„CLARIN-PL-Biz – technologie językowe dla nauki i biznesu II”

24-25 czerwca 2021, platforma ZOOM

CZWARTEK, 24.06.2021	
9:00-9:15	**CLARIN-PL – infrastruktura badawcza nauk humanistycznych i społecznych** (M. Piasecki, PWr) *Wprowadzenie: czym jest CLARIN-PL, jakie wyzwania stały przed zespołem realizującym projekt, jak się z tych zadań wywiązujemy.*
9:15-9:45	**CLARIN-PL – model współpracy z infrastrukturą** (J. Wieczorek, PWr) *Przedstawienie wypracowanego modelu współpracy między użytkownikami infrastruktury a zespołem zapewniającym jej utrzymanie i rozwój.*
9:45-11:00	**Korpus Czterech Wieszczów – nowy wymiar dziedzictwa polskiego Romantyzmu (**M. Troszyński, IBL PAN; M. Oleksy, PWr, A. Mędrzecka, IBL; E. Mirkowska, IBL; T. Korpysz, UKSW) *Prezentacja koncepcji Korpusu Czterech Wieszczów – projektu, którego celem jest stworzenie nowoczesnego zasobu zawierającego pełną twórczość Wieszczów. Projekt jest sztandarowym przykładem współpracy specjalistów z zakresu historii literatury, językoznawców oraz informatyków.*
11:00-11:45	**Słowosieć i jej wykorzystanie w badaniach ekonomicznych** (K. Klimczak, PŁ; E. Rudnicka, PWr) *Prezentacja projektu dotyczącego ewolucji języka finansowego, w którym wykorzystano rzutowanie międzyjęzykowe Słowosieci na angielski Princeton WordNet. Celem projektu jest opracowanie metod analizy tekstów finansowych w różnych językach opartych o techniki stworzone dla języka angielskiego. Słowosieć pozwala na stworzenie narzędzi międzyjęzykowych.*
11:45-12:00	*PRZERWA*
12:00-12:45	**Nowe usługi przetwarzania tekstów** (T. Walkowiak, PWr) *Prezentacja nowych narzędzi do przetwarzania tekstów, m.in. Punctuatora – narzędzia do wprowadzania interpunkcji oraz nowego klasyfikatora tematycznego tekstów.*
12:45-13:30	**Prezentacja parsera zależnościowego COMBO wraz z zastosowaniami** (A. Wróblewska, IPI PAN) *Prezentacja COMBO, czyli systemu wstępnego przetwarzania języka, który przeprowadza analizę morfologiczną, tagowanie, lematyzację oraz parsowanie zależnościowe. Przetwarzając język systemem COMBO można wykorzystać gotowe modele dla ponad 40 języków, albo wytrenować nowy model dla dowolnego języka, dla którego istnieje bank drzew zależnościowych. Zaprezentowane zostanie działanie systemu w realnych zadaniach.*
13:30-14:15	**Migranci i pandemia na Twitterze – analiza wydźwięku** (O. Czeranowska, SWPS; K. Chlasta, SWPS; J. Kocoń, PWr) *Prezentacja metodologii i wstępnych wyników projektu „(IT)MOBILITY: Niemobilność mobilnych, mobilność niemobilnych – migranci w czasach pandemii i nowe technologie teleinformatyczne”. Celem projektu jest zbadanie funkcji nowych mediów w kontekście pandemii COVID-19 a także zmian społecznych spowodowanych pandemią COVID-19 w zakresie mobilności wirtualnej i nie-mobilności geograficznej.*
14:15-14:30	*PRZERWA*
14:30-15:45	**Dialog Obywatelski w dyskursie parlamentarnym – przykład integralnego zastosowania infrastruktury** CLARIN (A. Hess, UJ; M. Ogrodniczuk, IPI PAN; T. Walkowiak, PWr) *Prezentacja celu, uwarunkowań oraz wstępnych wyników badania użycia terminu "dialog obywatelski" oraz pojęć pokrewnych*

	w dyskursie parlamentarnym. Realizacja projektu opierała się na wykorzystaniu Korpusu Dyskursu Parlamentarnego oraz narzędzi modelowania tematycznego dostarczanych przez infrastrukturę CLARIN-PL.
15:45-16:30	**Korpusomat i jego zastosowanie w analizie i rozwoju Zintegrowanego Rejestru Kwalifikacji** (M. Będkowski, IBE; Ł. Kobyliński, IPI PAN) *Prezentacja aplikacji Korpusomat, która umożliwia tworzenie korpusów językowych na podstawie własnych zasobów tekstowych użytkowników. Aplikacja ta w bazowej postaci pozwala na przetwarzanie tekstów na wielu poziomach analizy językowej i wykorzystanie tych warstw do przeszukiwania korpusu. Pozwala ona również na prowadzenie analiz statystycznych i porównawczych między korpusami. W trakcie prezentacji omówione zostanie wdrożenie systemu, którego bazę stanowiła aplikacja Korpusomat, umożliwiającego analizę, grupowanie i klasyfikację zasobów Zintegrowanego Rejestru Kwalifikacji. Zaprezentowane zostaną przykłady zastosowania a) aplikacji Korpusomat do analizy podstawy programowej kształcenia ogólnego oraz b) systemu opartego na aplikacji Korpusomat do przeprowadzenia grupowań tekstów (opisów kwalifikacji) na potrzeby webowej aplikacji Kompas wspierającej osoby w ich wyborach edukacyjnych.*
16:30-17:00	**Dyskusja/Q&A – odniesienie się do zgłoszeń od użytkowników** *Syntetyczne podsumowanie potrzeb użytkowników zidentyfikowanych dzięki informacjom z formularza zgłoszeniowego oraz wskazanie dalszego kierunku działań w celu ich realizacji.*

PIĄTEK, 25.06.2021	
9:00-10:00	**CLARIN-PL Biz – infrastruktura badawczo-rozwojowa dla sztucznej inteligencji i ich zastosowań** (M. Piasecki, PWr; M. Tykierko, PWr) *CLARIN jako kluczowa infrastruktura świadcząca usługi w działaniach badawczo-rozwojowych odwołujących się do wielowymiarowej technologii sztucznej inteligencji. Prezentacja strategii rozwoju, celu funkcjonowania oraz ogólnej oferty infrastruktury.*
10:00-10:40	**Rozpoznawanie mowy i jego ocena** (D. Korzinek, PJATK; P. Szymański, PWr) *Prezentacja badań nad mową foniczną z perspektywy zjawiska akustycznego. Wykorzystanie danych wydobytych z zapisów języka (mowa) w badaniach społecznych i psychologicznych.*
10:40-11:15	**Anonimizator – narzędzie do automatycznej anonimizacji tekstów (**T. Walkowiak, PWr) *Prezentacja nowej usługi służącej do automatycznej anonimizacji tekstów oraz jego zastosowań.*
11:15-11:30	*PRZERWA*
11:30-12:00	**Biznesowe zastosowania wordnetów i ontologii: przykład współpracy z QTravel** (A. Dziob, PWr; A. Janz, PWr; P. Muzioł, Q&Q) *Prezentacja badań nad technikami semantycznego wyszukiwania opartymi na elementach uczenia maszynowego przy jednoczesnym wykorzystaniu wordnetów oraz ontologii dziedzinowych. Zaprezentowane techniki półautomatycznej konstrukcji ontologii, metody budowy rozszerzonych modeli językowych oraz podejścia do semantycznego wyszukiwania będą omawiane na przykładzie projektu kontekstowej wyszukiwarki ofert realizowanego we współpracy z QTravel.*
12:00-12:45	**Wydobywanie informacji z tekstów a ocena kwalifikacji pracowników: IBE** (W. Stęchły, IBE; T. Walkowiak, PWr) *Prezentacja wyników oraz uwarunkowania realizacji projektu dotyczącego klasyfikatora kwalifikacji pracowników. Poruszona zostanie tematyka ewolucji od grupowania do klasyfikacji, warianty metod branych pod uwagę przez realizatorów projektu, sposoby oceny tych metod a także praktyczne zastosowanie wyników – aplikacja wspierającą użytkowników w ich decyzjach dotyczących kształcenia.*
12:45-13:25	**Wykorzystanie korpusów CLARIN jako zasobów do uczenia sztucznej inteligencji – anotowany korpus dialogów** (M. Oleksy,

	PWr; P. Pęzik, UŁ) *Opis procesu tworzenia anotowanego korpusu dialogów od pozyskiwania konwersacji aż po ich wielowarstwową anotację związaną z ich pragmatycznym wymiarem (m.in. funkcje komunikacyjne poszczególnych segmentów).*
13:25-14:05	**Wykorzystanie korpusów CLARIN jako zasobów do uczenia sztucznej inteligencji – korpusy wielojęzyczne i modele językowe** (J. Kocoń, PWr; R. Roszko, IS PAN; P. Pęzik, UŁ) *Prezentacja wielojęzycznego korpusu MultiEmo anotowany wydźwiękiem oraz wielojęzyczny korpus MultiPar, wykorzystywany do podziału tekstu na akapity. Zaprezentowane zostaną przykłady użycia głębokich modeli tekstowych niezależnych od języka, takich jak LASER oraz LaBSE. Tematem wystąpienia będą również korpusy wielojęzyczne przygotowane przez zespół CLARIN-PL.*
14:05-14:45	**Chronopress – eksploracja diachronicznych korpusów prasowych** (A. Pawłowski, UWr) *Prezentacja Chronopressu – korpusu chronologicznego, stanowiącego modelową reprezentację polskiej prasy wydawanej w latach 1944-1965. Korpusowi prasy towarzyszy zaawansowany interfejs, umożliwiający tworzenie profili oraz szeregów czasowych w odniesieniu do poszukiwanych przez użytkownika słów. Chronopress jest przykładem zintegrowanego zasobu i narzędzia kluczowego dla polskiego prasoznawstwa, medioznawstwa i lingwistyki.*
14:45-15:00	*PRZERWA*
15:00-15:45	**Wykrywanie klauzul abuzywnych we wzorcach umownych – przykład współpracy z UOKiK** (Ł. Augustyniak, PWr) *Prezentacja zbudowanego wspólnie z UOKiK zbioru danych oraz modelu sztucznej inteligencji wykrywającego klauzule abuzywne, czyli niedozwolone postanowienia umowne pojawiające się w umowach konsumenckich.*
15:45-16:15	**CLARIN jako źródło modeli wektorowych** (A. Janz, PWr) *CLARIN jako rozbudowane repozytorium podstawowych i zaawansowanych modeli języka naturalnego. W prezentacji przedstawiony zostanie proces konstrukcji repozytorium z omówieniem poszczególnych rodzajów opracowywanych modeli językowych. Oprócz przedstawienia planów rozszerzenia standardowych modeli statycznych oraz dynamicznych (takich jak modele typu transformer) omówione zostaną również podejścia ukierunkowane na modelowanie słów, fraz, zdań, innych typowych całostek, modele ukierunkowane na rozkłady składniowe zdań oraz modele wzbogacane o wiedzę zewnętrzną, pozakorpusową. Całość zostanie podsumowana przedstawieniem wizji platformy do udostępniania i strojenia modeli na żądanie.*
16:15-16:45	**Pytania, dyskusja** *Podsumowanie warsztatów i dyskusja angażująca uczestników, pozwalająca poznać nieznane jeszcze marzenia badawcze naukowców w obszarze przetwarzania języka naturalnego. Dyskusja na temat przyszłości przetwarzania języka w kontekście nauk humanistycznych i społecznych, możliwości współpracy.*