CLARIN-PL-Biz

technologie językowe dla nauki i biznesu

Fundusze Europejskie
Inteligentny Rozwój

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego

CLARIN-PL-Biz

CLARIN-PL (Common Language Resources & Technology Infrastructure) to wynik współpracy jednostek naukowych, które budują elektroniczne zasoby językowe i narzędzia do pracy z dużymi zbiorami tekstów w języku polskim. Obecnie konsorcjum tworzą: Politechnika Wrocławska (lider Konsorcjum), Instytut Podstaw Informatyki Polskiej Akademii Nauk, Instytut Slawistyki Polskiej Akademii Nauk, Uniwersytet Łódzki oraz Uniwersytet Wrocławski. Członkiem naukowej części infrastruktury badawczej jest również Polsko-Japońska Akademia Technik Komputerowych, która wnosi niezwykle ważny wkład wiedzy o metodach analizy i rozpoznawania mowy. CLARIN-PL zrzesza więc jedne z największych polskich ośrodków badawczych w dziedzinie inżynierii języka naturalnego i jej powiązań z metodami sztucznej inteligencji.

Głównymi odbiorcami narzędzi CLARIN-PL byli do tej pory naukowcy badający dziedziny humanistyczne i społeczne. Pojawiły się jednak potrzeby przedstawicieli innych nauk, instytucji publicznych, organizacji i wreszcie firm komercyjnych. Dlatego zrodziła się idea rozszerzenia wersji CLARIN-PL i włączenia się w ogólny trend rozwoju sztucznej inteligencji. Odtąd projekt ma świadczyć usługi zarówno dla nauki, jak i dla biznesu.

CLARIN-PL-Biz ma na celu rozszerzenie infrastruktury badawczej CLARIN-PL, która stanie się **platformą badawczo-rozwojową** do przetwarzania języka naturalnego i eksploracji wielkich baz danych językowych. Połączona ona będzie z centrum technologicznym zapewniającym bazę sprzętową. Projekt realizują wyspecjalizowane zespoły złożone z programistów i lingwistów, specjalistów w zakresie przetwarzania języka naturalnego. Do konsorcjum projektowego dołączyły 22 firmy wraz ze swoim zapleczem badawczym. Osiem kolejnych firm polskich i zagranicznych wsparło projekt cennym wkładem w postaci baz danych i oprogramowania.

Przetwarzanie języka naturalnego, gospodarka oparta na danych i wiedzy, ogólnoświatowy kierunek rozwoju społeczeństw z dobrobytem osiąganym dzięki zaawansowanej technologii pracującej na dobro człowieka i nieinwazyjnej pod względem ekologicznym to już nie jest wizja przyszłości, lecz teraźniejszość. Dlatego tworzymy narzędzia służące badaniu tekstów mówionych i pisanych oraz tekstów multimodalnych za pomocą inżynierii języków naturalnych.

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

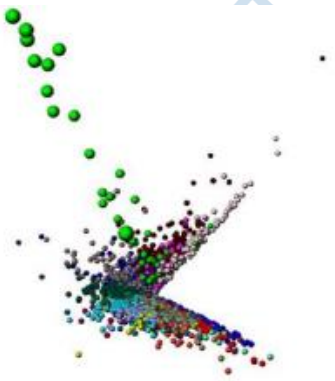
Beneficjenci CLARIN-PL mogą dogłębnie analizować teksty w językach naturalnych

- na uspójnionych technologiach, pozwalających dostosować programy do indywidualnych potrzeb;
- w środowisku informatycznym do tworzenia systemów dialogowych;
- bez konieczności posługiwania się kłopotliwymi instalacjami informatycznymi.

Dostarczone przez CLARIN-PL-Biz narzędzia i usługi umożliwią

- łatwy i wszechstronny dostęp do archiwum zasobów i technologii językowych oraz aplikacji dla użytkownika końcowego;
- kompleksowe wydobywanie informacji i wiedzy z danych tekstowych;
- rozwój gospodarki opartej na danych i wiedzy w Polsce, a dzięki temu wzrost konkurencyjności polskich przedsiębiorstw i nauki.

Zaledwie ok. 10 języków na świecie ma więcej niż 100 milionów użytkowników. Język polski, będąc na ok. 30. miejscu tej klasyfikacji (0,61% populacji świata), pozostaje poza głównym nurtem badań międzynarodowych. W dobie wspólnego rynku cyfrowego niezwykle ważna jest widoczność Polski, dlatego tak istotne jest stworzenie platformy gromadzącej dane, udostępniającej odpowiednią moc obliczeniową oraz algorytmy wyspecjalizowane dla języka polskiego i innych języków słowiańskich powiązanych za pośrednictwem angielskiego z zasobami światowymi. Dzięki nim powstanie szereg usług dla wszystkich sektorów rynku, wspierających budowę polskiej gospodarki opartej na wiedzy, innowacyjnej i konkurencyjnej.



Docelowo CLARIN-PL-Biz będzie oferować **usługi wysokiego poziomu**, które można podzielić na następujące grupy:

- gromadzenie i trwałe przechowywanie danych językowych, w tym tzw. *big data* (tekst, mowa, dane multimodalne oraz powiązane zasoby numeryczno-symboliczne);
- wydobywanie informacji i wiedzy z wielkich baz danych językowych, generowanie odpowiedzi na pytania na podstawie korpusu tekstów oraz automatyczne rozumowanie w oparciu o dane językowe;
- semantyczne indeksowanie i przeszukiwanie danych;
- środowisko informatyczne do tworzenia systemów dialogowych o różnej modalności, w tym generowanie tekstu i mowy, oraz rozpoznawanie mowy.

Historia

2012 Powstaje CLARIN ERIC, Polska podpisuje umowę o bezterminowym członkostwie

2013 Start pierwszego projektu CLARIN-PL, kierowanego do badaczy (zwłaszcza z zakresu nauk społecznych i humanistycznych), Politechnika Wrocławska liderem

2015 CLARIN-PL staje się certyfikowanym centrum repozytoryjnym typu B, powstaje Centrum Technologii Językowych (CTJ)

2017 CLARIN-PL staje się certyfikowanym centrum szerzenia wiedzy typu K, powstaje centrum PolLinguaTec

2020 Start projektu badawczo-rozwojowego CLARIN-PL-Biz, CLARIN-PL wychodzi z ofertą do sektora komercyjnego

Historia CLARIN-PL to wspólna historia – nasza i Państwa

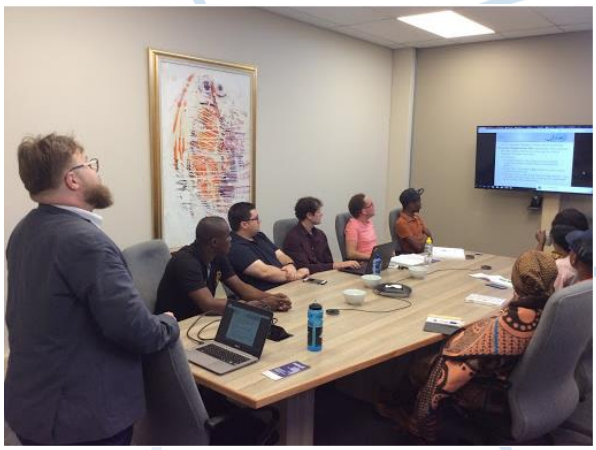
Helpdesk

Współpraca z użytkownikami musi być dostosowana do ich indywidualnych potrzeb. Dlatego w CLARIN-PL funkcjonuje **helpdesk**, którego zadaniem jest udzielanie informacji i porad; reagowanie na zgłaszane uwagi, sugestie i nieścisłości; wyjaśnianie sposobów korzystania z poszczególnych usług oraz narzędzi; pośrednictwo w kontaktach z ekspertami technicznymi.

Koordynator
Jan Wieczorek
jan.wieczorek(at)pwr.edu.pl

Helpdesk oferuje **bieżące wsparcie** wszystkich użytkowników: studentów, którzy potrzebują wsparcia w badaniach do prac dyplomowych, czy profesorów prowadzących badania własne. Nasza pomoc nie ogranicza się wyłącznie do analizy danych tekstowych w ramach prac naukowych. Wspomagamy również takie etapy badań, jak dostarczanie materiału, jego przetworzenie oraz interpretację wyników. Poza typowo badawczym charakterem działań, pracownicy helpdesku CLARIN-PL regularnie udzielają pomocy dydaktykom i naukowcom piszącym **wnioski grantowe**.

Wspieramy użytkowników również poprzez organizację **warsztatów i konsultacji**. Mamy doświadczenie wynikające z przygotowania kilkudziesięciu wydarzeń szkoleniowych: przekrojowych prezentacji infrastruktury CLARIN-PL (średnio 120 uczestników), mniejszych warsztatów dostosowanych do zapotrzebowania specjalistów rozmaitych dziedzin (medioznawców, psychologów, pracowników Kancelarii Prezesa Rady Ministrów) oraz indywidualnych konsultacji małych zespołów badawczych. Wielokrotnie wspieraliśmy również firmy szukające wiedzy z zakresu przetwarzania języka podczas procesu **przygotowania wniosków do NCBiR**.



Korzystanie z infrastruktury CLARIN-PL nie wymaga specjalistycznej wiedzy programistycznej, jednak poziom jej złożoności może utrudniać obsługę niedoświadczonemu użytkownikowi. By obniżyć poziom tych utrudnień, przygotowujemy **materiały szkoleniowe i informacyjne**: instrukcje, samouczki, tutoriale i ćwiczenia. Umożliwiają one indywidualną pracę w środowisku narzędzi CLARIN-PL bez konieczności każdorazowego kontaktu z helpdeskiem lub oczekiwania na termin warsztatów.

Poza aktywnym wspieraniem naszych użytkowników, staramy się wyciągnąć wnioski o tym, jak zmienia się zapotrzebowanie na technologie i zasoby językowe. Dzięki obserwacji zmieniających się potrzeb i trendów widzimy, że mamy do zaoferowania bardzo wiele nie tylko nauce, lecz także **sektorowi biznesowemu.**

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Centrum repozytoryjno-obliczeniowe

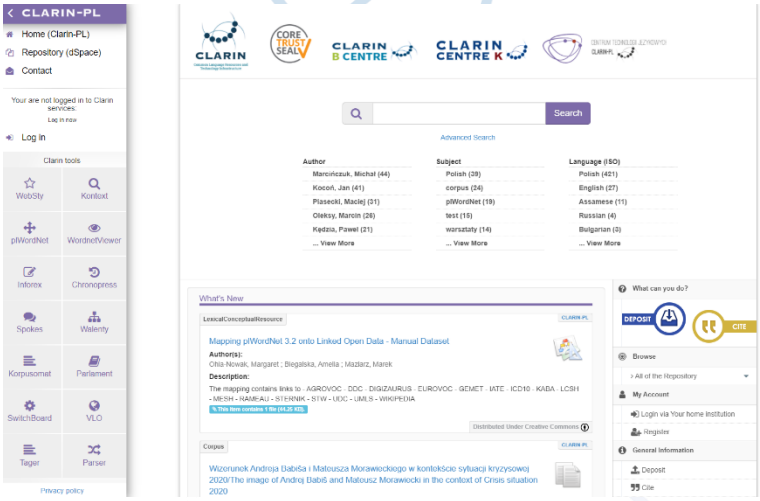
Obecnie CLARIN-PL dysponuje infrastrukturą techniczną do zastosowań związanych głównie z obszarem badawczym nauk humanistycznych i społecznych. Konieczność jej rozbudowy wynika z wychodzenia naprzeciw potrzebom nowych użytkowników.

Koordynator
Tomasz Naskręt
tomasz.naskret(at)pwr.edu.pl

Centrum Technologii Językowych (CTJ) dysponuje klastrem repozytoryjno-obliczeniowym, rozmieszczonym na dziewięciu serwerach. Posiada również zapasową infrastrukturę, umożliwiającą ciągły dostęp do kluczowych aplikacji i usług sieciowych.

DSpace jest cyfrowym systemem repozytoryjnym, oferującym proste wyszukiwanie danych i narzędzi oraz ich pobieranie. Zapewnia on także opis spójnym zestawem metadanych, co umożliwia integrację polskich narzędzi i zasobów z europejskimi oraz ułatwia ich prawidłowe cytowanie w publikacjach naukowych. Obecnie CLARIN-PL akceptuje wszelkie dane i narzędzia językowe oraz NLP, np. banki drzew, korpusy, zasoby leksykalne, modele językowe, parsery, tagery, serwisy językowe. System dSpace dostępny jest na stronie https://clarin-pl.eu/dspace/ .

Chmura repozytoryjna **CLARINCloud** znajduje się pod adresem https://nextcloud.clarin-pl.eu. Jej otwarta architektura pozwoliła na zintegrowanie z repozytorium dSpace, uruchomienie wspólnego systemu logowania użytkowników oraz integrację z innymi usługami działającymi w infrastrukturze (np. Inforexem, zob. Korpusy). Pliki CLARINCloud są przechowywane w konwencjonalnych strukturach katalogów.



Usługi CLARIN-PL są dostosowane do wymogów zintegrowanego narzędzia **CLARIN Language Resource Switchboard** (https://switchboard.clarin.eu/), które identyfikuje język tekstu oraz proponuje odpowiedni program do jego analizy.

Nowa infrastruktura sprzętowa ma zapewnić dużo większą wydajność obliczeniową oraz możliwości składowania danych. Planujemy usprawnić i ujednolicić metody przetwarzania danych. Dostępne będą również nowe usługi, takie jak tworzenie własnych potoków przetwarzania z dostępnych narzędzi czy wypożyczenie mocy obliczeniowej do prowadzenia własnych obliczeń np. trenowania modeli.

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

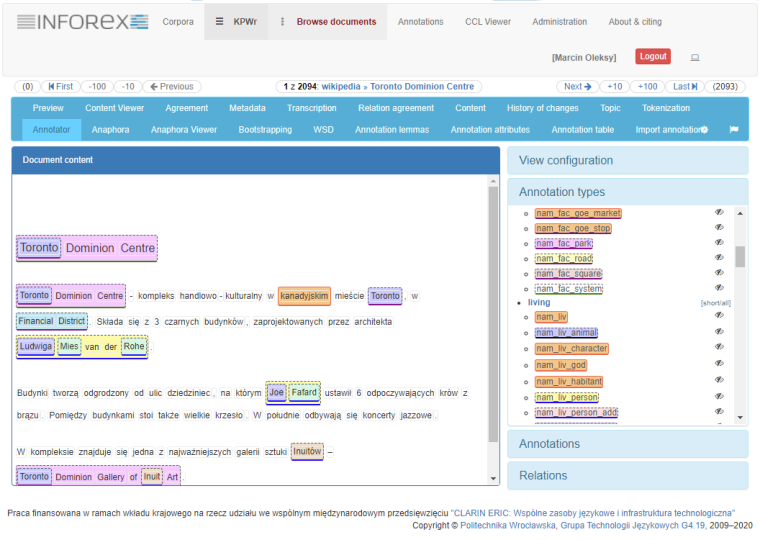
Zasoby i narzędzia korpusowe

Dane korpusowe to jedno z dwóch podstawowych źródeł, obok zasobów leksykalnych, do konstrukcji i uczenia inteligentnych algorytmów przetwarzania języka. Są one odpowiednio dobranymi zbiorami dokumentów tekstowych. O ich wartości decydują również dodatkowe informacje na temat słów, fraz, zdań i całych tekstów, które są wprowadzane zarówno w sposób automatyczny, jak i ręczny przez wykwalifikowanych specjalistów.

Koordynator
Marcin Oleksy
marcin.oleksy(at)pwr.edu.pl

KPWr (**Korpus Języka Polskiego Politechniki Wrocławskiej**) jest zbiorem dokumentów tekstowych dostępnych na licencji Creative Commons. Próbki do korpusu pobrano ze źródeł typu Wikipedia, Wikinews, portale informacyjne, dzieła literackie z domeny publicznej lub udostępnione na otwartej licencji. Dokumenty zostały automatycznie opisane informacjami gramatycznymi i ręcznie – innymi typami informacji, takimi jak: frazy składniowe (chunks), relacje między frazami składniowymi, jednostki identyfikacyjne (w przeważającej części nazwy własne), ujednoznacznione znaczenia słów, wyrażenia przestrzenne, czasowniki z podmiotem domyślnym, tekstowe słowa kluczowe, wyrażenia czasowe, sytuacje, role semantyczne i koreferencja.

Korpus w najnowszej wersji znajduje się pod linkiem: https://clarin-pl.eu/dspace/handle/11321/722. Jest on stale rozbudowywany i rozwijany w kierunku korpusu jeszcze lepiej zrównoważonego i zróżnicowanego gatunkowo. Szczegółowe statystyki znajdują się na stronie: http://nlp.pwr.wroc.pl/narzedzia-i-zasoby/zasoby/kpwr.



Inforex jest otwartym systemem sieciowym, służącym do zarządzania, przeszukiwania i analizowania zawartości wybranych korpusów, w tym również tworzonych zespołowo. Pozwala on na wyliczanie podstawowych statystyk i robienie list frekwencyjnych, ale przede wszystkim umożliwia tworzenie jakościowych danych językowych, które mogą posłużyć jako podstawa do tworzenia korpusów treningowo-testowych oraz do pogłębionej i usystematyzowanej analizy materiału badawczego. Z tego względu Inforex wykorzystywany jest przez szerokie grono badaczy. System umożliwia wprowadzanie różnorodnych poziomów opisu (anotacji), zarówno tych automatycznych, jak i dodawanych ręcznie przez użytkowników. Aplikacja dostępna jest na stronie: https://inforex.clarin-pl.eu/.

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

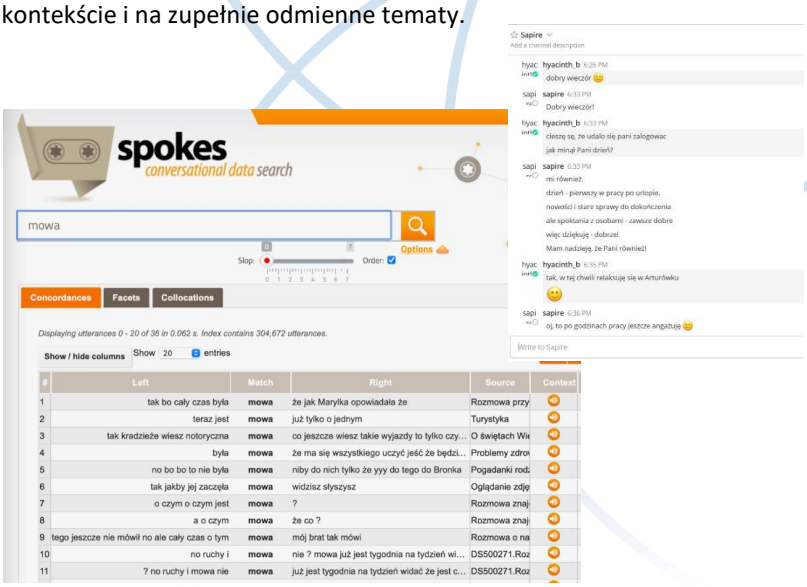
Zasoby i narzędzia korpusowe

Język konwersacyjny to pierwotna i najbardziej powszechna odmiana każdego żywego języka. Jednocześnie ze względu na koszty pozyskania nagrań autentycznych rozmów oraz ich transkrypcji, rejestr mówiony jest niedoreprezentowany w zasadzie we wszystkich korpusach referencyjnych.

SpokesPL to system, udostępniający unikalne korpusy polszczyzny konwersacyjnej. Oprócz dostępu przez wyszukiwarkę (http://spokes.clarin-pl.eu) oraz interfejs API, istnieje możliwość pobrania podkorpusu przetranskrybowanych i zdiaryzowanych nagrań (zob. http://docs.pelcra.pl/doku.php?id=spoken_offline_corpora). W **Korpusie Mowy** znajdzie się co najmniej 1000 godzin nagrań danych medialnych, wywiadów i rozmów formalnych i swobodnych. Jednym z jego unikalnych komponentów będzie baza min. 150 godzin przetranskrybowanych nagrań dialogów prowadzonych na łączach telefonicznych wg zróżnicowanych scenariuszy z branży telekomunikacyjnej, bankowej, ubezpieczeniowej, turystycznej, medycznej oraz energetycznej. Przetranskrybowane dialogi zostaną opisane anotacją na poziomie tematyki i przebiegu procesów biznesowych, segmentów funkcyjnych dialogów oraz intencji wyrażanych przez rozmówców.

Koordynator
Piotr Pęzik
piotr.pezik(at)uni.lodz.pl

Idiolekt to odmiana języka charakterystyczna dla danej osoby. Istnienie idiolektów leży u podstaw hipotezy o śladzie językowym, czyli możliwości identyfikacji autorów wypowiedzi ustnej lub pisemnej na podstawie językowych cech osobniczych w odpowiednio dużej próbie ich wypowiedzi lub tekstów. Podczas gdy biometria głosowa doczekała się wdrożeń np. w bankowości elektronicznej, skuteczność „biometrii" tekstowej jest nadal przedmiotem sporych kontrowersji. W szczególności wyzwaniem jest wielkoskalowa identyfikacja jednego z tysięcy lub milionów autorów na podstawie próbek ich odmiany języka z różnych kanałów, rejestrów i gatunków tekstów. Przykładem takiego zadania może być ustalenie autora zbiorów komentarzy na dużym forum internetowym na podstawie transkrypcji rozmów danej osoby prowadzonych w innym kontekście i na zupełnie odmienne tematy.



Korpus Idiolektów jest unikalną bazą zbieranych w kontrolowanych warunkach, wielokanałowych danych do ewaluacji metod ustalania autorstwa w języku polskim. Ponad 100 ochotników dostarcza próbki języka w min. trzech kanałach, takich jak wywiad swobodny, e-mail, czat na urządzeniu mobilnym, forum, wystąpienie publiczne itp. Dane te są następnie normalizowane i integrowane w centralnej bazie. Otwarty Korpus Idiolektów umożliwi rozwój i ewaluację wielokanałowych metod ustalania autorstwa. Automatyczna identyfikacja autorstwa ma ważne implikacje w obszarze bezpieczeństwa, prywatności (maskowanie idiolektu), a także do identyfikacji źródeł dezinformacji.

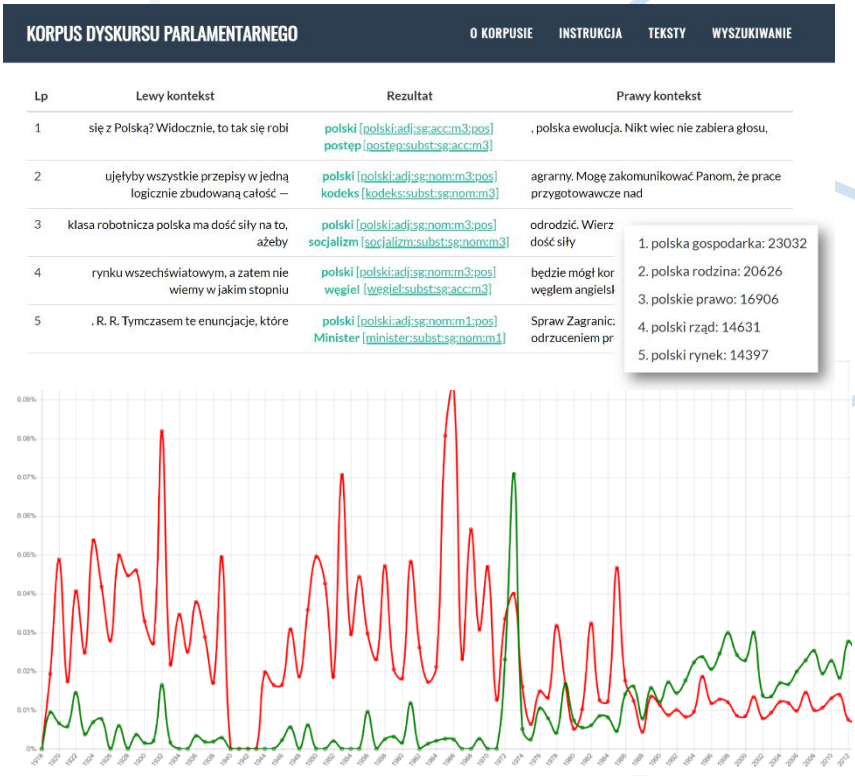
Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Zasoby i narzędzia korpusowe

Korpus Dyskursu Parlamentarnego to zbiór wypowiedzi polskich posłów i senatorów z posiedzeń plenarnych, posiedzeń komisji oraz interpelacji, zapytań i odpowiedzi od roku 1919 do chwili obecnej. Teksty korpusu są opisane metadanymi oraz przetworzone automatycznie narzędziami lingwistycznymi. Zawierają, wzorowaną na modelu Narodowego Korpusu Języka Polskiego, informację o własnościach słów, składni zdań, użytych nazwach osób, miejsc i instytucji. Korpus liczy obecnie ok. 800 mln słów i jest stale uzupełniany danymi bieżącymi i historycznymi. Jego materiały są dostępne do pobrania (http://clip.ipipan.waw.pl/PPC) i przeszukiwania (https://kdp.nlp.ipipan.waw.pl/).

Teksty korpusu zostały też włączone do Projektu **parlaMint** (https://www.clarin.eu/content/parlamint), tworzącego spójnie opisany zbiór wypowiedzi przedstawicieli 15 parlamentów narodowych w Europie. Poprzez po-równanie danych referencyjnych (sprzed listopada 2019 r.) z danymi z okresu pandemii możemy np. badać zmianę języka w czasie – w danym kraju i w perspektywie europejskiej.

Koordynator
Maciej Ogrodniczuk
maciej.ogrodniczuk(at)ipipan.waw.pl



Korpusomat to internetowe narzędzie umożliwiające użytkownikowi nieposiadającemu zaawansowanych umiejętności informatycznych samodzielne tworzenie i wykorzystywanie korpusów tekstów. Narzędzie pozwala na przesłanie zestawu plików tekstowych (lub binarnych) wraz z metadanymi, a następnie zlecenie ich automatycznej analizy lingwistycznej.

Wszystkie komponenty analityczne zainstalowane są na serwerze, który kompiluje korpus, a zadania użytkownika ograniczają się do przygotowania tekstów i wprowadzania zapytań przeszukujących korpus. Nie musi on znać żadnych szczegółów technicznych rozwiązań informatycznych wykorzystywanych do analizy tekstów.

Narzędzie jest dostępne pod adresem: https://korpusomat.pl/.

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Zasoby leksykalne

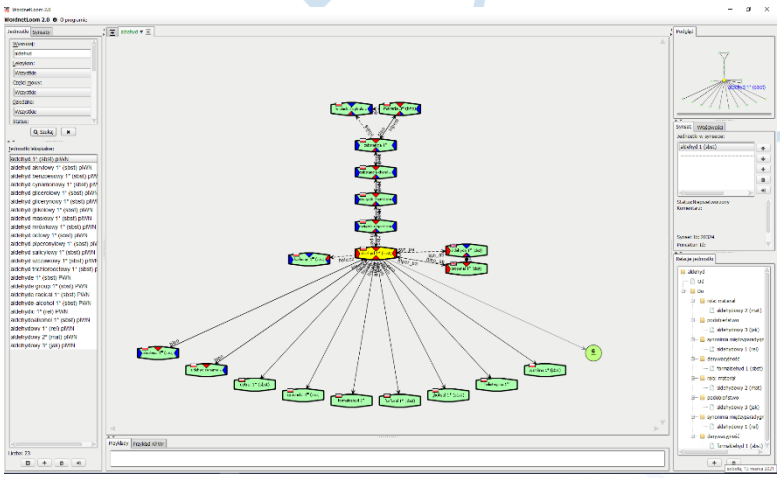
Zasoby leksykalne to jedno z dwóch podstawowych źródeł, obok danych korpusowych, do konstrukcji i uczenia inteligentnych algorytmów przetwarzania języka. Aby ich opis był łatwo przyswajalny dla komputera, są one dużo bardziej sformalizowane niż podobne zasoby tworzone dla człowieka. Dzięki temu mogą odzwierciedlać relacje, jakie zachodzą w świecie rzeczywistym, będąc dla maszyny źródłem informacji o nim.

Koordynatorka
Agnieszka Dziob
agnieszka.dziob(at)pwr.edu.pl

Słowosieć jest polskim wordnetem, czyli rodzajem elektronicznego tezaurusa, w którym znaczenia są opisywane za pomocą relacji leksykalno-semantycznych. Obejmuje całość słownictwa ogólnego języka polskiego (również najnowszego). Jest tworzona od 2005 roku, a od 2012 roku łączona relacjami międzyjęzykowymi z amerykańskim Princeton WordNetem. Stanowi unikalny, wysokiej jakości zasób polsko-angielski, będący „wyjściem na świat" dla języka polskiego w kierunku innych języków. Dzięki połączeniom międzyjęzykowym Słowosieć jest wpięta w światową sieć zasobów połączonych danych (Linked Open Data) i jest wykorzystywana m.in. przy klasyfikacji tekstów, wydobywaniu z nich informacji i wiedzy, budowie tzw. Semantic Web i opartych o nią narzędzi do tworzenia wyszukiwarek semantycznych.

100 000 znaczeń Słowosieci jest anotowanych wydźwiękiem emotywnym przez parę lingwista-psycholog, stanowiąc tym samym źródło do budowy narzędzi służących analizie wydźwięku. Zasób jest dostępny online na stronie http://plwordnet.pwr.edu.pl, skąd można też go pobrać w formacie bazy danych.

Słownik walencyjny języka polskiego **Walenty**, tworzony w Instytucie Podstaw Informatyki PAN, opisuje wymagania składniowe i semantyczne dla czasowników, rzeczowników, przymiotników i przysłówków. Jego warstwa semantyczna jest połączona ze Słowosiecią. Walenty znajduje się pod adresem: http://walenty.clarin-pl.eu/.



WordnetLoom to aplikacja do budowy Słowosieci, która, dzięki swojej modułowości, zyskała uniwersalne zastosowanie do tworzenia słowników przyswajalnych dla komputera oraz dla człowieka. Obecnie WordnetLoom jest wykorzystywana przez zespoły z Portugalii, Danii, RPA, Czech oraz Uniwersytetu Warszawskiego (do opisu języka jidysz). Aplikacja ta jest dostępna w postaci kodu źródłowego w repozytorium github-a pod adresem https://github.com/CLARIN-PL/WordnetLoom. Wersja pokazowa znajduje się pod adresem http://ws.clarin-pl.eu/public/WordnetLoom-Viewer.zip.

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

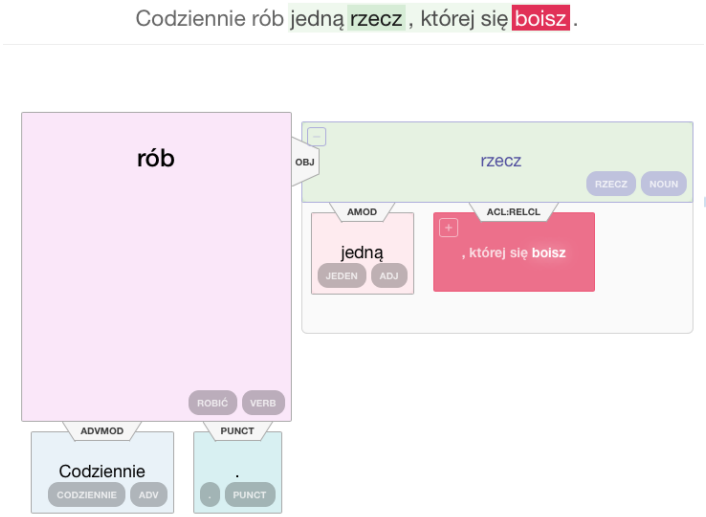
Podstawowe narzędzia analizy językowej

Rozwiązywanie zaawansowanych zadań NLP jest często poprzedzone etapem wstępnej **analizy morfoskładniowej** obejmującej m.in. podział na zdania i segmenty, analizę morfologiczną (przypisanie segmentom ich interpretacji gramatycznych), tagowanie (identyfikacja klas gramatycznych segmentów), lematyzację (identyfikacja kanonicznych form leksemów) czy parsowanie zależnościowe (predykcja struktury składniowej zdania). Współczesne narzędzia wstępnego przetwarzania języka opierają się na modelach trenowanych przy użyciu metod maszynowego uczenia na dużych zbiorach danych.

Koordynatorka
Alina Wróblewska
alina(at)ipipan.waw.pl

COMBO-pytorch to nowej generacji system do tagowania, analizy morfologicznej, lematyzacji i parsowania zależnościowego, czyli podstawowych zadań wstępnego przetwarzania języka. Zastosowana architektura typu *end-to-end* umożliwia trenowanie jednego modelu NLP do predykcji tagów, cech morfologicznych, lematów, krawędzi i etykiet zależnościowych. Model COMBO może być uczony na jednym zbiorze danych w dowolnym języku. Jakość predykcji zwracanych przez COMBO dla zdań w języku polskim jest zdecydowanie wyższa niż jakość predykcji systemów spaCy czy UDPipe.

COMBO jest łatwym do zainstalowania pakietem Pythonowym zbudowanym na platformie AllenNLP i bibliotece PyTorch. Może być używany jako biblioteka w dowolnym kodzie Pythonowym z możliwością auto-matycznego pobrania modelu albo jako niezależne narzędzie uruchamiane z linii poleceń. Umożliwia dużą swobodę konfiguracji podczas trenowania modelu, np. opcjonalne użycie pretrenowanych zanurzeń słów word2vec, fastText lub BERT. System COMBO powstał w IPI PAN, a obecnie jest rozwijany w projekcie CLARIN-PL-Biz https://gitlab.clarin-pl.eu/syntactic-tools/combo.



Polski Bank Drzew Zależnościowych (PDB, http://zil.ipipan.waw.pl/PDB) jest największym zbiorem ręcznie opisanych drzew dla języka polskiego. PDB zawiera zdania z ogólnej domeny językowej. Wersja banku drzew automatycznie przekonwertowana do formatu Universal Dependencies (PDB-UD) została włączona do największego na świecie repozytorium banków drzew zależnościowych UD. Bank PDB-UD został wykorzystany do wytrenowania modeli spaCy, Stanza, UDPipe i COMBO. PDB i PDB-UD powstawały w IPI PAN, w ramach CLARIN-PL planujemy rozbudować PDB o dwa dziedzinowe banki drzew zależnościowych.

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

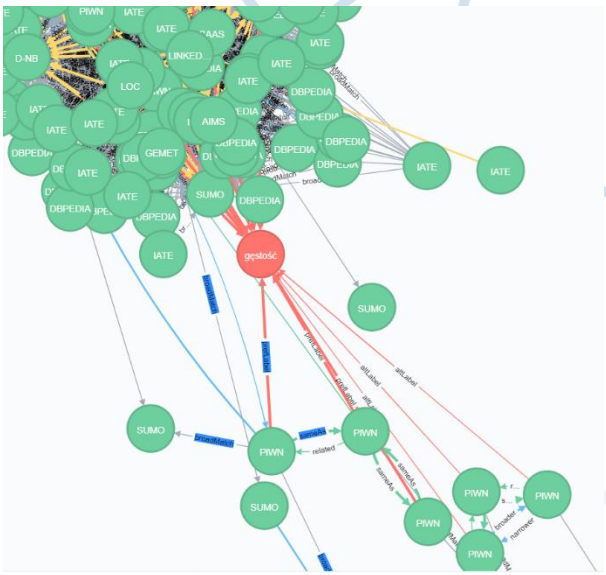
Analiza semantyczna

Ważnym obszarem przetwarzania języka naturalnego jest **ujednoznacznianie znaczeń leksykalnych** (z ang. WSD - Word Sense Disambiguation) oraz techniki rzutowania pojęć i bytów nazwanych na dostępne zasoby wiedzy (z ang. Entity Linking i Entity Disambiguation), które pozwalają odróżnić znaczenia słów o tej samej formie (np. zamek jako budowla i zamek jako narzędzie do zamykania drzwi, ale również *Jan Kochanowski* - polski poeta i *Jan Kochanowski* - historyk, rektor UW). Algorytmy WSD działają na Słowosieci wzbogaconej o połączenia z Linked Open Data, dostępne korpusy tekstów oraz definicje i przykłady użycia znaczeń ze Słowosieci.

Koordynator
Arkadiusz Janz
arkadiusz.janz(at)pwr.edu.pl

W narzędziu do ujednoznaczniania znaczeń leksykalnych **WoSeDon** wykorzystano techniki uczenia maszynowego oraz zasoby Słowosieci, połączonej z zasobami Linked Open Data, składającymi się z licznych słowników ogólnych i dziedzinowych oraz dostępnych ontologii, m.in.: Wikipedii, GeoWordnetu, DBpedii i ontologia YAGO oraz tezaurusów dziedzinowych, a także innych zasobów leksykalnych CLARIN-PL (np. słownika walencyjnego Walenty). W ostatnich pracach skoncentrowano się głównie na grupowaniu znaczeń leksykalnych w Słowosieci, aby zmniejszyć złożoność problemu ujednoznaczniania, oraz efektywnym wykorzystaniu definicji i przykładów użycia znaczeń. Podstawowa wersja systemu znajduje się pod adresem: https://clarin-pl.eu/dspace/handle/11321/540 lub http://ws.clarin-pl.eu/wsd.shtml (w formie aplikacji webowej).

Ze względu na złożoność przetwarzania i liczność dodatkowych zasobów wiedzy wersja webowa nie zawiera niektórych rozszerzeń. Trwają jednak prace nad optymalizacją algorytmów ujednoznaczniania, by w przyszłości móc udostępnić szybkie i skuteczne rozwiązania z licznymi rozszerzeniami również w ramach platformy webowej.



Drugi kierunek prac to wydobywanie informacji z tekstu na podstawie, między innymi, Słowosieci i LOD. Obejmują one rozszerzanie korpusów tekstów do uczenia i strojenia systemu WSD; rzutowanie tekstu na zasoby LOD poprzez wykorzystanie algorytmów WSD i połączeń pomiędzy Słowosiecią a LOD; generowania słów kluczowych dla tekstów; ujednoznacznianie dwujęzyczne dzięki dwujęzycznej Słowosieci. W ramach tego obszaru badań tworzone są metodami półautomatycznymi ontologie dziedzinowe, łączone następnie ze Słowosiecią i, za jej pośrednictwem, z siecią połączonych danych światowych.

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

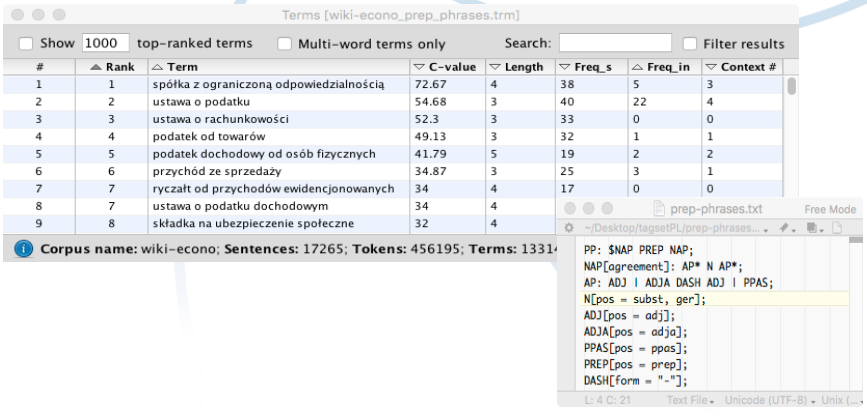
Wydobywanie informacji i wiedzy z tekstu

Wydobywanie wiedzy z tekstów składa się z etapów generowania informacji na podstawie danych zawartych w abstrakcyjnych modelach opisujących zależności oraz wnioskowania opartego na wiedzy z tych danych. Obejmuje swoim zakresem wydobywanie informacji (klasyfikacja wystąpień nazw własnych, odniesień do czasu i przestrzeni, nawiązań w tekście oraz opisów sytuacji wraz z ich strukturą), automatyczne streszczanie, podsumowywanie i znaczeniowe kompresowanie, upraszczanie komunikatu językowego oraz rozpoznawania relacji semantycznych pomiędzy fragmentami tekstów. Podstawą współczesnych metod wydobywania wiedzy są modele semantyki dystrybucyjnej, terminologia i sieci semantyczne. Nasze rozwiązania będą podstawą do systemów monitorowania i łączenia informacji, zarządzania wiedzą oraz wydobywania wiedzy z danych (Data Mining).

Koordynator
Maciej Piasecki
maciej.piasecki(at)pwr.edu.pl

PolDeepNer jest narzędziem, służącym do wydobywania nazw własnych z tekstów. Jego działanie opiera się na współczesnej, zaawansowanej architekturze transformera, korzystającej z najnowszych modeli językowych. Jego struktura pozwala na wielopoziomowe sprawdzenie tekstu dzięki wyznaczaniu nachodzących na siebie kategorii w nazwach własnych. Aktualnie jest to najlepsze narzędzie do tego zadania dla języka polskiego i jest rozwijane również dla innych języków, np. angielskiego, francuskiego, litewskiego. Program jest dostępny w repozytorium https://gitlab.clarin-pl.eu/information-extraction/poldeepner2.

TermoPL (Instytut Podstaw Informatyki PAN) to program do wydobywania terminologii z tekstów. Wyszukuje frazy rzeczownikowe będące kandydatami na terminy. Następnie je szereguje według wartości współczynnika C-value, wyliczanego na podstawie częstości występowania fraz, ich długości oraz liczby kontekstów. Początkowy fragment tak utworzonej listy zawiera terminy często używane w danej dziedzinie. Program działa dla różnych języków. Umożliwia także porównywanie dwóch list przy użyciu kilku współczynników wyrażających asymetrię występowania terminów w dwóch różnych korpusach tekstów. Zasób dostępny jest na stronie http://zil.ipipan.waw.pl/TermoPL, wersja pokazowa dostępna jest pod adresem https://ws.clarin-pl.eu/termopl.shtml.

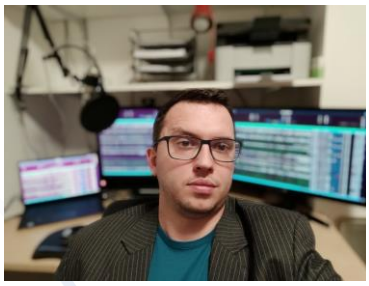


MeWeX jest programem do wydobywania kolokacji (częstych połączeń słów, powiązanych ze sobą gramatycznie lub semantycznie) z korpusów tekstów, a następnie budowania słownika wielowyrazowych słów (reprezentujących wielowyrazowe jednostki leksykalne). Umożliwia załadowanie dowolnej wielkości korpusu tekstów spakowanego do jednego pliku w formacie Zip. Wersja webowa dostępna jest pod adresem: http://ws.clarin-pl.eu/mewex.shtml. Wersja instalowana pozwala również na wydobywanie kolokacji dla języka angielskiego.

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

System analizy wydźwięku

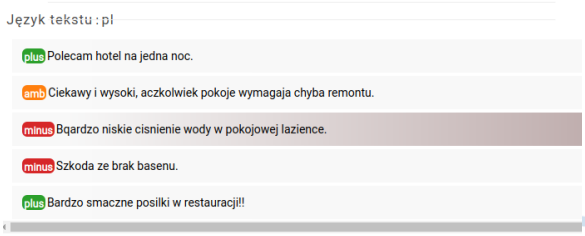
Wydźwięk emocjonalny tekstu to potencjał wzbudzania określonych emocji w jego odbiorcy. System analizy wykorzystuje jako dane treningowe teksty oznaczone przez człowieka polaryzacją wydźwięku i/lub nazwami emocji.



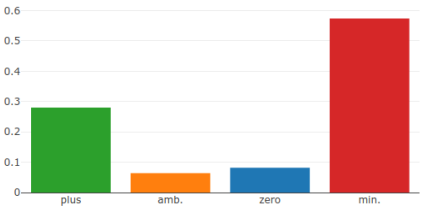
Koordynator
Jan Kocoń
jan.kocon(at)pwr.edu.pl

System analizy wydźwięku powstawał w oparciu o anotację emotywną znaczeń słów w Słowosieci. Intensywnie zaczął się jednak rozwijać pod wpływem zainteresowania użytkowników – głównie naukowców oraz firm. Leksykon wydźwięku był podstawą do konstrukcji aplikacji, pozwalającej na analizę tekstu na podstawie wydźwięku i emocji znaczeń słów pojawiających się w tekście https://ws.clarin-pl.eu/sentyment.shtml. Aktualna wersja jest oparta o wielojęzyczne modele językowe i głębokie sieci neuronowe. Umożliwia analizę wydźwięku na poziomie całego tekstu, paragrafu oraz zdania dla kilkudziesięciu języków: http://ws.clarin-pl.eu/multiemo.

Powszechnie wiadomo, że ludzie różnie reagują na te same bodźce. Istniejące metody i narzędzia do analizy wydźwięku często ignorują ten fakt i modelują wyłącznie najbardziej prawdopodobną bądź uśrednioną reakcję grupy osób. W CLARIN-PL zajmujemy się również badaniem **spersonalizowanych modeli wydźwięku i emocji**, które mogą wykorzystywać dodatkową wiedzę o człowieku i dzięki temu – dokładniej przewidzieć reakcje konkretnej osoby. Dzięki temu możliwe będzie ukierunkowanie sztucznej inteligencji na konkretną osobę i tym samym – uczynienie AI bardziej „ludzką".



Rozkład wydźwięku emocjonalnego dla zdania



Badanie opinii oparte jest o wydobywanie jej **aspektów**. Przykładowo w opinii dotyczącej hotelu może znaleźć się ocena wnętrza pokoju, wyposażenia, personelu, otoczenia i wielu innych elementów. Obecnie trwają prace związane z przygotowaniem narzędzia, które będzie w stanie wykrywać aspekty opinii, a także, na podstawie kontekstu występowania, ocenić polaryzację wydźwięku aspektu wraz z jej intensywnością.

Zespół badawczy zdobywał wcześniej swoje doświadczenie w projekcie **Sentimenti** (https://sentimenti.pl/), w którym psychologowie, językoznawcy, informatycy i specjaliści ds. zarządzania stworzyli automatyczny analizator emocji w tekście pisanym. Dane do przygotowania modeli powstały z udziałem ponad 20 tysięcy respondentów, od których pozyskano ponad 18 milionów anotacji dotyczących wydźwięku emocjonalnego i pobudzenia oraz 8 emocji podstawowych z modelu Roberta Plutchika.

Sentimenti

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

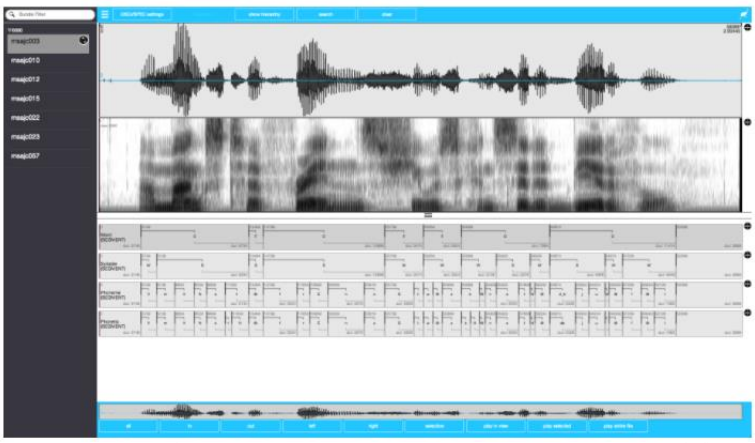
System przetwarzania mowy

Wiele danych wykorzystywanych w naukach humanistycznych i społecznych, ale nie tylko, jest przechowywanych w formie nagrań audio. Przykładami mogą być nagrania radiowe bądź telewizyjne, wywiady, przemowy, wykłady, filmy, literatura czytana i inne nagrania mowy. Problemami są czas, którego przetwarzanie mowy wymaga więcej, niż przetwarzanie tekstu pisanego, oraz konieczność posiadania przynajmniej podstawowej wiedzy na temat danych dźwiękowych.

Koordynatorka
Alicja Wieczorkowska
alicja(at)poljap.edu.pl

Korpusy mowy służą do wytrenowania określonych modeli oraz wydobywania z nich informacji. Celem Korpusu Danych Studyjnych (https://mowa.clarin-pl.eu/korpusy) było zebranie nagrań czytanych, jak najbardziej zróżnicowanych w zakresie wymowy i słownictwa. Ma on duże znaczenie dla analizy fonetycznej. Korpus Polskich Kronik Filmowych uwzględnia nagrania w formie audio oraz wideo z lat 1945-1962. Charakteryzuje się dawnym językiem oraz relatywnie niską jakością nagrań. Ze względów licencyjnych możliwe było jedynie udostępnienie transkrypcji, które znajdują się pod adresem https://clarin-pl.eu/dspace/handle/11321/426. Korpus Sejm, zawierający transkrypcję rozmów parlamentarnych jest z kolei korpusem powstałym dzięki współpracy Polsko-Japońskiej Akademii Technik Komputerowych i Senatu RP.

Narzędzie do automatycznego zrównoleglania tekstu, napisane w języku Python, jest przystosowane do przetwarzania dużych zbiorów tekstów (big data). Wersja serwerowa znajduje się pod linkiem: https://clarin-pl.eu/dspace/handle/11321/528, a uproszczona wersja narzędzia w formie aplikacji webowej pod linkiem: http://ld.clarin-pl.eu:9000. Wytrenowano również statystyczne modele dla wielu par języków powiązanych z polskim jak i angielskim. Korzystając z tychże modeli statystycznych, bogatych zasobów Wikipedii oraz wersji narzędzia znajdującej się w serwisie github pozyskano dane równoległe w dużej ilości dla różnych par języków. Pod linkiem: https://goo.gl/hFnRNV znajduje się wersja offline, której algorytm został wytrenowany na danych jedynie quasi-porównywalnych dla języków PL oraz EN, potrafiąca automatycznie rozpoznać języki dokumentów oraz mająca dodatkowe opcje filtrowania, głównie na potrzeby firm tłumaczeniowych.



System do transliteracji nagrań to narzędzie w postaci usługi online (http://mowa.clarin-pl.eu), dzięki której użytkownik może, po przesłaniu przez stronę WWW pliku audio, otrzymać transliterację mowy, czyli zapis mowy w postaci ortograficznej. Wykorzystuje ono do rozpoznawania mowy system Kaldi i modele z głębokimi sieciami neuronowymi. Pod adresem: https://github.com/danijel3/ClarinStudioKaldi znajduje się opis wersji korpusu przygotowanej do wytrenowania systemu rozpoznawania mowy skrypty pozwalające na samodzielne wytrenowanie systemu.

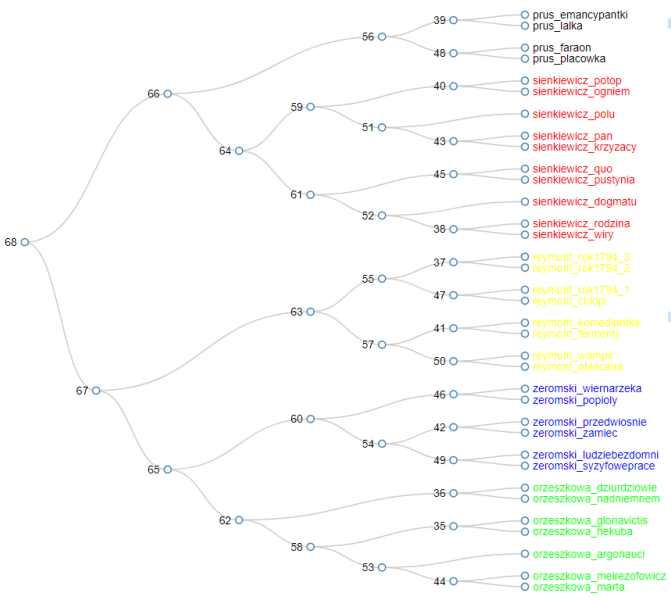
Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Analiza semantyczna i webserwis

CLARIN-PL działa w modelu SaaS (ang. *Software as a Service*, oprogramowanie jako usługa), tzn. dąży do tego, żeby większość aplikacji była dostępna w formie usługi sieciowej wraz z interfejsem **webowym**, umożliwiającym użytkownikom swobodny dostęp. W tym celu został utworzony serwis https://ws.clarin-pl.eu/, który łączy narzędzia do przetwarzania tekstów, wydobywania z nich informacji i wiedzy oraz zapewnia dostęp do wybranych zasobów poprzez aplikacje webowe.

Koordynator
Tomasz Walkowiak
tomasz.walkowiak(at)pwr.edu.pl

System **WebSty** umożliwia przeprowadzanie analiz stylometrycznych (np. atrybucja autorstwa, analiza stylu) oraz semantycznych (np. detekcja grup tematycznych czy analiza podobieństwa semantycznego) na dowolnych zbiorach tekstów poprzez interfejs webowy. Znajduje się pod adresem: https://ws.clarin-pl.eu/websty (wersja dla języka polskiego), https://ws.clarin-pl.eu/webstyml (wersja wielojęzyczna). Wersja wielojęzyczna obsługuje, poza polskim, języki: angielski, niemiecki, rosyjski, hiszpański, hebrajski oraz węgierski. Na prośbę użytkowników możliwa jest rozbudowa o kolejne języki. System automatycznego wyszukiwania podobnych tekstów z wykorzystaniem algorytmów WebSty służy do wykrywania cech charakteryzujących różnice i podobieństwa pomiędzy tekstami i korpusami tekstów. Metody te analizują cechy na różnych poziomach opisu języka naturalnego, począwszy od poziomu leksykalnego, poprzez wybrane elementy składni, po cechy semantyczne.



Narzędzie do analizy dystrybucji i frekwencji danych korpusowych **LEM - *Literacki Eksplorator Maszynowy*** (z ang. = *Literary Exploratory Machine*) zostało opracowane jako odpowiedź na potrzeby użytkowników. Z każdą jego funkcjonalnością wiąże się konkretna historia i konkretny problem z zakresu analizy tekstów w języku polskim, mniej lub bardziej złożony. Jest ono jednym z najczęściej wykorzystywanych narzędzi CLARIN-PL. Integruje w jednym miejscu podstawowe operacje na tekstach (np. wyznaczanie nazw własnych, pomiar tzw. mglistości tekstu, czyli stopnia jego skomplikowania) oraz umożliwia przetwarzanie korpusów i zwracanie wyników w prostych formatach (plik tekstowy, skoro-szyt Excel). LEM dostępny jest pod adresem: https://ws.clarin-pl.eu/lem.

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Systemy dialogowe

Systemy dialogowe, najczęściej w postaci czatbotów (ang. chatbot), weszły już do powszechnego użytku. Kluczem ich działania jest połączenie dobrej jakości rozpoznawania mowy z odpowiednim poziomem rozpoznawania istotnych elementów wypowiedzi rozmówcy. W wielu wypadkach czatboty są ukierunkowane na wybiórczą analizę treści i śledzenie dialogu. Często zastosowane metody uczenia maszynowego są silnie związane z określoną wąską dziedziną.

Koordynator
Maciej Piasecki
maciej.piasecki(at)pwr.edu.pl

Architektura budowanego w CLARIN-PL-Biz systemu dialogowego obejmie kolejne etapy: rozpoznawania mowy (otwarty na użycie rozwiązań od niezależnych dostawców); interpretację semantyczną tekstu; wnioskowanie; analizę struktury retorycznej dialogu (funkcje komunikacyjne i cele rozmówców, relacje pomiędzy fragmentami tekstu); generowanie wypowiedzi (w postaci tekstu oraz mowy). Główny nacisk zostanie położny na budowę modułów analizy treści i struktury, w których zaawansowane metody CLARIN-PL-Biz mogą przynieść największy postęp w stosunku do technologii dostępnej na rynku. Pozwoli to wyeliminować problemy wybiórczej analizy treści i śledzenia dialogu.

Planujemy opracowanie otwartego, podstawowego, szkieletowego, **modułowego systemu dialogu** dla języka polskiego oraz – na jego podstawie – wyspecjalizowanych rozwiązań dla wybranych dziedzin, dostrojonych na podstawie dziedzinowych bazy danych i metod głębokiego uczenia maszynowego. **Moduł analizy semantycznej treści** połączy narzędzia CLARIN-PL-Biz do analizy wyrażeń językowych, rozpoznawania nazw własnych i ich odniesień do baz wiedzy (np. sieci semantycznych), określeń czasu, pojęć i terminów, a także powiązań pomiędzy wyrażeniami np. sygnalizowanych przez zaimki. **Moduł analizy struktury dialogu** umożliwi powiązanie celów rozmówcy (np. klienta lub petenta) ze sposobem wyrażania funkcji komunikacyjnej, określonym typem sytuacji, której dotyczy dialog, formułowanymi intencjami oraz przekazywanymi bądź oczekiwanymi informacjami.

CLARIN
Common Language Resources and Technology Infrastructure

Dziedzinowe systemy dialogowe (np. dla obsługi klientów w bankowości i ubezpieczeniach, rezerwacji usług turystycznych, obsługi obywateli w urzędach) będą budowane jako rozwinięcie systemu ogólnego na podstawie zebranych specjalistycznych korpusów dialogowych, dziedzinowych baz wiedzy i dodatkowych rozszerzających modułów dziedzinowych, trenowanych w adaptacyjny sposób pod konkretne zbiory sytuacji, funkcji dialogowych i intencji.

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Na zakończenie należy dodać, że ani CLARIN-PL, ani CLARIN-PL-Biz nie mógłby istnieć bez **Biura Projektu**, które niezmiennie dba o wszystkie kwestie prawno-administracyjne i finansowe.

Kierowniczka Biura
Grażyna Podbielska
grazyna.podbielska(at)pwr.edu.pl

CLARIN-PL nie może istnieć również bez Państwa – odbiorców naszych działań, Beneficjentów, Klientów, Współpracowników, Partnerów.
Wiele z naszych zasobów i aplikacji powstało jako odpowiedź na Państwa konkretne potrzeby. Wiele jest rozwijanych w kierunku, który jest wypadkową Państwa pragnień i naszych możliwości.

Życzmy sobie więc kolejnych lat wspaniałej, rozwojowej i mobilizującej intelektualnie współpracy, do której już teraz zapraszamy.

Po więcej informacji na temat projektu i kontaktu z zespołem CLARIN-PL zapraszamy na stronę https://clarin.biz/ oraz na http://clarin-pl.eu/.

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

CLARIN

Konsorcjanci

Politechnika Wrocławska

IPI PAN

PJATK

SLAV

Uniwersytet Łódzki

Uniwersytet Wrocławski

netguru

W3

Silver Bullet Solutions

Funmedia

Technicenter Sp. z o.o.

avra software

pap

Qtravel

VoiceLab

Literacka technologie

TECHMO Voice Technologies

sages

feeCOMPASS

FINDWISE Search Driven Solutions

MakoLab

SmartNet Research & Solutions

Equilibrium Solutions Sp. z o.o.

Press Service monitoring mediów

stermedia

itmatica

sentione

Firmy wspierające

elp IT solutions & services

EVENTREGISTRY

sta

IQS

lingventa

Polska Press Grupa

Fundusze Europejskie Inteligentny Rozwój

Rzeczpospolita Polska

Unia Europejska Europejski Fundusz Rozwoju Regionalnego