PRZEGLĄD RUSYCYSTYCZNY 2022, nr 1 (177)

DOI 10.31261/pr.12764

ROMAN ROSZKO
ORCID: https://orcid.org/0000-0002-2291-6939
Instytut Slawistyki Polskiej Akademii Nauk

RUSYCYSTYCZNE ZASOBY I NARZĘDZIA CLARIN-PL

RUSSIAN LANGUAGE RESOURCES AND TOOLS ON THE CLARIN-PL WEBSITE

In this article I present multilingual resources with Russian language already created and currently being constructed by CLARIN-PL consortium. I also introduce the tools and services offered by this consortium for users interested in analysing Russian-language resources.

1. WSTĘP

Zespoły badawcze Instytutu Slawistyki Polskiej Akademii Nauk od lat 70. minionego wieku prowadzą badania kontrastywne języków słowiańskich i bałtyckich. Początkowo badania materiałowe były poprzedzane żmudną ręczną ekscerpcją zasobów drukowanych z wybranych do analizy języków. Wiązało się to z brakiem jakichkolwiek wielojęzycznych kolekcji cyfrowych, tym bardziej narzędzi do ich przetwarzania. Tę niedogodność próbowano na różne sposoby pokonać. Już w XX wieku powstał pierwszy *Eksperymentalny polsko-litewski korpus tekstów dwudziestowiecznych*[1] autorstwa Danuty Roszko i Romana Roszki, który znalazł zastosowanie w wielu studiach nad okre-

[1] Szczegółowy opis projektu jest zawarty w D. Roszko, R. Roszko, *Polsko-litewskie korpusy IS PAN i CLARIN-PL*, w: N. Birgiel, D. Roszko (red.), *Prace Bałtystyczne — 7. Język. Kultura. Literatura*, Uniwersytet Warszawski, Warszawa 2018, s. 185–205. oraz D. Roszko, R. Roszko, *Korpusy wielojęzyczne wkładem Instytutu Slawistyki Polskiej Akademii Nauk w rozwój infrastruktury CLARIN-PL: Przykłady analizy korpusowej nad wołaczem*, w: J. L. Banasiak, A. Kiklewicz, J. Mazurkiewicz-Sułkowska (red.), *Języki słowiańskie dziś — w kręgu kategorii, struktur i procesów*, Instytut Slawistyki PAN — Wydawnictwo Uniwersytetu Łódzkiego, Warszawa — Łódź 2021, s. 281–313.

ślonością-nieokreślonością, modalnością, aspektem i temporalnością w językach polskim i litewskim. Ten udany eksperyment, potwierdzony wieloma publikacjami, doczekał się rozwinięcia w *Eksperymentalnym bułgarsko-polsko-litewskim korpusie* (*Bulgarian–Polish–Lithuanian Corpus*[2]) tworzonym praktycznie przez czteroosobowy polsko-bułgarski zespół w składzie: Ludmila Dimitrova, Violetta Koseska-Toszewa, Danuta Roszko i Roman Roszko. Oba te zasoby nigdy nie doczekały się publikacji ze względu na brak stosownych licencji. Jednak ich część została włączona do nowo powstających polsko-bułgarsko-rosyjskiego i polsko-litewskiego korpusów, konstruowanych przez Zespół Semantyki i Lingwistyki Korpusowej IS PAN (ZSiLK) w strukturach CLARIN-PL. Polskie Konsorcjum CLARIN-PL[3] tworzą zespoły badawcze sześciu instytucji naukowych: Instytutu Podstaw Informatyki PAN, Instytutu Slawistyki PAN, Politechniki Wrocławskiej (lider konsorcjum), Polsko-Japońskiej Akademii Technik Komputerowych, Uniwersytetu Łódzkiego oraz Uniwersytetu Wrocławskiego. Celem Konsorcjum CLARIN-PL jest utworzenie polskiej infrastruktury badawczej na rzecz rozwoju nauk humanistycznych i społecznych w Polsce w tych obszarach badawczych, których podstawą jest analiza danych językowych. W ramach wsparcia badaczy szeroko rozumianych nauk humanistycznych i społecznych Konsorcjum Clarin-PL utworzyło spójną infrastrukturę, którą nieustannie rozwija i dostosowuje do zmieniających się standardów światowych. Jest to otwarta sieć, dostępna dla wszystkich badaczy, zapewniająca prowadzenie badań z wykorzystaniem nowoczesnych metod opartych na nieprzerwanie udoskonalanych technologiach przetwarzania języka naturalnego. Finansowanie działalności Konsorcjum CLARIN-PL rozpoczęło się w roku 2013. W latach 2013–2021 Konsorcjum zrealizowało trzy duże granty dotowane przez Ministerstwo Edukacji i Nauki, tworząc wiele narzędzi i zasobów językowych. Na początku 2020 roku Konsorcjum CLARIN-PL uzyskało kolejne dofinansowanie w Programie Operacyjnym Inteligentny Rozwój 2014–2020 (Priory-

[2] Szerzej na ten temat: L. Dimitrova, V. Koseska-Toszewa, D. Roszko, R. Roszko, *Bulgarian–Polish–Lithuanian Corpus: Current development*, w: C. Vertan, S. Piperidis, E. Paskaleva, M. Slavcheva (red.), *International Workshop: Multilingual resources, technologies and evaluation for Central and Eastern European languages held in conjunction with the International Conference RANLP–2009: PROCEEDINGS*, Borovets 2009, s. 1–8. L. Dimitrova, V. Koseska-Toszewa, D. Roszko, R. Roszko, *Trilingual Aligned Corpus: Current state and new applications*, "Cognitive Studies / Études cognitives" 2014, vol. 14, s. 13–20.

[3] http://clarin-pl.eu/ (11.11.2021).

tet IV: Zwiększenie potencjału naukowo-badawczego, Działanie 4.2: Rozwój nowoczesnej infrastruktury badawczej sektora nauki) na realizację nowych zadań.

W odniesieniu do tytułowych rusycystycznych akcentów w infrastrukturze CLARIN-PL należy przede wszystkim wymienić ukończone korpusy równoległe: polsko-bułgarsko-rosyjski, polsko-rosyjski (v.1 i v.2), rosyjsko-bułgarski, rosyjsko-litewski i rosyjsko-ukraiński oraz konstruowany obecnie zupełnie nowy polsko-rosyjski zasób w ramach szerszego zadania budowy ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusów równoległych.

2. *POLISH–BULGARIAN–RUSSIAN PARALLEL CORPUS*

W latach 2013–2016 ZSiLK w składzie Anna Kisiel (do IX 2015), Violetta Koseska-Toszewa, Natalia Kotsyba, Joanna Satoła-Staśkowiak i Wojciech Sosnowski pracowali nad bazą tekstów współczesnych w językach polskim, bułgarskim i rosyjskim. Wynikiem ich prac jest zasób, któremu nadano nazwę *Polish–Bulgarian–Russian Parallel Corpus*[4]. Objętość tego korpusu wyniosła 6 479 367 słowoform. Łącznie zawiera on 162 jednostki (są to pojedyncze utwory lub większe zbiory), w tym 78 plików to utwory beletrystyczne, 72 — teksty unijne reprezentujące język prawny, 6 — artykuły naukowe, po 3 — teksty prawnicze i religijne. Początkowy zamiar utworzenia korpusu, który składałby się tylko z utworów wzajemnie tłumaczonych (w ramach tej trójki języków) nie został osiągnięty. Ostatecznie tylko 24 jednostki bazują na wzajemnych tłumaczeniach, pozostałe w liczbie 138 są tłumaczeniami z języków trzecich (głównie z języka angielskiego). Zbiory będące wynikiem przeprowadzonych prac zostały opatrzone spójnym opisem metadanych w formacie CMDI i opublikowane w cyfrowym systemie repozytoryjnym *dSpace*[5]. Każdy użytkownik infrastruktury CLARIN-PL może pobrać wszystkie lub tylko wybrane pozycje z tej bazy w formacie TMX[6].

---

[4] A. Kisiel, V. Koseska-Toszewa, N. Kotsyba, J. Satoła-Staśkowiak, W. Sosnowski, 2016, *Polish-Bulgarian-Russian Parallel Corpus*, CLARIN-PL digital repository 2016, http://hdl.handle.net/11321/308 (11.11.2021).

[5] https://clarin-pl.eu/dspace/handle/11321/308 (11.11.2021).

[6] TMX (< ang. *Translation Memory eXchange* / pol. *wymiana pamięci tłumaczeniowej*) — jest to jeden z najlepiej rozpoznawalnych standardów zapisu plików wymiany pamięci tłumaczeniowej (TM < ang. *Translation Memory* / pol. *pamięć tłumaczeniowa*).

3. *POLISH–RUSSIAN PARALLEL CORPUS* (2018)

Po reorganizacji ZSiLK w okresie 2016–2018 skonstruowano pięć dwujęzycznych korpusów równoległych z językiem polskim jako węzłowym, w tym polsko-rosyjski: *Polish–Russian Parallel Corpus*[7]. Zasób ten został opublikowany na stronach CLARIN-PL jednocześnie w cyfrowym systemie repozytoryjnym *dSpace*[8] oraz w wielofunkcyjnym narzędziu *KonText*[9]. Czym jest wspomniany *KonText*? Jest to zaawansowane i nieustannie modernizowane narzędzie, oferujące użytkownikowi zarówno tworzenie własnych (dla języka polskiego automatycznie znakowanych) korpusów, jak i przeszukiwanie wielu udostępnionych w nim zasobów jedno- i wielojęzycznych. Sposoby przeszukiwania zbiorów są bardzo rozbudowane, a zastosowany w zapytaniu język CQL[10] obsługuje atrybuty, operatory, wyrażenia regularne, lematy, tagi, klasy i fleksemy słów, kategorie gramatyczne i metaanotację. Interfejs *KonTextu* jest przejrzysty. Początkujący użytkownik bez problemu uzyska interesujące go profilowanie (sortowanie, filtrowanie) wyszukiwanych form, wyrażeń, konstrukcji oraz konkordancję zarówno w lewym jak i prawym kontekście od wybranej formy szukanej[11]. *KonText* umożliwia automatyczne obliczanie wielu miar. Na stronach Clarin-PL jest dostępna instrukcja wyszukiwarki korpusowej *KonText*[12] oraz tzw. ściągawka instrukcji stosowanych w *KonText*(-cie)[13].

W odniesieniu do opisanej w punkcie 2. wielojęzycznej bazy ten polsko-rosyjski korpus jest jej pewnym rozwinięciem dostosowanym do wymogu zbioru dwujęzycznego. W pracach nad jego powstaniem

---

[7] R. Roszko, W. Sosnowski, M. Duszkin, D. Roszko, R. Tymoshuk, *Polish-Russian Parallel Corpus*, CLARIN-PL digital repository 2018, http://hdl.handle.net/11321/534 (11.11.2021).

[8] https://clarin-pl.eu/dspace/handle/11321/534 (11.11.2021).

[9] Adres wyszukiwarki: https://kontext.clarin-pl.eu/run.cgi/first_form (11.11.2021). Opis *KonText-u*: T. Machálek, *KonText: Advanced and flexible corpus query interface*, w: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, European Language Resources Association 2020, s. 7003–7008.

[10] CQL (< ang. *Corpus Query Language* / pol. *język zapytań korpusowych*) — jest to stworzony z myślą o efektywnym przeszukiwaniu zasobów korpusowych język zapytań.

[11] KWIC (< ang. *Key Word In Context* / pol. słowo kluczowe/węzłowe w kontekście).

[12] https://nextcloud.clarin-pl.eu/index.php/s/fzAZg9xbxA4YEdu (11.11.2021).

[13] https://nextcloud.clarin-pl.eu/index.php/s/IsIriR9v5Hopaml#pdfviewer (11.11.2021).

uczestniczył ZSiLK w składzie: Maksim Duszkin, Danuta Roszko, Roman Roszko, Wojciech Sosnowski i Roman Tymoshuk. Objętość wyniosła 5 615 274 słowoform. Całość korpusu liczy 152 jednostki. Zrównoważenie zasobów względem bazy wyjściowej zostało zmienione. Dołączono nowe teksty reprezentujące rejestr zbliżony do języka mówionego, mianowicie dialogi filmowe w liczbie 101. Język prawny jest reprezentowany przez 26 jednostek, beletrystyka — 12. Pozostałe pliki stanowią teksty naukowe, religijne i techniczne (razem 13). Zasoby nie zostały oznakowane morfologicznie.

Opublikowane w roku 2018 równoległe korpusy CLARIN-PL zostały zastosowane przez wielu badaczy[14] i zespoły badawcze (polsko-bułgarski, polsko-bułgarsko-ukraiński)[15]. Znane są zastosowania tych zasobów w pracach licencjackich (Katedra Językoznawstwa Ogólnego, Migowego i Bałtystyki UW), magisterskich (Instytut Lingwistyki Stosowanej UW, Katedra Językoznawstwa Ogólnego, Migowego i Bałtystyki UW, Wydział Polonistyki UW), doktorskich (UW, UJ) i habilitacyjnych (Instytut Slawistyki PAN).

4. *POLISH–RUSSIAN PARALLEL CORPUS „2"* (2022) ORAZ INNE NOWE KORPUSY DWUJĘZYCZNE Z JĘZYKIEM ROSYJSKIM (2022)

4.1. *POLISH–RUSSIAN PARALLEL CORPUS „2"*

W latach 2018–2021 opisany w punkcie 3. korpus podlegał dalszemu rozwojowi. Zadania były realizowane przez ZSiLK w składzie: Maksim Duszkin, Danuta Roszko, Roman Roszko we współpracy z Jewgieniją Żejmo[16]. W tym okresie zwrócono szczególną uwagę na wstępne prze-

[14] Między innymi warto przywołać prace Jakuba Banasiaka nad semantyką, por. J. Banasiak, *Built-in argument positions in Bulgarian and Polish*. "Cognitive Studies / Études cognitives", 2021(21).

[15] Wypada wskazać zespoły międzyuczelniane i międzynarodowe prowadzące prace nad leksyką w językach słowiańskich, por. W. Sosnowski, J. Satoła-Staśkowiak, *A contrastive analysis of feminitives in Bulgarian, Polish and Russian*. "Cognitive Studies / Études cognitives" 2019 (19), Article 1922; D. Blagoeva, M. Jaskot, W. Sosnowski, *A lexicographical approach to the contrastive analysis of Bulgarian and Polish phraseology*, „Cognitive Studies / Études cognitives" 2019 (19), Article 1923.

[16] Szerzej na ten temat por. M. Duszkin, D. Roszko, R. Roszko, *New parallel corpora of Baltic and Slavic languages — Assumptions of corpus construction*, w: K. Ekštein, F. Pártl, M. Konopík (red.), *Lecture Notes in Artificial Intelligence LNAI 12848: Text, Speech, and Dialogue TSD 2021,* Springer International Publishing, Cham 2021, s. 173–183.

tworzenie już gotowych zasobów, w którego procesie zlikwidowano wiele błędów pisowni, uzgodniono wersje tłumaczeń oraz wyróżniono najmniejsze jednostki (tzw. tokeny). Ponadto w odpowiedzi na propozycje użytkowników CLARIN-PL zmodyfikowano tak normy zrównoleglenia zasobów, by poszczególne wydzielane jednostki były możliwie najmniejsze. Przyjęta na wcześniejszych etapach projektowych zasada jednoczesnego wiązania dwóch równoległych tekstów na poziomie segmentów zdaniowych spełniających warunek zupełności komunikatywnej, doskonale sprawdzająca się w opisie jednego języka, została rozwinięta. We wcześniejszej wersji korpusów CLARIN-PL przyjmowano, że zarówno początek zdania, jak i jego koniec mają następujące formalne wykładniki: wielką literę (jako początek zdania) i kropkę lub wykrzyknik, lub pytajnik, lub wielokropek (jako znak kończący zdanie). Ze względu na ograniczoną objętość artykułu pomijam dodatkowe warunki, które muszą zostać spełnione, by te elementy formalne mogły zostać uznane za początek lub koniec zdania. Zgodnie z tymi założeniami każdy ciąg rozpoczęty wielką literą i zakończony jednym ze wspomnianych znaków końca był uznawany za zdanie, któremu przypisywano zupełność komunikatywną. Jednoczesna segmentacja na poziomie zdania — jak wiadomo — wiązała się ze wskazaniem najmniejszych równoległych partii tekstu, dla których byłby spełniony warunek jednoczesnego zachodzenia w obu językach początku i końca zdania oraz zachowana zgodność treści odczytywanych na płaszczyźnie semantycznej. W rezultacie takich założeń uzgodnione międzyjęzykowe segmenty mogły zawierać nawet po kilka-kilkanaście zdań. W przypadku niektórych utworów beletrystycznych faktycznie cały utwór wypełniał znamiona jednego segmentu zdaniowego, na przykład opowiadanie *Miesto laikrodis* litewskiego pisarza Algisa Kuklysa jest jednym długim zdaniem liczącym około 1500 wyrazów. W związku z takim stanem rzeczy analizowano równoległe zasoby wielojęzyczne i starano się znaleźć inne formalnie wyrażone charakterystyczne miejsca, które warunkowo można uznać za koniec zdania. Wśród nowych znaczników, które w ściśle określonych warunkach mogą być potraktowane jako wskaźnik końca zdania, znalazły się następujące znaki interpunkcyjne: przecinek i nawiasy (szczególnie w tekstach prawnych), dwukropek i średnik (te głównie w utworach beletrystycznych), por. tab. 1–2, a szczególnie w tab. 2 korespondujące ze sobą w rosyjskiej i bułgarskiej wersji tekstu różne znaki przestankowe kropki i średnika.

Tabela 1. Propozycja rozbicia segmentu na kilka mniejszych z wykorzystaniem znaków interpunkcyjnych «;» i «,» w utworze beletrystycznym

Język rosyjski	**Język bułgarski**
Дождь был косой, ниспадала хлещущая сплошная завеса;	*Дъждът плющеше и полегатите му струи ту се засилваха, ту отслабваха;*
изредка перед глазами Джима вставали грозно надвигающиеся волны,	*от време на време пред очите на Джим се появяваха страшно надигащи се вълни;*
маленькое суденышко металось у берега;	*малкото корабче се мяташе край брега;*
неподвижные строения вырисовывались в плавучем тумане;	*неподвижните постройки се открояваха сред плаващата мъгла;*
тяжело раскачивались широкие паромы на якоре,	*тежко се люшкаха закотвените широки фериботи;*
поднимались и опускались огромные пристани, задушенные брызгами.	*издигаха се и се спускаха обширните кейове, задавени от пръски.*

Tabela 2. Propozycja rozbicia segmentu na kilka mniejszych z wykorzystaniem znaków interpunkcyjnych «;» w języku rosyjskim, formalnie odpowiadających «.» w bułgarskim (utwór beletrystyczny).

Język rosyjski	**Język bułgarski**
Глаза твои голубиные под кудрями твоими;	Очите ти са гълъбови под твоите къдри.
волосы твои — как стадо коз, сходящих с горы Галаадской;	Косата ти е като стадо кози, кога слиза от Галаатската планина.
зубы твои — как стадо выстриженных овец, выходящих из купальни, из которых у каждой пара ягнят, и бесплодной нет между ними;	Зъбите ти като стадо овце, кога излиза от къпалня, от които всяка с по две агънца, и ялова няма помежду им.

W tekstach prawnych i prawniczych zdecydowano się na uwzględnienie w segmentacji miejsc wyznaczonych strukturą dokumentu, por. tab. 3.

Tabela 3. Propozycja rozbicia zdania na segmenty zgodne ze strukturą dokumentu prawnego (języki polski i litewski)

Język polski	**Język litewski**
Dziennik Urzędowy Unii Europejskiej	Europos Sąjungos oficialusis leidinys
C 336 / 19	C 336 / 19
z dnia 28 czerwca 2011 r.	2011 m. birželio 28 d.
w sprawie E–18 / 10	byloje E–18 / 10
Urząd Nadzoru EFTA przeciwko Królestwu Norwegii	ELPA priežiūros institucija prieš Norvegijos Karalystę
(Niewykonanie wyroku Trybunału stwierdzającego uchybienie zobowiązaniom — artykuł 33 porozumienia o nadzorze i Trybunale — środki niezbędne do wykonania wyroku Trybunału)	(Įsipareigojimų neįvykdymą konstatuojančio Teismo sprendimo neįvykdymas — Institucijos ir Teismo susitarimo 33 straipsnis — Priemonės Teismo sprendimui įvykdyti)
2011 / C 336 / 13	2011 / C 336 / 13

Zgodnie z nowo opracowanymi zasadami segmentacji na poziomie zdania dokonano ręcznej korekty/aktualizacji zrównoleglenia zasobów. Następnie wszystkie tokeny oznakowano w automatycznym procesie. Całość polskich tekstów otagowano narzędziem *MorphoDiTa-PL*[17], rosyjskie zaś — *UDPipe* (*russian-gsd-ud–2.6–200830*)[18], por. przykład [1]. Poszczególne jednostki opatrzono takimi metadanymi jak: text.txtype (zaszeregowanie tekstu do odpowiedniego rejestru), text.srclang (język oryginału), text.original (oznaczenie rozróżniające teksty oryginalne i tłumaczenia, przewidziano również wartość "unknown"), text.author (autor tekstu), text.translator (autor tłumaczenia), text.srctitle (nazwa tekstu w języku tekstu pierwotnego), text.title (nazwa utworu w danym języku), text.year (data publikacji utworu), text.period (zaszeregowanie utworu do wyróżnionych okresów), text.url (adres internetowy), por. przykład [2].

[17] https://ws.clarin-pl.eu/morphoDiTa.shtml (11.11.2021). M. Piasecki, W. Walentynowicz, *MorphoDiTa-based tagger adapted to the Polish language technology*, w: *Proceedings of Human Language Technologies as a Challenge for Computer Science and Linguistics*, LTC 2017, Poznań 2017 s. 377–381.

[18] http://lindat.mff.cuni.cz/services/udpipe/ (11.11.2021). M. Straka, J. Straková, *UDPipe*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Prag 2016, http://hdl.handle.net/11234/1–1702 (11.11.2021).

[1] Oznakowanie segmentu 4 w języku rosyjskim: *Международный инструментальный конкурс «ЗОЛОТЫЕ КЛАВИШИ» — ONLINE*[19].

```
# sent_id = 4
1	Международный	международный	ADJ	JJL	Case=Nom/Degree=Pos/Gender=Masc/Number=Sing	_	_	_	_
2	инструментальный	инструментальный	ADJ	JJL	Case=Nom/Degree=Pos/Gender=Masc/Number=Sing	_	_	_	_
3	конкурс	конкурс	NOUN	NN	Animacy=Inan/Case=Nom/Gender=Masc/Number=Sing	_	_	_	_
4	«ЗОЛОТЫЕ	«золотой	ADJ	JJL	Case=Nom/Degree=Pos/Number=Plur	_	_	_	_
5	КЛАВИШИ»	клавиши»	NOUN	NN	Animacy=Inan/Case=Nom/Gender=Fem/Number=Plur	_	_	_	_
6	–	–	PUNCT	(	_	_	_	_	_
7	ONLINE	ONLINE	X	FW	Foreign=Yes	_	_	_	_
```

Widoczne w przykładzie [1] zapisy nie są widoczne dla użytkownika, jednak mają niebagatelne znaczenie podczas przeszukiwania korpusu. Wszystkie przypisane formom parametry mogą być częścią zapytania. Spójrzmy na 4. słowoformę (*ЗОЛОТЫЕ*) w przykładzie [1]. W jej opisie pojawia się lemat *золотой* zgodny z formą hasłową w słowniku, następnie po nim następują tagi, które wskazują na wartości morfologiczne przypisane tej słowoformie (*ЗОЛОТЫЕ*) w danym kontekście: część mowy = przymiotnik / przypadek = mianownik / stopień = równy / liczba = mnoga.

Poniżej zamieszczony zostaje plik informacyjny z metadanymi opisującymi tekst zawierający przedstawione w przykładzie [1] zdanie nr 4 (# sent_id = 4).

[2] Przykładowy plik z metadanymi do tekstu rosyjskojęzycznego.

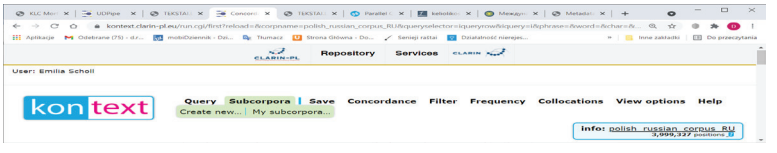
```
text.txtype"journalism"
text.srclang"bg Bulgarian"
text.original"no"
text.author"European Association of Folklore Festivals"
text.translator"European Association of Folklore Festivals"
text.srctitle"ЗЛАТНИ КЛАВИШИ"
text.title"ЗОЛОТЫЕ КЛАВИШИ"
```

[19] W pozycji 5. zastosowany przez nas do znakowania zasobów rosyjskich program *UDPipe* (russian-gsd-ud–2.6–200830) wygenerował lemat *клавиши* w formie liczby mnogiej zamiast spodziewanego *клавиша*. Wariant lematu w liczbie pojedynczej jest generowany przez *UDPipe* (russian-taiga–2.6–200830). Wybór zastosowanego modelu do znakowania był poprzedzony testami, które wskazywały na statystycznie wyższą trafność russian-gsd-ud-2.6-200830.

text.year"2021"
text.period"after 1990"
text.url"https://eaff.eu/ru/championships/1011–2811-international-instrumental-competition-golden-keys-online"

Parametry zawarte w metadanych również mogą być częścią zapytania. Na przykład użytkownik zajmujący się użyciem przymiotników po roku 1990 w języku rosyjskim powinien skorzystać z dwu możliwości zawężenia przeszukiwanych plików do datowanych na rok 1990 i późniejsze. Pierwszym sposobem jest utworzenie własnego podkorpusu składającego się z plików spełniających wybraną wartość "after 1990" dla parametru text.period, por. przykład [3], drugim zaś — każdorazowe zaznaczanie poniżej pola zapytania w części text.period wartości "after 1990".

[3] KonText: Menu tworzenia i wyboru własnych podkorpusów.



Powracając do badań nad rosyjskim przymiotnikiem w okresie po 1990 roku użytkownik ponadto winien w polu wpisywania zapytania zamieścić właściwą formułę, por. przykład [4].

[4] Zapytanie o wyświetlenie wszystkich zdań/segmentów zawierających przymiotnik w zasobach rosyjskojęzycznych.
[tag="ADJ.*"]

W przykładzie [4] nawiasy kwadratowe wyznaczają granicę słowoformy, której przypisano jeden parametr *tag* o wartości *ADJ* (przymiotnik). Następujące po *ADJ* znaki .* oznaczają, że inne doprecyzowujące wartości są obojętne. Jak można zauważyć, wartość parametru (w tym przypadku *tag*) jest umieszczana w cudzysłowie (tu: "ADJ.*").

W poniższym przykładzie [5] zostaje podana składnia zawężająca parametry wyszukiwania. Załóżmy, że badacz jest zainteresowany użyciem przymiotników rosyjskich rozpoczynających się tylko na literę э. W tym celu rozbudowuje formułę o parametr lexem (lemat), któremu w cudzysłowie przypisuje wartość э.* (tj. "э.*") i dodaje operator koniunkcji &, by obie wartości wymienionych parametrów zachodziły jednocześnie.

[5] Rozbudowa zapytania z przykładu [4] zawężająca wynik do zdań/segmentów zawierających przymiotnik rozpoczynają się od litery *э* (np. *этнологический*, *экономические*, *эстетического* itd.) w języku rosyjskim.
[lexem="э.*" & tag="ADJ.*"]

W przykładzie [6] zostaje przedstawione zapytanie o dwuwyrazowe wyrażenia złożone zawierające przymiotnik rozpoczynający się od litery *э*, po którym bezpośrednio następuje rzeczownik rozpoczynają się od litery *ф*.

[6] Rozbudowa zapytania z przykładu [5] zawężająca wynik do zdań/segmentów zawierających dwuwyrazowe złożenia przymiotnika (rozpoczynającego się od litery *э)* i rzeczownika (rozpoczynającego się od litery *ф*) (np. *этнологического факультета*, *экономический форум*, (*на*) *эстетическом фоне* itd.) w języku rosyjskim.
[lexem="э.*" & tag="ADJ.*"] [lexem="ф.*" & tag="NOUN.*"]

Wspomniane w tej części artykułu wstępne przetworzenie zasobów objęło również wiele kolejnych istotnych dla użytkownika korekt. Wprowadzono znaki umowne, które informują badacza o braku ekwiwalentu w jednym z języków korpusu. W tym celu zastosowano skrót [—], por. przykład [7]. Kolejny wprowadzony znak [...] informuje o skróceniu tekstu o nieistotne z punktu widzenia wielojęzycznej analizy fragmenty, por. przykład [8]. Ten drugi znak początkowo stosowano w celu ukrycia danych wrażliwych. W obecnie opracowywanych korpusach (o których poniżej w punkcie 5.) odstąpiono od tak przeprowadzanej anonimizacji na rzecz inteligentnej anonimizacji, która w pełni zachowuje strukturę tekstu, por. przykład [9].

[7] Użycie znaku umownego [—] na oznaczenie braku w którymkolwiek języku ekwiwalentnego segmentu.

PL	RU
Chcesz w pysk, to chodź!	Тебе нужны неприятности?
[—]	Сейчас они будут!
Podoba ci się?	Ну что, доволен?

[8] Użycia znaku umownego [...] na oznaczenie pominiętych fragmentów na etapie wstępnego przetworzenia (przykład polski).

Tekst na wejściu	Tekst po przetworzeniu
REPUBLIKA LITEWSKA	REPUBLIKA LITEWSKA
LT	LT

Instytucja Państwowa „Regitra”	Instytucja Państwowa „Regitra”
DOWÓD REJESTRACYJNY	DOWÓD REJESTRACYJNY
Свидетелство за регистрация / Permiso de circulation / Osvědčení o registraci / Registreringsattest / Zulassungsbescheinigung / Rcgistreerimistunnistus / Πιστοποιητικό εγγραφής / Registration certificate / Certificat d’immatriculation / Prometna dozvola / Teastas Claráithe / Carta di circolazione / Reģistrācijas apliecība / Forgalmi engedély / Čertifikat ta’ Reġistrazzjoni / Kentekenbewijs / Dowód Rejestracyjny / Certificado de matrícula / Certificat de Înregistrare / Osvedčenie o evidencii / Prometno dovoljenje / Rekisteröintitodistus / Registreringsbeviset	[…]

[9] Inteligentna anonimizacja w języku polskim.

Tekst na wejściu	**Tekst ze zidentyfikowanymi nazwami wrażliwymi**	**Tekst po przetworzeniu (po inteligentnej anonimizacji)**
13 grudnia 1982 roku w *Warszawie* w *Alejach Jerozolimskich* został zatrzymany uczestnik strajku *Jan Kowalczyk* z narzeczoną *Joanną Nowak*.	[DATE] roku w [MIEJSCE] w [MIEJSCE] został zatrzymany uczestnik strajku [OSOBA] z narzeczoną [OSOBA].	*21 grudnia 1900* roku w *Arkadii* na *ulicy Cichej* został zatrzymany uczestnik strajku *Ananiasz Baran* z narzeczoną *Elżbietą Jerzy*.

W okresie 2018–2021 obok prac nad dopracowaniem już zebranych zasobów, podjęta została rozbudowa korpusu polsko-rosyjskiego o kolejne teksty współczesne. Zgodnie z sugestiami dotychczasowych i potencjalnych użytkowników głównie skupiono się na pozyskaniu nowych zasobów reprezentujących rejestr jak najbardziej zbliżony do mowy, czyli na dialogach filmowych, oraz na dokumentach prawnych i naukowych. Szczególną uwagę zwrócono na jakość i wzajemną międzyjęzykową adekwatność części nieautoryzowanych tłumaczeń

list dialogowych. Podczas sprawdzania zrównoleglenia korektor wychwytywał niezgodności między obiema wersjami językowymi (na poziomie segmentów zdaniowych i leksykalnym) oraz niekonsekwencje tłumaczenia w każdym języku niezależnie[20]. Na przykład występujący w wersji polskiej *płaszcz przeciwdeszczowy* w rosyjskiej wersji listy dialogowej raz był nazwany *шинель* innym razem — *пальто*. Należy zauważyć, że w wersji rosyjskiej dochodzi do nazywania tego samego przedmiotu dwoma różnymi leksemami, które w żadnym stopniu nie są leksykalnymi ekwiwalentami polskiego *płaszcz przeciwdeszczowy*. Podobną rozbieżność na poziomie leksykalnym odnotowujemy w formach polskiej *kamizelka kuloodporna* i rosyjskiej *свитерок*. Dużo uwagi poświęcono ujednoliceniu zapisu nazw własnych, na przykład w wersji polskiej trzy formy imienia *Brian*, *Brajan*, *Brajen* zastąpiono jedną *Brian*, podobnie leksemy *Samanta* i *Samantha* zastąpiono jednym *Samantha*. W przypadku podwójnego zapisu *izraelici* i *Izraelici* wskazano postać *Izraelici* jako zgodną z normami języka polskiego. Ponadto korygowano rodzaj gramatyczny zaimków osobowych, by odpowiadał płci bohatera, wyrównywano rozbieżności w użyciu liczby pojedynczej i mnogiej[21]. W przykładzie [10] zebrano kilka egzemplifikacji rozbieżności formalnych i merytorycznych między językami polskim i rosyjskim oraz przedstawiono decyzję korektorów-tłumaczy. W wielu niejednoznacznych kontekstowo zapisach sprawdzano trafność tłumaczenia z wersją oryginalną listy dialogowej.

[10] Przykłady różnych strategii tłumaczeniowych w językach polskim i rosyjskim oraz (nie)wprowadzone zmiany w polsko-rosyjskich zasobach.

	PL	**RU**	**Decyzja korektora**
1	Nasza matka nas nie urodziła. // Zrobiła to taka inna pani.	Наша мама — не наша родная мама. // Другая женщина — наша родная мама.	Pozostawiono bez zmian.
2	Zostało mi jeszcze na jedną.	У меня хватит индейки еще на один сэндвич.	Pozostawiono bez zmian.
3	Stacja metra *State and Balboa*.	Станция метро, *переход со Стейт на Балбоа*.	Pozostawiono bez zmian ze względu na odmienne realia polskie i rosyjskie związane z infrastrukturą metra.

[20] Błędy pisowni zostały wyeliminowane podczas wstępnego przetwarzania. Błędy gramatyczne (np. użycie niewłaściwego przypadka), niektóre błędy interpunkcyjne były korygowane przez korektora-tłumacza.

[21] W języku rosyjskim często należało zaimek *они* zamieniać na *он* lub *она*.

4	Ona też w tym siedzi.	Ты был прав по поводу этого дерьма.	Pozostawiono bez zmian.
5	Tego chcesz?	Нет.	Utworzono dwa segmenty, dodając dwukrotnie znak [—]: *Tego chcesz?* / [—] ‖ [—] / *Нет.*
6	Co *tu* się dzieje?	Что *со мной* происходит?	Zmiana zdania rosyjskiego jako wynik sprawdzenia oryginału: *Co tu się dzieje?* / *Что здесь происходит?*
7	Proszę, *podejdź*.	Пожалуйста, *проходи*.	Zmiana zdania rosyjskiego jako wynik sprawdzenia oryginału: *Proszę, podejdź.* / *Пожалуйста, подойди.*
8	Musiał użyć *defibrylatora*.	Врач был вынужден применить *электрошокер*.	Zmiana zdania rosyjskiego jako wynik sprawdzenia oryginału: *Musiał użyć defibrylatora.* / *Ему пришлось воспользоваться дефибриллятором.*
9	Ty *pierwszy*.	Ты *первая*.	Zmiana zdania rosyjskiego powodowana koniecznością zachowania zgodności odniesienia do współrozmówcy-mężczyzny: *Ty pierwszy.* / *Ты первый.*
10	*Marzyłem* o tobie.	Ты мне часто *снилась*.	Zmiana treści zdania rosyjskiego na bardziej odpowiadającą treści filmu. W angielskim oryginale został użyty leksem *dream*: *Marzyłem o tobie.* / *Я мечтал о тебе.*
11	Płaskie czy *na obcasach*? (buty)	Плоские или *на платформе*?	Zmiana leksemu w zdaniu rosyjskim po obejrzeniu sceny: *Płaskie czy na obcasach?* / *Плоские или на каблуках?*

12	*Idź*, kochanie.	*Продолжай*, дорогая.	Zmiana treści zdania rosyjskiego, by ją uzgodnić z kontekstem: *Idź, kochanie.* / *Иди, дорогая.*
13	Dlaczego?	Ты только приехал, что не так?	Zdecydowano się na wprowadzenie znaku [—] w języku polskim: [—] *Dlaczego?* / *Ты только приехал, что не так?*
14	Chcę, byś włożył kurczaka do piekarnika i ustawił na *375*.	Мне нужно, чтобы ты поставила курицу в духовку на *375* градусов.	Zmiana wskazań temperatury w obu językach, by jej wartość została właściwie odczytana przez polskich i rosyjskich odbiorców (zastosowano wskazania w stopniach Celsjusza): *Chcę, byś włożył kurczaka do piekarnika i ustawił na 190.* / *Мне нужно, чтобы ты поставила курицу в духовку на 190 градусов.*
15	Wirus po 24 godzinach doprowadza do zgonu.	Вирус станет инертным через 24 часа.	Zmiana wprowadzona w zdaniu polskim, by odpowiadała realiom przedstawianym w filmie: *Wirus staje się obojętny po 24 godzinach.* / *Вирус станет инертным через 24 часа.*
16	*Być czym?*	*Где?*	Usunięto niepowiązane z kontekstem segmenty (nie znaleziono oryginału tej listy).

[11] Przykład spektakularnej zmiany dokonanej przez korektora-tłumacza w zasobach polsko-ukraińskich.

PL	**UK**
Wersja zgodna z zapisem w liście dialogowej	
Byliśmy na wycieczce w Universal Studio, a potem w woskowym muzeum Hollywood.	Ми ходили на екскурсію по кіностудії Довженка, а потім Чарлі водив мене у музей воскових фігур Верховної Ради.

Wersja po zmianach korektora-tłumacza[23]	
Byliśmy na wycieczce w Universal *Studios*, a potem w *Muzeum Figur Woskowych* w Hollywood.	Ми ходили на екскурсію по *Студії Universal*, а потім Чарлі водив мене в *Голлівудський музей воскових фігур*.

Korekta list dialogowych (głównie polsko-rosyjskich i sporadycznie polsko-ukraińskich) dotyczyła tylko części z nich[22]. Autoryzowane przekłady nie podlegały jakimkolwiek zmianom. Informacja o wprowadzeniu tego typu poprawek w wybranych tekstach została zawarta w skorelowanych z nimi plikach metadanych. Decyzję o korekcie części list dialogowych podjęto z myślą o uczeniu maszynowym. Otóż złe dane wejściowe prowadzą do wytworzenia błędnych międzyjęzykowych modeli, a te później zastosowane w translatorach generują błędne tłumaczenia.

4.2. KORPUSY ROSYJSKO-* -BUŁGARSKI, -LITEWSKI I -UKRAIŃSKI (2022)

Równolegle z pracami nad polsko-rosyjskim korpusem prowadzono konstrukcję dziesięciu nowych korpusów, w tym trzech z językiem rosyjskim: rosyjsko-litewskiego, rosyjsko-bułgarskiego i rosyjsko-ukraińskiego. Nie są to duże zasoby, zwłaszcza rosyjsko-litewskie. Zrównoważenie tych kolekcji jest różne. Najlepsze udało się osiągnąć dla pary rosyjsko-bułgarskiej, która zawiera 44 jednostek, w tym 27 reprezentujących rejestr prawny, 11 — beletrystykę (wśród nich jeden utwór literatury dziecięcej), 3 — rejestr prawniczy, 2 — rejestr tekstów naukowych, 1 — rejestr dziennikarski. Zrównoważenie części rosyjsko-ukraińskiej jest następujące: 25 jednostek reprezentujących język prawny, 3 — to teksty naukowe i 1 utwór beletrystyczny. Korpus rosyjsko-litewski zawiera tylko kilka tekstów prawniczych. Wszystkie te zasoby zostały wstępnie przetworzone, ręcznie zrównoleglone, automatycznie oznakowane i opatrzone metadanymi. Do

[22] Podczas konstrukcji korpusów rozważaliśmy pominięcie niedoskonałych przekładów, jednak szczegółowa analiza tych „amatorskich" tłumaczeń dialogów filmowych wykazała użycie w nich wielu form, wyrażeń i konstrukcji typowych dla języka potocznego. Uznaliśmy więc, że warto niektóre z tych tekstów „uratować" poprzez naniesienie niezbędnej korekty językowej. Wprowadzane przez nas zmiany są wręcz konieczne w przypadku wykorzystania tych zasobów w uczeniu, testowaniu i konstrukcji nowych narzędzi lingwistycznych.

[23] Pochyleniem zaznaczono zmieniane fragmenty zdań w obu językach. Korekta w języku ukraińskim jest o wiele dalej idąca. Została ona podyktowana miejscem, w którym toczy się akcja filmu (Hollywood). Należy tu podkreślić, że przyjęta przez ukraińskiego tłumacza strategia jest niezrozumiała.

znakowania tekstów rosyjskojęzycznych zastosowano znane już narzędzie *UDPipe*[18], dla litewskich — *KLC Morfologijos servisas*[24], bułgarskich — *CL@RK System*[25] i ukraińskich *UDPipe* (*ukrainian-iu-ud–2.6200830*)[18].

W 2022 roku jest planowana publikacja tych korpusów na stronach CLARIN-PL w przeglądarce *KonText* oraz cyfrowym repozytorium *dSpace*.

5. RĘCZNIE ZRÓWNOLEGLONY I RĘCZNIE ZNAKOWANY POLSKO-ROSYJSKI KORPUS TEKSTÓW RÓWNOLEGŁYCH

Od 2020 roku jest realizowany projekt CLARIN-PL-BIZ w ramach „Programu Operacyjnego Inteligentny Rozwój 2014–2020" w osi priorytetowej IV „Zwiększenie potencjału naukowo-badawczego" i działaniu 4.2 „Rozwój nowoczesnej infrastruktury badawczej sektora nauki". Jego celem jest utworzenie platformy badawczo-rozwojowej do przetwarzania języka naturalnego i eksploracji dużych zasobów danych językowych. Obejmuje ona wiele zadań, w tym:

— utworzenie Centrum Technologicznego (CTech) jako bazy dla technologii eksploracji danych językowych;

— zastosowanie w CTech rozbudowanych technologii językowych do inteligentnego przetwarzania dużych różnorodnych danych na niewspieranych dotychczas płaszczyznach;

— opracowanie i wdrożenie właściwych standardów konstrukcji zasobów i narzędzi językowych;

— zaprojektowanie i wytworzenie nowych narzędzi analizy danych językowych;

— przygotowanie i dostarczenie nowych danych ręcznie znakowanych do badań i trenowania narzędzi językowych[26], a wśród nich również polsko-rosyjskiego korpusu[27].

---

[24] Przygotowując korpusy, uzyskaliśmy dostęp do najnowszej wersji tagera języka litewskiego *KLC Morfologijos servisas*, opracowanego przez Centrum Lingwistyki Komputerowej Uniwersytetu Witolda Wielkiego w Kownie. W momencie przygotowania tego artykułu wspomniany tager nie został jeszcze udostępniony.

[25] K. Simov, A. Simov, P. Osenova, *An XML architecture for shallow and deep processing*, w: *The Proceedings of the ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP*, ESSLLI 2004, s. 51–60.

[26] Szerzej na temat projektu por. R. Roszko, *O nowych ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusach równoległych oraz ich zastosowaniach*, „Acta Baltico-Slavica" 2021, t. 45, artykuł 2576 oraz https://clarin.biz/ (11.11.2021).

[27] Trwają prace nad czterema tego typu korpusami. Obok polsko-rosyjskiego tworzone są korpusy polsko-słoweński, polsko-litewski i polsko-bułgarski.

Odbiorcami ręcznie zrównoleglonych i znakowanych zasobów równoległych będą przede wszystkim programiści, informatycy zajmujący się projektowaniem i wytwarzaniem narzędzi do przetwarzania języka naturalnego (NLP)[28] oraz badacze reprezentujący szeroko rozumiane nauki humanistyczne i społeczne. Należy zwrócić uwagę, że ręcznie zrównoleglone i znakowane korpusy (w tym polsko-rosyjski) znajdą zastosowanie w definiowaniu algorytmów projekcji znaczeń z jednego języka na drugi. Szczególnie chodzi tu o projekcję tych znaczeń, które w jednym języku są jednoznacznie wyrażane na płaszczyźnie formalnej, w drugim zaś — dochodzi do tzw. niedopowiedzenia językowego[29]. Jak wiadomo, ręcznie zrównoleglone i znakowane korpusy równoległe odgrywają dużą rolę w dopracowaniu narzędzi do automatycznego wyrównywania zasobów dwujęzycznych, co w konsekwencji przekłada się na wzrost liczby i jakości nowych kolekcji wielojęzycznych generowanych automatycznie. Nie można nie wspomnieć o znaczeniu tych zasobów w rozwoju rekurencyjnych sieci neuronowych, mających zastosowanie w przekładzie maszynowym i rozwoju sztucznej inteligencji.

5.1. ZASADY KONSTRUKCJI RĘCZNIE ZRÓWNOLEGLONEGO I ZNAKOWANEGO POLSKO-ROSYJSKIEGO KORPUSU RÓWNOLEGŁEGO

W realizowanym projekcie CLARIN-PL-BIZ na przełomie lat 2020/2021 przeprowadzono gruntowną analizę istniejących zasobów cyfrowych dla języków słowiańskich i bałtyckich[30], a następnie wypracowano metodologię konstrukcji poszczególnych ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusów równoległych. Dużą wagę przywiązano do selekcji utworów, w tym oceny poprawności językowej tekstów oraz wewnętrznego zrównoważenia

[28] Przykłady narzędzi NLP: systemy do przechowywania i udostępniania danych językowych, wyszukiwarki korpusowe, analizatory cech gramatycznych, składniowych, stylometrycznych, anotatory znakujące/kodujące zasoby językowe (np. tagery, lematyzatory), syntezatory mowy, systemy do przetwarzania mowy itd.

[29] V. Koseska, R. Roszko, *On semantic annotation in CLARIN-PL parallel corpora*, "Cognitive Studies / Études cognitives" 2015, nr 15, s. 211–236. https://doi.org/10.11649/cs.2015.016 (11.11.2021).

[30] Wyniki badań przedstawiono w czterech obszernych raportach CLARIN-PL-BIZ autorstwa Jakuba Banasiaka (IS PAN), Pawła Kowalskiego (IS PAN), Danuty Roszko (UW) i Romana Roszko (IS PAN). Szerzej na temat raportów por. R. Roszko, *O nowych ręcznie zrównoleglonych i znakowanych…*

zasobów. Przewidziano korektę pisowni, w przypadku danych wrażliwych także „inteligentną" anonimizację, por. wyżej przykład [9]. Segmentacja na poziomie zdania (zrównoleglanie/wyrównywanie) oraz tokenów (znakowanie/tagowanie) będzie przeprowadzana dwuetapowo. W pierwszym etapie dwa niezależne zespoły dokonają ręcznej segmentacji i znakowania. W drugim zaś — zespół superanotatorów w przypadkach rozbieżności wskaże właściwą segmentację i znakowanie. Wszystkie pliki włączone do zasobu będą posiadały rozbudowane metadane.

Rozważane jest pozyskiwanie quasi-równoległych danych w automatycznym przeszukiwaniu zasobów sieciowych z wykorzystaniem narzędzi LASER[31] czy LAMBERT[32]. Decyzja o włączeniu tego typu danych do polsko-rosyjskiego korpusu jeszcze nie zapadła. Quasi-równoległe dane na pewno uzupełniłyby rejestr dziennikarski, który głównie tworzyłyby krótkie doniesienia prasowe oraz wypowiedzi znanych w świecie przywódców duchowych czy polityków. Należy jednak zważyć, że tego typu dane to przede wszystkim krótkie/hasłowe informacje prasowe o quasi-równoległym charakterze. Ponadto takie dane zamierza w tym projekcie gromadzić zespół prof. Piotra Pęzika, którego quasi-równoległe dane wraz z ręcznie zrównoleglanymi i znakowanymi dwujęzycznymi korpusami będą stanowić podstawę do budowy międzyjęzykowych modeli na rzecz maszynowego przekładu i dalszego rozwoju sztucznej inteligencji.

6. *MULTIEMO* — NARZĘDZIE DO WIELOJĘZYCZNEJ ANALIZY SENTYMENTU

MultiEmo[33] jest nowym, modelowym zbiorem danych utworzonym na cele wielojęzycznej analizy sentymentu. *MultiEmo* wywodzi się

[31] V. Chaudhary, Y. Tang, F. Guzmán, H. Schwenk, P. Koehn. *Lowresource corpus filtering using multilingual sentence embeddings*, w: O. Bojar, i in. (red.), *Proceedings of the Fourth Conference on Machine Translation (WMT)*, Association for Computational Linguistics 2019.

[32] Ł. Garncarek, R. Powalski, T. Stanisławek, B. Topolski, P. Halama, M. Turski, F. Graliński, *LAMBERT: Layout-aware language modeling for information extraction*, w: J. Lladós, D. Lopresti, S. Uchida (red.), *Document Analysis and Recognition — ICDAR 2021*, Springer International Publishing 2020, s. 1–16.

[33] J. Kocoń, P. Miłkowski, K. Kanclerz, *MultiEmo: Multilingual, Multilevel, Multidomain Sentiment Analysis Corpus of Consumer Reviews*, w: M. Paszynski, D. Kranzlmüller, V.V. Krzhizhanovskaya, J.J. Dongarra, P.M.A. Sloot (red.), *Computational Science — ICCS 2021. ICCS 2021. Lecture Notes in Computer Science*, t. 12743, Springer, Cham 2021.

PolEmo (v. 1), zawierającego 57 466 zdań w 8216 różnych dokumentach/recenzjach/opisach. Dane *PolEmo* zostały ręcznie anotowane w warstwie sentymentu zarówno na poziomie całego tekstu, jak i poszczególnych zdań. Podczas tworzenia *MultiEmo* dane *PolEmo* zostały automatycznie przetłumaczone na dziesięć języków, w tym na język rosyjski. W wyniku zastosowanej anotacji dane korpusowe mogą być analizowane na poziomie całego tekstu lub jego części — zdań i akapitów.

Korpusy *PolEmo* i *MultiEmo* zawierają zbiory recenzji konsumenckich z zakresu usług medycznych, hotelarskich, obsługi konsumenckiej i obszaru powiązanego z działalnością uniwersytecką. Narzędzie *MultiEmo* jest publicznie dostępne na licencji Creative Commons Attribution 4.0 International[34].

7. *TAGERML*

Bezpośrednio na stronie CLARIN-PL użytkownik uzyskuje dostęp do wielojęzycznego anotatora *TagerML*[35] z zaimplementowany funkcjami obsługi tekstów rosyjskojęzycznych. Ten tager odwołuje się do *UDPipe*[36].

8. *INFOREX*

Również na stronie CLARIN-PL znajduje się kolejne narzędzie *Inforex*[37], które służy do komputerowego wspomagania ręcznego anotowania własnych zasobów. Tworzący rosyjskojęzyczne korpusy mogą w procesie znakowania rosyjskich tekstów skorzystać z oferowanej w *Inforexie* funkcji automatycznego przypisywania poszczególnym formom sugerowanych morfosyntaktycznych znaków. W tym celu *Inforex* komunikuje się z tym samym co *TagerML* narzędziem *UDPipe*.

---

[34] Zasoby do pobrania: J. Kocoń, K. Kanclerz, P. Miłkowski, B. Bojanowski, M. Zaśko-Zielińska, *PolEmo 1.0 + MultiEmo-Test 1.0 Multilingual Sentiment Analysis Dataset for KES2020*, CLARIN-PL digital repository 2020, http://hdl.handle.net/11321/737 (11.11.2021), J. Kocoń, P. Miłkowski, K. Kanclerz, *MultiEmo: Multilingual, Multilevel, Multidomain Sentiment Analysis Corpus of Consumer Reviews*, CLARIN-PL digital repository 2021, http://hdl.handle.net/11321/798, (11.11.2021).

[35] http://ws.clarin-pl.eu/tagerml.shtml (11.11.2021).

[36] http://lindat.mff.cuni.cz/services/udpipe/ (11.11.2021), por. też M. Straka, J. Straková, *UDPipe*...

[37] https://inforex.clarin-pl.eu/.

9. KORPUSY WŁASNE UŻYTKOWNIKÓW PLATFORMY CLARIN-PL

CLARIN-PL wspomaga użytkowników w tworzeniu własnych zasobów. Każdy użytkownik infrastruktury CLARIN po załadowaniu swojego jednojęzycznego korpusu na stronę *Repozytorium dSpace* uzyskuje po wstępnym przetworzeniu danych dostęp do swoich zbiorów w przeglądarce webowej *KonText*. W przypadku wielojęzycznych zbiorów niezbędna jest konsultacja z Centrum Wiedzy PolLinguaTec CLARIN-PL, które udziela aktywnego wsparcia merytorycznego wszystkim użytkownikom w planowaniu badań, wyborze właściwych metod i narzędzi.

Infrastruktura CLARIN-PL jest nieustannie integrowana z wieloma światowymi repozytoriami narzędzi i zasobów typu Linked Open Data (otwartych danych połączonych). Dla badacza oznacza to możliwość natychmiastowego dostępu do danych połączonych z CLARIN-PL jedną wielką siecią bez konieczności przeszukiwania przepastnych zasobów internetowych.

10. INNE PLANOWANE ZADANIA

Obok konstrukcji nowych korpusów (ręcznie znakowanych i zrównoleglonych) będą prowadzone prace nad utrzymaniem już utworzonych zasobów, które obejmą korektę dostrzeżonych błędów oraz nieustanne przystosowywanie parametrów do zmieniających się standardów. Ponadto planowane jest utworzenie konstruktora zapytań — wspólnego dla wszystkich języków. Utworzenie takiego narzędzia ułatwi zadawanie złożonych pytań dla każdego reprezentowanego w zasobach wielojęzycznych CLARIN-PL języka bez konieczności uczenia się i stosowania różnie zdefiniowanych zbiorów tagów poszczególnych języków.

11. PODSUMOWANIE

W artykule opisano już dostępne na stronie CLARIN-PL i aktualnie tworzone zasoby z językiem rosyjskim (tj. korpusy wielojęzyczne i pamięci TMX) oraz narzędzia, które umożliwiają analizę zbiorów rosyjskojęzycznych. Dostęp do korpusów wielojęzycznych CLARIN-PL wymaga uprzedniej rejestracji na stronie. Narzędzia *TagerML* i *MultiEmo* są dostępne bez wymaganej rejestracji. Warto zwrócić uwagę,

że zarejestrowani użytkownicy uzyskują dojście do znacznie większej liczby zasobów (https://clarin-pl.eu/index.php/zasoby/), narzędzi i usług (https://clarin-pl.eu/index.php/uslugi/). Mogą też skorzystać z oferty *Repozytorium dSpace* (https://clarin-pl.eu/dSpace/) i chmury *Clarin Cloud* (https://nextcloud.clarin-pl.eu/). Każdy użytkownik, deponując w repozytorium lub chmurze własne materiały, określa ich widoczność dla innych osób oraz przypisuje im właściwe prawa autorskie. Wszelkie zamieszczone w repozytorium cyfrowym *dSpace* zasoby (np. korpusy badawcze) mogą „za kliknięciem" użytkownika zostać automatycznie przetworzone i być dostępne w przeglądarce webowej *KonText* (dot. zasobów jednojęzycznych). W przypadku korpusów wielojęzycznych zachodzi konieczność kontaktu z Centrum Wiedzy PolLinguaTec, które służy natychmiastową pomocą użytkownikom. Na stronach CLARIN-PL są przystępnie opisane wszystkie zasoby, narzędzia i usługi. W *Mediatece* (https://clarin-pl.eu/index.php/mediateka/) zamieszczono praktyczne instrukcje obsługi narzędzi, materiały warsztatowe, publikacje i prezentacje. Również tam znajdzie użytkownik linki do wielu projektów z zakresu e-humanistyki w Polsce. Konsorcjum CLARIN-PL publikuje na kanale YouTube filmy instruktażowe oraz nagrania z konferencji i warsztatów (https://www.youtube.com/channel/UCqrhEITxu8_MIWPnFdYomPw/videos).

REFERENCES

Banasiak, Jakub. "Built-in argument positions in Bulgarian and Polish." *Cognitive Studies / Études cognitives*, 2021(21), Article 2558, https://doi.org/10.11649/cs.2558, 2021.

Blagoeva, Diana, and Jaskot, Maciej, and Sosnowski, Wojciech. „A lexicographical approach to the contrastive analysis of Bulgarian and Polish phraseology." *Cognitive Studies / Études cognitives*, 2019(19), Article 1923. https://doi.org/10.11649/cs.1923, 2019.

Chaudhary, Vishrav, and Tang, Yuqing, and Guzmán, Francisco, and Schwenk, Holger, and Koehn, Philipp. "Lowresource corpus filtering using multilingual sentence embeddings." *Proceedings of the Fourth Conference on Machine Translation (WMT)*. Bojar, Ondřej i in. Eds. Florence: Association for Computational Linguistics, 2019.

Dimitrova, Ludmila, and Koseska-Toszewa, Violetta, and Roszko, Danuta, and Roszko, Roman. "Bulgarian-Polish-Lithuanian Corpus: Current development." *International Workshop: Multilingual resources, technologies and evaluation for Central and Eastern European languages held in conjunction with the International Conference RANLP–2009: Proceedings*. Vertan, Cristina, and Piperidis, Stelios, and Paskaleva, Elena, and Slavcheva, Milena. Eds. Bulgaria. Borovets, 2009: 1–8.

Dimitrova, Ludmila, and Koseska-Toszewa, and Violetta, Roszko, and Danuta, and Roszko, Roman. "Trilingual Aligned Corpus: Current state and new applications." *Cognitive Studies / Études cognitives* 2014, no. 14: 13–20.

Duszkin, Maksim, and Roszko, Danuta, and Roszko, Roman. "New parallel corpora of Baltic and Slavic languages — Assumptions of corpus construction." *Lecture Notes in Artificial Intelligence LNAI 12848: Text, Speech, and Dialogue TSD 2021*. Ekštein, Kamil, and Pártl, František, and Konopík, Miloslav. Eds. Cham: Springer International Publishing, 2021: 173–183. DOI: https://doi.org/10.1007/978–3-030–83527–9_15.

Garncarek, Łukasz, and Powalski, Rafał, and Stanisławek, Tomasz, and Topolski, Bartosz, and Halama, Piotr, and Turski, Michał, and Graliński, Filip. "LAMBERT: Layout-aware language modeling for information extraction." *Document Analysis and Recognition — ICDAR 2021*. Lladós, Josep, and Lopresti, Daniel, and Uchida, Seiischi. Eds. Cham: Springer International Publishing, 2020: 1–16.

Kisiel, Anna, and Koseska-Toszewa, and Violetta, Kotsyba, and Natalia, Satoła-Staśkowiak, Joanna, and Sosnowski, Wojciech. *Polish-Bulgarian-Russian Parallel Corpus*. CLARIN-PL digital repository, 2016, http://hdl.handle.net/11321/308 (11.11.2021).

Kocoń, Jan, and Miłkowski, Piotr, and Kanclerz, Kamil. "MultiEmo: Multilingual, Multilevel, Multidomain Sentiment Analysis Corpus of Consumer Reviews." *Computational Science — ICCS 2021. ICCS 2021. Lecture Notes in Computer Science*, vol. 12743, Paszynski, Maciej, and Kranzlmüller, Dieter, and Krzhizhanovskaya, Valeria V., and Dongarra, Jack J., and Sloot, Peter M.A. Eds. Cham: Springer International Publishing, 2021.

Kocoń, Jan, and Kanclerz, Kamil, and Miłkowski, Piotr. *MultiEmo: Multilingual, Multilevel, Multidomain Sentiment Analysis Corpus of Consumer Reviews*. CLARIN-PL digital repository, 2021, http://hdl.handle.net/11321/798, (11.11.2021).

Kocoń, Jan, and Kanclerz, Kamil, and Miłkowski, Piotr, and Bojanowski, Bartosz, and Zaśko-Zielińska, Monika. *PolEmo 1.0 + MultiEmo-Test 1.0 Multilingual Sentiment Analysis Dataset for KES2020*. CLARIN-PL digital repository, 2020, http://hdl.handle.net/11321/737 (11.11.2021)

Koseska, Violetta, and Roszko, Roman. "On semantic annotation in CLARIN-PL parallel corpora." *Cognitive Studies / Études cognitives* 2015, no. 15: 211–236. https://doi.org/10.11649/cs.2015.016 (11.11.2021).

Machálek, Tomáš. "Advanced and flexible corpus query interface." *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association, 2020: 7003–7008.

Piasecki, Maciej, and Walentynowicz, Wiktor. "MorphoDiTa-based tagger adapted to the Polish language technology." *Proceedings of Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznań: LTC 2017, 2017: 377–381.

Roszko, Danuta, and Roszko, Roman. "Polsko-litewskie korpusy IS PAN i CLARIN-PL." *Prace Bałtystyczne* vol. 7. *Język. Kultura. Literatura*. Birgiel, Nijola, and Roszko, Danuta (eds.). Warszawa: Uniwersytet Warszawski, 2018: 185–205.

Roszko, Danuta, and Roszko, Roman. "Korpusy wielojęzyczne wkładem Instytutu Slawistyki Polskiej Akademii Nauk w rozwój infrastruktury CLARIN-PL: Przykłady analizy korpusowej nad wołaczem." *Języki słowiańskie dziś — w kręgu kategorii, struktur i procesów*. Banasiak, Jakub, and Kiklewicz, Aleksander, and Mazurkiewicz-Sułkowska, Julia. Eds. Warszawa–Łódź: Instytut Slawistyki PAN — Wydawnictwo Uniwersytetu Łódzkiego, 2021: 281–313.

Roszko, Roman. “O nowych ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusach równoległych oraz ich zastosowaniach.” *Acta Baltico-Slavica* 2021, no. 45, article 2576.

Roszko, Roman, and Sosnowski, Wojciech, and Duszkin, Maksim, and Roszko, Danuta, and Tymoshuk, Roman. *Polish-Russian Parallel Corpus*, *CLARIN-PL* digital repository, 2018, http://hdl.handle.net/11321/534 (11.11.2021).

Simov, Kiril, and Simov, Alexander, and Osenova, Petya. “An XML architecture for shallow and deep processing.” *The Proceedings of the ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP*. ESSLLI, 2004: 51–60.

Sosnowski, Wojciech, and Satoła-Staśkowiak, Joanna. “A contrastive analysis of feminitives in Bulgarian, Polish and Russian.” *Cognitive Studies / Études cognitives*, 2019 (19), Article 1922. https://doi.org/10.11649/cs.1922, 2019.

Straka, Milan and Straková, Jana. *UDPipe*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Prag 2016, http://hdl.handle.net/11234/1–1702 (11.11.2021).

Artykuł został częściowo wsparty przez CLARIN — Common Language Resources and Technology Infrastructure, nr projektu POIR.04.02.00-00C002/19.