O różnych korpusowych metodach badawczych – próba krytycznej refleksji

Article · December 2016

CITATIONS
2

READS
1,167

2 authors, including:

Milena Hebal-Jezierska
University of Warsaw
24 PUBLICATIONS **29** CITATIONS

SEE PROFILE

KOMUNIKACJA SPECJALISTYCZNA

TOM 11.

Warszawa 2016

Recenzenci	-	prof. dr hab. Elżbieta Tabakowska-Muskat prof. dr hab. Ewa Wolnicz-Pawłowska dr hab. Roman Hajczuk dr hab. Luba Biesiekirska dr hab. Grażyna Mańkowska dr Marek Łukasik dr Włodzimierz Majewski dr Adam Szynal
Redaktor naczelny	-	dr Łukasz Karpiński
Sekretarz redakcji	-	mgr Wojciech Drajerczak
Rada Programowa	-	prof. dr hab. Barbara Z. Kielar prof. dr hab. Łarisa I. Korniejewa (Ekaterynburg) prof. dr hab. Jerzy Lukszyn prof. dr hab. Zenon Weigt prof. dr hab. Wanda Zmarzer prof. dr hab. Aleksandr W. Zubow (Mińsk) dr hab. Václav Cvrček (Praga)
Redaktorzy językowi	-	dr hab. Krzysztof Fordoński dr Dorota Muszyńska-Wolny mgr Marta Ehrenkreutz-Jasińska mgr Oliwia Tortorelo
Redaktorzy tomu		dr Łukasz Karpiński dr Piotr Michałowski
Adres redakcji	-	Instytut Komunikacji Specjalistycznej i Interkulturowej WLS UW 02-678 Warszawa, ul. Szturmowa 4 tel. 22 55 34 200 e-mail: ks.kjs@uw.edu.pl

Opracowanie graficzne	-	Łukasz Karpinski
Druk	-	Zakład Graficzny Uniwersytetu Warszawskiego

ISSN 2080-3532

ŁUKASZ GRABOWSKI
Uniwersytet Opolski, Wydział Filologiczny
MILENA HEBAL-JEZIERSKA
Uniwersytet Warszawski, Wydział Polonistyki (Instytut Slawistyki Zachodniej i Południowej)

O RÓŻNYCH KORPUSOWYCH METODACH BADAWCZYCH – PRÓBA KRYTYCZNEJ REFLEKSJI

1. Wstęp

Pojęcie korpusu znane było już w starożytności. Ewolucję tego terminu przedstawia w swojej publikacji Francis [2000: 169-181][49]. Pierwotnie terminem *korpus* nazywano nielingwistyczne zbiory tekstów, czego przykładem może być *Corpus Iuris Civilis*; powstał w VI wieku z inicjatywy cesarza Justyniana I Wielkiego. Był to zbiór wczesnego prawa rzymskiego, ilustrowany przykładami, wraz z interpretacją nowego prawa, które miało dopiero zacząć obowiązywać.

W VIII wieku *korpus* oznaczał już lingwistyczny zbiór tekstów. Jednym z nich jest łaciński *Corpus Glossary*. Zgromadzono w nim ułożone alfabetycznie trudne słowa, po których następowały synonimy bardziej zrozumiałe oraz ich odpowiedniki staroangielskie, głównie z dialektów używanych w Mercji i Kent [Kuhn 1939]. Następny etap rozwoju korpusów jest związany z nowymi metodami leksykograficznymi, jakie pojawiły się w osiemnastowiecznej Anglii. Zamiast korzystać z wcześniejszych słowników, próbowano gromadzić materiał do celów leksykograficznych, wykorzystując wydane druki lub przeprowadzając ankiety. Kompilacja korpusu była zatem etapem przygotowawczym do tworzenia słowników. Oto w jaki sposób tworzył go Samuel Johnson:

> "Do tej trudnej pracy wynajął dom na Gough Square, gdzie w jednym pomieszczeniu ustawił stoły i inne rzeczy dla pomocników, których było

[49] W przytoczonym artykule Francis zajmuje się wyłącznie korpusami języka angielskiego.

> pięciu lub sześciu, których miał pod kontrolą. Do kopii słownika Baileya wkładał artykuły, które gromadził, czytając twórczość najlepszych pisarzy angielskich. [...] Wybrane przez siebie wyrazy przekazywał swoim pomocnikom, żeby je wkładali na swoje miejsca." [Francis 2000: 171]

Powstały w ten sposób słownik wywołał kontrowersje w środowisku lingwistycznym; wytknięto mu m.in. niedostateczny zasób leksykalny. W rezultacie opracowano projekt nowego słownika. W tym wypadku korpus miał zawierać każde słowo istniejące od 1000 r., wszystkie jego formy, sposoby wymowy i znaczenia, zarówno historyczne, jak i współczesne. Zwrócono uwagę także na reprezentatywność wyboru literatury. Zbieraniem materiału zajmowały się tysiące ochotników. Porządkował je James A. H. Murray, stąd nazwa *Murray's Corpus*. Bardziej profesjonalny korpus stworzono w Stanach Zjednoczonych na potrzeby słownika *Webster's New International Dictionary* [wyd. trzecie 1961]. Wykorzystał on doświadczenie Murraya, ale tu twórcami byli filolodzy. Był to ostatni słownik powstały przy pomocy korpusu nieskomputeryzowanego. [Francis 2000: 171-173].

Korpusy językowe w znaczeniu zbiorów tekstów zebranych według wcześniej ustalonych kryteriów, a następnie wykorzystanych do celów badawczych były znane nie tylko badaczom zajmującym się angielszczyzną. Warto tu także przywołać badania Kaedinga [1897] nad konwencjami niemieckiej ortografii, gdzie wykorzystano zbiór tekstów z 11 milionami wyrazów tekstowych, a także badania nad językiem pamiętników pisanych przez dzieci [Preyer 1889]; oba przykłady przytaczają McEnery i Wilson [1996: 10].

Następnym etapem w rozwoju myśli lingwistyki korpusowej jest epoka strukturalizmu w Stanach Zjednoczonych. Dla niektórych lingwistów, np. dla Harrego i Hilla, korpus przedstawiał dostatecznie dużą liczbę danych językowych, osadzonych w naturalnym kontekście, przy czym intuicja lingwistyczna została zepchnięta na dalsze miejsce, a nieraz nawet odrzucana [Leech 2000: 39]. Dopiero reakcja Chomsky'ego zmieniła kierunek rozwoju lingwistyki korpusowej. Stwierdził on bowiem, że każdy naturalny korpus jest niekompletny. Niektóre zdania w nim nie wystąpią, bo są oczywiste lub niepoprawne, lub niestosowane. Powyższa wypowiedź rozpoczęła nową erę. Już po dwóch latach rozpoczęli działalność założyciele nowej szkoły lingwistyki korpusowej. Randolph Quirk ogłosił pod nazwą *Survey of English Usage* (SEU) swój projekt stworzenia korpusu mówionej i pisanej angielszczyzny. Krótko po tym, Nelson Francis i Henry Kučera stworzyli grupę, dzięki której powstał pierwszy korpus, szerzej znany jako *Brown Corpus*, składający się z miliona wyrazów zebranych w tekstach reprezentujących amerykańską prozę i prasę tamtego okresu [Francis i Kučera 1964].

Z prawdziwą ekspansją korpusów językowych mamy do czynienia wraz z nadejściem tzw. rewolucji informatycznej w latach 80. XX wieku, kiedy to rozpoczęto prace nad Brytyjskim Korpusem Narodowym (BNC), aż do chwili obecnej, w której to można cieszyć się dostępem przez Internet do dużych korpusów narodowych (np. NKJP, CNK, NKJR), a także do specjalistycznego oprogramowania (np. wyszukiwarek, konkordanserów, kolokatorów). Korzystając z zaawansowanych narzędzi informatycznych używanych do analizy tekstu, badacze coraz częściej opracowują własne korpusy językowe na potrzeby prowadzonych badań. Ze względu na ograniczenia związane z prawami autorskimi wiele takich korpusów nie jest udostępnianych publicznie.

Warto zauważyć, że rozwój lingwistyki korpusowej jest przede wszystkim uzależniony od wyboru narzędzi badawczych, a te są ściśle powiązane z rozwojem badań nad przetwarzaniem języka naturalnego (NLP). Czeski Korpus Narodowy wydaje się być jednym z modelowych przykładów wykorzystywania najnowszych technologii oraz badań z dziedziny NLP. Dla pierwszych korpusów języka czeskiego (udostępnionych przez Instytut Czeskiego Korpusu Narodowego już w 2000 r.) opracowano wielofunkcyjną wyszukiwarkę, pozwalającą badaczowi na prowadzenie wielowymiarowych badań. Niemniej jednak, jak tylko jego twórcy znaleźli informację o lepszych testach eksplorujących kolokacje niż informacja wzajemna (MI), niezwłocznie wprowadzili do swoich wyszukiwarek inne testy, m.in. log Dice, loglog. Na podkreślenie zasługuje także poprawienie jakości tagowania morfologicznego, które także jest związane z rozwojem badań nad przetwarzaniem języka. Różnica jakości pomiędzy znakowaniem pierwszych udostępnionych korpusów a następnymi jest w przypadku znacząca. Warto dodać, że taki rozwój jest możliwy dzięki współpracy z innymi ośrodkami (Instytut Lingwistyki Formalnej i Lingwistyki Stosowanej - zespół Jana Hajiča oraz Instytut Lingwistyki Teoretycznej i Komputerowej - zespół Vladimira Petkeviča oraz Alexandra Rosena), zajmującymi się teorią języka i automatyzacją danych. Najnowsze korpusy języka mówionego zintegrowane z Czeskim Korpusie Narodowym wykorzystują natomiast najnowsze badania nad przetwarzaniem mowy. Korpusy języków mówionych (Oral 2013) pozwalają na odsłuchanie określonych fragmentów. Innym przykładem może być program KWords, pozwalający na wyszukanie wyrazów kluczowych z wykorzystaniem różnych metod pomiaru kluczowości (*keyness*). Obecnie jest on nie najlepszej jakości. W miarę możliwości jednak ma zostać ulepszony, a docelowo nawet wbudowany w wyszukiwarki.

Niemniej jednak, należy postawić pytanie, czy wraz z rozwojem korpusów i badań językowych oraz rosnącą ich popularnością istnieje świadomość możliwości i ograniczeń, które kryją się za badaniem języka na

tzw. sposób korpusowy. W krajach, w których lingwistyka korpusowa rozwija się od wielu lat, wraz z budową i rozwojem korpusów zwiększa się świadomość poprawnego z nich korzystania. W Polsce w wielu przypadkach tego brakuje.

2. Korpusowe metody badawcze

W niniejszej sekcji zwięźle zostanie przybliżona specyfika językoznawstwa korpusowego, zwłaszcza jego istota oraz metody badawcze.

Językoznawcy traktują korpusy jako eksperymentalną bazę materiałową, służącą do weryfikacji istniejących teorii językoznawczych lub opracowania nowych hipotez badawczych. Językoznawcy korpusowi koncentrują się przede wszystkim na częstości występowania jednostek językowych i ich klas. W tym podejściu zamiast przykładów zdań lub wypowiedzeń stworzonych specjalnie na potrzeby potwierdzenia lub odrzucenia danej hipotezy, badacze wykorzystują autentyczne teksty powstałe w rzeczywistych sytuacjach komunikacyjnych.

W związku z tym, że korpusy językowe stanowią materiał ilustrujący zastosowanie języka w różnorodnych sytuacjach komunikacyjnych, traktowane są jako zasoby czy też źródła informacji (zbiory danych językowych), które pozwalają na odnalezienie odpowiedzi na najróżniejsze, często nieoczekiwane pytania dotyczące użycia języka. Korpusy pozwalają zatem na wyjście poza ograniczoną intuicję językową rodzimych użytkowników języka. W ten sposób można spojrzeć na język z „lotu ptaka", tj. na dużą liczbę danych językowych. Ilustracją powyższego stwierdzenia mogłoby być zbadanie naraz wszystkich książek przechowywanych w danej bibliotece.

Metodologia badań korpusowych jest wysoce niejednorodna. W literaturze specjalistycznej można znaleźć liczne omówienia różnic między sposobami analizy danych korpusowych [m.in. Clear i in. 1996, McEnery i Wilson 1996, Ooi 1998, Tognini-Bonelli 2001, McEnery i in. 2006, Xiao 2008, McEnery i Gabrielatos 2008, Lee 2008, McEnery i Hardie 2011, Górski 2012]. Spośród nich wyłania się podział na dwa główne podejścia, tj. *corpus-based* i *corpus-driven*, który zaproponowała po raz pierwszy Tognini-Bonelli [2001].

Najważniejsze różnice między nimi dotyczą statusu danych językowych, kierunku analizy, a także roli intuicji samego badacza. W podejściu sterowanym korpusem (*corpus-driven approach*), analizie poddaje się tzw. surowe dane językowe, czyli słowoformy nieujęte w ramy jakiejkolwiek teorii języka. Wszelkie hipotezy badawcze formułuje się właśnie w oparciu o częstość występowania i dystrybucję wyrazów lub ich sekwencji w teks-

tach (np. n-gramów, ram frazowych[50]). W takim indukcyjnym podejściu badacze sięgają do wybranej teorii języka lub proponują nowe konstrukty teoretyczne dopiero po ekstrakcji danych językowych z tekstów.

W podejściu *corpus-based* kierunek postępowania jest odwrotny. Badacz, korzystając z wybranej teorii języka, formułuje hipotezę badawczą, a następnie dokonuje jej weryfikacji w oparciu o dostępne dane językowe. W ten sposób można dokonać weryfikacji różnych konstruktów teoretycznych (np. grupy rzeczownikowej, wyrażenia przyimkowego, przydawki, idiomu), które pojawiły się na gruncie językoznawstwa, zanim korpusy językowe stały się powszechne. Wspomniane różnice wpływają na postrzeganie językoznawstwa korpusowego przez badaczy. Podejście sterowane korpusem uważane jest za autonomiczny paradygmat badań językoznawczych, a dedukcyjne podejście corpus-based traktuje się bardziej jako metodologię badawczą, czyli określony sposób postępowania z danymi językowymi [Fuster-Marquez i Clavel 2010: 54]. Jednak w praktyce wielu badaczy stosuje metody mieszane [Ooi 1998: 52] lub nazywa swoje badania *corpus-based*, nawet jeśli korzystają z podejścia indukcyjnego (*corpus-driven*) [McEnery i Hardie 2011: 151; Górski 2012: 293].

Inne metody to modyfikacje podziału dokonanego przez Tognini-Bonelli [2001] lub propozycje łączenia danej metodologii korpusowej (*corpus-based* lub *corpus-driven*) z innymi procedurami badawczymi. Na przykład Partington [2006] w swoim badaniu łączy podejście *corpus-driven* z metodami niekorpusowymi (ankietyzacja), nazywając swój sposób postępowania *corpus-assisted approach*.

W literaturze specjalistycznej można się także spotkać z takimi terminami, jak *corpus-induced approach* [Lee 2008: 88] oraz *corpus-illustrated* / *corpus-informed approach* [Lee 2008:88, Górski 2012: 292]. Pierwszy z nich, tj. *corpus-induced approach*, polega na wykorzystaniu danych korpusowych do szeroko zakrojonych analiz ilościowych, najczęściej automatycznych, których wyniki można wykorzystać do różnych celów praktycznych i komercyjnych [Lee 2008: 88]. Natomiast drugi z użytych tu terminów, tj. *corpus-illustrated* / *corpus-induced*, wydaje się być częścią wyżej opisanej metody *corpus-based*, której jednym z zadań jest egzemplifikacja. Innymi słowy, podejście to polega na przytaczaniu przykładów użycia języka na podstawie danych korpusowych (zamiast tworzenia przykładów sposobem *ad hoc,* tj. specjalnie na potrzeby potwierdzenia lub odrzucenia danej hipotezy).

Warto zauważyć, że w Polsce w przeważającej większości językoznawcy określają swoje badania jako *korpusowe*. Nie wiadomo zatem dokładnie, którą z metod stosują.

Każdy z opisanych powyżej sposobów postępowania z danymi korpusowymi niesie za sobą zarówno możliwości, jak i ograniczenia. W kolejnej

[50] N-gramy to sekwencje n wyrazów (np. *I think that*); ramy frazowe (phrase frames) to zbiory wariantów n-gramów różniących się jednym wyrazem wchodzącym w ich skład (np. *it is * to*) [Fletcher 2002-2007].

sekcji spróbujemy je zilustrować wybranymi studiami przypadku powstałymi na podstawie pracy z korpusem języka polskiego (NKJP), czeskiego (CNK), słowackiego, rosyjskiego (NKJR) i angielskiego (BNC).

3. Na co należy uważać podczas prowadzenia badań korpusowych?

Ze względu na ograniczoną objętość artykułu, w niniejszym tekście chcielibyśmy skupić się na kilku wybranych aspektach pracy z korpusami elektronicznymi. Opisywane problemy dotyczyć będą poprawności anotacji danych korpusowych, szczegółowości tagsetu, braku poświadczeń danego zjawiska w korpusie, precyzyjnego określania zapytań, dużej liczby poświadczeń w korpusie, ograniczeń wyszukiwarek korpusowych i programów konkordacyjnych, a także celowości stosowania testów istotności statystycznej przy porównywaniu wyników przeszukiwania różnych korpusów.

W przypadku podejścia *corpus-based* badacze często korzystają z korpusów anotowanych[51], tj. takich, w których teksty wzbogacono o dodatkowe informacje metajęzykowe (opisujące różne cechy gramatyczne), metatekstowe (opisuję mikro- i makrostrukturę tekstów) oraz metasytuacyjne (opisujące źródło danych, a także czas i okoliczności powstania tekstu) [Piotrowski 2003, Lewandowska-Tomaszczyk 2005]. Ze względu na dużą liczbę danych językowych większość anotacji, zwłaszcza metajęzykowej, dokonywana jest w sposób automatyczny przy pomocy odpowiednich narzędzi służących do znakowania tekstów (np. tagerów morfosyntaktycznych). Niemniej jednak automatyczny sposób anotacji nie jest sposobem niezawodnym. Oznacza to, że jeśli poprawność anotacji 100-milionowego korpusu oscyluje na poziomie 95% to 5 milionów wyrazów tekstowych jest w nim opisane błędnie.[52]

Przejdźmy zatem do konkretnych przykładów. To co wydaje się szczególnie niebezpieczne z punktu widzenia analizy korpusowej to oznaczenie morfosyntaktyczne istniejących, choć nienormatywnych, form językowych, np. *przyszłem*. W korpusie NKJP[53] jest ona oznakowana jako nieznana (*ign*), co oznacza, że w przypadku wyszukiwania według tagów nie zostanie odnaleziona. To pociąga za sobą także brak przypisanego lematu.

a skąd coś ty ***przyszłem*** *[przyszłem:ign] .. nie jest ?*

[51] Oczywiście prowadzone są również liczne badania, w których badacze, stosując podejście *corpus-based*, korzystają z korpusów nieanotowanych.

[52] Bardziej szczegółowe omówienie konkretnych problemów związanych z błędną adnotacją można znaleźć w pracach Hebal-Jezierskiej [2013: 22-23], czy też Piotrowskiego i Grabowskiego [2013: 65-70]. W artykule Piotrowskiego i Grabowskiego [2013: 62-65] znajdziemy również szczegółowy opis problemów związanych z normalizacją wyników wyszukiwania do wspólnej podstawy (najczęściej jest nią 1 milion wyrazów tekstowych).

[53] W poniższych przykładach wykorzystano wyszukiwarkę korpusową Poliqarp.

*ale dopuszczalna jest forma **przyszłem** [przyszłem:ign] słuchaj i ja tak słucham*

Formy nienormatywne w NKJP są dość często oznaczane tagiem ign. Kolejnym przykładem może być forma włanczać:

*ano trzeba się **wyłanczać** [wyłanczać:ign] robisz swoje i gra muzyka*

Podobna sytuacja miała miejsce na początku istnienia Czeskiego Korpusu Narodowego (CNK). Badając końcówki fleksyjne leksemów zakończonych na *-eta*, *-ota*, można było dojść do wniosku przy wyszukiwaniu za pomocą tagów, że końcówka *-i* nie istnieje. W tagsecie bowiem końcówka ta była oznaczona jako nieznana. Dzięki poprawie systemu tagowania oraz możliwości zgłaszania błędów wyeliminowano zarówno tę, jak i wiele innych pomyłek.

W przypadku języka czeskiego, w którym mamy dwie normy (literacką i nieliteracką), interpretacja dowodu negatywnego wydaje się jeszcze bardziej skomplikowana. Za przykład mogą nam posłużyć chociażby przytoczone w dalszej części artykułu czeskie form narzędnikowe *kluky*/*klukama*. W korpusie tekstów mówionych, ORAL2016, pierwsza z wymienionych form nie występuje. Interpretacja dowodu negatywnego w tym wypadku jako formy nieistniejącej w języku czeskim byłaby błędna.

Kolejnym problemem może być nadmierna szczegółowość tagsetu. Wydawałoby się, że dokładniejszy tagset ułatwia pracę badacza korpusowego. Z naszych doświadczeń wynika jednak, że zbytnia szczegółowość tagsetu może znacząco utrudnić ekscerpcję danych. Innymi słowy – im więcej kategorii gramatycznych rozróżnianych w tagsecie, tym łatwiej można popełnić błąd w wyszukiwaniu, które jest wstępem do dalszych badań. Do tego dochodzą potencjalne błędy w anotacji, których nie da się zupełnie wyeliminować nawet przy zastosowaniu nowoczesnych metod anotacyjnych. Błędy w anotacji są nieuniknione w przypadku w pełni automatycznego tagowania tekstów. Na przykład w Narodowym Korpusie Języka Rosyjskiego (NKJR), gdzie do automatycznego znakowania semantycznego zastosowano bardzo szczegółowy tagset, przypisuje się wyrazom tekstowym wiele potencjalnych cech semantycznych, które nie realizują się w danym kontekście. W poniższych przykładach przymiotnik *Волынский* (‘Wołyński’), składający się na toponim *Владимир Волынский* (‘Włodzimierz Wołyński[54]’), poprawnie otagowano jako nazwę własną (tag: r:propn) i toponim (tag: t: topon), ale jednocześnie niepoprawnie jako nazwisko (tag: t: famn) i osobę (t: hum). Temu samemu przymiotnikowi, występującemu jako antroponim *А. П. Волнński* (‘A. P. Wołyński’), również przypisano cechy toponimu. Brak ręcznej weryfikacji tagowania semantycznego, o czym eksplicytnie piszą twórcy korpusu na stronie internetowej[55], sprawia, że użytkownik powinien dokładnie przeanalizować

[54] Miasto na Ukrainie, w obwodzie łuckim.

[55] Więcej można przeczytać na stronie: http://ruscorpora.ru/corpora-sem.html

wystąpienia danego wyrazu tekstowego, aby odnaleźć poświadczenia w poszukiwanym znaczeniu.

> *Если же принять во внимание, что до 1528 года Радзивиловская рукопись все еще не была переплетена, то ее местонахождение в епископальной библиотеке позволяло предполагать, что Владимир* ***Волынский [der:s, dt:topon, r:propn, r:rel, t:famn, t:hum, t:topon]*** *и был местом ее создания в 90-х годах XV века*. [Андрей Никитин. Поиски и находки: Тайны Радзивиловской летописи // «Знание — сила», 2003]

> *Жестокую расправу учинил над ним кабинет-министр А.П.* ***Волынский [der:s, dt:topon, r:propn, r:rel, t:famn, t:hum, t:topon]***. [Элла Кричевская. Литературная жизнь старого Петербурга (2003) // «Вестник США», 2003.06.25]

Na podobne błędy możemy natknąć się zwłaszcza w wypadku występowania synkretyzmów, homonimii lub kategorii granicznych. Takim przykładem może być próba odróżnienia rzeczownika od gerundium. W korpusie NKJP formy gerundialne są lematyzowane formą podstawową czasownika i opisane jako gerundium (tag: ger), natomiast formy rzeczowne są lematyzowane formą podstawową rzeczownika i opisywane jako rzeczowniki (tag: subst). Aby lepiej zilustrować problem przyjrzyjmy się bliżej leksemowi *zachowanie*, który jest otagowany w zależności od kontekstu jako rzeczownik (lemat: *zachowanie*) lub czasownik (lemat: *zachować*). Poniżej podajemy dwa zdania, w których *zachowanie* wydaje się występować jako gerundium, podczas gdy w rzeczywistości tagowanie jest różne.

> *... nie boi się o* ***zachowanie*** *[zachować:ger:sg:acc:n:perf:aff] męża*

> ***zachowanie*** *[zachowanie:subst:sg:acc:n] się Dziewanowskiego przy telefonie wydało.*

Innym przykładem mogą być przymiotniki, które w zależności od funkcji w zdaniu są tagowane jako rzeczowniki lub przymiotniki. Jednak ze względu na pojawiające się błędy to poprawne założenie twórców korpusów bywa w praktyce bardziej przeszkodą niż pomocą dla użytkownika, np.

> *W dowolnie* ***wybranym*** *[wybrany:subst:sg:loc:m1]* ***momencie*** *[moment:subst:sg:loc:m3] ... Nawet teraz*

> *Jesteśmy* ***wybranym*** *[wybrany:subst:sg:loc:m1]* ***narodem*** *[naród:subst:sg:inst:m3]*

Ponadto niektóre z podobnych leksemów mogą być imiesłowami i wtedy są lematyzowane kanoniczną formą czasownika. Takim przykładem jest *znany*, np.

*Ze zmęczonych murów ulatniał się **znany** [znać:ppas:sg:nom:m1:imperf:aff] mi dobrze zapach zmiksowany z [...].*

*którzy nagle wchodzą w dobrze **znany** [znany:adj:sg:acc:m3:pos] film i oglądają go*

*Wyraz specjalistyczny. **Znany** [znany:subst:sg:nom:m1] **specjalistom** [specjalista:subst:pl:dat:m1]*

Przytaczając powyższe przykłady, chcemy wyraźnie podkreślić, że naszym celem nie jest wytykanie błędów twórcom korpusów. Podobne pomyłki, jak opisane powyżej, znajdziemy bowiem w każdym korpusie. Chcemy tylko pokazać, że dobre zamiary twórców korpusów, wyrażające się w tworzeniu jak najszczegółowszego tagsetu nie zawsze przekładają się na jakość pracy z korpusem podczas prowadzenia badań językoznawczych. Dla porównania, tagsety stosowane w CNK lub SNK wydające się na pierwszy rzut oka proste, a nawet prymitywne, są jednak bardziej przejrzyste dla użytkownika (choć klasyczny podział na kategorie gramatyczne, np. części mowy, jest z pewnością niedoskonały w swoim charakterze). Warto jednak w tym miejscu odnotować, że różnorodność i/lub zbytnia szczegółowość tagsetów sprawiają, że należy zachować szczególną ostrożność, zwłaszcza podczas prowadzenia badań porównawczych.

Przy interpretacji dowodu negatywnego należy wziąć pod uwagę przede wszystkim typy i proweniencję tekstów zgromadzonych w korpusie. Na przykład w Narodowym Korpusie Języka Polskiego (NKJP) aż 56% tekstów to publicystyka i informacje prasowe, natomiast w Brytyjskim Korpusie Narodowym (BNC) odsetek tekstów publicystycznych wynosi jedynie 33%; w BNC znajdziemy teksty powstałe w latach 1960-1993, NKJP zawiera teksty powstałe od początku XX wieku, z kolei w Rosyjskim Korpusie Narodowym (NKJR) znajdują się teksty pochodzące nawet z 2. połowy XVIII wieku. Wspomniane różnice wynikają z odmiennych kryteriów reprezentatywności i zbilansowania tekstów [Górski & Łaziński 2012][56], co sprawia, że porównywanie danych zaczerpniętych z różnych korpusów narodowych jest utrudnione.

Z tym wiąże się także problem normy językowej w danych korpusowych. Przykładem mogą być polskie formy *przyszłem* versus *przyszedłem* oraz czeskie formy narzędnikowe *kluky* versus *klukama*. Analizując oba zapytania jako twory niezwiązane z sytuacją językową, otrzymujemy następujące dane. W czeskim korpusie SYN2015 (korpus tekstów pisanych) stosunek literackiej formy *kluky* do nieliterackiej *klukama* wynosi 408 (3.38 ipm[57]) do 331 (2.74 ipm). Bez głębszej analizy i nieznajomości rozwarstwienia językowego[58] można by dojść do błędnych

[56] W przypadku NKJP jako główne kryterium przyjęto strukturę czytelnictwa w Polsce.
[57] Z ang. *instances per million* (words).
[58] W języku czeskim norma języka literackiego znacznie różni się od normy języka ieliterackiego. Dotyczy to nie tylko słownictwa, ale także poziomu fonetycznego i morfologicznego.

wniosków prowadzących do uznania opisywanych form jako równorzędnych czy wariantywnych. Tymczasem formy nieliterackie *klukama* są reprezentacją normy nieliterackiej, które w korpusie często występują w dialogach lub monologach.

> *Tak jsem celý dny popíjel pivo v krámku na rohu, hulil trávu a kecal s* ***klukama*** *ze čtvrti o kravinách.*

Pytanie o formę klasyfikowaną przez polskich normatywistów za błędną, *przyszłem*|*Przyszłem* zwraca 133 poświadczenia (w pełnym korpusie NKJP – 1800M segmentów). Należy wskazać, że w stosunku do formy *przyszedłem*|*Przyszedłem* (ponad 1000 poświadczeń, dokładna liczba nie jest podana) jest jej mniej, ale 133 wystąpienia nie powinny zostać zignorowane. Głębsza analiza pozwoli jednak na weryfikację warstwy językowej oraz odrzucenie części poświadczeń reprezentujących przymiotnik *przyszły* w postaci np. *w przyszłem roku*.

> *Jedyne pocieszenie, że* ***w*** *[w:prep:loc:nwok]* ***przyszłem*** *[przyszłem:ign] roku ma być gruntowna modernizacja*

Innym problemem, z którym stykają się badacze, jest dokładne określenie zapytania w korpusie. Po pierwsze niezależnie od tego, czy stosujemy podejście *corpus-based* czy też *corpus-driven*, należy pamiętać nie tylko o regułach tagowania, lematyzacji, składni zapytań, ale także o tym, że zapytanie o daną treść znaczeniową (denotat) wymaga najczęściej wielokrotnych i długotrwałych przeszukiwań korpusu. Na przykład, gdybyśmy chcieli zbadać językowy obraz Niemca czy też Rosjanina, korzystając z korpusu NKJP, to w celu dokonania pełniejszej analizy nie powinniśmy ograniczyć naszego wyszukiwania do leksemu *Niemiec* lub *Rosjanin*. Wyniki takiej analizy będą ograniczone, ponieważ nie uwzględnią one tych tekstów, w których autorzy posługują się również synonimami o różnym nacechowaniu, pełniącymi funkcje ekspresywne (np. *helmut, szkop, Niemiaszek*, czy też *kacap, Rusek*), peryfrazami (np. *sąsiedzi zza Odry, sąsiedzi zza Buga*) lub zaimkami. Innym problemem jest to, że wyrazy ekspresywne mogą nie być frekwentowane w korpusach ogólnych. Może się też zdarzyć, że przy wyszukiwaniu nazwy podstawowej, trudno będzie nakreślić obraz wybranego leksemu. Przy badaniu kolokacyjnego obrazu Węgra w polszczyźnie, okazało się, że za pomocą kolokatów leksemu *Węgier*, wyekscerpowanych z NKJP, nie można zrekonstruować obrazu tej narodowości. [por. Bańko M., Doliński I., Duda J., Hebal-Jezierska M. 2012]. Natomiast kolokaty ekwiwalentów leksemu Węgier wyszukiwane w korpusach języka czeskiego oraz słowackiego dostarczyły tak wielu danych, że analiza kolokatów dodatkowych leksemów musiałaby zostać opisana w oddzielnym artykule.

Powyższe przykłady dotykają również innego ważnego ograniczenia korpusów językowych. Korpus jest obiektem skończonym, tj. ma ograniczony rozmiar. W praktyce oznacza to, że brak potwierdzenia jakiegoś faktu językowego w korpusie nie oznacza, że takowy fakt nie ma miejsca w rzeczywistości językowej [Górski 2012: 292]. I tutaj niezmiernie ważna jest intuicja badacza, o czym wspomina większość językoznawców korpusowych, analizując problematykę braku dowodu w korpusie.

Z powyższego można wysnuć wniosek, że przed rozpoczęciem przeszukiwań korpusu należy sformułować precyzyjne pytania badawcze i odpowiednio je zoperacjonalizować dla potrzeb przeszukiwania korpusów, a także upewnić się, czy dane badanie można przeprowadzić. W tym miejscu warto zwłaszcza zastanowić się nad tym, czy korpus, który chcemy wykorzystać, jest odpowiedni w odniesieniu do celów i zakresu planowanego badania.

Kolejny częsty problem, z którym mamy do czynienia niezależnie od tego, czy stosujemy podejście *corpus-based* czy *corpus-driven*, to duża liczba danych otrzymanych w wyniku przeszukiwania korpusów. Za przykład posłuży tu czasownik *ogarnąć/ogarniać*, który zostanie wyszukany w zrównoważonym podkorpusie NKJP (z 240 192 461 wyrazów tekstowych). Celem jest zbadanie, w którym znaczeniu występuje on najczęściej we współczesnej polszczyźnie. Internetowy *Słownik Języka Polskiego PWN* odnotowuje sześć znaczeń wspomnianego czasownika, tj.

> 1) ‘opasać ramionami’; 2) ‘otoczyć ze wszystkich stron’; 3) ‘objąć swoim zasięgiem’, 4) ‘opanować (uczucia, wrażenia)’; 5) ‘zdać sobie z czegoś sprawę’; 6) *pot.* ‘niezbyt dokładnie posprzątać’.

Można również zauważyć, że obecnie leksem ten, zwłaszcza w języku potocznym, używany jest również w znaczeniu innym, choć bliskim znaczeniu nr 5, ‘zrozumieć coś, nauczyć się czegoś, przyswoić coś’, np.

> (…) *ja tego nie* ***ogarniam*** *dlatego z tego nie korzystam wiesz po co mi aparat tym telefonie w sensie fotograficzny jakiejś wysokiej klasy* (PELCRA_7203010000054).

Wyniki przeszukiwania NKJP[59] to 8263 akapity pasujące do zapytania *ogarn**. Szczegółowa analiza wszystkich konkordancji będzie w takim przypadku niezwykle czasochłonna. Aby uzyskać pewien obraz tendencji użycia pary aspektowej czasownika *ogarnąć/ogarniać* zostanie przeanalizowana próbka 100 lub 200 poświadczeń, które – najprościej – można wybrać na przykład metodą próbkowania losowego lub w sposób bardziej systematyczny (tj. metodą próbkowania systematycznego). W tym drugim przypadku analizie zostaną poddane poświadczenia znajdujące się w równych odstępach od siebie, tj. przebadanie próbki 200 wystąpień z czasow-

[59] Do przeszukiwania korpusu NKJP użyliśmy wyszukiwarki PELCRA [Pęzik 2012].

nikiem *ogarnąć* oznacza mniej więcej co 40. poświadczenie. Wspomniane metody próbkowania w kontekście m.in. językoznawstwa szczegółowo opisują m. in. Oakes [1998], Rowntree [2000: 26-27], czy też Babbie [2013: 226]. Pobieżny przegląd konkordancji pokazuje, że większość wystąpień czasownika *ogarnąć/ogarnąć* dotyczy znaczenia nr 4, tj. 'opanować (uczucia, wrażenia)'.

Niemniej jednak stosowanie opisanych powyżej metod próbkowania pozwala na otrzymanie jedynie przybliżonego obrazu tendencji użycia wspomnianego czasownika. Uzyskanie reprezentatywniejszych wyników wymagałoby zastosowania zaawansowanych metod statystycznych polegających na przebadaniu losowo wybranej próbki konkordancji (np. 100 lub 200) oraz ekstrapolacji wyników względem wszystkich poświadczeń leksemu *ogarnąć/ogarniać*. Taką metodę opisują bardziej szczegółowo Hoffmann, Evert, Smith, Lee i Berglund Prytz [2008: 79-83]. Za przykład posłuży nam ponownie czasownik *ogarnąć/ogarniać*, a będziemy starali się oszacować, ile z 8263 poświadczeń analizowanego leksemu występuje w znaczeniu nr 1, tj. 'opasać ramionami'. W tym celu zostanie przebadana losowa próbka np. 100 konkordancji (np. pierwszych 100 poświadczeń zwróconych przez wyszukiwarkę korpusową). Po odnotowaniu, ile razy ów czasownik występuje w określonym znaczeniu, należy obliczyć przedział ufności (*confidence interval*), który mierzy stopień precyzji naszego oszacowania, najczęściej na poziomie 95%[60]. W tym celu warto skorzystać ze specjalnie do tego celu opracowanego narzędzia Corpus Frequency Wizard[61] [Baroni & Evert 2008], które jest dostępne w Internecie. Pobieżna analiza wykazała, że w losowej próbce 100 poświadczeń[62] w NKJP czasownik *ogarnąć/ogarniać* używany jest w poszukiwanym znaczeniu zaledwie trzy razy (przykłady wraz z identyfikatorem tekstu podajemy poniżej); co ciekawe, tylko raz w przebadanej próbce czasownik ten został użyty w znaczeniu potocznym 'niezbyt dokładnie posprzątać'

> *Dobrze, że chociaż obiady gotuje i trochę* ***ogarnie*** *w domu. Bo tak tobym jeszcze się musiał po różnych stołówkach żywić*. (PWN_2002000000179);

Znakomita większość poświadczeń w przebadanej próbce odnosi się natomiast do znaczenia nr 4, tj. 'opanować (uczucia, wrażenia)'.

> *Rysuję Twoje ręce na krzyżu umyślnie za długie niech* ***ogarną*** *ludzi, najwięcej rany grubsze, stopy za ogromne* (PWN_2002000000062)

[60] Oznacza to 5% ryzyko, że znaczenie poświadczeń czasownika *ogarnąć/ogarniać* jest w danym przedziale ufności inne od poszukiwanego.

[61] Opisywane narzędzie jest dostępne na stronie internetowej: http://sigil.collocations.de/wizard.html

[62] W innej losowo wybranej próbce wyniki wyszukiwania byłyby zapewne inne, dlatego też zaleca się tzw. łączenie próbek (wtedy analizie poddaje się większą liczbę konkordancji), co pozwoli na zmniejszenie przedziału ufności i zneutralizuje efekt zróżnicowania zawartości próbek [Hoffmann i in. 2008: 82].

*(…) Henryczek Fichtelbaum dostrzegał tylko łagodność na jej twarzy, a nieznaczny rys szyderstwa wziął za wyzwanie, zachętę, kuszenie, kiedy więc skończył jeść i pić, zbliżył się do kobiety, **ogarnął** ją ramieniem, a lewą dłoń położył na jej piersi.* (PWN_2002000000126)

*Podszedł bliżej brzegu, przeżegnał się na wszelki wypadek, wziął głęboki oddech i rzucił czarnej wdowie w objęcia. **Ogarnęło** go jej ciało, falowanie krwi i już płynął w jej ramionach, wyrzucał przed siebie ręce, a ona przyjmowała łagodnie każdy gest.* (PWN_2002000000184)

Korzystając z narzędzia Corpus Frequency Wizard [Baroni & Evert 2008], można stwierdzić, że na podstawie analizy losowych 100 poświadczeń przedział ufności dla wystąpień czasownika *ogarnąć/ogarniać* użytych w znaczeniu ‘opasać ramionami’ waha się w granicach 0.77%-9.15% wszystkich poświadczeń (8263). Innymi słowy, oznacza to tyle, że czasownik ten został użyty w poszukiwanym znaczeniu w przedziale od 63 do 756 wystąpień. Podobne szacunki dotyczą znaczenia potocznego ‘niezbyt dokładnie posprzątać’ (4 do 515 wystąpień). Warto w tym miejscu wspomnieć, że zwiększenie wielkości próbki (np. do 200 lub 300 losowo wybranych poświadczeń) pozwoli odpowiednio zmniejszyć przedział ufności i tym samym precyzyjniej oszacować liczbę wystąpień danego czasownika w wyszukiwanym znaczeniu [Hoffmann i in. 2008: 82]. Należy przy tym zauważyć, że w przypadku sprawdzenia poświadczeń każdej formy fleksyjnej (wyrazu tekstowego) oddzielnie, precyzja wyników wyszukiwania będzie bardzo niska.[63] Ponadto w omawianym przykładzie sprawdzaliśmy poświadczenia pary aspektowej wspomnianego czasownika, nie zwracając szczególnej uwagi na jego dystrybucję w różnych typach i gatunkach tekstów.

Truizmem jest stwierdzenie, że ogromne rozmiary współczesnych korpusów językowych sprawiają, że nie da się przeanalizować „ręcznie” tysięcy wyekscerpowanych poświadczeń; do tego niezbędne są zaawansowane narzędzia. Można obecnie zaobserwować, że językoznawcy coraz częściej zdają sobie sprawę z ograniczeń niektórych narzędzi komercyjnych, np. programów opracowanych na potrzeby analiz korpusowych (np. WordSmith Tools, AntConc). W związku z tym, jak pisze Gries [2015: 93], badacze coraz częściej wykorzystują specjalne skrypty opracowane we współpracy z informatykami. Wymaga to jednak umiejętności programowania, np. w języku R lub Python. Takie podejście umożliwia przeprowadzenie m.in. analiz wielowymiarowych pozwalających zbadać zależności pomiędzy wieloma zmiennymi i ich wpływem na frekwencję

[63] W tym miejscu można postawić pytanie, czy niektóre formy fleksyjne, występujące w różnych strukturach gramatycznych, mają tendencję do występowania częściej lub rzadziej w określonym znaczeniu danego leksemu.

i dystrybucję danych językowych, a także zweryfikować otrzymane wyniki w oparciu o bardziej zaawansowany aparat statystyczny [Gomez 2013: 89-133], np. zbadać zależności pomiędzy frekwencją różnych jednostek języka i ich klas (zmienne zależne) w różnych tekstach bądź typach i gatunkach tekstu (zmienne niezależne). Takie metody, których celem - jak piszą Baayen [2008: 127] i Gomez [2013: 90], - jest najczęściej dokonanie klasyfikacji lub opracowanie systemu klasyfikacji danych językowych, są obecnie z powodzeniem stosowane w badaniach z dziedziny stylometrii komputerowej oraz atrybucji autorstwa [np. Craig 2004, Craig & Kinney 2009], również na materiale tekstów napisanych w języku polskim [np. Eder 2014, Rybicki 2014], czy też analizy zróżnicowania rejestrów języka [Biber 1988, 1995]. Stosowaniu metod wielowymiarowych przyświecają inne cele niż tradycyjnemu czy też najbardziej powszechnemu podejściu do danych korpusowych, kiedy to sprawdzamy frekwencję danej jednostki języka w określonym zbiorze tekstów (wtedy mamy do czynienia z danymi dwuwymiarowymi, tj. frekwencją wyrazu *x* w tekście/korpusie *y*).

Na koniec zostanie jedynie zasygnalizowany problem celowości wykorzystania testów istotności statystycznej przy porównywaniu frekwencji danych językowych. W takim przypadku jest sprawdzane, czy różnice między frekwencją danej cechy językowej w różnych korpusach są istotne statystycznie. W praktyce jednak testy istotności statystycznej nie dostarczają nam informacji, jak znacząca jest to różnica (np. niewielka różnica we frekwencji danej jednostki w porównywanych korpusach, nawet jeśli istotna ze statystycznego punktu widzenia, może niewiele mówić o różnicach w zakresie stylu, gramatyki itp.). Rodzi to konieczność pomiaru tzw. wielkości efektu (*effect size*). Problem ten bardziej szczegółowo opisują, m.in. Coe [2002], Neill [2008], Corder i Foreman [2009], Cumming [2012] czy też Sulivan i Feinn [2012]. W przypadku dużych zbiorów danych językowych nawet niewielka różnica może okazać się istotna statystycznie; taki wynik jednak nie zawsze niesie relewantną wartość poznawczą. Ponadto siła efektu jest niezależna od wielkości próbki/korpusu. Obecnie zaleca się w pracach naukowych podawanie nie tylko wyników testów istotnosci statystycznej, ale także odpowiedniej miary siły efektu [Thompson 2000]. Dla przykładu miarą wielkości efektu dla testu chi-kwadrat może być współczynnik phi lub V Cramera [Corder & Foreman 2009: 169]. Ponadto, jak pisze Neill [2008, cyt. w Biddix [2015]], w przypadkach, gdy nie ma potrzeby uogólniania różnic między danymi, np. w porównywanych korpusach[64], badacz nie ma potrzeby korzystania z testów istotności statystycznej. Po drugie, jak pisze Neill [2008],

[64] Mamy tu na myśli sytuacje, w których badacza interesują wyłącznie wyniki uzyskane na podstawie przebadanej próbki tekstów. Natomiast nie ma on potrzeby wyciągania na ich podstawie wniosków uogólniających, dotyczących innych tekstów niż przebadane.

stosowanie testów istotności statystycznej może prowadzić do wyciągnięcia błędnych wniosków, zwłaszcza gdy porównywane są duże próbki danych. Jak wspomnieliśmy wcześniej, w takich przypadkach niewielkie i mało znaczące różnice mogą okazać się bowiem istotne statystycznie.

4. Podsumowanie

W niniejszym artykule podjęto próbę przedstawienia różnic między różnymi podejściami do analizy i interpretacji danych językowych zgromadzonych w korpusach tekstowych. Podniesione zostały także niektóre problemy związane z korzystaniem z tzw. korpusów narodowych, np. poprawność anotacji danych korpusowych, szczegółowość tagsetu, brak lub duża liczba poświadczeń, precyzyjne określanie zapytań, ograniczenia wyszukiwarek korpusowych i programów konkordacyjnych, a także celowość stosowania testów istotności statystycznej. Z omówionych kwestii wynika, że metodologia korpusowa jest niejednorodna, a samo korzystanie z korpusów w badaniach językoznawczych jest związane nie tylko z dużymi możliwościami, ale także wieloma ograniczeniami i pułapkami. Warto o nich pamiętać mając na uwadze cele i zakres planowanych badań.

Bibliografia

Baayen R. H., **2008**, *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Babbie E., 2013, *The Basics of Social Research*. Belmont: Wadsworth, Cengage Learning.

Bańko M., Doliński I., Duda J., Hebal-Jezierska M., 2012, *Collocation Images of Hungarians in Slavonic Languages – Research Report*. [w:] A. Obrębska (red.), *Practical Applications of Linguistic Research*, Łódź 2012.

Baroni M. & Evert, S., 2008, *SIGIL: Corpus Frequency Test Wizard*, dostęp 22.08.2016, http://sigil.collocations.de/wizard.html

Biber D., 1988, *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber D., 1995, *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biddix P., 2015, *Research Rundowns/Quantitative methods*, dostęp 22.08.2016, https://researchrundowns.files.wordpress.com/2009/07/rreffectsize_71709.pdf

Clear J., Fox G., Francis G. & Krishnamurthy R., 1996, *COBUILD: The State of the Art*, [w:] «International Journal of Corpus Linguistics», 1 (2), John Benjamins Publishing Company, 303-314.

Coe R., 2002, *It's the effect size, stupid: what "effect size" is and why it is important*. Paper presented at the 2002 Annual Conference of the British Educational Research Association, University of Exeter, Exeter, Devon, England, September 12–14, 2002, dostęp 25.08.2016, http: //www.leeds.ac.uk/educol/documents/00002182.htm (cited in Sulllivan & Feinn 2012).

Corder G., Foreman D., 2009, *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. Londyn: John Wiley & Sons.

Craig H., 2004, *Stylistic analysis and authorship studies*. In S. Schreibman, R. Simens & J. Unsworth (Eds.), *A Companion to Digital Humanities*. Oxford: Blackwell, 273–288.

Craig H. & Kinney A. F. (Eds.), 2009, *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Cumming G., 2012, *Understanding the new Statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Eder M., 2014, *Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu*, [w:] «Teksty drugie», 2: 90-105.

Fletcher, W. 2002–2007. *KfNgram*. Annapolis: USNA, dostęp 20 listopada 2011 r., http://www.kwicfinder.com/kfNgram/kfNgramHelp.html

Francis W., 2000, *Jazykové korpusy "před naším počítačovým letopočtem"* (Language Corpora B.C.), přeložil V. Petkevič. [w:] „Studie z korpusové lingvistiky", Univerzita Karlova v Praze, Karolinum, 169-181.

Francis W., Kucera H., 1964, *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Department of Linguistics, Brown University, Providence, Rhode Island, USA.

Fuster-Marquez M. & Clavel, B., 2010, *Corpus Linguistics and its Applications in Higher Education*. *Revista Alicantina de Estudios Ingleses*, 23, 51-67, dostęp 26.08.2016, http://rua.ua.es/dspace/bitstream/10045/17423/1/RAEI_23_04.pdf

Gomez P., 2013, *Statistical Methods in Language and Linguistic Research*. London: Equinox.

Górski R., 2012, *Zastosowanie korpusów w badaniu gramatyki.* [w:] A. Przepiórkowski, M. Bańko, R. Górski & B. Lewandowska-Tomaszczyk (red.), *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN, 291-300.

Górski R., Łaziński M., 2012, *Reprezentatywność i zrównoważenie korpusu*. In A. Przepiórkowski, M. Bańko, R. Górski & B. Lewandowska-Tomaszczyk (red.), *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN, 25-36.

Gries S., 2015, *Some Current Quantitative Problems in Corpus Linguistics and a Sketch of Some Solutions*, [w:] «Language and Linguistics», 16(1): 93-117, Dostęp: 27.07. 2016 [DOI: 10.1177/1606822X14556606], http://www.ling.sinica.edu.tw/files/publication/j2015_1_5_3464.pdf.

Hebal-Jezierska M., 2013, *Podstawowe zasady korzystania z korpusów przy badaniu języka*, [w:] W. Chlebda (Ed.), *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*. Opole. Wydawnictwo Uniwersytetu Opolskiego, pp. 17-30.

Hoffmann S., Evert S., Smith N., Lee D. & Berglund Prytz Y., 2008, *Corpus Linguistics with BNCweb: A Practical Guide*. Frankfurt am Main: Peter Lang.

Kuhn S., 1939, *The Dialect of Corpus Glossary*, [w:] «PMLA», 54 (1), 1-19.

Lee D., 2008, *Corpora and discourse analysis*. In V. Bhatia, J. Flowerdew & R. Jones (Eds.), *Advances in Discourse Studies*. London: Routledge, 86-99.

Leech G., 2000, *Korpusová lingvistika - současný stav oboru* (The State of Art in Corpus Linguistics), přeložil V. Petkevič. In. *Studie z korpusové lingvistiky*, Univerzita Karlova v Praze, Karolinum, 39-56.

Lewandowska-Tomaszczyk B. (red.), 2005, *Podstawy językoznawstwa korpusowego*, Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

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

McEnery T. & Gabrielatos C., 2008, *English Corpus Linguistics* In B. Aarts & A. McMahon (Eds.), *The Handbook of English Linguistics*. Oxford: Blackwell Publishing, 33-71.

McEnery T. & Hardie A., 2011, *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Neill J. T., 2008, *Why use effect sizes instead of significance testing in program evaluation?* Dostęp: 24.08.2016, http://wilderdom.com/research/effectsizes.html.

Oakes M., 1998, *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Ooi V., 1998, *Computer Corpus Lexicography*. Edinburgh: University Press.

Partington A., 2006, *The Linguistics of Laughter. A Corpus-Assisted Study of Laughter-Talk*.

Oxon: Routledge Studies in Linguistics.

Pęzik P., 2012, *Wyszukiwarka PELCRA dla danych NKJP*, [w:] A. Przepiórkowski, M. Bańko, R. Górski & B. Lewandowska-Tomaszczyk (Eds.), *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN, 253-274.

Piotrowski T., 2003, *Językoznawstwo korpusowe – wstęp do problematyki*, [w:] S. Gajda (red.), *Językoznawstwo w Polsce. Stan i perspektywy*, Opole: Wyd. Uniwersytetu Opolskiego, s. 143-154.

Piotrowski T., Grabowski Ł., 2013, *Interpretacja danych frekwencyjnych z korpusów*

językowych: opis pewnych problemów (na kilku przykładach z życia wziętych). [w:] W. Chlebda (red.), *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*. Opole. Wydawnictwo Uniwersytetu Opolskiego, s. 59-71.

Rowntree D., 2001, *Statistics Without Tears. An Introduction for Non-Mathematicians*. London: Penguin Books.

Rybicki J., 2014, *Pierwszy rzut oka na stylometryczną mapę literatury polskiej*, [w:] «Teksty drugie», 2: 106-128.

Sulivan G. & Feinn R., 2012, *Using Effect Size-or Why the P Value is Not Enough*, [w:] «Journal of Graduate Medical Education», 4 (3): 279-282.

Thompson B., 2000, *A suggested revision to the forthcoming 5th edition of the APA publication manual: Effect size section*. Available at: http://people.cehd.tamu.edu/~bthompson/apaeffec.htm (accessed 24 August 2016)

Tognini-Bonelli E., 2001, *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Xiao R., 2008, *Theory-driven corpus research: Using corpora to inform aspect theory*. In: **A. Lüdeling & Kytö M., (Eds.),** *Corpus linguistics: An international handbook Volume 2*. Berlin and New York: Walter de Gruyter, 987-1007.

Słownik Języka Polskiego PWN (www.sjp.pwn.pl)

Narodowy Korpus Języka Polskiego (www.nkjp.pl)

Narodowy Korpus Języka Rosyjskiego (ruscorpora.ru)

Brytyjski Korpus Narodowy (www.natcorp.ox.ac.uk/)

Czeski Korpus Narodowy (CzNK) www.korpus.cz

Słowacki Korpus Narodowy (SNK) www.korpus.sk