# Warsztaty “CLARIN-PL w praktyce badawczej” XIII Warszawa, IPI PAN 11-12 maja 2023

# Opis bloków warsztatowych

# Wykłady

Zajęcia dostępne dla wszystkich zarejestrowanych uczestników

## Analiza tematyczna oraz semantyczna przy użyciu narzędzi CLARIN-PL

Przedstawimy narzędzia służące do identyfikowania topików (wiązek słów powiązanych semantycznie), grupowania tekstów według kryteriów semantycznych oraz wyodrębniania jednostek tekstowych charakterystycznych dla analizowanego materiału. Narzędzia z tej grupy są bardzo przydatne

## Analiza wydźwięku emocjonalnego oraz wykrywanie mowy nienawiści w tekstach

Podczas prezentacji przedstawimy najnowsze narzędzia, których zadaniem jest identyfikacja i ocena wydźwięku emocjonalnego oraz jednostek tekstowych zawierających mowę nienawiści. Przedstawimy możliwości i ograniczenia narzędzi funkcjonujących w ramach infrastruktury CLARIN-PL oraz plany dotyczące ich rozszerzenia.

## Analiza jakości usług translatorycznych online oraz korpusy równoległe polsko-słowiańsko-bałtystyczne

Podczas prezentacji zostanie przedstawiona analiza istniejących usług translatorycznych oraz oferta zasobów korpusów równoległych wchodzących w skłąd infrastruktury CLARIN-PL. Korpusy równoległe są nieustannie rozwijanym zasobem dwujęzycznym. Korpusy zawierają ręcznie zrównoleglone współczesne teksty:

polskie i bułgarskie (o łącznej objętości >27,5 mln słowoform); polskie i litewskie (o łącznej objętości >16,5 mln słowoform); polskie i ukraińskie (o łącznej objętości >1,2 mln słowoform); polskie i rosyjskie o łącznej objętości >5,6 mln słowoform.

## Jak rozumieć znaczniki fleksyjne stosowane w narzędziach IPI PAN

Przetwarzanie języka naturalnego jest dziedziną interdyscyplinarną, która angażuje głównie wiedzę z zakresu informatyki i językoznawstwa. Zrozumieniu zasad działania, ale też możliwości i ograniczeń narzędzi NLP, sprzyja poznanie niektórych istotnych ustaleń mieszczących się w ww. obszarach nauki. Jedną z kluczowych koncepcji w tym kontekście jest system znaczników fleksyjnych stosowanych w korpusach PAN. To dzięki tym znacznikom nasze narzędzia są w stanie odczytać formy gramatyczne wyrazów.

## Korpus Dyskursu Parlamentarnego i korpus ParlaMint

Korpus Dyskursu Parlamentarnego jest zbiorem stenogramów posiedzeń plenarnych Sejmu i Senatu RP, tekstów interpelacji oraz sprawozdań posiedzeń komisji od roku 1919 do chwili obecnej. Dokumenty te zostały opisane metadanymi oraz przetworzone automatycznie narzędziami lingwistycznymi i udostępnione do przeszukiwania oraz pobrania. Korpus ParlaMint rozszerza ten schemat na dokumenty parlamentarne z kilkunastu krajów europejskich, także zapisanych we wspólnym formacie i udostępnionym do badań porównawczych. Podczas wykładu zaprezentujemy oba korpusy i zachęcimy do ich wykorzystania.

## Korpusomat.eu – tworzenie korpusów nie tylko w języku polskim

Przedstawimy aplikację internetową Korpusomat.eu pozwalającą na tworzenie korpusów lingwistycznych w wielu językach. Korpusomat.eu wykorzystuje dane udostęnione przez użytkownika do stworzenia korpusu, nie wymaga przy tym żadnej wiedzy technicznej. Omówimy funkcje aplikacji - m.in. przeszukiwanie korpusu, profile słów, grupowanie wyników, wizualizację drzew zależnościowych. Wykorzystanie funkcji zostanie omówione na przykładzie case study analizy regulacji UE.

## Korpus Zależności Referencyjnych i korpus CorefUD

Korpus Zależności Referencyjnych jest zbiorem tekstów z Narodowego Korpusu Języka Polskiego anotowanych lingwistycznie relacjami koreferencji nominalnej. Został on włączony do wielojęzycznego korpusu CorefUD, stosującego wspólny schemat koreferencji dla 12 języków europejskich. Podczas wykładu opowiemy o modelu relacji referencyjnych stosowanych w tych korpusach, ich konstrukcji, metodach dostępu i możliwości ich użycia.

## Słownik jednostek wielowyrazowych

Podczas wykładu zaprezentowany zostanie "Słownik gramatyczny jednostek wielowyrazowych", tworzony obecnie w ramach projektu CLARIN-PL. Słownik powstał z połączenia wcześniejszych dzieł: SaWA, SEJF, Sejfek i Verbel. Wcześniejsze opracowania i aktualnie powstające różni jednak zmiana technologiczna (nowy słownik jest aplikacją online) oraz poszerzone możliwości przeszukiwania danych językowych. O szczegółach korzystania ze słownika opowiedzą jego twórcy: Monika Czerepowicka i Marcin Woliński.

## Słownik walencyjny Walenty

Słownik walencyjny ***Walenty*** zawiera zależności walencyjne występujących w języku polskim predykatów, w przeważającej liczbie czasowników. Zależności walencyjne to inaczej ograniczenia narzucane na sposób, w jaki poszczególne wyrazy (a właściwie ich klasy paradygmatyczne zwane leksemami) wiążą się ze swoimi podrzędnikami. Reguły takiego wiązania zwane są walencją przez analogię do wiązań chemicznych.\

## Słowosieć - wielki słownik relacyjny języka polskiego

Słowosieć (z ang. wordnet) – to relacyjny słownik semantyczny, który odzwierciedla system leksykalny języka polskiego. Obecnie zawiera 191 tys. słów, 285 tys. znaczeń oraz ponad 600 tys. relacji. Jest największym istniejącym wordnetem na świecie i nieustannie się rozrasta. Pojedyncze znaczenia w Słowosieci połączone są wzajemnymi relacjami, tworząc sieć semantyczną. Wyrazy definiowane są poprzez odniesienie do innych wyrazów połączonych znaczeniowo, np. *samochód* to hiponim dla *pojazdu drogowego* i *dwuśladu*, hiperonim dla *sedana* czy *kabrioletu* oraz synonim dla *fury* i *bryki*, tworzy kolokację z przymiotnikiem *terenowy*, a jego meronimy to *silnik*, *spryskiwacz* czy *podwozie*.

Podczas prezentacji przedstawimy zasady tworzenia Słowosieci, jej aktualny stan oraz zastosowania naukowe - zrealizowane i potencjalne.

## TermoPL – omówienie narzędzia do ekstrakcji terminologii dziedzinowej

TermoPL to narzędzie, służące do automatycznego wydobycia terminologii dziedzinowej z tekstów. Poza terminologią, narzędzie może służyć do ekstrakcji jednostek wielowyrazowych. Wydobycie terminologii jest mechanizmem pomocnym m.in. w tworzeniu słowników dziedzinowych, zasobów do tłumaczenia tekstów oraz streszczeń dokumentów, w opracowywaniu ontologii danej dziedziny, w anotacji dokumentów i wspomaganiu wyszukiwania odpowiedzi na pytania.

## Zasoby i narzędzia do automatycznego odpowiadania na pytania (QA)

W ramach prac w projekcie CLARIN-BIZ powstają narzędzia do automatycznego odpowiadania na pytania. Jednym z zasobów przygotowanych w toku prac, jest wzorowany na anglojęzycznym SQuADzie 2.0 zbiór PoQuAD: zestaw 70 tys. kontekstów pochodzących z Wikipedii, anotowanych pytaniami i odpowiedziami. Zasób ten posłużył do treningu modeli realizujących ekstraktywne, oraz generatywne strategie odpowiadania na pytania. Zostały one inkorporowane do aplikacji, pozwalającej na indeksowanie źródeł, i uzyskiwanie na ich podstawie odpowiedzi na pytania. Podczas wykładu przedstawimy możliwości, jak i ograniczenia wypracowanych strategii.

#

# Ćwiczenia / laboratoria

Na tę grupę będą obowiązywały zapisy dodatkowe prowadzone wśród osób już zarejestrowanych.

5 maja zostaną wysłane mejle z linkiem do zapisu.

## Analiza tematyczna oraz semantyczna (warsztat)

Podczas warsztatów uczestnicy będą mogli osobiście przetestować narzędzia, które zostały omówione podczas prezentacji na tożsamy temat.

## CLARIN-PL w paradygmacie Open Science

Infrastruktura CLARIN ERIC dostarcza europejskiej społeczności naukowej ogromnych ilości danych na wolnej licencji. Pomagamy również w przygotowaniu procesu deponowania danych, ich właściwym opisie, zarządzaniu i udostępnianiu. Uwzględnienie zasad FAIR (findability, accessibility, interoperability, and reusability) w zarządzaniu danymi naukowymi jest coraz częstszym wymaganiem stawianym przez rozmaitych grantodawców (w tym ERC oraz NCN). Celem warsztatów jest zaznajomienie uczestników z tym, czym są open science i zasady FAIR w perspektywie naukowca planującego badania finansowane ze środków publicznych.

## COMBO – narzędzie do wstępnego przetwarzania języka (1)

Na warsztatach zostanie przedstawione COMBO — narzędzie do wstępnego przetwarzania języka naturalnego (ang. natural language preprocessing). Testując możliwości narzędzia, uczestnicy będą mieli okazję przejść przez kolejne etapy wstępnego przetwarzania języka, tj. segmentację, lematyzację, analizę morfologiczną wraz z tagowaniem częściami mowy, oraz parsowanie zależnościowe. Ponadto uczestnicy poznają proces trenowania modelu, który stanowi podstawę działania COMBO, dla nowego języka, oraz metody weryfikacji jakości predykcji zwracanych przez system.

## COMBO – narzędzie do wstępnego przetwarzania języka (2)

j.w.

## Dewulgaryzator – automatyczna redakcja przekleństw

Dewulgaryzator to narzędzie umożliwiające automatyczną redakcję wyrażeń wulgarnych w polskich tekstach. Zostało przygotowane w oparciu o zbiór zdań wulgarnych pochodzących z filmów oraz ich niewulgarnych odpowiedników, w których niecenzuralne wyrażenia zostały zastąpione eufemistycznymi zamiennikami. W trakcie zajęć zaprezentowane zostanie narzędzie a także materiał wykorzystany do jego przygotowania, następnie uczestnicy będą mieli możliwość samodzielnego przetworzenia wybranych przez siebie tekstów.

## Jak wykorzystać TermoPL we własnej pracy

Warsztaty posłużą demonstracji możliwości narzędzia TermoPL, które może być wykorzystywane do identyfikowania i ekstrakcji terminologii, jednostek wielowyrazowych oraz słownictwa charakterystycznego dla określonej grupy tekstów. Narzędzie może być wykorzystywane wszechstronnie m.in. do tworzenia słowników dziedzinowych, analizy dyskursu, grupowania semantycznego tekstów.

## Korpusomat.eu

Zajęcia warsztatowe będą służyły przetestowaniu funkcji Korpusomat.eu, odpowiedzi na pytania oraz rozszerzeniu wiedzy uczestników na temat aplikacji. Warsztat będzie miał charakter komplementarny do części wykładowej jednak wcześniejsza obecność na wykładzie nie jest wymagana.

## Korpus Dyskursu Parlamentarnego w praktyce

Podczas warsztatów przećwiczymy język zapytań do korpusu wykonując kilka ćwiczeń w wyszukiwarce.

## Narzędzia analizy stylometrycznej tekstu

Stylometria jest dziedziną wiedzy, która opisuje teksty przy wykorzystaniu miar statystycznych odnoszących się do warstwy językowej komunikatu. Celem warsztatu jest zaznajomienie uczestników z narzędziami, które umożliwiają użytkownikom identyfikację cech formalnych (językowych) tekstu, grupowanie (klasyfikację) tekstów według kryteriów formalnych i dalszą pracę z tak przygotowanymi wynikami. Narzędzia stylometryczne stosowane są dziś z powodzeniem w analizie dyskursu, naukach o literaturze, językoznawstwie kryminalistycznym, medioznawstwie i badaniach marketingowych.

## Pozyskiwanie dobrej jakości danych językowych

Warsztat poświęcony będzie kwestiom związanym z pozyskiwaniem dobrej jakości danych językowych wykorzystywanych w pracy badacza-humanisty. Uczestnicy zostaną zapoznani ze źródłami, z których można czerpać teksty dostępne na wolnej licencji, a także z zagadnieniami dotyczącymi selekcji materiału badawczego pod względem jego przydatności do analizy. Poruszone zostaną problemy opisu materiału za pomocą metadanych; ponadto uczestnicy będą mieli okazję poznać praktyczne wykorzystanie narzędzi służących do gromadzenia oraz przygotowania danych językowych do analizy.

## WieszczGPT: laboratorium ChatGPT dla nauk humanistycznych i społecznych

Celem warsztatów jest przedstawienie możliwości, jakie oferuje ChatGPT - zaawansowany model języka stworzony przez OpenAI, a także zaprezentowanie idei powstającego w infrastrukturze CLARIN-PL modelu językowego o nazwie WieszczGPT - modelu ukierunkowanego na potrzeby użytkowników zainteresowanych naukami humanistycznymi i społecznymi. W ramach warsztatu przedstawione zostaną techniki posługiwania się narzędziem ChatGPT do rozwiązywania różnorodnych problemów powstających przy pracy z tekstem. Zaprezentujemy również plany zespołu CLARIN-PL dotyczące stworzenia polskiego modelu językowego. Serdecznie zapraszamy do uczestnictwa w warsztatach i zachęcamy do włączenia się we wspólne budowanie nowej technologii językowej!

## Wykrywanie zależności referencyjnych

Przedstawione zostanie narzędzie programistyczne do wykrywania zależności referencyjnych w języku polskim oraz case study wykorzystania koreferencji w analizie danych tekstowych. W ramach ćwiczeń zostanie udostępniony interfejs w postaci strony internetowej do przetestowania działania narzędzia.