W ramach Letniej Szkoły Humanistyki Cyfrowej odbędzie się

III cykl wykładów i warsztatów

CLARIN-PL w praktyce badawczej. Narzędzia cyfrowe do analizy języka w naukach humanistycznych i społecznych

17 - 19 czerwca 2015 roku

Uniwersytet Pedagogiczny w Krakowie, ul. Podchorążych 2

Organizatorzy:

CLARIN-PL

Wydział Filologiczny Uniwersytetu Pedagogicznego w Krakowie

Instytut Języka Polskiego Polskiej Akademii Nauk

Prowadzący: dr hab. Maciej Eder, dr Wojciech Jaworski, mgr inż. Paweł Kędzia, mgr inż. Jan Kocoń, dr hab. Krzysztof Marasek, dr inż. Michał Marcińczuk, dr Marek Maziarz, dr Marcin Oleksy, dr Piotr Pęzik, dr inż. Maciej Piasecki, dr hab. Adam Przepiórkowski, dr Ewa Rudnicka

Program warsztatów CLARIN-PL

Ze względu na bardzo duże zainteresowanie warsztatami CLARIN-PL zajęcia będą przebiegać równolegle w dwóch tokach.

środa 17 czerwca	
GRUPA A sala 214	**GRUPA B sala 424**
9:30 - 11:00	
Centrum Technologii Językowych CL-PL: gromadzenie, deponowanie, anotowanie i udostępnianie korpusów	Korpusy mowy i narzędzia do ich przetwarzania
11:30 - 13:00	

Narzędzia do automatycznej analizy odniesień w tekstach	Centrum Technologii Językowych CL-PL: gromadzenie, deponowanie, anotowanie i udostępnianie korpusów

14:30 - 16:00

Korpusy mowy i narzędzia do ich przetwarzania	Narzędzia do automatycznej analizy odniesień w tekstach

czwartek 18 czerwca

9:30 - 11:00

Słowosieć 3.0 – leksykalna sieć semantyczna języka polskiego i jej zastosowanie w analizie znaczeń	Narzędzia do automatycznego wydobywania słowników kolokacji i do oceny połączeń wyrazowych

11:30 - 13:00

Narzędzia do automatycznego wydobywania słowników kolokacji i do oceny połączeń wyrazowych	Parsowanie semantyczne i jego zastosowania

14:30 - 16:00

Parsowanie semantyczne i jego zastosowania	Słowosieć 3.0 – leksykalna sieć semantyczna języka polskiego i jej zastosowanie w analizie znaczeń

piątek 19 czerwca

9:30 - 11:00

Rejestr konwersacyjny polszczyzny, czyli dyskurs w czasie rzeczywistym na podstawie danych Spokes	System do klasyfikacji tekstu i analizy stylometrycznej

11:30 - 13:00

System do klasyfikacji tekstu i analizy stylometrycznej	Rejestr konwersacyjny polszczyzny, czyli dyskurs w czasie rzeczywistym na podstawie danych Spokes

Centrum Technologii Językowych CLARIN-PL: gromadzenie, deponowanie, anotowanie i udostępnianie korpusów (45 min. wykład + 45 min. warsztaty)

Prowadzący: dr Marcin Oleksy, mgr inż. Jan Kocoń, dr inż. Maciej Piasecki

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: ustalaniem odpowiedniej licencji, wyborem właściwego formatu, opisem meta-danymi, możliwościami przetwarzania i znakowania korpusów w systemie Inforex, użyciem 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.

Narzędzia do automatycznej analizy odniesień w tekstach (45 min. wykład + 45 min. warsztaty)

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. Tematem wykładu będzie prezentacja tych narzędzi oraz 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.

Korpusy mowy i narzędzia do ich przetwarzania (45 min. wykład + 45 min. warsztaty)

Prowadzący: dr hab. Krzysztof Marasek i mgr inż. Danijel Korzinek

W ramach CLARIN-PL opracowano szereg narzędzi wspomagających prace z nagraniami mowy polskiej. Obejmują one możliwość transkrypcji fonetycznej tekstu, detekcji mowy w sygnale audio, wyszukiwania specyficznych zjawisk akustycznych (np. muzyki) oraz podziału nagranych wypowiedzi na wypowiedziane przez poszczególnych mówców. Istnieje także możliwość czasowego dopasowania transkrypcji do nagrania, co umożliwia dokładną analizę fonetyczną. W ramach warsztatów uczestnicy zapoznają się z opracowanymi narzędziami i sposobami ich użycia.

Słowosieć 3.0 – leksykalna sieć semantyczna języka polskiego i jej zastosowanie w analizie znaczeń (45 min. wykład + 45 min. warsztaty)

Prowadzący: dr Marek Maziarz, mgr inż. Paweł Kędzia, dr inż. Maciej Piasecki, dr Ewa Rudnicka

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. Będą mogli również śledzić frekwencję znaczeń wybranych przez siebie wyrazów w korpusie stenogramów sejmowych (Sejmu Rzeczypospolitej ostatnich kadencji), jak również wygenerować listę frekwencyjną znaczeń wyrazów (np. w konkretnym okresie).

Narzędzia do automatycznego wydobywania słowników kolokacji i do oceny połączeń wyrazowych (45 min. wykład + 45 min. warsztat)

Prowadzący: dr Marek Maziarz, dr inż. Maciej Piasecki

W ramach CLARIN-PL opracowane zostało narzędzie, które rozpoznaje w tekstach kolokacje – potencjalne jednostki leksykalne (zestawienia, 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 z korpusu testowego kolokacje i za pomocą dostępnego systemu stworzą własny słownik połączeń wyrazowych. Uczestnicy warsztatów nauczą się wydobywać jednostki wielowyrazowe z korpusu testowego i za pomocą dostępnego systemu stworzą własny słownik.

Parsowanie semantyczne i jego zastosowania (45 min. wykład + 45 min. warsztaty)

Prowadzący: dr Wojciech Jaworski, dr hab. Adam Przepiórkowski

Parsowanie semantyczne polega na automatycznym uzyskaniu reprezentacji znaczenia danego zdania lub - ogólniej - tekstu. Wykład zostanie poświęcony przedstawieniu wstępnej wersji rozwijanego obecnie parsera języka polskiego. Omówiona zostanie przyjęta reprezentacja semantyczna i jej wizualizacja w postaci grafów semantycznych. Krótko przedstawiony zostanie także proces uzyskiwania takich reprezentacji dla zdań wejściowych. Podczas warsztatów uczestnicy będą mogli bliżej zapoznać się z reprezentacjami składniowymi i semantycznymi. Wykorzystany zostanie system INESS oferujący wizualizację struktur składniowych i możliwość wyboru poprawnej z być może wielu zaproponowanych przez parser. Wspólnie zbadane zostaną także reprezentacje semantyczne wybranych zdań polskich, oraz ich użyteczność w zadaniach związanych z humanistyką cyfrową.

Rejestr konwersacyjny polszczyzny, czyli dyskurs w czasie rzeczywistym na podstawie danych Spokes (45 min. wykład + 45 min. warsztaty)

Prowadzący: dr Piotr Pęzik

Korpus udostępniony przez wyszukiwarkę 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 odnotowanych w korpusie. Uczestnicy warsztatów zapoznają się z wyszukiwarką Spokes (http://spokes.clarin-eu.pl) oraz z metodami badań języka mówionego z wykorzystaniem danych korpusowych.

System do klasyfikacji tekstu i analizy stylometrycznej (45 min. wykład + 45 min. warsztaty)

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. 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, które charakteryzują klasy semantyczne zdefiniowane w tekstach.