NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Nikolai Karpov, Vera Sibirtseva, Dmitriy Bogdanov, Anna Dmitrieva, Elena Elian, Eugeniy Kleshnin, Ekaterina Markina, Tatiana Teplukhina, Lubov Violentova

DEVELOPMENT OF MODERN ELECTRONIC TEXTBOOK OF RUSSIAN AS A FOREIGN LANGUAGE: CONTENT AND TECHNOLOGY

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: HUMANITIES
WP BRP 06/HUM/2012

SERIES: *HUMANITIES*

Nikolai V. Karpov[1]*, Vera G. Sibirtseva*[2]*, Dmitriy Bogdanov, Anna Dmitrieva, Elena Elian, Eugeniy Kleshnin, Ekaterina Markina, Tatiana Teplukhina, Lubov Violentova* [3]

DEVELOPMENT OF MODERN ELECTRONIC TEXTBOOK OF RUSSIAN AS A FOREIGN LANGUAGE: CONTENT AND TECHNOLOGY

The paper considers the features of selecting the teaching illustrative material for the theoretical part of a multimedia textbook on Russian as a foreign language, and describes the peculiarities of compiling a set of exercises on the basis of the National Corpus of the Russian Language. The author(s) analysed in detail the difficulties caused by working with the National Corpus of the Russian Language for educational purposes and problems emerged in the process of working on a project aimed at creating an electronic textbook "The Russian verb. Word formation". Technology of electronic textbook was created, which is used for learning foreign language. The helpful tool tip was added in the textbook. It was compounded with grammatical information and English translation for each separated word.

JEL Classification: Z19.

Keywords: RNC, electronic textbook, Russian as a foreign language, lexical and grammatical exercises, multimedia information systems, intelligent Internet-technology, semantic modeling, distance learning

[1] National Research University Higher School of Economics. Faculty of Business Informatics and Applied Mathematics; senior lecturer; E-mail: nkarpov@hse.ru

[2] National Research University Higher School of Economics. Department of Applied Linguistics and Intercultural Communication, associate professor; E-mail: vsibirtseva@hse.ru

[3] National Research University Higher School of Economics. Faculty of Business Informatics and Applied Mathematics, students

Introduction

The form of distance education has existed since the late 19th century, and with the advent of Internet technologies available it is firmly established in our lives.

Electronic course materials are becoming more popular than printed, and under certain conditions may not only improve the efficiency of the educational process, but also turn it into a new format. In recent years there has been an active penetration of information technologies in very traditional, well-established fields of education, one of which is teaching Russian language, in particular, Russian as a foreign language (hereinafter – RAF).

The quality of e-learning in general and e-books in particular, depends on the technologies used, because the complexity of the electronic textbook, not only impedes the perception of educational material, but also cause a certain rejection by the student. In the development of an electronic textbook on the basis of the Russian National Corpus (hereinafter - the RNC) not paid electronic publishing was preliminary analyzed, but existing network grammar manuals, available for free use.

The most interesting on-line tutorials are available at the Moscow Financial University (http://www.dofa.ru/open/book/1_russ/u123.htm), People's Friendship University (http://www.webmetod. Narod.ru / ddd / GlagDvig / Index.htm), an unknown developer (http://russianmentor.net/Ru_xx/STARTHERE.HTML), of The Lexiteria Corporation (http://www.alphadictionary.com/rusgrammar/index.html).

Because the developers’ plans included the use of the RNC material and software corresponding to specific linguistic requirements, as a result we had to make a pilot fragment of a textbook on RAF, aimed at the advanced level of training, work with which is free and remote.

Thus, the development of the project was carried out on two fronts: content filling and structure of the textbook.

The content of the textbook

As a fragment of the textbook the theme "The Russian verb. Prefixal word-formation" was chosen, as one of the most difficult for the assimilation by foreign students. It is assumed that in the future the electronic textbook will be supplemented with materials on other parts of speech. However, at this stage it was important to put a really achievable goal and create a product balanced from the point of view of linguistics and programming. Address to the RNC materials allowed us to avoid repetitive examples in the exercises and really focus on the living, actual word usage. At the same time, the richness of lexical composition of RNC determines the highest level of exercise difficulty, after the creation of which additional examination of linguistic data and stylistic differentiation of the material is required.

The Russian Corpus is characterized by deep, diverse, and constantly improving marking (metatext, morphological, syntactic, semantic and accent), it greatly simplified the work of the compilers in the selection of illustrative material. The main tasks in the preparation of textbooks included the selection of RNC examples to illustrate the rules and make up exercises. The complexity of any textbooks creation related to the exhausting search of examples, requiring considerable time costs and to the limited imagination of the author in selection of exercises.

For selection of examples in the theoretical part of the textbook it is appropriate to appeal to the RNC custom search. The portal provides a detailed instruction "How to use the corpus", but the search page is intuitively understandable and comes with handy pop-up prompts, and the search is organized logically enough, so the choice of examples with specific word forms did not cause difficulties.

For educational purposes the feature of giving results in Excel is useful, you can work with the illustrative material for one linguistic phenomenon, containing about 2,000 examples. The more well-constructed and limited request is, the less side information is contained in the received examples. However, in the process of working with the RNC we had to deal with some difficulties.

The Corpus provides the ability to search for several words, and semantic-grammatical, so you can get a clear enough limited request. The apparent advantage of the request for RNC is the ability to change the settings of the delivery of examples, although with every new request to the search page, these settings are not saved and they have to be re-entered, as RNC does not provide saving of search criteria by default.

The required examples are not always displayed with accents, although there was the possibility in search settings. Because the accent in the Russian language is non-fixed (as opposed to, for example, French or Czech languages), this feature in a textbook for foreign students is obligatory. In further work on the textbook the issue of accent was successfully solved by the programmers, as it is reported in detail in the next section of the article.

Another disadvantage of RNC is the inability to set a search parameter, like the words with a particular morpheme in the middle of the word (for example, the suffix -ну-), which in some cases has a significant influence on the meaning of the word. Also, there is no search on prefixes, for example, specifying in the request "пере-" and grammatical features: verb, indicative mood, we received examples of words beginning with "пёр" (the verb "прут").

As for the filtering of examples in Excel, then, as it turned out, the words are only in a certain order in which the initial text search is performed (user changes of the settings are not fully taken into account), namely, the search system finds some text and selects from it all possible variants that match the specified criteria. Then it moves to the next, and analogically presents the examples. In this scheme, there is undoubtedly a positive thing: if the subject matter of the text from which the example was taken, does not fit initially, you can skip a number of examples of the same text and move to the next. But there can be many such texts, and no ability to specify the preferred, and sometimes even after consideration of all the examples that are saved in Excel, 30 verbs (needed to accomplish the original problem) were not gained. We had to think out the possible cases by analogy, limit search parameters, and view examples of specific verbs. A major drawback of RNC was a significant slowing down the server, directly proportional to the amount of simultaneously set required categories of verb (mood, tense, transitivity / intransitivity, reflexivity, the presence of semantic features).

Despite these difficulties, the search for examples for the theoretical part of the electronic textbook based on the RNC can solve two major problems faced by developers of teaching materials: first, the possibility of adjusting the search on grammatical and semantic criteria, and the subsequent issuance of the results in Excel format allows to automatically obtain a sufficient number of examples, from which we can select the required manually. Secondly, the texts that are accessed by RNC, and really modern and only the developers’ sense of proportion and taste is a criterion for the selection of the material.

If the main purpose of the selection of examples for the rules is to demonstrate the use of the verb in the required situation, in the preparation of exercises you should look for examples that are appropriate to a particular rule, and bear in mind the possibility of restoring a verb or its part (for example, the prefix) in the context.

One of the major advantages is the possibility of making the exercises, which are based on the marked video clips from Soviet and Russian feature films (RNC multimedial subcorpus), which are available on video-hosting "Yandex Video". Fragments, close to living communication, and not read by one speaker, transmitting at the same time the wealth of intonational structures of the Russian language in a professional performance - a real godsend for compilers of textbooks. Multimedia corpus permits requests providing us with sufficient material for the exercises. Each video clip is provided with decoding a replica with metamarking and marked accents. Implementation of this work by hand, without reference to the Corpus materials, is not possible. Most of the time was occupied by listening to each movie, because it was necessary to assess the overall sound quality, the breadth of the context, the presence of background noise in the replica or the dialogue.

The negative moment in the sentences selected with the help of the RNC and received in Excel-format, is that the source of many of the texts are online forums where users communicate in spoken language or jargon, often incorrectly and illiterately (the words are not always used correctly for a particular context , there are mistakes in spelling). Content aspect of this kind of examples also sometimes wears a primitive character. These drawbacks suggest that for the use in the exercises RNC examples should be thoroughly pre-filtering.

The developers’ focusing on only one, rather narrow aspect of study of the verb in the Russian language, has played a positive role. The Corpus materials are so varied and diverse (for example, we can mention a variety of subcorpuses: accentologic, news, dialects, multimedia, training, parallel, poetic, syntax, oral) that only a clear phased work allows a detailed investigation of all its features and use them optimally in developments. In the long term - exercises not only in morphology but also in syntax, stylistics, lexicology, historical morphology and other aspects of language. The plans of the developers of the project also include compilation of exercises with more extensive use of multimedia RNC subcorpus.

The above features of work with using RNC materials relate only to the content, linguistic aspects of the project, difficulties and discoveries, faced by the members of the project - programmers, whose main task was to create a structure of an electronic textbook on RAF, are no less fascinating and important.

The structure of the textbook

To link the RAF materials we use the technology for creating an electronic textbook that supports the format of the SCORM (Sharable Content Object Reference Model) to implement it in any modern system of LMS (Learning Management System). A feature of this textbook is the ability to view real-time actual examples from the RNC, and a tooltip containing the translation of words into English, the primary form of the word and the corresponding grammatical information.

As a platform for filling of the electronic textbook in the format of SCORM course we chose a XHTML (Extensible Hypertext Markup Language) editor for e-learning materials eXeLearning, distributed free. This editor is a tool for teachers and researchers, allowing to design, develop and prepare the publication of educational and training materials in electronic form.

During the development of the textbook several problems have been solved: we designed a tree-like structure of the electronic resource corresponding to the structure of educational material, with the help of eXeLearning editor we organized theoretical and practical parts of the textbook, where the basic information elements were added.

The electronic textbook developed during the project, had to have benefits such as interactivity, visualization and control – possibility of independent view of the material and doing the assignments in any order. The theoretical part of the textbook is filled with materials pre-selected from the RNC. The main component of the theoretical part of the textbook is the rules for the use of prefixes to the verb, and a number of examples that are available after clicking the appropriate button.

Additional element that improves understanding of the use of examples in the textbook is a tooltip containing the translations and grammatical information of each word. The data are generated by server software and transmitted to the user's browser, where there is only a drawing of a pop-up window with the acquired information. Obtaining the necessary information is produced on a particular signal of a student or, in terms of the program, during a certain event in the user's browser.

A program running in the user’s browser has access to every single word as to the DOM element in an HTML page over the selector ".tooltip". To implement the tips it’s necessary to make a link in the page to the JavaScript file and put each word within the tag with the class "tooltip". To do this, we developed a special program «Tagger» in Java, which marks HTML text and adds the specified tag to every single word.

This method allows you to modify the contents of the pop-up window and the logic of the program, located on the server, without modifying the text of the textbook.

As the source of the translation the service MyMemory with open API (Application Program Interface) was chosen. With the help of JavaScript the textbook can do a HTTP (Hyper Text Transfer Protocol) request, the answer to which comes in the format of JSON (JavaScript Object Notation), from which the data are easily extracted by standard methods.

For the development JavaScript library - jQuery was chosen. Library is suitable for interaction with the DOM (Data Object Model) - page elements and building HTTP requests. Formation of HTTP requests is done with the help of the standard function ajax (), library jQuery, working on the basis of the same name technology. To synchronize the query the option «async: false» was specified.

The formation of the correct grammatical information is on the server side using the program of morphological analysis of the Russian language for noncommercial use - Mystem. This is a separate module that runs as a console application. Commercially Mystem is not distributed and will not be distributed. The acquired information is processed by a specially designed server-side application written in C language. It is used to ensure communication with the server via standard CGI (Common Gateway Interface), and to extract relevant data and convert them to JSON format. In the architecture of the developed application system is running a separate process, which calls the console program Mystem. The output of the program is recorded and processed. The acquired information is divided into the token itself and grammatical information to this token. The information is structured for its further use in accordance with the format of data exchange JSON.

As a result, we developed a functioning tooltip that appears when you single-click the right mouse button on a selected word, and contains a translation of the word in the English language and its grammatical characteristics in Russian.

The main problem in the selection of examples was the lack of stress in most words. In the end, the placement of stress in words was made in all the examples. Here we consider several ways of setting an accent mark. Accent mark "́" - a sign of non-alphabetic spelling of Russian writing, in other terms - one of the superlinear diacritical marks, is placed above the vowel corresponding to the stressed sound.

In the project we used a complementary diacritical sign encoded in Unicode «́», which is encoded as U +0301 (769 in decimal, that is in the text of html-document you should write &#769;) The same as combining characters like it, if it comes after the stressed letter, it unites with this letter. That is, the emphasis will be on the letter, and not after it.

The advantage of this method is that this way you can put the accent over any letter. Programs that cannot correctly display the symbol, still understand that it is complementary symbol and just skip it. This is important for automatic translation of the word, and the allocation of grammatical information.

The practical part of the electronic textbook is called "The use of prefixed verbs in speech. Exercises. " To this section we added a variety of exercises on a single / multiple choice, filling gaps, specially designed video exercises, for which we selected the most relevant video clips from the multimedia RNC subcorpus posted on the video-hosting "Yandex Video", and multimedia elements of HTML-text are integrated into the exercises.

The exercises are used for working out the studied rules of the use of prefixes that have been learned in the theoretical part of the textbook, reinforcement of knowledge and expanding a student's passive vocabulary. Also final control tests were added to the book, you can use them to check the quality of learning material.

In the process of exercise making we revealed a drawback of the program eXeLearning, connected with the choice of the size of the text fields to fill the gaps. The program chose the length of the field equal to the number of letters in the missed fragment, which gave a significant hint to a student. To solve this problem, the function of creation tasks with gaps has been changed (the formation of built-in element instructional device CaseStudyDevice) so that the size of the gaps is not dependent on the length of the missing prefix.

In the theoretical part of the electronic textbook, in addition to the possibility to analyze the effect of rules on the sentences selected by the compiler and written in the book, the opportunity was organized to address RNC by pressing the button and pick out "live" examples. These "live" sentences are extracted in real time from the constantly renewing corpus and, therefore, always include various and actual examples of use of words. The accuracy of the semantic meaning of an example of the use of a prefix depends on the ability of a linguist - compiler of the textbook, to formulate a query to search for the RNC as narrowly as possible, to cut off possible getting of sentences in which the search word is used in another meaning.

The program, selecting actual examples of RNC for the electronic textbook of Russian, should work on the server and correspond to the standard CGI. The task of implementing a server application using the Python language that does not use third-party libraries, was split into separate subtasks.

The first sub-task was getting results from the RNC site; it was solved by the transfer of the specified HTTP-request to the server of the Russian language and parse any incoming response from it.

We illustrate the mechanism described by a query that extracts examples of the use of verbs with the prefix "в" and indicating the orientation inside (вбить, втолкнуть, внести, воткнуть, всунуть). To illustrate the rules we should find verbs that govern accusative with the preposition "в" (в комнату). To prevent getting of participles and verbs in the subjunctive and the imperative, in the lexical-grammatical search restriction by grammatical features is given: verb - indicative mood. The step of searching a preposition from the verb is set (1-2 words), as often a direct object is after the verb before the preposition (втолкнул ЕГО/МАЛЬЧИКА в комнату).

Content of request was developed and optimized so that we could always get the answer, the most relevant to the given section of the textbook. At the same time a small percentage of examples, that got into the textbook mistakenly, is possible. This is the downside of using a "live" search of examples in the Russian National Corpus. We decided to level this drawback with the help of a multipage view of search results in multiple requests to the server.

In the second stage, when the results are produced, examples are sorted. For this we use the settings: select giving a single example of one author, order the location of the sentences, for example, by creation date, indicate the number of examples on the page and request the variant with accents. The latter does not always work, if the examples are taken not from the main corpus, but from the newspaper subcorpus, the accents are usually absent. Concerning the normativity of the examples, the most suitable subcorpus is the teaching subcorpus, as it is the most completely marked, but its volume is very limited and not always the required examples are present in sufficient quantity.

The response from RNC to the transmitted request is stored and transmitted to the input of the state machine by means of which the second task is implemented - the analysis of the results to create dynamic examples. The machine is designed to select examples, key words for search and to cut off unnecessary information. In order to implement each part we use a separate pair of states of the state machine. The alphabet of the state machine is HTML-tags, and a number of symbols of any human language in the selected encoding.

The third sub-task - issuing of search results into structured HTML text has been implemented due to the fact that the state machine distinguishes different types of words, for every of which techniques of presentation in HTML are implemented. These ideas are combined and added to the page template, and then the created page is sent to the client's computer.

Conclusion

The subject of development of electronic textbooks is now, of course, important, since with the advent of distance education in electronic form, its active implementation to the universities and the corporate sector, new approaches to provide knowledge, fast and cheap, are required. Analysis of existing electronic textbooks of Russian as a foreign language showed that they are designed for elementary and basic level of language learning and, accordingly, contain low-level exercises. Identified gap in the basis of textbooks on RAF presented on the Internet and on digital media (namely, the lack of exercises on grammar and word formation for students of second and third certification level of language), convinced the developers of the textbook that a need has formed for using quality linguistic manuals that are devoid of drawbacks typical for this type of product.

The approach to the content side of the textbook and the developed technology can be useful for creating an electronic textbook for any foreign language. Many of the world's languages have marked national corpuses, the material of which can be used for educational purposes. The time required for selection and sorting of examples, is significantly reduced when referring to the corpuses, and the relevance of examples may be given initially, by limiting the time frame of the search. A large volume of corpus (for example, the RNC has more than 300 million words) allows you to create manuals on syntax, stylistic differentiation of speech, use of certain language constructs.

The technology of creating this product uses a free XHTML editor for the formation of the text of the book and a set of server applications for the implementation of the selected function. Applications were written in Python and C without the use of third-party libraries. The program running in the user's browser and interacting with and server applications is developed in JavaScript using the open-source library jQuery.

The use of deeply marked-up linguistic corpus, rich in all kinds of texts of different genres, helps to make the study of linguistic phenomena and the use of words more understandable to the student.

In this scientific paper we used the results obtained in the course of the study "The Russian verb: development of an electronic textbook "Russian as a Foreign Language" at the RNC material, made in the framework of the" Research Fund HSE "in 2012, grant № 12-05-0027.

References

1. API technical specifications. // MyMemory (http://mymemory.translated.net/doc/spec.php)
2. Dobrushina Nina. Corpora metods in education Russian // Russian National Corpus 2006-2008, 2009. SPb (in Russian).
3. How to use accent? // Slovomania. (http://www.slovomania.ru/dnevnik/2007/08/11/how-to-use-stress-sign/) (in Russian).
4. Levinson Anna. Using the RNC in Teaching "Rhetoric" in high school // Russian National Corpus and the problems of humanitarian education. Moscow, 2007 (in Russian).
5. Technologies. About Mystem // Company Yandex. (http://company.yandex.ru/technologies/mystem/) (in Russian).
6. The eXe project: eXeLearning. (http://exelearning.org/).
7. What is the Corpus? // Russian National Corpus (http://www.ruscorpora.ru/corpora-intro.html) (in Russian).

Vera G. Sibirtseva
National Research University Higher School of Economics. Department of Applied Linguistics and Intercultural Communication, associate professor;
E-mail: vsibirtseva@hse.ru; Tel. +7 (952) 767-25-20

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.