DOI 10.20535/2617-5339.2019.4.189753

UDC 81.11’255.2:62

Valeriia Havrylenko
Lecturer
National Technical University of Ukraine
“Igor Sikorsky Kyiv Polytechnic Institute”
Kyiv, Ukraine.
ORCID ID 0000-0001-6873-093X
lera.aveo@gmail.com

COMPARISON OF AUTOMATIC SYSTEMS OF TERMS’ EXTRACTION

Abstract. Nowadays the processes of translation become more unified, and translators depend not only on their knowledge and sense of language, but also on various software, which facilitate the process of translation. The following article is devoted to one branch of such software, the systems of automatic extraction, which are an essential part in the process of lexicographic sources development of translation of text, which include a variety of terms. Consequently, the necessity to choose among the variety of different programs arose and the results of this research i.e. the comparison of functions of different programs, are described in our article. Several criteria, by which the quality of terms extraction can be measured, have been compared, e.g., the speed of extraction, the “purity” of the output list of terms, whether the extracted lexical material corresponded to the requirements to terms, the quality of irrelevant choices, extracted by automatic extraction systems, and the factors, influencing this quality, etc. The advantages and disadvantages of cloud and desktop services have been investigated and compared. It was noted that the main difficulty is that programs still are not able to distinguish between word forms, thus the texts that undergo the extraction process, require auxiliary procedures such as POS-marking, lemmatization and tokenization. The other obstacle was the inability of certain programs to distinguish between compound terms and simple word combinations. The key points of the research may be used in the course of translation studies, in researches devoted to “smart” or electronic lexicography and by translators in general as they may use these systems of terms extraction during the process of translation for the purpose of forming or unifying the required glossary.

Keywords: terminology; terms extraction; automatic extraction systems; extraction software; terms extraction software; translation of terms.

1 INTRODUCTION

In modern conditions, that are changing constantly, application of automatic means to various industrial and working processes plays a crucial role. The more automatic the operations are, the more competitive the specialist is as their work is of a higher quality. Every sphere of human activity is turning to or blending with automatics and the translation industry is no exception. Linguists need to constantly keep an eye on technological advances in order to remain successful. Since the appearance of CAT-tools there were no revolutions in the sphere of translation, but evolution of existing means is constant. This article is devoted to only one aspect of translation activity – terminology, but not terminology as such but to a certain aspect of working with it – its extraction (or retrieval). Correctness and consistency of terminology are some of the key quality indicators of any translation. It all comes to wording, and no matter how intricate the text is and how sophisticated the style may be, incorrect or heterogeneous terminology may spoil the outcome. That’s where extraction plays the key part as it provides the possibility to pay attention to the most essential lexical material, represented in the text.

Primary **objective** of our research is to describe and organize the terminology of sustainable development both in Ukrainian and English. With the view to achieve this task we need to organize two different lists of terms which can be described; to achieve this we have selected several software types, by means of which this can be achieved, and further we will describe the outcome of their application.

Extraction of terms is the task which occupies a special place in linguistic studies. It’s connected with the fact that terms are the object of study of various branches of linguistics, but not only terminology as such, but also such branches as lexicology, lexicography, morphology, translation studies, etc. Our current research lies in between three branches – lexicography, terminology and translation studies. It results from the fact that the scientists nowadays distinguish between three branches of terminology: the first deals with standardization of language, the second – with terms as a special layer of lexis and the third is called "terminology for translation". Terminology-for-translation deals with searching for best possible ways of automatic extraction of terms, with the ways of terms' description, their organization into glossaries, applicable for usage in CAT-tools, and organization of electronic dictionaries (Candel Mora, & Carrio-Pastor, 2014, p. 167).

2 MATERIALS AND METHODS

The basis of our work was a substantial article (39 pages), devoted to issues of sustainable development, called "Sustainable development estimation methodology in the context of human life quality and safety", which describes the main indicators of sustainable development and how they can be actually calculated. The text was not processed via tokenization, was not lemmatized and was processed in its original form. For this comparative study we have selected several types of software, both desktop and cloud ones, free of charge and those, which are distributed for fee. Firstly, these are memoQ and SDL MultiTerm Extract, which are the representatives of the most popular CAT-tools in the industry, which means that the user coverage by freelance translators and translation agencies is substantial. Secondly, this is a separate program SynchroTerm, deisgned by Terminotix, which was developed specifically to extract terminology; it is not included in any known software packages and has very positive reviews on the Internet. SynchroTerm is not provided free of charge, so it was interesting to find out how expedient it is to purchase it, providing the availability more affordable solutions. Thirdly, three representatives of cloud solutions for extracting terminology were analyzed, namely, the Prospector service designed by Logrus Global, Tilde Terminology from Tilde and SketchEngine (OneClick Terms application – in particular), developed by Lexical Computing Limited. We can't but mention that this is not a definitive list of possible solutions, and in each of the categories you can find various other options, which might suit some very specific requirements.

3 RESULTS AND DISCUSSION

The first program we have tried was MemQ. It is used in translation projects both for organizing and delivering single projects and for coordinating the work of several translators through a server. The program provides the possibility to retrieve terminological units directly from the text. Among the advantages of the program we should mention the option of making a so-called "stop list", which contains samples of lexical units, which should be left out during the retrieval process of terminological lexical units. The user can specify the maximum number of words in a compound, the minimum number of characters in a term for one-word terms, the minimum frequency of occurrence of terms, etc. (Term Extraction Editor). The final selection of terms showed a lot of irrelevant choices, i.e., introductory phrases, adjectives without terminological meaning, nouns and verbs without terminological meaning. It became clear that the program does not apply semantic and distributive analyses, but extracts terms basing on a statistical method. So, the presence of the stop-list doesn't affect the situation to a required extent. If a particular term was encountered quite often, namely – a predefined lower threshold of occurrence frequency was exceeded, the program extracts the lexical unit. Otherwise, the term remains unnoticed. An important advantage of the MemQ program is its availability – the majority of functions are provided for free and only some of the functions needed mostly for managers' work require a prepaid access.

The opportunity to use the MultiTermExtract program was possible within the academic course of "Practice of Translation", in the trial version. In this case the functionality is not limited in its amount, but is limited in time. It was possible to select only the "noise" level (the same "irrelevant choices", but not in the manually selected stop list) in the settings (*SDL Multiterm User Guide*, 2011). The program showed the level of work which is approximately the same as MemQ; changing the level of the “noise” did not help; the list of terms increased or decreased due to repetitions, as well as different forms of the same terminological structure, for example, *human life quality* and *quality of human life*, *population quantity* and *quantity of people*, *natural resources* and *natural and ecological resources*. It should be noted that both programs select primarily the words and phrases, where noun plays the role of a semantic core, and almost no verbs are retrieved in the list.

Given the fact that one of the global goals of our study is the harmonization of two terminologies at once, it should be noted that both programs are not suitable for working with the Ukrainian language. MultiTerm has the option of working with the Russian language, but this is not suitable for the purposes of the main study.

SynchroTerm is the only program we have tried in our comparison that was created specifically for the purpose of terminology extraction and this program is not distributed for a fee. It seems fair to assume that the quality of extraction should be higher than in more multifunctional software, otherwise it would be no point in purchasing a separate product. Advantages of the program are the following: the set of settings in SynchroTerm is larger than in memoQ and MultiTerm Extract, but the basic settings are almost the same as in memoQ. The feature, deserving attention, is the option of retrieval of nouns only. The performance of SynchroTerm program is comparable to the one MultiTerm Extract level. Moreover, in the extracted terms, there were also irrelevant choices and fragments of phrases. However, it should be mentioned that the quality of retrieved terms was significantly higher than in those extracted by memoQ and SDL MultiTerm Extract, since SynchroTerm managed to extract the actual terms and did it rather correctly. In addition, the program interface is especially designed for terms’ processing. It is easier to analyze the context, select the appropriate terms and add them to the glossary.

SynchroTerm is more effective than memoQ and SDL MultiTerm Extract. The total amount of irrelevant lexical units was less and the number of extracted terms is bigger. Consequently, it takes less time to prepare glossaries. In addition, we can’t but mention that SynchroTerm supports both monolingual and bilingual extraction. But still the program does not work with the Ukrainian language. It is also worth noting that the program is distributed for a fee and the hardware must meet a number of installation requirements.

The process of application of these software proved the necessity of splitting the source text into segments and then extracting duplicate word forms i.e. lemmatization (there are free online tools for lemmatization, for example, CST's Lemmatiser). Also, before lemmatization, you need to perform POS-marking of the text.

Among the desktop programs, we tried to extract the terms; there was also the Simple Concordance Program app. The results of working with this app were unsatisfactory. The software works only with internal projects or files with the.txt. extension. In the process of work it also became clear that in case of text encodings’ mismatch (if the encoding was newer in Microsoft Word than in the program), the program simply “does not see” the loaded text. Among the advantages we should mention that the program is distributed for free and the presence of option to set up a personal “stop list” (a list of words that will not be included into the list of intended terms and terminological compounds).

In the course of our study, we compared several cloud services, namely Tilde Terminology and SketchEngine (OneClickTerm) and Prospector. All three services are shareware.

Tilde Terminology is a cloud service developed by Tilde. Among the peculiar advantages we can single out the possibility of choosing the topic of translation (which should somehow contribute to improving the quality of the terms retrieved) (*Tilde Terminology*). But“Sustainable development” was absent in the list of proposed topics. We tried several related topics, “Ecology”, “Economics” and “Social Studies” but this didn’t affect the quality of the retrieved terms to a significant extend. It is also worth noting that the when the topic “Social Studies” was selected, many phrases with the word “*human*”, which did not belong to terminological units, were extracted. Among the disadvantages of the program we should mention that in case of free use, the translator can deal with only one translation project. It is allowed to start a new one each time, but in case the researcher or translator needs to process several token sets sequentially, it might be inconvenient.

The paid subscription provides the possibility to conduct up to 25 projects. The option to work with the Ukrainian language is absent.

Logrus Global's "Prospector" service is a cloud service focused solely on working with terminology. However, it is important to mention that it works with English terminology only (*Logrus Global*). This service proved to be very efficient – in a set of 400 terms, 80 percent was retrieved correctly. Last but not least – the service extracts not only 1-word terms, but also the ones consisting of 2-3 words, and even complex terms, consisting of 4 words, which is especially impressive in case of un-processed text. The final glossary can be obtained in Excel file. The service is free, but the registration process can take certain time since a potential user needs to submit an application to the website of the developer company in order to gain access.

The "SketchEngine" service, and one of its utilites, OneClickTerm, is a cloud service for working with terminology. It is possible to work with parallel text corpora, as well as with bilingual terminology (however, the service does not work with the Ukrainian language). Also, the service developed by Lexical Computing Limited, has tools for processing of words and texts at different levels. Among the functions there is the option of setting the level of "noise", the frequency of the words encountered in the set, the maximum number of words that can be within the same phrase. The final list is represented in Excel file. The free version of the program is suitable only if a small text is supposed to be processed since the service "blurs" the names of the selected terms after the first ten and shows some other, but in random order. The number of irrelevant choices was at an acceptable level, although the final glossary should still be redefined manually. To disclose all the functions and obtain a full glossary, there are several purchase options - a monthly subscription and an annual one. There is also the option of an academic subscription, where the institution pays for use without limiting the number of projects from one account, which makes SketchEngine suitable also for academic purposes (for example, for studying in courses of in practice of translation or translation theory).

4 CONCLUSIONS AND SCOPE FOR FURTHER RESEARCH

We have compared 7 programs for automatic extraction of terms, in particular 4 desktop programs and 3 cloud services. The worst one for working with texts was "Simple Concordance Program", but among its advantages there was the ability to specify the initial list of words that would not be assigned as terms or parts of term compounds. Two cloud services, Prospector and SketchEngine, demonstrated better results than the others. But Prospector works with English terminology only, which is not suitable in the course of work on bilingual lexicographic resources. The application of all 7 programs showed the necessity of preliminary processing of the text with the view to achieve maximum results during extraction. In our opinion, the SketchEngine service will be promising for work on compiling monolingual and bilingual lexicographic resources, since it can be used both for practical purposes (directly for working on a dictionary/glossary), as well as in a study-and-work combination. It also has a number of additional tools for working with text and terms, which facilitates and systematizes the process of working with terms.

REFERENCES

Candel-Mora, M., & Carrio-Pastor, M. (2014). Terminology Standardization Strategies towards the Consolidation of the European Higher Education Area. *Procedia – Social and Behavioral Sciences*. 116, 166–171. DOI: 10.1016/j.sbspro.2014.01.187.

Computing. (n.d.). One Click Terms. Term Extractor. Retrieved November 20, 2019, from https://terms.sketchengine.eu/

CST's Lemmatiser. (n.d.). Retrieved November 20, 2019 from https://cst.dk/online/lemmatiser/uk/

Logrus Global. (n.d.). Prospector. Retrieved November 20, 2019 from https://logrusglobal.com/prospector.html

Sdl Multiterm Extract Tools User Guide. (2011). Retrieved from http://downloadcenter.sdl.com/T2011/Docs/SDL_MultiTerm_2011_Extract_User_Guide.pdf

Term Extraction Editor. (n.d.). Retrieved November 20, 2019 from https://help.memoq.com/current/en/Places/term-extraction-editor.html

Tilde Terminology. (n.d.). Retrieved November 20, 2019 from https://term.tilde.com/

Валерія Гавриленко. Порівняння автоматичних систем вилучення термінів. Процес перекладу стає дедалі більш уніфікованим і перекладачі мають покладатись не лише на власні знання і відчуття мови, а й на різноманітне програмне забезпечення, яке полегшує процес перекладу. В статті подаються результати дослідження однієї з галузей такого програмного забезпечення (ПЗ) – програм для автоматичного видобутку термінів – які є невід'ємною частиною процесу укладання лексикографічних джерел, а також перекладу текстів, насичених термінологічними одиницями. Наслідком такої різноманітності є необхідність порівняння різних типів програмного забезпечення з метою визначення їх переваг та недоліків. Проаналізовано критерії, за якими можна порівнювати програми автоматичного видобутку термінів: швидкість видобутку, «чистота» кінцевого списку термінів, відповідність відібраних одиниць критеріям термінологічності, якість і значення нерелевантних відборів та фактори, які впливають на якість відібраних одиниць. Були досліджені і порівняні переваги і недоліки десктопних програм і хмарних сервісів. Окремо було зазначено, що програми у більшості випадків не відрізняють словоформи, а тому тексти, які опрацьовуються, мають також проходити первинні процедури (а саме – POS-маркування, лематизацію і токенізацію). Також в статті обґрунтовується необхідність здійснення комплексного відбору, не лише автоматичними засобами, через нездатність ПЗ відрізняти складні термінологічні сполуки від простих словосполучень. Основні положення статті можуть бути використані під час перекладознавчих студій, в дослідженнях, присвячених «смарт» або електронній лексикографії, і спеціалістами-перекладачами в цілому, оскільки останні можуть використовувати системи автоматичного видобутку термінів для уніфікації робочих глосаріїв.

Ключові слова: термінологія; видобуток термінів; автоматичний видобуток термінів; ПЗ для видобутку термінів; переклад термінів.

Received: November 03, 2019
Accepted: November 18, 2019