Editing an automatically-generated index with K Index Editing Tool

Kseniya Egorova

K Dictionaries, Tel-Aviv, Israel
E-mail: kseniya.a.egorova@gmail.com

Abstract

This paper presents the editing process of a new Russian–English index using dedicated software. The initial index was generated automatically from the semi-bilingual Password English learner's dictionary for speakers of Russian and the editing was carried out with K Index Editing Tool (KIET). Initially, the editor was provided with the raw index produced according to a set of pre-established principles. It contained all the Russian translations from the Password database, converted to potential Russian headwords arranged in alphabetical order and accompanied by the part of speech of the original English equivalents. The revision process then consisted of modifying, removing or adding headwords, confirming or amending their automatically associated part of speech, and matching and re-ordering links to their English equivalents. At the final stage the index was proofread line by line for spelling and grammar mistakes, resulting in a change in index size from 31,666 to 29,039 headwords with 45,929 senses. The paper also demonstrates the main features of KIET and highlights some of the problem areas and major challenges we faced while revising the index.

Keywords: Russian–English index; automatically-generated index; editorial tool

1. Technical description

K Index Editing Tool (KIET) is a new editorial software for creating indices of *Password* semi-bilingual English dictionaries for any language. The initial bilingual list is automatically generated according to a set of pre-established editorial principles, so the Russian target language (TL) translations from the dictionary database are reversed into headwords and the original English source language (SL) headwords are converted into their potential translation equivalents. The automatic generation of the index consists of several steps including XML data parsing and building basic SQLite tables. First of all, the software searches the database for all translations, which are known as translation containers in XML. Subsequently, each translation container is linked with the sense set, which includes several elements: a definition, examples and a headword with part of speech label. The main parameter used for creating basic tables for each language is the definitions, constituting the main attributes of the linked sense, and sense identifiers. Next, the software uses the resulting tables for further parsing. At this step, it identifies translations, which contain commas and semicolons inside the text, and automatically parses them into several parts, divided by these characters. Subsequently, these parts are also turned into separate headwords. The newly-built raw index has the following elements:

- TL translation (turned into headword)
- part of speech
- SL definition
- SL examples (if needed)
- SL senses

Finally, the software links all the sense sets associated with a TL headword. See Figure 1 for the microstructure of a TL entry.

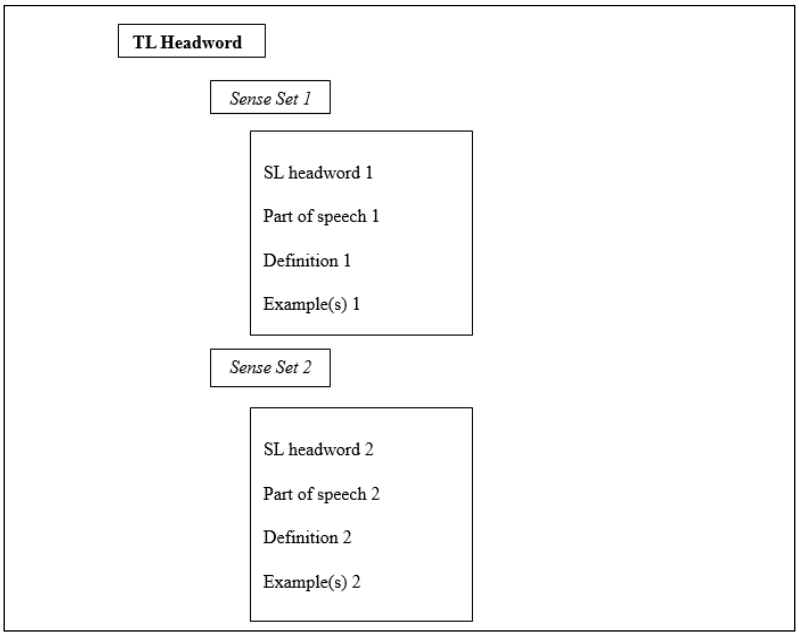

Figure 1: Microstructure of a TL entry

Sorting of the generated index is performed according to the TL alphabet. Subsequently, the editor is provided with the initial index for further editing in KIET.

2. Description of the editing process

The main editing task was to keep the entire structure simple and shape it into a cohesive and comprehensive unit. As the index was intended for Russian speakers, it was important to provide, in one entry, links to all possible English equivalents ('senses') associated with the Russian headword and to make them easily accessible. The entries are displayed in a simple way: corresponding English senses are ordered in a flat structure and followed by definitions (see Figure 2). Examples are not visible in this section. However, when needed, examples of usage and other additional dictionary data can be looked up in full entries.

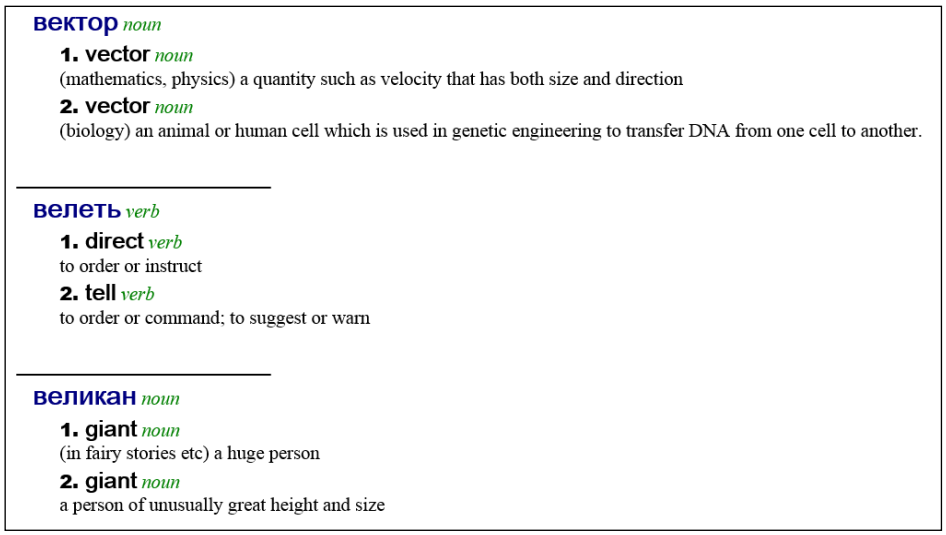

Figure 2: Preview of the index

In brief, the editing process of the Russian–English index can be described in four steps:

(1) modifying, removing and adding the Russian headwords

(2) adjusting part of speech labels

(3) revising and reordering the list of related senses

(4) exporting and proofreading the final index

The following sections of the paper will detail each of these editing stages. First, however, it is necessary to provide a short overview of the tool's functionality. The majority of the editing was performed in the KIET main screen, which consists of three main parts (see Figure 3). On the left is the list of all headwords. In the middle, the editor can view the list of related senses associated with the headword. The current entry structure is displayed in a dictionary-like form in the entry preview window (on the right). The examples are visible only to the editor to assist in decisions regarding the senses. The icons at the bottom of the main screen (from the left to the right), are used to perform the following actions with the headword list:

1. Edit current Headword
2. Duplicate Headword
3. Add new Headword
4. Remove current Headword
5. Restore current Headword

6. Save changes made to the Headword list

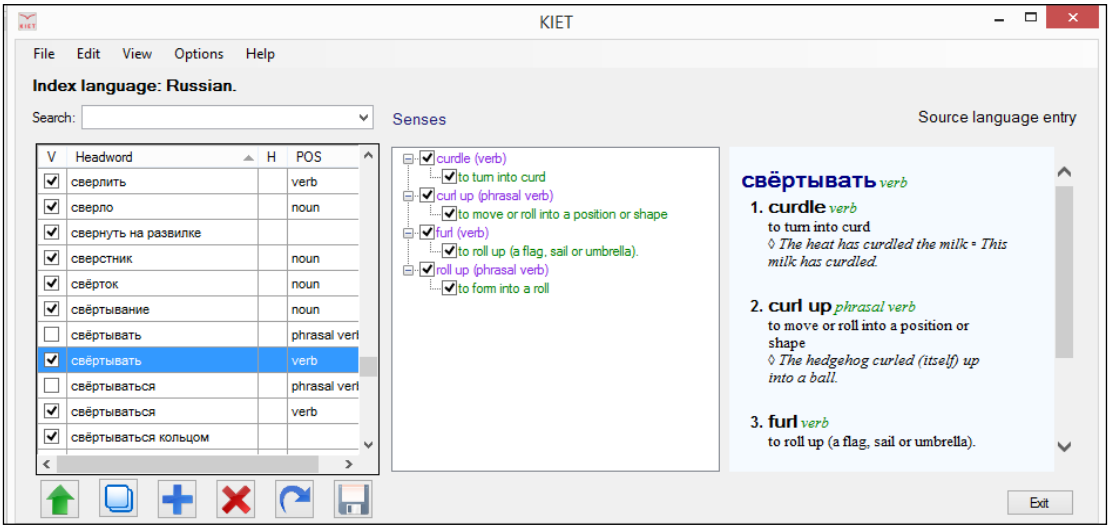

Figure 3: the KIET main screen and its functional buttons.

These functional buttons are used during various stages of editing.

2.1. Modifying, removing and adding headwords

The first editing task concerned reviewing the automatically-generated Russian headword list to check the translations-turned-into-headwords for accuracy and comprehensiveness. The editing was performed in KIET by choosing Select/Unselect a Headword (in the main screen on the left) and checking or unchecking the checkbox preceding it to determine whether or not the headword will be displayed in the dictionary index. In other words, each headword may be set as visible or invisible in the list of selected headwords (e.g. as applied to the redundant headword '**свёртываться**' (curdle) displayed in Figure 3). Editorial revision at this stage included taking decisions about which headwords should remain unmodified, be modified in different ways, or be removed altogether (buttons 'Edit entry' and 'Remove current entry', respectively). With KIET it is not possible to physically remove any headword from the initial database but rather it is indicated for later automatic removal by the software from the dataset once editing is complete. It also enables the editor to add new headwords to the headword list if appropriate (buttons 'Add new entry' and 'Duplicate'). In case a newly-modified or added headword happens to already exist elsewhere in the index, KIET displays it to the editor for further consideration.

As the lexical structure of the headword list depended on the *Password* dictionary translation database, there were several types of automatically-formed headwords:

(1) Direct translations

(2) Approximate translations

(3) Explicative definitions which served as descriptions when there were no equivalents in the TL

Namely, particular challenges were encountered with the second and the third type of translations, in cases when the candidate Russian headword stemmed from them. Such headwords had to be rephrased or shortened into a multi-word expression (if possible) or had certain elements extracted as new headwords to suit the full framework of the edited index and to be comprehensive for its users.

It is important to note that due to the KIET pre-settings the editor was not able to make any corrections in the SL (English) 'part' of the dictionary database (including the original source language headword, their part of speech labels, examples and definitions). Only the TL 'part' of the database could be edited and modified.

2.1.1. Lexical types of headwords

The Russian headword list consisted of the following types of items: *simple words, abbreviations, partial words*, and *multi-word expressions* (MWEs). *Simple words* included both *lexical words* (nouns, adjectives, verbs, adverbs and interjections) and *grammatical words* such as prepositions, conjunctions, pronouns, numerals and particles. *Partial words* (productive affixes and combining forms) were also given headword status as many of them are frequently used in Russian: e.g. ***про-*** (pro-), ***недо-*** (under-), ***дву-*** (bi-), ***авто-*** (auto-), etc.

MWEs[1] included collocations, fixed and semi-fixed phrases, similes, phrasal idioms, greetings and phatic phrases. Below we give some examples of MWEs from the headword list. As Anokhina (2010) points out, when compiling a bilingual dictionary it is difficult to distinguish between fixed or semi-fixed phrases and collocations, especially those with unconventional translations (even more so for the Russian language, though this is not covered in this paper). Thus, we put first three types of MWEs into one group here:

(1) Collocations and fixed or semi-fixed phrases: e.g. ***оказывать влияние*** (to bias), ***проводить кампанию*** (to campaign), ***дурное предчувствие*** (misgiving, foreboding)

(2) Similes: e.g. ***холодный как лёд***[2] (stone-cold, stone-dead, stone-deaf), **как бешеный** (like fury), ***словно живой*** (lifelike)

[1] Here we follow the classification of multi-word expressions given by Atkins & Rundell (2008: 166–171).
[2] Russian similes may be (and usually are) translated with the English equivalents belonging to other types of MWE or even to single-word units.

(3) Phrasal idioms: e.g. *буря в стакане воды* (a storm in a teacup), *лезть на рожон* (to stick one's neck out), *сводить концы с концами концами* (to make (both) ends meet)

(4) Greetings: e.g. *Добрый день!* (Good afternoon!), *Здравствуйте!* (hello, hallo)

(5) Phatic phrases: e.g. *всего хорошего!* (Cheers!), *не беспокойтесь* (never mind)

The bulk of the headwords were common words, but a limited number of proper names was included as well, e.g. *Восток* (the Orient, the East), *Венера* (Venus), *Телец* (Taurus), *Ханука* (Hanukkah), etc.

2.1.2. Homograph headwords

The editorial revision of the headword list included treatment of homographs, since it turned out that the Russian homographs were not identified in the automatic parsing, so it was decided to treat homographs as separate entries. There were two types of homographs to deal with:

(1) Same spelling but different meaning and pronunciation

e.g. **атлас**[1] (with the stress on the second syllable) (satin) and **атлас**[2] (with the stress on the first syllable) (a book of maps)

(2) Same spelling and pronunciation but different meaning and capitalization

e.g. **Весы** (sign of the Zodiac) and **весы** (a weighing machine)

As a result, homographs with the same spelling but different meaning and pronunciation were duplicated and distinguished by the symbol # and an Arabic numeral (1, 2, etc.). This was performed in KIET by means of clicking on the 'Duplicate' button and making the necessary changes in the list of related senses. As shown in Figure 4, the inappropriate sense (a book of maps) was unchecked from **'атлас#1'**. That sense was linked to the duplicated entry '**атлас#2**' with this meaning. Figure 5 shows a preview of the two entries after changes were made.

The initial processing of the SL translations also did not differentiate between capitalized and non-capitalized homographs with the same part of speech, and these corrections followed manually. If their meanings were different they were also treated as separate headwords but with no homograph number distinction. The capitalization served as a sign that meanings were different (see Figure 6).

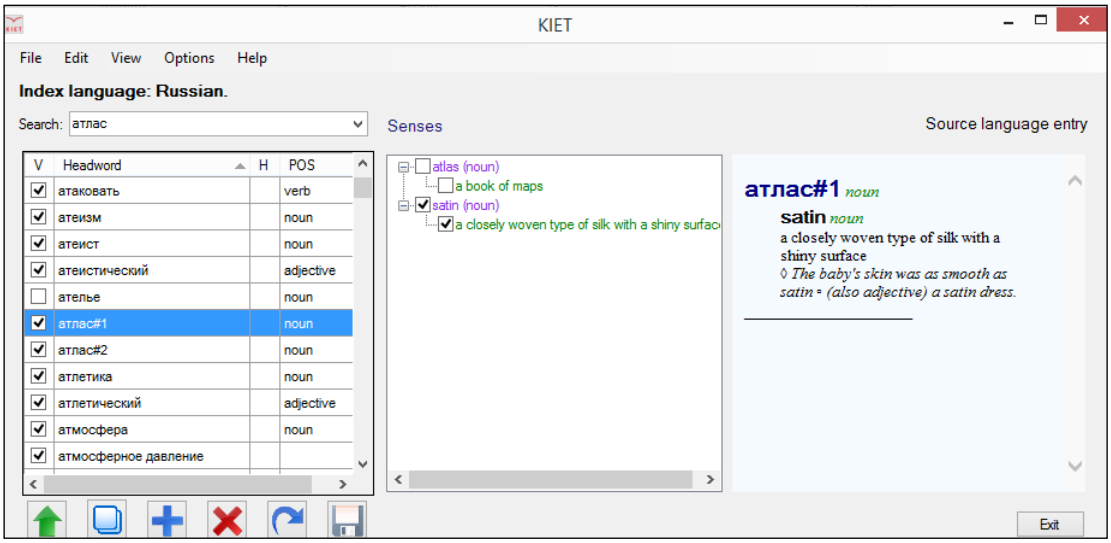

Figure 4: Entry '**атлас**[1] preview

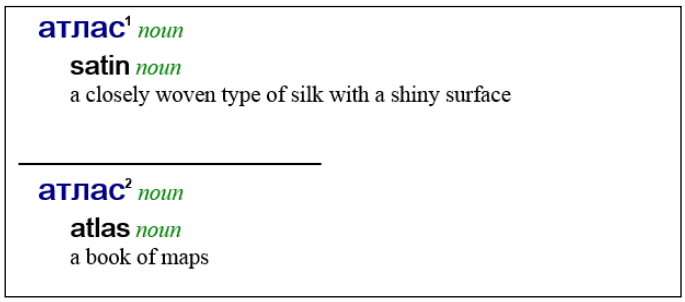

Figure 5: Entries '**атлас**[1]', '**атлас**[2]' preview

For those cases when it was difficult to differentiate homonymy from polysemy – whether it was a plurality of meanings or 'meaning' from 'shade of meanings' – the headwords were not treated as separate entries. In the case of such difficult decisions, other bilingual and Russian monolingual dictionaries were consulted.

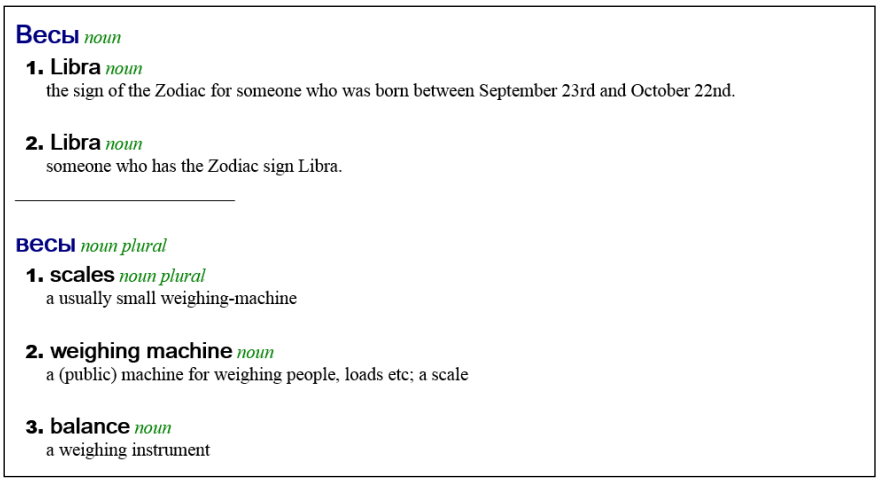

Figure 6: Entries '**Весы**', '**весы**' preview

2.1.3. Making one headword out of several parts

During automatic index generation preceding editing, TL translations that contained commas and semicolons inside the text were parsed by the software and divided by their punctuation settings into separate headwords. This worked well for the translations where a comma or semicolon were used to separate items in a series (e.g. when several synonyms denoting the same thing or object were listed), with each item becoming an independent headword. However, when these punctuations served to introduce a clause in a translation, this rule made a mess. In such cases the translation, which consisted of a complex sentence, was split into two parts that made no sense when used separately. For example, in the translation database the noun *achiever* was translated into Russian as *'человек, добивающийся успеха успеха в жизни'* (literally, 'a person who achieves success in life'). The second part of the translation, separated by a comma, is a participial phrase, which starts with a Russian present participle '*добивающийся*'. As a result of the automatic parsing, there appeared two headwords in the index, '*человек*' (person) and '*добивающийся успеха в жизни'* (someone who achieves success in life), neither of which makes any sense on its own. Subsequently, while revising the headword list, the editor's task was to find and identify such 'nonsense' or inappropriate headwords and reunify the split parts into the corresponding headword ('*человек, добивающийся успеха в жизни*').

2.2. Adjusting the part of speech labels

As explained with regards to the lexical structure of the headword list in 2.1.1, both lexical and grammatical words were included in the index. They belonged to the following word-class categories: nouns, adjectives, verbs, adverbs, interjections, prepositions, conjunctions, pronouns, numerals and particles. Due to the overall simplicity of the structure, we did not add grammatical subcategorization in the index. Thus, indications of verb transitivity/intransitivity, of their perfective/imperfective aspects or of various types of pronouns (reflexive, demonstrative, possessive, etc.) were not provided.

According to the pre-established principles, the software automatically attributed the original SL part of speech label to the TL headwords. Subsequently, if the SL equivalent did not belong to the same word-class, the part of speech had to be modified in line with the edited Russian headword or to be removed in the case of MWEs as headwords, which are not labelled at all. In the screen 'Edit Headword' the POS label may be changed by selecting from the drop-down menu the necessary word-class marker (see Figure 7). After introducing the changes, the 'Update' button was clicked to accept them.

Indeed, in most cases the Russian and English parts of speech did not correspond to each other due to several reasons.

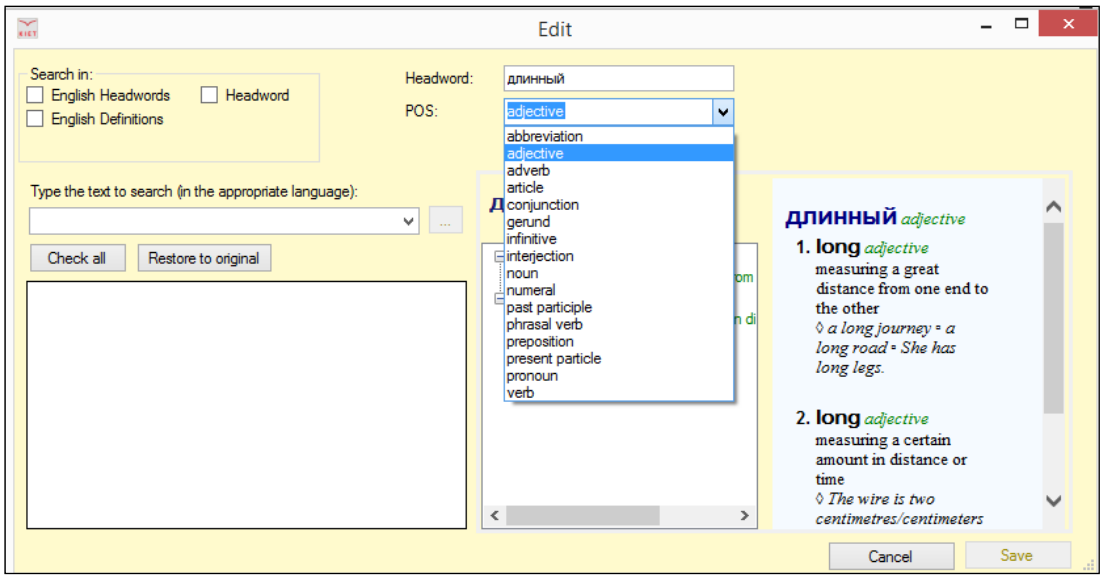

Figure 7: 'Edit Headword' screen with a part of speech drop-down menu

First, many English headwords were initially translated into Russian as a MWE or (more rarely) by a different word class. For example, the noun *bookshop* was translated into Russian as ***книжный магазин***, which is an adjective + noun fixed phrase (or collocation). Another example is the noun *intermarriage,* which is impossible to translate into Russian as a single-word unit. The typical translation is a phrase of five words of different word-class categories (N. + Prep. + N. + Adj. + N.) such as '***брак между людьми разных национальностей/рас***' depending on the context.

Secondly, some English grammatical categories do not exist in Russian (e.g. articles, gerunds and phrasal verbs). If a headword was automatically attributed this kind of 'foreign' word-class marker it had to be adjusted according to Russian grammar. For instance, additional editing was done with '*phrasal verb*' labels, which appeared frequently. Phrasal verbs are usually translated into Russian as verbs with semantically meaningful verbal prefixes (though also depending on the context, see e.g. Yatskovich, 1999; Mudraya et al., 2005). For example, in the dictionary database the phrasal verb *to wake up* was translated as ***разбудить*** (a single verb with a prefix ***раз-***). When the TL translation (***разбудить***) was converted into a headword it still retained the original English-derived part of speech label (*phrasal verb*) and had to be modified into a '*verb*' label. The editor considered all these 'phrasal verb' cases in the index and made any necessary changes.

2.3. Revising and reordering the list of related senses

Another main task of the editorial process consisted of attributing the appropriate English equivalents ('senses') for each Russian headword and re-arranging them in order. This involved not only fitting the right English translation(s) to the Russian headword, but actually linking the headword to each specific sense of English

polysemous entries that corresponded to it.

If a particular sense was not in the list, the full database was searched. KIET enables the editor to search among the original English entries and definitions or other Russian headwords in the index. A new sense is added by ticking the checkbox that precedes it and the result appears automatically in the preview section.

According to the predefined entry microstructure, the headword senses were presented in a simple flat structure and numbered 1, 2, 3, and so on. The order of the senses could be changed using the mouse to drag the selected sense and drop it in place. This could be done either in the 'Edit entry' screen or in the main screen (in the section showing the list of related senses). As Atkins and Rundell point out "…'dictionary senses' in a bilingual dictionary are not really senses of the headword at all, but simply the most user-friendly way to structure the material. Bilingual dictionary senses are predicated more on the TL than on the actual meaning of the SL headword" (Atkins & Rundell, 2008: 500). At this stage of editing we stuck to these rules and tried to lay out the senses in a user-friendly way, based on the presumption of which sense the user will look up first. Therefore, we chose the semantic order, putting first the 'core' or most common meaning, as judged intuitively. We did not follow the frequency order, as this required a parallel corpus and a frequency analysing software which we lacked. As a rule, the commonest meaning usually consisted of the direct translation of the Russian headword or the most neutral word (in style and register) when selecting among several translation variants from the database. Figure 8 shows the headword '**вверх дном'** (*upside down*) linked to three English senses. The first two are synonyms and the last one is a contextual, indirect translation that was linked with the Russian TL translation in the dictionary database. Therefore, we placed the 'safest' meaning (*upside down*) first followed by the less common or stylistically different variants.

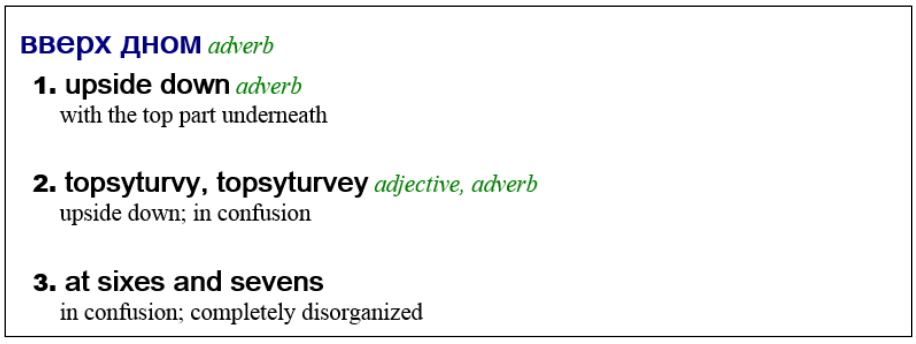

Figure 8: Entry '**вверх дном**'

In cases when senses that were linked to the headword happened to be regional variants, they were also ordered in the same way. For instance, the Russian '*багажная тележка'* was formed from two 'senses' – *luggage cart* in British English and *baggage cart* in American English – that were in fact derived from a single entry.

They were subsequently numbered sense 1 and sense 2, with preference given according to the editorial style guide to the American variant. As a result of this rearrangement, this entry appeared as in Figure 9:

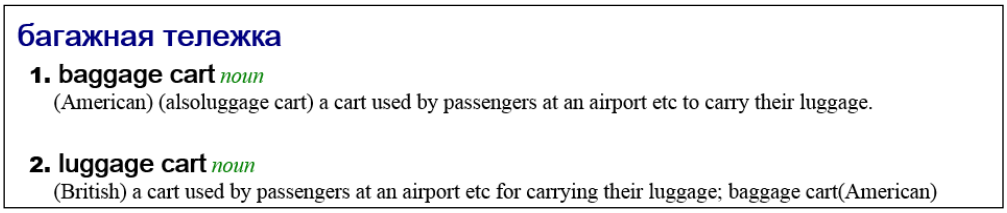

Figure 9: Entry '**багажная тележка**'

The entries that consisted of the full translation equivalent and its contracted or abbreviated form were also presented in a flat structure with the full form always first and the contraction/abbreviation after. For instance, as two English equivalents were linked with the headword '*суббота*' (*Saturday* and *Sat.*), we rearranged their order using the drag-and-drop function and listed *Saturday* as sense 1 and *Sat.* as sense 2 (see Figure 10).

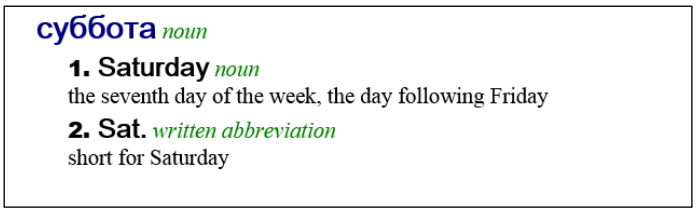

Figure 10: Entry '*суббота*'

2.4. Exporting and proofreading the final index

Finally, after all changes had been saved in the database, the edited index was exported from KIET and the export files were sent for processing. The features of KIET also enable the editor to create HTML files and see all the performed changes and the final result in a user-friendly format. When the data had been processed, the entire index was proofread line by line (in HTML-format on a screen) for spelling and grammar mistakes. The POS-labels and the linked senses were double-checked once again.

3. Conclusion

This paper gave an overview of the functions of KIET that are used for automatic generation of bilingual indices. After editing and proofreading was completed, the size of the Russian–English index changed from 31,666 to 29,039 headwords with 45,929 senses. In other words, at least 2,627 raw headwords were removed altogether (especially explicative definitions, due to their wordiness and a low probability of being looked up). Another part was paraphrased and shortened and some of the

headwords, which were split parts of single translation units, were combined into a single headword. While revising the headword list, we did not add many new headwords; where added, they were basically duplicated entries for the homograph headwords we discussed above.

Editing the Russian–English index was an interesting, challenging and thought-provoking task. Some of the challenges, no doubt, are language-specific and may be explained by the peculiarities and complexity of the Russian language. Major problem areas (such as part of speech tagging) were reported to the KIET technological developers and solved on the run by means of export adjustments in initial data processing. New export algorithms were added to the latest version of KIET. It would be interesting to investigate if the main challenges and problem areas discussed in this paper are also relevant to the editing of other language pairs, and to compare the results of other *Password* indices.

4. Acknowledgements

The author wishes to thank Natalia Kustovinov, of the KIET development team, for her help with the technical description of the editing tool.

5. References

Adamska-Sałaciak, A. (2013). Equivalence, synonymy, and sameness of meaning in a bilingual dictionary. *International Journal of Lexicography,* 26(3), pp. 329-345.

Anokhina, J. (2010). Lingvo Universal English-Russian Dictionary: Making a Printed Dictionary from an Electronic One. In A. Dykstra & T. Schoonheim (eds.). *Proceedings of the 14 Euralex International Congress. 6-10 July 2010. Leeuwarden/Ljouwert, the Netherlands: Fryske Akademy,* pp. 539-548

Apresjan, J. D. (1973). 'Regular Polysemy', *Linguistics* 12 (142), pp. 5-39.

Apresjan, J. D. (2003). 'Lexicographic Concept of the New English Russian Comprehensive Dictionary'. In Apresjan J.D. (ed.) *New English-Russian Comprehensive Dictionary.* Moscow: Russky yazik, v.1., pp. 6-17 [Leksikograficheskaya kontseptsiya Novogo bol'shogo anglo-russkogo slovar'ya]

Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

K Index Editing Tool (KIET) User Guide, KIET version 1.0.0.0., Copyright © 2004-2014, K Dictionaries Ltd.

Mudraya, O., Piao, S., Lofberg, L., Rayson, P., &Archer, D. (2005). English-Russian-Finnish cross-language comparison of phrasal verb translation equivalents. Accessed at: http://comp.eprints.lancs.ac.uk/1061/1/phraseology05.pdf

Yatskovich, I. (1999). Some ways of translating English phrasal verbs into Russian. *Translation Journal,* Vol.3 (3), July, 1999. Accessed at: http://translationjournal.net/journal/09russ.htm

Dictionaries:

Abbyy Lingvo-Online English-Russian Dictionary. Accessed at: http://www.lingvo-online.ru/en

Apresjan, J.D. (2003) *New English-Russian Comprehensive Dictionary.* Moscow: Russky yazik. [Novyy bol'shoy anglo-russkiy slovar']

Katzner's English-Russian, Russian - English Dictionary (1994). Rev. and expanded ed. John Wiley & Sons, Inc.

Oxford Russian Dictionary (1996). Revised and updated by C. Howlett. Oxford: Oxford University Press.

Password English Dictionary for Speakers of Russian. Accessed at: http://www.kdictionaries.com