ROCZNIKI HUMANISTYCZNE
Tom LXXIII, zeszyt 2 – 2025
ZESZYT SPECJALNY
DOI: https://doi.org/10.18290/rh25732.7s

BARBARA MITRENGA
KINGA WĄSIŃSKA

DIALOGICZNE TEKSTY DAWNE – OD DIGITALIZACJI ŹRÓDEŁ HISTORYCZNYCH DO OPERATYWNEGO KORPUSU

Abstrakt. Artykuł ma na celu ukazanie specyfiki dawnych tekstów dialogowych w kontekście ich digitalizacji i możliwości wykorzystania w korpusach. Przedstawia etapy tworzenia Korpusu Dawnych Polskich Tekstów Dramatycznych (1772–1939). W pracy zwrócono uwagę na różnice między starszymi i nowszymi wydaniami tych samych tekstów. Wskazano na niedostatki technik OCR podczas sczytywania tekstów dialogicznych, w tym m.in. brak możliwości automatycznego oddzielenia didaskaliów od tekstu głównego. Przedstawiono także szczegóły związane z przygotowaniem danych do prawidłowej pracy narzędzi cyfrowych. Zgromadzone teksty wymagały m.in. rozpisania nazw postaci biorących udział w scenie i usunięcia kropek kończących zdania w didaskaliach. W efekcie przedstawionych zabiegów wyodrębniono przeszukiwalną część materiału językowego oraz obudowę tekstu o funkcji strukturalnej. W wyniku dalszych prac teksty zanotowano konkretnymi wartościami trzech wyznaczników socjolingwistycznych (płeć, wiek i status społeczny postaci). Zastosowane rozwiązania mogą być cenne dla autorów innych korpusów i narzędzi cyfrowych wykorzystujących teksty tzw. zapisanej mówioności.

Słowa kluczowe: narzędzia cyfrowe; korpus językowy; polski dramat dawny; teksty dialogiczne; anotacja socjolingwistyczna; *open science*

DIALOGICAL OLD TEXTS: FROM DIGITALIZATION OF HISTORICAL SOURCES TO AN OPERATIONAL CORPUS

Abstract. The article aims to show the specificity of old dialogue texts in the context of their digitalization and the possibilities of their use in corpora. It presents the stages of creating the Corpus of

Dr hab. BARBARA MITRENGA, profesor UŚ – Uniwersytet Śląski, Instytut Językoznawstwa; adres do korespondencji: ul. Uniwersytecka 4, 40-007 Katowice; e-mail: barbara.mitrenga@us.edu.pl; ORCID: https://orcid.org/0000-0001-9389-8152.

Dr KINGA WĄSIŃSKA – Uniwersytet Śląski, Instytut Językoznawstwa: adres do korespondencji: ul. Uniwersytecka 4, 40-007 Katowice; e-mail: kinga.wasinska@us.edu.pl; ORCID: https://orcid.org/0000-0002-1731-9347.

Polish Dramatic Texts (1772–1939). The article highlights the differences between older and newer editions of the same texts. It also indicates the shortcomings of OCR techniques in reading dialogue texts, including the inability to automatically separate stage directions from the main text. Then, details related to preparing the data for the correct operation of digital tools are presented. The collected texts required, among other things, writing down the characters' names, taking part in the scene and removing dots ending the sentences of stage directions. As a result of the presented procedures, the searchable part of the linguistic material and the text casing with a structural function were isolated. The texts were also annotated considering specific values of three sociolinguistic determinants (gender, age and social status of the characters). The applied solutions may be valuable for authors of other corpora and digital tools using so-called speech-related text.

Keywords: digital tools; corpus linguistic; old Polish drama; dialogue texts; sociolinguistic annotation; open science

W lingwistyce korpusowej, a szerzej rzecz ujmując – we współczesnej humanistyce cyfrowej i ogólnie w nauce – standardem stało się wdrażanie idei *open science*, zakładającej dostępność oraz ponowne wykorzystanie zgromadzonych danych, umożliwiającej testowanie i weryfikowanie zastosowanych technik badań oraz zwiększającej transparentność nauki. W kontekście tej idei warto wspomnieć również o tzw. zasadzie FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*) wskazującej na cechy optymalnie zarządzanych danych badawczych, opublikowanym w 2001 r. przez Komisję ds. Etyki w Nauce PAN w dokumencie „Zbiór zasad i wytycznych pt. «Dobre obyczaje w nauce»"[1] oraz zaleceniach UNESCO „Recommendation on Open Science" z 2021 r.[2] Wszystkie wymienione dokumenty i wytyczne nakładają na naukowców i twórców narzędzi cyfrowych obowiązek udostępniania danych oraz zamieszczania w publikacjach danych służących swobodnemu przepływowi informacji naukowej.

Niniejszy artykuł wpisuje się w ideę *open science*, gdyż jego celem jest prezentacja kolejnych etapów pracy nad opracowaniem od podstaw pierwszego w Polsce anotowanego socjolingwistycznie otwartego zbioru danych, jakim jest Korpus Dawnych Polskich Tekstów Dramatycznych (1772-1939) (dalej cyt. KorTeDa), stworzonego w Instytucie Językoznawstwa Uniwersytetu Śląskiego w Katowicach pod kierunkiem Magdaleny Pastuch[3] i przy technicznym wsparciu

[1] „Zbiór zasad i wytycznych pt. «Dobre obyczaje w nauce»", Komitet Etyki w Nauce Polskiej Akademii Nauk (2001), https://ken.pan.pl/images/2001_ZbiorWytycznych.pdf [dostęp: 29.04.2025].

[2] „Recommendation on Open Science", UNESCO (2021), https://unesdoc.unesco.org/ark:/48223/pf0000379949 [dostęp: 29.04.2025].

[3] Autorkami koncepcji całego korpusu oraz przyjętych merytorycznych rozwiązań są członkinie zespołu „Pragmalingwistyki historycznej": Magdalena Pastuch, Barbara Mitrenga i Kinga Wąsińska.

konsorcjum CLARIN-PL[4]. Należy podkreślić, że KorTeDa to otwarty zbiór danych, ponieważ jest publicznie i bezpłatnie dostępny pod adresem: https://korteda.clarin-pl.eu, a korzystanie z korpusu nie wymaga instalowania dodatkowego oprogramowania czy logowania się. Warto również nadmienić, że KorTeDa jest korpusem specjalistycznym, który wymagał wdrożenia autorskich koncepcji oraz dostosowania istniejących narzędzi do założonego celu badawczego.

Artykuł ma zatem charakter narzędziowy i metodyczny. Przedstawiono w nim nie tylko działania podjęte w celu stworzenia korpusu, lecz także trudności i wyzwania wynikające z opracowania korpusu tekstów dialogicznych. Omówiono zastosowane narzędzia oraz przyjęte rozwiązania umożliwiające powstanie operatywnego korpusu. Zasadnicza treść artykułu obejmuje cztery części: w pierwszej scharakteryzowano zasady opracowania bazy materiałowej i cele projektu, w drugiej opisano etapy prac nad tekstami, w trzeciej zaprezentowano narzędzie Inforex i sposób przygotowania danych do anotacji socjolingwistycznej, natomiast w czwartej przedstawiono wypracowane w projekcie rozstrzygnięcia merytoryczne i techniczne. Zaprezentowane kwestie dotyczą realizacji korpusu dialogicznych tekstów dawnych, jednak zastosowane rozwiązania mogą być cenne dla autorów innych korpusów i narzędzi cyfrowych wykorzystujących teksty tzw. zapisanej mówioności[5].

CHARAKTERYSTYKA I CELE PROJEKTU

Charakterystyka zawartości tekstowej KorTeDa oraz przyjętych zasad anotacji została przedstawiona w odrębnych publikacjach[6]. Na potrzeby niniejszego

[4] Szczegółowe informacje o zespole CLARIN-PL zaangażowanym w tworzenie KorTeDa znajdują się na stronie korpusu w zakładce „O korpusie".

[5] Jonathan Culpeper, Merja Kytö, „Data in Historical Pragmatics: Spoken Interaction (Re)Cast as Writing", *Journal of Historical Pragmatics* 1 (2000): 175–99. https://doi.org/10.1075/jhp.1.2.03cul. Odwołując się do rozstrzygnięć terminologicznych na temat mówioności i oralności w językoznawstwie polskim, warto zaznaczyć, że KorTeDa gromadzi teksty przydatne do badań nad językiem mówionym używanym w konkretnych sytuacjach komunikacyjnych przez osoby piśmienne (por. badania na temat oralności pierwotnej w staropolskich apokryfach, czyli funkcjonowania języka w kulturze nieznającej pisma lub wykorzystującej go w ograniczonym zakresie). Zob. Aleksandra Deskur, „Oralność czy mówioność? O śladach mowy w tekstach staropolskich", *LingVaria* 16, nr 1 (31) (2021): 73–83, https://doi.org/10.12797/LV.16.2021.31.06; Małgorzata Kita, „Koncepcja oralności W.J. Onga a poglądy polskich badaczy języka mówionego", w *W kręgu języka polskiego. Śląsko-poznańskie kolokwia lingwistyczne*, red. Ewa Jędrzejko (Katowice: Wydawnictwo Uniwersytetu Śląskiego, 2001), 116–30.

[6] Magdalena Pastuch, Barbara Mitrenga, Kinga Wąsińska, „Anotacja socjolingwistyczna w Korpusie dawnych polskich tekstów dramatycznych (1772-1939)", *LingVaria* 19, 1(37) (2024):

opracowania warto jedynie wskazać, że KorTeDa zawiera 50 oryginalnie polskich dramatów o tematyce m.in. miłosnej, rodzinnej i obyczajowej z lat 1772-1939, o czym świadczą przykładowe tytuły dzieł, takie jak: *Zabawy, czyli życie bez celu* Feliksa Oraczewskiego; *Anarchia domowa* Adama Siemońskiego; *W małym domku* Tadeusza Rittnera, *Dla szczęścia* Stanisława Przybyszewskiego, *Kochanek bez serca, czyli guwernantka zalotna* Grzegorza Broniszewskiego[7]. W wersji cyfrowej baza tekstowa to ponad 800 tys. tokenów. Tak dobrany zestaw tekstów i sposób anotacji stanowią odpowiedź na potrzebę stworzenia narzędzia umożliwiającego badania naukowe nad dawną polszczyzną potoczną. Aby było to możliwe, zrodziła się potrzeba opracowania korpusu gromadzącego teksty, które w możliwie największym stopniu byłyby zbliżone do dawnego języka mówionego. Zdecydowano się na teksty zapisane, dialogowe, gdyż odmiana potoczna języka używana jest przede wszystkim w dialogach. Warto również podkreślić, że język w odmianie mówionej jest genetycznie i komunikacyjnie językiem pierwszym[8]. Zanim pojawiły się nagrania, teksty mówione utrwalała pamięć ludzka, a następnie pismo. Chociaż magnetofony pojawiły się w latach 50. i 60. XX wieku, to wykorzystywanie ich do nagrań współczesnego języka mówionego w celach naukowych wymagało upływu kolejnych lat[9]. Nagrania sprzed kilkudziesięciu lat również obecnie są cennym materiałem badawczym[10].

W odniesieniu do badań o charakterze historycznym w ogóle nie było możliwości utrwalenia ustnych wypowiedzi poza ich zapisem. Jak pisze Marek Osiewicz, badacz fonetyki historycznej „wnioskować o niej może tylko w sposób pośredni na podstawie jedynego źródła, którym są teksty pisane, a ściślej – ich pisownia”[11]. Chęć poznania cech dawnej komunikacji potocznej motywuje do poszukiwania śladów mówioności w tekstach pisanych, a także do gromadzenia

151–70; Barbara Mitrenga, Magdalena Pastuch, Kinga Wąsińska, „Możliwości i ograniczenia historycznych badań pragmalingwistycznych”, w *W kręgu dawnej polszczyzny*, t. 7, red. Maciej Mączyński, Ewa Horyń, Ewa Zmuda (Kraków: Wydawnictwo Naukowe Akademii Ignatianum, 2021), 175–76.

[7] Wykaz wszystkich utworów tworzących zasób KorTeDa wraz z ich danymi bibliograficznymi znajduje się na stronie https://korteda.clarin-pl.eu/, w zakładce „Teksty w korpusie”.

[8] Janina Labocha, „Pragmatyczne mechanizmy składni języka mówionego”, *Slavia Occidentalis* 69 (2012): 139.

[9] Monika Zaśko-Zielińska, Anna Majewska-Tworek, Marta Śleziak, Artur Tworek, *Od rozmowy do korpusu, czyli jak zbierać i archiwizować dane mówione* (Wrocław: Uniwersytet Wrocławski–Quaestio, 2020), 7.

[10] Anna Majewska-Tworek, Monika Zaśko-Zielińska, Piotr Pęzik, „«Polszczyzna mówiona miast» – kontynuacja badań z lat 80. XX wieku z wykorzystanie narzędzi lingwistyki cyfrowej”, *Forum Lingwistyczne* 7 (2020): 71-87.

[11] Marek Osiewicz, „Kierunki przemian polszczyzny w zakresie fonetyki (propozycja rozdziału podręcznika do nauczania treści historycznojęzykowych na studiach I stopnia)”, *Kwartalnik Językoznawczy* 2 (2010): 60.

różnych typów tekstów, które wykazują odrębne formy polskiego języka mówionego. Przykładowo w korpusie referencyjnym współczesnego języka francuskiego[12] oprócz danych mówionych i pisanych umieszczono teksty pseudo-mówione (ang. *pseudo-spoken*), które obejmują m.in. teksty dramatów, scenariusze filmowe, napisy do filmów, seriali, wiadomości tekstowe. Dawne teksty dramatyczne – choć nie są zapisem autentycznej mowy – dają wyobrażenie o mowie Polaków w przeszłości. Jak przekonują badacze, dopiero zróżnicowany korpus, który będzie pokazywał użycie polszczyzny mówionej w różnych sytuacjach komunikacyjnych, daje możliwość pełnego oglądu danych mówionych[13]. Korpus stanowi zasób, który w kolejnych latach złoży się na jeden duży korpus główny, pokazujący użycie dawnej polszczyzny mówionej w różnych sytuacjach komunikacyjnych. Obecnie ukończone zostały prace nad pierwszym subkorpusem, opartym na wyselekcjonowanych tekstach dramatycznych. W przyszłości planowana jest rozbudowa KorTeDa o inne i – co warto podkreślić – zróżnicowane gatunkowo teksty oddające cechy dawnego języka potocznego (rozmówki z podręczników do nauki języka polskiego, listy prywatne czy zapiski sądowe).

ETAPY OPRACOWANIA BAZY TEKSTOWEJ

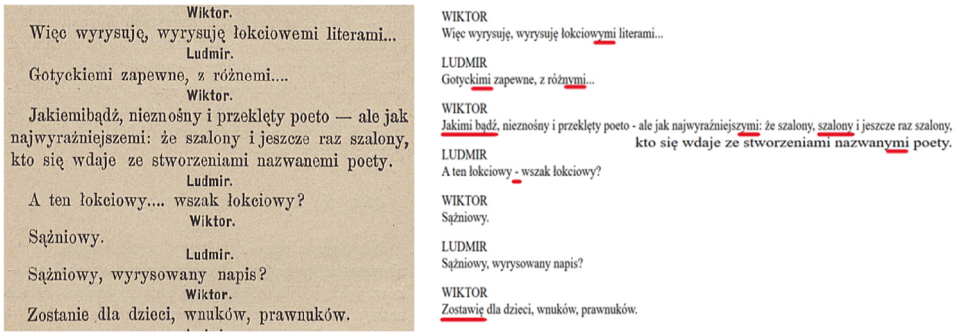
Pierwszym etapem prac było zgromadzenie tekstów spełniających określone przez nas kryteria[14]. Dramaty są dziełami 34 autorów. Zostały wydane w ośmiu ośrodkach, m.in. w Warszawie, Krakowie i Wilnie. Wykorzystano najstarsze dostępne cyfrowo w roku 2020 wydania dramatów, głównie z Biblioteki Narodowej POLONA (https://polona.pl), w mniejszym zakresie także ze Śląskiej Biblioteki Cyfrowej (https://sbc.org.pl/)[15]. Nie korzystano z wersji dostępnych w Wirtualnej Bibliotece Literatury Polskiej (https://literat.ug.edu.pl) czy w Wolnych Lekturach (https://wolnelektury.pl/), ponieważ pozyskany materiał miał w jak największym stopniu oddawać język badanego przedziału czasowego. Tym samym nie korzystano z nowszych wydań, uwzględniających zmiany edytorskie i poprawki naniesione w nowszych wersjach tego samego utworu. Przykładowe porównanie wersji uwzględnionej w korpusie z wersją dostępną w Wirtualnej Bibliotece Literatury Polskiej oraz Wolnych Lekturach przedstawiają rysunki 1 i 2.

[12] Dirk Siepman, „Dictionaries and Spoken Language: A Corpus-Based Review of French Dictionaries", *International Journal of Lexicography* 28, nr (2) (2015): 143.

[13] Zaśko-Zielińska, Majewska-Tworek, Śleziak, Tworek, *Od rozmowy do korpusu*, 16.

[14] Mitrenga, Pastuch, Wąsińska, „Możliwości i ograniczenia historycznych badań pragmalingwistycznych", 175-76.

[15] 48 utworów pobranych zostało z POLONY, natomiast dwa ze Śląskiej Biblioteki Cyfrowej.

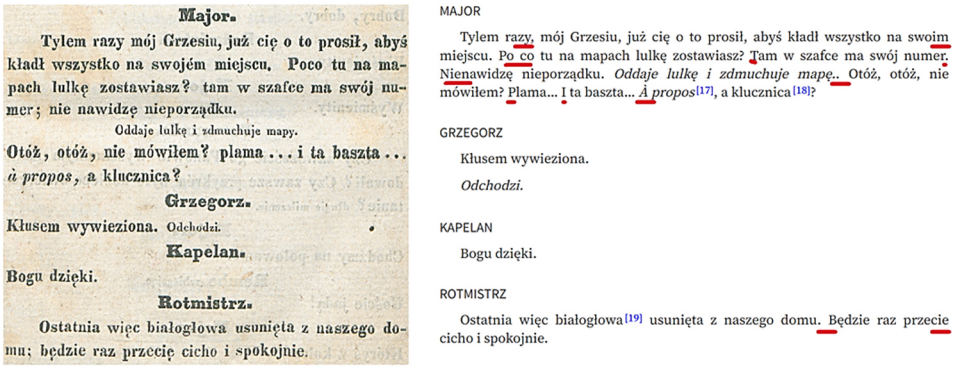


Wiktor.
Więc wyrysuję, wyrysuję łokciowemi literami...
Ludmir.
Gotyckiemi zapewne, z różnemi....
Wiktor.
Jakiemibądź, nieznośny i przeklęty poeto — ale jak najwyraźniejszemi: że szalony i jeszcze raz szalony, kto się wdaje ze stworzeniami nazwanemi poety.
Ludmir.
A ten łokciowy.... wszak łokciowy?
Wiktor.
Sążniowy.
Ludmir.
Sążniowy, wyrysowany napis?
Wiktor.
Zostanie dla dzieci, wnuków, prawnuków.

WIKTOR
Więc wyrysuję, wyrysuję łokciowymi literami...
LUDMIR
Gotyckimi zapewne, z różnymi...
WIKTOR
Jakimi bądź, nieznośny i przeklęty poeto - ale jak najwyraźniejszymi: że szalony, szalony i jeszcze raz szalony, kto się wdaje ze stworzeniami nazwanymi poety.
LUDMIR
A ten łokciowy - wszak łokciowy?
WIKTOR
Sążniowy.
LUDMIR
Sążniowy, wyrysowany napis?
WIKTOR
Zostawię dla dzieci, wnuków, prawnuków.

Rys. 1. Porównanie warstwy językowej dawnej i współczesnej edycji *Pana Jowialskiego* Aleksandra Fredry (fragment aktu I, sceny I)

Źródło: POLONA, wydanie z 1880 r. [https://polona.pl/item-view/aa8485c0-90f3-45e8-acf0-567f8541fa9c?page=83] oraz Wirtualna Biblioteka Literatury Polskiej, na podstawie wydania z 1956 r. [https://literat.ug.edu.pl/jwlski/index.htm]



Major.
Tylem razy mój Grzesiu, już cię o to prosił, abyś kładł wszystko na swojém miejscu. Poco tu na mapach lulkę zostawiasz? tam w szafce ma swój numer; nie nawidzę nieporządku.
Oddaje lulkę i zdmuchuje mapy.
Otóż, otóż, nie mówiłem? plama ... i ta baszta ... *à propos*, a klucznica?
Grzegorz.
Kłusem wywieziona. Odchodzi.
Kapelan.
Bogu dzięki.
Rotmistrz.
Ostatnia więc białogłowa usunięta z naszego domu; będzie raz przecię cicho i spokojnie.

MAJOR
Tylem razy, mój Grzesiu, już cię o to prosił, abyś kładł wszystko na swoim miejscu. Po co tu na mapach lulkę zostawiasz? Tam w szafce ma swój numer. Nienawidzę nieporządku. *Oddaje lulkę i zdmuchuje mapę.* Otóż, otóż, nie mówiłem? Plama... I ta baszta... *A propos*[17], a klucznica[18]?
GRZEGORZ
Kłusem wywieziona.
Odchodzi.
KAPELAN
Bogu dzięki.
ROTMISTRZ
Ostatnia więc białogłowa[19] usunięta z naszego domu. Będzie raz przecie cicho i spokojnie.

Rys. 2. Porównanie warstwy językowej dawnej i współczesnej edycji *Dam i huzarów* Aleksandra Fredry (fragment aktu I, sceny II)

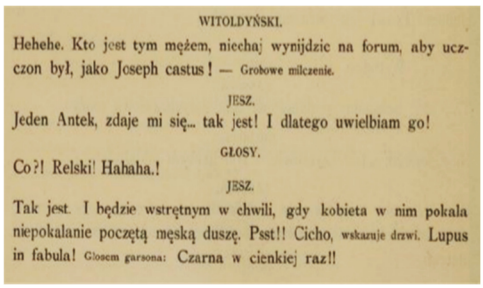
Źródło: POLONA, wydanie z 1839 r. [https://polona.pl/item-view/fed3a6f1-affd-4d4c-b030-ec1127accc76?page=4] oraz Wolne Lektury [https://wolnelektury.pl/katalog/lektura/fredro-damy-i-huzary.html]

Oba zestawienia pokazują nie tylko zmiany fleksyjne (zastąpienie końcówek *-em* i *-emi* odpowiednio *-im*/*-ym* oraz *-imi*/*-ymi*)[16], lecz także różnice dotyczące łącznej czy rozłącznej pisowni, podziału na zdania, zasad interpunkcji, zmiany

[16] Zagadnienie użycia końcówek *-em* lub *-im*/*-ym* oraz *-emi* lub *-imi*/*-ymi* w mowie i w piśmie w polszczyźnie XVIII-wiecznej stanowi osobne zagadnienie badawcze. Rozstrzygnięcie, która forma była bliższa mowie Polaków, wymagałoby odrębnego szczegółowego komentarza, uwzględniającego zależności między ówczesnym uzusem językowym a zaleceniami normatywnymi wynikającymi z tzw. reguły Kopczyńskiego.

w paradygmacie czasowników (*zostanie* versus *zostawię*) czy dodania elementów nieobecnych w oryginale (np. *szalony i jeszcze raz szalony* versus *szalony, szalony i jeszcze raz szalony*).

Drugi etap opracowania bazy materiałowej do korpusu polegał na poddaniu zebranych tekstów procedurze OCR (*optical charakter recognition*), mającej na celu rozpoznanie znaków i ich konwersję na plik tekstowy. Prace te wykonali pracownicy Centrum Informacji Naukowej i Biblioteki Akademickiej w Katowicach. Pozyskana warstwa tekstowa wymagała szczegółowej korekty. Problemy wynikały m.in. z niewystarczającego dostosowania oprogramowania OCR do tekstów dawnych (część znaków została rozpoznana błędnie lub nierozpoznana). W 2020 r. nie dysponowano narzędziem umożliwiającym zautomatyzowanie korekty, dlatego wykonano ją ręcznie, przez porównanie sczytanego tekstu z oryginałem. Nie uwspółcześniono pisowni, nie poprawiono literówek występujących w oryginalnym tekście. Pozostawiono dawne zapisy, np. odmienne od dzisiejszej normy poprawnościowej użycia liter *i*, *j* oraz *y* (np. *iako*, *iuż*, *y*, *poydę*, *przynaymniey*, *przyiaciołka*), niekonsekwentne zapisy (np. *być* i *bydź*), zapis łączny i rozłączny wyrazów (np. *wierzyćto*, *nawetbym*, *niema*). Ingerencję w tekst ograniczono do minimum, przede wszystkim zrezygnowano z oznaczania samogłosek pochylonych[17] oraz z zapisu kursywą obcych elementów językowych (np. *à propos*), zastąpiono literę ſ literą s czy uzgodniono zapis wielokropka (jako trzy kropki). Na potrzeby operatywnego korpusu konieczne okazało się również ujednolicenie sposobu zapisu didaskaliów – ustalono zapis w nawiasach okrągłych.



WITOLDYŃSKI.
Hehehe. Kto jest tym mężem, niechaj wynijdzie na forum, aby uczczon był, jako Joseph castus! — Grobowe milczenie.
JESZ.
Jeden Antek, zdaje mi się... tak jest! I dlatego uwielbiam go!
GŁOSY.
Co?! Relski! Hahaha.!
JESZ.
Tak jest. I będzie wstrętnym w chwili, gdy kobieta w nim pokala niepokalanie poczętą męską duszę. Psst!! Cicho, wskazuje drzwi. Lupus in fabula! Głosem garsona: Czarna w cienkiej raz!!

WITOLDYŃSKI. Hehehe. Kto jest tym mężem, niechaj wynijdzie na forum, aby uczczon był, jako Joseph castus! – Grobowe milczenie.
JESZ. Jeden Antek, zdaje mi się... tak jest! I dlatego uwielbiam go!
GŁOSY. Co?! Relski! Hahaha!
JESZ. Tak jest. I będzie wstrętnym w chwili, gdy kobieta w nim pokala niepokalanie poczętą męską duszę. Psst!! Cicho, wskazuje drzwi. Lupus in fabula! Głosem garsona: Czarna w cienkiej raz!!

Rys. 3. Fragment *Karykatur* Jana Augusta Kisielewskiego z nierozpoznanymi przez OCR didaskaliami

Źródło: POLONA, wydanie z 1903 r. [https://polona.pl/item-view/0525ef60-0891-45c2-86b1-5b8ef0995fef?page=0]

[17] Temat rozróżniania w piśmie samogłosek jasnych i pochylonych wiąże się ściśle z praktykami drukarskimi. W wybranych dramatach pochylenia były zaznaczane niekonsekwentnie i nie dotyczyły wszystkich tekstów, dlatego zrezygnowano z oznaczania samogłosek pochylonych.

Niejednokrotnie didaskalia trzeba było „wydobyć" z tekstu głównego, gdyż po procedurze OCR didaskalia i tekst główny zlały się ze sobą. Dotyczyło to sytuacji, w których w oryginale didaskalia były zapisywane mniejszą czcionką lub kursywą, ale bez nawiasów. Na etapie ręcznej korekty plików TXT błędy związane z niewłaściwym rozpoznaniem didaskaliów zostały poprawione.

INFOREX – OPIS MOŻLIWOŚCI I OGRANICZEŃ WYKORZYSTANEGO NARZĘDZIA

Przygotowane po korekcie pliki TXT zostały zaimplementowane do narzędzia Inforex[18] (https://clarin-pl.eu/tools/inforex), które umożliwiło naniesienie kilku warstw anotacji. Istotna była możliwość oznaczenia określonych części tekstu etykietami unikatowymi dla zgromadzonych tekstów – takie rozwiązanie oferował Inforex.

Inforex jest narzędziem bezpłatnym. Dostęp do systemu jest możliwy z dowolnej przeglądarki internetowej obsługującej Java i nie wymaga pobierania oraz instalacji dodatkowego oprogramowania[19]. System przechowuje wszystkie dane (tekst i anotacje) w zwykłym formacie tekstowym lub HTML w centralnym repozytorium zespolonym z narzędziem. Z systemem zintegrowane są dwa tokenizatory: TaKIPI-WS[20] oraz Maca[21]. W przypadku błędów możliwa jest modyfikacja granic tokenów i zdań. Tokeny są przechowywane jako pary wartości reprezentujące indeksy pierwszego i ostatniego znaku tokenów, tak samo rejestrowane są zestawy cech reprezentujące informacje morfosyntaktyczne tokenów. Anotacje utworzone przez użytkownika są przechowywane w taki sam sposób, jak tokeny. Każda modyfikacja zawartości jest śledzona przez system, a różnica

[18] Michał Marcińczuk, Jan Kocoń, Bartosz Broda, „Inforex – A Web-Based Tool for Text Corpus Management and Semantic Annotation", w *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, red. Nicoletta Calzolari, Khalid Choukri, Theirry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis (Istambul, Turkey: European Language Resources Association (ELRA), 2012), 225–30.

[19] Marcińczuk, Kocoń, Broda, „Inforex", 224 n.

[20] Bartosz Broda, Michał Marcińczuk, Maciej Piasecki, „Building a Node of the Accessible Language Technology Infrastructure", w *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, red. Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, Daniel Tapias (Valletta, Malta: European Language Resources Association (ELRA), 2010.

[21] Adam Radziszewski, Tomasz Śniatowski, „Maca – A Configurable Tool to Integrate Polish Morphological Data", w *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation, Barcelona, Spain, 20-21 January 2011*, red. Felipe Sánchez-Martinez, Juan Antonio Pérez-Ortiz (Barcelona: Universitat Oberta de Catalunya, 2011), 29-36.

w stosunku do poprzedniej wersji jest generowana i przechowywana w bazie danych. Dodatkowo aplikacja ma przyjazny dla użytkownika interfejs w formie dynamicznych stron HTML (technologia AJAX).

Rozwiązania wdrożone w Inforexie pozwoliły wykonać anotację i usprawnić proces korekty. Umożliwiły również jednoczesną pracę kilku osób nad korpusem i zarządzanie ich zadaniami. Narzędzie Inforex było najbardziej odpowiednie do zakładanych planów badawczych, ale okazało się, że wymaga ono dostosowania ze względu na specyficzny charakter dawnego materiału dialogowego. W toku prac konieczne było dopasowanie narzędzia do widoku wypowiedzi dialogowych.

W Inforexie została naniesiona automatyczna anotacja morfosyntaktyczna – zastosowano analizator morfologiczny MorphoDiTa. Ponadto wprowadzono dodatkową anotację socjolingwistyczną, która objęła trzy kategorie (wiek, płeć i status społeczny postaci)[22]. Na podstawie przygotowanych formatek system przypisał każdej wypowiedzi konkretną wartość w każdej z trzech kategorii socjologicznych charakteryzujących nadawcę wypowiedzi. W ten sposób zanotowane zostały wszystkie wypowiedzi w korpusie – dialogi i monologi, z wyłączeniem wypowiedzi zbiorowych[23].

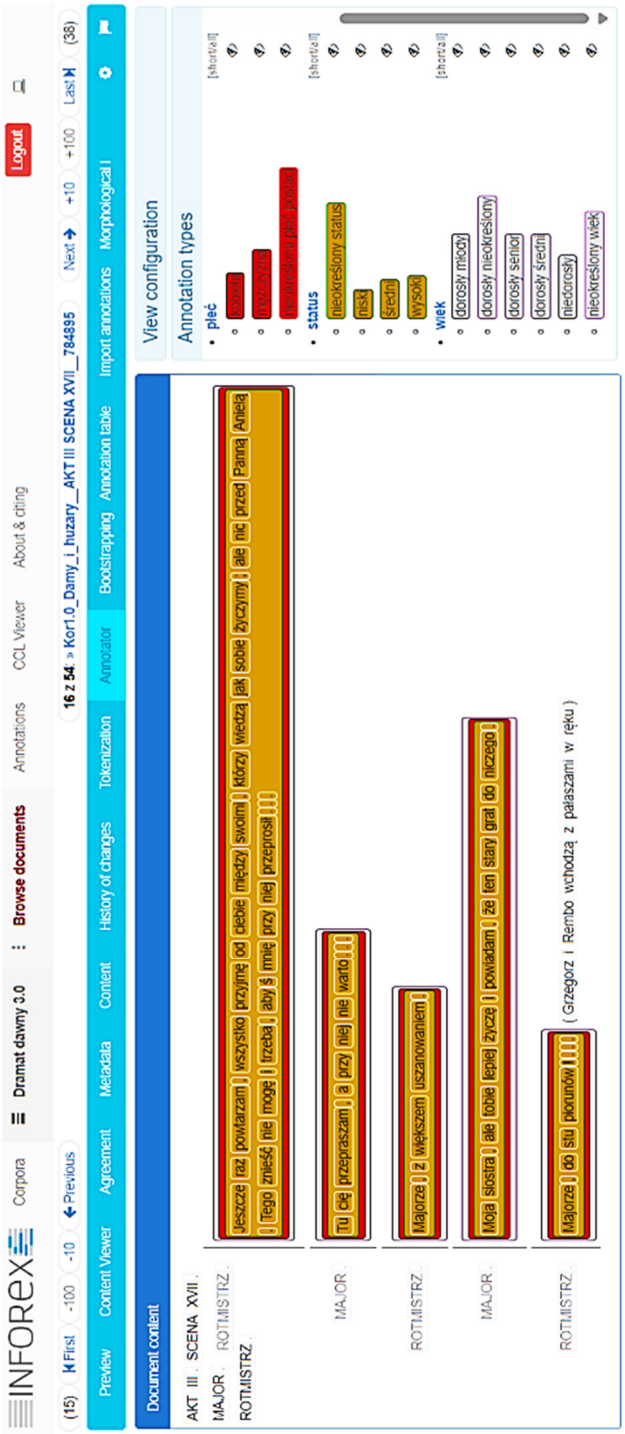
	A	B	C	D	E
1	**Postać z opisem**	**Nazwa postaci w tekście głównym**	**Płeć**	**Status społeczny**	**Wiek**
2	HRABIA JAN. przyjaciel Doktora Karola; krewny Kassyldy; oficer j	HRABIA	mężczyzna	wysoki	dorosły średni
3	WŁADYSŁAW. przyjaciel Doktora Karola; oficer jazdy polskiej.	WŁADYSŁAW	mężczyzna	średni	dorosły średni
4	SŁUŻĄCY.	SŁUŻĄCY	mężczyzna	niski	dorosły nieokreślony
5	DOKTOR KAROL.	DOKTOR	mężczyzna	średni	dorosły średni
6	DOKTOR KAROL.	JEZDZIEC	mężczyzna	średni	dorosły średni
7	KASSYLDA. (majątek/wysoki status materialny)	KASSYLDA	kobieta	wysoki	dorosły młody
8	EUSTACHY. brat Kassyldy. (majątek/wysoki status materialny)	GŁOS MĘZKI	mężczyzna	wysoki	dorosły młody
9	EUSTACHY. brat Kassyldy. (majątek/wysoki status materialny)	EUSTACHY	mężczyzna	wysoki	dorosły młody
10	ROMUALD. aptekarz.	ROMUALD	mężczyzna	średni	dorosły średni
11	HELENA. córka aptekarza; Jakob do niej mówi Panniczka	HELENA	kobieta	średni	dorosły młody
12	JAKOB. służący Doktora Karola; od siedmiu lat służy u Karola	JAKOB	mężczyzna	niski	dorosły średni
13	DWAJ LEKARZE. Taka postać jest w spisie, w tekście nie ma				
14	PIERWSZY LEKARZ	PIERWSZY LEKARZ	mężczyzna	średni	dorosły średni
15	DRUGI LEKARZ	DRUGI LEKARZ	mężczyzna	średni	dorosły średni
16	ELEW MEDYCYNY.	ELEW	mężczyzna	średni	dorosły młody
17	DOZORCA SZPITALU.	DOZORCA	mężczyzna	niski	dorosły nieokreślony
18	UBOGA KOBIETA.	KOBIETA	kobieta	niski	dorosły średni

Rys. 4. Fragment formatki sporządzonej dla dramatu *Sztuka i miłość* Maurycego Manna

Źródło: opracowanie własne.

[22] Pastuch, Mitrenga, Wąsińska, „Anotacja socjolingwistyczna", 151–70.

[23] Wypowiedzi zbiorowe często nie mają określonego autora, np. Głosy, Ludzie, Wieśniacy, Danserzy, lub są wypowiadane przez kilka osób o różnej płci, wieku i statusie społecznym, co przy anotacji mnożyłoby liczbę znaczników i powodowało nieczytelność anotacji.



Rys. 5. Prawidłowa anotacja socjolingwistyczna w narzędziu Inforex

Źródło: Inforex [https://inforex-work.clarin-pl.eu/index.php?page=report&corpus=158&subpage=annotator&id=875627]

Niejednokrotnie formatki wymagały wnikliwej lektury dramatu, aby jak największej liczbie postaci przypisać przyjęte wartości w obrębie anotowanych kategorii. Na rys. 4 widoczny jest fragment rozbudowanej formatki dramatu *Sztuka i miłość* Maurycego Manna, liczącej 29 postaci (w tym różnie nazywanych w dramacie, pojawiających się w scenach bezpośrednio lub jako głos, z postaciami zbiorowymi, które czasem wypowiadają się indywidualnie). Przypisanie odpowiednich wartości wymagało szczegółowej analizy informacji zawartych w spisie postaci, w didaskaliach rozproszonych w całym tekście, a także treści dramatu.

Zaprezentowany na rys. 5 widok jest docelowym efektem, który udało się osiągnąć w wyniku kolejnych prac. Automatyczne naniesienie wartości socjolingwistycznych wymagało jednak weryfikacji, gdyż nie przyniosło w pełni satysfakcjonujących efektów. Pojawiały się bowiem wcześniej nieprzewidziane problemy, głównie związane z zaburzeniem diaryzacji (rozbicie tekstu dramatu na kategorie <author> i <content>; przekroczenia granic obszarów anotowanych w ramach kategorii <author>, <content> i <message>; braku wydzielenia didaskaliów). Zdiagnozowano błędy:

– szerokiej anotacji, gdy anotacja przekroczyła granicę <author> i <content>, czego wynikiem było naniesienie wartości socjolingwistycznych na nazwę postaci w polu <author>;

– głębokiej anotacji – błąd anotacji polegający na przekroczeniu obszaru anotowanego granicy <author>, <conent> lub <message>. W wyniku tego błędu niektóre wypowiedzi bądź ich fragmenty nie miały naniesionej anotacji socjolingwistycznej;

– anotacji didaskaliów – didaskalia zostały oznaczone jako zwykły tekst z przypisanymi wartościami socjolingwistycznymi.

Na tym etapie prac konieczne okazało się wykonanie dodatkowej korekty technicznej plików.

ROZSTRZYGNIĘCIA MERYTORYCZNE I TECHNICZNE

W toku prac nad stworzeniem dialogicznego korpusu tekstów dawnych, oprócz wykonanych korekt tekstowych i technicznych, konieczne okazało się również wyeliminowanie zjawisk strukturalnych, które powodowały, że kolejne etapy znakowania nie przynosiły założonych efektów. Korekta plików HTML w narzędziu Brackets zakładała wprowadzenie nowych oznaczeń na wyodrębnione dane. Zdecydowano się na ingerencję w elementy struktury dramatów w stosunku do oryginału, by ulepszyć prace narzędzi korpusowych. Co istotne dla badań, wprowadzone zmiany nie wpłynęły na treść samych wypowiedzi w zakresie leksyki, składni i fleksji. Przykładowo na początku sceny,

w której biorą udział te same osoby, które były w scenie poprzedniej, zastąpiono zapis *ciż sami*, *poprzedniejsi*, *wcześniejsi*, przez podanie nazwy osób, np. *Zosia. Leon. Hrabia*. Do spisu postaci dopisano osoby, które wypowiadają się w danej scenie, choć nie zostały wymienione w oryginale. W celu przypisania wartości socjolingwistycznych każdej wypowiedzi doprecyzowano także nazwę osoby wypowiadającej tekst, np. *jego dziecko* zamieniono na *Olek*. Usunięto kropki kończące zdanie zapisane w didaskaliach, ponieważ burzyło to automatyczną identyfikację didaskaliów. Jeśli w oryginale zdanie kończy się przecinkiem (taki zapis może być wynikiem błędu drukarskiego), to zamieniony został on na kropkę. Z kolei przy braku kropki na końcu wypowiedzi, wstawiano ją. Występujące w tekście głównym oryginałów nawiasy okrągłe – zarezerwowane w korpusie jako znacznik didaskaliów – zamieniono na nawiasy kwadratowe. Ponadto do korpusu nie włączono przypisów dolnych, przedmów, wstępów, podziękowań, dedykacji, wykazów aktorów teatralnych, z tego względu, że nie są to partie mówione, ale komentarze odautorskie lub wydawnicze.

Wprowadzenie wymienionych zmian pozwoliło wyodrębnić przeszukiwalną część materiału językowego oraz obudowę tekstu o funkcji strukturalnej. Jak już wspomniano, każda pojedyncza wypowiedź została opatrzona informacją o wieku, płci i statusie społecznym postaci, która ją wypowiada. Dzięki tej pracy, oprócz informacji o socjolingwistycznym profilu nadawcy, stworzono bazę tekstów przeszukiwalną po konkretnych parametrach. Korzystając z Inforexa, można było zatem analizować wszystkie wypowiedzi kobiet zgromadzone w korpusie albo wszystkie wypowiedzi osób młodych czy też osób o wysokim, średnim lub niskim statusie społecznym. Jednak znaczącym ograniczeniem systemu Inforex okazał się brak możliwości krzyżowania kryteriów wyszukiwania. Stworzenie korpusu o precyzyjnych możliwościach doboru parametrów przeszukiwanych tekstów wymagało zbudowania nowego narzędzia dedykowanego takim badaniom. Należało rozwiązać również problemy związane z nieciągłą anotacją oraz zaprojektować osobne przeszukiwanie wypowiedzi zbiorowych oraz didaskaliów. Nowa wyszukiwarka wymagała zbudowania sekcji z frekwencją, która zastąpiłaby niemiarodajne dla tekstów dialogicznych dane liczbowe i statystyki dostępne w Inforexie[24].

Równocześnie z prowadzonymi zadaniami korpusowymi trwały prace nad przekształceniem plików tekstowych do formatu TEI. Pozwoliło to włączyć teksty 50 polskich dramatów do międzynarodowego projektu Programmable Corpora[25].

[24] Inforex do liczby wyrazów wliczał nazwy postaci dramatu, opisy zamieszczone w didaskaliach oraz wyrazy należące do obudowy tekstowej, np. „akt" i „scena". Ponadto nie pozwalał na generowanie statystyk dotyczących liczby poszczególnych wyrazów.

[25] Frank Fischer, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christpher Kittel, Carsten Milling, Peer Trilcke, „Programmable Corpora: Introducing DraCor, an Infrastructure for the Research

Kolekcji tekstów sztuk teatralnych na platformie DraCor: Polish Drama Corpus (PolDraCor, https://dracor.org/pol). Utworzony korpus, wzbogacony narzędziami do analiz stylometrycznych[26], pozwala na prowadzenie kolejnych badań, tym razem o nachyleniu stylistycznym. Co ważne, platforma projektu umożliwia zestawianie ze sobą korpusów, a tym samym prowadzenie badań o charakterze komparatywnym. Przykładowo można porównywać charakterystykę sieci wypowiedzi[27] poszczególnych postaci różnych dramatów z wielu języków, które powstały w podobnym lub odległym od siebie czasie. Realizacja PolDraCor, podobnie jak opracowanego korpusu, stanowi przykład ponownie wykorzystanych danych tekstowych, a tym samym potwierdza znaczenie otwartej nauki w rozwoju narzędzi cyfrowych istotnych do dalszych badań – nie tylko językoznawczych.

*

Stworzony korpus dawnych tekstów dialogicznych jest nowatorski i w pełni operatywny, o czym świadczą jego cechy:

– jest przeszukiwalny na kilku poziomach,
– użytkownik ma możliwość krzyżowania parametrów wyszukiwania,
– anotacja morfosyntaktyczna i socjolingwistyczna zostały dostosowane do potrzeb projektu,
– oferuje zaawansowane filtrowanie danych,
– umożliwia łatwe kopiowanie wyników zapytań,
– ma szybko działający oraz intuicyjny interfejs,
– zawiera dokładną i rozbudowaną informację frekwencyjną.

BIBLIOGRAFIA

Broda, Bartosz, Michał Marcińczuk, Maciej Piasecki. „Building a Node of the Accessible Language Technology Infrastructure". W *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, red. Nicoletta Calzolari, Khalid Choukri, Bente Maegaard,

on European Drama", w *Proceedings of DH2019: "Complexities", Utrecht University, 2019*, https://doi:10.5281/zenodo.

[26] Więcej na temat narzędzia Stylo i samych badań stylometrycznych można przeczytać w: Rafał L. Górski, Magdalena Król, Maciej Eder, *Zmiana w języku. Studia kwantytatywno-korpusowe* (Kraków: Instytut Języka Polskiego PAN, 2019).

[27] System generuje graf współwystąpień wypowiedzi postaci, które pojawiają się w tej samej scenie lub akcie i rozmawiają ze sobą. Wizualizacja pokazuje liczbę segmentów w wypowiedzi, gęstość wymian dialogowych, średnią długość wypowiedzi postaci i współczynnik występowania podobnych wzorców w danej sztuce.

Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, Daniel Tapias. Valletta, Malta: European Language Resources Association (ELRA).

Culpeper, Jonathan, Merja Kytö. „Data in Historical Pragmatics: Spoken Interaction (Re)Cast as Writing". *Journal of Historical Pragmatics* 1 (2000): 175–99. https://doi.org/10.1075/jhp.1.2.03cul.

Deskur, Aleksandra. „Oralność czy mówioność? O śladach mowy w tekstach staropolskich". *LingVaria* 16, nr 1(31), (2021): 73–83. https://doi.org/10.12797/ LV.16.2021.31.06.

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christpher Kittel, Carsten Milling, Peer Trilcke. „Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama". W *Proceedings of DH2019: "Complexities", Utrecht University, 2019*. https://doi:10.5281/zenodo.4284002.

Kita, Małgorzata, „Koncepcja oralności W.J. Onga a poglądy polskich badaczy języka mówionego". W *W kręgu języka polskiego. Śląsko-poznańskie kolokwia lingwistyczne*, red. Ewa Jędrzejko, 116–30. Katowice: Wydawnictwo Uniwersytetu Śląskiego, 2001.

Labocha, Janina. „Pragmatyczne mechanizmy składni języka mówionego". *Slavia Occidentalis* 69 (2012): 139–45.

Majewska-Tworek, Anna, Monika Zaśko-Zielińska, Piotr Pęzik. „«Polszczyzna mówiona miast» – kontynuacja badań z lat 80. XX wieku z wykorzystaniem narzędzi lingwistyki cyfrowej". *Forum Lingwistyczne* 7 (2020): 71-87.

Marcińczuk, Michał, Jan Kocoń, Bartosz Broda. „Inforex – A Web-Based Tool for Text Corpus Management and Semantic Annotation". W *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, red. Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis, 225--30. Istanbul, Turkey: European Language Resources Association (ELRA), 2012.

Mitrenga, Barbara, Magdalena Pastuch, Kinga Wąsińska. „Możliwości i ograniczenia historycznych badań pragmalingwistycznych". W *W kręgu dawnej polszczyzny*, t. 7, red. Maciej Mączyński, Ewa Horyń, Ewa Zmuda, 163–82. Kraków: Wydawnictwo Naukowe Akademii Ignatianum, 2021.

Osiewicz, Marek. „Kierunki przemian polszczyzny w zakresie fonetyki (propozycja rozdziału podręcznika do nauczania treści historycznojęzykowych na studiach I stopnia)". *Kwartalnik Językoznawczy* 2 (2010): 58-75.

Pastuch, Magdalena, Barbara Mitrenga, Kinga Wąsińska. „Anotacja socjolingwistyczna w Korpusie dawnych polskich tekstów dramatycznych (1772–1939)". *LingVaria* 19, nr 1(37) (2024): 151–70. https://doi.org/10.12797/LV.19.2024.37.10.

„Recommendation on Open Science". UNESCO, 2021. https://unesdoc.unesco.org/ark:/48223/pf 0000379949 [dostęp: 29.04.2025].

Radziszewski, Adam, Tomasz Śniatowski. „Maca – A Configurable Tool to Integrate Polish Morphological Data". W *Proceedings of the Second International Workshop on Free/Open-Sour ceRule-Based Machine Translation*. Barcelona, Spain, 20-21 January 2011, red. Felipe Sánchez-Martínez, Juan Antonio Pérez-Ortiz, 29–36. Barcelona: Universitat Oberta de Catalunya, 2011.

Siepman, Dirk. „Dictionaries and Spoken Language: A Corpus-Based Review of French Dictionaries". *International Journal of Lexicography* 28, nr 2 (2015): 139–68.

Zaśko-Zielińska, Monika, Anna Majewska-Tworek, Marta Śleziak, Artur Tworek. *Od rozmowy do korpusu, czyli jak zbierać i archiwizować dane mówione*. Wrocław: Uniwersytet Wrocławski–Quaestio, 2020.

„Zbiór zasad i wytycznych pt. «Dobre obyczaje w nauce»". Komitet Etyki w Nauce Polskiej Akademii Nauk, 2001. https://ken.pan.pl/images/2001_ ZbiorWytycznych. pdf [dostęp: 29.04.2025].