Compiling a Specialised Corpus for Translation Research in the Environmental Domain

Anastasiia Laktionova
University of Wolverhampton
laktionovaaa@gmail.com

Abstract

The present study is an ongoing research that aims to investigate lexico-grammatical and stylistic features of texts in the environmental domain in English, their implications for translation into Ukrainian as well as the translation of key terminological units based on a specialised parallel and comparable corpora. The research will comprise the process of creation of the English-Ukrainian parallel corpus and a Ukrainian comparable corpus and the exploration of the resource from a translational perspective.

1 Introduction

The research focuses on compiling a parallel English-Ukrainian corpus comprising environmental texts, conducting an analysis to identify genre-specific lexico-grammatical and stylistic features of the texts in the corpus using the corpus management tools, identifying key terminological units in the English corpus and their translations and qualitative analysis of the pairs in order to identify potential inconsistencies and errors, as well as multiple and null translation equivalents associated with specific English terminological units, compiling a comparable Ukrainian corpus of texts based on the key terms of the parallel corpus and conducting quantitative and qualitative lexico-grammatical and stylistic analysis of translated and non-translated (original Ukrainian) texts in the two corpora.

For the corpus, international conventions, protocols and agreements in the environmental domain are collected to serve as a base for analyzing specific lexico-grammatical and stylistic features of from the translation perspective, observing possible differences between translated and native use of language in the domain as well as capturing possible inconsistencies in the translation of terminological units: such as mistranslations, non-translations or multiple translations of terms.

2 Methodology

The methodology that underpins the study is based upon the latest developments in the field of corpus linguistics, which has become an extremely common paradigm for studying various aspects of language, including translation. A corpus can be defined as a collection of authentic texts held in electronic form and assembled according to specific design criteria (Laviosa, 2013, p.228) or a form of linguistic data which represents a collection of written texts (or transcripts of speech) that can be searched by a computer using specialized software (Brezina, 2018, p.1). Corpora can contain texts in several languages, representing either parallel or comparable types. Parallel corpora are corpora that contain native language (L1) source texts and their translations (L2). Comparable corpora contain texts that are collected using the same sampling frame (the same proportions of the texts of the same genres in the same domains etc.) in different languages.

In order to apply a corpus to translation studies it must be compiled in several stages in line with a number of criteria. Scholars outline the following steps in the compilation process: the selection of texts, followed by preprocessing, annotation and alignment, after which the search and retrieval of information will be conducted, and the analysis of

Proceedings of the Student Research Workshop associated with RANLP-2021, pages 94–98,
held online, Sep 1–3, 2021.
https://doi.org/10.26615/issn.2603-2821.2021_014

findings. An important feature of corpus design, applicable for parallel corpora, is text alignment, which means the connection of "the parallel or translational relationship between the source texts and target texts at different levels" (Hu, 2016, p.37). The alignment can be performed at lexical, sentence, paragraph or text level. This process is quite technically challenging, but is crucial for further corpus analysis. Alignment is carried out on segment level and realization of alignment at the sentential level is necessary for the investigation of translation equivalents. At present, some corpus management tools already have in-built alignment functions (for example, SketchEngine (Kilgarriff et al. 2014)), and for this process it is only required to save corpus data in a vertical format. However, in order to make corpus resources even more useful and insightful, it is necessary to perform corpus annotation (manual or automatic) at various levels (phonological, morphological, lexical (POS tagging, lemmatization, semantic annotation), syntactic (parsing, treebanking), and sometimes at discoursal level (adding coreference annotation) depending on the purpose of the research.

For this research, a specialised parallel corpus and a comparable corpus is compiled. The texts in the corpora represent legal environmental documents, such as environmental conventions, protocols, guidelines on legislation approximation and agreements available in English and Ukrainian. The size of the parallel corpus will be around 100,000 tokens and around 50,000 tokens for the comparable corpus in Ukrainian. The size is limited due to the available translations of the English texts: a larger number of documents is adapted, and not translated. It is planned to carry out tokenization of texts in Python, followed by lemmatization. Other most common annotation types include part-of-speech tagging and syntactic parsing, however resources for these types of annotation are not highly developed for Ukrainian and will not necessarily be used for annotation of the corpora. Corpus analysis tools include frequency lists, concordances, keywords and n-grams, as well as statistical methods (frequency, dispersion etc.).

3 Related work

There is a substantial number of monolingual corpus resources for many languages, and a great number of parallel and comparable corpora have been compiled for various languages, for example the Europarl parallel corpus, consisting of the proceedings of the European Parliament, the OPUS (open parallel corpus) which contains texts in 40 languages, the "Oslo Multilingual Corpus", the "ACTRES Parallel Corpus" (P-ACTRES), which contains English source texts and their Spanish target texts, EUR-Lex corpus with documents in the official languages of the EU, English-Norwegian Parallel Corpus and many other. This approach is applied to an increasing number of language pairs and domains.

A specialised corpus contains texts of a particular type, e.g. computer manuals, medical package inserts etc. whereas a general corpus is normally large, balanced, containing as many text types as possible and more representative of a language in general. An example of a specialised corpus in the environmental domain is the EcoLexicon English Corpus (EEC), which contains 23.1 million words and consists of contemporary environmental texts in a wide range of genres (León-Araúz, San Martín and Reimerink, 2018). It was built by the LexiCon research group for the development of EcoLexicon – a terminological knowledge base for the environment. Specialised parallel and comparable corpora have proven to be very useful in domain-specific translation research, terminology extraction and other practical applications.

In case of a general corpus, the size has to be as large as possible, in order for it to represent the variety of a language. However, in a specialized, domain-specific corpus it is easier to achieve representativeness of this particular language type based on a smaller-sized corpus. McEnery, Xiao and Tono (2006), for example, highlight that the representativeness of general and specialised corpora should be measured in different ways: for a general corpus it is related to sampling from a wide range of genres, for a specialised one – it can be measured by the level of 'saturation' or 'closure' at the lexical level. According to the authors, saturation or closure for a linguistic feature (for example, the size of lexicon) of a language type or variety (for example, environmental conventions) means that "the feature appears to be finite or is subject to very limited variation beyond a certain point". It can be concluded, that specialised corpora in a certain subject area have high saturation or concentration of vocabulary that represents this area and it affects the potentially sufficient size of such corpora.

As for the Ukrainian language, there are several large general corpora compiled in the recent years. One of the publicly available corpora is "GRAC" – the General Regionally Annotated Corpus of Ukrainian, which contains more than 400,000,000 tokens and represents most genres of written texts. A specific feature of the corpus is the regional annotation, which means that about half of the texts are attributed with regard to the different regions of Ukraine or countries of the diaspora (Shvedova 2020, p.489). Another available corpus is the Ukrainian Web Corpus of the Leipzig University, which contains a little over 1,5 billion tokens. It contains internet texts from the year 2014 and it is only possible to search word forms to see textual examples or collocations; the corpus also has a feature of graphs, visualizing frequencies of word forms co-occurred in a sentence. As for specialised corpora, there has recently been developed a corpus of texts in the medical domain called UKRMED (Cherednichenko et. al 2020). An example of a specialised comparable corpus is a corpus of political media discourse containing texts in English and Ukrainian (400 texts from 2014-2017) (Romanyshyn, 2020).

Available resources also include some general parallel corpora: Parallel Ukrainian-Russian and Russian-Ukrainian corpora within the Russian National Corpus (6,5 million tokens) and some parallel corpora developed by The Corpus Project of the Laboratory of Ukrainian, that include Polish, French, German, Spanish, Portuguese and a 1,5-million-token English-Ukrainian parallel corpus (bidirectional, mostly containing translation of fiction). There is morphological tagging in parallel corpora, made automatically with the Universal Dependencies system. This parallel corpus is based on the NoSketchEngine Platform and is publicly available online for searching. At present, corpus resources are widely present for English and other common European languages. Despite some active work on development of corpora, they are still not available in sufficient volume for Ukrainian, and especially scarce are bilingual and specialised corpora.

4 Environmental translation

Translation of texts in the environmental domain is challenging due to the relative recency of the environmental science as a specialized field, its multidisciplinary nature, and its fluid terminology, which includes both single-word terms and multiword expressions. Another factor is the urgent social message in many texts, which are fueled by the assumption that time is running out. (Faber and León-Araúz, 2021 p.589). There is research indicating inconsistencies in environmental terminology translation from English. For example, a study by Krimpas and Karadimou (2018) shows evidence for a number of terminological issues in official translations of international environmental conventions translation into Greek. A few studies by Ukrainian researchers also point out potential difficulties in translation of environmental texts focusing on the lexico-semantic aspect (Chervonetsky and Chervonetska, 2015).

In light of the European integration of Ukraine and the respective approximation of Ukrainian legislation to the EU, including the areas of environment and energy, the Ukrainian national government as well as local municipalities constantly adapt various European environmental conventions and use materials translated from English to create according legislation and development plans, implemented in many Ukrainian cities. The materials are actively disseminated to the population and main novel concepts are even introduced to the life in the country and implemented into governments' policies. In view of that, adequate translation of texts in the environmental domain is extremely important, especially for developing countries such as Ukraine. Moreover, it has been pointed out by researchers, that for lesser-used languages, terminology transfer from English can often be observed due to the import of technical advances from European countries where terms are mostly coined in the widely-used English language (Krimpas and Karadimou, 2018, p.22).

5 Research questions

The motivation behind the research is the limited availability of multilingual specialised corpus resources which include the Ukrainian language and the choice of the environmental domain is conditioned by this field being one of the most often communicated and crucial for the developing countries. In the course of the research the following research questions need to be addressed: What are the specific features of international conventions, protocols and agreements in the environmental domain? Is there a difference between translated and native use of

language in the domain? How do translations and original texts written in the target language differ in terms of their lexico-grammatical and stylistic features? Are there inconsistencies in the translation of terminological units: are there mistranslations, non-translations or multiple translations of a term? Compilation of a parallel corpus and a comparable corpus will serve as a basis for the investigation.

6 Conclusions

The research focuses on building a parallel and comparable corpus of environmental texts which can be further used for a number of applications, such as terminology or information extraction, research on differences and similarities between the English language and Ukrainian language, or between native and non-native language speakers' output (the translationese or textual fit), language-specific features, universal features, and any typological or cultural differences and potentially for MT. Analysing potential inconsistencies in official translation of terminology in the domain can be applied in terminology resources for translators, specialists developing and enhancing the terminology database or for creating a specialised terminology database.

Acknowledgements

I would like to express my gratitude to my supervisors Dr Maria Stambolieva (New Bulgarian University) and Dr Sara Moze (University of Wolverhampton) for their encouragement, help and valuable feedback for my research. I would also like to thank Prof Ruslan Mitkov, the EM TTI coordinator, for encouraging students to submit their research to the conference.

References

Baker, Mona. 1996. Corpus-based translation studies: the challenges that lie ahead. In H. Somers (ed.) *Terminology, LSP, and Translation: Studies in Language Engineering in Honour of Juan C. Sager*. Philadelphia/Amsterdam: John Benjamins, pages 175-186.

Biel, Lucja and Giczela-Pastwa, Justyna. 2016. Metody korpusowe w analizie gatunków specjalistycznych – założenia, perspektywy i ograniczenia. In *Pod pretekstem slow. Ksiega jubileuszowa dla Profesora Wojciecha Kubinskiego*, Gorszczynska, P. and Karwacka, W. (eds.), Gdansk, Częsci Proste.

Brezina, Vaclav. 2018. Introduction: Statistics Meets Corpus Linguistics, in *Statistics in Corpus Linguistics*, pages 1–37.

Cherednichenko, Olga *et al.* 2020. Collection and processing of a medical corpus in Ukrainian. In *CEUR Workshop Proceedings* (Volume 2604), pages 272–282. Available at: http://ceur-ws.org/Vol-2604/paper21.pdf

Chervonetsky, Volodymyr and Chervonetska, S. 2015. Problems of text translation in the field of ecology from the English language into Ukrainian: lexico-semantic aspect. In Ivashchenko V. (ed.) *Terminolohichnyi visnyk (Collected papers)*, vol.3(2), Kyiv, National Academy of Sciences of Ukraine, pages 94-101.

Cronin, Michael. 2013. Translation and globalization. In Millán C. and Bartrina F. (eds.) *The Routledge Handbook of Translation Studies*. New York: Routledge, pages 491-502.

Faber, Pamela and León-Araúz, Pilar. 2021. Designing Terminology Resources for Environmental Translation. In Ji M. and Laviosa, S. (eds.) *The Oxford Handbook of Translation and Social Practices*. New York: Oxford University Press, pages 587-615.

Hu, Kaibao. 2016. *Introducing Corpus-based Translation Studies, Introducing Translation Studies*. Shanghai: Springer.
doi: 10.4324/9781315691862.

Ji, Meng. (ed.). 2019. *Translating and Communicating Environmental Cultures*, New York: Routhledge.

Kilgarriff, Adam *et al.* 2014. The Sketch Engine: ten years on. In *Lexicography*, (1), pages. 7–36. Available at: https://www.sketchengine.eu/wp-content/uploads/The_Sketch_Engine_2014.pdf.

Krimpas, Panagotis G. and Karadimou, Despina. 2018. Terminological issues in official translations of international environmental conventions. In *Parallèles*, 30(1), pages. 21–35.
doi: 10.17462/para.201

Lapshinova-Koltunski, Ekaterina. 2015. Variation in translation: Evidence from corpora. In *New directions in corpus-based translation studies, Translation and Multilingual Natural Language Processing*,
Fantinuoli, C. and Zanettin, F. (eds.), Berlin, Language Science Press, pages 93–113.

Laviosa, Sara. 2013. Corpus linguistics in translation studies. In Millán, C. and Bartrina, F. (eds.) *The Routledge Handbook of Translation Studies*. London and New York: Routhledge, pages 228–240.
doi: 10.4324/9780203102893.ch17.

León Araúz, Pilar and García Melania C. 2020. Term and translation variation of multi-word terms. In: Mogorrón Huerta, Pedro (ed.) *Multidisciplinary Analysis of the Phenomenon of Phraseological Variation in Translation and Interpreting*. MonTI Special Issue 6, pages. 210-247.
doi: 10.6035/MonTI.2020.ne6.7

León-Araúz, Pilar, San Martín, A. and Reimerink, A. 2018. The EcoLexicon English Corpus as an Open Corpus in Sketch Engine. In *XVIII EURALEX*

International Congress: Lexicography in Global Contexts, Ljubljana, pages 893–901.
McEnery, Tony and Hardie, Andrew. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
McEnery, Tony, Brezina, Vaclav, Gablasova, Dana and Banerjee, Jayanti. 2019. Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use. In *Annual Review of Applied Linguistics*, (39): pages.74–92.
Romanyshyn, Natalia. 2020. 'Application of corpus technologies in conceptual studies (based on the concept Ukraine actualization in English and Ukrainian political media discourse). In *CEUR Workshop Proceedings*, 2604, pages 472–488.
Sinclair, John. 2004. Corpus and Text — Basic Principles in *Developing Linguistic Corpora: a Guide to Good Practice*. Edited by M. Wynne. Available at: https://users.ox.ac.uk/~martinw/dlc/chapter1.htm.