Beyond Philology No. 14/1, 2017
ISSN 1732-1220, eISSN 2451-1498

CLARIN-PL workshops in Łódź, 3-4 February 2017

JOANNA REDZIMSKA
DANUTA STANULEWICZ
MAGDALENA WAWRZYNIAK-ŚLIWSKA

Received 12 March 2017,
accepted 8 May 2017.

1. CLARIN-PL

CLARIN-PL is part of CLARIN (Common Language Resources and Technology Infrastructure), a European research infrastructure whose aim is to facilitate work with large collections of texts in the areas of the humanities and social sciences (see <http://clarin-pl.eu>, <http://clarin-pl.eu/en/home-page/>). CLARIN-PL's partners are the following institutions in Poland:

- Grupa Technologii Językowych G4.19 Politechniki Wrocławskiej, Wrocław University of Science and Technology;
- The Linguistic Engineering Group, Polish Academy of Sciences;
- Polish-Japanese Academy of Information Technology, Institute of Computer Science;
- Instytut Slawistyki PAN, Polish Academy of Sciences;
- University of Łódź;
- University of Wrocław.

CLARIN-PL hosts a number of language resources, including, *inter alia*:

- Paralela, a Polish-English parallel corpus, available at <http: //paralela.clarin-pl.eu/>;
- ChronoPress: Chronologiczny Korpus Polskich Tekstów Prasowych (1945-1954), a corpus of Polish press texts 1945-1954, available at <http://chronopress.clarin-pl.eu>;
- Słowa Dnia (Words of the Day), the most frequent words in the Polish press, available at <http://slowadnia.clarin-pl.eu/#/ default/1060>;
- Słowosieć (Pl Wordnet), available at <http://plwordnet.pwr. wroc.pl/wordnet/>;
- Walenty, a valency dictionary of the Polish language, available at <http://walenty.ipipan.waw.pl/>;
- Spokes, conversational data resources, available at <http:// spokes.clarin-pl.eu/>,

as well as tools for analyzing language, including, *inter alia*,

- Nowy Morfeusz, a tool used for morphological analysis, available at <http://sgjp.pl/morfeusz/>;
- Inforex, a system used for editing of annotated corpora, available at <https://inforex.clarin-pl.eu/>;
- Mapa Literacka, used to recognize references of geographical names, available at <http://litmap.clarin-pl.eu/>;
- Transkrypcja Fonetyczna, a tool used for phonetic transcription, available at <http://mowa.clarin-pl.eu/transcriber/>;
- Chunker, a tool used for syntactic analysis, available at <http://ws.clarin-pl.eu/chunker.shtml>;
- Mowa, a speech processing tool, available at <http://ws. clarin-pl.eu/chunker.shtml>
- Parser, used to analize the Polish language, available at <http://ws.clarin-pl.eu/parser.shtml>.

2. The lectures and workshops

On 3–4 February 2017, Institute of English, University of Łódź, hosted the workshops "CLARIN-PL w praktyce badawczej: Cyfrowe narzędzia do analizy języka w naukach humanistycznych i społecznych [CLARIN-PL in research practice: IT tools used in language analysis in the humanities and social sciences]".

The participants had an opportunity to attend a number of workshops and lectures offered by specialists in natural language processing and IT technology.

2.1. Day 1

On the first day the participants were warmly welcomed by the hosts, Maciej Piasecki and Piotr Pęzik. Then they were introduced to the topic of the workshops and listened to two lectures presenting the possibilities offered by CLARIN-PL:

(1) "CLARIN – infrastruktura naukowa technologii językowych – wprowadzenie [CLARIN – the research infrastructure of language technologies: An introduction" by Maciej Piasecki (Wrocław University of Science and Technology);

(2) "Repozytorium Centrum Technologii Językowych: deponowanie i upowszechnianie zasobów i narzędzi językowych, gromadzenie korpusów tekstowych [The repository of the Centre of Language Technologies: The storing and sharing of language tools and resources, the collecting of text corpora]" by Marcin Pol, Tomasz Walkowiak and Marcin Oleksy (Wrocław University of Science and Technology).

The lectures and workshops that followed were divided into two sections. In the first section, the presenters concentrated on language corpora – their creating, managing and annotating as well as on analyzing corpus texts:

(3) "Zarządzanie i anotowanie korpusów tekstowych w systemie Inforex [Managing and annotating text corpora in Inforex]" by Marcin Oleksy and Michał Marcińczuk (Wrocław University of Science and Technology);

(4) "Tworzenie przeszukiwalnych korpusów języka polskiego za pomocą Korpusomatu [Creating searchable corpora of the Polish language using Korpusomat]" by Łukasz Kobyliński, Witold Kieraś and Maciej Ogrodniczuk (Institute of Computer

Science of the Polish Academy of Sciences);

(5) "Badanie prasy narzędziami Clarin-PL. Przykład korpusu ChronoPress [Investigating the press with the tools of Clarin-pl: The case of the ChronoPress corpus]" by Adam Pawłowski (University of Wrocław);

(6) "WebSty – otwarty sieciowy system do analizy stylometrycznej tekstu [WebSty: an open source web system for stylometric text analysis]" by Maciej Piasecki and Tomasz Walkowiak (Wrocław University of Science and Technology).

The second section was devoted to parsing and speech processing tools as well as to conversational corpora:

(7) "Parsowanie składniowe LFG i bank struktur LFG [LFG syntactic parsing and the LFG bank of structures]" by Agnieszka Patejuk and Adam Przepiórkowski (Institute of Computer Science of the Polish Academy of Sciences);

(8) "Parsowanie semantyczne wypowiedzi w języku polskim z użyciem parsera ENIAM [Semantic parsing of Polish utterances using the ENIAM parser]" by Wojciech Jaworski (Institute of Computer Science of the Polish Academy of Sciences);

(9) "Narzędzia do przetwarzania mowy [Speech processing tools]" by Danijel Koržinek and Łukasz Brocki (Polish-Japanese Academy of Information Technology);

(10) "Dyskurs w czasie rzeczywistym, czyli korpusy konwersacyjne PL i EN [Real time discourse or Polish and English conversational corpora]" by Piotr Pęzik (University of Łódź).

2.2. Day 2

On the second day the participants were again offered lectures and workshops in two sections. Section 1 concerned the use of Wordnet and of a valency dictionary as well as information extraction and automantic semantic analysis:

(11) "Słowosieć 3.0 – leksykalna sieć semantyczna języka polskiego i jej zastosowania [Słowosieć 3.0 / Wordnet 3.0 – lexical semantic net of the Polish language and its use]" by Maciej Piasecki and Agnieszka Dziob (Wrocław University of Science and Technology);

(12) "Elektroniczny słownik walencyjny Walenty [The electronic valency dictionary Walenty]" by Elżbieta Hajnicz and Tomasz Bartosiak (Institute of Computer Science of the Polish Academy of Sciences);

(13) "Narzędzia do ekstrakcji informacji z tekstu [Tools for information extraction from text]" by Michał Marcińczuk (Wrocław University of Science and Technology);

(14) "Automatyczna analiza semantyczna zbiorów tekstów na poziomach leksykalnym i fragmentów tekstu (WSD, statystyki znaczeń, semantyka dystrybucyjna – narzędzia i produkty, relacje semantyczne, klasyfikacja semantyczna, tagowanie semantyczne, przykłady aplikacji, np. Mapa Literacka [Automatic semantic analysis of text collections at lexical and text fragments levels (WSD, sense statistics, distributional semantics – tools and products, sematic relationships, semantic classification, semantic tagging, selected applications, e.g. Literary Map)]" by Paweł Kędzia, Michał Marcińczuk, Maciej Piasecki and Tomasz Walkowiak (Wrocław University of Science and Technology).

In section 2 the participants had an opportunity to learn how to investigate fixed phrases in corpora and use bilingual – Polish-English – Wordnet as well as how to extract collocations and specialist terms:

(15) "Badania frazeologii na podstawie korpusów referencyjnych (NKJP, BNC) i równoległych (Paralela) [Investigating fixed phrases with the use of reference corpora (NKJP, BNC) and parallel corpora (Paralela)]" by Piotr Pęzik (University of Łódź);

(16) "Dwujęzyczna Słowosieć – możliwości wykorzystania w pracy tłumacza i analizie porównawczej [Bilingual SłowoSieć / Wordnet – possible applications in translation and comparative analysis]" by Ewa Rudnicka (Wrocław University of Science and Technology);

(17) “Narzędzia do automatycznego wydobywania słowników kolokacji i do oceny połączeń wyrazowych [Tools for automatic extraction of collocation dictionaries and for evaluation of phrases]” by Agnieszka Dziob, Marek Maziarz and Maciej Piasecki (Wrocław University of Science and Technology);

(18) “Ekstrakcja terminologii z korpusów dziedzinowych [Terminology extraction from specialist corpora]” by Małgorzata Marciniak (Institute of Computer Science of the Polish Academy of Sciences).

The presentations are available in the pdf format on the CLARIN-PL website: <http://clarin-pl.eu/pl/materialy-iv-cykl-wykladow-i-warsztatow-clarin-pl/>.

3. A final word

The lectures and workshops offered by CLARIN-PL in Łódź provided their participants with an excellent opportunity to get acquainted with the different language resources and tools freely available on its website. Undoubtedly, these resources and tools are nowadays indispensable in language study.

Joanna Redzimska
Instytut Anglistyki i Amerykanistyki
Uniwersytet Gdański
ul. Wita Stwosza 51
80-308 Gdańsk
Poland
joared@wp.pl

Danuta Stanulewicz
Instytut Anglistyki i Amerykanistyki
Uniwersytet Gdański
ul. Wita Stwosza 51
80-308 Gdańsk
Poland
danuta.stanulewicz@gmail.com

Magdalena Wawrzyniak-Śliwska
Instytut Anglistyki i Amerykanistyki
Uniwersytet Gdański
ul. Wita Stwosza 51
80-308 Gdańsk
Poland
magdalenaws@ug.edu.pl