NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Olga Vinogradova, Tatyana Pitra

STUDENT VOCABULARY EXPANSION WITH THE HELP OF A LEARNER CORPUS

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: LINGUISTICS
WP BRP 53/LNG/2017

SERIES: *LINGUISTICS*

Olga Vinogradova[1]*, Tatyana Pitra*[2]

STUDENT VOCABULARY EXPANSION WITH THE HELP OF A LEARNER CORPUS[3]

In courses of general English, and in preparation for examination in particular, students need much feedback when they submit written works. Besides tips concerning content, coherence and cohesion, grammatical range and accuracy, all of which are standardly included in instructor's comments, more efficient advice touches upon directions towards lexical improvements. However, such feedback requires a huge amount of time and effort on the part of the instructor, whose workload is heavy enough to make any extra effort undesirable. Besides, the wealth of student texts in the learner corpus allows researchers to make use of the many samples of student writing by applying certain computer tools. As a result, the authors set themselves a task of developing a system of automated lexical inspection of student works. Initially, we used essays in the corpus to work out which formal parameters in the essays demonstrate in what ways essays that have been evaluated as good by the examination experts can be distinguished, then we applied those parameters in the process of automated inspection, after which we proceeded to checking the correlation between the inspection results and the traditional grading. Finally, after a thorough analysis a system of lexical inspection of student essays was established, which paves way to the development of automated lexical feedback in order to orient students in how to improve the lexical variety in their writing.

Keywords: learner corpus; lexical proficiency; corpus research; feedback on student essay; readability tests.

JEL Classification Code: Z19

[1] National Research University Higher School of Economics, School of Linguistics, Assistant Professor (olgavinogr@gmail.com)

[2] National Research University Higher School of Economics, School of Foreign Languages, Lecturer (tpitra@hse.ru)

[3] The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2016 (grant № 16-05-0057 «Learner corpus REALEC: Lexicological observations») and by the Russian Academic Excellence Project "5-100".

More than twenty years of research in the field of learner corpora have firmly established the fact that access to a learner corpus makes the process of L2 acquisition more efficient for both learners and instructors alike (as proved by the collection [Granger, Gilquin, and Meunier, 2013] and in particular by the summary in [Granger, 2012]). This can account for the fact that EFL instructors teaching at the computer linguistics department at the Higher School of Economics set up a learner corpus of their own.

REALEC, the corpus set up at the School of Linguistics, is the first collection of English texts written by Russian students learning English easily available in the open access at (http://www.realec.org/). What makes REALEC particularly interesting among other learner corpora is, first, the fact that all errors made by students are pointed out with special tags by experts annotators (EFL instructors, as a rule), and, second, that students' essays are uploaded in the corpus as a part of the English course, and students work with categorization of the errors outlined by their instructors. Moreover, the annotations in these essays are monitored by the research team consisting of EFL specialists and computer linguists (the name of the team is "REALEC for Real Words" - https://realec-nug.wikispaces.com/). Researchers in the project team are responsible for suggestions, changes and directions throughout the development of REALEC. The authors of this paper are HSE instructors and at the same time members of this research group, and as such we were in a perfect position to look at the lexical range and some other lexical features in the best essays of the past examination (pairs of essays in IELTS format[4] graded higher than others by the professional examiners) and compare them with the average corresponding features of the average essays among the examination essays. In this, the research undertaken is similar to the one reported in [Cobb & Horst, 2015].

Methods and statistics

The aim of the experiment was to pave the way towards automated evaluation of lexical content of Bachelor students' essays written in IELTS format. After the examination, the essays were evaluated by experts and assigned a grade for both tasks in the percentage points. For the purposes of the experiments, the essays were divided into two groups – those that the experts graded at 75% or higher, and the rest of the essays. Essays in either group were subjected to the procedure of automatic evaluation of the lexical resource, and this procedure was called lexical inspection. The results of this inspection in two groups were compared with each other. Such criteria as task response, coherence and cohesion, grammatical range and accuracy were not analyzed in this experiment. The choice of the parameters to be included in evaluation is discussed, for example, in [Lavallée & McDonough, 2015]. The adjacent filed - comparisons of student texts with authentic academic texts - were reported by Canadian researchers from University of Grenoble-II Benoît Lemaire and Philippe Dessus in their work which presents Apex, a system for automatic assessment of a student essay based on the use of Latent Semantic

[4] IELTS (International English Language Testing System) is a test of English language proficiency for non-native speakers of English. IELTS certificates are recognized in more than 120 countries round the world. HSE Bachelor students at the end of their 2nd year have been taking an examination similar to Academic IELTS in its format for the last eight years. It covers all four language skills - listening, reading, writing and speaking. The writing part of this examination includes two tasks each requiring that a testee write an essay – one about 150 words long, the other about 250 words long, both within the period of one hour. Both essays are evaluated by the following criteria: task response, coherence and cohesion, lexical resource, grammatical range and accuracy.

Analysis ([Dessus & Lemaire, 2001]). Both reported experiments presented the results of readability of the texts, explored in detail in the HSE Master's thesis of Konstantin Druzhkin ([Druzhkin, 2015] and [Druzhkin, 2016]).

The objective of our experiment was to establish the correlation between the marks that were given by experts and the automated evaluation of lexical content on the basis of certain criteria.

Our hypothesis was that the criteria applied in the developed application would be sufficient for a valid preliminary evaluation of lexical variability of written papers. For the purposes of our experiment the comparisons were drawn across the following features:

1. Length of words

2. Length of sentences (cf. [McCarthy & Jarvis, 2010] about the choice of criteria 1 and 2)

3. Distribution of words across the Common European Framework scale levels (A1-C2)[5]

4. Frequency of each word in the Corpus of Contemporary American English [6] (for the justification of this parameter in lexical evaluation cf. [Crossley, Cobb, & McNamara, 2013] and [Vongpumivitch, Huang, & Chang, 2009])

5. Use of academic vocabulary from the two lists - the Coxhead Academic Word List in [Coxhead, 2000] and [Coxhead, 2011]) and in the Corpus of Contemporary American English[7]

6. Repetitions

7. Use of linking words

8. Use of collocations (as attested by the presence on the Pearson academic collocation list)[8]

The works in the experiment consisted, on the one hand, of 45 sets (2 essays in each set – an argumentative essay and a description of a diagram or diagrams), - those that were marked by experts at 75% (out of 100%) and higher, and in the second group 900 sets (the same types of essays) marked below 75%. The results of lexical inspection application in the two groups were analysed in comparison. It was revealed that certain characteristics (for example, the average sentence length, the number of words from academic vocabulary lists) are significantly different in the two groups, which proves that grades assigned by experts do have certain correlations with parameters that can be evaluated by a software application. For one, there are more words on average in "good" texts.

The following are the main numerical results of the comparative analysis:

---

[5] CEFR – policies: http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf; CEFR lists - British Council Word Family Framework of General English (the lexical lists) https://www.learnenglish.org.uk/wff/index.html; CEFR explained - http://www.englishprofile.org/

6 COCA (The Corpus of Contemporary American) - http://corpus.byu.edu/coca/ and http://www.wordfrequency.info/ and http://www.academicvocabulary.info/

7 http://www.academicvocabulary.info/

8 http://pearsonpte.com/research/academic-collocation-list/

There are 188 words on average in a diagram description in a “good” essay versus exactly the same number of words in a diagram description from the big collection of essays.
There are 294 words on average in a “good” argumentative essay versus 268 words on average in an argumentative essay from the big collection.
On average, sentences are longer in “good” essays: there are 21.15 words in a “good” diagram description versus 18.4 words in a diagram description from the entire collection, and 20.34 words in a “good” argumentative essay versus 18.1 words in an argumentative essay from the entire collection.
The maximum sentence length is longer in “good” texts: it is 37.5 words in a “good” diagram description versus 33.14 words in a diagram description from the entire collection, and 37.4 words in a “good” argumentative essay versus 35.75 words in an argumentative essay from the big collection.
At the same time, the average word length is approximately the same for both “good” and “ordinary” essays, both argumentative and diagram descriptions. It means that the word length is not a factor to consider when giving feedback to students because it is not among the significant characteristics.
The same stands true for the longest words in the papers and the number of word repetitions. Both “good” and “ordinary” student texts have approximately the same number of word repetitions.
The number of linking words. “Good” diagram descriptions have more linking words, though not overwhelmingly so - 3.6 versus 3.23. However, “good” argumentative essays demonstrate a significant difference: 8.97 versus 6.33.
The number of collocations from Pearson’s list (with repetitions) for diagram descriptions: 1.35 in “good” texts versus 0.4 in “ordinary” descriptions, and 1.62 in “good” argumentative essays versus 0.71 in “ordinary” argumentative essays.
The number of collocations without repetitions: for diagram descriptions - 0.88 in “good” ones versus 0.71 in “ordinary” descriptions, and for argumentative essays - 1.46 in “good” versus 0.67 in “ordinary”.
In general, “good” essays have more CEFR scale words at each level, but not much more. This is rather due to the fact that the good papers have more words altogether. So, these figures will not be given here.
The same stands true for COCA frequencies. The “good” essays on the whole have more words at each level.
At the same time, there are notably more words from academic vocabulary lists in “good” essays: for diagram descriptions it is 43 pieces of academic vocabulary on average versus 36 (with repetitions), and 28 versus 22 (without repetitions), and for argumentative essays, the average number of academic vocabulary items is 70 versus 56 (with repetitions), and 51 versus 40 (without repetitions).

Fig. 1 below gives the synopsis of the significant differences between essays scored highly and the rest of the essays.

Parameters for automated lexical inspection	Essays scored as high as 75% and higher		Essays scored lower than 75% by experts	
	Task 1	Task 2	Task 1	Task 2
1) Number of words in the essay	188	294	188	268
2) Average length of a sentence in the essay	21.15	20.34	18.4	18.1

3) Length of the longest sentence in the essay	37.5	37.4	33.14	35.75
4) Number of academic words in the essay (with repetitions/without repetitions)	43/36	70/56	28/22	51/40
5) Number of linking words and expressions in the essay	3.6	8.97	3.23	6.33
6) Number of collocations from the Pearson academic collocation list in the essay (with repetitions/without repetitions)	1.35/0.88	1.62/1.46	0.4/0.71	0.71/0.67

Fig. 1 Parameters of significant difference between "good" essays and the rest of the essays

Development of the application

The results of the comparisons between "good" and average essays have allowed us to set up an automated application called REALECInspector[9], which carries out observations over an essay (argumentative or description of diagrams) uploaded to the corpus and provides some statistical information based on the comparison of its formal features with the average figures for an essay of this type collected in REALEC, as well as offers some recommendations for improvements.

The stages of work with this application are the following.

There is an input window on its homepage with the "inspect" button to open the page for the lexical analysis.

[9] The pilot version of the script for the application was adopted by Ilya Golubev, Master's programme student at the NRU HSE, from a similar application for Russian texts and can be seen at http://realec-inspector.appspot.com/

Fig.2 Window for lexical inspection

The first thing on the page that appears after pressing **Inspect** button is the essay itself. Then comes the short statistics on:

1) Number of words in the essay

2) Average length of a sentence in the essay

3) Length of the longest sentence in the essay

4) Average length of word in the essay

5) Length of the longest word in the essay

6) Number of words of each level of CEFR in the essay

7) Number of words from the COCA frequency lists

8) Number of academic words in the essay with repetitions and without them

9) Number of repetitions of words used in the essay. The word most frequently repeated.

10) Number of linking words and expressions in the essay

11) Number of collocations from the Pearson academic collocation list in the essay

Statistical summary

Number of words: 290
Average sentence length: 18.875 words.
Max sentence length: 32 words.
Average word length: 5.10104529617 letters.
Max word length: 18 letters.
CEFR
A1: 49
A2: 16
B1: 11
B2: 7
C1: 1
C2: 0
Unclassified: 38
Stopwords: 36
Frequency:
1-500: 39
501-3000: 36
>3000: 47
Academic words: 71 (51 unique)
Word repetitions: 44 (('children', 6) is the most repeated)
Linking phrases: 12
Pearsons collocations: 7 (5 unique)

Fig 3. List of statistics for the essay under inspection

After this list of short statistics, each figure gets detailed comments and the necessary diagrams. For the histogram of CEFR words distribution (Fig. 4), Word Family Framework was used (the possibility to use English Vocabulary Profile instead has been reserved), and each word is lemmatized with the help of NLTK. Stopwords (153 on the list) are excluded. Words that the system was unable to relate with a particular CEFR level are categorized as "Unclassified" (some misspelled words are among them).

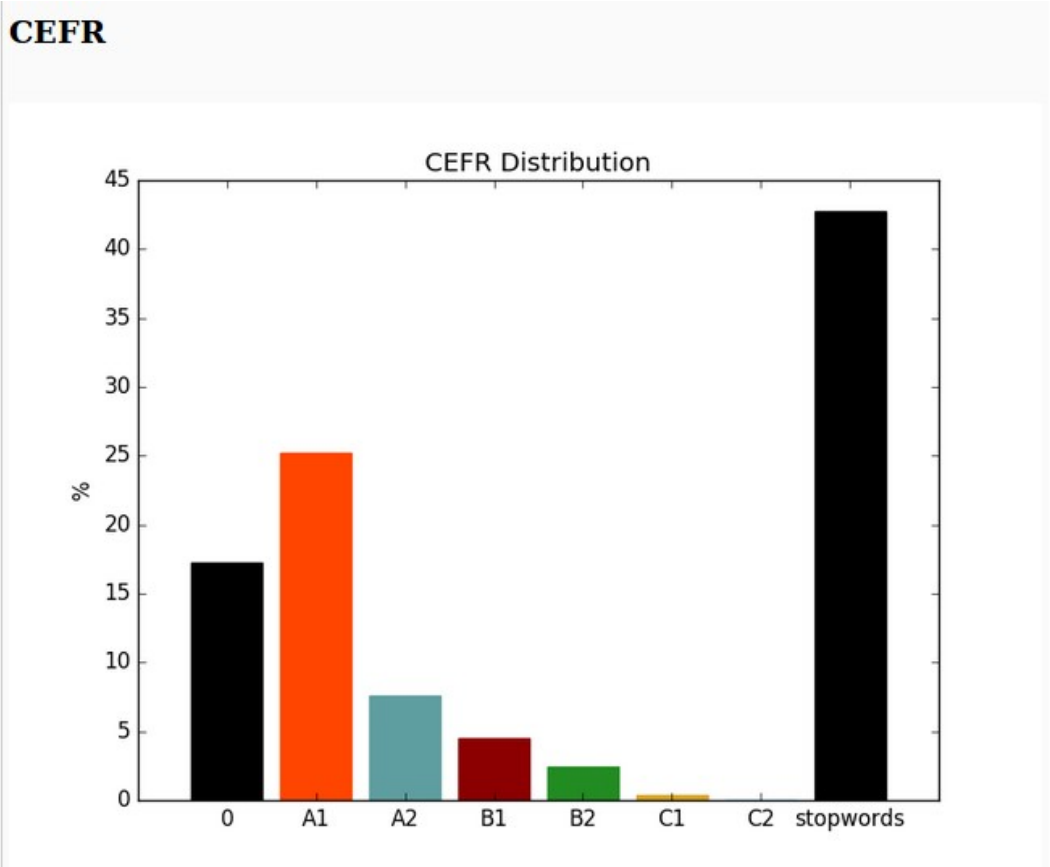

Fig 4 Sample picture of distribution of CERF-level words

The author of the text then gets the list of words from the essay that are among the 500 most frequent words in COCA, and then those that are among the 3000 most frequent words in COCA. Stop-words are again excluded.

The next comment is on the occurrence of academic words from the list which is a combination of two - the Academic Word List Coxhead and the Corpus of Contemporary American English. As a result, if a word belongs to either of these lists, it will be considered academic.

In the next section the author will see two diagrams. The first is the distribution of average sentence length in the corpus (Fig. 5), and the second, the distribution of average word length in the corpus. The red line on both diagrams marks the average index in the essay under inspection for the author to compare with other essays. The comparison can also be numerical, as percentage is given here as well. For example, if under the diagram with the average sentence length there is a figure of 90%, it means the average sentence length is longer than in 905 of all essays in the corpus. More often than not, it is a feature of a good essay, as it implies that sentences are more sophisticated than in the majority of essays. On the contrary, low percentage number is the result of oversimplified sentence structure.

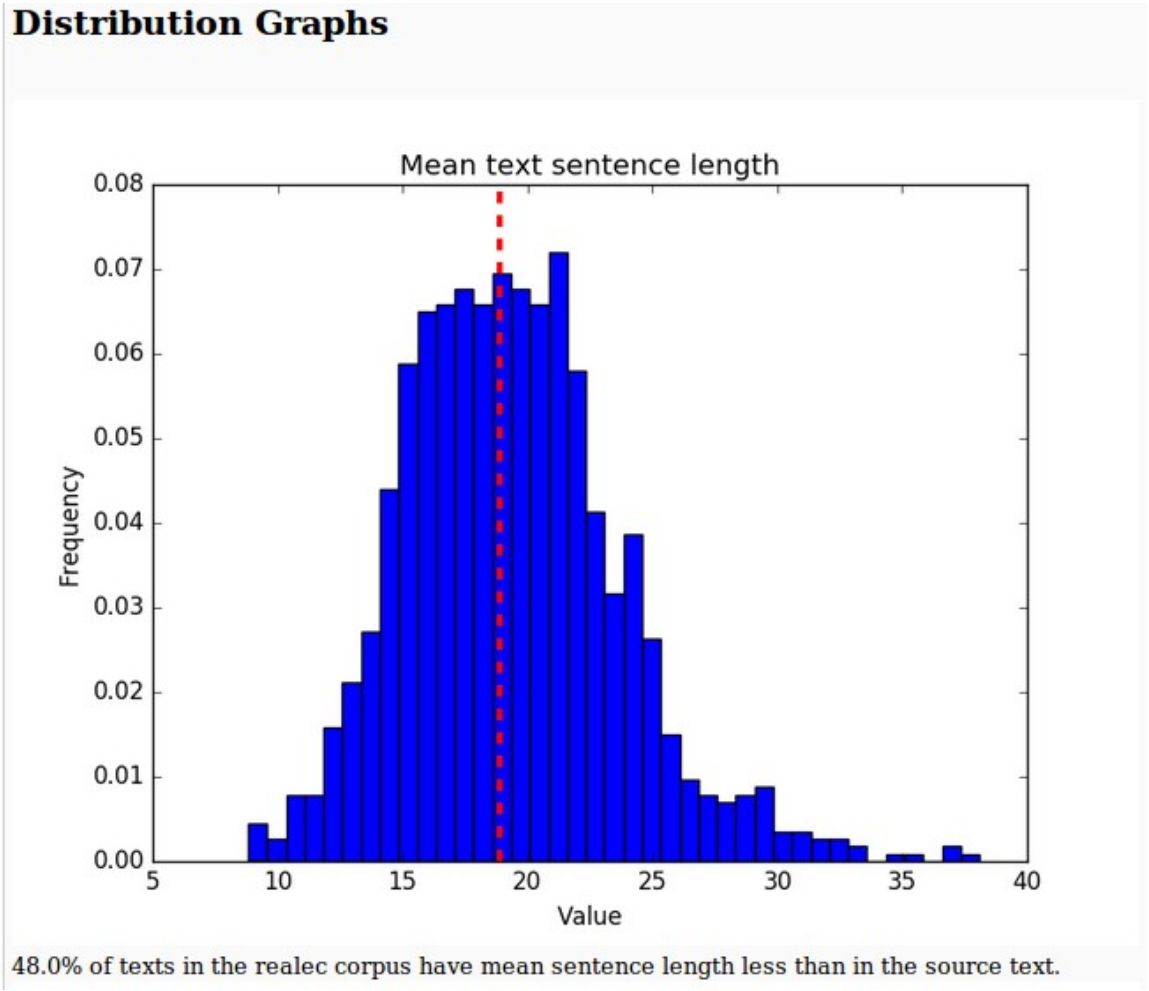

Fig 5 Sample distribution of sentence length on the background of its average index

Five most frequently repeated words are shown next (Fig. 6) (no stop-words). The need for demonstrating the ability to paraphrase can be emphasized here.

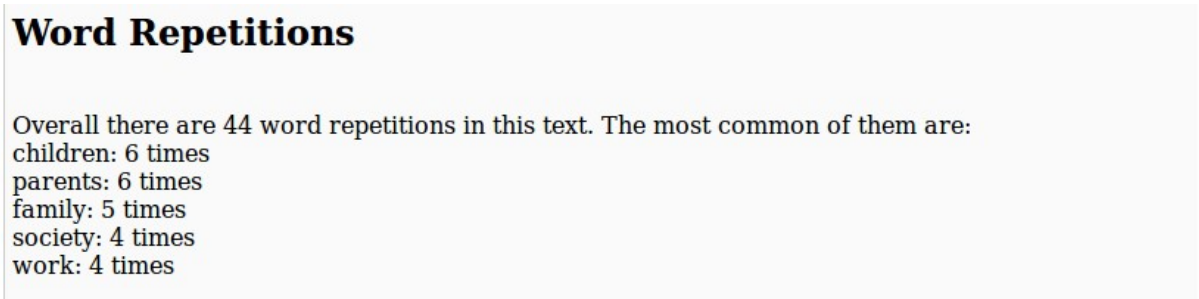

Fig. 6 Sample list of repetitions in the essay

The number and the list of linking words and introductory expressions used in the essay are accompanied next by the indication of their categories (Comparison, Time and sequence, Addition, Cause and Effect, Conclusion and summary, Examples, Concession, Repetition, Giving reasons, explanations, Contrast (Fig. 7).

Linking Phrases

There are 12 introductory phrases.
Comparison: 0
Time and sequence: 5
then: 2
now: 2
nowadays: 1
Addition: 4
also: 3
moreover: 1
Cause and Effect: 0
Conclusion and summary: 1
in conclusion: 1
Examples: 1
for example: 1
Concession: 0
Repetition: 0
Giving reasons, explanations: 0
Contrast: 1
however: 1

Fig. 7 Sample list of linking words in the essay

The comparison of the use of linking phrases in the essays under inspection with all other essays in the corpus can be presented to the author as is shown in Figure 8.

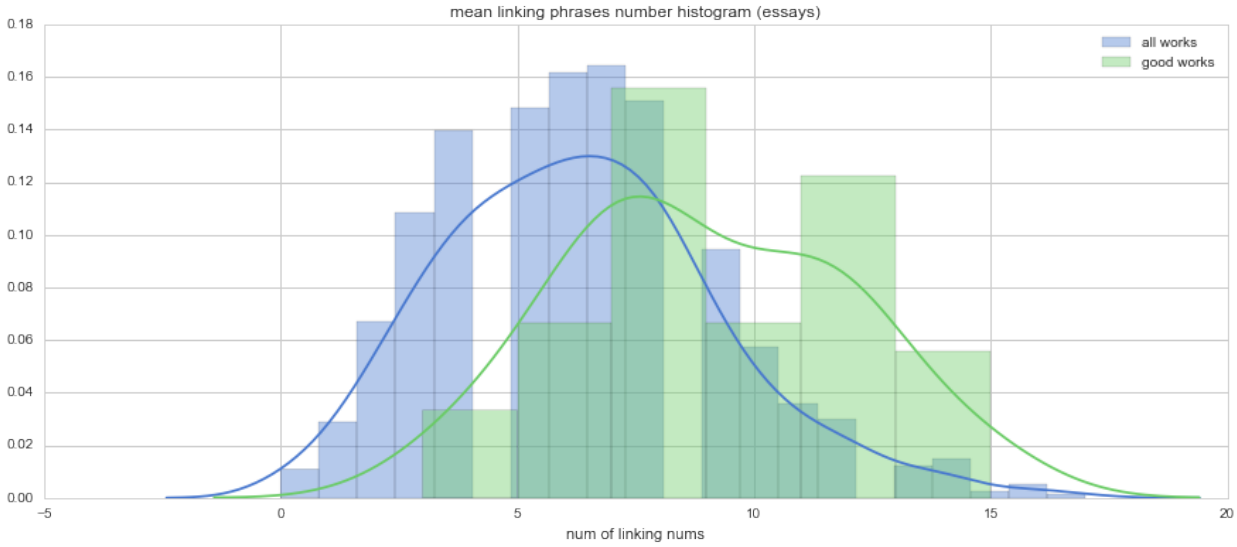

Fig. 8 Sample distribution of the number of linking words for "good" essays and all essays in the corpus

The inspector then gives the number and the list of collocations from the essay if they are on the Pearson Academic Collocation List (Fig. 9).

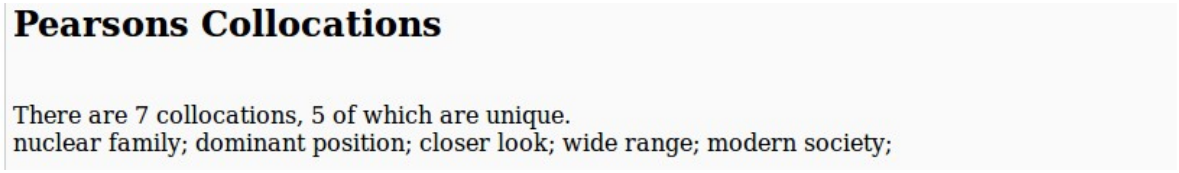

Fig. 9 Sample list of collocation in the essay

Below these the author gets the text of the essay three times with the following different visual features:

1) with words of different CEFR levels presented in different colours

2) with words of different COCA frequencies presented in different colours

3) with academic words highlighted.

As a result, the author gets recommendations on the basis of the parameters listed above.

Conclusions

The observations over the features of many student essays in the learner corpus have confirmed the following conclusions important for working out approaches to automated evaluation of student writing:

1. Overall, sentence length and numbers of linking words, collocations, and academic words are larger in essays highly evaluated by experts.

2. Word length and number of repetitions are insignificant as indicators of the writing proficiency.

3. The numbers of words at each CEFR level and of those with high COCA frequency are greater in essays highly evaluated by experts.

4. All parameters in automated inspection except word length are valid in distinguishing essays scored highly by experts, so the application can work as the preliminary stage in evaluating writing proficiency, but not instead of expert evaluation. Nevertheless, REALECInspector makes up a good suggestion for students' independent work on how to expand their writing potential.

With independent training in mind, we are thinking of introducing a few more computer tools that will be of use in the process of writing an essay in the corpus, which will increase the convenience of writing in the corpus in future – time management system, instant demonstration of low-profile features like superfluous repetitions, misspelled wirds, etc.

References

CEFR, 2001 - The Common European Framework of Reference for Languages: Learning, Teaching, Assessment https://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Cobb & Horst, 2015 - Cobb, T., & Horst, M. "Learner corpora and lexis." In S. Granger, G. Gilquin, & F. Meunier (eds.) The Cambridge Handbook of Learner Corpus Research, Cambridge, UK: Cambridge University Press. Pp. 185–206.

Coxhead, 2000 – Coxhead, A. "A new academic word list." TESOL Quarterly, 34, no 2, pp. 213–238

Coxhead, 2011. – Coxhead, A. "The academic word list 10 years on: Research and teaching implications." TESOL Quarterly, 45, no 2, pp. 355–362.

Crossley, Cobb, & McNamara, 2013 - Crossley, S. A., Cobb, T., & McNamara, D. S. "Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications." System, 41, no 4, pp. 965–981.

Druzhkin, 2015 – Дружкин, К. Ю. "Readability: онлайн-сервис." Online resource at http://web-corpora.net/wsgi3/readability/index.

Druzhkin, 2016 – Дружкин, К. Ю. Метрики удобочитаемости для русского языка. Master's Thesis, NRU HSE, Moscow, 2016. Online publication at: https://www.hse.ru/edu/vkr/184791276.

Granger, 2012 – Granger, S. "How to use Foreign and Second Language Learner Corpora." In A. Mackey and S. M. Gass (eds), Research Methods in Second Language Acquisition: A Practical Guide Blackwell, Oxford (2012). Ch.2, pp. 5-29

Granger & Meunier, 2013 - Granger, S., Gilquin, G. and Meunier, F. (Eds.) "Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead:" Proceedings of the First Learner Corpus Research Conference. Vol. 1. Presses universitaires de Louvain, 2013.

González-López & López-López, 2015 - González-López, S., López-López, A "Lexical analysis of student research drafts in computing"- in Computer Applications in Engineering Education 23, no 4, 2015, pp. 638–644.

Lavallée & McDonough, 2015 - Lavallée, Maxime & McDonough, Kim "Comparing the Lexical Features of EAP Students' Essays by Prompt and Rating" TESL Canada Journal, 2015, Vol 32, no 2, pp. 30-44

McCarthy & Jarvis, 2010 - McCarthy, P. M., & Jarvis, S. "MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment." In Behavior Research Methods, 42, no 2, pp. 381–392.

Vongpumivitch, Huang, & Chang, 2009 - Vongpumivitch, V., Huang, J.-Y., & Chang, Y.-C. "Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. " In English for Specific Purposes, 28 no 1, pp. 33–41.

Olga I. Vinogradova
National Research University Higher School of Economics (Moscow, Russia). School of Linguistics. Associate Professor;
E-mail: ovinogradova@hse.ru, olgavinogr@gmail.com, Tel. +7 (916) 6385362

Tatiana G. Pitra

National Research University Higher School of Economics (Moscow, Russia). School of Foreign Languages, Lecturer;

E-mail: tpitra@hse.ru, Tel. +7(915) 0305713