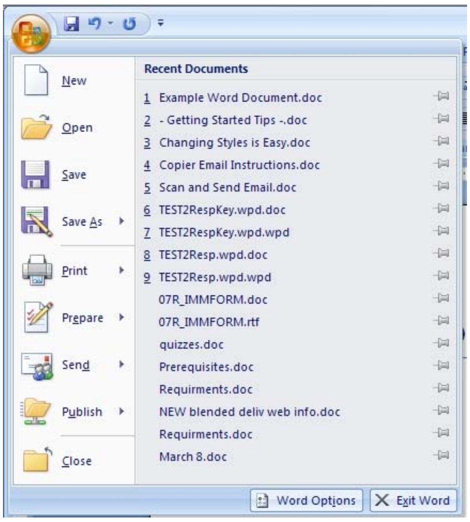
You might be surprised to find this topic in our book wondering what value it has for translation and if there is anything, you do not yet know about the product, as you have been a Word user since middle school. However, this wonderful software is a truly vital tool for any translator and having all its tricky features at your fingertips is a must. Besides by reading about MS Word functions in English is sure to enrich your vocabulary with English equivalents for well-known Russian words related to IT. So, let's get started.

Microsoft Word Menus

File

When you open a new file in Word, the first tab you will see is File. Look carefully – you may miss it because it is a different color than the other tabs (depending on what color scheme you have your desktop set on). The File tab has just what you would think: options related to the entire file, such as save, print, share, and open.





Recent Documents

1 Example Word Document.doc
2 - Getting Started Tips -.doc
3 Changing Styles is Easy.doc
4 Copier Email Instructions.doc
5 Scan and Send Email.doc
6 TEST2RespKey.wpd.doc
7 TEST2RespKey.wpd.wpd
8 TEST2Resp.wpd.doc
9 TEST2Resp.wpd.wpd
07R_IMMFORM.doc
07R_IMMFORM.rtf
quizzes.doc
Prerequisites.doc
Requirments.doc
NEW blended deliv web info.doc
Requirments.doc
March 8.doc

Instead of the File button, you may see The Office Logo Clicking the Office logo at the top left of the screen will provide most of the items formerly found under the file menu including those listed above. Beside the logo, you also will find a disk icon to save your document as well as the undo and redo buttons.

As you can see here, the Office logo opens up listing your options for new, open etc. and also contains a list of your recent documents for quick opening. Any of the items listed with an arrow beside them will replace the recent documents on the right with the options associated with the menu item. You should also see at the bottom right of this menu a button for exiting Word and changing Word's options.

Save as

The Save as option will provide you with the most common file formats to save your document in. The common ones are Word Document, Word Template, and Word 97-2003 document. However, some of the programs we will need for the course of Translation with Technology will require the text to be in a different format, like .txt or .pdf. Some software may need to change the encoding it may be done using the same Save as option.

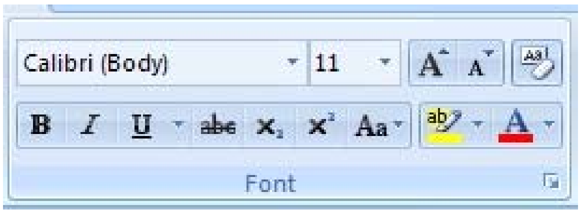
The Home Tab

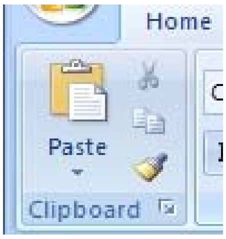
The Home tab has the most commonly used features, especially as they relate to modifying text. In the home tab you can select your font, size, color, attributes (bold, italics, underline), and alignment (left, center, right). You can also select a style, which is a predetermined text made to fit certain document parts, such as headings, subtitles, and text.

Clicking the down arrow beside any of the icons here will drop down more options for that tool. Each section also contains an arrow in the bottom right corner, which will open a window containing the options found in that section.

Clipboard

The Clipboard allows you to cut, copy, paste and copy formatting from one place to another.





Font

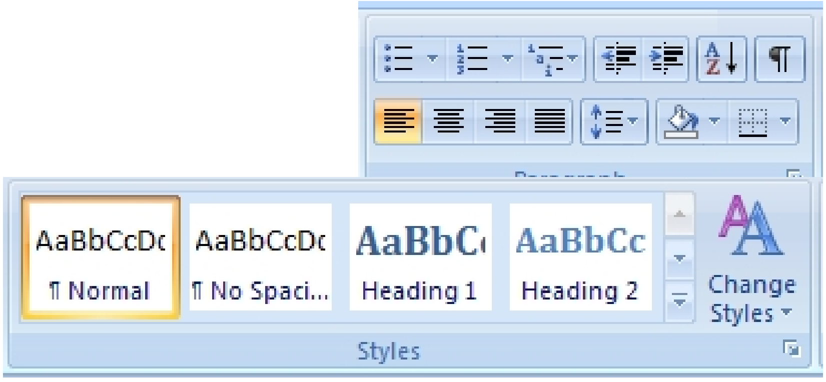
The font section of the ribbon provides a section to handle the basic text formatting. Items such as bold, underline, strikethrough, highlight and font type can be changed here. Some items from this section and some items from the paragraph section are also available by highlighting text and moving your cursor slightly above the highlighted text. This saves having to move your cursor all the way to the top of the screen for some common formatting items.

Paragraph

The paragraph section provides icons for bullets, lists, justify, line spacing, indents and borders.

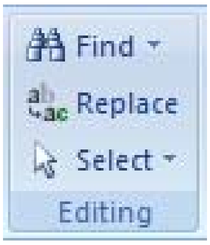
Styles

The styles section allows you to quickly change the formatting of a section of text by choosing one of the predefined styles. You can also create a new style based on the formatting of your selected text for use later in other sections of your document.



Editing

The Editing section of the toolbar allows you to find, replace and select items. The select option gives you the ability to select all, select objects or select text with similar formatting. This last option gives you the ability to quickly change everything in your document with one style to another style without having to manually find all of that text and change each area separately.



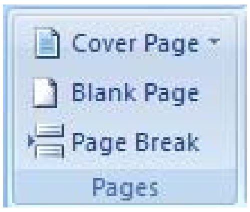
Insert Tab

Under the Insert tab, there are a number of choices. In a Word document, there are many types of visual aids and highlights you can add to a file to help summarize and present information. It's in the insert ribbon tab where you can find options for graphics, charts, hyperlinks, page breaks, headers, footers, textboxes, and reference information, such as date and time, comments, page numbers, and bookmarks.

The insert tab has seven sections for inserting most types of objects. The sections are pages, tables, illustrations, links, header and footer, text and symbols.

Pages

The pages section is where you can go to insert an cover page, blank page or page break. The cover page drop down offers a selection of predefined cover pages for your document that have sections for title, date and author. You can also select text in your document and choose to save the selected text to the cover page gallery for use in future documents.



Tables

The table section only has a drop down menu which offers a grid to create a new table, insert table, draw table, convert text to table for selected text, Excel spreadsheet, and some predefined “Quick Tables” that have formatting already setup for you. When working on a table you will have two additional tabs along the top of the ribbon, the design and layout tabs. There are screenshots of both directly below.

Table Design

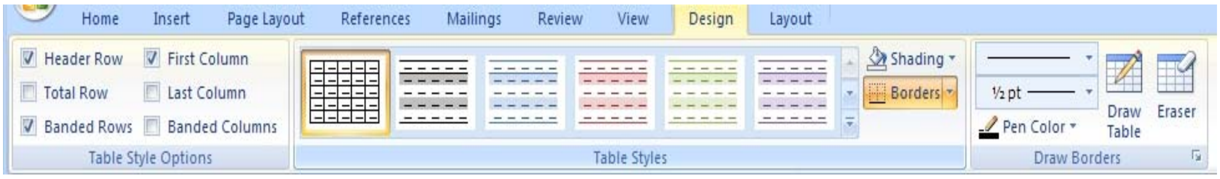
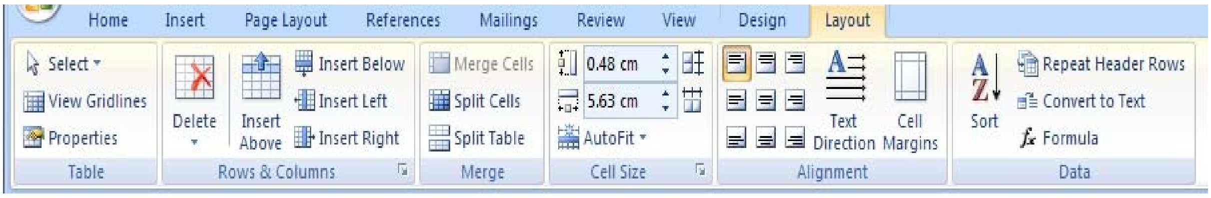
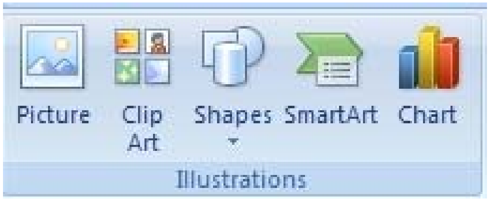


Table Layout



Illustrations

The Illustrations section allows you to insert pictures, clipart, shapes, SmartArt and charts.



After inserting or selecting a picture you are provided with a new toolbar along the top shown here. This toolbar gives you the ability to change the brightness, contrast, shape, position, text wrapping and other options for the picture.

Clicking off the picture or on one of the other tabs will take you back to the standard toolbars. The Shapes option of the Illustrations section allows you to insert lines, arrows, boxes, basic flowchart shapes.

Design Tab

The Design tab can be either very useful or hardly used, depending on your own understanding of Word. Most of the space in the design tab is taken up by examples of document designs that you can select, such as documents with centered titles, offset headings, and left justified text. However, in addition to those less popular tools, the design tab also includes watermarks, page color, and page borders, which may be used by advanced Word users.

Page Layout

The Page Layout ribbon is an important tab to determine how your document looks. This is the tab that has the options to modify margins, page orientation, paper size, columns, indents, spacing, page breaks, and the arranging of any parts of the document, such as text and graphics or tables.

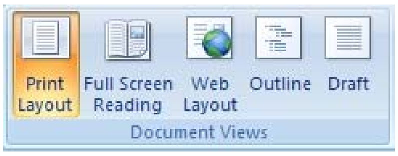
References

The References tab is one that you may never use, or may be used heavily, depending on the type of work you do in Word. For students, the references tab is the easiest way to insert citations and references into the Word document. It can help with creating the reference page, table of contents, footnotes, and sources.

View Tab

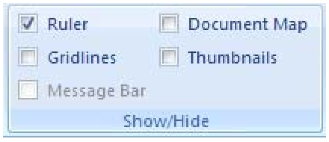
The view tab offers five sections which include document views, show/hide, zoom, window and macros.

Document Views



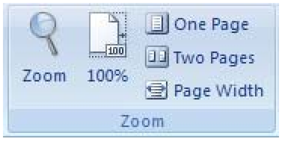
The document views section switches you between print layout, full screen reading, web layout, outline and draft views. Print layout is the default view. Full screen view removes all but a couple of tools from the top of the screen and the rest of the screen is your document. Web layout will take away the empty space on either side of the document if there is any and fill the window as if it were a web page. Outline view changes the look of your document into an almost point form style which may help with reviewing main points. Draft view takes away most of your formatting and images and just shows the text. It also fills the window with your text similar to web layout.

Show/Hide


☑ Ruler □ Document Map
□ Gridlines □ Thumbnails
□ Message Bar

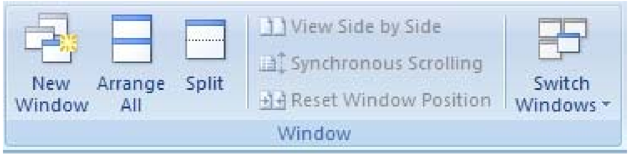
The show/hide section will toggle certain tools on or off the screen including rulers, gridlines, message bar, document map and thumbnails. The rulers will show along the top and left side of the screen. Gridlines will cover your entire document inside the margins. They will be visible on screen but don't print. The message bar can only be displayed when there is a message to be displayed. One common reason for the message bar to display is when macros have been enabled or disabled. The document map and thumbnails will show along the left hand side of the screen.

Zoom



The zoom section provides tools to zoom into or out of the document. You can choose your own zoom factor or use one of the predefined zoom factors of 100 %, one page, two pages (side by side), or page width which causes the document to zoom in or out so it fills your window.

Window



The new window button will open your current document in a new window. The arrange all button will take your currently open windows and stack them one on top of the other. The split button will take your current document and show it in two frames within the window one on top of the other. This will allow you to look at something you wrote on page one while working on page twenty. View side by side allows you to view two windows side by side, once in side by side view you can turn on synchronous scrolling so both side scroll at the same time. Also while in side by side mode if you resized either window you can click the reset window position button to have them share the screen equally again. The switch window drop down will allow you to switch between open windows.

Macros

The macros section provides the tools required to work with and create basic macros. You can view existing macros or record your own. Choose record macro from the drop down and then perform the functions you do often, like change the page layout, and style of the document. Once you have done those tasks then stop recording. You will be able to use that macro over again to shorten the steps you need to take every time you need to perform that set of tasks.

SELF-CHECK TASKS

1. Give English equivalents for the following words and word combinations**:** вставка, разметка страницы, ссылки, рецензирование, вырезать, вставить, поля, ориентация, разрыв страницы, отступ, оглавление, сноска, создать примечание, исправление, принять/отклонить исправление, режим структуры, черновик, линейка, масштаб, по ширине страницы.

2. Open Microsoft Word and describe the interface using English equivalents of all menu options.

PRACTICE TASKS

1. a) Read and translate the guidelines below.

b) Create a text document (or modify an existing document) with a table, a figure, footnotes and a list of references according to the following guidelines.

Submission guidelines

1. Please submit your manuscript as an **.RTF file**.

2. **Page size**: A4 (210×297 mm). Portrait layout.

3. **File name**: your last name – underscore – section number/mc (for master class)/pp (for poster presentation). For example, Smith_4 or Jones_mc.

4. **Margins**: 3 cm at left, 2.5 cm at right, 2.5 at top and 3.0 at bottom.

5. **Font**: Times New Roman; **font size** 12, 1.5 line spaced.

6. **Maximum length** of the manuscript: 2100 words.

7. **Paragraphs**: **indented,** without extra spaces between them.

8. **Center text justification**: no hyphenation.

9. **All illustrations** (charts, drawings, diagrams, pictures, etc.) are to be captioned as Picture or Table and numbered in the manuscript. All illustrations should be placed directly in the manuscript.

10. **All examples** are to be **italicized**. For more emphasis use **boldface**.

11. **Definitions** and translations: in **single quote marks** ('…'). **Double quote marks** ("……") should be used around quotations.

12. **Pages**: no numbering.

13. **Footnotes** are not to be used.

14. **Format** of the manuscript.

Your first and last name: bold, italicized, right margin justification.

- The manuscript title: in capital letters bold, centered.
- The information about the grant that supported your research should be given below the manuscript title, center justification, italicized.
- An abstract of the manuscript (100 – 120 words) should summarize the key points of the paper. State the purpose of your research, its important findings and conclusions.
- Key words in bold (5 – 6 in alphabetical order) separated by commas and a full stop at the end of the line, indented.
- The manuscript text.
- **References**: the subtitle in bold, left justified, indented, with a colon.
- **About the author**: the subtitle in bold, left justified, with a colon.

On a new line below write your name in bold. On a new line below write your degree (if any), post, affiliation (or the word 'freelance' if applicable), town/city, country, e-mail.

15. **References**: in alphabetical order. For quotes, write [Stevens 2010: 45] in the text. To refer to several sources, write [see Brown 1999; Robertson 2010; Smith 1957 and others]. All the references should be made in APA publication style.

(from https://inno-conf.mgimo.ru/info-eng.php)

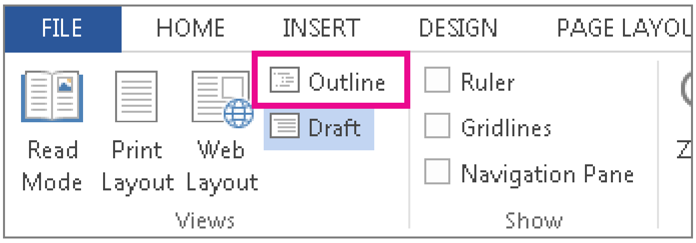
TUTORIAL

When working with a multi-page document you may find it difficult to navigate for specific information and creating a list of contents manually may turn into a real challenge. To simplify the work Microsoft Word offers you a number of helpful options. This tutorial will introduce you to some of them available in **Outline View**.

Create a document outline in Outline View.

1. To enter Outline view, click the View tab, and in the Views group, click the Outline button. The document's presentation changes to show Outline view, and the Outlining tab appears on the Ribbon, as shown. Now you can focus on the document structure.

2. Use Outline View to create or edit headings, adjust heading levels, and rearrange the content until everything is right where you want it. Click View > Outline. This automatically generates an outline and opens the Outlining tools.



If your document has headings (any heading levels from H1 – H9), you'll see those headings organized as a list. By moving between the parts create the document structure.

3. Create list of contents. As soon as you are ready with the document outline you can create an automated table of contents. Microsoft Word can scan your document and find everything in the Heading 1 style and put that on the first level of your table of contents, put any Heading 2's on the second level of your table of contents, and so on.

4. If you want an automatic table of contents you need to label all of your chapter titles and front matter headings in the style Heading 1. All major headings within your chapters should be labeled Heading 2. All subheadings should be labeled Heading 3, and so on.

5. If you have used Heading styles in your document, creating an automatic table of contents is easy. Place your cursor where you want your table of contents to be. On the References Ribbon, in the Table of Contents Group, click on the arrow next to the Table of Contents icon, and select Insert Table of Contents.

6. If you want to change the style of your table of contents (e.g. you want more space between the items on level 1 and level 2 of your table of contents, or you want all your level 1 items to be bold), click on the Modify button, select the TOC level you want to change, then click the Modify button to do so.

7. If you want to change which headings appear in your table of contents, you can do so by changing the number in the Show levels: pulldown. Click OK to insert your table of contents. The table of contents is a snapshot of the headings and page numbers in your document. At any time, you can update it by right-clicking on it and selecting Update field. Notice that once the table of contents is in your document, it will turn gray if you click on it. This indicates that it is getting information from somewhere else.

For formatting rules see Appendix 1, for report structure refer to Appendix 2, a sample tutorial report can be found in Appendix 3.

Unit 2

SEARCH ENGINES

Pre-reading tasks

Make sure you know Russian equivalents of the following words and word combinations.

Envision the invention of hypertext, proceed, open client-server design, protocol, carry payloads, encoded files, markup language, client, application, send a request, transmit data, domain, root of the hierarchy, path, fetch content, transmit data, send a request, markup, formatting rules, crawling, indexing, search engine, directory.

Introduction

The Web is unprecedented in many ways: unprecedented in scale, unprecedented in the almost-complete lack of coordination in its creation, and unprecedented in the diversity of backgrounds and motives of its participants. Each of these contributes to making web search different – and generally far harder – than searching "traditional" documents.

The invention of hypertext, envisioned by Vannevar Bush in the 1940's and first realized in working systems in the 1970's, significantly precedes the formation of the World Wide Web (which we will simply refer to as the Web), in the 1990's. Web usage has shown tremendous growth to the point where it now claims a good fraction of humanity as participants, by relying on a simple, open client-server design: (1) the server communicates with the client via a protocol (the http or hypertext transfer protocol) that is lightweight and simple, asynchronously carrying a variety of payloads (text, images and – over time – richer media such as audio and video files) encoded in a simple markup language called HTML (for hypertext markup language); (2) the client – generally a browser, an application within a graphical user environment – can ignore what it does not understand. Each of these seemingly innocuous features has contributed enormously to the growth of the Web, so it is worthwhile to examine them further.

The basic operation is as follows: a client (such as a browser) sends an http request to a web server. The browser specifies a URL (for Uniform Resource Locator) such as *http://www.stanford.edu/home/atoz/contact.html*. In this example URL, the string http refers to the protocol to be used for transmitting the data. The string *www.stanford.edu* is known as the domain and specifies the root of a hierarchy of web pages (typically mirroring a filesystem hierarchy underlying the web server). In this example, */home/atoz/contact.html* is a path in this hierarchy with a file contact.html that contains the information to be returned by the web server at *www.stanford.edu* in response to this request. The HTML-encoded file *contact.html* holds the hyperlinks and the content (in this instance, contact information for Stanford University), as well as formatting rules for rendering this content in a browser. Such an http request thus allows us to fetch the content of a page, something that will prove to be useful to us for crawling and indexing documents.

(from https://nlp.stanford.edu/IR-book/html/htmledition/background-and-history-1.html)

Search tools

When you're just getting started using the web, it can be quite overwhelming to understand exactly what tools are best to use to find what you may be looking for. The web is definitely a two-edged sword; while the availability of information is astonishing, it also can be quite intimidating if you do not know how to access it in a way that makes sense.

That is where basic tools come in that can help you organize information on the web into more meaningful channels. There are three basic types of search tools that most people use to find what they are looking for on the web (there's more than this, but these are the basics that everyone should start with):

Search Engines
Subject Directories
Meta Search Tools

None of these search tools allows you to search the entire web; that would be an almost impossible task. However, you can use these web search tools to scour different parts of the web, obtain different types of information, and broaden your web search horizons.

Search engines

To get a better idea of a search engine operation we will introduce you to their main functions

Crawling – A crawler, or web spider, is the part of a search engine that collects as much information about websites as they can. They search the Internet to find website addresses, content, and relevant keywords and links. It collects all this information and stores it in the search engine database.

Indexing – Once the search engine has gathered all the information, it indexes it according to specific keywords. Keywords are what people usually type in to find something or someone. It organizes the information and keywords for quick access. Search engines use algorithms to search the web according to certain parameters like keywords. The newest Google algorithm searches for more than just keywords and this is what SEO, SEM service providers take into consideration when designing websites, and marketing plans.

Storage – The information collected from websites is stored in the search engine database. This is important to make searching the web fast and easy. The size of the storage will determine how much information is available to Internet users.

Results – Results are what you get when typing in specific keywords in search of something. The crawler runs through its index selecting websites that match that keyword. The search engine algorithms look for the most relevant content, links, and keywords and rank the results accordingly. Different search engines will give different results because they do not use the same algorithm.

However, search results gathered by these search engines are not always relevant to the topic since these engines are limited by depth of indexing and due to the efforts of SEO and SMM specialists. This is where advanced search techniques come handy (see Boolean Search, and Advanced Search sections).

(from http://semadvisory.com/4-functions-of-internet-search-engines-to-help-you-understand-their-importance)

Subject Directories

Subject directories, in general, are smaller and more selective than search engines. They use categories to focus your search, and their sites are arranged by categories, not just by keywords. Subject directories are handy for broad searches, as well as finding specific websites. Most subject directories' main purpose is to be informational, rather than commercial. A translator may find useful specialized web directories focused on one field or subject for example *http://www.auto.mmt.ru/ (motor vehicles and transport), https://hro.org (human rights), http://samod.chat.ru/ (astronomy).*

Metasearch Engines

Metasearch engines process your request and send it to multiple search engines to get their search results from all of them. They use indexes built by other search engines, aggregating and often post-processing results in unique ways. Users will receive the best hits to their keywords from each search engine. Metasearch tools are a good place to start for very broad results but do not (usually) give the same quality results as using each search engine and directory. An example of metasearch engine is *https://nigma.eu/.*

Search in Translation

Internet search is an integral part of translation and interpreting. Whenever one gets an order, search is often the first step. One may need to get deeper into the topic, find out details about equipment mentioned, geographical objects and a good deal of other things to get background knowledge that enables you to make the translation exact and sound natural to the target audience.

Besides purely background value, web search is a handy tool to find equivalents. The best source of information on official data about companies, their structure and units, positions and job titles relevant to the sphere a company works with is their web sites easily accessible by simple search. Most of organizations and companies have their web sites in at least two languages, which makes it a treasure trove for a skilled translator.

Let us assume that you have used the web site to the full and have all the “official” equivalents at your fingertips. Now you have to do the hardest job translating the rest. No doubt, there will be hundreds of situations where you will be in two minds about the translation of a word or a word combination and you will need somebody or something to help you decide. Here millions of internet users may be helpful. By punching in each of the translation options in the target language you can see the number of results and which is even more valuable the context and situation where they are used. However, don't rely on this tool too much, especially when the choice deals with language norms. Sometimes the majority may be mistaken.

The like checkups are often necessary to make sure you have chosen a good word from a long list offered by a dictionary or verify any hypothesis you may have when translating.

There are sure to be times when you will feel "lost in translation" or simply lost, with no worthy idea coming up about some intricate case. Then you can check out translator's forums for similar cases or address the question to the translators' community. If you are not yet a regular with one, you can find them using web-search. Besides forums web-search can equip you with glossaries on specific topics and other useful linguistic resources that may come handy.

However to make an efficient query you should know some of the search tips and tricks.

Search techniques

1. Use unique, specific terms related to the subject you are researching.

2. Use the minus operator (-) to narrow the search.

Terms with multiple meanings can return many unwanted results. The rarely used but powerful minus operator, equivalent to a Boolean NOT, can remove many unwanted results. For example, when searching for the insect caterpillar, references to the company Caterpillar, Inc. will also be returned. Use Caterpillar -Inc to exclude references to the company or Caterpillar -Inc -Cat to further refine the search.

3. Use quotation marks for exact phrases.

One can often remember parts of phrases one has seen on a Web page or part of a quotation one wants to track down. Using quotation marks around a phrase will return only those exact words in that order. It's one of the best ways to limit the pages returned. Example: “Be nice to nerds”.

4. Don't use common words and punctuation.

Common terms like **a** and **the** are called stop words and are usually ignored. Punctuation is also typically ignored. But there are exceptions. Common words and punctuation marks should be used when searching for a specific phrase inside quotes. There are cases when common words like **the** are significant. For instance, Raven and The Raven return entirely different results.

5. Capitalization.

Most search engines do not distinguish between uppercase and lowercase, even within quotation marks. The following are all equivalent: technology, Technology, TECHNOLOGY, “technology”, “Technology”.

6. Drop the suffixes.

It is usually best to enter the base word so that you do not exclude relevant pages. For example, bird and not birds walk and not walked. One exception is if you are looking for sites that focus on the act of walking, enter the whole term walking.

7. Use browser history.

Many times, I will be researching an item and scanning through dozens of pages when I suddenly remember something I had originally dismissed as being irrelevant. If you can remember the general date and time of the search you can look through the browser history to find the Web page.

8. Set a time limit – then change tactics.

Sometimes, you never can find what you are looking for. Start an internal clock, and when a certain amount of time has elapsed without results, stop beating your head against the wall. It is time to try something else. Use a different search engine, like Yahoo!, Bing, Startpage, or Lycos or ask a peer. You can try calling support or asking a question in the appropriate forum.

9. Customize your searches.

There are several other less well-known ways to limit the number of results returned and reduce your search time.

The plus operator (+). As mentioned above, stop words are typically ignored by the search engine. The plus operator tells the search engine to include those words in the result set.

E.g.: tall +and short will return results that include the word and.

The tilde operator (~). Include a tilde in front of a word to return results that include synonyms. The tilde operator does not work well for all terms and sometimes not at all.

The wildcard operator (*). Google calls it the fill in the blank operator. For example, amusement * will return pages with amusement and any other term(s) the Google search engine deems relevant. In some cases, you can't use wildcards for parts of words. So for example, amusement p* may be invalid.

The OR operator (|). Use this operator to return results with either of two terms. For example happy joy will return pages with both happy and joy, while happy | joy will return pages with either happy or joy.

Numeric ranges. You can refine searches that use numeric terms by returning a specific range, but you must supply the unit of measurement. Examples: Windows XP 2003…2005, PC $700 $800.

Site search. Many Web sites have their own site search feature, but you may find that Google site search will return more pages. When doing research, it is best to go directly to the source, and site search is a great way to do that. Example: site: www.intel.com rapid storage technology.

Related sites. For example, related: www.youtube.com can be used to find sites similar to YouTube.

Change your preferences. Search preferences can be set globally by clicking on the gear icon in the upper-right corner and selecting Search Settings. I like to change the Number of Results option to 100 to reduce total search time.

Forums-only search. Under the Google logo on the left side of the search result page, click More | Discussions or go to Google Groups. Forums are great places to look for solutions to technical problems.

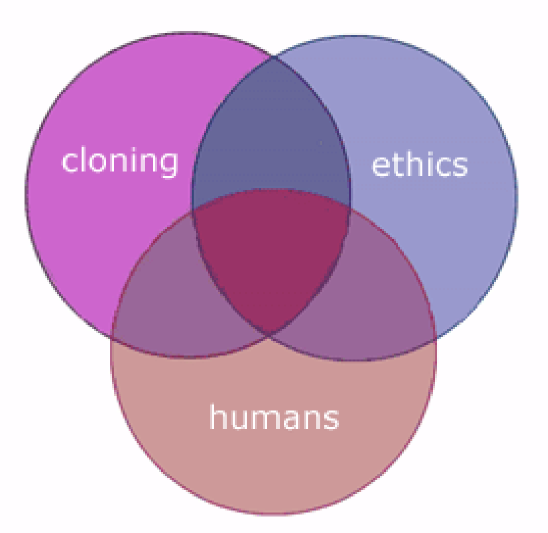
Advanced searches. Click the Advanced Search button by the search box on the Google start or results page to refine your search by date, country, amount, language, or other criteria.

(from https://www.techrepublic.com/blog/10-things/10-tips-for-smarter-more-efficient-internet-searching)

Boolean operators

Boolean operators form the basis of mathematical sets and database logic. They connect your search words together to either narrow or broaden your set of results. The three basic Boolean operators are: **AND**, **OR**, and **NOT**.

Use **AND** in a search to narrow your results. By using **AND** you tell the database that ALL search terms must be present in the resulting records.



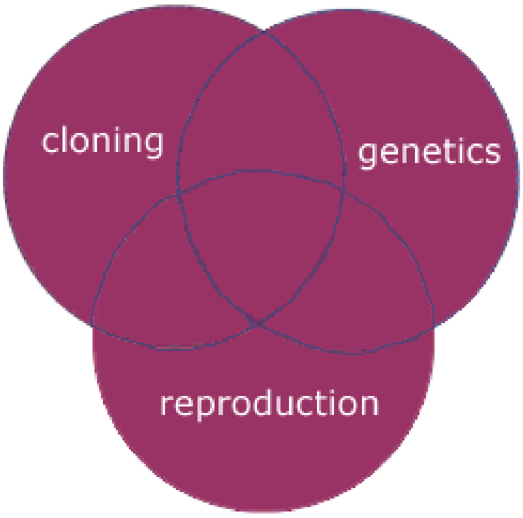
E.g.: cloning **AND** humans **AND** ethics.

The purple triangle in the middle of the diagram represents the result set for this search. It is a small set using **AND**, the combination of all three search words.

Be aware! In many, but not all, databases, the **AND** is implied. For example, Google automatically puts an **AND** in between your search terms. Though all your search terms are included in the results, they may not be connected together in the way you want.

E.g.: "college students test anxiety" is translated to "college **AND** students **AND** test **AND** anxiety". The words may appear individually throughout the resulting records, which is not exactly what you want.

To specify your request you can type in “college students” **AND** “test anxiety”. This way, the phrases show up in the results as you expect them to be.



Use **OR** in a search to connect two or more similar concepts (synonyms). It broadens your results, telling the database that ANY of your search terms can be present in the resulting records.

E.g.: cloning **OR** genetics **OR** reproduction.

All three circles represent the result set for this search. It is a big set because any of those words are valid using the **OR** operator.

Use **NOT** in a search to exclude words from your search thus narrowing your search, telling the database to ignore concepts that may be implied by your search terms. E.g.: cloning **NOT** sheep.

Databases follow commands you type in and return results based on those commands. Be aware of the logical order in which words are connected when using Boolean operators. Databases usually recognize AND as the primary operator, and will connect concepts with AND together first. If you use a combination of AND and OR operators in a search, enclose the words to be together in parentheses. E.g.: ethics AND (cloning OR reproductive techniques).

Truncation

Truncation, also called stemming, is a technique that broadens your search to include various word endings and spellings. To use truncation, enter the root of a word and put the truncation symbol at the end. The database will return results that include any ending of that root word.

E.g.: child* = child, child’s, children, children’s, childhood;
genetic* = genetic, genetics, genetically.

Common truncation symbols include: *, !, ?, or #

Wildcards substitute a symbol for one letter of a word. This is useful if a word is spelled in different ways, but still has the same meaning.

E.g.: wom!n = woman, women; colo?r = color, colour.

(from https://libguides.mit.edu/c.php?g=175963&p=1158679)

SELF-CHECK TASKS

***1. Give English equivalents for the following words and word combinations*:** отправить запрос, браузер, передавать данные, быстрый доступ, хранилище, тематический каталог, знать как свои пять пальцев, сужать поиск, расширять поиск, кавычки, заглавная буква, строчная буква, логические символы, настраивать поиск, расширенный поиск, специальный символ, выделение основы, звездочка, решетка, тильда, двоеточие, точка с запятой, многоточие, восклицательный знак, вопросительный знак, скобки.

2. Refer to ex. 1 and speak about the functions of each Boolean operator mentioned, its functions and use.

3. Open a search engine and describe the interface using English equivalents of all menu options.

4. Answer the questions.

1. When was the Web, as we know it today invented? What preceded its invention?
2. What is a client-server design?
3. How does a web-browser operate?
4. What information does a web-link normally contain?
5. What search tools do you know?
6. What are the functions of a search engine?
7. What is a metasearch engine?
8. What is a web directory?
9. How can a translator/interpreter use search engines?
10. What ways can help you to narrow your search?
11. How can one broaden the search enquiry?
12. What Boolean operators do you know and what is their meaning?

PRACTICE TASKS

1. ***Using web search, find English equivalents for the following word combinations:*** Департамент надзора за системно значимыми кредитными организациями; Отделение по Владимирской области Главного управления Центрального банка Российской Федерации по Центральному федеральному округу; председатель Комитета Государственной Думы по экономической политике, промышленности, инновационному развитию и предпринимательству; заместитель председателя правления Газпрома, генеральный директор службы корпоративной защиты; Аварийно-технический центр Минатома России, ФГУП, Дирекция единого заказа оборудования для АЭС (АО «ДЕЗ»).

2. When you translate you should convert measurements into units the target audience is used to. It depends on the country of origin or residence rather than on language. Fill in the table below. Use the web to convert the following.

Russia	UK	USA	Germany	Austria
+10 °C				
0 °C				
–10 °C				
101 см (диагональ ТВ)				
30 м				
100 м^2				
50 га				
90 км				
500 мл				
3 л				
5 кг				
2 т				
10 000 руб.				
38 (размер обуви)				
25 (размер детской обуви)				
44 (размер женской одежды)				
50 (размер мужской одежды)				
128 (размер детской одежды)				

3. Compile a list of specialized web directories for the following fields:

- soil study;
- medical equipment;
- machine tools;
- environmental studies.

TUTORIAL

1. Choose any two search engines and compare their advanced search options and Boolean operators. Practice search, analyze results and provide conclusions.

2. Translate the text below using search engines to clarify new notions (make notes).

3. Translate the text below. It is a call for papers for a conference. Use advanced search where necessary.

Московский государственный университет имени М. В. Ломоносова совместно с Международной ассоциацией преподавателей русского языка и литературы (МАПРЯЛ) проводит

V Международный конгресс исследователей русского языка

«Русский язык: исторические судьбы и современность»

18 – 21 марта 2019 г. на филологическом факультете МГУ имени М. В. Ломоносова

Состав Оргкомитета

Председатель Оргкомитета Конгресса – ректор МГУ академик РАН **Виктор Антонович Садовничий**

Заместитель Председателя Оргкомитета, Председатель Программного Комитета – декан филологического факультета МГУ, зав. кафедрой русского языка профессор **Марина Леонтьевна Ремнева**

Заместитель Председателя Оргкомитета – зав. лабораторией общей и компьютерной лексикологии и лексикографии профессор **Анатолий Анатольевич Поликарпов**

Ученый секретарь Оргкомитета – старший научный сотрудник **Суровцева Екатерина Владимировна**

Ответственный секретарь Оргкомитета по техническому сопровождению – **Александр Александрович Варламов**

Тематические направления работы конгресса

Актуальные теоретические аспекты современного развития русского языка

Русский язык в его истории и предыстории

Проблемы компьютерного и математического анализа русского языка

Русский язык в художественной литературе, деловой речи, Интернете и других родах и видах словесности

Русский язык как средство международного и межнационального общения в современных геополитических условиях

Лингвистические аспекты анализа традиционного и современного русского фольклора и обыденного языка

Проблемы преподавания русского языка как родного, неродного, иностранного

Проблемы лингвистической экспертизы русских текстов

Сроки подачи тезисов

Регистрация тезисов начинается с 20 августа 2018 г.

Тезисы принимаются до 11 ноября 2018 г. (включительно) по электронной почте rlc2014@philol.msu.ru.

Рецензирование присланных тезисов осуществляется по тезисам, авторство которых для рецензентов неизвестно. Для этого присылается два файла: в первом содержится название доклада и текст тезисов, но без указания их авторства («файл с текстом»); во втором указаны авторы и название доклада («файл с указанием автора, но без текста тезисов доклада» (он же – регистрационная форма)).

Решение о принятии или об отклонении тезисов на основе решения Программного Комитета будет сообщено к 12 декабря 2018 г.

Требования к оформлению тезисов

1) «Файл с текстом»

Тезисы присылаются по электронной почте только в виде приложения.

Форматы файлов: *.doc, *.docx или *.rtf.

Объем тезисов – не более 2 страниц (приблизительно 6500 печатных знаков).

Шрифт – Times New Roman, кегль – 12 pt.

Поля документа – 2,5 см (все четыре).

Междустрочный интервал – одинарный.

«Файл с текстом» оформляется следующим образом (каждый пункт – новый абзац):

Название доклада

Ключевые слова (не более 5 слов и/или словосочетаний)

Краткая аннотация на английском языке (1 – 4 фразы, до 500 печатных знаков)

Далее следует текст тезисов.

2) Образец «Файла с указанием автора, но без текста» (он же – регистрационная форма) можно скачать здесь.

Язык тезисов – русский. Тезисы доклада должны быть обязательно снабжены краткой аннотацией на английском языке.

Рабочие языки представления докладов на Конгрессе – русский, английский, немецкий, французский, испанский.

Публикация тезисов

Сборник тезисов будет опубликован к началу Конгресса и будет в распоряжении зарегистрированных участников Конгресса.

Добро пожаловать!

Заместитель Председателя Оргкомитета Конгресса, Председатель Программного Комитета Декан филологического факультета МГУ М. Л. Ремнева

Адрес Оргкомитета Конгресса

119991, Москва, ГСП-1, Ленинские горы, МГУ им. М. В. Ломоносова,
1-й корпус гуманитарных факультетов, комната 935
Тел./факс: +7 (495) 939-31-78
Электронный адрес: rlc2014@philol.msu.ru

For formatting rules see Appendix 1, for report structure refer to Appendix 2, a sample tutorial report can be found in Appendix 3.

Unit 3

ELECTRONIC DICTIONARIES

Pre-reading tasks

Make sure you know Russian equivalents of the following words and word combinations.

Headword, inflected languages, intransitive verbs, conjugator, explanatory dictionaries, pronunciation dictionaries, dictionary of idioms, PED, durable casing material, desktop dictionaries, cloud-based dictionaries.

An electronic dictionary is a dictionary whose data exists in digital form and can be accessed through a number of different media. Electronic dictionaries can be found in several forms, including software installed on tablet or desktop computers, mobile apps, web applications, and as a built-in function of E-readers. They may be free or require payment**.**

Most of the early electronic dictionaries were, in effect, print dictionaries made available in digital form: the content was identical, but the electronic editions provided users with more powerful search functions. But soon the opportunities offered by digital media began to be exploited. Two obvious advantages are that limitations of space (and the need to optimize its use) become less pressing, so additional content can be provided; and the possibility arises of including multimedia content, such as audio pronunciations and video clips.

Electronic dictionary databases, especially those included with software dictionaries are often extensive and can contain up to 500,000 headwords and definitions, verb conjugation tables, and a grammar reference section. Bilingual electronic dictionaries and monolingual dictionaries of inflected languages often include an interactive verb conjugator, and are capable of word stemming and lemmatization.

Publishers and developers of electronic dictionaries may offer native content from their own lexicographers, licensed data from print publications, or both, as in the case of Babylon offering premium content from Merriam Webster, and Ultralingua offering additional premium content from Collins, Masson, and Simon & Schuster, and Paragon Software offering original content from Duden, Britannica, Harrap, Merriam-Webster and Oxford.

(from https://en.wikipedia.org/wiki/Electronic_dictionary)

Today electronic dictionaries are in abundance, they come in various types and sizes, they differ in form, functions and types. So the choice may be quite challenging and should be made with regard to the purpose and task.

Just like conventional paper dictionaries, EDs can be monolingual and bi-lingual. The former are represented by dictionaries of synonyms and antonyms, explanatory dictionaries, pronunciation dictionaries, etc. The latter include general and specialized dictionaries containing equivalents for one language pair (= two languages). Unlike paper dictionaries which have two different sections for each language pair (e.g.: English-Russian and Russian-English), electronic dictionaries are made up of cross-coded entries which work equally fast in both directions and you don't need to switch.

If we take into account the form EDs come in, we can distinguish between the following types:

PED – portable (hadheld) electronic dictionaries. They are miniature laptop computers with full keyboards and LCD screen battery-powered and made with durable casing material. Besides translation their features include stroke order animations; voice output; handwriting recognition; language-learning programs; organizer functions; encyclopedias, time zone and currency converter. PEDs used to be popular, however due to advance of smartphone technologies and applications they have lost their popularity.

Smartphone apps – applications containing dictionary database which can be downloaded and installed on your phone. They may come in two formats: on-line and off-line. The former usually feature a good database with multiple options; however require internet access for operation. The latter require more storage capacity on your phone but work offline and are available anytime and anywhere.

Desktop dictionaries are installed on your PC. They can be downloadable or come on CD- or DVD-ROMs. Typically desktop dictionaries have good customizing options, allow to create user's dictionaries for different projects; contain grammar module to show all forms of a word; include both general and specialized dictionaries and have a learner's module to practice and memorize new words and word combinations.

Online dictionaries are so varied and numerous that can hardly be described without speaking about each sub-type individually.

Online dictionaries from prominent lexicographers and publishing houses. Most monolingual dictionaries contain explanation, examples; provide voice output alongside transcription, and some etymology data. Besides they have thesaurus which provides synonyms, antonyms and related words.

Monolingual online dictionaries

https://www.collinsdictionary.com/dictionary/english

https://www.lexico.com/en

https://www.oxfordlearnersdictionaries.com

https://www.merriam-webster.com/

Bilingual online dictionaries offer less options compared to monolingual ones and hardly ever contain etymology information.

https://dictionary.cambridge.org/ru

Dictionaries available free from non-commercial publishers

https://wooordhunt.ru/

Multi-dictionary web-sites

www.multitran.com

A new type of dictionaries incorporate bilingual dictionaries for a number of language pairs with search technologies that allow crawlers work with parallel texts, i.e. texts existing in two languages.

One example of such service is Reverso Context (*https:// context.reverso.net*) which is based on data gathered from millions of real-life texts (official documents, movie subtitles, product descriptions) in both languages. These texts are processed with powerful "big data" algorithms and machine learning to provide you the best results. Besides offering equivalents in a target language, it provides grammar advice as well as synonyms and offers opportunities for learning words with flashcards and games.

Another example is Linguee (*https://www.linguee.com*). Linguee uses specialized webcrawlers to search the Internet for appropriate bilingual texts and to divide them into parallel sentences. The paired sentences identified undergo automatic quality evaluation by a human-trained machine-learning algorithm that estimates the quality of translation. The user can set the number of pairs using a fuzzy search, access, and the ranking of search results with the previous quality assurance and compliance is influenced by the search term. Users can also rate translations manually, so that the machine learning system is trained continuously. In addition to serving the bilingual Web, Patent translated texts as well as the EU Parliament protocols and laws of the European Union (EUR-Lex) as sources. In addition to officially translated text from EU sources, its French language service relies on translated texts from Canadian government documents, websites, and transcripts, along with Canadian national institutions and organizations, which often provide bilingual services.

SELF-CHECK TASKS

1. Open a dictionary and describe the interface using English equivalents of all menu options.

2. Answer the following questions.

1. What is an electronic dictionary?
2. How did electronic dictionaries develop?
3. What typed of dictionaries do you know? Comment on their advantages and disadvantages.

PRACTICE TASKS

1. Register a user account in Multitran, explore its functions. Participate in forum discussions.

2. Use dictionaries to compile a glossary for the text below. Use search engine to clarify the meaning of terms. Use Translit for the List of references.

ПРИЕМЫ ОСНОВНОЙ ОБРАБОТКИ ПОД МНОГОЛЕТНИЕ ТРАВЫ В УСЛОВИЯХ ПОЧВЕННОЙ НЕОДНОРОДНОСТИ СЕРЫХ ЛЕСНЫХ ПОЧВ

В условиях значительной **внутрипольной пестроты пологоволнистого рельефа** почвенного покрова **Владимирского ополья** проведены исследования по изучению эффективности **приемов основной обработки** в **зернотравяном севообороте** под **многолетние травы** (**клевер** первого года пользования), их влияния на плодородие и урожайность. **Запасы продуктивной влаги** в метровом слое серых лесных и серых лесных со вторым **гумусовым горизонтом** почв в течение вегетации многолетних трав первого года пользования (клевер + **тимофеевка**) не зависели от глубины, системы приемов обработки почвы. В почве со вторым гумусовым горизонтом отмечено увеличение запасов продуктивной влаги. **Засоренность** клевера первого года пользования перед первым **укосом** (начало цветения клевера) не зависела от системы приемов основной обработки. **Общая засоренность** колебалась по вариантам опыта от 61 до 93 шт./м^2. Если для вариантов, расположенных на **серой лесной почве**, засоренность была на уровне 61 – 86 шт./м^2, то на серой лесной почве со вторым гумусовым горизонтом она составила 76 – 93 шт./м^2. На всех вариантах уровень засоренности превышал **порог экономической вредоносности**. Системы приемов основной обработки на почвенных разностях не обеспечивали **подавление сорняков** в посевах культуры ниже **уровня экологической вредоносности**. На серой лесной почве и серой лесной почве со вторым гумусовым горизонтом **урожайность** сена составила соответственно 42,6 – 49,4 ц/га и 50,8 – 55,5 ц/га. Наиболее высокие показатели урожая клевера как при первом, так и при втором укосе отмечались на вариантах, расположенных на серой лесной почве со вторым гумусовым горизонтом. Повышение урожайности культуры на серой лесной почве со вторым гумусовым горизонтом определяется ее более высоким **исходным плодородием**, формирующим благоприятные **водно-физические и биологические свойства**.

Список литературы

1. Звягинцев Д. Г., Полянская Л. М., Гончиков Г. Г., Корсунов В. М. Биомасса микроорганизмов в почвах Забайкалья // Почвоведение. 1999. № 9. С. 1132 – 1139.

2. Благовещенская Г. Г., Духанина Т. М. Микробные сообщества почв и их функционирование в условиях применения средств химизации // Агрохимия. 2004. № 2. С. 140 – 152.

3. Смагин А. В., Азовцева Н. А., Смагина М. В., Степанова А. Л., Мягкова А. Д., Курбатова А. С. Некоторые критерии и методы оценки экологического состояния почв в связи с озеленением городских территорий // Почвоведение. 2006. № 5. С. 603 – 615.
4. Балюк С. А., Мирошниченко Н. Н., Фатеев А. И. Принципы экологического нормирования допустимой антропогенной нагрузки на почвенный покров Украины // Почвоведение. 2008. № 12. С. 1501 – 1509.
5. Зинченко М. К., Зинченко С. И. Ферментативный потенциал агроландшафтов серой лесной почвы Владимирского ополья // Успехи современного естествознания. 2015. № 1 – 8. С. 1319 – 1323.
6. Зинченко М. К., Стоянова Л. Г. Реакция почвенной микрофлоры серой лесной почвы на длительное применение разных по уровню интенсификации систем удобрения // Достижения науки и техники АПК. 2016. Т. 30, № 2. С. 21 – 23.
7. Зинченко М. К., Шаркевич В. В., Федулова И. Д. Микробиологические аспекты адаптивно-ландшафтного земледелия в зоне Владимирского ополья // Владимирский земледелец. 2018. № 1. С. 14 – 19.
8. Теппер Е. З., Шильникова В. К., Переверзева Г. И. Практикум по микробиологии. М. : Дрофа, 2004. 255 с.

TUTORIAL

Describe and compare dictionaries of different types according to the following plan:

1) interface of home page;
2) structure of a dictionary entry (what information is provided, in what order, any hyperlinks to other resources);
3) search (advanced search) options;
4) settings and customizing options;
5) information it provides for a word;
6) additional functions and options.

For formatting rules see Appendix 1, for report structure refer to Appendix 2, a sample tutorial report can be found in Appendix 3.

Unit 4

CORPUS LINGUISTICS

Pre-reading tasks

Make sure you know Russian equivalents of the following words and word combinations.

On par with, naturally occurring, applied linguistics, obtain data, foreign language acquisition research, made-up examples, draw on large amounts of authentic, naturally occurring language data, prosodic marking, mark-up, tagging, term extraction, parallel corpora.

Introduction

A corpus (plural corpora, German “das Korpus”, not “der”) is a collection of texts used for linguistic analyses, usually stored in an electronic database so that the data can be accessed easily by means of a computer. Such corpora generally comprise hundreds of thousands to billions of words and are based on authentic naturally occurring spoken or written usage.

Based on the above definition of a corpus, corpus linguistics is the study of language by means of naturally occurring language samples; analyses are usually carried out with specialised software programmes on a computer. Corpus linguistics is thus a method to obtain and analyse data quantitatively and qualitatively rather than a theory of language or even a separate branch of linguistics on a par with e.g. sociolinguistics or applied linguistics.

The corpus-based approach can be used to describe language features and to test hypotheses formulated in various linguistic frameworks. To name but a few examples, corpora recording different stages of learner language (beginners, intermediate, and advanced learners) can provide information for foreign language acquisition research; by means of historical corpora it is possible to track the development of specific features in the history of English, such as changes in the use of the modal verb must and the emergence of alternatives such as have to or have got to;

the emergence of the modal verbs "gonna" and "wanna"; or sociolinguistic markers of specific age groups, such as the use of like as a discourse marker, can be investigated for purposes of sociolinguistic or discourse-analytical research.

The great advantage of the corpus-linguistic method is that language researchers do not have to rely on their own or other native speakers' intuition or even on made-up examples. Rather, they can draw on large amounts of authentic, naturally occurring language data produced by a variety of speakers or writers in order to confirm or refute their own hypotheses about specific language features on the basis of a robust and solid empirical foundation.

The majority of present-day corpora are "balanced" or "systematic". This means that the texts are collected ("compiled") according to specific principles, such as different genres, registers, or styles of English (e.g. written or spoken English, newspaper editorials or technical writing); these sampling principles do not follow language-internal but language-external criteria. For example, the texts for a corpus are not selected because of their high number of relative clauses, but because they are instances of a predefined text type, say broadcast English, magazine or newspaper texts.

(from https://www.anglistik.uni-freiburg.de/ seminar/abteilungen/sprachwissenschaft/ls_mair/corpus-linguistics)

Background

The use of collections of text in language study is not a new idea. In the Middle Ages work began on making lists of all the words in particular texts, together with their contexts – what we today call "concordancing". Other scholars counted word frequencies from single texts or from collections of texts and produced lists of the most frequent words. Areas where corpora were used include language acquisition, syntax, semantics, and comparative linguistics, among others. Even if the term "corpus linguistics" was not used, much of the work was similar to the kind of corpus based research we do today with one great exception – they did not use computers. Today, corpus linguistics is closely connected to the use of computers; so closely, actually, that the term "Corpus Linguistics" for many scholars today means “the use of collections of computer-readable text for language study”.

The Brown Corpus

The first modern, electronically readable, corpus was *The Brown Corpus of Standard American English*. The corpus consists of one million words of American English texts printed in 1961. To make the corpus a good standard reference, the texts were sampled in different proportions from 15 different text categories: Press (repotage, editorial, reviews), Skills and Hobbies, Religious, Learned/scientific, Fiction (various subcategories), etc.

Today, this corpus is considered small, and slightly dated. The corpus is, however, still used. Much of its usefulness lies in the fact that the Brown corpus lay-out has been copied by other corpus compilers. *The LOB, Lancaster-Oslo-Bergen, Corpus* (British English) and the *Kolhapur Corpus* (Indian English) are two examples of corpora made to match the Brown corpus. They both consist of 1 million words of written language (500 texts of 2,000 words each), sampled in the same 15 categories as the Brown Corpus. For a long time, the Brown and LOB corpora were the only easily available computer readable corpora. Much research within the field of corpus linguistics has therefore been based on these corpora.

The London-Lund Corpus of Spoken British English

Another important “small” corpus is the *London-Lund Corpus of Spoken British English* (LLC). The corpus was the first computer readable corpus of spoken language, and it consists of 100 spoken texts of appr. 5,000 words each. The texts are classified into different categories, such as spontaneous conversation, spontaneous commentary, spontaneous and prepared oration, etc. The texts are orthographically transcribed and have been provided with detailed prosodic marking.

BoE and BNC

The first generation corpora, of 500,000 and 1 million words, proved to be very useful in many ways and have been used for a number of research tasks. It soon turned out, however, that for certain tasks, larger collections of text were needed. Dictionary makers, for example wanted large, up-to-date collections of text where it would be possible to find not only rare words but also new words entering the language.

In 1980, COBUILD started to collect a corpus of texts on computer for dictionary making and language study. The compilers of the Collins Cobuild English Language Dictionary (1987) had daily access to a corpus of approximately 20 million words. New texts were added to the corpus, and in 1991 it was launched as the *Bank of English* (BoE). New material is constantly added to the corpus to make it reflect the mainstream of current English today. A corpus of this kind, which by the new additions 'monitors' changes in the language, is called "a monitor corpus". Some people prefer not to use the term "corpus" for text collections that are not finite but constantly changing/growing.

The British National Corpus (BNC) was originally created by Oxford University press in the 1980s – early 1990s. It is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written. The latest edition is the BNC XML Edition, released in 2007.

The written part of the BNC (90 %) includes, for example, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text. The spoken part (10 %) consists of orthographic transcriptions of unscripted informal conversations (recorded by volunteers selected from different age, region and social classes in a demographically balanced way) and spoken language collected in different contexts, ranging from formal business or government meetings to radio shows and phone-ins.

Work on building the corpus began in 1991, and was completed in 1994. No new texts have been added after the completion of the project but the corpus was slightly revised prior to the release of the second edition BNC World (2001) and the third edition BNC XML Edition (2007). Since the completion of the project, two sub-corpora with material from the BNC have been released separately: the BNC Sampler (a general collection of one million written words, one million spoken) and the BNC Baby (four one-million word samples from four different genres).

BNC can be described with the following characteristics.

Monolingual. It deals with modern British English, not other languages used in Britain. However non-British English and foreign language words do occur in the corpus.

Synchronic. It covers British English of the late twentieth century, rather than the historical development which produced it.

General. It includes many different styles and varieties, and is not limited to any particular subject field, genre or register. In particular, it contains examples of both spoken and written language.

Sample. For written sources, samples of 45,000 words are taken from various parts of single-author texts. Shorter texts up to a maximum of 45,000 words, or multi-author texts such as magazines and newspapers, are included in full. Sampling allows for a wider coverage of texts within the 100 million limit, and avoids over-representing idiosyncratic texts.

(from http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=creation)

The Corpus of Contemporary American English (COCA)

https://www.english-corpora.org/coca/

It is the only large, genre-balanced corpus of American English. COCA is probably the most widely-used corpus of English. The corpus contains more than one billion words of text (25+ million words each year 1990 – 2019) from eight genres: spoken, fiction, popular magazines, newspapers, academic texts, TV and movies subtitles, blogs, and other web pages.

There are four main ways to search the corpus.

First, you can browse a frequency list of the top 60,000 words in the corpus, including searches by word form, part of speech, ranges in the 60,000 word list, and even by pronunciation. This should be particularly useful for language learners and teachers.

Second, you can search by individual word, and see collocates, topics, clusters, websites, concordance lines, and related words for each of these words. Note that some of these searches are unique to COCA and iWeb.

Third you can input entire texts and then use data from COCA to get detailed information on the words and phrases in the text.

Fourth, you can search for phrases and strings. And because the corpus is optimized for speed, searches for substrings (*ism, un*able) and phrases are very fast, e.g.: got VERB-ed, BUY * ADJ NOUN, "gorgeous" NOUN – and even high frequency phrases like: from ADJ to ADJ, phrasal verbs, or NOUN NOUN.

You might pay special attention to the comparisons between genres and years and virtual corpora, which allow you to create personalized collections of texts related to a particular area of interest.

Specialized corpora

Historical corpora. The use of collections of text in the study of language is, as we have seen, not a new invention. Among those involved in historical linguistics were some that soon saw the potential usefulness of computerized historical corpora. A diachronic corpus with English texts from different periods was compiled at the University of Helsinki. *The Helsinki Corpus of English Texts* contains texts from the Old, Middle and Early Modern English periods, 1.5 million words in total.

Another historical corpus is the recently released *Lampeter Corpus of Early Modern English Tracts*. This collection consists of Pamphlets and tracts published in the century between 1640 and 1740 from six different domains. The Lampeter Corpus can be seen as one example of a corpus covering a more specialized area.

Corpora for Special Purposes. The corpora described above are general collections of text, collected to be used for research in various fields. There is a large, and growing, amount of highly specialized corpora that are created for a special purpose. Many of these are used for work on spoken language systems. Examples of such are, for example, the *Air Traffic Control Corpus, ATCC*, created to be used "in the area of robust speech recognition in domains similar to air traffic control" and the *TRAINS Spoken Dialogue Corpus* collected as part of a project set up to create "a conversationally proficient planning assistant" (railroad freight system).

A number of highly specialized corpora are held at the Centre for Spoken Language Understanding, CSLU, in Oregon. These corpora are specialized in a different way to the ones mentioned above. They are not restricted to be used within a particular subject field, but are called specialized because their content. Many of the corpora/databases consist of recordings of people asked to perform a particular task over the telephone, such as saying and spelling their name or repeating certain words/phrases/numbers/letters.

International/multilingual Corpora

As we have seen above, there is a great variety of corpora in English. So far much corpus work has indeed concerned the English language, for various reasons. There are, however, a growing number of corpora available in other languages as well. Some of them are monolingual corpora – collections of text from one language. Here the Oslo Corpus of Bosnian text and the Contemporary Portuguese Corpus can be mentioned as two examples.

A number of multilingual corpora also exist. Many of these are parallel corpora; corpora with the same text in several languages. These corpora are often used in the field of Machine Translation. The English-Norwegian Parallel Corpus is one example, the English Turkish Aligned Parallel Corporaanother.

Ongoing projects

ICE: the International Corpus of English. In twenty centres around the world, compilers are busy collecting material for the ICE corpora. Each ICE corpus will consist of 1 million words (written and spoken) of a national variety of English. The first ICE corpus to be completed is the British component, ICE-GB. On their own, the ICE corpora will be a small but valuable resources to exploit in order to learn about different varieties of English. As a whole, the 20 corpora will be useful for variational studies of various kinds. You can learn more about the ICE project at the ICE-GB site.

ICLE: the International Corpus of Learner English. Like ICE (see above) ICLE is an international project involving several countries. Unlike ICE, however, the ICLE corpora do not consist of native speaker language. Instead they are corpora of English language produced by learners in the different countries. This will constitute a valuable resource for research on second language acquisition.

(from https://www1.essex.ac.uk/linguistics/external/ clmt/w3c/corpus_ling/content/history.html)

To sum up we can come with the following list of some of the most common types of corpora is provided.

a) General corpora, such as the British National Corpus, contain a large variety of both written and spoken language, as well as different text types, by speakers of different ages, from different regions and from different social classes.

b) Synchronic corpora, such as F-LOB and Frown, record language data collected for one specific point in time, e.g. written British and American English of the early 1990s.

c) Historical (or diachronic) corpora, such as ARCHER and the Helsinki corpus, consist of corpus texts from earlier periods of time. They usually span several decades or centuries, thus providing diachronic coverage of earlier stages of language.

d) Learner corpora, such as the International Corpus of Learner English and the Cambridge Learner Corpus, are collections of data produced by foreign language learners, such as essays or written exams.

e) Corpora for the study of varieties, such as the International Corpus of English and the Freiburg English Dialect Corpus, represent different regional varieties of a language

f) Specialized corpora, e.g. the Michigan Corpus of Academic Spoken English (MICASE), are useful for various types of research (cf. e.g. http://www.helsinki.fi/varieng/CoRD/corpora/index.html).

(from https://www.anglistik.uni-freiburg.de/seminar/abteilungen/ sprachwissenschaft/ls_mair/corpus-linguistics)

The list of corpora presented in this section is not in the least a complete one. You can access a number of other highly valuable corpora here: *https://www.english-corpora.org*

Modern Corpora

The Russian National Corpus covers primarily the period from the middle of the 18th to the early 21st centuries. This period represents the Russian language of both the past and the present in a wide range of sociolinguistic variants: literary, colloquial, vernacular, in part dialectal.

The Corpus includes original (non-translated) works of fiction (prose, drama and poetry) of cultural importance which are interesting from a linguistic point of view. Apart from fiction, the Corpus includes a large volume of other sources of written (and, for the later period, spoken) language: memoirs, essays, journalistic works, scientific and popular scientific literature, public speeches, letters, diaries, documents, etc.

The RNC includes primarily original prose representing standard Russian (from the middle of the 18th century) but also, albeit in smaller volumes, translated works (parallel with the original texts) and poetry, as well as texts, representing the non-standard forms of modern Russian: spoken (recordings of oral speech, spontaneous and public) and dialectal.

The main corpus

The main corpus, which includes texts representing standard Russian, can be subdivided into 3 parts, each of which has its distinguishing features: modern written texts (from the 1950s to the present day), a subcorpus of real-life Russian speech (recordings of oral speech from the same period), and early texts (from the middle of the 18th to the middle of the 20th centuries). By default, the search is carried out in all the three sub-groups. It is possible to choose one of them and add search parameters on the "customize your corpus" page.

Every text included in the main corpus is subject to meta tagging and morphological tagging. Morphological tagging is carried out by computer programs for automated morphological analysis. In a small part of the main corpus (currently around 5 million tokens; this figure is set to increase with time) homonyms are disambiguated by hand and the results of automated morphological analysis corrected. This part is the model morphological corpus and serves as a testing ground for various search algorithms and programs of morphological analysis and automated processing. It can also be used for research on modern Russian morphology that requires particular preciseness. Examples of this subcorpus are annotated as "disambiguated" ("омонимия снята"). Disambiguated texts are automatically supplied with indicators of stress (from the Grammatical dictionary of Russian). Stress annotation may be turned off for printing or saving the search results.

Modern written texts

The representative corpus of morphologically tagged modern texts is the main and the largest of the subcorpora. The planned volume of the corpus is 100 million tokens. The corpus includes various types of texts representing modern standard (written) Russian:

- modern fiction of various genres;
- modern drama;
- memoirs and biographies;
- journalism and literary criticism;
- scientific, popular scientific and teaching texts;
- religious and philosophical texts;
- technical texts;
- business and jurisprudence texts;
- day-to-day life texts, including texts not intended for publication (letters, diaries, etc.).

Texts are represented in proportion to their share in real-life usage. For example, the share of fiction (including drama and memoirs) does not exceed 40 %.

The sources of book, magazine and newspaper texts included in the Corpus are usually proof-read electronic versions supplied by their respective publishers and the texts are used with publishers' permission.

The search can be limited to modern texts in the Date of creation field of the Customize your corpus page.

Mid-18th to mid-20th century texts

Texts from the middle of the 18th century to the middle of the 20th century are also included in the Corpus and represent various genres (fiction, scientific texts, journalism, letters) but due to limited availability of such texts in electronic form or in modern reprints the proportion of fiction for this period is much higher than for the main corpus. Pre-1918 texts are given in modern orthography; peculiarities of their original orthography preserved in modern academic editions are also preserved in the Corpus.

Deeply Annotated Corpus. This subcorpus of the RNC contains texts augmented with morphosyntactic annotation. Besides the morphological information ascribed to each word in the text, every sentence has its syntax structure marked up.

The Deeply Annotated Corpus (DAC) uses dependency trees as its annotation formalism. Nodes in such a tree are words of the sentence, while its edges are labeled with names of syntax relationships. This way of representing the syntax structure originates from “Meaning ⇔ Text” linguistic model by Igor A. Mel’čuk and Alexander K. Zholkovsky. The repertory of syntactic relationships for the DAC, as well as other specific linguistic decisions on how to represent the syntax of Russian sentences, has been developed in the Laboratory for Computational Linguistics, Institute for Information Transmission, Russian Academy of Sciences that compiled the DAC.

Unlike the morphologically annotated portion of the RNC, the DAC only contains fully disambiguiated annotations (i.e. both morphological and syntax ambiguity is resolved).

Parallel text corpus. The parallel text corpus is a special type of corpus where a text in Russian is complemented by its translation into a different language, and vice versa. The units of the original and the translated texts (usually, a unit is a sentence) are matched through a procedure known as “leveling”. A leveled parallel corpus is an important tool for various type of research, including studies on the theory of translation; it can also be used as a language teaching tool.

This site contains the following parallel text corpora: English-Russian, Russian-English, German-Russian, Ukrainian-Russian, Russian-Ukrainian, Belorussian-Russian, Russian-Belorussian, and multilingual.

Dialectal corpus. The dialectal corpus contains recordings of dialectal speech (presented in loosely standardized orthography) from different regions of Russia. There is no intention to present the phonetic variation, but morphological, syntactic and lexical peculiarities of these texts are preserved. The subcorpus employs special tags for specifically dialectal morphological features (including those absent in standard language); moreover, purely dialectal lexemes are supplied with commentary.

Poetry corpus. At the moment the poetry corpus covers the time frame between 1750 and 1890s, but also includes some poets of the 20th century; currently, works of drama composed in poetry are not included. Apart from the usual morphological tagging (identical to that available for the non-disambiguated corpus), there is a number of tags adapted for poetry. For example, it is possible to search for texts written in various poetic meters such as amphibrach.

Educational corpus. The educational corpus is a small disambiguated corpus adapted for the Russian educational program, including works of fiction on the school reading list and several additional morphological features.

Corpus of Spoken Russian

The Corpus of Spoken Russian includes the recordings of public and spontaneous spoken Russian and the transcripts of the Russian movies. To record the spoken specimens the standard spelling was used. The lexical, morphological and semantic queries are practicable. The building of the user's sub-corpora is available (for this purpose the usage of the sociological parameters is also possible). The corpus contains the patterns of different genres/types and of different geographic origins (Moscow, Sanct-Peterburg, Saratov, Ulyanovsk, Taganrog, Ekaterinburg, and so on). The corpus covers the time frame from 1930 to 2007.

Corpus managers

In order to analyse a corpus and search for certain words or phrases (strings), you can either access the data via an online user interface or, if none is provided, need to use special software – so-called concordancers or corpus managers like AntConc.

Corpus manager (corpus browser or corpus query system) is a tool for multilingual corpus analysis, which allows effective searching in corpora.

A corpus manager usually represents a complex tool that allows one to perform searches for language forms or sequences. It may provide information about the context or allow the user to search by positional attributes, such as lemma, tag, etc. These are called concordances. Other features include the ability to search for Collocations, frequency statistics as well as metadata information about the processed text. The narrower meaning of corpus manager refers only to the server side or the corpus query engine, whereas the client side is simply called the user interface. A corpus manager can be software installed on a personal computer or it might be provided as a web service.

(*from https://en.wikipedia.org/wiki/Corpus_manager)*

Sketch Engine

https://www.sketchengine.eu

Sketch Engine is a multifunctional toolkit for processing big corpora.

Sketch Engine is a corpus manager and text analysis software developed by Lexical Computing Limited since 2003. Its purpose is to enable people studying language behaviour (lexicographers, researchers in corpus linguistics, translators or language learners) to search large text collections according to complex and linguistically motivated queries. Sketch Engine gained its name after one of the key features, word sketches: one-page, automatic, corpus-derived summaries of a word's grammatical and collocational behaviour. Currently, it supports and provides corpora in 90+ languages.

Features

- Word sketches – a one-page automatic derived summary of a word’s grammatical and collocational behaviour
- Word sketch difference – compares and contrasts two words by analysing their collocation
- Distributional Thesaurus – automated thesaurus finding words with similar meaning or appearing in the same/similar context
- Concordance search – finds examples of a word form, lemma, phrase, tag or complex structure
- Collocation search – word co-occurrence analysis displaying the most frequent words (to a search word) which can be regarded as collocation candidates
- Word lists – generates frequency lists which can be filtered with complex criteria
- n-grams – generates frequency lists of multi-word expressions

- Terminology/Keyword extraction (both monolingual and bilingual) – automatic extraction key words and multi-word terms from texts (based on frequency count and linguistic criteria)
- Diachronic analysis (Trends) – detecting words which undergo changes in the frequency of use in time (show trending words)
- Corpus building and management – create corpora from the Web or uploaded texts including part-of-speech tagging and lemmatization which can be used as data mining software
- Parallel corpus (bilingual) facilities – looking up translation examples (EUR-Lex corpus, Europarl corpus, OPUS corpus, etc.) or building parallel corpus from own aligned texts

(from https://en.wikipedia.org/wiki/Sketch_Engine)

Translators can make use of many features of Sketch Engine related term extraction, translation lookup, grammar and usage checking.

Term extraction for translators

Translators can easily fill their CAT tool term base with terminology for maintaining consistency and quality across all translation jobs from the same area or the same client. Sketch Engine's sophisticated term extraction or bilingual term extraction uses state-of-the-art natural language processing know-how to identify terminology. Statistical analysis is aided by comparing the use of words in the user's text to their use in a large multi-billion word corpus of general text. The results include both single- and multi-word units and can be exported into your CAT tool in a widely supported TBX format.

Term extraction for interpreters

You can prepare for your interpreting job even if the client did not provide any supporting materials. You can have Sketch Engine search the internet for you and download texts related to the topic you specify. Then use the term extraction feature to extract terminology. This process can be repeated many times to create a large domain-specific corpus and extract as many terms as needed.

Translation lookup

Get inspiration from other translators. Use parallel corpora to look up translations of words or phrases as they were translated by others and chose the option that fits your context best. Useful especially for idiomatic expressions or any other situations when bilingual dictionaries fail to provide a solution.

Grammar and usage checking

Use the concordance feature to find examples of the word or phrase in context and see how it is used by real users of the language. Compare how frequently different alternatives appear in the corpus to assess which is the most natural to use in your context.

Use Word Sketch for any high-frequency expressions to get a one-page overview of words frequently used together with the word. The collocations are sorted according to grammatical categories such as modifiers, objects of verbs etc. for easy use.

Alternative word suggestions

A printed thesaurus is always limited by space. The automatic distributional thesaurus in Sketch Engine can provide synonyms, antonyms and related words for almost any word in a language (provided the corpus is large enough to contain enough occurrences of the word but Sketch Engine contains some of the biggest corpora in many languages). Use the thesaurus to quickly view alternatives of the word you are about to use.

Additional content for the term base

Upload your translation memory and usen-grams to identify multi-word expressions which may not be normally labelled as terms but frequently appear in the texts you translate. Some examples in English might be as a matter of fact, at the beginning of etc. Add them to your term base to have them translated automatically by your CAT tool.

(from https://www.sketchengine.eu/user-guide/ translators-term-extraction/)

Just the word

http://www.just-the-word.com

A simple application based on BNC to fetch collocations and word combinations with a search word. Just the Word is a handy tool to check up combinability of a word, choose a proper preposition suitable for a specific context. By clicking on a word combination you are redirected to BNC to see the context of each word combination. You can also see statistics and general representation of combinability models.

SELF-CHECK TASKS

***1. Give English equivalents for the following words and word combinations*:** язык в его естественном виде, количественный и качественный анализ, корпусный подход, овладение иностранным языком, отследить развитие, появление вариантов, маркеры дискурса, придуманные примеры, подтвердить или отвергнуть гипотезу, рационально организованный, принципы отбора, собственно языковые критерии, экстралингвистические критерии, составление алфавитного списка слов, частотность, разметка, словоупотребление (2), представительный корпус, снята неоднозначность.

2. Answer the questions.

1. What is a corpus?
2. What is corpus linguistics?
3. What for can we use corpus-based approach?
4. What historical/specialized/multilingual corpora do you know? What are they famous for?
5. What sort of corpus is the BNC?
6. What is a corpus manager?
7. What are the main features of a corpus manager?
8. What is concordance?
9. What is an n-gram?
10. In what ways can corpora and corpus-based approach be used in translation?

PRACTICE TASKS

1. Analyze the use of nerd vs. geek. Compare the use and collocates of nerd and geek in COCA and BNC. When comparing mention frequency, common collocates, topics and types of discourse they are used in. Suggest differences in meaning.

2. Use GloWbE to determine in which varieties of English the following words or expressions are the most common. Can you see traces of borrowing into other varieties? Can you figure out their meaning from the context?

- lah
- vex
- dunny
- tuque or toque
- lekker
- dinkum
- good on (pronoun)
- sleveen
- speed money
- tai tai

TUTORIAL

1. In this tutorial you are going to learn creating our own corpora using AntConc. Find and download a file with the programme. Before doing the tasks of the laboratory work, watch tutorials by Anthony Lawrence, the main developer of this software.

1) https://www.youtube.com/watch?v=9TsqFVrUYO0 – basic features

2) https://www.youtube.com/watch?v=_z9wwX7eR-Y – how to start your work in AntConc

3) https://www.youtube.com/watch?v=uAYCA8dYbr4 – how to use concordance tool

4) https://www.youtube.com/watch?v=2rvsBaM6W8Y – advanced features of concordance tool

5) https://www.youtube.com/watch?v=Zb71yaBP_lI – word list tool

6) https://www.youtube.com/watch?v=JvasTvQY7kU – key word list tool

2. The task for the tutorial is to create a glossary based on a created corpus using a corpus manager. In order to do it you should follow the steps below.

1. Create a corpus using the texts attached:

a) download the texts;

b) convert them into .txt format;

c) make sure the encoding is Юникод (UTF-8);

d) when you upload the texts in AntConc, check if it works by inserting any word in the box Search Term (for example почв*). If you see in the concordance box all the contexts for the search word, you have successfully made a corpus.

2. Work with the Word list tool to find the most frequently used words and word combinations:

a) go to the Word list tab and create a word list;

b) you will see that on the top of the list are articles, prepositions, conjunctions, etc. that are no value for our glossary;

c) in order to get rid of them click on the Tool Treferences, then choose the category Word list and add the most frequent but unnecessary words to the stoplist (choose Use the stoplist below): add at least 10 words, after entering every word click Add. When you have entered all the 10 (or more) words click Apply;

d) click Start in the Word List tab. If you see the word list without words you have added to the stoplist, congratulations!!!

3. Work with the most frequent words which are on the top of the list in the Word list tab. Concordance tab to make up a list of terms and word combinations to compile a glossary:

a) by clicking on each word you will see all the contexts in the Concordance tab;

b) choose the most frequent word combinations and copy-paste them into your glossary (a .doc or .docs file);

c) use search engines to find explanations and definitions of the words and terms for your glossary.

4. Use dictionaries to find equivalents. For word combinations you cannot find equivalents in dictionaries, make your own hypothesis and use search engines or corpora to verify your hypothesis or choose between several ones.

5. Make a glossary of at least 30 (probably more) terms and word combinations as a table with 3 columns (source term – equivalent – explanation or example). Make sure that explanations as well as examples are taken from originally English web-sites.

6. Make up a report based on the work.

7. Attach the resulting glossary as Appendix.

For formatting rules see Appendix 1, for report structure refer to Appendix 2, a sample tutorial report can be found in Appendix 3.

Unit 5

TRANSLATION MEMORY

Pre-reading tasks

Make sure you know Russian equivalents of the following words and word combinations.

Ensure the consistency, a match, table entry content, fuzzy matches, configurable settings, keypunched, text alignment, retrieval, concordance search, UI, in-house solution, autosuggest, customizing.

Introduction

A translation memory (TM) is basically a database of previously translated segments for CAT (computer-assisted translation) tools. Different CAT tools use different TM formats, but most of them can be converted. It’s typically one source language and one or more target languages.

TMs work at the sentence level. CAT tools break down source documents into their component segments. The segment is the smallest reusable chunk of text. A segment can be equal to a sentence, or it can be a heading, an element in a list or table entry content. Words are not used for this purpose because different contexts require different translations.

As the translator works, the current segment is compared to those in the translation memory, and if it has something very similar, the CAT tool will automatically show this to the translator. Identical source segments are called 100 % matches. This means that somebody in the past had already translated that exact segment. There are also 101 % and 102 % matches, which means that not only the current segment, but also one or both of those before and/or after it are the same as stored in this TM entry. Matches below 100 % are called fuzzy matches. These are ranked from 0 % to 99 %. A 99 % match means that the segments differ by at least one character. Matches below 70 % are often considered useless and might not show up, depending on the settings.

"Repetitions" are identical segments within one document that have no translation in the translation memory yet. Most CAT tools scan for repetitions before the translator starts working. After the translator is done with the first occurrence of these, all others will get filled in automatically.

Translation memory statistics is a breakdown of how many 100 %+ matches, fuzzy matches, repetitions, and new text the file has. Typically, each category has a different price applied to it. For example, a 99 % match can be priced at 10 % of the default per-word rate, while a 75 % match can be priced at 40 %. These settings are configurable in most CAT tools. Words "discounted" according to such percentages are sometimes referred to as "weighted" words. The rationale behind this is that a translator works less on matching segments, thus a lowered per-word rate nevertheless results in the same or higher per-hour rate. This is considered a fair deal and is practiced by almost every translation agency and customer nowadays.

Translation memories reduce the price for clients when there is a lot of repetitive content across their past translations while reducing the time required to finish a project.

But most importantly, translation memories ensure the consistency and hence quality of translation. Besides matches as such, you can use your existing translation memories for concordance searches – where you investigate if the translation memory has a certain term was translated before, if the segment as a whole wasn't. Note that for even more consistent terminology management it might be better to use glossaries instead of translation memories.

There are a lot of translation memory tools. These include Wordfast, Trados, Smartcat, Memsource, MemoQ, Across, etc. Some language service providers even make their own translation tools, but these are typically of lower quality and only provide the most basic functions.

(from https://www.smartcat.ai/articles/what-is-translation-memory/)

Historical Background

In 1966, the Automatic Language Processing Advisory Committee (ALPAC) published an influential report called "Language and Machines", which concluded that future prospects for machine translation were limited. In this same report, there is a short description of a system used by the European Coal and Steel Community (CECA) that seems to qualify as an early TM system. The report describes CECA's system as

> automatic dictionary look-up with context included. […] [T]he translator indicates, by underlining, the words with which he desires help. The entire sentence is then keypunched and fed into a computer. The computer goes through a search routine and prints out the sentence or sentences that most clearly match (in lexical items) the sentences in question. The translator then retrieves the desired items printed out with their context and in the order in which they occur in the source[1].

What is interesting about the system is its bilingual output. Although this particular system was intended primarily for terminological research, the process includes the elements of text alignment, automatic matching and retrieval, and keeping terms in their contexts, thus anticipating many of the essential features of modern systems.

In 1978, Peter Arthern filled in some of the blanks by expanding the idea of a "translation archive". His vision included the storage of all previously translated texts the ability to retrieve quickly and insert into

[1] ALPAC Report. Language and Machines – Computers in Translation and Linguistics. A Report by the Automatic Language Processing Advisory Committee, Washington, DC, 1966 [Электронный ресурс]. URL: https://www.nap.edu/resource/alpac_lm/ARC000005.pdf (дата обращения: 30.04.2020).

new documents as required. The quick retrieval of any parts of any text does not presuppose that the translator would be limited to looking up lexical units or even sentences, and the ability to insert matches into new documents adds a new dimension of usefulness for the translator.

In 1980, Martin Kay called for a complete re-evaluation of the relationship between translators and computers with his proposal of a *translator's amanuensis*. In his view, machines will gradually take over certain functions in the overall translation process and little by little, they will approach translation itself, his main idea being modesty and reliability at each stage. "Little steps for little feet!" he called it.

Kay offers text editing and dictionary look-up as examples of easily mechanizable tasks that are likely to increase a translator's productivity. He then describes an outline of a TM-type system:

> The translator might start by issuing a command causing the system to display anything in the store that might be relevant to [the text to be translated]. This will bring to his attention decisions he made before the actual translation started, statistically significant words and phrases, and a record of anything that had attracted attention when it occurred before. Before going on, he can examine past and future fragments of text that contain similar material[1].

The idea was not quite as trivial as dictionary look-up, and it seemed to fall into the area of functions "approach[ing] translation itself". However, Kay considered this task more mechanizable and more achievable in the shorter term than machine translation proper.

Alan Melby picked up this theme again two years later with his "translator's workstation", functioning on three levels. The first level includes all functions that can be completed in the absence of an electronic source text, including word processing, telecommunications and terminology management. Melby's second level assumes the availability of the source text in electronic format for such functions as text analysis,

[1] Kay, Martin. The Proper Place of Men and Machines in Language Translation [Электронный ресурс]. URL: https://link.springer.com/article/10.1023%2FA%3A1007911416676 (дата обращения: 01.05.2020).

dictionary look-up, and synchronized bilingual text retrieval, while the third level refers to machine translation.

Melby's description of synchronized bilingual text retrieval in his 1992 paper, "The translator workstation", begins to approach quite closely the current incarnations of TM tools:

> When a document had been translated and revised, the final version, as well as its source text, would be stored in such a way that each unit of source text was linked to a corresponding unit of target text. The units would generally be sentences, except in cases where one sentence in the source text becomes two in the target text or vice versa. The benefits of synchronized bilingual text retrieval are manifold with appropriate software. A translator beginning a revision of a document could automatically incorporate unmodified units taken from a previous translation into the revision with a minimum of effort[1].

Melby's use of the words "synchronized" and "linked" is important. They refer to the concept of alignment, an important element in the design of effective TM tools. It was the appearance of TM tools like ALPS (1981) and ETOC (1988) during the 1980s that pushed translation memory beyond the realm of mere academic speculation, but it was during the 1990s that TM developers were finally able to incorporate advances in corpus alignment research. Hutchins sees this as an essential step toward the viability of a TM as a useful tool for translators.

The late 1980s and early 1990s saw the development of another interpretation of "synchronized bilingual text retrieval", namely bilingual concordancing. A bilingual concordancing tool is used to search for patterns in a bitext, also called a parallel corpus, which is made up of aligned source and target texts. It retrieved the requested patterns in their immediate contexts along with their corresponding translations. One such tool is RALI's TransSearch. Users of TransSearch must define and enter search patterns themselves, but attempts have since been made to automate the look-up process.

[1] Melby, A. [K.] The Translator Workstation // Computers in Translation : A Practical Appraisal. London : Routledge, 1992. P. 147 – 165.

During the early- to mid-1990s, translation memory made the next leap, from research to commercial availability. TRADOS, a German translation company, released the terminology management system MultiTerm in 1990, later following it with its TM tool Translator's WorkBench. Atril released its in-house TM tool, Déjà Vu, in 1993 (*www.atril.com*). Transit, by STAR, and Translation Manager/2, by IBM, were also released around the same time. Translation companies and freelance translators are now faced with the question of whether to embrace the new technology, which still requires a significant investment of resources, and if yes, which among the growing selection of competing brands to choose.

(from "Metrics for Evaluating Translation Memory Software"
by Francie Gow)

In 1990, TRADOS launched their first product, MultiTerm (a terminology database, now known as SDL MultiTerm) into the market place. The first version of Translator's Workbench was later released in 1992. Trados also started expanding as a company in the mid-nineties. Matthias Heyn, a computational linguist from the University of Stuttgart joined the company and developed the first alignment tool on the market (T Align, later to become WinAlign, one of the applications available in SDL Trados 2007 Suite). In addition, TRADOS began to open a network of global offices, including Brussels, Virginia, UK and Switzerland.

The nineties saw a marked increase of development in translation software technology. Many freelance translators were benefiting from the increasing sophistication and affordability of personal computers, meaning that CAT tools were becoming more and more commonplace. As well as the time-saving and quality benefits of using translation memory tools at a desktop level, the Internet paved the way for enhanced productivity through the real-time sharing of translation assets via server technology. This helped to rapidly accelerate the rate at which content could be localized, enabling organizations to enter new marketplaces and communicate their messaging in the language of their customers.

The acquisition of TRADOS by SDL in 2005, enabled the two market leaders to leverage their respective product and technical knowledge to expand functionality and features for their customers. The product release of SDL Trados 2007 Suite, combined robust technology, with innovative new features, including automated translation (beta), to help further increase speed and productivity within the translation process.

2009 saw the launch of the Studio family of SDL Trados products, the next generation in translation memory software which revolutionized the way localization professionals worked. SDL Trados Studio 2009, not only merged the best of TRADOS and SDLX, but this release benefitted from the culmination of 25 years translation software expertise and over $100 million investment in R&D.

Following on from the release of SDL Trados Studio 2009, the SDL OpenExchange was launched to industry acclaim as the first application store for the translation industry. SDL OpenExchange is a unique, open industry platform, which enables 3rd party developers and translators to build and market apps and plug-ins, the SDL OpenExchange now has more than 40 apps available to translators. During 2011 the journey of CAT tool evolution continued and SDL Trados Studio 2011 was launched to the market. Which allowed more opportunities for localization professionals.

(from https://www.sdltrados.com/about/history.html)

Further development of TMs involved the development of features and functions common for most of contemporary software. The innovations brought:

- about going online with cloud-based technologies;
- increased personalization and customizing options along with a more user-friendly UI;
- MT engine built in (often a self-learning one);
- autosuggest feature;
- project related functions (allowing project management, collaborations, payment, etc);
- increased variety of formats.

Most common TMs

When speaking about different TM software we should mention two types of programmes – desktop and cloud-based. Which growing internet availability cloud-based software has been gaining popularity. They are easy-to-use, do not take up your computer's storage space, and allow shared access to projects as well as on-line management. Besides you can work with you project from any location and any device that has access to the internet. The weak point in this case is security and confidentiality. Companies offering cloud based solutions have been working really hard to solve the problems and claim to have protected the data from unauthorized access. Desk top versions are installed on your computer and do not need internet. It is safer and provides due confidentiality however most of the advantages of cloud-based solutions are not available here.

Smartcat

https://ru.smartcat.ai/

Smartcat is a cloud based TM tool which was originally developed as a CAT tool in 2012–2015 as an in-house solution by ABBYY Language Solutions, a linguistic service provider within the ABBYY group of companies. The impetus for its development was that ABBYY LS had "felt constrained by translation technologies that had existed for the last 15 years" and wanted a solution that would let them "manage projects with dozens of collaborators, including project managers, translators, editors, and so on" while being "intuitive, cloud-based, scalable, and powerful."

(from https://en.wikipedia.org/wiki/Smartcat)

Features

- unlimited number of translators and proof readers can work on the same project
- project management functions, automated payment, opportunities of finding a projects for free-lancers and finding translators for a project
- supports MS Office formats
- supports files in PDF, JPG, JPEG, TIF, TIFF, BMP, PNG, GIF, CX, PCX, JP2, JPC, JFIF, JB2, DJVU и DJV after they are processed by OCR – optical character recognition – for additional fee

- previous translation memory files can be uploaded to and downloaded from Smartcat

Linguistic functions

- dictionary (by default - ABBYY Lingvo)
- glossaries (can be created, uploaded used repeatedly in new projects by adding them to the project resources)
- two free machine translation options: Microsoft Translator and Яндекс.Translator (GoogleTransalte and Lilt. require a fee)
- TM import and export (.tmx and .sdl iles)

MemoQ

MemoQ is a proprietary computer-assisted translation software suite which runs on Microsoft Windows operating systems. It is developed by the Hungarian software company memoQ Fordítástechnológiai Zrt. (memoQ Translation Technologies), formerly Kilgray, a provider of translation management software established in 2004 and cited as one of the fastest growing companies in the translation technology sector in 2012 and 2013. MemoQ provides translation memory, terminology, machine translation integration and reference information management in desktop, client/server and web application environments.

As of 2018, all supported memoQ editions contained these principal modules.

File statistics. Word counts and comparisons with translation memory databases, internal content similarities and format tag frequency. MemoQ was the first translation environment tool to enable the weighting of format tags in its count statistics to enable the effort involved with their correct placement in translated documents to be considered in planning. Another innovation introduced for file statistics was the analysis of file homogeneity for identifying internal similarities in a file or a group of files which might affect work efforts. Previously such similarities had only been identified in the form of exact text segment repetitions or in comparisons with translation unit databases (translation memories) from previous work.

File translation and editing grid. A columnar grid arrangement of the source and target languages for translating text, supported by other information panes such as a preview, difference highlighting with similar information in reference sources and matches with various information sources such as translation memories, stored reference files, terminology databases, machine translation suggestions and external sources.

Translation memory management. Creation and basic management of databases for multilingual (in the case of memoQ, bilingual) translation information in units known as "segments". This information is often exchanged between translation management and assistance systems using the file format TMX. MemoQ is also able to import translation memory data in delimited text format.

Terminology management. Storage and management of terminology and meta information about the terminology to assist in translation or quality assurance. MemoQ is able to import terminology data in TMX and delimited text formats and export it in delimited text and an XML format. MemoQ also includes an integrated facility for statistical terminology extraction from a chosen combination of documents to translate, translation memory databases and reference corpora. The stopword implementation in the terminology extraction module includes special position indicators to enable blocked terms to be included at the beginning, in the body or at the end of multi-word phrases, which distinguishes this terminology extraction approach from most others available in this type of application.

Reference corpus management. Also known by the trademarked name "LiveDocs", this is a diverse collection of information types, including aligned translations, bitext files from various sources, monolingual reference information in many formats and various types of media files as well as any other file types users choose to save for reference purposes. File types not known by the memoQ application are opened using external applications intended to use them. A distinguishing characteristic of bilingual text alignments in memoQ is automated alignment which need not be finalized and transferred to translation memory databases before it can be used as a basis for comparison with new texts to translate, and alignments can be improved as needed in the course of translation work. In practice this often results in much less effort to maintain legacy reference materials.

Quality assurance. This is for verifying the adherence to quality criteria specified by the user. Profiles can be created to focus on specific workflow tasks, such as format tag verification or adherence to specified terminology.

There are also other supporting features integrated in the environment such as spelling dictionaries, lists of nontranslatable terms, autocorrection rules and "auto-translation" rules which enable matching and insertion of expressions based on regular expressions.

Supported source document formats. MemoQ 2015 supports dozens of different file types, including: various markup and tagged formats such as XML, HTML, XLIFF, SDLXLIFF (SDL Trados Studio's native format for translation), OpenDocument files; plain text files; Microsoft Word, Excel, and PowerPoint; and some Adobe file formats, such as PSD, PDF and InDesign. To know more about supported formats and languages in memoQ, see this link: Languages and file formats.

Handling of translation memories and glossaries. The translation memory (TM) format of memoQ is proprietary and stored as a group of files in a folder bearing the name of the translation memory. External data can be imported in delimited text formats and Translation Memory eXchange format (TMX), and translation memory data can be exported as TMX. MemoQ can also work with server-based translation memories on the memoQ Server or, using a plug-in, other external translation memory sources. MemoQ Translation memories are bilingual.

In translation work, translation segments are compared with translation units stored in the translation memory, and exact or fuzzy matches can be shown and inserted in the translated text.

Glossaries are handled by the integrated terminology module. Glossaries can be imported in TMX or delimited text formats and exported as delimited text or MultiTerm XML. Glossaries can include two or more languages or language variants. Term matching with glossary entries can be based on many different parameters, taking into consideration capitalization, partial or fuzzy matches and other factors. Terms to be avoided can be marked as "forbidden" in the properties of a particular glossary entry.

Integration of machine translation and postediting. MemoQ has integrated machine translation and postediting into its translation workflow. With the selection of appropriate conditions and a plug-in for machine translation, machine-generated translation units (TUs) will be inserted if no match is found in an active translation memory.

The translator can then post-edit the machine translation in the attempt to make sense of it. MemoQ currently includes plug-ins which support the following MT systems: Omniscien Technologies (formerly Asia Online), Globalese, iTranslate4.eu, KantanMT, Let's MT!, Systran MT, Google Translate, Microsoft Translator and a pseudotranslation engine. Other MT systems can be integrated via the application programming interface (API).

(from https://en.wikipedia.org/wiki/MemoQ)

SELF-CHECK TASKS

***1. Give English equivalents for the following words and word combinations*:** ранее переведенные сегменты, язык перевода, заголовок, список, содержимое ячейки таблицы, исходный сегмент, нечеткие соответствия, полные совпадения, повторы, значение по умолчанию, оплата по количеству слов, повторяющееся содержание, словоупотребление, управление терминологией, вбивать вручную на клавиатуре, извлекать, выравнивание, совместное использование переводческих ресурсов, запуск, облачное решение, единство терминологии.

2. Answer the questions.

1. What is translation memory?

2. What is a unit of TM? What can it be equal to?

3. What do you remember about the evolution of the technology? Whose ideas laid the grounds for modern software?

4. What tools are integrated in a modern TM?

5. What are the benefits and downsides of using TM in translation?

PRACTICE TASKS

Describe any TM. Characterize the UI, customizing options, functions, etc.

TUTORIAL

In this tutorial you will get to know Omega T by translating a children's poem "This is the house that Jack built..." by Mother Goose. Omega T is the simplest software that is purely based on translation memory technology without additional features. But once you have mastered Omega T you are sure to quickly deal with any other modern software.

1. Before doing the task watch a tutorial following the link *https://www.youtube.com/watch?v=3Wv79R9Sp6E*

2. Download Omega T (https://omegat.org/download) and create a project. The text for the project will be a children's poem "This is the house that Jack built..." by Mother Goose.

3. Before translating open all tabs and settings and describe the interface the program. When preparing a report, attach screenshots that show different elements of it.

4. Upload the text in the newly created project and start translating it segment after segment.

5. In the course of translation you will come across new words. Make up a glossary inside the project. Make sure the glossary works well and in the next segment the words from the glossary are highlighted and in the glossary window you can see the Russian equivalent.

6. As you translate, you will see, that with every new segment, the number of variants in the Fuzzy Match (нечеткие совпадения) window increases. You can choose the match which suits your translation needs best. Provide a screenshot showig how the fuzzy matching box and the glossary box work together.

7. As soon as you have completed translation you should go to the tab *Проект* and choose *Создать переведенные документы*.

8. Go to the project folder and open the subfolder Target and file the resulting translation. Attach the file together with the report.

9. Go to the project folder and open the subfolder Glossary and file the resulting translation. Attach the file together with the report.

10. Prepare a report based on the laboratory work. Files to be submitted: 1) Report; 2) Target text; 3) Glossary.

For formatting rules see Appendix 1, for report structure refer to Appendix 2, a sample tutorial report can be found in Appendix 3.

Unit 6

MACHINE TRANSLATION

Pre-reading tasks

Make sure you know Russian equivalents of the following words and word combinations.

Computer-aided translation, customization, output quality, generative linguistics and transformational grammar, semantic ambiguity, interlingua, PEMT, MAHT, source language, target language.

Introduction

Machine translation, sometimes referred to by the abbreviation MT (not to be confused with computer-aided translation, machine-aided human translation (MAHT) or interactive translation) is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another.

On a basic level, MT performs simple substitution of words in one language for words in another, but that alone usually cannot produce a good translation of a text because recognition of whole phrases and their closest counterparts in the target language is needed. Solving this problem with corpus statistical, and neural techniques is a rapidly growing field that is leading to better translations, handling differences in linguistic typology, translation of idioms, and the isolation of anomalies.

Current machine translation software often allows for customization by domain or profession (such as weather reports), improving output by limiting the scope of allowable substitutions. This technique is particularly effective in domains where formal or formulaic language is used. It follows that machine translation of government and legal documents more readily produces usable output than conversation or less standardised text.

Improved output quality can also be achieved by human intervention: for example, some systems are able to translate more accurately if the user has unambiguously identified which words in the text are proper names. With the assistance of these techniques, MT has proven useful as a tool to assist human translators and, in a very limited number of cases, can even produce output that can be used as is (e.g., weather reports).

Historical Background

In the mid-1930s the first patents for "translating machines" were applied for by Georges Artsrouni, for an automatic bilingual dictionary using paper tape. Russian Peter Troyanskii submitted a more detailed proposal that included both the bilingual dictionary and a method for dealing with grammatical roles between languages, based on the grammatical system of Esperanto. This system was separated into three stages: stage one consisted of a native-speaking editor in the source language to organize the words into their logical forms and to exercise the syntactic functions; stage two required the machine to "translate" these forms into the target language; and stage three required a native-speaking editor in the target language to normalize this output. Troyanskii's proposal remained unknown until the late 1950s, by which time computers were well-known and utilized.

The first set of proposals for computer based machine translation was presented in 1949 by Warren Weaver, a researcher at the Rockefeller Foundation, "Translation memorandum". These proposals were based on information theory, successes in code breaking during the Second World War, and theories about the universal principles underlying natural language.

A few years after Weaver submitted his proposals, research began in earnest at many universities in the United States. On 7 January 1954 the

Georgetown-IBM experiment was held in New York at the head office of IBM. This was the first public demonstration of a machine translation system. The demonstration was widely reported in the newspapers and garnered public interest. The system itself, however, was no more than a “toy” system. It had only 250 words and translated 49 carefully selected Russian sentences into English – mainly in the field of chemistry. Nevertheless, it encouraged the idea that machine translation was imminent and stimulated the financing of the research, not only in the US but worldwide.

Early systems used large bilingual dictionaries and hand-coded rules for fixing the word order in the final output which was eventually considered too restrictive in linguistic developments at the time. For example, generative linguistics and transformational grammar were exploited to improve the quality of translations. During this period operational systems were installed. The United States Air Force used a system produced by IBM and Washington University, while the Atomic Energy Commission and Euratom, in Italy, used a system developed at Georgetown University. While the quality of the output was poor it met many of the customers' needs, particularly in terms of speed.

At the end of the 1950s, Yehoshua Bar-Hillel was asked by the US government to look into machine translation, to assess the possibility of fully automatic high quality translation by machines. Bar-Hillel described the problem of semantic ambiguity or double-meaning, as illustrated in the following sentence:

Little John was looking for his toy box. Finally he found it. The box was in the pen.

The word “pen” may have two meanings: the first meaning is “something used to write in ink with”; the second meaning is “a container of some kind”. To a human, the meaning is obvious, but Bar-Hillel claimed that without a “universal encyclopedia” a machine would never be able to deal with this problem. At the time, this type of semantic ambiguity could only be solved by writing source texts for machine translation in a controlled language that uses a vocabulary in which each word has exactly one meaning.

Research in the 1960s in both the Soviet Union and the United States concentrated mainly on the Russian-English language pair. The objects of translation were chiefly scientific and technical documents, such as articles from scientific journals. The rough translations produced were sufficient to get a basic understanding of the articles. If an article discussed a subject deemed to be confidential, it was sent to a human translator for a complete translation; if not, it was discarded.

A great blow came to machine-translation research in 1966 with the publication of the ALPAC report. The report was commissioned by the US government and delivered by ALPAC, the Automatic Language Processing Advisory Committee, a group of seven scientists convened by the US government in 1964. The US government was concerned that there was a lack of progress being made despite significant expenditure. The report concluded that machine translation was more expensive, less accurate and slower than human translation, and that despite the expenditures, machine translation was not likely to reach the quality of a human translator in the near future. The report recommended, however, that tools be developed to aid translators – automatic dictionaries, for example – and that some research in computational linguistics should continue to be supported.

The publication of the report had a profound impact on research into machine translation in the United States, and to a lesser extent the Soviet Union and United Kingdom. Research, at least in the US, was almost completely abandoned for over a decade. In Canada, France and Germany, however, research continued. In the US the main exceptions were the founders of Systran (Peter Toma) and Logos (Bernard Scott), who established their companies in 1968 and 1970 respectively and served the US Department of Defense. In 1970, the Systran system was installed for the United States Air Force, and subsequently by the Commission of the European Communities in 1976. The METEO System, developed at the Université de Montréal, was installed in Canada in 1977 to translate weather forecasts from English to French, and was translating close to 80,000 words per day or 30 million words per year until it was replaced by a competitor's system on 30 September 2001.

While research in the 1960s concentrated on limited language pairs and input, demand in the 1970s was for low-cost systems that could translate a range of technical and commercial documents. This demand was spurred by the increase of globalisation and the demand for translation in Canada, Europe, and Japan.

By the 1980s, both the diversity and the number of installed systems for machine translation had increased. A number of systems relying on mainframe technology were in use, such as Systran, Logos, Ariane-G5, and Metal.

As a result of the improved availability of microcomputers, there was a market for lower-end machine translation systems. Many companies took advantage of this in Europe, Japan, and the USA. Systems were also brought onto the market in China, Eastern Europe, Korea, and the Soviet Union.

During the 1980s there was a lot of activity in MT in Japan especially. With the fifth generation computer Japan intended to leap over its competition in computer hardware and software, and one project that many large Japanese electronics firms found themselves involved in was creating software for translating into and from English (Fujitsu, Toshiba, NTT, Brother, Catena, Matsushita, Mitsubishi, Sharp, Sanyo, Hitachi, NEC, Panasonic, Kodensha, Nova, Oki).

Research during the 1980s typically relied on translation through some variety of intermediary linguistic representation involving morphological, syntactic, and semantic analysis.

At the end of the 1980s, there was a large surge in a number of novel methods for machine translation. One system was developed at IBM that was based on statistical methods. Makoto Nagao and his group used methods based on large numbers of translation examples, a technique that is now termed example-based machine translation. A defining feature of both of these approaches was the neglect of syntactic and semantic rules and reliance instead on the manipulation of large text corpora.

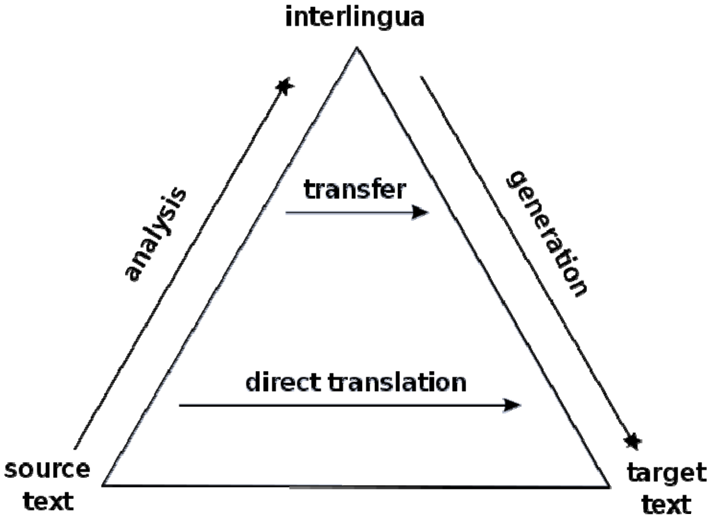
During the 1990s, encouraged by successes in speech recognition and speech synthesis, research began into speech translation with the development of the German Verbmobil project. The Forward Area Language Converter (FALCon) system, a machine translation technology designed by the Army Research Laboratory, was fielded 1997 to translate documents for soldiers in Bosnia.

There was significant growth in the use of machine translation as a result of the advent of low-cost and more powerful computers. It was in the early 1990s that machine translation began to make the transition away from large mainframe computers toward personal computers and workstations. Two companies that led the PC market for a time were Globalink and MicroTac, following which a merger of the two companies (in December 1994) was found to be in the corporate interest of both. Intergraph and Systran also began to offer PC versions around this time. Sites also became available on the internet, such as AltaVista's Babel Fish (using Systran technology) and Google Language Tools (also initially using Systran technology exclusively).

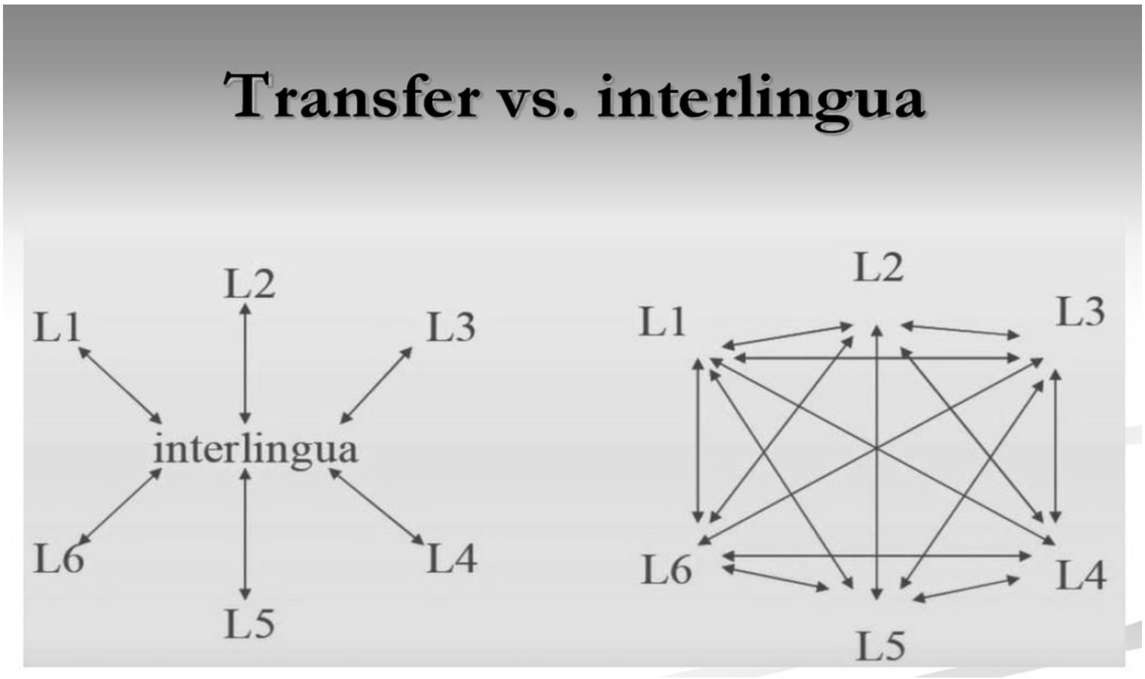
The field of machine translation has seen major changes in the last few years. Currently a large amount of research is being done into statistical machine translation and example-based machine translation. In the area of speech translation, research has focused on moving from domain-limited systems to domain-unlimited translation systems.

Approaches

Machine translation can use a method based on linguistic rules, which means that words will be translated in a linguistic way – the most suitable (orally speaking) words of the target language will replace the ones in the source language. It is often argued that the success of machine translation requires the problem of natural language understanding to be solved first.



Generally, rule-based methods parse a text, usually creating an intermediary, symbolic representation, from which the text in the target language is generated. According to the nature of the intermediary representation, an approach is described as interlingual machine translation or transfer-based machine translation. These methods require extensive lexicons with morphological, syntactic, and semantic information, and large sets of rules.



Transfer vs. interlingua

Given enough data, machine translation programs often work well enough for a native speaker of one language to get the approximate meaning of what is written by the other native speaker. The difficulty is getting enough data of the right kind to support the particular method. For example, the large multilingual corpus of data needed for statistical methods to work is not necessary for the grammar-based methods. But then, the grammar methods need a skilled linguist to carefully design the grammar that they use.

To translate between closely related languages, the technique referred to as rule-based machine translation may be used.

Rule-based machine translation. The rule-based machine translation paradigm includes transfer-based machine translation, interlingual machine translation and dictionary-based machine translation paradigms. This type of translation is used mostly in the creation of dictionaries and grammar programs. Unlike other methods, RBMT involves more information about the linguistics of the source and target languages, using the morphological and syntactic rules and semantic analysis of both languages. The basic approach involves linking the structure of the input sentence with the structure of the output sentence using a parser and an analyzer for the source language, a generator for the target language, and a transfer lexicon for the actual translation. RBMT's biggest downfall is that everything must be made explicit: orthographical variation and erroneous input must be made part of the source language analyser in order to cope with it, and lexical selection rules must be written for all instances of ambiguity. Adapting to new domains in itself is not that hard, as the core grammar is the same across domains, and the domain-specific adjustment is limited to lexical selection adjustment.

Transfer-based machine translation. Transfer-based machine translation is similar to interlingual machine translation in that it creates a translation from an intermediate representation that simulates the meaning of the original sentence. Unlike interlingual MT, it depends partially on the language pair involved in the translation.

Interlingual machine translation. Interlingual machine translation is one instance of rule-based machine-translation approaches. In this approach, the source language, i.e. the text to be translated, is transformed into an interlingual language, i.e. a "language neutral" representation that is independent of any language. The target language is then generated out of the interlingua. One of the major advantages of this system is that the interlingua becomes more valuable as the number of target languages it can be turned into increases. However, the only interlingual machine translation system that has been made operational at the commercial level is the KANT system, which is designed to translate Caterpillar Technical English (CTE) into other languages.

Dictionary-based machine translation. Machine translation can use a method based on dictionary entries, which means that the words will be translated as they are by a dictionary.

Statistical machine translation. Statistical machine translation tries to generate translations using statistical methods based on bilingual text corpora, such as the Canadian Hansard corpus, the English-French record of the Canadian parliament and EUROPARL, the record of the European Parliament. Where such corpora are available, good results can be achieved translating similar texts, but such corpora are still rare for many language pairs. The first statistical machine translation software was CANDIDE from IBM. Google used SYSTRAN for several years, but switched to a statistical translation method in October 2007. In 2005, Google improved its internal translation capabilities by using approximately 200 billion words from United Nations materials to train their system; translation accuracy improved. Google Translate and similar statistical translation programs work by detecting patterns in hundreds of millions of documents that have previously been translated by humans and making intelligent guesses based on the findings. Generally, the more human-translated documents available in a given language, the more likely it is that the translation will be of good quality. Newer approaches into Statistical Machine translation such as METIS II and PRESEMT use minimal corpus size and instead focus on derivation of syntactic structure through pattern recognition. With further development, this may allow statistical machine translation to operate off of a monolingual text corpus. SMT's biggest downfall includes it being dependent upon huge amounts of parallel texts, its problems with morphology-rich languages (especially with translating into such languages), and its inability to correct singleton errors.

Example-based machine translation. Example-based machine translation (EBMT) approach was proposed by Makoto Nagao in 1984. Example-based machine translation is based on the idea of analogy. In this approach, the corpus that is used is one that contains texts that have already been translated. Given a sentence that is to be translated, sentences from this corpus are selected that contain similar sub-sentential components. The similar sentences are then used to translate the sub-sentential components of the original sentence into the target language, and these phrases are put together to form a complete translation.

Hybrid machine translation. Hybrid machine translation (HMT) leverages the strengths of statistical and rule-based translation methodologies. Several MT organizations claim a hybrid approach that uses both rules and statistics. The approaches differ in a number of ways.

Rules post-processed by statistics. Translations are performed using a rules based engine. Statistics are then used in an attempt to adjust/correct the output from the rules engine.

Statistics guided by rules. Rules are used to pre-process data in an attempt to better guide the statistical engine. Rules are also used to post-process the statistical output to perform functions such as normalization. This approach has a lot more power, flexibility and control when translating. It also provides extensive control over the way in which the content is processed during both pre-translation (e.g. markup of content and non-translatable terms) and post-translation (e.g. post translation corrections and adjustments).

More recently, with the advent of Neural MT, a new version of hybrid machine translation is emerging that combines the benefits of rules, statistical and neural machine translation. The approach allows benefitting from pre- and post-processing in a rule guided workflow as well as benefitting from NMT and SMT. The downside is the inherent complexity which makes the approach suitable only for specific use cases. One of the proponents of this approach for complex use cases is Omniscien Technologies.

Neural machine translation. A deep learning based approach to MT, neural machine translation has made rapid progress in recent years, and Google has announced its translation services are now using this technology in preference to its previous statistical methods.

(from https://en.wikipedia.org/wiki/Machine_translation)

Other recent developments include generic, customizable and adaptive MT.

Generic MT usually refers to platforms such as Google Translate, Bing, Yandex, and Naver. These platforms provide MT for ad hoc translations to millions of people. Companies can buy generic MT for batch pre-translation and connect to their own systems via API.

Customizable MT refers to MT software that has a basic component and can be trained to improve terminology accuracy in a chosen domain (medical, legal, IP, or a company's own preferred terminology). For example, WIPO's specialist MT engine translates patents more accurately than generalist MT engines, and eBay's solution can understand and render into other languages hundreds of abbreviations used in electronic commerce.

Adaptive MT offers suggestions to translators as they type in their CAT-tool, and learns from their input continuously in real time. Introduced by Lilt in 2016 and by SDL in 2017, adaptive MT is believed to improve translator productivity significantly and can challenge translation memory technology in the future.

Ethics for Translation Providers using MT

Confidentiality. Content translated by free MT platforms such as Google Translate and Microsoft Translator is not confidential. It is stored by the platform owners and may be reused for later translations.

Notifying the Client about MT Use. It's a point of debate in the industry if a translation company should notify clients about use of MT on their projects. Many pundits are in favor of informing the customer of MT usage and others may not disclose the use of MT. Be sure to ask your provider if you have questions about MT usage.

(from https://www.gala-global.org/what-machine-translation)

SELF-CHECK TASKS

***1. Give English equivalents for the following words and word combinations*:** нейротехнологии, формализованный язык, точность, экономичная система машинного перевода, распознавание речи, метод перевода на основе правил, метод перевода на основе переноса, исходное предложение, переведенное предложение, неоднозначность, машинный перевод, основанный на примерах, постредактирование.

2. Answer the questions.

1. What is machine translation?
2. What types of texts are suitable for MT? Why?
3. What are the strong and weak points of MT as compared to TM?
4. How did the technology evolve?
5. What types of MT do you know? Which of them are newer ones?
6. What is the ethics of using MT in professional settings?

PRACTICE TASKS

1. Find examples of software for different types of machine translation.

2. Compare two MT systems. Speak about their strong and weak points. What types of texts are suitable for each?

TUTORIAL

The topic of this laboratory work is machine translation and integrated cloud-based CAT tools.

The aim is to learn post-editing of MT and get familia with SmartCat, an integrated cloud-based CAT tool.

The work has 2 tasks:

1) to identify the weak points of statistical machine translation;

2) to practice post-editing.

In order to do it follow the steps below.

1. Register an account with Smartcat at https://ru.smartcat.ai/
2. Create a new project, enable machine translation option.
3. Download the glossary you did in LW 3 for soil-study texts.
4. Choose one of the texts attached to LW 3 and upload it into the project.
5. Do machine translation and download the resulting text without any editing (it will be Appendix 1 of your laboratory work report), provide screenshots when you describe the procedure.
6. Do the post-editing by gooing from segment to segment, use your glossary, provide screenshots.
7. Download and check the resulting tranlation (prodive the text with your corrections in Appendix 2).
8. Make conclusion about the most common mistakes of MT engine, comment on your experience with SmartCat.

For formatting rules see Appendix 1, for report structure refer to Appendix 2, a sample tutorial report can be found in Appendix 3.

AFTERWORD

You have reached the final destination in your trip to the world of information technology in translation. Now look around! You can see a whole lot of new things to discover, new tools to master you translation skills and new horizons for your professional development. Never stop learning and practicing, stay hungry… hungry for new knowledge and opportunities.

Keep calm and carry on!

REFERENCES

1. http://www.just-the-word.com
2. http://www.natcorp.ox.ac.uk
3. http://www.stanford.edu
4. https://dictionary.cambridge.org/ru
5. https://en.wikipedia.org
6. https://wooordhunt.ru/
7. https://www.anglistik.uni-freiburg.de/
8. https://www.collinsdictionary.com/dictionary/english
9. https://www.english-corpora.org/bnc/
10. https://www.gala-global.org
11. https://www.laurenceanthony.net
12. https://www.lexico.com/en
13. https://www.merriam-webster.com/
14. https://www.oxfordlearnersdictionaries.com
15. https://www.sdltrados.com
16. https://www.sketchengine.euhttps://www.smartcat.ai
17. https://www.techrepublic.com
18. https://www1.essex.ac.uk
19. www.atril.com
20. www.multitran.com

APPENDICES

Appendix 1

Tutorials formatting guidelines

1. Please submit your manuscript as an **.RTF** or a **.DOC** file.
2. **Page size**: A4 (210×297 mm). Portrait layout.
3. **File name**: your last name – underscore – tutorial number. For example, Smith_4.
4. **Margins**: 3 cm at left, 2.5 cm at right, 2.5 at top and 3.0 at bottom.
5. **Font**: Times New Roman; **font size 12**, 1.5 line spaced.
6. **Paragraphs**: **1.25 indented,** without extra spaces between them.
7. **Text justification**: evenly between the margins; no hyphenation.
8. **All illustrations** (charts, drawings, diagrams, pictures, etc.) are to be captioned as Picture or Table and numbered. All illustrations should be placed directly in the manuscript.
9. **All lists automatically numbered.**
10. **All examples** are to be **italicized**. For more emphasis use **boldface**.

Appendix 2

Tutorial Report Structure

1. Front sheet.

2. Topic.

3. Aim and Task.

4. Software used.

5. Theoretical background (specify terms, definitions, software features, etc.) – at least half a page.

6. Reference sources (all web-sites, books, articles used – numbered in alphabetical order).

7. Procedure description (a stepwise report about the work illustrated by screenshots cut appropriately to show the feature or idea described). All problems and solutions are to be reported here as well.

8. Conclusion (a well thought over summary about the experience gained in the tutorial, as well as advice and recommendations for the efficient use of software in question).

9. Appendices (all linguistic materials used for the tutorial, as well as the resulting text; the source and the target texts are separate appendices).

Appendix 3

Sample Tutorial

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Владимирский государственный университет
имени Александра Григорьевича и Николая Григорьевича Столетовых»

Гуманитарный институт
Кафедра «Иностранные языки профессиональной коммуникации»

Отчет
о лабораторной работе
по дисциплине
«Перевод с применением современных технологий»
студента … курса … группы …
ФИО

Направление 45.03.02 Лингвистика

Преподаватель
доцент кафедры ИЯПК О. А. Селиверстова

Владимир, 2020

Laboratory work № 6

Topic: machine translation.

Aim: learn how to use machine translation.

Task:

1) identify weak points of statistical machine translation;
2) practice post-editing of machine-translated text.

Software:

1) search engine Google;
2) text editor Microsoft Office Word 2007;
3) cloud-based CAT-tool Smartcat.

Theoretical background:

Machine translation is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another. MT conducts simple substitution of words in one language for words in another, which is not enough for good translation of a text, because recognition of whole phrases and their closest counterparts in the target language is needed. Nowadays the field solving the stated problem with corpus statistical and neural techniques grows rapidly. It leads to improvements of translation, to handling differences in linguistic typology, translation of idioms, and the isolation of anomalies.

The first patents for translating machine were already applied in mid-1930s, and researches have been conducted since then.

There are different approaches to machine translation, for example, the rule-based one which uses morphological and syntactic rules and semantic analysis of both languages and links the structures of input and output sentences, the statistical one which tries to generate translations using statistical methods based on bilingual text corpora, the example based one which implies the use of already translated texts, and many others.

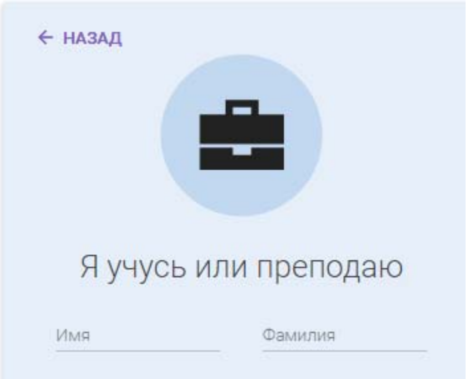
Reference sources:

1) https://ru.smartcat.ai/
2) https://www.multitran.com/
3) https://dic.academic.ru/dic.nsf/bse/68042/%D0%91%D0%B5%D0%B7%D0%BE%D1%82%D0%B2%D0%B0%D0%BB%D1%8C%D0%BD%D0%B0%D1%8F
4) https://www.google.ru

Procedure:

1. First of all, I register an account on the official site of Smartcat. I choose the status of alumnus and fill in the gaps.

2. The second step is to create a glossary. One can upload files with .xlsx format, but our glossary are in .docx format, so I create a new glossary by hand: I type in words and their equivalents.

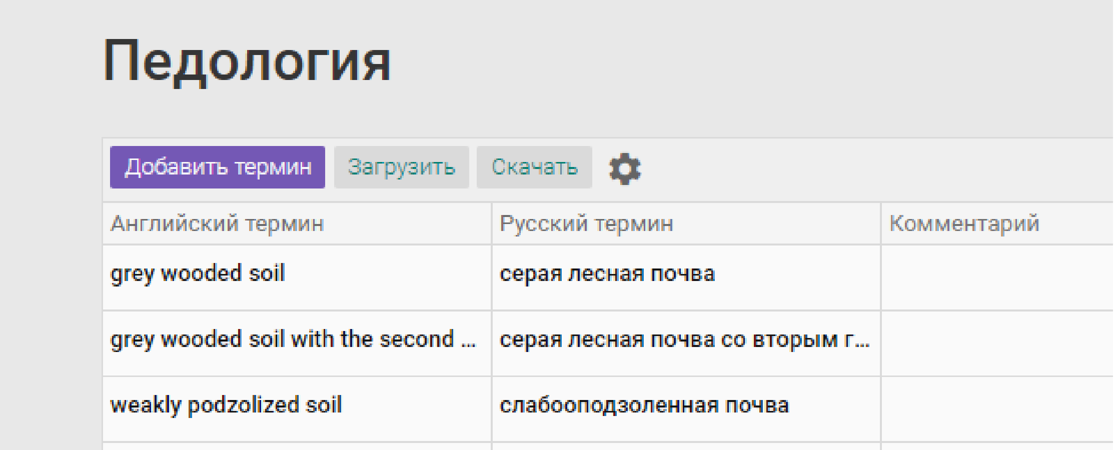


← НАЗАД

Я учусь или преподаю

Имя

Фамилия

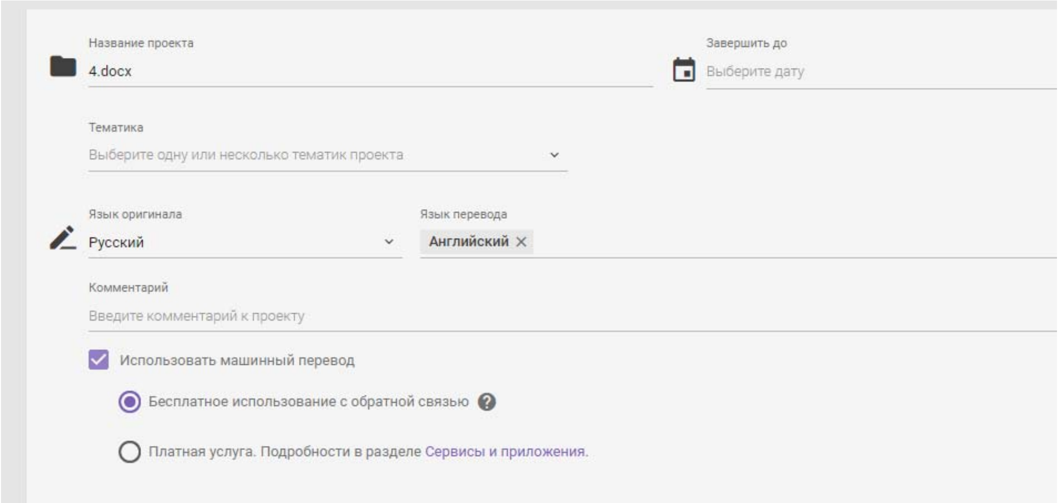


Педология

Английский термин	Русский термин	Комментарий
grey wooded soil	серая лесная почва	
grey wooded soil with the second ...	серая лесная почва со вторым г...	
weakly podzolized soil	слабооподзоленная почва	

3. The third step is to start a new project. I choose the file number 4 from the documents attached to the Laboratory Work 3, upload it

into the new project, choose the source and the target languages, click the button «use machine translation» and begin to work.



Название проекта
4.docx

Завершить до
Выберите дату

Тематика
Выберите одну или несколько тематик проекта

Язык оригинала
Русский

Язык перевода
Английский ×

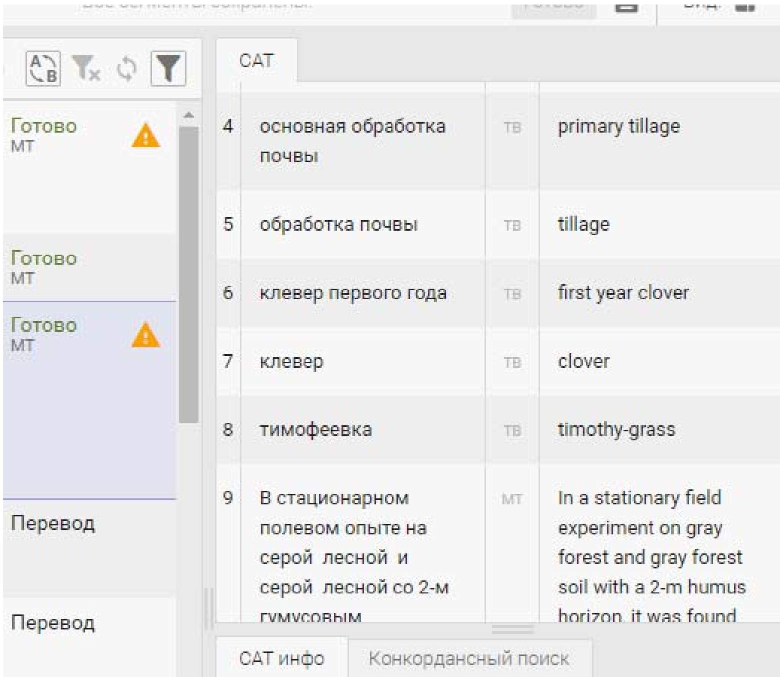
Комментарий
Введите комментарий к проекту

☑ Использовать машинный перевод

● Бесплатное использование с обратной связью

○ Платная услуга. Подробности в разделе Сервисы и приложения.

The translation conducted by machine is shown in the upper right corner after the words contained in the glossary. I work with each segment separately: I click on the empty one and press buttons «Cltr+number» in order to paste machine-translated piece; the number coincides with the one that indicates machine translation (MT) in the list as in a screenshot below.



Готово MT

Готово MT

Готово MT

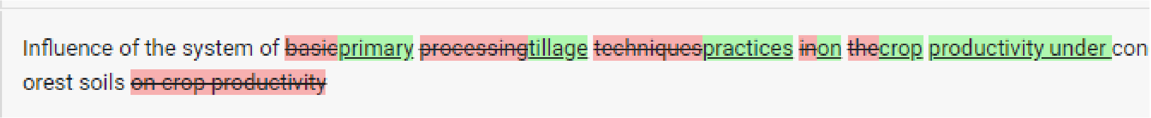
Перевод

Перевод

4	основная обработка почвы	TB	primary tillage
5	обработка почвы	TB	tillage
6	клевер первого года	TB	first year clover
7	клевер	TB	clover
8	тимофеевка	TB	timothy-grass
9	В стационарном полевом опыте на серой лесной и серой лесной со 2-м гумусовым	MT	In a stationary field experiment on gray forest and gray forest soil with a 2-m humus horizon, it was found

After I have filled all the segments with machine-translated text, I download the file and attach the unedited translation to the present Laboratory Work (see Appendix 1).

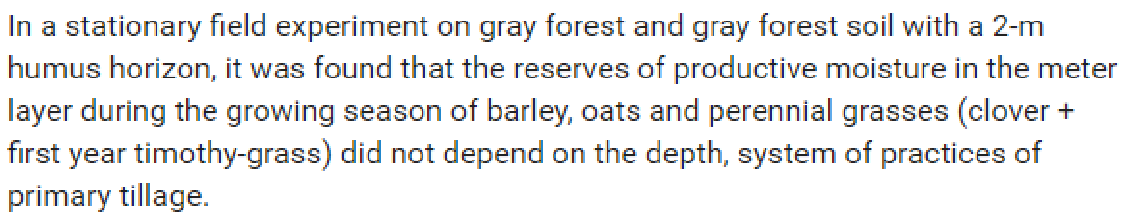
4. The next step to edit the segments of the machine-translated document. The first segment undergoes great changes. First of all, the word-order offered my machine translation makes the whole sentence vague and unclear. I transfer the ending of the sentence closer to the beginning, in order to make the structure «influence of ... on ...» sound more natural. Then I change the phrase «the system of basic processing techniques» for «the system of primary tillage practices», in accordance with the previously created glossary. Besides, I change the preposition «in» for «under» in the expression «in the conditions». So the whole sentence looks as follows: Influence of the system of primary tillage practices on crop productivity under conditions of heterogeneity of grey forest soils.



Influence of the system of ~~basic~~primary ~~processing~~tillage ~~techniques~~practices ~~in~~on ~~the~~crop productivity under con
orest soils ~~on crop productivity~~

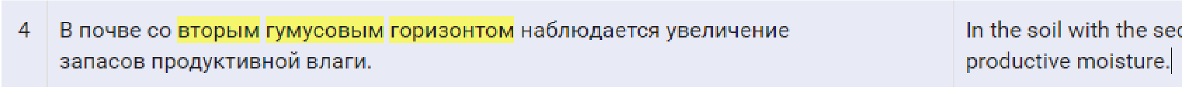
5. The next segment contains the only word «report». I change it for «abstracts».

6. The segment number three is not transformed so radically. Having consulted the online dictionary Multitran, I only change the name of the plant, so «Timofeevka» becomes «timothy-grass». Besides, the glossary created for the previous work on pedology contains an expression «first year clover» (клевер первого года). So I conduct translation using that example and put «first year» before «timothy-grass» instead of «Timofeevka of the first year of use», which is translated word-by-word, without preservation of form utilized by native speakers and without analysis of the meaning which helps to conduct translation. Using search engine Google, I check the correctness of expressions «perennial grasses» and «reserves of productive moisture» and figure out that they are utilized in the given manner. The edited text can be seen in the screenshot below.



In a stationary field experiment on gray forest and gray forest soil with a 2-m humus horizon, it was found that the reserves of productive moisture in the meter layer during the growing season of barley, oats and perennial grasses (clover + first year timothy-grass) did not depend on the depth, system of practices of primary tillage.

7. The fourth segment is remained unchanged due to its correctness. The words in yellow indicate expressions fixed in the glossary.

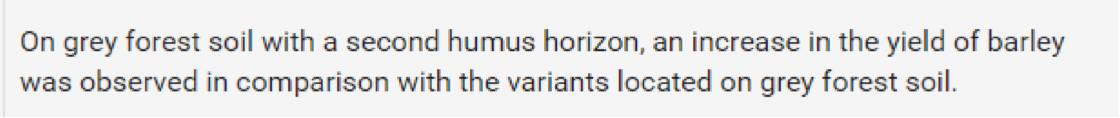


4	В почве со вторым гумусовым горизонтом наблюдается увеличение запасов продуктивной влаги.	In the soil with the sec productive moisture.

8. The segment number five contains a big number of uncertainties. I consult the online dictionary Academic in order to check the meaning of the word-combination «безотвальная обработка почвы», then I look for the presumptive English equivalent of the phrase in the dictionary Multitran. And finally, having utilized Google search engine, I run into the book «Selected water resources. Abstracts» published in Ohio which contains the needed expression. After that I type «no-till treatment» instead of «fallow treatment».

The next expression to find is «ярусная вспашка». In a similar way described above I find out its equivalent: «layer plowing» instead of «long-line plowing».

9. The sixth segment is correct lexically, but contains unnecessary inversion, when an adverbial construction stands before the subject.

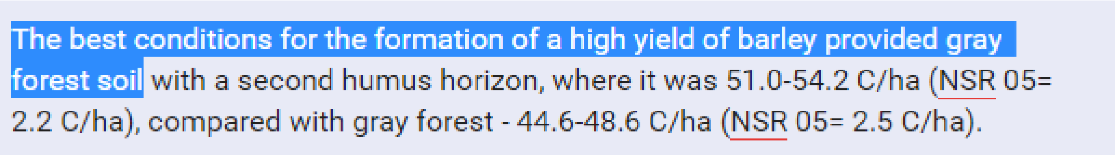


On grey forest soil with a second humus horizon, an increase in the yield of barley was observed in comparison with the variants located on grey forest soil.

So I transfer the subject to the beginning: «An increase in the yield of barley on grey forest soil with a second humus horizon was observed in comparison with the variants located on grey forest soil».

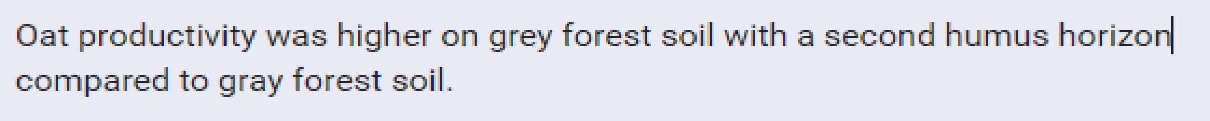
10. The segment number 7 does not undergo extensive changes. But as I have established in the given translation a norm «no-till treatment», I maintain it correcting the word-combination «non-tillage treatment».

11. The eighth segment is correct from the lexical point of view. However, it contains a deviation from the common word-order of the declarative sentence: the object here is placed at the beginning (like in the source text), then goes the predicate followed by the subject.

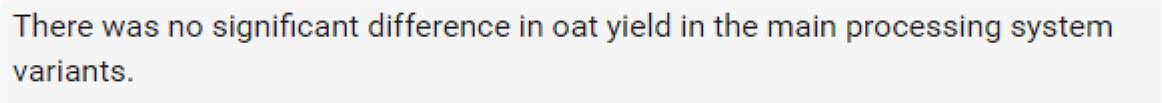

The best conditions for the formation of a high yield of barley provided gray forest soil with a second humus horizon, where it was 51.0-54.2 C/ha (NSR 05= 2.2 C/ha), compared with gray forest - 44.6-48.6 C/ha (NSR 05= 2.5 C/ha).

So I put the subject on the first place, the predicate keeps the second position, and the object takes the third place: «Grey forest soil with a second humus horizon provided the best conditions for the formation of a high yield of barley».

12. The segment number nine contains a mistake described in the paragraph 9: adverb is placed at the beginning of the sentence which is a quite typical feature of Russian syntax. I reconstruct the direct word-order and the sentence takes the following form:

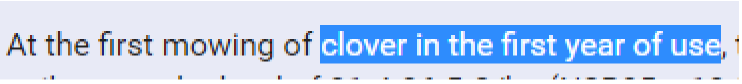

Oat productivity was higher on grey forest soil with a second humus horizon compared to gray forest soil.

13. The tenth segment has completely lost its sense after machine translation. The Russian sentence «По вариантам системы основной обработки существенной разницы в урожае овса не отмечено» cannot be decoded from the English equivalent offered by the software:


There was no significant difference in oat yield in the main processing system variants.

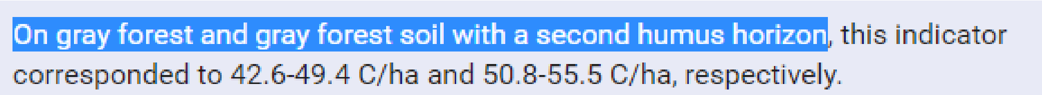
The main idea of the source phrase is not conveyed through the machine-translated piece. It is that despite the distinctions of tillage systems, the yields do not differ. I make certain changes, so the sentence looks in the following way: «There was no significant difference of oat yield despite the differences of the systems of primary tillage».

14. The segment number 11 contains only one expression to correct: I change «clover in the first year of use» for «first year clover» based on the glossary. The machine-translated expression «клевер первого года пользования» is again conducted word-by-word:



At the first mowing of clover in the first year of use,

15. The following segment again contains ungrounded violation of direct word-order.

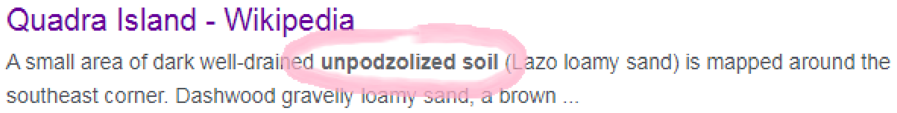


On gray forest and gray forest soil with a second humus horizon, this indicator corresponded to 42.6-49.4 C/ha and 50.8-55.5 C/ha, respectively.

Having worked on the sentence, I manage to put its parts in a well-established manner, so it takes the following form: «This indicator corresponded to 42.6 – 49.4 C/ha and 50.8 – 55.5 C/ha on grey forest and grey forest soil with a second humus horizon, respectively».

16. The thirteenth segment is correct both from lexical and syntactical points of view: the use of vocabulary is unmistakable, the word-order is direct, the meaning of the sentence in Russian is preserved in its translation. So I make no changes.

17. The segment fourteen contains certain expressions requiring attention. The word «neopodzloennaya» I change for «unpodzolized», constructed following the example of a word from the glossary – «podzolized» – and checked using Wikipedia.

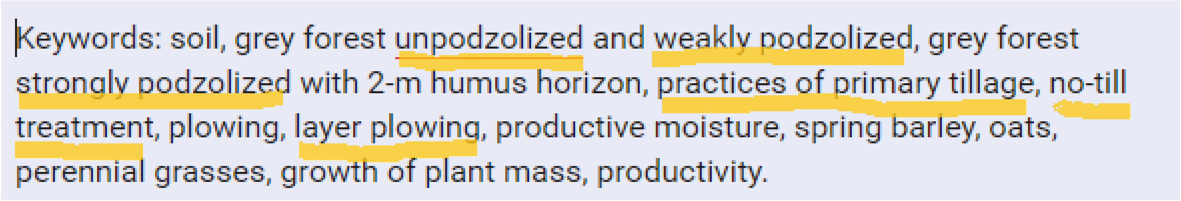


Quadra Island - Wikipedia

A small area of dark well-drained **unpodzolized soil** (Lazo loamy sand) is mapped around the southeast corner. Dashwood gravelly loamy sand, a brown ...

The word combination transliterated into English by the software – «slaboopodzolennaya», I it turn into «weakly podzolized», according to the glossary. The rest of the changes can be seen in pairs below: «methods of basic processing» – «practices of primary tillage», «non-fallow processing» – «no-till treatment», «low-line plowing» – «layer plowing».

The resulting sentence looks the following way:



Keywords: soil, grey forest unpodzolized and weakly podzolized, grey forest strongly podzolized with 2-m humus horizon, practices of primary tillage, no-till treatment, plowing, layer plowing, productive moisture, spring barley, oats, perennial grasses, growth of plant mass, productivity.

18. The very last step is to download the translation, format it and paste into the Laboratory Work.

Conclusion:

The aim of the present laboratory work was to learn about machine translation, its advantages and disadvantages. During the accomplishment of the tasks one has mastered the skills of creating translation projects using the cloud-based computer assisted tool Smartcat and has learnt a convenient way of conducting written translation with the help of machine translation programme. The advantage of such programme is that it converts source text into another language a lot faster than a translator could do it by hand. It proves to be a great convenience given the fact that translators always work under conditions of limited time. Besides, machine translation can successfully deal with simple sentences which one meets in the text. However, during the given work I ran across two out of fourteen sentences which were machine-translated correctly. It means that machine translation has many weak points. First of all, it does not have a sustained system of terms it employs: having checked the text, I found that the same phrase «безотвальная обработка» was translated as «non-tillage» at the beginning of the text and as «non-fallow processing» at the end. Secondly, it translates sentences word-by-word, so the whole segment often sounds unnatural or even senseless in the target language. Thirdly, machine translation does not change the structure of sentences to meet the standards of the target language syntax, in this case – of English, which has a system of strict patterns of sentences. The given programme transfers the word order from the source sentence to the target one and this sometimes violates the rules of English grammar. All in all, one can come to conclusion that machine translation is a tool which helps to convert files from one language into another within a minimum time, but still it leaves very much room for translator's work and interpretation.

Appendix 1

Unedited translation

Influence of the system of basic processing techniques in the conditions of soil heterogeneity of gray forest soils on crop productivity

Report. In a stationary field experiment on gray forest and gray forest soil with a 2-m humus horizon, it was found that the reserves of productive moisture in the meter layer during the growing season of barley, oats and perennial grasses (clover + Timofeevka of the first year of use) did not depend on the depth, system of methods of basic soil treatment. In the soil with the second humus horizon, there is an increase in the reserves of productive moisture. On gray forest soil, a high yield of barley was obtained on variants with annual non-tillage of 20 – 22 cm (48.2 C/ha), with long-line plowing of grasses at 28 – 30 cm and subsequent non-tillage of 6 – 8 cm (47.7 C/ha – 48.6 C/ha). On the gray forest soil with the second humus horizon, an increase in the yield of barley was observed in comparison with the variants located on the gray forest soil. The maximum yield was obtained on the variant with an annual non-fall processing of 6 – 8 cm (54.2 C/ha). The best conditions for the formation of a high yield of barley provided gray forest soil with a second humus horizon, where it was 51.0 – 54.2 C/ha (NSR 05 = 2.2 C/ha), compared with gray forest – 44.6 – 48.6 C/ha (NSR 05 = 2.5 C/ha). On gray forest soil with a second humus horizon, oat productivity was higher compared to gray forest soil. There was no significant difference in oat yield in the main processing system variants. At the first mowing of clover in the first year of use, the yield of hay on gray forest soil was at the level of 31.4 – 36.5 C/ha (NSR 05 = = 10.0 C/ha), the second mowing – 11.2 – 13.3 C/ha (NSR 05 = 3.1 C/ha). On gray forest and gray forest soil with a second humus horizon, this indicator corresponded to 42.6 – 49.4 C/ha and 50.8 – 55.5 C/ha, respectively. The highest rates of clover yield, both at the first and second mowing, were observed on variants located on gray forest soil with a second humus horizon.

Keywords: soil, gray forest neopodzolennaya and slaboopodzolennaya, gray forest strongly podzolistaya with 2-m humus horizon, methods of main processing, non-fallow processing, plowing, long-line plowing, productive moisture, spring barley, oats, perennial grasses, growth of plant mass, productivity.

Appendix 2

Edited translation

Influence of the system of primary tillage practices on crop productivity under conditions of heterogeneity of grey forest soils

Abstracts. In a stationary field experiment on gray forest and gray forest soil with a 2-m humus horizon, it was found that the reserves of productive moisture in the meter layer during the growing season of barley, oats and perennial grasses (clover + first year timothy-grass) did not depend on the depth, system of practices of primary tillage. In the soil with the second humus horizon, there is an increase in the reserves of productive moisture. On grey forest soil, a high yield of barley was obtained on variants with annual no-till treatment of 20 – 22 cm (48.2 C/ha), with layer plowing of grasses at 28 – 30 cm and subsequent no-till treatment of 6 – 8 cm (47.7 C/ha – 48.6 C/ha). An increase in the yield of barley on grey forest soil with a second humus horizon was observed in comparison with the variants located on grey forest soil. The maximum yield was obtained on the variant with an annual no-till treatment of 6 – 8 cm (54.2 C/ha). Grey forest soil with a second humus horizon provided the best conditions for the formation of a high yield of barley, here it was 51.0 – 54.2 C/ha (NSR 05 = 2.2 C/ha), compared with gray forest – 44.6 – 48.6 C/ha (NSR 05 = 2.5 C/ha). Oat productivity was higher on grey forest soil with a second humus horizon compared to gray forest soil. There was no significant difference of oat yield despite the differences of the systems of primary tillage. At the first mowing of first year clover, the hay yield on gray forest soil was at the level of 31.4 – 36.5 C/ha (NSR 05 = 10.0 C/ha), at the second mowing – 11.2 – 13.3 C/ha (NSR 05 = 3.1 C/ha). This indicator corresponded to 42.6 – 49.4 C/ha and 50.8 – 55.5 C/ha on gray forest and gray forest soil with a second humus horizon, respectively. The highest rates of clover yield, both at the first and second mowing, were observed on variants located on gray forest soil with a second humus horizon.

Keywords: soil, gray forest unpodzolized and weakly podzolized, gray forest strongly podzolized with 2-m humus horizon, practices of primary tillage, no-till treatment, plowing, layer plowing, productive moisture, spring barley, oats, perennial grasses, growth of plant mass, productivity.

Учебное издание

СЕЛИВЕРСТОВА Оксана Александровна

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ПЕРЕВОДЕ

Учебно-практическое пособие

Редактор Т. В. Евстюничева
Технический редактор Е. А. Лебедева
Корректор Н. В. Пустовойтова
Корректор иностранного языка Н. А. Авдеева
Компьютерная верстка Е. А. Кузьминой
Выпускающий редактор А. А. Амирсейидова

Подписано в печать 22.12.20.
Формат 60×84/16. Усл. печ. л. 5,58. Тираж 50 экз.
Заказ
Издательство
Владимирского государственного университета
имени Александра Григорьевича и Николая Григорьевича Столетовых.
600000, Владимир, ул. Горького, 87.