DOI: 10.19195/2082-8322.9.13

Adrian Zasina

ORCID: 0000-0001-9348-5833

Uniwersytet Karola, Praga

Językoznawstwo korpusowe. Empiryczne podejście w badaniach humanistycznych

1. Wstęp

Empiryczne podejście w nauce jest przede wszystkim kojarzone z naukami ścisłymi, które opierają swoje badania na obserwacji zjawisk, a postawione hipotezy weryfikuje się na podstawie przeprowadzonych doświadczeń. Takie podejście jest również możliwe w naukach humanistycznych, głównie w językoznawstwie, które od ubiegłego wieku w ramach węższej dziedziny, zwanej językoznawstwem korpusowym, poświadcza wyniki badań autentycznymi tekstami zgromadzonymi w elektronicznej bazie danych. Zainteresowanie empirycznym podejściem w obszarze językoznawstwa systematycznie rośnie, jednak, jak zauważa Sampson¹, nadal dużą popularnością cieszą się prace introspekcyjne oparte na spekulacjach czy intuicji badacza, ponieważ wciąż mają swych odbiorców. Ten chwilowy stan prawdopodobnie wkrótce zmieni się, ustępując pola empirycznemu poznaniu języka².

Na przełomie lat siedemdziesiątych i osiemdziesiątych ubiegłego wieku wyodrębniła się nowa dyscyplina naukowa nazwana językoznawstwem korpusowym. Była to

¹ G. Sampson, *The empirical trend: Ten years on*, „International Journal of Corpus Linguistics” 18 (2), 2013, s. 281–289.

² *Ibidem*, s. 288.

bez wątpienia reakcja na postulaty Noama Chomsky'ego³, oparta na wrodzonej intuicji rodzimego użytkownika języka w odniesieniu do terminów gramatyczności i akceptowalności zdań. Według Chomsky'ego gramatyka to urządzenie, które generuje wszystkie zdania gramatyczne i wyłącznie zdania gramatyczne. Innym przyczynkiem do prac korpusowych były badania amerykańskiego językoznawcy Williama Labova⁴ z lat sześćdziesiątych XX wieku. Labov w spojrzeniu na gramatykę wykorzystał pojęcie prawdopodobieństwa. Opisywał system komunikacji społecznej, co pozwoliło stwierdzić, że pewne formy i konstrukcje w języku są częstsze od innych⁵.

Problematyczną kwestią dotyczącą metodologicznych aspektów językoznawstwa korpusowego jest różnica pomiędzy podejściem *corpus-based* a *corpus-driven* w badaniu danych językowych. Różnica między nimi powstaje na podstawie stopnia autonomii danych językowych przy prowadzeniu analizy korpusowej⁶. Niniejszą dysfunkcję wprowadza Elena Tonini-Bonelli⁷. Podejście *corpus-based* próbuje wyjaśnić, przetestować i zilustrować teorie oraz analizy, które powstały przed pojawieniem się obszernych korpusów językowych⁸. Natomiast podejście *corpus-driven* systematycznie identyfikuje kategorie językoznawcze na podstawie powtarzających się wzorców i częstotliwości wystąpień (dystrybucji frekwencyjnej), które ujawniają się bezpośrednio w kontekście⁹. Możemy zatem w uproszczeniu stwierdzić, iż pierwsze podejście opiera się na pewnych założeniach językowych i stara się zweryfikować teorię, podczas gdy drugie podejście próbuje konstruować nowe teorie na podstawie obserwowanych danych w korpusie.

1.1. Definicja językoznawstwa korpusowego

Znając uwarunkowania historyczne, zapoznajmy się z definicją językoznawstwa korpusowego: „Językoznawstwo korpusowe jest jedną z części językoznawstwa komputerowego i zajmuje się analizą języka zgromadzonego w korpusach językowych”¹⁰.

1.2. Pierwsze istotne korpusy

Pierwszy powszechnie używany korpus języka powstał w USA. Został zebrany w latach sześćdziesiątych XX wieku przez Henry'ego Kučerę i Nelsona Francisa¹¹ i za-

³ N. Chomsky, *Aspects of the Theory of Syntax*, Cambridge, Mass. 1965.

⁴ W. Labov, *The Social Stratification of English in New York City*, Washington 1966.

⁵ *Podstawy językoznawstwa korpusowego*, red. B. Lewandowska-Tomaszczyk, Łódź 2005, s. 9–10.

⁶ Ł. Grabowski, *A Corpus-Driven Study of Translational and Non-Translational Texts: The Case of Nabokov's Lolita*, Opole 2012, s. 23.

⁷ E. Tognini-Bonelli, *Corpus Linguistics at Work*, Amsterdam-Philadelphia 2001.

⁸ *Ibidem*, s. 65.

⁹ *Ibidem*, s. 87.

¹⁰ *Podstawy językoznawstwa korpusowego...*, s. 11.

¹¹ H. Kučera, W.N. Francis, *Computational Analysis of Present-Day American English*, Providence 1967.

wierał zaledwie jeden milion słów. Współcześnie korpusy posiadają obszerną bazę tekstów zawierającą setki milionów wyrazów. Komputery umożliwiły gromadzenie i analizowanie znacznie większej ilości materiału językowego niż przy wykorzystaniu tradycyjnych metod językoznawczych¹².

Pierwsze prace nad korpusem języka słowiańskiego sięgają początku lat dziewięćdziesiątych XX wieku. Był to mały korpus języka serbskiego. Prace w większym zakresie podjęto dopiero w roku 1994 w Republice Czeskiej (więcej o tym w podpunkcie 2.2). Od tego momentu rozpoczęto prace nad wielkimi referencyjnymi korpusami w innych krajach słowiańskich: Chorwacji, Polsce, Rosji, Słowacji czy Słowenii. Odkrycia czeskich naukowców były przełomowe dla dalszych badań korpusowych, gdyż automatyczna analiza morfologiczna i składniowa typowa dla języków flekcyjnych postawiła technologii większe wyzwanie¹³.

2. Korpus językowy

Zanim przejdziemy do opisu poszczególnych korpusów, skupimy się na definicji korpusu językowego. Jak podaje *Słownik języka polskiego* PWN (<https://sjp.pwn.pl/szukaj/korpus.html>), korpus to „teksty, dane itp. zgromadzone ze względu na swą reprezentatywność, stanowiące podstawę do analizy naukowej”. Inne źródło podaje, iż korpus to „zestaw tekstów językowych zebranych w celu badania jego systemu lub podsystemu”¹⁴. Dariusz Tkaczewski natomiast uważa, iż „jest to zespół udokumentowanych dowodów autentycznego użycia języka naturalnego, rozległy zespół elektronicznych tekstów celowo zgromadzonych jako referencyjne źródło”¹⁵. Możemy zatem uogólnić, że korpus językowy jest to zbiór różnych rodzajów tekstów pisanych i mówionych (w postaci przetranskrybowanych nagrań dźwiękowych) zgromadzonych w komputerowej bazie danych, która umożliwia automatyczne wyszukiwanie jednostek tekstowych w ich naturalnej strukturze gramatycznej i stylistycznej, co pozwala na sprawniejszą analizę językoznawczą.

2.1. Typologia korpusów

Ze względu na różne aspekty naukowe tworzone są różne typy elektronicznych korpusów językowych, które pełnią specyficzne funkcje. Możemy wyróżnić: k o r p u s y o g ó ł n e — zajmujące się w szerokim zakresie językiem (służą tworzeniu słowni-

¹² B. Lewandowska-Tomaszczyk et al., *Narodowy Korpus Języka Polskiego: Geneza i dzień dzisiejszy*, [w:] *Narodowy Korpus Języka Polskiego*, red. A. Przepiórkowski et al., Warszawa 2012, s. 3–10.

¹³ *Ibidem*, s. 5.

¹⁴ *Encyklopedia językoznawstwa ogólnego*, red. K. Polański, Wrocław 1999, s. 319.

¹⁵ D. Tkaczewski, *Český národní korpus — internetové źródlo standaryzacji i weryfikacji języka czeskiego oraz nowoczesne narzędzie dydaktyczne*, „Bohemistyka” 8, 2008, s. 363–378.

ków) — bądź korpusy specjalistyczne — skupiające się na jednej wybranej cesze, na podstawie której można wydzielić kolejne podtypy, jak na przykład korpus autorski (korpus dzieł wybranego autora), korpus określonego gatunku, korpus dialektu (na przykład hanackiego). Wziąwszy za kryterium historyczny punkt widzenia, wyróżniamy korpusy synchroniczne (opisujące obecny stan języka naturalnego, zawierające informację o zjawiskach językowych i przykłady używania wyrazów w odpowiednich kontekstach) i diachroniczne (zajmują się językiem na przestrzeni czasu, ukazują rozwój języka, przy czym gromadzenie tekstów jest tu znacznie trudniejsze). Ze względu na sposób komunikacji wyróżnia się korpusy języka mówionego (trudne i pracochłonne w gromadzeniu materiałów, gdyż do ich utworzenia zbiera się nagrania, przepisuje się i przetwarza w teksty lingwistyczne) i korpusy języka pisanego (bazują na tekstach książek, czasopism i gazet, ale i te trzeba poddać obróbce strukturalnej i lingwistycznej). Korpusy można podzielić również na jednojęzyczne i wielojęzyczne, a te z kolei mogą być porównywalne¹⁶ bądź paralelne¹⁷. Wykorzystuje się w nich specjalne programy zestawiające teksty w dwóch językach, tak by wyrazy nawiązywały do siebie (z ang. *alignment* — wiązanie). Takie korpusy są przydatne w translatoryce, gdyż oferują znacznie większy zakres ekwiwalentów słów w różnych związkach zdaniowych. Ich zasób leksykalny i frazeologiczny staje się coraz częściej źródłem pozwalającym stworzyć doskonalsze translatory komputerowe¹⁸. Ułatwiają one przekład, jednak nadal nie są tak dokładne jak przekłady wykonane przez człowieka, aczkolwiek stale pracuje się nad udoskonalaniem ich funkcjonowania. Jednym z powszechnie znanych i często używanych jest *Google translator*. Z dzisiejszego punktu widzenia możemy również mówić o korpusach tradycyjnych i internetowych gromadzących w swoich zasobach teksty pochodzące z Internetu.

2.2. Korpusy języków słowiańskich

Pierwszy korpus języka słowiańskiego powstał już na początku lat dziewięćdziesiątych XX wieku. Był to mały korpus języka serbskiego, jednakże z przyczyn politycznych i ekonomicznych przerwano prowadzone nad nim prace¹⁹. Za najbardziej doświadczonych w dziedzinie słowiańskiej lingwistyki korpusowej są uważani Czeši, którzy jako pierwsi poradzili sobie z problemem komputerowego przetworzenia języ-

¹⁶ Zob. V. Benko, *Aranea: Yet Another Family of (Comparable) Web Corpora*, [w:] *Text, Speech and Dialogue*, red. P. Sojka et al., Brno 2017, s. 257–264. Korpusy porównywalne mogą być również jednojęzyczne — zob. <http://wiki.korpus.cz/doku.php/en:cnk:jerome>.

¹⁷ Zob. E. Kaczmarek, A. Rosen, *Praktyczny przewodnik po korpusie równoległym InterCorp*, [w:] *Praktyczny przewodnik po korpusach języków słowiańskich*, red. M. Hebal-Jezińska, Warszawa 2014, s. 207–231.

¹⁸ D. Tkaczewski, *op. cit.*, s. 368–370.

¹⁹ M. Šulc, *Korpusová lingvistika: První stup*, Praha 1999, s. 42.

ka fleksyjnego²⁰. Ich dokonania były pomocne przy tworzeniu pozostałych korpusów słowiańskich. Obecnie w Internecie są dostępne korpusy referencyjne między innymi następujących języków słowiańskich: chorwackiego (<http://filip.ffzg.hr>), polskiego (<http://nkjp.pl/>), rosyjskiego (<http://www.ruscorpora.ru/>), czeskiego (<https://www.korpus.cz/>), słowackiego (<http://korpus.juls.savba.sk/>), słoweńskiego (<http://www.gigafida.net/>), które posiadają morfologiczną anotację umożliwiającą wyszukiwanie danych na podstawie informacji lingwistycznych (na przykład na podstawie kategorii gramatycznych). Pozostałe języki słowiańskie starają się w swych działaniach dorównać wynikom czeskich naukowców, a w niektórych obszarach nawet ich wyprzedzają. W następnych punktach skupimy naszą uwagę na trzech korpusach słowiańskich: języka rosyjskiego, polskiego oraz czeskiego²¹.

2.2.1. Narodowy Korpus Języka Rosyjskiego

Korpus współczesnego języka rosyjskiego (<http://www.ruscorpora.ru/>) w ramach Narodowego Korpusu Języka Rosyjskiego (NKJR) składa się w całości z ponad 300 milionów tekstowych słów²². Został opublikowany w roku 2004. Opracowywany jest w Instytucie Języka Rosyjskiego im. W.W. Winogradowa Rosyjskiej Akademii Nauk²³. Korpus jest dostępny dla wszystkich zainteresowanych językiem rosyjskim — językoznawców, nauczycieli, uczniów oraz studentów, a także obcokrajowców uczących się języka rosyjskiego. Korpus języka rosyjskiego jest reprezentatywny i otagowany. Wykorzystuje pięć typów anotacji: metatekstową, morfologiczną, syntaktyczną, akcentową i semantyczną. W przyszłości twórcy korpusu planują wprowadzić słowo-twórczą anotację. Stale pracuje się nad ulepszeniem sposobu anotacji.

Teksty NKJR obejmują głównie okres od połowy XVIII do początku XXI wieku. Okres ten ukazuje jednocześnie język poprzednich epok i współczesny język w różnych odmianach socjolingwistycznych — pisanych, kolokwialnych, potocznych, częściowo dialektalnych. W korpusie znajdują się oryginalne dzieła literackie (proza, dramaty, poezja), które mają znaczenie kulturowe oraz są interesujące z językowego punktu widzenia. Co więcej, NKJR to nie tylko korpus literatury pięknej. Prócz tekstów literackich obejmuje wiele innych gatunków tekstów: memuary, teksty prasowe, eseje, popularnonaukowe i naukowe teksty, mówione wypowiedzi, korespondencje, dzienniki, dokumenty itd.

Obecnie NKJR w swoich zbiorach posiada następujące korpusy: korpus główny (podstawowy) — obejmuje prozę (włącznie z dramatami) tekstów pisanych z okresu

²⁰ B. Lewandowska-Tomaszczyk *et al.*, *op. cit.*, s. 5.

²¹ Więcej informacji o pozostałych korpusach języków słowiańskich zawiera *Praktyczny przewodnik po korpusach języków słowiańskich*.

²² Aktualna statystyka korpusu jest dostępna na <http://www.ruscorpora.ru/corpora-stat.html>.

²³ Ł. Grabowski, *Korpusy w pracy idiomatykarzy: Wybrane studia przypadku na przykładzie Narodowego Korpusu Języka Rosyjskiego*, [w:] *Na tropach reproduktów: W poszukiwaniu wielowyrazowych jednostek języka*, red. W. Chlebda, Opole 2010, s. 47–58.

od XVIII do XXI wieku; korpus składniowy (dokładnie otagowany) — dla każdego zdania jest wytworzona kompletna morfologiczna i składniowa struktura (bank drzew); korpus tekstów prasowych (korpus współczesnych mediów) — obejmuje artykuły z mediów od roku 1990 do 2000; korpus równoległy — obecnie są dostępne następujące języki: angielski, niemiecki, francuski, hiszpański, włoski, ukraiński, polski i białoruski; korpus dialektalny — obejmuje dialekty różnych regionów Rosji, zachowuje ich cechy gramatyczne, posiada specjalną wyszukiwarkę biorącą pod uwagę morfologię dialektu; korpus poezji — możliwe jest wyszukiwanie nie tylko według anotacji gramatycznej i leksykalnej, ale również według typowych charakterystyk poezji, jak na przykład typy rymów; korpus uczniowski; korpus języka mówionego — zawiera zapisy magnetofonowych nagrań oficjalnego i prywatnego języka oraz zapisy filmów; korpus akcentowy (historyczny korpus zmian rosyjskiego akcentu) — teksty noszące informację o historii rosyjskiego akcentu, które można wyszukiwać na podstawie miejsca akcentu i prozodycznej struktury słowa; korpus multimedialny — obejmuje nagrania wideo, również fragmenty filmów z lat 1930–2000, gdzie można wyszukiwać nie tylko mówione teksty, ale też gesty (kiwanie głową, poklepanie po plecach itd.) oraz typy aktów mowy (zgoda, ironia itp.).

2.2.2. Narodowy Korpus Języka Polskiego

Prace korpusowe były skupione w różnych ośrodkach w Polsce. W latach dziewięćdziesiątych powstało kilka niezależnych od siebie ośrodków, które zajmowały się przetwarzaniem języka naturalnego (NLP). Językoznawcze, informatyczne, leksyko-graficzne grupy gromadziły jednocześnie dane korpusowe, a także tworzyły narzędzia do ich przetwarzania, między innymi w zespole Instytutu Podstaw Informatyki Polskiej Akademii Nauk (<http://zil.ipipan.waw.pl/>), zespole Instytutu Języka Polskiego Polskiej Akademii Nauk w Krakowie (<http://www.ijp-pan.krakow.pl/>), zespole korpusowym PELCRA (<http://pelcra.pl/>) w Katedrze Języka Angielskiego i Językoznawstwa Stosowanego na Uniwersytecie Łódzkim oraz zespole korpusowym w Redakcji Słowników Języka Polskiego Wydawnictwa Naukowego PWN (<http://korpus.pwn.pl/>).

W roku 2006 z inspiracji Prezydium Komitetu Językoznawstwa Polskiej Akademii Nauk i przy wsparciu ówczesnego przewodniczącego Stanisława Gajdy utworzono konsorcjum Narodowy Korpus Języka Polskiego, w którego składzie znaleźli się: IPI PAN (Adam Przepiórkowski), IJP PAN (Rafał Górski), UŁ (Barbara Lewandowska-Tomaszczyk) oraz PWN (Mirosław Bańko i Marek Łaziński). Rok później zespół uzyskał grant (od grudnia 2007 do grudnia 2010, przedłużony do czerwca 2011), którego celem było stworzenie NKJP oraz opracowanie narzędzi do jego obsługi. Koordynatorem projektu został IPI PAN, a kierownikiem Adam Przepiórkowski²⁴.

²⁴ B. Lewandowska-Tomaszczyk *et al.*, *op. cit.*, s. 8–9.

Obecnie do dyspozycji użytkowników pozostaje zrównoważona część NKJP (<http://www.nkjp.pl/>), która zawiera ponad 250 milionów słów (300 milionów tokenów), oraz cały korpus zawierający 1800 milionów tokenów. Do wyszukiwania służą dwie oficjalnie dostępne wyszukiwarki korpusowe: wyszukiwarka PELCRA oraz wyszukiwarka Poliqarp, w której są dostępne jeszcze inne korpusy, jak na przykład ręcznie otagowany milionowy korpus, dwa korpusy IPI, a także Korpus Słownika Frekwencyjnego. Wzorcowy skład tekstów w korpusie dla NKJP był opracowany przez Rafała L. Górskiego i Marka Łazińskiego²⁵. Naukowcy skupili się nie tylko na typach tekstów, ale również na zrównoważeniu tematycznym, jednakże nie starali się zapewnić reprezentatywności korpusu pod względem tematyki. Korpus tworzą następujące teksty: publicystyka i krótkie wiadomości prasowe (50,0%), literatura piękna (16,0%), literatura faktu (5,5%), typ informacyjno-poradnikowy (5,5%), typ naukowo-dydaktyczny (2,0%), inne teksty pisane (3,0%), książka niebeletrystyczna nieklasyfikowana (1,0%), teksty konwersacyjne, mówione medialne i quasi-mówione razem (10,0%), teksty internetowe statyczne i dynamiczne razem (7,0%).

2.2.3. Czeski Korpus Narodowy

Wiosną 1992 roku w Pradze grupa naukowców²⁶ założyła stowarzyszenie Počítačový fond češtiny. Jego celem była koordynacja pracy i zapewnienie komunikacji i współpracy specjalistom, którzy są zainteresowani komputerowym przetwarzaniem języka czeskiego. Ostatecznie starania badaczy zostały docenione i wsparte grantem na projekt *Počítačový korpus českých psaných textů*, w którym od roku 1993 brali udział naukowcy z Uniwersytetu Karola, Uniwersytetu Masaryka i Ústavu pro jazyk český. Projekt zakończono sukcesem w roku 1995²⁷. Formalny początek lingwistyki korpusowej w Republice Czeskiej datuje się jednak na rok 1994, kiedy to doszło do powstania na Wydziale Filozoficznym Uniwersytetu Karola wielodziałowego Ústavu Českého národního korpusu pod kierownictwem Franciszka Čermáka. Tak rozpoczął się rozwój prac nad korpusem narodowym. Wśród innych korpusów narodowych Czeski Korpus Narodowy (CKN) zajmuje nadrzędną pozycję w Europie²⁸. Pod koniec wieku XX wyklarowały się silne, o znaczeniu europejskim, ośrodki lingwistyki korpusowej w Republice Czeskiej, takie jak: Uniwersytet Karola, Uniwersytet Masaryka oraz Instytut Języka Czeskiego (ÚJČ AV ČR).

²⁵ R.L. Górski, M. Łaziński, *Typologia tekstów w NKJP*, [w:] *Narodowy Korpus Języka Polskiego...*, s. 13–23.

²⁶ Do tejże grupy należeli naukowcy z FF UK w Pradze: František Čermák, Jan Hajič, Eva Hajičová, Jan Králík, Karel Pala, Klára Osolobě, Věra Schmiedtová i inni.

²⁷ K. Osolobě, *Matematická lingvistika*, [w:] *Kapitoly z dějin české jazykovědné bohemistiky*, red. J. Pleskalová, Praha 2007, s. 447–467.

²⁸ F. Čermák, *Korpusová lingvistika*, [w:] *Kapitoly z dějin české jazykovědné bohemistiky...*, s. 468–472.

Pierwszy korpus w ramach CKN został opublikowany w roku 2000 (<http://korpus.cz/>). Był to korpus SYN2000 — zrównoważony korpus zawierający w przeważającej części teksty z lat 1990–1999. Od tej pory regularnie co pięć lat jest wydawany nowy ogólny reprezentatywny korpus synchroniczny czeszczyzny pisanej serii SYN, powstają również korpusy synchroniczne pisanej publicystyki skupiające się na tekstach prasowych serii PUB. Najnowszy korpus ogólny SYN2015 składa się z beletrystyki (33,33%), literatury naukowej (33,33%) i tekstów prasowych (33,33%) i zawiera 100 milionów słów. Prócz powyższych korpusów synchronicznych języka pisanego w CKN znajdziemy korpusy specjalistyczne: FSC2000 — referencyjne źródło Frekwencyjnego Słownika Języka Czeskiego, CZESL-PLAIN — korpus uczniowski zawierający teksty nierodzimych użytkowników języka czeskiego, CZESL-SGT — korpus uczniowski z automatyczną anotacją, KSK-DOPISY — Korpus Prywatnej Korespondencji, JEROME — porównywalny korpus przekładów, LINK — korpus naukowych tekstów z dziedziny językoznawstwa, ORWELL — korpus powieści Orwella 1984, SKRIPT2012 — korpus prac pisemnych uczniów szkół podstawowych, FicTree — ręcznie składowo anotowany korpus tekstów beletrystycznych.

Obok korpusów języka pisanego w CKN znajdziemy również korpusy języka mówionego. Pierwszym z nich był Praski Korpus Języka Mówionego (PMK) opublikowany w 2001 roku; rok później powstał Brneński Korpus Języka Mówionego (BMK). Kolejnym korpusem jest korpus SCHOLA2010, który zawiera teksty mówione godzin lekcyjnych, korpusy serii ORAL oraz nowej serii ORTOFON, a także korpusy specjalistyczne: DIALEKT, LIDSEI_CZ i SPEECHES.

Innymi projektami CKN są: korpus diachroniczny DIAKORP, korpus równoległy InterCorp, obcojęzyczne DOTKO, HOTKO, lEstRepublicain, internetowe korpusy serii WaC i inne²⁹.

3. Zastosowanie korpusów

Niewątpliwie silną stroną komputerowych korpusów jest syntagmatyka — kombinatoryczne możliwości występowania jakiejkolwiek wyszukiwanej formy. Dowolnie rozległy kontekst pozwala badaczowi odnaleźć informację nie tylko, jak dana forma się z nim łączy, ale również to, jak się jej zwyczajnie i prymarnie używa w różnych kontekstach. Ma to swoje zasadnicze skutki dla językoznawstwa i jego metodologii w poszerzaniu wiedzy o charakterze syntagmatycznym oraz dla praktyki leksykologicznej³⁰.

²⁹ Pełna lista korpusów wraz z ich parametrami jest dostępna na <https://wiki.korpus.cz/doku.php/cnk:uvod>.

³⁰ F. Čermák, *op. cit.*, s. 469.

Korpus umożliwia wykonywanie wielu operacji. Posługując się nim, możemy wyszukać: konkretną formę wyrazową wraz z jej kontekstem, jednostki wielowyrazowe (na przykład wyrażenia przymkowe, frazeologizmy), leksem lub hasło kluczowe, części mowy, kombinację części mowy i lematu. Jest to narzędzie, które umożliwia weryfikację słów nie tylko badaczom języka, lecz również miłośnikom poprawności językowej. Każdy ma łatwy dostęp do charakterystyki statystycznej i frekwencyjnej wyrazów swego języka ojczystego (na przykład najpopularniejsze czy najdłuższe wyrazy). Sprzyja to niewątpliwie poprawności językowej zgodnej z uzusem semantycznym. Szybko możemy sprawdzić poprawność nowych form językowych, których jeszcze drukowane słowniki nie zdążyły uwzględnić. Poszczególne jednostki leksykalne jesteśmy w stanie badać pod względem ich łączliwości (kolokacji) z innymi jednostkami³¹, weryfikować związki frazeologiczne³², walencję wyrazową³³, rekację czasownika³⁴. Dla dydaktyków to dobra pomoc naukowa, aby przygotować wiele ćwiczeń, nie tylko leksykalnych³⁵. Nieocenionym zastosowaniem CKN jest pomoc przy sporządzaniu wielu rodzajów słowników³⁶.

4. Podsumowanie

Niniejsza praca miała na celu przedstawienie metodologii badań korpusowych w językoznawstwie. Zapoznaje czytelnika z podstawowymi zagadnieniami językoznawstwa korpusowego oraz skupia swoją uwagę na korpusach języków słowiańskich. Charakteryzuje trzy wybrane korpusy narodowe: języka czeskiego, polskiego i rosyjskiego. Informuje również o możliwym zastosowaniu korpusów językowych w takich dziedzinach, jak na przykład leksykografia i leksykologia, frazeologia czy semantyka.

Tak naprawdę językoznawstwo korpusowe zaczyna się dopiero rozwijać, możemy więc spodziewać się wielu innowacji, które umożliwią analizę i statystykę językową w ramach różnych dyscyplin językoznawczych. Dzięki łatwemu dostępowi do korpusów za pomocą Internetu, każdy z nas może z nich korzystać, aby weryfikować poprawność swojego języka i otrzymywać statystyczne wyniki. Dla badaczy jest to

³¹ P. Baker, *Using Corpora in Discourse Analysis*, London-New York 2006, s. 95–120.

³² M. Kopřivová, *Contribution Towards a Corpus-Based Phraseology Minimum*, [w:] *Computational and Corpus-Based Phraseology*, red. R. Mitkov, Cham 2017, s. 220–231.

³³ A. Čermáková, *Valence českých substantiv*, Praha 2009.

³⁴ M. Pastuchowa, M. Siuciak, *Zmiany rekacji czasowników jako przejaw stałej tendencji w dziejach polszczyzny (na przykładzie dopełniacza i biernika)*, „Forum Lingwistyczne” 1, 2014, s. 77–87.

³⁵ A. J. Zasina, *Korpusy językowe w nauczaniu języków obcych — metoda, narzędzia, praktyka*, [w:] *Specyficzne potrzeby studentów szkół wyższych a nauczanie języków obcych*, t. 2. *Praktyczne narzędzia*, red. M. Jedynak, Wrocław 2018, s. 110–123.

³⁶ D. Tkaczewski, *op. cit.*, s. 375–377.

nowy sposób na sprawdzanie rezultatów swoich badań. Warto dalej śledzić rozwój tej dziedziny i czekać na to, co nowego przyniesie.

Corpus linguistics. An empirical approach to humanities research

Summary

The aim of the article is to shed light on the methodology of corpus research in the humanities, primarily in linguistics. Corpus linguistics emerged in the late 1970s and early 1980s, focusing on electronic language corpora. Corpora are collections of various types of texts (written and spoken) gathered in a computer database which makes it possible to automatically search for text units in their natural context. There are various types of corpora depending on the type of study. The first corpora were compiled for the English language, although more and more languages are acquiring their national corpora, like the National Corpus of the Polish Language, the Czech National Corpus or the National Corpus of the Russian Language.