INITIAL STEP OF SPECIALIZED CORPORA BUILDING: CLEANING PROCEDURES

Chief of the Center Vera Yakubson[1]
Assoc. Prof. Dr. Victor Zakharov[2]
[1]Peter the Great St. Petersburg Polytechnic University, Russia
[2]St. Petersburg State University, Russia

ABSTRACT

This paper deals with the specialized corpora building, specifically academic language corpus in the biotechnology field. Being a part of larger research devoted to creation and usage of specialized parallel corpus, this piece aims to analyze the initial step of corpus building. Our main research question was what procedures we need to implement to the texts before using them to develop the corpus.

Analysis of previous research showed the significant quantity of papers devoted to corpora creation, including academic specialized corpora. Different sides of the process were analyzed in these researches, including the types of texts used, the principles of crawling, the recommended length of texts etc. As to the text processing for the needs of corpora creation, only the linguistic annotation issues were examined earlier. At the same time, the preliminary cleaning of texts before their usage in corpora may have significant influence on the corpus quality and its utility for the linguistic research.

In this paper, we considered three small corpora derived from the same set of academic texts in the biotechnology field: "raw" corpus without any preliminary cleaning and two corpora with different level of cleaning. Using different Sketch Engine tools, we analyzed these corpora from the position of their future users, predominantly as sources for academic wordlists and specialized multi-word units.

The conducted research showed very little difference between two cleaned corpora, meaning that only basic cleaning procedures such as removal of reference lists are can be useful in corpora design. At the same time, we found a significant difference between raw and cleaned corpora and argue that this difference can affect the quality of wordlists and multi-word terms extraction, therefore these cleaning procedures are meaningful.

The main limitation of the study is that all texts were taken from the unique source, so the conclusions could be affected by this specific journal's peculiarities. Therefore, the future work should be the verification of results on different text collections.

Keywords: *specialized corpora, text cleaning, academic wordlist, term extraction, corpus building*

INTRODUCTION

Specialized corpora are widely used in modern linguistics. The main advantage of using such corpora is linguistic material consistent with specific language register. Specialized corpora is very valuable in linguistic research but can also find application in information retrieval, natural language processing, and ontology building [1].

One of the principal applications of corpora has been the compilation of vocabulary lists for English language learning. The wordlists derived from corpora have been used both in general English and in academic English learning, but usually they were not disciplinary specialized [2]. To date, there are several examples of compiling academic wordlists in specialized fields [3], [4], [5], [6]; nevertheless, all of them concern individual words, not multi-word units.

To date, researchers can easily compile comprehensively large corpus using up-to-date software and "web as a corpora" technologies. There are several publications discussing the preparation stage of the corpus design [7], but to the best of authors' knowledge, the language cleaning procedures in this initial stage has not been a concern for researchers.

The corpus quality is an ambiguous concept depending on the wide range of criteria. The basic properties of every corpus are representativeness, balance, homogeneity, coverage (size). These characteristics were actively discussed in the literature [8], [9], [10], but there is no widely used requirements for compiling specialized corpora, as we know.

In this paper we aim to look closely at the initial step of corpora building and answer the question, what cleaning procedures need to be implemented to guarantee the corpus quality. We strive to find the balance between the desired clarity of data and the sensible time spared to the text preparation for corpus building. In other words, how much time it is rational to spend in preparing texts for a corpus?

To understand this, we analyzed a real case of specialized corpus building and implemented different preparation procedures to find out which cleaning steps are valuable and which take too much time or do not affect the quality of the corpus.

MATERIALS AND METHODS

This research is based on the corpora in the biomedical field derived from the Russian scientific journal Acta Naturae. This work is a part of a larger research

devoted to the construction of parallel English-Russian biomedical corpus. The journal was chosen based on this goal, because this is a bilingual journal with two versions in Russian and English.

For the research of cleaning procedures we used only the Russian version of the journal and took a small massive of papers that constitutes the 2015 year volume of the journal, 58 papers.

Based on this massive, we compiled three corpora in Sketch Engine with different degree of files preparation :

- "AN_pdf" is a raw corpus derived from pdf files without any preliminary cleaning;
- "AN_clean" is a cleaned corpus with minimum preparation of files including elimination of reference lists, tables, author information and hyphens;
- "AN_very_clean" is a cleaned corpus with more sophisticated preparation procedures – in addition to the above mentioned the running titles were eliminated, as well as breaks in words and sentences. These breaks appear when an image or a table is embedded in the text.

The table 1 shows the principal features of compiled corpora. The substantive variation in number of tokens and sentences between raw corpus and cleaned corpora can be explained primarily by elimination of paper elements such as reference lists and author information.

Reference lists and author information were removed from AN_clean and AN_very_clean because these parts of an article include a large number of proper nouns and numeric symbols such as page numbers, addresses and so on. The tables were excluded because in the text format they mispresent the co-occurrence range of words which follow each other but do not have an association.

Table 1. Principal features of the corpora

	Number of tokens	**Number of words**	**Number of sentences**
AN_pdf	400,518	255,275	20,438
AN_clean	272,988	198,247	9,765
AN_very_clean	275,714	199,916	9,580

Table 2 presents time expenditures for preparation of two cleaned corpora.

Table 2. Time spared for preparation of AN_clean u AN_very_clean corpora

	Procedures	Total time for corpus cleaning	Average time spared for 1 article
AN_clean	Removal of hyphens, tables, reference lists, author information	87 min	1,5 min
AN_very_clean	Removal of running titles, breaks in words and sentences	232 min	4,0 min

To evaluate the usefulness of preliminary cleaning we need to compare the results obtained on three corpora. We chose several parameters that are important for specialized corpora and analyzed them using Sketch Engine:

1. The lists of the most frequent words using the "wordlist" tool.
2. Keywords and multi-word units.
3. Comparison of corpora using "compare corpora" tool.

RESULTS

Wordlist Analysis

For the first stage, we analyzed the first 100 words by frequency in three corpora using the tool "wordlist" in Sketch Engine. The wordlists were very similar in case of corpora AN_clean and AN_very_clean: there was slight difference in frequencies of meaningful words, but it was not sufficient enough to affect its rank in the lists. At the same time, wordlist of AN_pdf appeared very different from other two: it includes the same meaningful words, but their rank is much lower because of the noise included. For example, the word "белок" (protein) has the highest frequency in the first corpus, but the lower rank (17 instead of 13) than it has in cleaned corpora.

To understand the nature of additional "words" in the AN_pdf corpus, we looked at the first three of them that did not appear in other two corpora: p., v. and j. The concordance analysis for lemma "p." showed that only 46 entries from 2832 were in the article texts being a part of bacteria name (e.g. "P. aeruginosa"). All others were from reference lists being a reduction for "page" or authors initials. For "v." and "j." there were no entries from article texts, only from reference lists as a reduction for "volume" and "journal" accordingly or as authors initials.

Based on this short analysis, we can summarize that additional words in AN_pdf corpus are in fact noisy and do not add any value for potential users of the corpus.

Another issue could be the difference in word frequencies. We can see that for meaningful words the frequencies vary in all three corpora. Nevertheless, it seems that this variation is not significant. To test this hypothesis, we calculated the variation coefficient for first five nouns by rank in the AN_pdf corpus using the following formula: $V = \frac{\sigma}{\acute{x}} * 100\%$, where σ– standard deviation, $\acute{x}$ – mean value for frequencies of each word in three corpora. The results were the following:

- “клетка” (cell) V=0,9 %;
- “белок” (protein) V=0,8 %;
- “экспрессия” (expression) V=0,7%;
- “активность” (activity) V=1,6 %;
- “результат” (result) V=0, 1 %.

Therefore, our hypothesis was justified: the frequencies of meaningful words vary insignificantly in all three corpora.

Nevertheless, the difference in ranks looks significant. For example, considering the first 100 words by rank we can find several non-words – predominantly Latin symbols: in the AN_pdf corpus the quantity of such elements are 27 per 100, and in the AN_clean and AN_very_clean 6 and 7 respectively. These non-words do not have value for linguistic analysis and can be considered as noise.

Keyword and Term Extraction

The examined corpora are intended to serve as a source for extracting keywords and terms. To evaluate their ability to play this role, we used the Sketch Engine tool “Keywords”. This tool extracts both keywords and multi-word expressions from a focus corpus using comparison of its frequencies with a reference corpus. For Russian language the Russian Web 2011 corpus is used as a reference corpus. It is Russian general language web corpus crawled by SpiderLing in 2011, cleaned, deduplicated, consists of 18,280 million tokens [11].

Table 3. The first 10 single-word terms obtained by the Keywords tool for three corpora, ranked by score

No.	AN_pdf	AN_clean	AN_very_clean
1.	NATURAE	экспрессия	экспрессия
2.	ACTA	мрнк	мрнк
3.	ЭКСПЕРИМЕНТАЛЬНЫЕ	PARP	PARP
4.	Biol	ррнк	ррнк
5.	экспрессия	pestis	NeuN
6.	mol	NeuN	транскрипция
7.	PARP	транскрипция	her
8.	мрнк	her	Ala
9.	Chem	промотора	рчБуХЭ
10.	Cell	ИПСК	промотора

From the table 3 we can see, as in the wordlist case, that AN_clean and AN_very_clean are very similar by keywords, at the same time AN_pdf list demonstrates a big difference. It includes the title of the journal “NATURAE”, “ACTA”, standard section names “ЭКСПЕРИМЕНТАЛЬНЫЕ” (experimental), reductions of journal titles from reference lists “Biol”, “Chem”, “Cell”.

Besides single keywords, extraction of multi-word terms is very important function for users of specialized corpora. We can see in the table 4 an example of this extraction for our three corpora. Here we can observe that there is no clear difference between cleaned and raw corpora: the lists are almost the same except for “список литературы” in AN_pdf which means “references”. For the first 100 multi-word terms the difference is still insignificant: in AN_pdf there are only 5 occurrences that are not present in cleaned corpora – all of them are the universities’ names given in author information.

Table 4. The first 11 multi-word terms obtained by the Keywords tool for three corpora, ranked by score

No.	AN_pdf	AN_clean	AN_very_clean
1.	уровень экспрессии	уровень экспрессии	уровень экспрессии
2.	дикий тип	дикий тип	дикий тип
3.	экспрессия генов	экспрессия генов	аминокислотный остаток
4.	аминокислотный остаток	молочная железа	экспрессия генов
5.	молочная железа	культура клеток	молочная железа
6.	культура клеток	настоящее время	культура клеток
7.	настоящее время	аминокислотный остаток	настоящее время
8.	фактор роста	фактор роста	фактор роста
9.	продолжительность жизни	продолжительность жизни	продолжительность жизни
10.	сайт связывания	сайт связывания	сайт связывания
11.	список литературы	экспрессия гена	экспрессия гена

Corpora Comparison

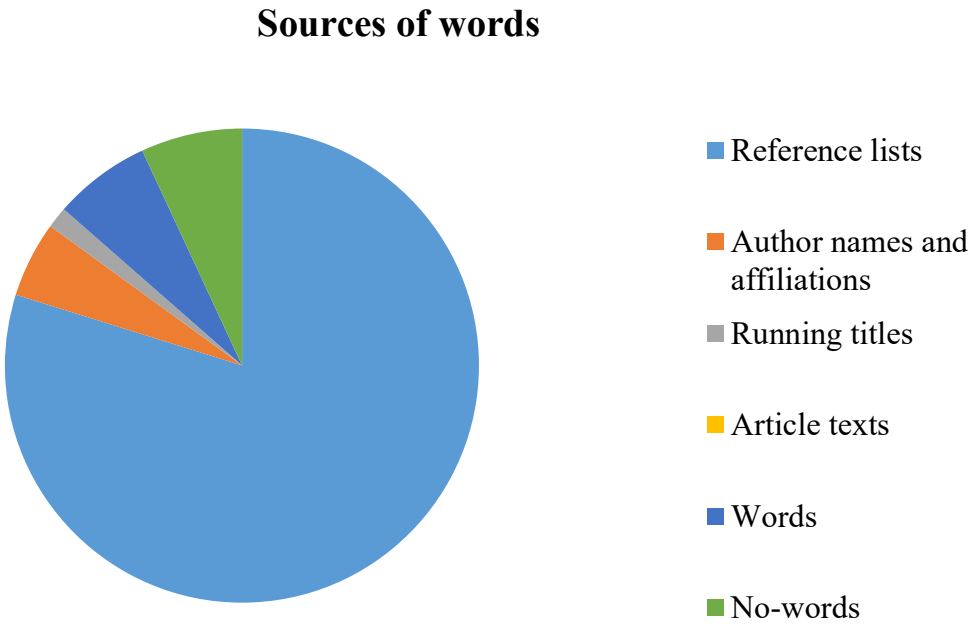
In previous chapters we looked only at first results, i.e. at the high frequency words and collocations in corpora. Nevertheless, less frequent words can also be important for the research results. To evaluate the differences between corpora in quantitative manner, we used the “compare corpora” tool in Sketch Engine [12], [13]. We received a predictably big value for pairs “AN_pdf vs. AN_clean” and “AN_pdf vs. AN_very_clean” (1.74 and 1.73 respectively) and smaller value for pair “AN_clean vs. AN_very_clean” (1.18).

To understand the nature of these differences, we firstly analyzed the wordlist made by comparison of corpora AN_pdf and AN_clean (pic.1). It consisted of 358 words, the majority of them had null frequency in corpus AN_clean. The

percentage of different types appeared in this list is shown on picture 1. We can see form here that reference list is the most significant source of these words, 77 % – these are reductions from journal titles and popular surnames and initials, there are no meaningful words perceived from this source.

Secondly, there are some occurrences from authors' information (surnames, addresses, 5%) and from running titles (repeated titles of sections, 1%).

And lastly, we received 13% of this wordlist from actual article texts, where we performed a deeper analysis. From these 47 occurrences 24 were elements of formulae, chemical structures etc. that are not relevant for linguistic analysis. Others 23 (6,5%) were meaningful words extracted predominantly from tables and pictures that were removed in AN_clean corpus.



Pic. 1. Percentage of words appeared in the comparison list AN_pdf and AN_clean, ranked by sources

We should mention here that some of these words in fact presented in the second corpus. Due to the methodology applied in corpora comparison in Sketch Engine, only first 5000 words are included in the analysis, and then the algorithm chooses only top 500 words according to score in each corpus separately. This may be the reason of the situation with the word "minzdrav" that have 2 occurrences in corpus AN_clean but in comparison list there are 0 occurrences from this corpus.

DISCUSSION

Analysis of wordlists and single-word keywords showed very small difference between AN_clean and AN_very_clean that, to our understanding, cannot affect the research results of corpus users. At the same time, AN_pdf demonstrated significant difference in both cases. For example, the differences in ranks of meaningful words seems important for us as it can influence the usability of corpus as a source of wordlists.

The majority of additional words and keywords appeared in the raw corpus were extracted from reference lists and had no value for text analysis. One could argue that reference list analysis could be useful to specify the most valuable authors for the field, but there are specific bibliographic tools for it, like Scopus and Web of Science, and there is no need to use corpora here. Based on this, we can assume that removal of reference lists from scientific texts could be valuable procedure as it decreases the noise significantly.

Other cleaning procedures performed in AN_clean and AN_very_clean corpora seem to not have big influence on their performance, at least on the most frequent words and keywords. At the same time, corpora comparison showed some influence of tables removal – it leaded to disappearance of some meaningful words that presented in the raw corpus.

Automatic extraction of multi-word terms showed very little difference between all three corpora.

CONCLUSIONS

1. The sophisticated software for corpora building and annotation such as Sketch Engine let researchers to create corpora without any manual preparation of texts.

2. Any preliminary cleaning of texts aiming to reduce the noise in the results could be sensible only if they are not time-consuming.

3. Based on the results of this research, we consider as valuable procedures only removal of reference lists and author information.

4. The removal of tables do not affect the multi-word term extraction significantly, but it removes some meaningful words from the wordlists. So this procedure cannot be recommended.

5. More complex procedures, such as removal of running titles, breaks in words and sentences, take too much time to perform; at the same time, they do not have significant effect on the corpus performance.

In this research we only considered Russian language corpus in biotechnology area. Furthermore, we used the unique source of texts – the Acta Naturae journal. These factors can obviously affect the results. For further research, we plan to enlarge our analysis using other sources and including English language.

ACKNOWLEDGEMENTS

The authors acknowledge Victor Glukhov and the National Digital Library (https://www.elibrary.ru/) for providing full text articles for the corpus building.

This work was supported by the Russian Foundation for Basic Research [grant No. 17-29-09159 "Quantitative grammar of Russian prepositional constructions"].

REFERENCES

[1] Mehdi M., Okoli C., Mesgari M., Nielsen F.Å., Lanamäki A., Excavating the mother lode of human-generated text: A systematic review of research that uses the wikipedia corpus, Inf Process Manag, 2016, vol. 53(2), pp 505–29;

[2] Smith S., DIY corpora for Accounting & Finance vocabulary learning, English Specif Purp, 2020, Jan 1;57, pp 1–12;

[3] Liu J., Han L., A corpus-based environmental academic word list building and its validity test, English Specif Purp, 2015, Jul 1;39, pp 1–11;

[4] Lei L., Liu D., A new medical academic word list: A corpus-based study with enhanced methodology, J English Acad Purp, 2016, Jun 1;22, pp 42–53;

[5] Yang M.N., A nursing academic word list, English Specif Purp, 2015, vol. 37(1), pp 27–38;

[6] Gilmore A., Millar N., The language of civil engineering research articles: A corpus-based approach, English Specif Purp, 2018, Jul 1;51, pp 1–17;

[7] Wynne M., Developing Linguistic Corpora: a Guide to Good Practice, 2005;

[8] Ramos F.P., Cerutti G., Guzmán D., Building representative multi-genre corpora for legal and institutional translation research The LETRINT approach to text categorization and stratified sampling, Transl Spaces(Netherland), 2019, vol. 8(1), pp 93–116;

[9] Biber D., Representativeness in corpus design, Lit Linguist Comput, 1993, 8(4), pp 243-257;

[10] McEnery T., Hardie A. Corpus linguistics: Method, theory and practice, 2011, pp 1–294;

[11] ruTenTen: Corpus of the Russian Web [Internet], Sketch Engine, Available from: https://www.Sketch Engine.eu/rutenten-russian-corpus/;

[12] Compare corpora [Internet], Sketch Engine, Available from: https://www.Sketch Engine.eu/guide/compare-corpora/;

[13] Kilgarriff A., Comparing Corpora, Int J Corpus Linguist, 2001, vol. 6(1), pp 97–133.