Jarosław Pacuła
Uniwersytet Śląski w Katowicach
https://orcid.org/0000-0002-9972-7925
jaroslaw.pacula@us.edu.pl

Forum Lingwistyczne
nr 7, 2020
ISSN 2450-2758 (wersja elektroniczna)

DOI: http://doi.org/10.31261/FL.2020.07.09

Było, (nie) minęło. Przypomnienie tekstu Tony'ego McEnery'ego i Nicka Ostlera *A New Agenda for Corpus Linguistics – Working with All of the World's Languages*

It's (Not) All Just Water under the Bridge. The Reminder of the Text of Tony McEnery and Nick Ostler *A New Agenda for Corpus Linguistics – Working with All of the World's Languages*

Abstract: In this text I refererate and comment the work from two decades ago. I recall an article important for corpus linguistics – the work of Tony McEnery and Nick Ostler from 2000. In their paper they argue that corpus linguistics needs to expand to cover a wider set of languages. The challenge of building a multilingual corps is admittedly current. The development of corpus linguistics still depends on technical capabilities, but today we see significant changes in this field. In the article I not only present the reflections of British linguists, but I also show current and outdated problems.

Key words: digital linguistics, corpus linguistics, Tony McEnery, Nick Ostler, retrospective article

Abstrakt: W tekście referuję i komentuję opracowanie sprzed dwóch dekad. Przypominam w nim artykuł ważny dla językoznawstwa korpusowego – autorstwa Tony'ego McEnery'ego i Nicka Ostlera z 2000 roku. W swoim artykule postulują oni potrzebę wyjścia poza dotychczasowe działania lingwistyki korpusowej na szerszym zestawie języków. Perspektywa czasowa pozwala na obiektywne, zdystansowane komentarze. Wyzwanie związane z budowaniem wielojęzycznego korpusu pozostaje aktualne; owszem, rozwój językoznawstwa korpusowego nadal zależy od możliwości technicznych, jednak dziś widzimy także znaczące zmiany w dziedzinie lingwistyki cyfrowej. W artykule nie tylko prezentuję refleksje brytyjskich lingwistów, ale także pokazuję aktualność/nieaktualność wymienianych przez nich problemów.

Słowa kluczowe: lingwistyka cyfrowa, językoznawstwo korpusowe, Tony McEnery, Nick Ostler, artykuł retrospektywny

Mija właśnie dwadzieścia lat od ukazania się jednego z istotniejszych dla humanistyki cyfrowej tekstu poświęconego gromadzeniu, przetwarzaniu i wykorzystaniu informacji językowej (a pośrednio – językoznawczej) – w 2000 roku Tony McEnery i Nick Ostler, lingwiści z Foundation for Endangered Languages, na łamach prestiżowego czasopisma "Article in Literary and Linguistic Computing" (No 15, s. 403–420) opublikowali artykuł *A New Agenda for Corpus Linguistics – Working with All of the World's Languages*. Chociaż sam tytuł wspomnianego opracowania zawiera odniesienie do „lingwistyki korpusowej", celowo wcześniej użyłem określenia „humanistyka cyfrowa", ponieważ w świetle założeń przedstawionych w tekście brytyjskich językoznawców znajdziemy wiele uwag, które

z jednej strony dotyczą cyfryzacji humanistyki (zob. m.in. Berry, 2012: 1–20; Bomba, 2013: 57–71), z drugiej natomiast – wskazują możliwości wykorzystania korpusów językowych w badaniu grup kultur i społeczeństw (mniejszych bądź większych wspólnot komunikatywnych). W samej jedynie perspektywie językoznawczej artykuł trzeba potraktować nie tylko jako manifest lingwistyki korpusowej, ale także jako apel skierowany do środowisk językoznawców o zintensyfikowanie i poszerzenie działań z zakresu lingwistyki korpusowej oraz ukierunkowanie ich na budowanie korpusów dotyczących różnych języków. Słowo *język* wyróżniam, by zaznaczyć, że mam na myśli termin, który w kontekście artykułu ma sens dość szeroki – otóż McEnery'emu i Ostlerowi nie chodzi wyłącznie o język ogólny, ale również o odmiany środowiskowe i funkcjonalne języka.

Autorzy omawianego artykułu, pisząc o potrzebie poszerzenia zasobów korpusowych poszczególnych języków, podkreślają, że o ile przedsięwzięcia takie mają niepodważalnie dużą wartość dla rozwoju nauki oraz posiadają istotne zalety społeczne i historyczne, o tyle największą trudnością w ich realizacji są ograniczenia techniczne. Językoznawcy nie poprzestają jednak na przedstawieniu tej opinii (z dzisiejszej perspektywy nieco anachronicznej, przybierającej charakter truizmu), ale prezentują także potencjalne korzyści dla lingwistyki płynące z gromadzenia danych korpusowych.

Przez długi czas językoznawstwem korpusowym zajmowali się przeważnie badacze brytyjscy i – co oczywiste – skupiali się oni na języku angielskim. Niemniej jednak lingwistyka korpusowa u swoich źródeł ma charakter wielojęzyczny. Otóż jeszcze przed pojawieniem się technologii komputerowej utrwalano ginące języki w postaci baz papierowych, natomiast w drugiej połowie XX wieku, kiedy w nauce wyłoniła się informatyka humanistyczna (mechanolingwistyka), wiele danych językowych zostało przekonwertowanych z wersji papierowej na formę elektroniczną (chodzi tutaj m.in. o „dygitalizację" tekstów łacińskich, aramejskich i nabatejskich, jakiej podjął się Roberto Busa (1980: 83–90), ale także o prace Alphonse'a Juillanda i Eugenia Chang-Rodrigueza nad maszynowym odczytywaniem tekstów chińskich, hiszpańskich, francuskich i rumuńskich (1964; 1965)). Wspomniane tutaj kwestie McEnery i Ostler jedynie sygnalizują, nie poświęcając im większej uwagi prawdopodobnie dlatego, że mają one wyłącznie dawać pewne tło dla dalszych, kluczowych problemów artykułu i niejako przy okazji argumentować sens kontynuacji działań podjętych kilkadziesiąt lat wcześniej (zresztą szeroką prezentację historii badań korpusowych zawiera opublikowana w 1996 r. i wznowiona kilka miesięcy po ukazaniu się tekstu *A New Agenda for Corpus Linguistics* współautorska książka McEnery'ego i Adrew Wilsona *Corpus Linguistics. An Introduction*). Zwracają jednak uwagę na istotną sprawę – według nich dobry punkt oparcia dla rozwoju lingwistyki cyfrowej tworzą osiągnięcia językoznawców amerykańskich, które – choć w dużej mierze zniszczone przez rewolucję metodologiczną towarzyszącą publikacjom prac Noama Chomsky'ego – przetrwały w Europie Północnej, zwłaszcza w Wielkiej Brytanii, dzięki dużemu zainteresowaniu językoznawców przyswajaniem (akwizycją, nauczaniem) języka z wykorzystaniem danych korpusowych.

Rzecz jasna, rzetelność naukowa wymagała od McEnery'ego i Ostlera wskazania nie tylko walorów budowania korpusów językowych, ale także wyzwań związanych z tworzeniem takich zbiorów. Jeśli spojrzymy na postęp, jaki miał miejsce w ostatnich dziesięcioleciach w dziedzinie technologii informacyjno-komputerowej, to jest oczywiste, że w momencie powstawania tekstu badacze mieli powody do utyskiwań na temat technicznych/technolo-

gicznych uwarunkowań projektowanych działań językoznawczych. W zasadzie przeszkody te już nie istnieją. Aktualne pozostają natomiast inne wyzwania, natury merytorycznej, bardzo często sygnalizowane również na gruncie polskiej nauki. Otóż także dziś najważniejszym wyzwaniem dla autorów korpusów językowych pozostaje brak jednolitego sposobu budowania baz, zwłaszcza doboru tekstów i konstruowania parametrów wyszukiwania materiału językowego. Niestety, z reguły korpusy powstają w ramach innych projektów badawczych lub są ukierunkowane na konkretny cel. Tak więc nadal ścierają się teorie dotyczące gromadzenia danych: po pierwsze, dotyczą one wielkości (ilości) materiałów (Czy korpus powinien być „otwarty", systematycznie uzupełniany, a może należy założyć skończoną liczbę słów i po osiągnięciu zaplanowanej wielkości należy zakończyć dodawanie materiału?); po drugie, wiążą się z jakością i reprezentatywnością danych (Czy korpus ma być zbiorem autentycznej, najbardziej naturalnej komunikacji ludzi, czy też powinien zawierać notacje normatywne i „standardowe"? Czy powinien zawierać dane dotyczące odmian regionalnych, terytorialnych i socjolektalnych? Jeśli tak, to w jakich proporcjach?); po trzecie, odnoszą się do formatu danych, ich przetwarzania i – co za tym idzie – narzędzi ich wyszukiwania i prezentowania (Czy, a jeśli tak, to w jakich zakresach, według jakich kryteriów, możliwe jest maszynowe przetwarzanie, selekcjonowanie danych?).

W kontekście wspomnianych zagadnień wypada osadzić ideę tworzenia korpusu wielojęzycznego. Po upływie dwóch dekad widać w tym zakresie niemały postęp, jeśli chodzi o korpusy dotyczące kilku języków (powstaje coraz więcej korpusów zarówno paralelnych, jak i porównywalnych[1]). Niestety, przypadki te w mniejszym stopniu dotyczą polskiego językoznawstwa – na rodzimym gruncie pomysł brytyjskich językoznawców wciąż jest w fazie wstępnej. Inna sprawa, że w skali światowej wspomniany postęp nie jest aż tak znaczący, co zapewne wynika z faktu, że nadal nie istnieje jeden wspólny model budowy korpusów dla poszczególnych języków świata. Co więcej, dalekie od jednolitości bywają korpusy odnoszące się do konkretnych języków, bo też – jak wspomniano – koncepcje budowania baz wciąż podporządkowane są raczej wąskim celom, wyznaczanym przez obszar i okres badań. Tymczasem zaproponowany przez McEnery'ego i Ostlera pomysł opracowania korpusów obejmujących większą liczbę języków świata, opiera się na założeniu, że korpusy powinny być gromadzone w taki sposób, aby umożliwiały przyjęcie zmian metodologicznych w badaniach nad poszczególnymi językami, w tłumaczeniach (zob. m.in. Kredens, 2005: 270–279; Grabowski, 2011: 89–112; Szela, 2016: 210–226) i w nauce języków obcych. Owszem, brytyjscy językoznawcy zdają sobie sprawę z faktu, że zawsze będą istnieć dostosowania charakterystyczne dla jakiegoś języka i posługującej się nim grupy społecznej, ale według nich twórcy korpusów powinni zmierzać do minimalizowania takich ograniczeń.

[1] Wprawdzie McEnery i Ostler w swoim artykule nie wprowadzają rozróżnienia na wspomniane rodzaje korpusów (nie posługują się więc samymi terminami *parallel corpus* i *comparable corpus*), niemniej jednak pierwszy z autorów wespół z innymi specjalistami wyjaśnia koncepcje obu typów korpusów: korpus porównywalny jest dwu- lub wielojęzyczny i składa z tekstów, które nie są własnymi translatami (teksty są dobrane według jednakowych parametrów stylistycznych, gatunkowych, tematycznych, chronologicznych itd.), z kolei korpus paralelny składa się z tekstów oryginalnych oraz ich przekładów na jeden lub więcej języków. Zob. McEnery, Xiao, Tono, 2006.

Widać dzisiaj aktualność także innych utrudnień, o których wspominają McEnery i Ostler. Nadal niewielki procent języków świata to języki urzędowe; tylko w samej Europie mamy sporo takich sytuacji, bo na przykład w Norwegii językiem urzędowym jest język norweski, ale lokalnie także języki lapońskie, w Hiszpanii – w skali kraju język kastylijski, a w skali regionu także kataloński czy baskijski, w Rosji – rosyjski, natomiast lokalnie między innymi ałtajski, czeczeński, czukocki, jakucki, kałmucki, komi czy tatarski[2]. Mało tego, nadal nie jest rozstrzygnięty status wielu dialektów (języków regionalnych), jak na przykład walijski czy kaszubski oraz śląski. Ponadto istnieją języki, którymi posługuje się duża liczba osób, ale języki te nie mają oficjalnego statusu, jak na przykład język Sylheti w Bangladeszu; i odwrotnie: istnieją języki narodowe z niewielką liczbą użytkowników, jak islandzki. Inne wyzwania: Jak potraktować między innymi odmiany języka chińskiego istniejące wyłącznie w wersji mówionej? Jaką pozycję nadać sposobom komunikacji wizualno-przestrzennej, zwłaszcza językom migowym (przecież występują między nimi spore różnice, a szacuje się, że istnieje ich na świecie nawet 300 i nie zawsze posługiwanie się nimi pokrywa się z granicami państw: mamy na przykład PJM [polski język migowy] obowiązujący w Polsce[3], ale już BSL [brytyjski język migowy] jest zupełnie różny od ASL [amerykańskiego języka migowego], używanego nie tylko w Stanach Zjednoczonych, ale również w Meksyku czy Kanadzie)? Oczywiście, podniesione tu kwestie mają znaczący wpływ na sposób budowania i wykorzystywania korpusów językowych. Niestety, problemy podjęte przez brytyjskich językoznawców w znacznym zakresie nadal nie zostały rozwiązane.

McEnery i Ostler zwracają również uwagę na ekscerpcję tekstów mających stanowić korpusy językowe. Entuzjastycznie wyrazili się na temat Internetu, wskazując go jako dobre źródło, z którego można czerpać niemal nieograniczoną liczbę egzemplifikacji. Z dzisiejszej perspektywy koncepcja ta wydaje się nie dość przekonująca – współczesność nie pozostawia wątpliwości: tworząc korpusy, nierzadko do zbiorów notacji włącza się wypowiedzi użytkowników Internetu, a przecież dzisiaj w Sieci można znaleźć wszystko i zarazem nic (ogrom form językowych zadziwia i zaskakuje jednocześnie, a sam fakt ich obecności nie może konstytuować faktu językowego i pozwalać na uogólnienia; zresztą, znane są praktyki, że jeśli czegoś w Sieci nie ma, a jest komuś potrzebne na potwierdzenie jakiejś tezy, zawsze można to coś zamieścić), a więc z konstruowaniem korpusu wiązałaby się żmudna praca oddzielania ziarna od plew (zob. uwagi zamieszczone w: Loewe, 2006: 93–103; Podhajecka, 2006: 338–347; Piotrowski, Grabowski, 2013: 59–71; Zabawa, 2019: 211–232). Owszem, Sieć oferuje również wiele innej wartości danych – zdygitalizowane teksty literackie, naukowe czy publicystyczne, które mogą być kanwą korpusu językowego, w większym stopniu

[2] Zob. m.in.: *Urzędowy wykaz nazw państw i terytoriów niesamodzielnych* (Komisja Standaryzacji Nazw Geograficznych poza Granicami Rzeczypospolitej Polskiej, wyd. 4, 2017); TNS Opinion & Social: *Europeans and their Languages* [online: https://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_243_en.pdf; data dostępu: 12.11.2019]; *О языках народов Российской Федерации* (с изменениями и дополнениями) Закон РФ от 25 октября 1991 г. N 1807-I; Zob. Footitt, Kelly, 2012.

[3] Na marginesie warto dodać, że w rodzimej przestrzeni istnieje już *Korpus Polskiego Języka Migowego* i oparty na nim *Korpusowy słownik polskiego języka migowego* – dzieła opracowane przez Pracownię Lingwistyki Migowej Uniwersytetu Warszawskiego w ramach realizowanego od 2011 roku projektu badawczego „Ikoniczność w gramatyce i leksyce polskiego języka migowego (PJM)" [online: http://www.slownikpjm.uw.edu.pl/; data dostępu: 12.11.2019].

zapewniając zasadę równowagi i reprezentatywności. Dane takie mają także tę zaletę, że procedura ich zbierania jest i tania, i stosunkowo prosta, podobnie jak łatwe jest ich maszynowe odczytywanie.

W założeniu autora szkic niniejszy nie ma być „rozliczeniem" koncepcji McEnery'ego i Ostlera ogłoszonej przed dwiema dekadami, ale jedynie przypomnieniem ich apelu, na który w wielu aspektach środowiska językoznawców odpowiedziały z entuzjazmem i – co więcej – ideę brytyjskich lingwistów wciąż realizują. Gwoli ścisłości wypada jednak zaznaczyć, że choć w swoim tekście McEnery i Ostler przywołują przykłady istnienia w świecie zachodnim pewnych podwalin dla realizacji ich pomysłu, w 2000 roku mających już nawet kilkadziesiąt lat, to nie powinno się zapominać, że również na gruncie rodzimego językoznawstwa – choć nieco później i na mniejszą skalę – w drugiej połowie XX wieku były podejmowane działania w zakresie inżynierii językoznawczej (dość wspomnieć tutaj o zapoczątkowanej w latach 70. na Uniwersytecie Warszawskim współpracy informatyków i lingwistów) (zob. np.: Lewandowska-Tomaszczyk, 2005; Świdziński, 2006: 23–34), a obecnie możemy się poszczycić między innymi wieloma korpusami równoległymi (zob. artykuły w: Gruszczyńska, Leńko-Szymańska, 2016) i narzędziami służącymi automatycznemu przetwarzaniu wielkich korpusów tekstów polskich (Szafran, 1997: 51–64; Rudolf, 2004; Rudolf, Świdziński, 2006: 31–43). Niemniej jednak do osiągnięcia celu wyznaczonego przez naukowców z Foundation for Endangered Languages, a więc zrealizowania przedstawionej przez nich koncepcji budowania wielojęzycznych korpusów na podstawie jednolitych zasad, jeszcze dość daleko. Dlatego też tekst Ostlera i McEnery'ego, mimo że ma charakter archiwalny, może i powinien być przypomniany jako zachęta i inspiracja do zintensyfikowania działań w tym zakresie.

Literatura

Berry D.M., 2012: *Introduction: Understanding the Digital Humanities*. In: Berry D., ed.: *Understanding Digital Humanities*. London, s. 1–20.

Bomba R., 2013: *Narzędzia cyfrowe jako wyznacznik nowego paradygmatu badań humanistycznych*. W: Radomski A., Bomba R., red.: *Zwrot cyfrowy w humanistyce*. Lublin, s. 57–71.

Busa R., 1980: *The Annals of Humanities Computing. The Index Thomisticus*. „Computers and the Humanities" XIV, nr 2, s. 83–90.

Footitt H., Kelly M., 2012: *Languages at War. Policies and Practices of Language Contacts in Conflict*. London.

Grabowski Ł., 2011: *Korpusy dwu- i wielojęzyczne w służbie tłumacza, leksykografa i badacza: poszukiwanie ekwiwalentów przekładowych w świetle hipotez dotyczących istnienia uniwersaliów tłumaczeniowych*. W: Chlebda W., red.: *Na tropach translatów. W poszukiwaniu odpowiedników przekładowych*. Opole, s. 89–112.

Ikoniczność w gramatyce i leksyce polskiego języka migowego (PJM) [online: http://www.slownikpjm.uw.edu.pl/; data dostępu: 12.11.2019].

Juilland A., Chang-Rodriguez E, 1964: *Frequency Dictionary of Spanish Words*. The Hague.

Juilland A., Chang-Rodriguez E., 1965: *Frequency Dictionary of Rumanian Words*. The Hague.

Kredens K., 2005: *Korpusy językowe w językoznawstwie sądowym*. W: Lewandowska-Tomaszczyk B., red.: *Podstawy językoznawstwa korpusowego*. Łódź, s. 270–279.

Lewandowska-Tomaszczyk R., red., 2005: *Podstawy językoznawstwa korpusowego*. Łódź.

Loewe I., 2006: *Internet i jego zasoby w polskich badaniach lingwistycznych. Rekonesans*. „Biuletyn Polskiego Towarzystwa Językoznawczego" LXII, s. 93–103.

McEnery T., Wilson A., 1996: *Corpus Linguistics*. Edinburgh.

McEnery T., Xiao R., Tono Y., 2006: *Corpus-Based Language Studies. An Advanced Resource Book*. London–New York 2006.

Piotrowski T., Grabowski Ł., 2013: *Interpretacja danych frekwencyjnych z korpusów językowych: opis pewnych problemów (na kilku przykładach z życia wziętych)*. W: Chlebda W., red.: *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*. Opole, s. 59–71.

Podhajecka M., 2006: *Kilka uwag o wykorzystaniu zasobów internetowych do analiz korpusowych języka*. „Język Polski" LXXXVI, z. 5, s. 338–347.

Rudolf M., 2004: *Metody automatycznej analizy korpusu tekstów polskich*. Warszawa.

Szafran K., 1997: *Automatyczne hasłowanie tekstu polskiego*. „Polonica" XVIII, s. 51–64.

Szela M., 2016: *O wykorzystaniu Angielsko-Polskiego Korpusu Równoległego Tekstów Prawnych w badaniu cech języka tekstów tłumaczonych*. W: Gruszczyńska E., Leńko-Szymańska A., red.: *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora*. Warszawa, s. 210–226.

Świdziński M., 2006: *Lingwistyka korpusowa w Polsce – źródła, stan, perspektywy*. „LingVaria" I, nr 1, s. 23–34.

Świdziński M., Rudolf M., 2006: *Narzędzia informatyczne obsługi wielkich korpusów tekstów: wyszukiwarka Holmes*. „Biuletyn Polskiego Towarzystwa Językoznawczego" LXI, s. 31–43.

TNS Opinion & Social: *Europeans and their Languages* [online: https://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_243_en.pdf; data dostępu: 12.11.2019].

Zabawa M., 2019: *O roli samodzielnie przygotowanych korpusów w badaniach językoznawczych (na przykładzie korpusu wykorzystującego zasoby internetowe)*. „Półrocznik Językoznawczy Tertium. Tertium Linguistic Journal" IV, nr 1, s. 211–232.

О языках народов Российской Федерации (с изменениями и дополнениями), Закон РФ от 25 октября 1991 г. N 1807-I.