

Koncepcja Rozproszonego Korpusu Dyskursu Akademickiego

KAMIL WABNIC*

CITATION: Wabnic K., 2025, Koncepcja Rozproszonego Korpusu Dyskursu Akademickiego, „Stylistyka” XXXV: 417–430, <https://doi.org/10.25167/Stylistyka35.2025.21>

1. Wstęp

Rozproszony Korpus Dyskursu Akademickiego (RKDA) to zasób badawczy powstający w ramach badań nad polskim dyskursem akademickim prowadzonych w Zakładzie Lingwistyki Korpusowej Uniwersytetu Wrocławskiego. Potrzeba szeroko zakrojonych badań tekstów należących do dyskursu akademickiego wynika z kilku przyczyn. Badania takie umożliwiłyby zaktualizowany opis tekstów akademickich i naukowych, których wzorce gatunkowe i stylistyczne podlegają stałym zmianom. Ustalenie tych wzorców umożliwi lepszą realizację zadań dydaktycznych związanych z nauczaniem pisania akademickiego. Dodatkowo badania tekstów akademickich mogą wnieść istotny wkład w rozwój działań antyplagiatowych oraz proinnowacyjnych – w odniesieniu do tekstów zarówno studentów, jak i naukowców. Odrębną kwestią pozostają potencjalne wnioski płynące z badań, które mogą dotyczyć upowszechniania

* <https://orcid.org/0000-0001-9762-3420>, Uniwersytet Wrocławski, Wrocław, Polska, kamil.wabnic@uwr.edu.pl



myśli naukowej za pomocą języka polskiego – jest to szczególnie istotne w kontekście promowania tworzenia przez naukowców tekstów anglojęzycznych kosztem tekstów w języku polskim (por. Warchał 2024). Oczywiście takie działanie nie dziwi i jest uzasadnione, ponieważ tekst opublikowany w języku angielskim może mieć wielokrotnie więcej odbiorców i może zapewnić autorowi większą liczbę cytowań. Daleką, ale możliwą, konsekwencją odchodzenia od użycia języka polskiego w tekstach naukowych może być zubożenie kultury (por. np. Bartmiński 2014: 17–21). Jest to oczywiście problem związany z jeszcze romantyczną wizją języka jako podstawy tożsamości narodowej i wykracza on poza językoznawstwo, aczkolwiek może być istotny w kontekście kreowania polityki językowej, która zdaniem ekspertów jest w Polsce zaniedbana (Dubisz 2020). Wnioski płynące z badań tekstów akademickich mogą być zatem wielowymiarowe i mogą wykraczać poza problemy językoznawcze.

2. Stan badań nad tekstami akademickimi

Prace badawcze nad dyskursem akademickim są stale prowadzone na świecie ze względu na możliwości oceny skuteczności komunikacji naukowej oraz naukowo-dydaktycznej, a także potrzebę weryfikowania np. autentyczności autorstwa tekstów naukowych zarówno w odniesieniu do prac naukowców, jak i do prac studenckich. Z różnymi tradycjami badawczymi wiążą się odmienne podejścia do badania komunikacji akademickiej. W Polsce teksty naukowe i naukowo-dydaktyczne są przede wszystkim przedmiotem zainteresowania stylistyki. Można zaryzykować twierdzenie, że badanie tekstów naukowych stało się podstawą rozwoju tej subdyscypliny językoznawstwa – przede wszystkim za sprawą prac Stanisława Gajdy (zob. np. 1974, 1982, 2008, 2018). Na tym gruncie kategoria stylu ściera się z kategorią gatunku (por. Witosz 1999). Dość wyczerpującą bibliografię dotyczącą badań stylistycznych tekstów naukowych zgromadziła Agnieszka Jasińska w artykule *Od stylu naukowego do idei języka polskiego jako obcego do celów akademickich* (2023). Oprócz publikacji dotyczących stylistyki należy wskazać badania dotyczące dyskursu naukowego. Szczególnie istotne ze względu na międzynarodową recepcję są tutaj teksty Anny Duszak (1994, 1998). W tradycji anglosaskiej badania tekstów naukowych związane są właśnie z analizą dyskursu (zob. np. Hyland 2005, 2012) lub genologią (zob. np. Swales 1990,

1996), co ma swoje odbicie również w badaniach nad polskim dyskursem naukowym (zob. np. Rejter 2023a).

W badaniach korpusowych polskojęzyczny dyskurs akademicki pojawiał się jako przedmiot zainteresowania pośrednio od początku rozwoju dyscypliny w Polsce. W ramach prac prowadzonych nad frekwencją leksyki polszczyzny lat 70. XX wieku jednym z badanych stylów funkcjonalnych był styl popularnonaukowy (Kamińska-Szmaj 1990; Kurcz i in. 1974, 1990). Także w późniejszym czasie w 1982 roku w ramach projektu badań nad polszczyzną mówioną miast (Nieckula 1990) były prowadzone szeroko zakrojone prace związane z gromadzeniem nagrań tekstów mówionych – wśród tych nagrań znalazły się m.in. nagrania *Zebranie zakładu* oraz *Dyskusja naukowa*. Badania nad tymi tekstami były kontynuowane w XXI wieku za pomocą nowych metod komputerowych (Majewska-Tworek i in. 2020).

Oprócz badań nad dyskursem akademickim prowadzonych przy okazji szerszych projektów należy wskazać studia poświęcone wyłącznie komunikacji akademickiej. Już w XXI wieku Wydział Filologiczny UWr uczestniczył w projekcie GeWiss, który dotyczył badań kontrastywnych niemieckich, angielskich i polskich tekstów mówionych w stylu naukowym (zob. np. Fandrych i in. 2009, 2017). Zgromadzony wówczas polskojęzyczny zasób obejmował nagrania ok. 20 godzin rozmów egzaminacyjnych i był wykorzystywany w badaniach także po zakończeniu projektu (zob. np. Rogoziński, Zarzeczny 2013; Śmiełowska, Wabnic 2019).

Inne badania nad dyskursem akademickim prowadzone na polskojęzycznych tekstach współczesnych (z XXI wieku) nie miały charakteru systematycznego oraz korpusowego. Badania tekstów naukowych były prowadzone w kontekście nauczania języka polskiego jako obcego (zob. np. Rudziński 2004) czy przekładoznawstwa (zob. np. Hanus, Szwed 2014; Bączkowska 2019), a także popularyzacji nauki (zob. np. Barańska-Szmitko 2022) – na temat tego ostatniego problemu w 2002 roku opublikowano obszerną monografię wieloautorską *O trudnym łatwo* (Miodek, Zaśko-Zielińska 2002). Liczne prace dotyczą też różnych elementów struktury tekstów naukowych, np. tytułów (zob. np. Dawidziak-Kładocznna 2014; Pawłowski, Łach 2020), przedmów do monografii (zob. np. Pałucka-Czeriak 2002; Mikołajczak 2011) czy całych gatunków (zob. np. Bulisz 2017). Istnieją również liczne badania dotyczące poszczególnych jednostek języka w kontekście stylu naukowego (zob. np. Kananowicz 2023).

Na uwagę zasługuje ponadto obszerny zasób badań diachronicznych dotyczących tekstów, które również pośrednio wpisują się w zakres dyskursu akademickiego. Oczywiście nie można bezpośrednio przekładać współczesnego postrzegania świata na rzeczywistość przeszłą, ale różne specjalistyczne odmiany polszczyzny dawnej dały początek współczesnemu dyskursowi naukowemu, który stanowi element dyskursu akademickiego. Warto w tym obszarze badawczym wskazać liczne prace dotyczące tekstów specjalistycznych, które ze współczesnej perspektywy można uznać za element dyskursu naukowego. Badaczy zajmowały różne zjawiska, przede wszystkim leksykalny poziom języka (zob. np. Jankowiak 2009, 2011; Olędzka 2018; Paluszak-Bronka 2023; Szczaus 2013, 2016), ale też zagadnienia tekstologiczne i pragmatyczne czy dyskursologiczne (Ostaszewska 2001; Rejter 2023b, 2023c), językowy obraz świata (zob. np. Ostaszewska, Sławek 1999; Kowalska 2021; Pasternak 2024) czy nawet problem stylu osobniczego (Bieńkowska, Umińska-Tytoń 2011) – co może być szczególnie istotne z punktu widzenia potencjału aplikacji wyników badań w kontekście działań antyplagiatowych. Nie wskazuję tutaj oczywiście wszystkich prac, ponieważ niniejszy artykuł nie ma na celu wskazania kompletnego stanu badań.

Na gruncie międzynarodowym przede wszystkim istotne są prace poświęcone anglojęzycznemu dyskursowi akademickiemu. Badania te mają charakter korpusowy i dotyczą np. stylu (zob. np. Hyland 2012), sposobów prowadzenia argumentacji (zob. np. Freddi 2005) czy też akademickich gatunków mówionych (zob. np. Evison 2013). Wagę badań dotyczących dyskursu akademickiego podkreśla fakt istnienia czasopism: „English for Specific Purposes” i „Journal of English for Academic Purposes”. Pierwsze z nich pośrednio dotyczy dyskursu akademickiego, a drugie w całości jest mu poświęcone.

3. Cel korpusu

Celem przedsięwzięcia jest zatem przede wszystkim zebranie materiału do lingwistycznych badań nad polskojęzycznymi tekstami naukowymi oraz akademickimi tekstami dydaktycznymi. Teksty zbierane w ramach zasobu będą należały do szeroko pojętego dyskursu akademickiego. Dyskurs rozumiem za Marią Wojtak jako „całokształt praktyk komunikacyjnych związanych z określoną dziedziną ludzkiej aktywności” (Wojtak 2010: 81), a zatem teksty wchodzące

w skład korpusu będą realizacjami gatunków powstających w związku z aktywnością akademicką. Dyskurs akademicki można rozumieć bardzo szeroko, a gatunki z nim związane mogą – przede wszystkim ze względu na kontekst, czyli m.in. intencje nadawcy oraz funkcje komunikacji (zob. Enkvist 1973: 54; Gajda 1982: 68; van Dijk 2001: 27) – funkcjonować w ramach różnych odmian stylistycznych języka. Należy tutaj wskazać przede wszystkim konteksty działalności pozajęzykowej, z którymi nierozzerwalnie związana jest działalność językowa: naukowy, dydaktyczny i popularnonaukowy. Te trzy podstawowe obszary składają się na aktywność akademicką, chociaż właściwie tworzą one układ hierarchiczny, ponieważ prymarne są aktywność naukowa oraz poziomy transfer wiedzy¹ (do innych naukowców), który dokonuje się za pomocą tekstów naukowych. Wobec działalności naukowej sekundarna jest aktywność umożliwiająca pionowy transfer wiedzy: do adeptów nauki (działalność i teksty dydaktyczne) oraz do ogółu społeczeństwa (działalność i teksty popularnonaukowe). W odrębnym planie – niejako równoległym – należy również wskazać wszelką działalność administracyjną, która ma na celu umożliwienie prowadzenia nauki. Gatunki tekstów administracyjnych związanych z prowadzeniem działalności akademickiej mogą charakteryzować się różnym stopniem specyficzności. Mogą być zupełnie tożsame z tekstami z innych obszarów, ale również mogą być unikatowe – przynajmniej w odniedzinie gatunkowej, np. wniosek o grant.

Projektowany zbiór będzie miał formę korpusu tekstów. Oznacza to, że na materiał językowy, który będzie wchodził w skład zasobu, będą składały się autentyczne teksty w całości lub we fragmentach. Teksty te będą wyposażone w stosowne metadane – pozwalające sortować teksty oraz porównywać je ze sobą według wybranych kryteriów. Wszystkie teksty obligatoryjnie będą także znakowane morfosyntaktycznie. Tak skonstruowany zbiór spełnia podstawową definicję korpusu językowego (Wabnic, Krzywdziński 2024: 87).

Szczególnie istotny jest również rozproszony charakter korpusu. Termin *korpus rozproszony* pojawia się w artykule Tadeusza Piotrowskiego i oznacza tam korpus „utrzymywany na wielu komputerach” (Piotrowski 2003: 147). W wypadku RKDA rozproszenie korpusu dotyczy nie tyle sposobu jego przechowywania (wszak obecnie praktycznie każdy korpus jest w pewien sposób

¹ Przekazywanie idei i przekonań jest jednym z podstawowych aspektów dyskursu według van Dijka (2001: 10).

rozproszony²), ile sposobu gromadzenia i przetwarzania materiału. Na RKDA będą składały się subkorpusy tworzone niezależnie od siebie. Każdy subkorpus wchodzący w skład RKDA będzie odrębnym zasobem specjalistycznym³ (zob. Zaśko-Zielińska 2013: 25–26; Hunston 2022: 21–23). Subkorpusy będą przygotowywane według odrębnych założeń, dostosowanych do charakterystyki zbieranych tekstów – będą różniły się zatem zarówno metadanymi, jak i sposobem przetwarzania i anotacji.

4. Aktualny stan prac nad RKDA

Obecnie⁴ trwają prace nad dwoma subkorpusem: subkorpusem abstraktów oraz subkorpusem wykładów.

4.1. Subkorpus abstraktów

Subkorpus abstraktów będzie obejmował abstrakty artykułów opublikowanych w polskich czasopismach naukowych w 2023 roku. Estymowana⁵ objętość subkorpusu to 13 tys. abstraktów. Na zasób mają składać się wyłącznie teksty opublikowane w czasopismach z listy ministerialnej obowiązującej w 2023 roku. Czasopisma powinny być wydawane w Polsce (według danych w bazach na podstawie strony punktoza.pl) i w języku polskim. Subkorpus gromadzi wyłącznie abstrakty artykułów⁶. Zasób docelowo ma zawierać wszystkie abstrakty artykułów wybranych według wskazanych kryteriów, co zapewni zrównoważenie i reprezentatywność korpusu.

² Korpusy są obecnie przechowywane na wielu różnych komputerach ze względu na sposób ich przetwarzania. Podczas wykorzystywania różnych narzędzi analitycznych dane pochodzące z korpusu są przetwarzane np. na różnych serwerach zewnętrznych. Zwykle również gdzie indziej przechowywane są surowe – nieprzetworzone dane i metadane itp.

³ Specjalistyczność subkorpusów może polegać np. na ich jednogatunkowości.

⁴ Druga połowa 2024 roku.

⁵ Objętość została oszacowana na podstawie zgromadzonego dotychczas materiału. Obecnie zgromadziliśmy 607 abstraktów. Abstrakty pozyskujemy z czasopism posortowanych malejąco według ich punktacji. Dotychczas przeszukaliśmy 5% czasopism punktowanych wydawanych w Polsce. Potencjalnie liczba zgromadzonych abstraktów może być wyższa niż zakładane 13 tys., ponieważ wraz ze spadkiem punktacji możemy spodziewać się, że rosnąć będzie liczba czasopism publikujących artykuły w języku polskim.

⁶ Teksty są przypisywane do gatunku na podstawie informacji z czasopisma, np. umieszczone w sekcji „artykuły”.

Oprócz tekstów abstraktów na zasób będą się składać również metadane. Metadane będą gromadzone w sposób, który pozwala na uzyskanie możliwie największej liczby informacji o tekście oraz o autorze – następnie będą one dostosowywane do potrzeb narzędzi, za pomocą których zasób będzie przetwarzany. Takie działanie jest konieczne ze względów metodologicznych – różne metadane są niezbędne, ponieważ umożliwiają porównywanie części subkorpusu ze sobą. Na zestaw metadanych składają się:

- numer identyfikacyjny dokumentu;
- autorzy tekstu;
- afiliacja autorów;
- stopień/tytuł;
- tytuł artykułu;
- tytuł czasopisma;
- tom/numer czasopisma;
- dyscypliny czasopisma;
- punktacja czasopisma;
- słowa kluczowe artykułu;
- ORCID autora;
- DOI artykułu;
- rok;
- tekst wieloautorski;
- URL artykułu;
- dyscyplina autora.

Pozyskiwanie danych językowych oraz metadanych będzie odbywało się półautomatycznie za pomocą skryptów w języku Python, które umożliwi parsowanie struktury kodu HTML. Odrębny problem, który również zostanie zautomatyzowany, to uzupełnienie metadanych o informacje dotyczące autorów tekstów. Są one pozyskiwane z systemu RAD-on za pomocą API tego systemu. Gromadzone dane i metadane będą ręcznie weryfikowane oraz uzupełniane. W kolejnym etapie prac – po zgromadzeniu materiału – teksty zostaną przetworzone w taki sposób, aby były one wyposażone w podstawową anotację morfosyntaktyczną. W tym celu subkorpus zostanie przetworzony za pomocą analizatorów i tagerów: Stanza (Qi i in. 2020) lub spaCy (Honnibal, Montani 2017). W takiej postaci subkorpus będzie mógł być przetwarzany za pomocą zewnętrznych narzędzi lub wzbogacany o inne warstwy anotacji.

Planujemy przeprowadzenie podstawowego opisu ilościowego zgromadzonego materiału. Opis będzie obejmował: frekwencyjne listy rangowe lematów oraz n-gramów, a także dystrybucję poszczególnych n-gramów oraz lematów. Analiza zostanie przeprowadzona za pomocą narzędzia AntConc (Anthony 2024) lub innego o podobnych możliwościach. Z myślą o dalszej analizie – także jakościowej – zasób zostanie umieszczony w systemie Korpusomat.eu oraz – z myślą o bardziej złożonych analizach – w systemie Sketch Engine. Umożliwi to także opracowanie hipotez badawczych, których zweryfikowanie będzie możliwe w ramach dalszych badań. W takiej formie subkorpus będzie mógł być – zależnie od potrzeb – udostępniany badaczom. Ze względu na prawo autorskie udostępnianie korpusu będzie ograniczone.

4.2. Subkorpus wykładów

Subkorpus wykładów powstanie w ramach projektu studenckiego realizowanego w trakcie seminarium licencjackiego na filologii polskiej Uniwersytetu Wrocławskiego w roku akademickim 2024/2025 i przy współpracy z Pracownią Polszczyzny Mówionej UWr. Koncepcja subkorpusu będzie ostatecznie opracowywana przez seminarzystów, jednakże zasób powinien obejmować kilkanaście godzin multimodalnych nagrań (dźwięk: słowa wykładowcy oraz obraz: prezentacja lub tablica) wykładów prowadzonych w Instytucie Filologii Polskiej UWr dotyczących zagadnień z zakresu literaturoznawstwa i językoznawstwa. Otwartość zasobu będzie zależała od zgód wyrażonych przez nadawców tekstów, ale planujemy udostępnienie przynajmniej części zgromadzonego materiału badawczego.

Zestaw metadanych wymaga opracowania, ale powinien być dostosowany do specyfiki tekstu multimodalnego. Metadane mogą być wzorowane np. na tych wykorzystywanych w ramach projektu *Biografii językowych Dolnoślązaków* (Zaśko-Zielińska i in. 2020: 67–73) czy opracowanych jako profil CMDI *DigitalMediaText* w ramach konsorcjum CLARIN (*DigitalMediaText*). Same nagrania muszą zostać odpowiednio przetworzone – w tym celu zostaną najpierw przetranskrybowane automatycznie za pomocą modelu językowego, np. Whisper (Radford i in. 2023), a następnie transkrypcje zostaną ręcznie zweryfikowane i poprawione, np. w programie Elan⁷ (Sloetjes, Wittenburg

⁷ <https://archive.mpi.nl/tla/elan>.

2008). Transkrypcje w takiej formie mogą być dalej przetwarzane stosownymi narzędziami w zależności od potrzeb badaczy.

Na zgromadzonym w ten sposób materiale planowane są w ramach seminarium prace licencjackie dotyczące m.in.: frekwencji i dystrybucji lematów, jakości i dystrybucji n-gramów, struktury gatunku wykładu, aspektów związanych ze stylistyką wykładu, zjawisk pragmatycznych z zakresu teorii aktów mowy, struktury definicji itp. Oprócz prac licencjackich planowana jest również publikacja zbiorowa z najważniejszymi fragmentami prac.

5. Podsumowanie

Projektowany korpus będzie zasobem danych językowych, który ze względu na swój rozproszony charakter umożliwi gromadzenie różnorodnych tekstów w ramach specjalistycznych subkorpusów. Każdy subkorpus przygotowywany jest niezależnie według założeń dostosowanych do potrzeb konkretnego materiału badawczego (zróżnicowanego pod względem stylu, gatunku czy medium). Obecnie nasze prace koncentrują się na przygotowaniu dwóch subkorpusów: abstraktów artykułów oraz wykładów. Kolejne etapy rozwoju RKDA będą obejmowały rozszerzenie materiału o inne subkorpusy tekstów naukowych, dydaktycznych i popularnonaukowych. Nie wykluczamy także włączenia do zasobu akademickich tekstów administracyjnych – wspomnianych wcześniej wniosków o grant.

Literatura

- Anthony L., 2024, *AntConc* (Version 4.3.0) [Computer Software], Tokyo, Japan: Waseda University. Available from <https://www.laurenceanthony.net/software>.
- Barańska-Szmitko A., 2022, Wyznaczniki tabloidyzacji popularnonaukowego dyskursu historycznego w kanale serwisu YouTube „Historia bez Cenzury”, *„Stylistyka”*, 31, s. 137–175.
- Bartmiński J., 2014, Język w kontekście kultury. – *Współczesny język polski*, red. J. Bartmiński, Lublin: Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, s. 13–22.
- Bączkowska A., 2019, Dyskurs medyczny a przekład naukowych tekstów medycznych, *„Konińskie Studia Językowe”*, 7(3), s. 299–315.
- Bieńkowska D., Umińska-Tytoń E., 2011, Indywidualne cechy stylu tekstów naukowych (na przykładzie wybranych prac Aleksandra Brücknera i Stanisława

- Rosponda), „*Studia Językoznawcze. Synchroniczne i diachroniczne aspekty badań polszczyzny*”, 10, s. 33–57.
- Bulisz E., 2017, O artykule naukowym słów kilka, „*Progress. Journal of Young Researchers*”, 1, s. 13–21.
- Dawidziak-Kładoczna M., 2014, Syntaktyczne ukształtowanie tytułów tekstów naukowych, „*Prace Naukowe Akademii im. Jana Długosza w Częstochowie. Filologia Polska. Językoznawstwo*”, 10, s. 87–103.
- DigitalMediaText, https://catalog.clarin.eu/ds/ComponentRegistry#!/?itemId=clarin.eu-%3Acr1%3Aap_1456409483204®istrySpace=public.
- Dijk van T.A., 2001, Badania nad dyskursem. – *Dyskurs jako struktura i proces*, red. T.A. van Dijk, przeł. G. Grochowski, Warszawa: Wydawnictwo Naukowe PWN, s. 9–44.
- Dubisz S., 2020, Polityka językowa, „*Poradnik Językowy*”, 778(9), s. 118–120.
- Duszak A., 1994, Academic discourse and intellectual styles, „*Journal of Pragmatics*”, 21(3), s. 291–313.
- Duszak A., 1998, Academic writing in English and Polish: Comparing and subverting genres, „*International Journal of Applied Linguistics*”, 8(2), s. 191–213.
- Enkvist N.E., 1973, *Linguistic stylistics*, Hague–Paris: Mouton.
- Evison J., 2013, A corpus linguistic analysis of turn-openings in spoken academic discourse: Understanding discursive specialisation, „*English Profile Journal*”, 3(1), s. 1–24.
- Fandrych Ch., Meißner C., Wallner F. (red.), 2017, *Gesprochene Wissenschaftssprache – digital: Verfahren zur Annotation und Analyse mündlicher Korpora*, Tübingen: Stauffenburg.
- Fandrych Ch., Tschirner E., Meißner C., Rahn S., Slavcheva A., 2009, Gesprochene Wissenschaftssprache kontrastiv: Deutsch im Vergleich zum Englischen und Polnischen. Vorstellung eines gemeinsamen Forschungsvorhabens, „*Studia Linguistica*”, 28, s. 7–30.
- Freddi M., 2005, Arguing linguistics: Corpus investigation of one functional variety of academic discourse, „*Journal of English for Academic Purposes*”, 4(1), s. 5–26.
- Gajda S., 1974, Uwagi o stylu naukowym, „*Studia Śląskie*”, 26, s. 95–103.
- Gajda S., 1982, *Podstawy badań stylistycznych nad językiem naukowym*, Warszawa: Państwowe Wydawnictwo Naukowe.
- Gajda S., 2008, Metafora w dyskursie naukowym, „*Studia Litteraria Polono-Slavica*”, 8, s. 29–42.
- Gajda S., 2018, Na tropach stylu (rozważania metastylistyczne), „*Poradnik Językowy*”, 753(4), s. 7–20.
- Hanus A., Szwed I., 2014, Przekład tekstu naukowego na tle lingwistycznych badań polonistycznych i germanistycznych, „*Stylistyka*”, 23, s. 357–376.

- Honnibal M., Montani I., 2017, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, „*To appear*”, 7(1), s. 411–420.
- Hunston S., 2022, *Corpora in applied linguistics. Second edition*, Cambridge–New York: Cambridge University Press.
- Hyland K., 2005, Stance and engagement: A model of interaction in academic discourse, „*Discourse Studies*”, 7(2), s. 173–192.
- Hyland K., 2012, Bundles in academic discourse, „*Annual Review of Applied Linguistics*”, 32, s. 150–169.
- Jankowiak L.A., 2009, Z dziejów terminologii medycznej. Słownictwo medyczne wieku XVI wobec XVII w. oraz pierwszej połowy XVIII w., „*Studia z Filologii Polskiej i Słowiańskiej*”, 44, s. 61–83.
- Jankowiak L.A., 2011, Polskie słownictwo medyczne w drugiej połowie XIX wieku oraz na przełomie XIX i XX wieku, „*Studia Językoznawcze. Synchroniczne i diachroniczne aspekty badań polszczyzny*”, 10, s. 97–109.
- Jasińska A., 2023, Od stylu naukowego do idei języka polskiego jako obcego do celów akademickich, „*Słowo. Studia językoznawcze*”, 14, s. 89–97.
- Kamińska-Szmaj I., 1990, *Różnice leksykalne między stylami funkcjonalnymi polszczyzny pisanej. Analiza statystyczna na materiale słownika frekwencyjnego*, Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego.
- Kananowicz T., 2023, Organizacja tekstu naukowego oglądana z perspektywy modeli zdaniowych (na materiale z obszaru humanistyki), „*Stylistyka*”, 32, s. 281–298.
- Kowalska D., 2021, Obraz zarazy w świetle traktatu medycznego Piotra Umiastowskiego pt. *Nauka o morowym powietrzu na czwory księgi przełożona* (1591), „*Roczniki Humanistyczne*”, 69(6), s. 85–104.
- Kurcz I., Lewicki A., Sambor J., Woronczak J., 1974, *Słownictwo współczesnego języka polskiego: listy frekwencyjne*, t. 1: *Teksty popularnonaukowe*, Warszawa: Instytut Języka Polskiego PAN.
- Kurcz I., Lewicki A., Sambor J., Szafran K., Woronczak J., 1990, *Słownik frekwencyjny polszczyzny współczesnej*, t. 1–2, Kraków: Instytut Języka Polskiego PAN.
- Majewska-Tworek A., Zaśko-Zielińska M., Pęzik P., 2020, „Polszczyzna mówiona miast” – kontynuacja badań z lat 80. XX wieku z wykorzystaniem narzędzi lingwistyki cyfrowej, „*Forum Lingwistyczne*”, 7, s. 71–87.
- Mikołajczak B., 2011, Interpersonalność jako cecha tekstów naukowych na przykładzie przedmowy do monografii naukowej, „*Lingwistyka Stosowana*”, 4, s. 137–148.
- Miodek J., Zaśko-Zielińska M., 2002, *O trudnym łatwo. Materiały sesji poświęconej popularyzacji nauki*, Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego.
- Nieckula F. (red.), 1990, *Polszczyzna mówiona Wrocławia*, cz. 1: *Studia wstępne*, Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego.

- Olędzka B., 2018, Dynamika rozwoju polskich nazw chorób w okresie od XVIII do XX wieku, „*Poradnik Językowy*”, 759(10), s. 67–81.
- Ostaszewska D., 2001, Relacje nadawczo-odbiorcze w tekście naukowym (perspektywa historyczna). – *Stylistyka a pragmatyka*, red. B. Witosz, Katowice: Wydawnictwo Uniwersytetu Śląskiego, s. 191–200.
- Ostaszewska D., Sławek E.J., 1999, Procesy nazwotwórcze a językowy obraz świata. Na materiale średniowiecznej terminologii botanicznej w słowniku Jana Stanki. – *Przeszłość w językowym obrazie świata*, red. A. Pajdzińska, P. Krzyżanowski, Lublin: Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, s. 149–162.
- Paluszak-Bronka A., 2023, Nazwy wybranych metod leczenia (związanych z humoralną koncepcją patogenезy) na podstawie rękopiśmiennego poradnika medycznego Józefa Zmijewskiego, „*Studia Językoznawcze. Synchroniczne i diachroniczne aspekty badań polszczyzny*”, 22, s. 79–94.
- Pałucka-Czerniak I., 2002, Językowe przejawy obecności nadawcy w przedmowach do książek naukowych i popularnonaukowych. – *O trudnym łatwo. Materiały sesji poświęconej popularyzacji nauki*, red. J. Miodek, M. Zaśko-Zielińska, Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego, s. 111–127.
- Pasternak S., 2024, Językowy obraz mężczyzny na podstawie dolegliwości opisanych w renesansowym Zielniku Szymona Syreńskiego. – *Obszary Polonistyki 7. (De)konstruowanie męskości. (De)mitologizacje, (dez)iluzje i (de)heroizacje męskości w literaturze, kulturze i języku*, red. K. Matuszko, M. Głos, K. Maciąg, B. Drozd, Rzeszów: Wydawnictwo Uniwersytetu Rzeszowskiego, s. 132–139.
- Pawłowski A., Łach A., 2020, Informacja czy perswazja? Tytuły publikacji naukowych w świetle badań empirycznych, „*Język Polski*”, 100(2), s. 5–19.
- Piotrowski T., 2003, Językoznawstwo korpusowe: wprowadzenie do problematyki. – *Językoznawstwo w Polsce. Stan i perspektywy*, red. S. Gajda, Opole: Polska Akademia Nauk – Komitet Językoznawstwa, Uniwersytet Opolski – Instytut Filologii Polskiej, s. 143–154.
- Qi P., Zhang Y., Zhang Y., Bolton J., Manning Ch. D., 2020, Stanza: A Python natural language processing toolkit for many human languages, „*Association for Computational Linguistics (ACL) System Demonstrations*”.
- Radford A., Kim J.W., Xu T., Brockmann G., McLeavey C., Sutskever I., 2023, Robust speech recognition via large-scale weak supervision, „*Proceedings of the 40th International Conference on Machine Learning*”, s. 28492–28518.
- Rejter A., 2023a, The scientific discourse of contemporary humanities: Towards postmodern de/normalization (?), „*Poznańskie Studia Polonistyczne. Seria Językoznawcza*”, 30(50), s. 141–151.
- Rejter A., 2023b, Wspólnotowy wymiar dawnego dyskursu naukowego (XVI–XVII wiek), „*Acta Universitatis Lodzensis. Folia Linguistica*”, 57, s. 33–44.

- Rejter A., 2023c, Zielnik jako gatunek dawnego dyskursu naukowego we współczesnych odsłonach, „*Prace Językoznawcze*”, 25(3), s. 75–88.
- Rogosiński Ł., Zarzeczny G., 2013, Mówca AD 1489 w świetle wybranych parametrów ilościowych. – *Sapientia ars vivendi. Księga Jubileuszowa ofiarowana Profesor Annie Dąbrowskiej*, red. A. Burzyńska-Kamieniecka, A. Libura, Wrocław: Atut, s. 317–333.
- Rudziński G., 2004, *Charakterystyka językowa tekstów naukowych a glottodydaktyka (na materiale prac socjologicznych)*, Łódź: Wydawnictwo Uniwersytetu Łódzkiego.
- Sloetjes H., Wittenburg P., 2008, Annotation by category – ELAN and ISO DCR, „*Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*”, s. 816–820.
- Swales J.M., 1990, *Genre analysis: English in academic and research settings*, Cambridge: Cambridge University Press.
- Swales J.M., 1996, Occluded genres in the academy: The case of the submission letter. – *Academic writing: Intercultural and textual issues*, red. E. Ventola, A. Mauranen, Amsterdam–Filadelfia: John Benjamins, s. 45–58.
- Szczaus A., 2013, *Leksyka specjalistyczna w „Informacyi matematycznej” Wojciecha Bystrzonowskiego z 1749 roku na tle polszczyzny XVIII wieku*, Szczecin: Wydawnictwo Naukowe Uniwersytetu Szczecińskiego.
- Szczaus A., 2016, Propozycje terminologiczne z zakresu architektury i budownictwa w *Budowie kościołów Piotra Aignera (1825)*, „*Studia Językoznawcze. Synchroniczne i diachroniczne aspekty badań polszczyzny*”, 15, s. 227–240.
- Śmiełowska N., Wabnic K., 2019, Rozmowa egzaminacyjna – wybrane cechy stylu wypowiedzi egzaminatora i egzaminowanego. Analiza korpusowa, „*Studia Linguistica*”, 38, s. 117–135.
- Wabnic K., Krzywdziński J., 2024, Słownik frekwencyjny w stulecie urodzin jego twórcy, „*Poradnik Językowy*”, 815(6), s. 86–95.
- Warchał K., 2024, Attitudes towards English for research publication in a multilingual context: The case of Polish linguists, „*Journal of English for Academic Purposes*”, 71, 101403.
- Witosz B., 1999, Czy gatunek i styl są we współczesnej stylistyce pojęciami konkurencyjnymi?, „*Stylistyka*”, 8, s. 37–52.
- Wojtak M., 2010, Styl dziennikarstwa prasowego w perspektywie dyskursywnej. – *Styl – dyskurs – media*, red. B. Bogólewska, M. Worsowicz, Łódź: Wydawnictwo Uniwersytetu Łódzkiego, s. 81–91.
- Zaśko-Zielińska M., 2013, *Listy pożegnalne. W poszukiwaniu lingwistycznych wyznaczników autentyczności tekstu*, Wrocław: Quaestio.

Zaśko-Zielińska M., Majewska-Tworek A., Śleziak M., Tworek A., 2020, *Od rozmowy do korpusu. Czyli jak zbierać i archiwizować dane mówione*, Wrocław: Quaestio.

The concept of the Dispersed Corpus of Academic Discourse

This article aims to present the concept of the Dispersed Corpus of Academic Discourse (Rozproszony Korpus Dyskursu Akademickiego – RKDA), still under construction. The corpus is a research project focused on collecting linguistic data to facilitate multifaceted studies of Polish scholarly, didactic, and popular science texts. Current efforts concentrate on the development of two subcorpora: one consisting of abstracts from academic articles published in 2023 in journals from the official ministerial list, and the other comprising academic lectures. The latter is at a more advanced stage and is expected to be completed in 2025. Ultimately, RKDA will include both written and spoken texts across various genres of academic discourse, ensuring broad applicability for research purposes. Data collection methods are tailored to the type of text, with separate metadata sheets customized for each subcorpus. The corpus includes basic morphosyntactic annotation (Stanza), which is further enriched with additional layers of annotation as analysis progresses. Depending on licensing agreements and author permissions, parts of the corpus will be made available via Korpusomat.eu and the Sketch Engine. RKDA's main goal is to enable systematic research on academic texts, including stylistic and genological analysis, as well as to support anti-plagiarism and pro-innovation initiatives within the academic community. The originality of the project lies primarily in its combination of written and spoken data, and in the creation of a partially open, multi-genre resource of academic discourse developed in stages and continually expanded.

Keywords: *Dispersed Corpus of Academic Discourse, academic discourse, scientific style, corpus linguistics*