Going parallel: using earlier translations as background for facilitating re-translation technique

Tetiana Anokhina[1] [orcidID0000-0002-8859-5568], Iryna Kobyakova [2][0000-0002-9505-2502], Svitlana Shvachko[3][0000-0002-2119-1884]

[1]National Dragomanov Pedagogical University, Kyiv, Ukraine
[2]Sumy State University, Sumy, Ukraine
[3]Sumy State University, Sumy
anokhina_mail@yahoo.com,kobyakova@ukr.net,shvachko.07@ukr.net

Abstract. As it goes for the re-translation technique, the modern data enable access and analysis of the previous translations for further re-translations based upon the previous empirical studies and earlier translations. Also, the students' re-translations are a perfect illustrated material to compile the little educational corpora being comparable due to the slight differences which make up the material of translation lacunae possible for analysis of the corpus-based studies. There are different tools for going parallel, including ParaConc, SketchEngine, tools applicable to the students (re-translation based on the previous translation) which illustrate their own ways to deal with difficulties of translation and lacunae. The paper provides an overview of tools such as AntConc for working with the re-translated words, MWUs and lacunae in translation corpus called ReTRans, the corpus of translator's re-translations. It is supposed to have texts of the original and published translation in order to build small self-made corpus of re-translations.

Keywords: students' translations, Wordfast, SketchEngine, AntConc Tool, bitext, re-translation technique.

1 Introduction

The corpus studies are applicable to many scientific spheres today. The linguistic areas of applied linguistic and translation studies are actively interacting with corpus linguistics giving new results approved statistically by the rich and overwhelming corpus material.

Among many problems of translation studies, the problem of finding the best matching equivalent is actual and popularized. To solve this problem and many others the corpus material and corpus tools appear to be the helping hand for translators and researchers.

The most popular modern tools for corpus analysis are concordancers which are effective tools for applied linguistics research and translation studies facilitating the learning of vocabulary, learning how to use fixed and non-fixed collocations in foreign languages. Among popular tools that are worth mentioning for further observations and discussions is Wordfast.

The Wordfast family are great translation tools that have the corpus match functioning. To prepare and use self-made parallel corpus the Wordfast aligner tool (https://www.wordfast.net/?go=align) can assist. There is a tendency in modern linguistics to have a corpus-based research. For the purpose of analysis he or she may use some tools ready to work with raw material and other requiring preparatory work for research and study.

If there are some txt files in the personal library, the AncConc tool (https://www.laurenceanthony.net/software/antconc/) will be serving for the analysis. If a researcher has at least the text of original and the text of translation the ParaConc (https://paraconc.com/) tool can be the needful choice. The tool working with both single units and multiword expressions might be helping to prepare the personal corpus along with using the existing corpora is SketchEngine (https://www.sketchengine.eu/).

Using the friendly interface of SketchEngine it is possible for linguists to compile their own corpus without writing any code. This popular tool helps to extract the searching items, one word or multiword units.

One can select well prepared annotated corpora in SketchEngine. It uses completely everything for compilation your own corpus (docs, pdfs, excel files, runs tmx files for setting up your parallel corpus) but it doesn't work with scans. Using SkectEngine our lab has started to compile ReTrans corpus comprising re-translations of popular and classical works made by students of our university.

Other tools serve differently but they all coincide in working with machine readable libraries, small and huge corpus data, give statistic information and may be used for corpus-based research and translation studies. AntConc, ParaConc, Wordfast aligner tool and SketchEngine are corpus managers and text analysis tools are able to work with numerous texts and corpora in various languages. They can provide collocation search and generate statistics related to the co-occurrences of collocations' constituents. A successful computational treatment of one word or multiword expressions leads to solving language ambiguity of the broad context. Word clusters and bundles tools search help in better collocations learning. The shifts of meaning of multiword units and idioms can be effectively learned relating the corpus-based approach to the interpreted data.

The idiom search is a effective in the corpus environment. We find the corpus tools are the needful tools today for linguistic research and translation studies. This study presents the study of corpus tools, their functioning and ability to analyze small or huge corpus data.

In our observation the software tools have shown its output based on their properties analysis. The above mentioned tools have been tested to have concluded the opinion of highly innovative tools with friendly interface serving for multifunctional approach and all worth trying and using.

2 The Ukrainian translation studies and corpus-based lacunology

Today translation students should not only learn the foreign language, enhancing writing, speaking, listening skills and improving their pronunciation and grammar [7].

It is the technological time when electronic texts and comparable corpora are accessible online; using the digitization forces translation students can include some additional skills to their CV.

As Veiga (2016) states the students should be working in the highly competitive profile of the translation market, to work with computer tools designed for linguistic analysis and translation [12].

It is important to analyze corpus-based technologies to develop techniques for identifying terms from a specialized field searching for them in internet and extracting from the comparable corpora, as our students English :: Ukrainian.

In this observation, we find the developed corpus-based tools used for students who are aiming to relate their research to the modern tendencies in applied and corpus linguistics for going parallel in their translation analysis and for better extracting elements with multiple realizations in contrasting languages (English and Ukrainian).

There are a variety of corpus-based studies that are applicable to many scientific spheres to confirm theories by the data available from corpora. The linguistic areas of applied linguistic and translation studies are actively interacting with corpus linguistics giving new results approved statistically by the rich and overwhelming corpus material. Among many problems of translation studies, the problem of finding the best matching equivalent is actual and popularized. To solve this problem and many others the corpus material and corpus tools appear to be the helping hand for translators and researchers.

The most popular modern tools for corpus analysis are concordances which are effective tools for applied linguistics research and translation studies facilitating the learning of vocabulary, learning how to use fixed and non-fixed collocations in foreign languages.

The modern linguistics go hand in hand with translation studies and corpus-based lacunology, the science with studies rare frequent words or non-equivalent words in other cultures as lacunae. These lacunae can be zero places in one language or difficulties of translations that can have different realization in multiple re-translations. The phenomenon of the lack of specific elements in the culture of one ethnic group against another in English termed as gap. Also, lacunae can be the aligned to the ethnographic elements, phraseologic units, multi words in the parallel translations resulting in either deletion of lacunar element or rendering “with the phrase, not by a word” [10].

We find the multi-word expressions (MWEs) to be one of the most heterogeneous phenomenona in the fields of Natural Language Processing (NLP) [2]. Today it is possible for find huge data comprising also MWEs in different corpora due to the electronic revolution.

Alghamdi and Atwell (2017) emphasize that the best computational practice related to MWEs processing is connected with new techniques and tools arise both with

machine translation developing and corpus studies. The both constituents help to find better extraction models and computational ways of study [2].

2.1 Lacuna as a unique phenomenon

Thus, lacuna is a unique phenomenon which mirrors zero reflection of non-equivalent vocabulary. Lacunae are quasi-comparable units that can refer to various referents of ethnic cultures. The comparable corpora return the lacunar queries in multiple variations. We trace lacunae as difficulties of translation.

There are many lacunae in the text of the original. To find some of them students can analyze the footnotes which eliminate lacunae by explanations given in the end of the text or in the mode "footnote at the page". Nowadays the footnotes are electronic which is very convenient for reading the electronic texts.

In order to work effectively with corpus material AntConc, ParaConc, Wordfast aligner tool and SketchEngine can be used. These are corpus managers and text analysis tools helping to work with numerous texts and corpora, for our purposes with parallel corpora in English – Ukrainian pair. Using these tools students can find the needful collocation searches and look for translation provided in the self-named corpus.

The extracted translation equivalents alone and translation memory file obtained after aligning translation pairs (bitext) can be used for re-translation purposes when students make their translation based on the re-translated text which they add as translation memory files (using Wordfast software).

Successful re-translation files are making the small corpus of students' re-translations (reTRans). This kind of activity is effective for learning unique translation pairs, helping students to find better results comparing their results with previously translated text and other variants in the reTRans corpus (the small self-made corpus of students translation) [6] which contains the original and different translation English – Ukrainian pairs. It is supposed to use SketchEngine tool in order to preload small or bigger corpora and use for analysis in the SketcthEngine environment.

As SkecthEngine is the most powerful tool it is highly recommended to use it for automatic extraction of needful words, lemmas and multiword units ("terms" in SketchEngine). It is also has POS (part of speech) search which enables to navigate and extract the needful data from other accessible corpora.

As students papers are corpus-oriented but deal with such linguistic material as adverbs/nouns/verbs, it is possible to add their data into ReTrans project based on Harry Potter series, Nebo, The Da Vinchi Code and other re-translations compiled from their smaller corpora into bigger one.

The special attention has been paid to corpus based search of translation variants. Due to the broad context search students can interpret the data. Some scholars claim [9] the lexical unit is very often longer than a single word [4]. As McEnery and Wilson (2001) state that one of the major trends in corpus linguistics over the past few years is the increased interest in very small, highly specialized corpora. Small corpora can be used for a great many different purposes [9].

The date of translation difficulties comprise single words (glossaries), multiword units (terms) and ideas hidden implicitly. To follow the latter task the translation corpus has been compiled to see how translation tasks are performed at the level of word, word unit (N-grams) and at the semantic level (meaning that the secondary text feels the most successful translation. Also, the needful phrase for translation may be found online or in the downloaded corpus by the idiom search.

As Pierre Colson (2017) believes that automatic extraction is possible from large web corpora in different languages. The use of a new algorithm based on metric clustering techniques made it possible to find long N-grams from the freely accessible web [1].

We find the translation corpora are giving more information that corpus-based dictionaries or glossaries of terms. If we are talking about re-translation corpus it is rich in contextual variants and can be effectively used for learning EFL purposes and for learning how to translate better, how to find more translation variants. For semi-automatic analysis the translation pairs are downloaded into separate files containing the unique translation making the subcorpus of the larger corpus of students' re-translations (ReTrans) (Figure 1).

Fig. 1 The SketchEngine interface of the translation corpora used for re-translation

We use the corpus tools in order to work more effectively with translation corpora (parallel corpora). The bellow mentioned needful tools serve the crucial role in translation studies. Our observation aims at describing the corpus tools for "going parallel", their functioning and ability to analyze small or huge corpus data. In our article we have mentioned the most popular software tools along with their output based on properties analysis.

There are plenty of tools to be tested with highly friendly interface serving for multifunctional purposes and worth trying and using. Among them there are some free and prepaid tools. Firstly, we regard free tools available for students with any paying

abilities. The highly innovative but prepaid tools such as TRADOS and SkecthEngine are regarded also for being currently the best as the market place.

3 Corpus Tools observation

It is highly important for the Ukrainian students to learn how they can make automatically the alignment of parallel texts (English and Ukrainian).

As a rule, in a translation class they make parallel texts when translating from Ukrainian and into Ukrainian. The problem is that there is only OPUS corpus available for parallel comparison which is not enough for educational purposes, so new translation corpora should arise.

We find it is to use comparable corpora from other languages into Ukrainian and English. This application area (the alignment of parallel texts from multilingual corpora) is highly important today. That is why we are making the small corpus of students translations based on the existing translations.

It may be considered helpful to use not only aligned 'translation pairs' from English – Ukrainian but also from other languages with parallel component (English or Ukrainian). Students had to find a txt file in English and in Ukrainian or scan and send the language pair to txt files (a very time consuming task).

If a researcher has at least the text of original and the text of translation the ParaConc tool can be the needful choice. Other software tools have the option to include the parallel corpora, such as SketchEngine. Wordfast tool generates translation memory files that can be used in other tools supporting the format.

Lacunae as difficult places to render in translation can be vividly observed in multiple re-translations performed by students. Lacunae thus will be found as aligned to the expanding explanation, footnote or omission or contextual change. We find all rare frequent units in the contrasting languages to be considered as lacunae. If the unit is very rare (less than 1 lemma entry per corpus) it is considered to be a hapax (law frequent unit/lacuna).

Lacunae, which are difficult to translate, are found in the contrast analysis of artistic translations and their re-translated versions, since they provide more complete information about the removal of lacunar elements in one or more versions of the translation (lacunar units) and contain extra text, preface, afterword, etc.

There are some resources available on-line, including free corpora of translated works are not systematized by the degree of extra-linguistic additions, commentary, or preface (Paraconc concordance; Ukrainian Language Corps; National corps of the Russian language).

Linguists and programmers are involved in creating independent databases, scanning and sorting data according to their tasks. There are open source platforms available for development [http://opus.nlpl.eu/].

Hapaxes (rare units or lacunae) are not usually included in traditional enclosures, as these bases are more useful for improving the work of electronic translators, for whom it is important to have contextually dynamic correspondences.

The minimum retrains corpus must contain n-versions including the original, its translation and re-translations (Table 1).

Table 1.

Original	Translation	Re-translation
Theft: a love story	Translation from English	Ukrainian
Harry Potter series	Translation from English	Ukrainian
The Da Vinci Code	Translation from English	Ukrainian
Alice in Wonderland	Translation from English	Ukrainian
Surely You're Joking, Mr. Feynman!	Translation from English	Ukrainian

3.1 AntConc as concordancer tool and text analyzer for the ReTRans corpus

We find the AntConc tool efficient for multiple re-translations versions analysis, as we had put the texts of translations in the manner text 1-n to analyse and compare re-translation results compiling the small comparable corpus of the printed and students' re-translations ReTRans.

The central tool used in most corpus analysis and translation studies we strongly recommend AntConc as your first concordancer. The texts of re-translations (even not annotated) give the information about difficulties of translation. The texts of re-translations can be downloaded to AntConc tool (each file named 1…n) and the difficult place (usually has a footnote in translation) can be quickly found in the other text translated by another student. It is the way how is the best variant is found.

This method can be used for effective assessment of students' translation (the figures1…n are used instead naming the file by the students' name).

AntConc can be used for this purpose. It is simple to use, it has a friendly interface and it is free of charge. All these make it to be the best choice corpus manager tool to work with raw texts in your compiled corpus of texts. Being used for research in translation and computational linguistics AntConc concordancer as Sun and Wang (2003) describe is a tool for learning vocabulary, collocations, and multiword units [4].

Also, if there is some or files in the personal library, the AntConc tool will be serving for the analysis working directly on the raw texts of the corpus. The idea to work with separate words is not working any more in linguistics and translation studies [13]. Word for word translation is the past of machine translation which is relying on editing and corpus data extraction (as Wordfast tools).

In AntConc, multiword units can be investigated using the Word Clusters Tool. As Lawrence (2004) states:

> *This tool displays clusters of words that surround a search frequency. The search term can be specified as a substring, word, phrase or regular expression as in the Concordancer, Plot and View File tools, and the number of additional words to the left and right of the search term can also be speci-*

fied. It is also possible to set a minimum frequency threshold for the clusters generated [13].

An alternative way to sear bundles [3], which are equivalent to N-grams, where n can vary usually between two and five words. Few corpus analysis programs offer this feature [5], but AntConc includes lexical bundle searches as an option in the Word Clusters Tool [13].

Lonfils and Vanparys (2001) explain how the AntConc works like a little but effective 'ant' as its logo shows, with most essential features of standard software applications to work with words in the contexts and set expressions, stable or free multiword units (MWEs) [8].

The needful function is bundles tool effective for translation studies of the unknown MWEs or word clusters, foreign collocations difficult for learners to acquire. Having one text or a book, the smallest corpus, the researcher can easily extract the searched item in the nearest context with enable its proper usage.

The program let us see the bundles two and five words) equivalents [3].

AntConc performs all operations directly on the raw texts of the corpus. This is useful in that the user is often switching or modifying the target corpus for a particular need, as the program does not need to do any pre-processing of the data, for example, creating an index. On the other hand, because AntConc does not use an index, it can only work effectively with small scale corpora [13].

Most corpus analysis programs offer users the ability to see collocates of a search term in a table, where the frequency of the most common words to the left or right of the search term is indicated. Learners often find such tables difficult to interpret and so the current version of AntConc offers no implementation of this feature. However, for advanced learners this can be a severe disadvantage that will be addressed in the next release of the program.

Students may use AntConc program [13] with statistics to run in the selected corpus, working both wit annotated and non-annotated data search results. Young learners have the helping advice using simple keyboard shortcuts which they can follow, namely: working with statistics just to copy and paste your results into a spreadsheet program for further analysis and research. The program enables to see the bundles as two, three, four or five words equivalents also called N-grams. Also, this free tool has offline concordancing feature.

AntConc supports different formats of annotated data and non-annotated data. It runs raw text, also it works with text in HTML/XML format with possibility to view or hide embedded tags used in HTML/XML [13] Student do not use all the possibilities at one time but as they get interested in what they can do with annotated corpora, they are trying to use their knowledge in their Bachelor' and Master' papers.

4 Translators' engines and CATs

Some free tools as ZExtractor are used for the automatic extraction of term candidates in a text or set of texts. The paid for Ukraine is a perfect program SketchEngine which

is a powerful linguistic data analysis system that has several features. Among them there is the Keywords feature to identify words in a text or set of texts.

In order to covert files into text formats, Notepad for Windows can be used. Also, in order to compile data we Microsoft Excel allowing filtering, calculate, compare and analyze data.

All these programs are helping young researchers to compile their own corpus (small or huge) for research or translation purposes, for instance for compiling the specialized dictionaries to use them by machine translation tools such as TRADOS, WORDFAST or other tools [12].

The popular tools worth mentioning and further observations and discussions are Wordfast family. The Wordfast tools are great translation tools that have the corpus match functioning. To prepare and use self-made parallel corpus the Wordfast aligner tool can assist. There is a tendency in modem linguistics to have a corpus-based research. It is the tool we are using to align our ReTRans corpus of students re-translations of the popular and classical masterpieces.

For the purpose of analysis translation students may use some tools ready to work with raw material and other requiring preparatory work for research and study. Short pieces of text (2 to 160 pages) are aligned.

4.1 The translation experiment

The mentioned above tools serve an educational purpose. The re-translation data is easily accessible for research and analysis. In class students are to compare what was done in translations and using the experience of many other translators in order to make the re-translation copy readable. Their practice includes editing the segments of the text, using machine translation, and creating the memory based on the existing translation. For educational purposes they translate the previously translated pieces.

Wordfast anywhere tool that can be used freely gives students to work with one memory at the time. It is important for students to find the text file in English and Ukrainian to align (usually it is classical masterpiece freely downloadable) or to buy e-text and its translation for the compilation translation work.

When this material is ready students are making their translation try (from one page to chapter) to compare and analyze with the original. The next stage is compiling MWUs to MWUs corpus [11].

5 Final remarks

In our studies we rely on the new parallel corpora based mostly on classical prose and best-sellers translated into English.

Step by step students are translating masterpieces into Ukrainian, making the small corpora of their secondary translations with the possibility to compare with the previous translations from the languages they can comprehend into Ukrainian.

The translations available from English and into English are making the huge corpora potential as they include worldwide classic and modern popular prose translated into English.

The idea is to facilitate their translation using the previous translation. Using the technological tools such as AntConc, SketchEngine as corpus tool and translation software Wordfast/Trados they align the translation pair into the translation memory before they translate the text making the new re-translation with their own footnotes and translation lacunae elimination.

References

1. Colson, J.-P.: The IdiomSearch Experiment: Extracting Phraseology from a Probabilistic Network of Constructions. Europhras, LNAI 10596, 16-28 (2017)
2. Alghamdi, A., Atwell, E.: Towards Comprehensive Computational Representations of Arabic Multiword Expressions. Europhras, LNAI 10596, 415-431(2017)
3. Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E.: Longman grammar of spoken and written English. Longman, London (1999)
4. Bowker, L., Pearson, J.:Working with Specialized Language: A Practical Guide to Using Corpora.Routledge, London / New York (2002)
5. Coniam, D.: Concordancing oneself: Constructing individual textual profiles. International Journal of Corpus Linguistics (2), 271-298 (2004)
6. Krajka, J.: Language Teachers Working with Text: Increasing Target Language Awareness of Student Teachers with Do-It-Yourself Corpus Research. In: Working with Text and Around Text in Foreign Language Environments. Springer, Cham (2016)
7. Lee, D., Swales, J.: A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. English for Specific Purposes \textbf{25}, 56-75 (2006)
8. Lonfils, C. and Vanparys, J.: How to design user-friendly CALL interfaces. Computer Assisted Language Learning (5), 405-417 (2001)
9. Mcenery, T. and Wilson, A.: Corpus Linguistics. An Introduction. Second edition. Edinburgh Press, Edinburgh (2001)
10. Markovina, I., Sorokin, Y. Kul'tura y tekst. Vvedenye v lakunolohyyu: ucheb. posobye [Culture and text: the textbook]. GEOTAR-Media, Moscow (2010)
11. Osenova, P., Simov, K.: Modelling multiword expressions in a parallel Bulgarian-English newsmedia corpus. In: Multiword expressions. Insights from a multi-lingual perspective (Phraseology and Multiword Expressions 1).Manfred Sailer & Stella Markantonatou (eds.), 247-271 (2018)
12. Veiga, A. Using corpus linguistics tools to help translation students create technical glossaries. In: ICICTE Proceedings, 341-347 (2016)
13. Lawrence, A. AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom. In: IEEE International Professional Communication Conference Proceedings, 729-737 (2003)
14. Sun, Y.-C., Wang, L.-Y. (2003). Concordancers in the EFL classroom: Cognitive approaches and collocation difficulty. In: Computer Assisted Language Learning, 16, 83-9