Józef Jarosz
ORCID: 0000-0002-7820-667X
Uniwersytet Wrocławski, Wrocław

*Linguistische Treffen in Wrocław*, Vol. 19, 2021 (I)
ISSN: 2084–3062, e-ISSN: 2657–5647
https://doi.org/10.23817/lingtreff.19-38
S. 537–541

# Sprawozdanie z wirtualnych warsztatów „CLARIN-PL w praktyce badawczej” 19–20.11.2020, Wrocław (Politechnika Wrocławska)

**Author**: Józef Jarosz, University of Wrocław, pl. Nankiera 15b, 50-140, Wrocław, Poland, e-mail: jozef.jarosz@uwr.edu.pl

**Received**: 1.12.2020 **Accepted**: 14.12.2020

W dniach 19–20. listopada 2020 r. odbyły się warsztaty z cyklu „CLARIN-PL w praktyce badawczej”, które po raz kolejny zorganizowane zostały przez Centrum Technologii Językowych CLARIN-PL oraz PolLinguaTec – Centrum Wiedzy CLARIN Technologii Językowej dla Języka Polskiego (Politechnika Wrocławska). Celem warsztatów, które adresowane były do pracowników naukowych i doktorantów, chcących w swoich pracach badawczych oraz działalności dydaktycznej wykorzystać technologię przetwarzania języka naturalnego, była popularyzacja narzędzi oraz prezentacja ich możliwości. Animatorami oraz moderatorami webinarium byli J. Wieczorek oraz P. Piasecki (Politechnika Wrocławska). Wydarzenie, w którym wzięło udział ok. 280 uczestników, odbyło się całkowicie w formie zdalnej.

Punkt ciężkości webinarium spoczywał na prezentacji wybranych narzędzi do przetwarzania języków naturalnych, które można wykorzystać w naukach humanistycznych i społecznych, a które dostępne są na platformie CLARIN. Zdecydowana większość referatów bazowała na badaniach dotyczących języka polskiego, jednak obecne były odniesienia do języka angielskiego, jidysz, języków słowiańskich i bałtyckich. Ponadto uczestnicy otrzymali wskazówki odnośnie innych języków, o które pytali na czacie w trakcie wystąpień zaproszonych referentów oraz podczas sesji plenarnych i dyskusji.

Obszerny i różnorodny program składający się ze starannie dobranych tematów zawierał zarówno wystąpienia o ogólnej treści, instruktażowe referaty pokazujące funkcjonalności wybranych instrumentów CLARIN i wreszcie jednostkowe przykłady konkretnego zastosowania na przykładzie bieżących lub zakończonych projektów.

Program dwudniowych warsztatów obejmował 22 referaty, które zostały ułożone w dwa ciągi (potok A i potok B) jednocześnie odbywających się wystąpień, które uzupełniały wykłady i spotkania plenarne dla wszystkich uczestników na jednym kanale.

Otwierające wykłady plenarne, które prowadzili gospodarze, tj. pracownicy Politechniki Wrocławskiej J. Wieczorek („Model współpracy i sposoby korzystania z infrastruktury") oraz M. Piasecki („Syntetyczny przegląd narzędzi i usług") adresowane były do uczestników, którzy po raz pierwszy zetknęli się z przedmiotową problematyką, i miały na celu wprowadzenie w zagadnienie przetwarzania korpusów językowych. Prelegenci przedstawili w zwięzły sposób ogólnoeuropejską infrastrukturę naukową CLARIN (Common Language Resources & Technology Infrastructure), która umożliwia badaczom z dziedziny nauk humanistycznych i społecznych zastosowanie szeregu narzędzi do pracy z bardzo dużymi zbiorami tekstów. Omówili ponadto zakres usług, sposób nawiązania współpracy z CLARIN-PL oraz dokonali zwięzłego przeglądu narzędzi i usług, sygnalizując tym samym profil tematyczny późniejszych referatów.

W ramach potoku A odbyły się trzy wystąpienia adresowane do różnych odbiorców. Przykład wykorzystania infrastruktury CLARIN w naukach ekonomicznych zaprezentował i szczegółowo przeanalizował J. Woźniczka (Uniwersytet Ekonomiczny, Wrocław) w wykładzie pt. „Obraz marketingu w mediach internetowych". Jego treścią były m.in. badanie wizerunku i obrazu świata, analiza dyskursu medialnego oraz analiza wydźwięku emocjonalnego tekstu. Propozycje zastosowania oraz możliwości techniczne ChronoCorpusu przedstawił A. Pawłowski (Uniwersytet Wrocławski). Omówione funkcjonalności pokazały, jak elektroniczny korpus prasowy można wykorzystać w diachronicznym badaniu dyskursu medialnego, i w jaki sposób można uzupełnić badania diachroniczne narzędziami do przetwarzania języków np. poprzez tworzenie statystyk oraz list frekwencyjnych słów i fraz dla tekstów w języku polskim. Ponadto referent w sposób przystępny i poglądowy poinstruował uczestników, jak tworzyć wizualizacje danych i jak je interpretować. Tworzenie i zarządzanie korpusami było przedmiotem wystąpienia przygotowanego przez M. Oleksego, W. Kierasia i Ł. Kobylińskiego (Instytut Podstaw Informatyki PAN). Wśród tematów wiodących pojawiły się m.in. praca nad surowym materiałem do korpusów, zasady anotacji (kodowanie i znakowanie) zbiorów tekstów, tworzenie statystyk opisujących zbiory tekstów (np. słownictwo, związki wyrazowe, konkordancje, listy frekwencyjne) oraz wprowadzenie w obsługę systemów DSpace, KonText, i Korpusomat. Na program potoku B pierwszego dnia warsztatów złożyły się cztery wystąpienia. Celem prezentacji M. Gajka (Uniwersytet Warszawski / Politechnika Wrocławska) było omówienie zastosowania rozwiązań wordnetowych na przykładzie badania zapożyczeń i reliktów leksykalnych (slawizmów) w języku jidysz. Prezentacja zawierała szereg odwołań do zastosowania infrastruktury CLARIN w językoznawstwie kontrastywnym oraz w tworzeniu i wykorzystaniu słowników relacyjnych. Wystąpienie M. Marciniak (Instytut Podstaw Informatyki PAN) miało wyraźny in-

struktażowy charakter i dotyczyło techniki ekstrakcji terminologii, fraz i jednostek wielowyrazowych z korpusów językowych. Prelegentka zademonstrowała ponadto sposób tworzenia własnych słowników, glosariuszy i indeksów, wydobywanie żądanego słownictwa z tekstów, identyfikację słownictwa charakterystycznego oraz badanie terminologii i języka specjalistycznego. Dwa następne referaty adresowane były do badaczy komunikacji ustnej. Możliwości analizy języka mówionego omówił P. Pęzik (Uniwersytet Łódzki). W szczególności przedstawiona została aplikacja Spokes, umożliwiająca przetwarzanie i analizę języka mówionego (np. zapis rozmów, dialogów, wywiadów, itp.), tworzenie statystyk i in. Analizie mowy i komunikatów akustycznych poświęcone było wystąpienie D. Korzinka (Polsko-Japońska Akademia Technik Komputerowych), na które składało się przedstawienie specyfiki badań nad językiem mówionym, analiza cech mowy zaburzonej, automatycznej transkrypcji mowy oraz wykorzystanie danych wydobytych z zapisów języka w badaniach społecznych i psychologicznych.

Drugi dzień warsztatów (piątek) w potoku A otworzyła prezentacja A. Dziob (Politechnika Wrocławska) dotycząca Słowosieci, tj. wielkiej relacyjnej bazy danych leksykalnych. Uczestników zapoznano z metodą korzystania ze słowników relacyjnych oraz możliwościami ich zastosowania. Omówiono m.in. tryb wyszukiwania przykładów słów na potrzeby dalszych badań, opcję tworzenia słowników, glosariuszy i indeksów oraz identyfikację cech i danych. Wykorzystanie polsko-angielskiej Słowosieci w pracy filologa było przedmiotem wystąpienia E. Rudnickiej (Politechnika Wrocławska). Referentka zaprezentowała mianowicie możliwości wykorzystania słowników relacyjnych w praktyce tłumaczeniowej oraz glottodydaktycznej, w badaniach translatorycznych i kontrastywnych oraz dalsze funkcjonalności omówione we wcześniejszych referatach (m.in. wyszukiwanie przykładów słów na potrzeby dalszych badań, tworzenie słowników, metody korzystania ze słowników relacyjnych). W. Świerczyńska-Głownia (Uniwersytet Jagielloński) zaprezentowała na przykładzie własnego badania dotyczącego analizy dyskursu medialnego o koronawirusie możliwości wprzęgnięcia infrastruktury CLARIN w badaniach społecznych, komunikologicznych i psychologicznych. Referentka wykazała, że dostępne narzędzia mogą być pomocne przy takich projektach badawczych jak badanie dyskursu medialnego, analiza obrazu świata i konceptu na podstawie danych językowych, modelowanie tematyczne tekstu lub korpusu tekstów. Autorzy prezentacji pod tytułem „Analiza stylometryczna" (T. Walkowiak, M. Piasecki) zapoznali uczestników z możliwością wykorzystania zasobów infrastrukturalnych CLARIN w badaniach tekstów pod kątem nacechowania stylistycznego. Więcej uwagi poświęcono m.in. zagadnieniom badania autorstwa tekstu, identyfikacji stylu autora lub cech stylu danego gatunku tekstów w badaniach genologicznych. Tematem referatu był ponadto ślad społeczno-kulturowy oraz tworzenie statystyk. Zasady przeprowadzenia analizy tematycznej (topic modeling) przybliżyło wystąpienie przygotowane przez T. Walkowiaka i M. Piaseckiego. Referenci podkreślili, że analizowane procedury polegające na identyfikacji określonych przez danego

badacza typów informacji w dużych kolekcjach tekstów przy zastosowaniu analiz ilościowych mogą być doskonale wykorzystane jako narzędzia wspierające badania semantyczne, tekstologiczne oraz dyskursologiczne.

Potok B otworzyły dwa referaty dotyczące korpusów wielojęzycznych. Ich specyfikę, obsługę przeglądarek korpusowych, korzystanie z zasobów wielojęzycznych oraz wykorzystanie w badaniach translatorycznych, kontrastywnych oraz semantycznych w ujęciu międzyjęzykowym przedstawił dla korpusów polsko-angielskich P. Pęzik (Uniwersytet Łódzki). R. Roszko (Instytut Slawistyki PAN) omówił z kolei korpusy równoległe polsko-słowiańskie i polsko-bałtyckie. Zastosowanie infrastruktury CLARIN w badaniach diachronicznych i genologicznych w zakresie historii języka i literatury było tematem wystąpienia M. Pastuch (Uniwersytet Śląski) pt. „Potoczność w dawnych polskich dramatach". Przykład tworzenia i przetwarzania specjalistycznego korpusu tekstów oraz opcje badania terminologii specjalistycznej były tematem wiodącym wystąpienia M. Ogrodniczuka (Instytut Podstaw Informatyki PAN), który przybliżył te zagadnienia na przykładzie korpusu tekstów z zakresu dyskursu parlamentarnego. Przegląd usług i opcji badań dla języków innych niż polski w badaniach przekładoznawczych, glottodydaktycznych i kontrastywnych był motywem przewodnim wystąpienia przygotowanego przez J. Wieczorka i E. Rudnicką (Politechnika Wrocławska). Zakres analizy wydźwięku emocjonalnego badanych tekstów omówił J. Kocoń, inżynier języka naturalnego z Politechniki Wrocławskiej. Wiodącym tematem referatu było omówienie zastosowania narzędzi CLARIN w badaniach nad emocjami (sentiment analysis), a w szczególności wyznaczanie wydźwięku emocjonalnego fragmentów tekstu, dokonanie charakterystyki polaryzacji analizowanego tekstu (negatywna, neutralna lub pozytywna) oraz wybrane aspekty analizy dyskursu medialnego. Drugi dzień webinarium zakończyła sesja plenarna i dyskusja.

Uczestnicy dwudniowych warsztatów otrzymali pokaźną dawkę wiedzy odnośnie aplikacji dostępnych w ramach sieci CLARIN, umożliwiających wykorzystanie opracowanych już zbiorów archiwów cyfrowych i korpusów językowych. Zajęcia warsztatowe wykazały, że istnieje możliwość opracowywania istniejących tekstów publikowanych w Internecie na bieżąco, takich jak informacje prasowe, artykuły, blogi, dokumenty i in. Ponadto istnieje możliwość analizy języka mówionego (wideoblogów, transmisji czy audycji), ponieważ dostępne są (lub wkrótce będą) stosowne aplikacje. Przedstawione prezentacje z pewnością przekonały wielu uczestników, że korzystanie z usług CLARIN nie wymaga wyjątkowo wysublimowanych kompetencji z zakresu specjalistycznej wiedzy informatycznej. Natomiast uświadomienie faktu, że zasady stosowane przy opracowaniu korpusu języka polskiego (CLARIN-PL) są w pełni zgodne z usługami innych europejskich centrów tego typu, co stwarza dalsze możliwości łączenia poszczególnych narzędzi w jedno- lub wielojęzyczne ciągi przetwarzania tekstów i wydobywanie z nich potrzebnych treści, było bardzo inspirujące i zachęcające do podjęcia współpracy.

Z powyższego sprawozdania wynika, że program warsztatów został bardzo dobrze przemyślany i ułożony w taki sposób, by uczestników o nikłych kompetencjach szybko w prowadzić w zagadnienia budowania i działań na dużych korpusach, poczynając od przekazania rudymentarnych instrukcji aż po przykłady o kompleksowym wykorzystaniu wielu możliwości w ramach jednego badania. Organizatorzy zadbali ponadto o to, by w czasie przerw obiadowych i kawowych stworzyć możliwość nawiązania kontaktu, zasięgnięcia porady, czy zadawania pytań organizatorom i gościom. Temu celowi służyły również spotkania plenarne i panel dyskusyjny pt. „Przyszłość przetwarzania języka w kontekście nauk humanistycznych i społecznych", który zakończył obrady. Przeprowadzone warsztaty w całości w formie online nie pozostawiły niedosytu z powodu braku kontaktu *face to face*. Zarówno układ programu zakładający progres merytoryczny, jak i doskonała organizacja oraz niezawodne aplikacje pozwoliły skoncentrować się na zawartości tematycznej i metodologicznej referatów, co zapewne zaowocuje wykorzystaniem omówionych narzędzi w badaniach i publikacjach badaczy nauk filologicznych i społecznych.

## ZITIERNACHWEIS:

Jarosz, Józef. „Sprawozdanie z wirtualnych warsztatów ‚CLARIN-PL w praktyce badawczej' 19–20.11.2020, Wrocław (Politechnika Wrocławska)", *Linguistische Treffen in Wrocław* 19, 2021 (I): 537–541. DOI: https://doi.org/10.23817/lingtreff.19-38.