Lingua Legis
nr 24, 2016, s. 91-95
lingualegis.ils.uw.edu.pl

Dariusz Koźbiał

Uniwersytet Warszawski

Ewa Gruszczyńska and Agnieszka Leńko-Szymańska (eds). *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora*. Warsaw: University of Warsaw. 2016. 280 pp. ISBN: 978-83-935320-4-9

On 30 March 2015, a conference under the title *Polish-Language Parallel Corpora* was held at the University of Warsaw. It was organized by the Institute of Applied Linguistics with the aim of providing an opportunity to exchange ideas and experience among academic researchers from a range of language-related disciplines dealing with corpora in multilingual and multicultural contexts as well as disseminating information about either ongoing or completed projects among the wider public.

The book entitled *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora*, edited by Ewa Gruszczyńska and Agnieszka Leńko-Szymańska from the Institute of Applied Linguistics, University of Warsaw, is a post-conference monograph published in 2016. It contains an introductory chapter and fourteen chapters (either in Polish or English) on the topic of corpora written by the conference participants, which are followed by notes on contributors. The general aim of the book was to present current projects related to compilation of parallel corpora including a Polish language component as well as to provide descriptions of studies conducted on parallel corpora which contain Polish data. In addition, the goal of the volume was also to encourage more scholars to compile their own corpus resources in order to contribute to the already fast-paced development of the field involving Polish language parallel corpora.

As Kenning (2010: 487) rightly observes, although corpus terminology has had enough time to settle since the beginnings of corpus linguistics in the late 1980s, there is still some inconsistency in how its key terminology is used. Both earlier and more contemporary works on corpora may include terms, such as *parallel* and *comparable corpus*, which are used interchangeably, depending on the researcher's approach (Kenning 2010: 487). Therefore, it is advisable to draw

a clear division between these two terms as their clear delimitation is crucial when reading the book under review and dealing with corpus linguistics in general.

A glossary entry under *parallel corpus* found in a book on corpus linguistics written by Tony McEnery and Andrew Hardie (2012: 248) defines a parallel corpus as:

> a corpus consisting of the same texts in several languages. This typically means a set of texts written in one language together with each text's translation into a second language (or into several other languages).

What distinguishes parallel from comparable corpora is that parallel corpora contain a common source text. In addition, parallel corpora may be either bilingual or multilingual as well as unidirectional and bi-directional (Kenning 2010: 488). Comparable corpora, on the other hand, contain texts chosen according to a common sampling frame with regard to time span, text size, text types, etc. (*ibid.*: 487-488). McEnery (2012: 240) defines a comparable corpus in the following way:

> a corpus containing two or more sections sampled from different languages or varieties of the same language in such a way as to ensure comparability. If more than one language is involved, this is a type of multilingual corpus.

As of today, corpora alongside corpus analysis tools are used by scholars from a wide range of disciplines, e.g. translation studies, contrastive linguistics, intercultural studies. Types of multilingual corpora that researchers find of use include parallel corpora and comparable corpora. Monolingual corpora are also found to have many uses: they may, for example, provide material against which the data obtained from previous studies can be assessed and later supplemented.

Some of the articles contained in the volume provide an answer to the question why parallel corpora are still scarce. This has to do mainly with the dearth of materials available in bilingual versions (i.e. originals and translations) and copyright issues. Furthermore, there are also other difficulties in the process of creating such corpora which make it even more time-consuming; these are, inter alia, interface construction, alignment, annotation and lemmatization.

In Chapter 2, Alexandr Rosen provides a description of the make-up of *InterCorp*, i.e. one of the largest available parallel synchronic corpora covering 39 languages. The corpus is compiled mostly by teachers and students of the Faculty of Arts at the Charles University in Prague. In Chapter 3, Milena Hebal-Jezierska,

Aleksandr Rosen and Elżbieta Kaczmarska provide an analysis of the challenges faced by compilers of *InterCorp* as well as an analysis of its users' needs; the same chapter also includes a detailed discussion on the Polish component of the corpus.

In Chapter 4, Piotr Pęzik provides a description of the *Paralela* corpus, i.e. a parallel Polish-English corpus. *Paralela* is one of the language tools and resources belonging to the CLARIN-PL infrastructure. The author of the chapter demonstrates the possibility of using *Paralela* to research idiomaticity in translations from English into Polish and vice versa.

Chapter 5 authored by Marek Łaziński and Magdalena Kuratczyk presents a Polish-Russian parallel corpus compiled at the University of Warsaw. It discusses the corpus design as well as cultural aspects which affected the choice of texts. The chapter presents two examples of corpus applications for studying translation equivalents and ends with a discussion on the significance of the project and its future.

In Chapter 6, Andreas Meger, Michał Woźniak and Ruprecht von Waldenfels describe another parallel corpus (aligned at the word level), which is currently being compiled under the auspices of the University of Mainz. Apart from describing the make-up of the corpus and its annotation schemes, the authors discuss in detail the development of its interface, which is based on the ParaVoz package, along with the query builder.

In Chapter 7, Danuta Roszko and Roman Roszko describe two parallel Polish-Lithuanian corpora compiled at the Institute of Slavic Studies of the Polish Academy of Sciences. The authors mention the issue of obtaining permission for texts which are to be present in a publicly available corpus such as, for example, CLARIN-PL Polish-Lithuanian parallel corpus. Another interesting topic raised in the chapter is semantic tagging, which is perfectly suited for Lithuanian thanks to its clarity of formal structures.

Chapter 8 by Natalia Kotsyba contains a detailed description of the steps undertaken in compiling a Polish-Ukrainian parallel corpus and the problems associated with it. The author discusses the process of creating a pilot corpus, PolUKR, as well as the process of creating an extended version, PolUKR2, which will be used in compiling a Polish-Ukrainian dictionary.

In Chapter 9, Marianne Petrincova addresses the subject of applying parallel corpora in lexicography, in her case a Polish-Slovak parallel corpus. Petrincova's project utilizes the on-line service called Sketch Engine as a management tool and interface for her aligned data. In the chapter, the author presents ways of obtaining Slovak translation equivalents for prefixed verbs and assessing their lexicographical potential.

Chapter 10 contains a discussion on the difficulty in obtaining parallel texts and compiling a parallel corpus. The authors, Krzysztof Wołk, Emilia Rejmund and Krzysztof Marasek, suggest a new methodology for extracting parallel sentences from comparable corpora with the Yalign tool.

In Chapter 11, Silvia Bonacchi and Mariusz Mela describe their project *MCCA: Multimodal Communication: Culturological Analysis* undertaken by the University of Warsaw and the University of Saarland in Saarbrücken with the aim of conducting a culturological and suprasegmental analysis of (im)politeness based on bilingual Polish-German corpora. The corpora described in the chapter consist of spoken data in the form of both recordings and transcripts.

Chapter 12 authored by Łucja Biel contains a presentation of a project aiming at describing and analysing the Eurolect, i.e. a new variety of the Polish language used in official contexts, which emerges under the influence of the influx of translations of EU documents. According to the author, a thorough study must make use of various kinds of multilingual and monolingual tools, including English-Polish parallel and comparable corpora as well as specialized and general Polish monolingual corpora. The project has been launched at the Institute of Applied Linguistics at the University of Warsaw.

The author of Chapter 13, Monika Szela, also stresses the need to use a variety of multilingual corpora to study translated legal language. In her chapter, Szela describes the comparable and parallel corpora used in her project aiming at exploring grammatical and lexical features of translated texts.

Chapter 14 provides a different perspective as it provides a report on a study based on corpus data, the aim of which was to discover the closest Polish translation equivalents of two semantically related verbs in the Czech language. Elżbieta Kaczmarska also presents her research aiming at establishing an algorithm facilitating the extraction of equivalents of verbs representing emotions, based on their syntactic behavior.

Chapter 15 describes a pilot project launched at the Institute of Applied Linguistics at the University of Warsaw, which aims at compiling a Swedish-Polish and Polish-Swedish parallel corpus of literary texts. The authors, Ewa Gruszczyńska, Agnieszka Leńko-Szymańska and Ruprecht von Waldenfels, describe its creation as well as present tools they used. The chapter gives information on how parallel corpora can provide invaluable data in exploring lexical units expressing the emotion of "fear" in Swedish and Polish in terms of the emotional loading.

In summary, it is undeniable that the ever-growing availability of well-compiled multilingual corpora enables scholars to draw better comparisons with regard to individual languages and cultures. As a result, all language users

can profit from that, either by being able to use better glossaries, dictionaries or receiving translations of better quality. Attendees of language studies can profit from the better design of translation courses thanks to the possibility of applying language corpora in translator education. In the case of translators, although there is no real substitute for traditional tools used by them, such as dictionaries and glossaries available either in paper or digital form, parallel corpora may be applied as bilingual dictionaries providing an immediate context of terms subject to analysis. Lastly, corpora may be used by machine translation developers to enhance MT systems based on either statistics or examples.

BIBLIOGRAPHY

Gruszczyńska E., Leńko-Szymańska A. (eds.) (2016) *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora*. Warsaw: University of Warsaw.

Kenning M. M. (2010) "What are parallel and comparable corpora and how can we use them?", in: A. O'Keeffe, M. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 487–500.

McEnery T., Hardie A. (2012) *Corpus Linguistics – Method, Theory and Practice*. Cambridge: Cambridge University Press.

Dariusz Koźbiał *is a PhD student at the Faculty of Applied Linguistics, University of Warsaw. In 2015, he graduated from the Institute of Applied Linguistics, University of Warsaw, with an MA in Applied Linguistics. Currently, he is involved as co-investigator in a research project "The Eurolect: An EU Variant of Polish and its Impact on Administrative Polish". In 2016, he completed a three-month-long translation traineeship in the Polish Translation Unit at the European Parliament in Luxembourg.*