Agnieszka Leńko-Szymańska
Uniwersytet Warszawski

Ewa Gruszczyńska
Uniwersytet Warszawski

Polskojęzyczne korpusy równoległe w Polsce i za granicą

Dane korpusowe stanowią niezastąpione źródło informacji dla badaczy reprezentujących szeroki wachlarz różnych dyscyplin naukowych zajmujących się językiem, począwszy od badań czysto teoretycznych, a skończywszy na przetwarzaniu języka naturalnego. W ciągu ostatnich trzydziestu lat udostępniono naukowcom z różnych dziedzin językoznawstwa i kulturoznawstwa ogromną liczbę wielkich korpusów. Rośnie też liczba języków, które zostały udokumentowane w postaci dużych, zrównoważonych i reprezentatywnych zbiorów tekstów pisanych i mówionych, stanowiących dokładne i całościowe odzwierciedlenie języków narodowych bądź regionalnych (np. Brytyjski Korpus Narodowy, Amerykański Korpus Narodowy, Czeski Korpus Narodowy). Język polski jest także reprezentowany w co najmniej dwóch tego typu korpusach: Narodowym Korpusie Języka Polskiego i Korpusie Języka Polskiego PWN. Oprócz ogólnych zbiorów powstaje także wiele korpusów mających skromniejsze cele, ograniczonych do poszczególnych odmian języka.

Coraz częściej pojawiają się także wielojęzyczne zbiory, co poszerza pole badawcze, na którym wykorzystuje się dane korpusowe. Podobnie, jak w przypadku korpusów jednojęzycznych, korpusy wielojęzyczne są przydatne do badań w wielu dziedzinach, a szczególnie tam, gdzie dokonuje się porównań dwóch lub więcej języków i kultur. Pole dociekań wykorzystujących wielojęzyczne korpusy obejmuje badania interkulturowe, kontrastywne badania dyskursu, językoznawstwo kontrastywne, przekładoznawstwo, automatyczne wyszukiwanie ekwiwalentów i tłumaczenie maszynowe. Zasoby wielojęzyczne obejmują zarówno korpusy porównywalne, tj. zbiory tekstów w dwóch lub więcej językach, które spełniają te same kryteria, np. reprezentują ten sam gatunek, temat, typ odbiorcy itp., jak i korpusy równoległe, w których zestawia się teksty oryginalne z przekładami na jeden lub więcej języków. Każdy z tych dwóch typów korpusów jest przeznaczony do rozwiązywania innego rodzaju problemów badawczych, ale rośnie liczba badań, w których wykorzystuje się

dane zarówno z korpusów porównywalnych jak i równoległych oraz dodatkowo wzbogaca się wyniki, stosując analizę korpusów jednojęzycznych.

Wśród wielojęzycznych zbiorów cyfrowych na świecie korpusy równoległe stanowią mniejszość z co najmniej dwóch powodów. Pierwszy wynika z tego, że do korpusu mogą zostać włączone tylko takie teksty, które istnieją w dwujęzycznej wersji (tekst oryginalny i jego przekład, ewentualnie dwa przekłady na różne języki tego samego tekstu wyjściowego), co znacznie ogranicza liczbę potencjalnych tekstów nadających się do takiego zbioru. Drugi powód wynika z charakteru procesu tworzenia korpusu równoległego, który jest o wiele trudniejszy pod względem technicznym, gdyż polega między innymi na zrównoleglenia obu dwujęzycznych tekstów na poziomie akapitów, zdań, a czasami nawet słów. Wymaga także skomplikowanego interfejsu umożliwiającego użytkownikowi dwujęzyczne symultaniczne wyszukiwanie informacji. Pierwsze korpusy równoległe często zawierały język angielski (np. Angielsko-Szwedzki Korpus Równoległy utworzony w 1993r.). Wynikało to ze statusu języka angielskiego jako powszechnie używanego i większej dostępności tekstów tłumaczonych z języka angielskiego i vice versa. Jednak z upływem lat zaczęło pojawiać się coraz więcej korpusów bazujących na innych parach języków. Ważnym źródłem danych paralelnych stały się oficjalne dokumenty krajów wielojęzycznych takich jak Kanada oraz dokumenty międzynarodowe (wydawane przez takie instytucje jak Organizacja Narodów Zjednoczonych, NATO i Unia Europejska) tłumaczone na kilkanaście lub kilkadziesiąt języków narodowych.

W ciągu ostatnich lat, zarówno w Polsce jak i za granicą, rozpoczęto wiele działań związanych z budowaniem polskojęzycznych korpusów równoległych. Niektóre z nich stanowią część większych zbiorów liczących wiele języków (np. InterCorp, ParaSol), inne są ograniczone tylko do jednej pary językowej (np. Polsko-Rosyjski i Rosyjsko-Polski Korpus Równoległy, Korpus Równoległy PELCRA). Projekty te znacznie różnią się od siebie zarówno pod względem założeń i objętości, jak i rozwiązań technicznych. Tabele 1. i 2. dołączone do niniejszego rozdziału zawierają aktualną (względem daty wydania książki) listę polskojęzycznych korpusów równoległych opisanych w literaturze przedmiotu wraz z afiliacją każdego z nich, nazwiskami wykonawców oraz w miarę dostępności adresami internetowymi.

Niniejszy tom jest odpowiedzią na rosnące zainteresowanie badaczy reprezentujących różne dyscypliny, którzy zajmują się analizowaniem języka polskiego i polskiej kultury w kontekście wielojęzycznym i wielokulturowym. Jego celem jest zaprezentowanie możliwie pełnego przeglądu bieżących projektów związanych z korpusami równoległymi z udziałem języka polskiego.

Zawiera sprawozdania z tworzenia takich korpusów jak również opisy badań przeprowadzonych na ich podstawie.

Tom składa się z rozdziału wstępnego oznaczonego numerem 1 oraz czternastu kolejnych rozdziałów opisujących projekty, które już zostały zakończone, a także takie, które są na etapie realizacji. W każdym rozdziale można znaleźć szczegółowy opis konkretnego korpusu równoległego zawierającego polski komponent. Prezentowane i poddane dyskusji są zarówno budowa korpusu, anotacja oraz zastosowany interfejs. Autorzy dostarczają także wielu przykładów badań opartych na danych uzyskanych z korpusów równoległych lub badań, które są planowane. Badania te potwierdzają wielość zastosowań cyfrowych zasobów paralelnych w językoznawstwie oraz kulturoznawstwie.

Rozdział 2. autorstwa Alexandra Rosena dotyczy jednego z największych wielojęzycznych korpusów równoległych InterCorp utworzonego na Uniwersytecie Karola w Pradze. Obejmuje zbiór tekstów w 39 językach z czeskim jako najlepiej reprezentowanym językiem głównym. W rozdziale tym przedstawiono strukturę korpusu, który został też porównany z innymi tego typu zasobami. Wyjaśniono także jego status organizacyjny oraz opisano proces kompilacji. Część końcowa to przegląd różnego rodzaju zapytań możliwych do zrealizowania za pomocą korpusowego interfejsu.

InterCorp zawiera komponent polski wielkości około 80 milionów słów, co przedyskutowano szczegółowo w rozdziale 3. Milena Hebal-Jezierska, Alexandr Rosen i Elżbieta Kaczmarska przeanalizowali wyzwania związane z oczekiwaniami użytkowników, przed jakimi stają twórcy korpusu. Autorzy przedstawiają problemy użytkowników, jakie pojawiają się przy korzystaniu z czesko-polskiej części korpusu oraz rodzaje zapytań, które nie są dobrze obsługiwane, jednak przytaczają także przykłady wyszukiwań, które zwracają bogate i relewantne dane.

W rozdziale 4. Piotr Pęzik prezentuje nowy polsko-angielski korpus równoległy zwany Paralela, który jest od niedawna dostępny jako część polskiej infrastruktury CLARIN-PL –zasobów i narzędzi do obsługi tekstów w języku polskim. Autor skrótowo przedstawia zawartość korpusu i opisuje specjalnie stworzoną dla niego wyszukiwarkę. Rozważania zawarte w rozdziale skupiają się wokół możliwości zastosowania Paraleli w badaniach idiomów występujących w angielsko-polskich przekładach. Autor dochodzi do wniosku, że tylko wielkie korpusy równoległe w połączeniu z wyspecjalizowanymi narzędziami służącymi do ich przeszukiwania, mogą dostarczyć odpowiednich danych do badań nad zjawiskiem globalnej ekwiwalencji frazeologicznej w przekładzie.

Rozdział 5. autorstwa Marka Łazińskiego i Magdaleny Kuratczyk dotyczy Polsko-Rosyjskiego Korpusu Równoległego utworzonego na Uniwersytecie Warszawskim. Zawiera on 30 milionów tokenów, jednak część korpusu nie jest ogólnie dostępna ze względu na prawa autorskie. Projekt był realizowany we współpracy z dwoma dużymi podmiotami: Narodowym Korpusem Języka Polskiego i Rosyjskim Korpusem Narodowym, których zasoby tekstowe oraz zasady anotacji częściowo wykorzystano. Autorzy omawiają proces budowy korpusu ze zwróceniem szczególnej uwagi na aspekty kulturowe przy doborze tekstów, proces tagowania oraz ujednoznaczniania, a także różne możliwości wyszukiwania. W rozdziale posłużono się dwoma przykładami zastosowania korpusu w badaniach ekwiwalentów przekładowych. Rozważania kończy dyskusja na temat znaczenia projektu oraz planów na przyszłość.

W rozdziale 6. Andreas Meger, Michał Woźniak i Ruprecht von Waldenfels opisują korpus równoległy, który jest obecnie tworzony pod auspicjami Uniwersytetu Jan Gutenberga w Moguncji. Istotną cechą tego korpusu jest zrównoleglenie nie tylko na poziomie zdań, ale także na poziomie słów. Na razie mały, pilotażowy korpus liczy około 1 miliona tokenów. Podobnie jak w przypadku innych korpusów opisanych w tym tomie, autorzy omawiają szczegółowo jego budowę i anotację. Szczególną uwagę poświęcają projektowi interfejsu, który jest oparty na pakiecie ParaVoz, oryginalnie stworzonym dla projektu ParaSol. Obsługuje on teksty równoległe w formacie CWB i działa poprzez przeglądarkę internetową. Graficzna wyszukiwarka oferuje różne możliwości: od prostego wyszukiwania tokenów do skomplikowanego wyszukiwania CQP, co sprawia, że korpus jest „przyjazny" w użytkowaniu zarówno dla laików jak i dla specjalistów w przetwarzaniu języka naturalnego.

Danuta Roszko i Roman Roszko (rozdział 7.) opisują dwa polsko-litewskie korpusy równoległe utworzone w Instytucie Slawistyki Polskiej Akademii Nauk. Starszy, korpus eksperymentalny, to wewnętrzny projekt zawierający ponad 2 miliony tokenów pochodzących z tekstów beletrystycznych i 14 milionów tokenów pochodzących ze współczesnych tekstów specjalistycznych w obu językach. Drugi korpus jest tworzony pod auspicjami konsorcjum CLARIN. Będzie także zawierał teksty beletrystyczne i teksty specjalistyczne, które znajdują się w wolnym dostępie, a jego objętość w roku 2016 planowana jest na około 6 milionów tokenów. Autorzy tłumaczą fakt istnienia dwóch korpusów równoległych prawami autorskimi. Polsko-litewski korpus równoległy CLARIN będzie w wolnym dostępie, więc może zawierać tylko teksty, dla których nie jest wymagana zgoda na upublicznienie, lub dla których taka zgoda została uzyskana. To ogranicza dobór tekstów, stąd decyzja o kontynuowaniu wewnętrznego projektu korpusowego, który ma służyć badaniom przeprowadzanym w macierzystej jednostce. Ciekawą cechą obu korpusów jest ich

anotacja, która zawiera tagowanie semantyczne. W drugiej części artykułu autorzy wyjaśniają, że dzięki przejrzystości struktur formalnych języka litewskiego oraz braku dwuznaczności przy mapowaniu na płaszczyźnie formalno-funkcyjnej język litewski jest szczególnie odpowiedni do automatycznej anotacji semantycznej. Zestawienie go z językiem polskim oraz innymi językami słowiańskimi, które są mniej regularne pod wyżej wymienionym względami, może ułatwić semantyczną anotację tych języków.

Rozdział 8., autorstwa Natalii Kotsyby, poświęcony jest opisowi kompilacji polsko-ukraińskiego korpusu równoległego oraz wyzwaniom, przed jakimi stanęli jego twórcy. Podobnie jak w przypadku korpusu polsko-litewskiego i oni także zdecydowali się na budowę dwóch korpusów. Pierwszy z nich, korpus pilotażowy PolUKR, powstawał w latach 2004-2009 w Instytucie Slawistyki Polskiego Akademii Nauk (podobnie jak wyżej omówione zasoby polsko-litewskie), a następnie na wydziale „Artes Liberales" Uniwersytetu Warszawskiego. Głównym celem tego projektu było sprawdzenie procedur oraz stworzenie i przetestowanie oprogramowania koniecznego w procesie kompilacji. Szczególny nacisk został położony na wypracowanie jednolitego morfosyntaktycznego systemu znaczników do anotacji obu języków, który obecnie jest częścią międzynarodowego projektu wielojęzycznego o nazwie MULTEXT-East. Wynikiem prac pilotażowych był niewielki oportunistyczny zbiór danych liczący około 600 tysięcy słów. Autorka pracuje obecnie nad powiększeniem zasobów. PolUKR2 zawiera już 6,5 miliona tokenów, a jego docelowa wielkość przewidziana jest na 10 milionów tokenów. Ma on służyć stworzeniu wielkiego słownika polsko-ukraińskiego.

Zastosowaniu równoległych zasobów w leksykografii jest poświęcony rozdział 9. Marianna Perincova opisuje krok po kroku tworzenie Polsko-Słowackiego Korpusu Równoległego zawierającego obecnie 1,3 miliona tokenów. Prezentuje zawartość korpusu, jak również sposoby pozyskiwania tekstów od autorów, tłumaczy i wydawców. W projekcie tym zdecydowano się na wykorzystanie komercyjnego pakietu online SketchEngine jako narzędzia do obsługi korpusu, a także jako interfejsu do zrównoleglonego materiału. Jest to wszechstronny system, który oprócz obsługi różnorodnych jednojęzycznych i równoległych korpusów, umożliwia także użytkownikowi tworzenie i obsługiwanie jego własnych zasobów. W drugiej części artykułu autorka prezentuje liczne przykłady pozyskiwania słowackich ekwiwalentów przekładowych dla czasowników prefiksalnych i ocenia ich leksykograficzną przydatność.

W rozdziale 10. poruszono problem trudności związanych z pozyskiwaniem tekstów i tworzeniem korpusów równoległych, który pojawiał się także w wyżej wzmiankowanych artykułach. Autorzy Krzysztof Wołk, Emilia Rejmund

i Krzysztof Marasek zaproponowali nową metodę pozyskiwania zdań równoległych z korpusów porównywalnych. Metoda ta polega na przeszukiwaniu sieci w celu zbudowania tematycznych korpusów porównywalnych, a następnie wyszukaniu w nich zdań prawdziwie równoległych za pomocą narzędzia Yalign. Narzędzie to zostało opracowane specjalnie do tego celu. Stosuje ono miernik podobieństw zdań (cyfra od 0 do 1), który wskazuje przybliżony stopień prawdopodobieństwa tego, że dwa zdania są swoimi tłumaczeniami. Autorzy dostarczają dowodów eksperymentalnych, świadczących o skuteczności tej metody.

Silvia Bonacchi i Mariusz Mela proponują inne spojrzenie na kompilację wielojęzycznych zasobów, w tym korpusów równoległych. W rozdziale 11. opisują dwujęzyczne korpusy polsko-niemieckie tworzone w ramach projektu MCCA: Multimodal Communication: Culturological Analysis, na Uniwersytecie Warszawskim i Uniwersytecie Kraju Saary w Saarbrücken. Celem zaprezentowanego projektu jest kulturologiczna i suprasegmentalna analiza (nie)grzeczności językowej. Oryginalność zgromadzonych dla celów badawczych zasobów polega na tym, że reprezentują mówioną odmianę obu języków, a udokumentowane są w postaci nagrań oraz tekstów transkrybowanych. Autorzy opisują trzy rodzaje danych ujętych w korpusie: rozmowy diadyczne na tematy ogólne zrealizowane w warunkach studyjnych, nagrania audio odgrywanych scenek, oraz nagrania z mediów, takie jak talk show i debaty. Sporządzono szczegółowe opisy każdego typu danych wraz z ich transkrypcją, anotacją i analizą. Rozdział kończy dyskusja na temat stopnia, w jakim zebrane dane spełniają kryteria definiujące korpusy porównywalne i równoległe.

W rozdziale 12. zaprezentowano projekt, którego celem jest opis oraz analiza eurolektu – nowej odmiany polszczyzny używanej w sytuacjach oficjalnych, która wykształciła się pod wpływem tłumaczeń ogromnej liczby unijnych dokumentów. Autorka, Łucja Biel, argumentuje, że dla dogłębnej analizy stopnia zróżnicowania tej odmiany języka polskiego niezbędne są zasoby zarówno jedno- jak i wielojęzyczne takie jak angielsko-polskie korpusy równoległe i porównywalne, a także specjalistyczne oraz ogólne korpusy języka polskiego. W artykule zaprezentowano skład i strukturę zasobów, które autorka zamierza opracować w projekcie realizowanym w Instytucie Lingwistyki Stosowanej Uniwersytetu Warszawskiego.

Monika Szela jest także zainteresowana badaniem cech charakterystycznych dla języka urzędowego w tekstach tłumaczonych i także podkreśla potrzebę posługiwania się wielojęzycznymi zasobami do prowadzenia tego typu badań. W rozdziale 13. opisuje zasoby porównywalne i równoległe utworzone na użytek projektu, którego celem jest przebadanie cech gramatycznych i leksykalnych przekładów oraz ich porównanie z tekstami oryginalnymi utworzonymi

przez tzw. użytkowników natywnych w języku docelowym. Korpus równoległy, który analizuje, składa się z unijnych tekstów aktów prawnych opublikowanych w języku angielskim i polskim. Każda z części zawiera 40 milionów tokenów w postaci niezrównoleglonych plików tekstowych bez anotacji. Autorka prezentuje wyniki prac wstępnych przeprowadzonych na podstawie pozyskanych danych włącznie z analizą list frekwencyjnych i słów kluczowych, a także kilku wybranych czasowników.

Podczas gdy większość rozdziałów w tej książce prezentuje szczegółowe opisy różnych korpusów równoległych, Elżbieta Kaczmarska przedstawia w rozdziale 14. badania oparte na danych pozyskanych z takiego korpusu. Celem eksploracji jest znalezienie i przeanalizowanie polskich ekwiwalentów dwóch bliskich sobie znaczeniowo czeskich czasowników. Autorka rozpoczyna od analizy znaczeń zawartych w tradycyjnym słowniku czesko-polskim, a następnie porównuje je z cytowaniami z InterCorp (opisanym szczegółowo w niniejszym tomie w rozdziałach 1. i 2.). Autorka podejmuje także próbę automatycznego profilowania odnalezionych ekwiwalentów i dochodzi do wniosku, że dane na tym etapie nie pozwalają jeszcze na zastosowanie do analizy narzędzia Word Sketch, dlatego zdecydowała się na analizę manualną. W ostatniej części artykułu autorka umieszcza swoją analizę w szerszej perspektywie i prezentuje swoje badania, których celem jest wypracowanie algorytmu ułatwiającego pozyskiwanie ekwiwalentów przekładowych dla czasowników będących językowymi wykładnikami emocji na podstawie ich charakterystyki składniowej.

Rozdział 15. dotyczy pilotażowego projektu realizowanego w Instytucie Lingwistyki Stosowanej UW, którego celem jest utworzenie Szwedzko-Polskiego i Polsko-Szwedzkiego Korpusu Równoległego współczesnych tekstów literackich. Ewa Gruszczyńska, Agnieszka Leńko-Szymańska i Ruprecht von Waldenfels opisują szczegółowo, jak powstawał liczący 750 tysięcy tokenów minikorpus i jakie narzędzia wykorzystywane są do jego obsługi. W drugiej części zaprezentowano wyniki studiów pilotażowych dotyczących analizy jednostek leksykalnych będących wykładnikami emocji *strach/skräck* oraz ich wzajemnych tłumaczeń. Wyniki badań pilotażowych wykazały, że przekłady z języka polskiego na język szwedzki i vice versa jednostek leksykalnych związanych z tą emocją znacznie różnią się od siebie pod względem siły nacechowania emocjonalnego. Potwierdziły także przydatność korpusu równoległego do tego typu badań.

Wielość przedsięwzięć związanych z różnymi korpusami równoległymi opisanymi w niniejszym tomie oraz różnorodność zagadnień naukowych związanych z przedstawionymi projektami są dowodem, że polskojęzyczne korpusy stały się niepodważalnym źródłem danych w badaniach lingwistycznych

i kulturowych. Redaktorzy mają nadzieję, że tom ten przyczyni się do rozpowszechnienia informacji na temat istniejących projektów i pozwoli na konsolidację społeczności akademickiej zainteresowanej polskojęzycznymi korpusami równoległymi. Mamy także nadzieję, że książka ta przyczyni się do rozwoju tej stosunkowo nowej dziedziny i zachęci kolejnych naukowców do tworzenia własnych zasobów równoległych. Rosnąca liczba dobrej jakości danych wielojęzycznych dostępnych za pomocą korpusów równoległych wpłynie z pewnością nie tylko na stopień dociekliwości i dokładności porównań między językami i kulturami, ale także na jakość glosariuszy, słowników i przekładów, które trafiają do odbiorców.

Agnieszka Leńko-Szymańska
Uniwersytet Warszawski

Ewa Gruszczyńska
Uniwersytet Warszawski

Polish-language parallel corpora in Poland and abroad

Corpus data constitute an indispensable source of information for scholars from a whole range of language-related disciplines, from purely theoretical studies to Natural Language Processing. In the last thirty years a multitude of large corpora have become available to researchers form different branches of linguistics and culture studies. An increasing number of world languages are being captured in large, balanced and representative collections of written and spoken text, some making claims to being an accurate reflection of a national or regional language as a whole (e.g. *British National Corpus*, *American National Corpus* and *Czech National Corpus*). Polish also has such a representation in at least two corpora: *National Corpus of Polish* and *PWN Corpus*. In addition to general collections, there is also a whole array of corpora compiled with more modest aims of representing a particular language variety.

More recently, multilingual language collections have become available, thus broadening the scope of research supported by corpus data. As in the case of monolingual resources, multilingual corpora are useful to researchers from the whole range of disciplines, interested in comparing and contrasting two or more languages and cultures. The fields of inquiry which benefit from multilingual corpus data include intercultural studies, contrastive discourse studies, contrastive linguistics, translation studies, automatic extraction of equivalents or machine translation. Multilingual resources include either comparable corpora, that is collections of texts in two or more languages which match one another on the number criteria such as genre, topic, audience etc., and parallel corpora which encompass original texts and their translation(s) into one or more languages. Each of these two different types of multilingual corpora is more suitable for addressing different types of research questions, but a growing number of projects draw their data from both comparable and parallel corpora and supplement their results with analyses of monolingual corpora.

Among multilingual resources around the world parallel corpora are less numerous for at least two reasons. First, texts to be included in them have to exist in at least a bilingual version (an original and its translation, or translations of the same text from another source language), thus drastically limiting the number of texts eligible for inclusion. Second, the process of compilation is technically more demanding as it involves aligning the bilingual content at the text, paragraph, sentence or sometimes even word level. It also requires a complex interface enabling users to query and display the bilingual information simultaneously. First parallel corpora often included English in their language pairs (e.g. *English-Swedish Parallel Corpus* launched in 1993). This was motivated by the status of English as the global language and consequently a larger availability of texts translated from and into English. However, with years an increasing number of corpora including other language pairs have started to emerge. An important source of parallel data have recently been official documents from multilingual countries such as Canada or international documents (issued by such official bodies as the United Nations, NATO or the European Union) translated into several national languages.

In recent years several ventures involving a compilation of parallel corpora including Polish have been launched in Poland and abroad. Some of them constitute sections of larger collections encompassing several languages (e.g. *InterCorp*, *ParaSol*), others are limited to one language pair (e.g. *Polish – Russian and Russian-Polish Parallel Corpus*, *PELCRA Parallel Corpus*). These projects vary greatly in their objectives and scope as well as in their technical solutions. Tables 1 and 2 at the end of this chapter contain an up-to-date (as of the publication date) list of parallel corpora including a Polish component described in the literature, together with their mother institutions, compilers' names and website addresses, if available.

This volume is an answer to a growing interest of researchers from various disciplines in analysing Polish language and culture in a multilingual and multicultural context. Its aim is to provide a fairly comprehensive review of current projects linked to parallel corpora with a Polish component. It includes reports on activities related to the compilation of such corpora as well as descriptions of studies based on Polish-language parallel data.

The volume consists of this introductory chapter (Chapter 1) and 14 chapters describing a variety of projects which have already been completed or which are currently under development. Each paper offers a detailed description of a parallel corpus including a Polish component. The composition of the corpora, their annotation schemes and query

interfaces are presented and discussed. The authors also present examples of studies based on parallel data which have been conducted or are planned to be conducted. These studies attest to the multitude of application of parallel resources in linguistic and cultural research.

Chapter 2 by Alexandr Rosen describes *InterCorp*, one of the largest multilingual parallel corpora, compiled at Charles University in Prague. This collection comprises texts in 39 languages, with Czech being its best-represented and pivot language. The chapter presents a detailed makeup of the corpus, and compares it to other resources of this kind. It also explains its organisational status and describes the compilation process. Finally, the paper briefly reviews the types of queries facilitated by the corpus interface.

InterCorp includes a sizeable Polish component of almost 80 million words, which is discussed in detail in Chapter 3. Milena Hebal-Jezierska, Aleksandr Rosen and Elżbieta Kaczmarska analyse the challenges facing the corpus compilers related to meeting users' needs. The authors demonstrate the problems users come across when using the Czech-Polish section of the corpus and the kinds of queries which are not well addressed by the corpus data. However, the chapter also presents examples of searchers which return rich and relevant data.

Piotr Pęzik (Chapter 4) presents a new parallel Polish-English corpus called *Paralela*, which has recently become available as part of the CLARIN-PL infrastructure of Polish language tools and resources. The author summarizes the contents of the corpus and describes its dedicated search engine. The chapter focuses on the applicability of Paralela in the study of idiomaticity in English-Polish translations. The author concludes that only large parallel corpora, in combination with specialized search tools, provide sufficient data for investigating the phenomenon of global phraseological equivalence in translation.

Chapter 5 by Marek Łaziński and Magdalena Kuratczyk presents a Polish-Russian parallel corpus compiled at the University of Warsaw. The collection consists of 30 million tokens but not all of it is publically available due to copyright restrictions. The project was run in cooperation with two large national corpora: *National Corpus of Polish* and *Russian National Corpus*, using some of their textual resources and the annotation schemes. The authors discuss the composition of the collection with special attention given to the cultural aspects governing the choice of texts included in it. The tagging and disambiguation processes are also described together with various search options. The chapter offers two

examples of applications of this corpus for research on translation equivalents and it ends with the discussion of the significance of the project and the outlook for the future.

In Chapter 6 Andreas Meger, Michał Woźniak and Ruprecht von Waldenfels describe another parallel corpus which is currently being compiled under the auspices of the University of Mainz. The interesting feature of this resource is that it is aligned not only at the sentence level but also at the word level. A small pilot corpus of 1 million tokens has already been completed. As with other text collections described in this volume, the authors provide the details of its composition and annotation schemes. Special attention in this chapter is given to the development of the interface which is based on the ParaVoz package, originally created for the *ParaSol* project. It works with parallel texts in CWB-format and operates through a web browser. The graphical query builder offers different options: from simplest token searches to complex CQP queries, which makes the corpus a user-friendly resource for both laymen as well as NLP specialists.

Danuta Roszko and Roman Roszko (Chapter 7) describe two parallel corpora of Polish and Lithuanian developed at the Institute of Slavic Studies, Polish Academy of Sciences. The earlier one, the experimental corpus, is an in-house project containing over 2 million tokens of fiction and 14 million tokens of contemporary specialist texts in the two languages. The other corpus is being compiled under the auspices of the CLARIN-PL consortium. It will also include fiction and specialist texts from the public domain and it is planned to reach the size of 6 million tokens in 2016. The authors explain the necessity of having two parallel corpora by copyright issues. The CLARIN-PL Polish-Lithuanian parallel corpus will be publically available, thus it can only contain texts for which permissions are not necessary or have been obtained. This limits the choice of texts, hence the decision was made to continue the in-house compilation project, which will only be used for internal research. An interesting feature of these two corpora is its annotation which will include semantic tagging. In the second part of the article the authors explain that due to the clarity of formal structures in Lithuanian and a lack of ambiguity in the form-function mappings Lithuanian is particularly suitable for automatic semantic annotation. Juxtaposing it with Polish and other Slavic languages, which are less regular in these respects, can facilitate automatic semantic annotation of these languages.

Chapter 8 by Natalia Kotsyba describes in detail the steps in building a Polish-Ukrainian parallel corpus and the challenges that the

compilers faced during this process. As in the case of the Polish-Lithuanian resources, two collections were created. The pilot corpus, *PolUKR*, was also compiled at the Institute of Slavic Studies, Polish Academy of Sciences in 2004-2009, and later at the faculty "Artes Liberales", University of Warsaw. The primary aim of this project was piloting the procedures and developing and testing software needed for the compilation process. Special attention was given to creating a morphosyntactic tagset for a uniform annotation of both languages, which is now part of the international multilingual project called MULTEXT-East. The result of the pilot project was a small and opportunistic resource of 600 thousand words. At the moment Kotsyba is working on extending the collection. *PolUKR2* already contains 6,5 million tokens and is planned to reach at least 10 million tokens. It will be used for compiling a great Polish-Ukrainian dictionary.

The application of a parallel collection in lexicography is addressed in Chapter 9. Marianne Petrincova reports on the subsequent steps in the creation of a Polish-Slovak parallel corpus containing over 1.3 million tokens. The contents of the corpus as well as ways of obtaining the data from authors, translators and publishers are presented. In this project the compiler decided to use the on-line service Sketch Engine as a management tool and an interface for her aligned data. It is a versatile on-line system which in addition to providing access to a variety of monolingual and parallel corpora allows users to upload and work with their own data. In the second part of the paper Petrincova presents several examples of obtaining Slovak translation equivalents for prefixed verbs and assessing their lexicographical potential.

Chapter 10 addresses the problem of difficulty in obtaining parallel texts and building a parallel corpus, already mentioned above. Krzysztof Wołk, Emilia Rejmund and Krzysztof Marasek propose a new methodology for extracting parallel sentences from comparable corpora. The new method involves first web crawling for compiling topic-aligned comparable corpora and then extracting from them truly parallel sentences with the help of Yalign tool. The tool was designed especially for his purpose. It applies a sentence similarity metric that produces a rough estimate (a number between 0 and 1) of the likelihood of two sentences being a translation of each other. The authors provide experimental evidence for a satisfactory performance of their method.

Silvia Bonacchi and Mariusz Mela offer a different perspective on the compilation of multilingual resources, including parallel corpora. In Chapter 11 they describe the bilingual Polish-German corpora they

compiled within the project *MCCA: Multimodal Communication: Culturological Analysis*, which was undertaken by the University of Warsaw and University of Saarland in Saarbrücken. Its aim is a culturological and suprasegmental analysis of (im)politeness. The originality of the collection created in the framework of this project lies in the fact that it consists of spoken data in the two languages in the form of both recordings and transcripts. The authors describe three types of data that were included in the corpus: dyadic conversations on topics of general interest recorded in a studio, audio recordings of acted situations, and media recordings such as talk shows and debates. The detailed description of each text type as well as of their transcription, annotation and analysis are provided. The authors finish the chapter with a discussion of the extent to which their data meet the criteria of comparable and parallel corpora.

Chapter 12 presents a project aimed at description and analysis of Eurolect, a new variety of Polish used in official contexts, which is emerging under the influence of translations of large number of EU documents. Łucja Biel argues that a thorough analysis of this language variety requires access to different kinds of multilingual and monolingual resources including English-Polish parallel and comparable corpora and specialised and general Polish monolingual corpora. The author presents the architecture of these resources which she intends to compile in the framework of the project just launched at the Institute of Applied Linguistics, University of Warsaw.

Monika Szela is also interested in research into the characteristics of the translated legal language and she also recognizes a need for a variety of multilingual resources necessary for this purpose. In Chapter 13 she describes comparable and parallel collections she compiled within her project whose aim is to explore the grammatical and lexical features of translated texts and compare them to texts produced originally by native speakers of the target language. Her parallel corpus consists of legal acts of the European Union published in English and Polish. Each of the two sections contains 40 million tokens. The corpus has the form of plain text files without annotation and alignment. Szela presents results of initial analyses of the collected data including analyses of frequency lists and keyword lists as well as of a few hand-picked verb forms.

While most of the chapters in this book offer detailed descriptions of various parallel resources, Elżbieta Kaczmarska' paper (Chapter 14) reports on a study based on the data drawn from such a corpus. The aim of the study was to find and examine the closest Polish translation

equivalents of two semantically related verbs in Czech. The author starts with the analysis of the equivalents found in a traditional Czech-Polish dictionary and then compares her results with the citations from *InterCorp*, described in detail in Chapters 1 and 2. The author also attempts to automatically profile the located equivalents and concludes that the data is not sufficient for applying the Word Sketch analysis, thus instead she conducts this analysis manually. In the last section of the chapter Kaczmarska puts her analysis in a lager perspective by presenting her research aiming at establishing an algorithm facilitating extration of translation equivalents of verbs being linguistic representations of emotions based on their syntactic behaviour.

Chapter 15 describes a pilot project launched at the Institute of Applied Linguistics, University of Warsaw and aimed at compiling the Swedish-Polish and Polish-Swedish parallel corpus of literary texts. Gruszczyńska, Leńko-Szymańska and von Waldenfels describe in detail the subsequent stages involved in the creation of a 750-thousand-token mini-corpus and the tools used for this purpose. The second part of the chapter presents the results of a pilot study into the expression of the emotion of 'fear' in the two languages. The results of this pilot study demonstrate that translations of lexical units connected with this emotion from Polish into Swedish and vice versa differ from each other in the intensity of emotional loading. They also confirm that the parallel corpus provides invaluable data in exploring this issue.

The multitude of corpus compilation ventures described in this volume as well as the variety of research questions addressed by these projects testify that Polish-language parallel corpora are becoming a well-established source of data in linguistic and cultural investigations. The editors hope that the volume will help disseminate the information about the existing projects and it will be a step forward in consolidating the research community interested in the analysis of Polish parallel data. It is also hoped that the volume will contribute to the development of this relatively new area of exploration and encourage more researches to engage in the compilation of their own resources. The growing availability of good quality multilingual corpus data will certainly have its influence not only on the depth and accuracy of comparisons between languages and cultures but will also be reflected in the excellence of glossaries, dictionaries and translations reaching their end-users.

Polskojęzyczne korpusy równoległe

Polish-language parallel corpora

Tabela 1. Korpusy dwujęzyczne i trójjęzyczne / Table 1. Bilingual and trilingual corpora

Nazwa i witryna projektu **Project name and website**	**Instytucja macierzysta** **Home institution**	**Kierownik projektu** **Project director**
PARALELA Angielsko-polskie teksty równoległe z zawansowana wyszukiwarką Polish-English parallel texts with an advanced search engine http://paralela.clarin-pl.eu/	CLARIN-PL	Pior Pęzik piotr.pezik@gmail.com
PELCRA (Polish and English Language Corpora for Research and Application) Korpusy równoległe PELCRA PELCRA parallel corpora http://pelcra.pl/new/	Instytut Anglistyki, Uniwersytet Łódzki	Barbara Lewandowska-Tomaszczyk, blt@uni.lodz.pl; Piotr Pęzik piotr.pezik@gmail.com
Polsko-Rosyjski i Rosyjsko-Polski Korpus Równoległy Polish-Russian and Russian-Polish Parallel Corpus http://pol-ros.polon.uw.edu.pl/	Instytut Języka Polskiego, Uniwersytet Warszawski Instytut Rusycystyki, Uniwersytet Warszawski	Marek Łaziński m.lazinski@uw.edu.pl Magdalena Kuratczyk m.kuratczyk@uw.edu.pl
Bułgarsko-Polsko-Rosyjski Korpus Równoległy Bulgarian-Polish-Russian Parallel Corpus	Instytut Slawistyki, Polska Akademia Nauk	Violetta Koseska amaz1312@gmail.com
EKorpPL-LT; KorpPL-LT_CLARIN Polsko-litewskie korpusy równoległe Polish-Lithuanian parallel corpora	Instytut Slawistyki, Polska Akademia Nauk	Roman Roszko roman.roszko@ispan.waw.pl

PolUKR; PolUKR2 Polsko-Ukraiński Korpus Równoległy Polish-Ukrainian Parallel Corpus http://domeczek.pl/~polukr	Instytut Slawistyki, Polska Akademia Nauk	Natalia Kotsyba natalia.kocyba@ipipan.waw.pl
Polsko-Słowacki Korpus Równoległy Polish-Slovak Parallel Corpus	Univerzita Palackého v Olomouci	Marianna Petrincová m_petrincova@yahoo.com
Polsko-Niemiecki i Niemiecko-Polski Korpus Równoległy Polish-German and German-Polish Parallel Corpus http://www.fb06.uni-mainz.de/polnisch/331.php	Johannes Gutenberg- Universität Mainz Uniwersytet Warszawski	Andreas Meger meger@uni-mainz.de Marek Łaziński m.lazinski@uw.edu.pl
Polsko-Węgierski i Węgiersko-Polski Korpus Równoległy Polish-Hungarian and Hungrian-Polish Parallel Corpus	Instytut Slawistyki, Uniwersytet w Pécsu	Robert Wołosz robert.wolosz@gmail.com
Polsko-Szwedzki i Szwedzko-Polski Korpus Równoległy Polish-Swedish and Swedish-Polish Parallel Corpus	Instytut Lingwistyki Stosowanej, Uniwersytet Warszawski	Ewa Gruszczyńska e.gruszczynska@uw.edu.pl
Polsko-Włoski Korpus Równoległy Polish-Italian Parallel Corpus	Katedra Językoznawstwa Ogólnego i Indoeuropejskiego, Uniwersytet Jagielloński	Dorota Sieroń dorota.sieron@uj.edu.pl
PL EUROLECT Korpusy równoległe i porównywalne polskiego i angielskiego unijnego języka urzędowego (eurolektu) Parallel and comparable corpora of Polish and English EU administrative language (Eurolect)	Instytut Lingwistyki Stosowanej, Uniwersytet Warszawski	Łucja Biel l.biel@uw.edu.pl
KRAN i KRPL Polsko-Angielski Korpus Równoległy Tekstów Prawnych Polish-English Parallel Corpora of Legal Texts	Wyższa Szkoła Filologiczna we Wrocławiu	Monika Szela monikaszela@gmail.com
MCCA (Multimodal Communication: Culturological Analysis) Niemieckie i polskie korpusy równoległe i porównywalne języka mówionego German and Polish parallel corpora of spoken language	Instytut Komunikacji Specjalistycznej i Interkulturowej, Uniwersytet Warszawski	Silvia Bonacchi s.bonacchi@uw.edu.pl

Tabela 2. Korpusy wielojęzyczne / Table 2. Multilingual corpora

Nazwa i witryna projektu **Project name and website**	**Instytucja macierzysta** **Home institution**	**Kierownik projektu** **Project director**
InterCorp Wielojęzyczny korpus równoległy Multilingual parallel corpus http://ucnk.ff.cuni.cz/intercorp/?lang=en	Filozofická fakulta, Univerzita Karlova v Praze	Alexandr Rosen alexandr.rosen@ff.cuni.cz; Michal Křen michal.kren@ff.cuni.cz
ParaSol Korpus równoległy zawierający wiele języków (głównie słowiańskich) Parallel corpus including multiple (mainly Slavic) languages http://www.slavist.de/	Humboldt-Universität zu Berlin	Ruprecht von Waldenfels ruprecht.waldenfels@gmail.com; Roland Meyer roland.meyer @sprachlit.uni-regensburg.de
Słowiański Korpus Równoległy Uniwersytetu w Amsterdamie Amsterdam Slavic Parallel Aligned Corpus http://www.uva.nl/over-de-uva/organisatie/medewerkers/content/b/a/a.a.barentsen/a.a.barentsen.html	Faculteit der Geesteswetenschappen, Capaciteitsgroep Slavische talen en culturen, Universiteit van Amsterdam	A.A. Barentsen A.A.Barentsen@uva.nl
Opus – an open source parallel corpus Zbiór wielojęzycznych korpusów równoległych przetłumaczonych tekstów dostępnych w Internecie A collection of multilingual parallel corpora of translated texts from the web http://opus.lingfil.uu.se/		Jörg Tiedemann jorg.tiedemann@helsinki.fi

JRC-Acquis; DGT-Acquis; DCEPT Wielojęzyczne korpusy równoległe tekstów prawnych UE (dostępne także przez OPUS) Multilingual parallel corpus of EU legislative texts (also available through OPUS) https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis https://ec.europa.eu/jrc/en/language-technologies/dgt-acquis https://ec.europa.eu/jrc/en/language-technologies/dcep	European Commission, Joint Research Centre	
Europarl Korpus równoległy transkrypcji obrad Parlamentu Europejskiego 1996-2011 (dostępny także przez OPUS) European Parliament Proceedings Parallel Corpus 1996-2011 (also available through OPUS) http://www.statmt.org/europarl/	Chair of Machine Translation, School of Informatics, University of Edinburgh	Philipp Koehn pkoehn@inf.ed.ac.uk
OpenSubtitles Corpus Zbiór przetłumaczonych napisów filmowych w wielu językach (dostępny także przez OPUS) A collection of translated movie subtitles in multiple languages (also available through OPUS) http://www.opensubtitles.org/	opensubtitles.org	admin@ opensubtitles.org
Korpus Równoległy Wykładów TED (dostępny także przez OPUS) TED Talk Parallel Corpus (also available through OPUS) http://www.casmacat.eu/corpus/ted2013.html	CASMACAT Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation	Mauro Cettolo

MULTEXT-East „1984” 4.0 Korpus równoległy tekstu powieści G. Orwella „1984” Parallel Corpus of Orwell's novel *1984* http://nl.ijs.si/ME/V4/ https://www.clarin.si/repository/xmlui/handle/11356/1043	Institut „Jožef Stefan”, Ljubljana	
Verne80days Wielojęzyczny korpus powieść J. Verne „W 80 dni dookoła świata” Multilingual edition of Verne's novel *Around the World in 80 Days* http://www.korpus.matf.bg.ac.rs/Verne80days/	META CESAR Multilingual Europe Technology Alliance Central and South-east European Resources	Duško Vitas vitas@matf.bg.ac.rs