Cykl wykładów i warsztatów

CLARIN-PL w praktyce badawczej. Cyfrowe narzędzia do analizy języka w pracy humanistów i tłumaczy

13 – 15 kwietnia 2015 roku
Warszawa, Pałac Staszica, ul. Nowy Świat 72, sala 144

Centrum Humanistyki Cyfrowej Instytutu Badań Literackich PAN
CLARIN-PL

Zajęcia prowadzą pracownicy naukowi Politechniki Wrocławskiej, Uniwersytetu Łódzkiego, Uniwersytetu Pedagogicznego im. KEN w Krakowie, Instytutu Podstaw Informatyki PAN

dr Anna Andrzejczuk, dr hab. Maciej Eder, mgr inż. Paweł Kędzia, mgr inż. Jan Kocoń, dr inż. Michał Marcińczuk, dr Marek Maziarz, dr Marcin Oleksy, dr Piotr Pęzik, dr inż. Maciej Piasecki, dr Ewa Rudnicka, dr inż. Tomasz Walkowiak, mgr inż. Michał Wendelberger, dr Marcin Woliński, dr Alina Wróblewska

Zapraszamy na warsztaty z praktycznego wykorzystania cyfrowych narzędzi do ilościowej analizy języka, skierowane do badaczy z obszaru nauk humanistycznych i społecznych, oraz do tłumaczy.

CLARIN-PL to polskie konsorcjum naukowe, będące częścią ogólnoeuropejskiej infrastruktury badawczej CLARIN (Common Language Resources & Technology Infrastructure), udostepniającej zasoby językowe oraz elektroniczne narzędzia do automatycznego przetwarzania języka, które mogą znaleźć zastosowanie w badaniach opartych na gromadzeniu i analizie dużych ilości tekstowych materiałów źródłowych.

Pierwsza część warsztatów będzie poświęcona wykorzystaniu narzędzi i zasobów językowych w pracach badawczych z zakresu nauk humanistycznych i społecznych. Zapraszamy pracowników naukowych do udziału w zajęciach, podczas których będą mogli zapoznać się m. in. z systemami do klasyfikacji tekstu, wspomagającymi tworzenie słowników dziedzinowych na podstawie tekstów oraz do badań związanych z nazwami własnymi i indeksami, które stanowią pomoc w pracach leksykograficznych. W zakres warsztatów wejdą takie zagadnienia, jak: gromadzenie i udostępnianie korpusów oraz możliwość wykorzystania

narzędzi CLARIN-PL w pracy humanisty (przegląd narzędzi, zasobów i aplikacji – potencjalne zastosowania).

Druga grupa zagadnień dotyczy wykorzystania korpusów językowych (oraz ekstrakcji i analizy frazeologii z korpusów) w pracy tłumaczy. Przyjrzymy się m.in. bazie równoległych tekstów polskich i angielskich, uczestnicy poznają pojęcia ekwiwalencji frazeologicznej oraz sposoby zastosowania korpusów do jej weryfikacji.

Organizując pierwsze w Polsce warsztaty CLARIN-PL dla humanistów, liczymy na udział wszystkich naukowców, których interesuje wykorzystanie nowych metod, technik i narzędzi w praktyce badawczej. Wcześniejsza znajomość zagadnień z zakresu lingwistyki korpusowej nie jest wymagana. Dostęp do opracowanych narzędzi i zasobów językowych oraz wykorzystanie technologii językowych w naukach humanistycznych otwierają nowe ścieżki działań w badaniach literaturoznawczych i językoznawczych oraz w pracach leksykograficznych i translatologicznych.

Osoby zainteresowane udziałem w warsztatach **prosimy o przesłanie zgłoszenia** na adres aleksandra.wojtowicz@ibl.waw.pl **do dnia 3 kwietnia 2015**. Warsztaty będą miały charakter praktyczny, niezbędne zatem będzie przyniesienie własnych laptopów. Jeżeli chcecie Państwo uczestniczyć tylko w wybranych dniach warsztatów, proszę je wskazać w zgłoszeniu.
Liczba miejsc jest ograniczona, pierwszeństwo mają pracownicy ora współpracownicy IBL PAN i CLARIN-PL.

PROGRAM

Pałac Staszica, sala 144

PONIEDZIAŁEK 13 kwietnia

Infrastruktura naukowa

9.00 – 10.00 Wykład
Centrum Technologii Językowych CLARIN-PL: deponowanie i upowszechnianie zasobów oraz narzędzi językowych dla języka polskiego
Prowadzący: **dr inż. Tomasz Walkowiak i dr inż. Maciej Piasecki**

Centrum Technologii Językowych CLARIN-PL, uruchomione na Politechnice Wrocławskiej, jest węzłem ogólnoeuropejskiej infrastruktury CLARIN ERIC, skierowanej do badaczy nauk humanistycznych i społecznych. Celem wykładu jest przegląd usług udostępnianych użytkownikom przez CLARIN-PL oraz pokazanie, w jaki sposób mogą oni wykorzystać

Centrum do deponowania i archiwizacji własnych zasobów językowych (np. korpusów, słowników). Omówione zostaną standardy metadanych stosowane w Centrum, a także system logowania w ogólnopolskiej federacji uwierzytelniania, gwarantującej bezpieczeństwo przechowywania danych i umożliwiającej logowanie na podstawie własnego konta z jednostki macierzystej (jeżeli przystąpiła ona do federacji).

Narzędzia korpusowe

10.00 – 10.45 Wykład
Gromadzenie, anotowanie i udostępnianie korpusów
Prowadzący: **dr Marcin Oleksy i mgr inż. Jan Kocoń**

Ważnym zadaniem Centrum Technologii Językowych CLARIN-PL jest przechowywanie i udostępnianie korpusów oraz dostarczenie narzędzi umożliwiających wygodne prace korpusowe. Podczas wykładu słuchacze zapoznają się z podstawowymi zagadnieniami dotyczącymi przechowywania w Centrum własnych korpusów, jak ustalenie odpowiedniej licencji, wybór właściwego formatu, opis metadanymi, możliwości przetwarzania i znakowania korpusów w systemie Inforex, czy wykorzystanie narzędzi do gromadzenia korpusów bezpośrednio ze źródeł internetowych.
W ramach zajęć warsztatowych uczestnicy samodzielnie zdeponują mały korpus testowy, wgrają go do systemu Inforex i poddadzą wstępnemu przetwarzaniu. Będą także anotować i przeszukiwać korpus (za pomocą systemu NoSketch) oraz wykonają statystyczną analizę anotacji i utworzą podstawowe listy frekwencyjne.

10.45-11.00 Przerwa na kawę

11.00 – 12.30 Warsztaty – Gromadzenie korpusów, anotowanie i udostępnianie

12.30 – 13.30 Wykład
Narzędzia do automatycznej analizy odniesień w tekstach
Prowadzący: **dr inż. Michał Marcińczuk, mgr inż. Jan Kocoń**

W ramach CLARIN-PL powstają narzędzia automatycznie rozpoznające w tekstach nazwy własne i wyrażenia określające relacje czasowe. Wykład poświęcony jest prezentacji tych narzędzi oraz kwestiom ich wykorzystania w automatycznym znakowaniu korpusów. Prowadzący pokażą, w jaki sposób przeglądać i poprawiać automatyczną anotację, jak zapisywać wyniki analizy, jak tworzyć słowniki najczęstszych wystąpień nazw własnych i wyrażeń czasowych.

Podczas warsztatów uczestnicy będą mogli wykorzystać zdobytą wiedzę do samodzielnej analizy korpusu testowego.

13.30 – 14.15 Przerwa obiadowa

14.15 – 15.45 Warsztaty – Narzędzia do automatycznej analizy odniesień w tekstach

15.45 – 16.15 Wykład
Zaawansowane narzędzie do analizy korpusu w oparciu o reguły
Prowadzący: **dr inż. Michał Marcińczuk**

Język WCCL służy do formalnego opisu konstrukcji językowych i pozwala samodzielnie tworzyć reguły znakowania korpusów. Podczas wykładu zaprezentowany zostanie system WCCL Match Tester, który pozwala uruchamiać i testować reguły zapisane w języku WCCL.
W ramach warsztatów uczestnicy będą mieli możliwość napisać proste reguły znakowania, a następnie wypróbować je na korpusie testowym.

16.15-16.30 Przerwa na kawę

16.30 – 17.30 Warsztaty – Zaawansowane narzędzie do analizy korpusu w oparciu o reguły

WTOREK 14 KWIETNIA

Narzędzia słownikowe

9.00 – 10.00
Wykład
Słowosieć 3.0 - leksykalna sieć semantyczna języka polskiego i jej zastosowanie w analizie znaczeń
Prowadzący: **dr Marek Maziarz, mgr inż. Paweł Kędzia, dr inż. Maciej Piasecki**

Słowosieć 3.0 to leksykalna sieć semantyczna języka polskiego i największy jak dotąd tego typu słownik (wordnet) na świecie, mający liczne i rozmaite zastosowania. Podczas wykładu słuchacze zapoznają się ze sposobem opisu znaczeń leksykalnych w Słowosieci. Zaprezentowany zostanie system WordnetLoom, który służy do przeglądania i edycji Słowosieci, oraz narzędzia działające w oparciu o Słowosieć, umożliwiające wyznaczanie miar podobieństwa znaczeniowego i automatyczne ujednoznacznianie znaczeń słów występujących w tekście.

Uczestnicy warsztatów zainstalują aplikację WordnetLoom i za jej pomocą będą przeglądać Słowosieć. Na korpusie testowym zastosują narzędzia ujednoznaczniające, przeprowadzą analizę statystyczną rozpoznanych znaczeń i stworzą ich słownik frekwencyjny.

10.00 – 11.00 Warsztaty – Słowosieć 3.0

11.00 – 11.15 Przerwa na kawę

11.15 – 11.45 Wykład
Dwujęzyczna Słowosieć - możliwości wykorzystania w pracy tłumacza
Prowadzący: **dr Ewa Rudnicka**

Znaczenia leksykalne w Słowosieci zostały połączone z odpowiadającymi im znaczeniami w sieci języka angielskiego - Princeton Wordnet. W ramach wykładu omówione zostaną różnice w sposobie opisu między obiema sieciami oraz przedstawiony zostanie system relacji międzyjęzykowych, wspierających pracę tłumacza. Podczas warsztatów uczestnicy zajmą się analizą konkretnych problemów tłumaczeniowych i spróbują je rozwiązać przy użyciu relacji międzyjęzykowych.

11.45 – 12.45 Warsztaty – Dwujęzyczna Słowosieć

12.45 – 13.30 Przerwa obiadowa

13.30 – 14.15 Wykład
Narzędzia do automatycznego wydobywania słowników związków frazeologicznych i terminów
Prowadzący: **mgr inż. Michał Wendelberger, dr Marek Maziarz**

W ramach CLARIN-PL opracowane zostało narzędzie, które rozpoznaje w tekstach wielowyrazowe jednostki leksykalne: terminy i związki frazeologiczne. Umożliwia ono (pół)automatyczne tworzenie (na podstawie dostarczonych korpusów tekstu) słowników takich jednostek, opisanych pod względem leksykalno-składniowym i semantycznym.
Uczestnicy warsztatów nauczą się wydobywać jednostki wielowyrazowe z korpusu testowego i za pomocą dostępnego systemu stworzą własny słownik.

14.15 – 15.15 Warsztaty – Narzędzia do automatycznego wydobywania słowników związków frazeologicznych i terminów

15.15 – 15.30 Przerwa na kawę

15.30 – 16.30
Wykład
Korpusy referencyjne i równoległe w warsztacie tłumacza
Prowadzący: **dr Piotr Pęzik**

Korpusy stanowią ważny element warsztatu tłumacza, a ich nieustanny rozwój stwarza coraz lepsze możliwości zastosowań. W ramach wykładu przedstawione zostaną: Narodowy Korpus Języka Polskiego oraz powstały w ramach CLARIN-PL polsko-angielski korpus równoległy Paralela. Wyjaśnione zostaną pojęcia ekwiwalencji frazeologicznej: syntagma, frazem, translat.
Podczas warsztatów uczestnicy zapoznają się z działaniem wyszukiwarki SlopeQ dla NKJP oraz dowiedzą się, jak stosować korpusy NKJP i Paralela do weryfikacji ekwiwalencji frazeologicznej.

16.30 – 18.00 Warsztaty – Korpusy referencyjne i równoległe w warsztacie tłumacza

ŚRODA 15 kwietnia

Narzędzia do badań nad tekstem

9.00 – 9.45 Wykład
Możliwości wykorzystania narzędzi CL-PL w pracy humanisty. Przegląd narzędzi, zasobów i aplikacji - potencjalne zastosowania
Prowadzący: **dr Marcin Woliński**

Wykład poświęcony zostanie analizatorom morfologicznym, które stanowią podstawę przetwarzania tekstów, w tym nowym możliwościom analizatora Morfeusz. Zaprezentowana zostanie także dostępna infrastruktura do tworzenia słowników dziedzinowych.

9.45 – 11.15 Warsztaty
Prowadząca: **dr Anna Andrzejczuk**

W ramach warsztatów uczestnicy zapoznają się z narzędziem Kuźnia i przy jego pomocy samodzielnie stworzą własny słownik.

11.15 – 11.30 Przerwa na kawę

11.30 – 12.30 Wykład
System do klasyfikacji tekstu i analizy stylometrycznej
Prowadzący: **dr hab. Maciej Eder, dr inż. Maciej Piasecki**

W ramach CLARIN-PL powstał system, który wspiera badania stylometryczne poprzez automatyczną klasyfikację tekstów oraz ich semantyczną anotację i analizę. Umożliwia między innymi zastosowanie znanego systemu Stylo (Maciej Eder i Jan Rybicki) za pośrednictwem strony WWW.
Celem wykładu jest prezentacja elementów systemu (od wydobywania cech tekstu po interpretację wyników analizy), wskazanie jego możliwości i ograniczeń oraz omówienie wybranych przykładów zastosowań.
Podczas zajęć warsztatowych uczestnicy wprowadzą do systemu przykładowy korpus, przeprowadzą analizy w oparciu o różne parametry i zinterpretują uzyskane wyniki. Przetestują także działanie przygotowanych wcześniej klasyfikatorów i przeanalizują cechy charakteryzujące zdefiniowane w tekstach klasy semantyczne.

12.30 – 13.30 Warsztaty – System do klasyfikacji tekstu i analizy stylometrycznej

13.30 – 14.15 Przerwa obiadowa

14.15 – 15.15 Wykład
Rejestr konwersacyjny - rzeczywistość i stylizacja na podstawie korpusu Spokes
Prowadzący: **dr Piotr Pęzik**

Korpus Spokes stanowi ważny zasób w badaniach nad rejestrem konwersacyjnym języka polskiego. Wykład poświęcony zostanie charakterystyce nieformalnej polszczyzny mówionej oraz wybranym aspektom stylistycznym na przykładzie formuł konwersacyjnych.
Uczestnicy warsztatów zapoznają się z wyszukiwarką Spokes (http://spokes.clarin-eu.pl) oraz z metodami badań języka mówionego z wykorzystaniem danych korpusowych.

15.15 – 16.15 Warsztaty – Rejestr konwersacyjny

16.15 – 16.30 Przerwa na kawę

16.30 – 17.30 Wykład
Parsowanie składniowe i jego zastosowania
Prowadzący: **dr Alina Wróblewska**

Parsowanie składniowe, czyli automatyczna analiza składniowa zdań, jest jednym z kluczowych elementów automatycznego przetwarzania języka naturalnego.

Wykład zostanie poświęcony parsowaniu zależnościowemu i składnikowemu, możliwościom i ograniczeniom obu tych metod oraz zastosowaniu parserów składniowych w aplikacjach NLP i w badaniu zjawisk składniowych w tekstach. Podczas warsztatów uczestnicy będą mogli przetestować parser zależnościowy dla języka polskiego w serwisie http://multiservice.nlp.ipipan.waw.pl oraz zapoznać się z dostępnymi systemami parsującymi. Poznają także możliwości pracy z bankami struktur składniowych, takimi jak Składnica i bank struktur LFG.

17.30 – 18.30 Warsztaty – Parsowanie składniowe i jego zastosowania