Victoria Kamasa
Uniwersytet im. Adama Mickiewicza w Poznaniu

Techniki językoznawstwa korpusowego wykorzystywane w krytycznej analizie dyskursu. Przegląd

Abstrakt Krytyczna analiza dyskursu (KAD) jako nurt z pogranicza socjologii i językoznawstwa skupia się na analizie roli dyskursu w umacnianiu i reprodukowaniu relacji władzy i dominacji, zaś badaniu empirycznemu podlegają przede wszystkim różnego rodzaju teksty. Właśnie koncentrację na tekście uznać można za jedną z przyczyn wzrastającego zainteresowania wykorzystaniem metod językoznawstwa korpusowego w takich analizach. Prezentacji tych metod poświęcony jest niniejszy artykuł. Przedstawione zostaną podstawowe strategie doboru korpusu do badań w ramach KAD oraz najczęściej wykorzystywane metody: analiza frekwencji, analiza słów kluczowych, analiza kolokacji i analiza konkordancji. Zaprezentowane zostaną także przykłady zastosowania poszczególnych metod w różnorodnych tematycznie badaniach. Przegląd metod podsumowany zostanie omówieniem korzyści wypływających z ich zastosowania oraz kosztów, które się z nimi wiążą.

Słowa kluczowe krytyczna analiza dyskursu, metody korpusowe, analiza kolokacji, analiza słów kluczowych, analiza konkordancji

Victoria Kamasa, dr, adiunkt w Instytucie Językoznawstwa na Uniwersytecie im. Adama Mickiewicza w Poznaniu. Jej zainteresowania badawcze mieszczą się na przecięciu tego, co językowe (doktorat z zakresu językoznawstwa stosowanego) i tego, co społeczne (magisterium z socjologii). Obecnie zajmuje się krytyczną analizą dyskursu Kościoła katolickiego w Polsce z wykorzystaniem narzędzi językoznawstwa korpusowego.

Adres kontaktowy:
Instytut Językoznawstwa, Collegium Novum
Uniwersytet im. Adama Mickiewicza w Poznaniu
al. Niepodległości 4, 61-874 Poznań
e-mail: vkamasa@amu.edu.pl

Ogromne zasoby językowe dostępne w sieci, możliwość digitalizacji tekstu, a także jego dalszej cyfrowej obróbki przyniosły znaczące zmiany dla wielu dyscyplin humanistyki i nauk społecznych. I tak, lingwiści zamiast samodzielnie tworzyć lub z trudem gromadzić zaledwie kilka przykładów dla omawianych przez siebie zjawisk, mogą dziś przy pomocy kilku kliknięć uzyskać dostęp do kilkuset, kilku tysięcy, a czasem nawet kilku milionów przykładów (Przepiórkowski i in. 2009). Historycy – dzięki możliwości geograficznego tagowania[1] tekstów źródłowych – mogą z niespotykaną dotąd łatwością obserwować przemieszczanie się wojsk czy rozprzestrzenianie się chorób (np. Gregory 2008). Socjolodzy mają możliwość korzystania z programów ułatwiających kodowanie analizowanych danych jakościowych, a także późniejsze dostrzeganie wzorców w badanym materiale (np. Schmidt, Skowrońska 2006). Zmiany technologiczne znacząco poszerzyły także możliwości dostępne dla analityków dyskursu, w tym badaczy pracujących w paradygmacie krytycznej analizy dyskursu (KAD). Tym możliwościom, przykładom ich wykorzystania, a także korzyściom i kosztom z nich wypływającym poświęcony jest niniejszy tekst.

[1] Oznaczanie wybranych informacji w tekście umożliwiające następnie automatyczne przeszukiwanie i porządkowanie tekstu wg tych informacji.

Wspomniana powyżej krytyczna analiza dyskursu jest szerokim nurtem badawczym, w którego centrum stoi zainteresowanie relacjami władzy i dominacji, a także rolą dyskursu w umacnianiu i reprodukowaniu tych relacji (van Dijk 2001; Rogers i in. 2005). Wśród celów, jakie stawia sobie KAD, wymienia się między innymi „opis, interpretację i wyjaśnianie relacji między językiem, praktyką społeczną i światem społecznym”[2] (Rogers i in. 2005: 376), „podnoszenie świadomości dotyczącej strategii używanych, by tworzyć, zachowywać i reprodukować (a)symetryczne relacje władzy, które to strategie realizowane są przy pomocy dyskursu” (Hidalgo Tenorio 2011: 184) czy „identyfikację i interpretację sposobów, w jakie ideologia funkcjonuje w dyskursie i poprzez dyskurs” (Breeze 2011: 520). Z perspektywy niniejszego artykułu szczególnie istotne jest założenie o zasadniczej społecznej roli tekstu będącego manifestacją dyskursu (Wodak 2001), a także praktyka badawcza, w której analizie poddaje się właśnie teksty, które – choć pochodzą z różnych źródeł (od prywatnych listów, przez artykuły prasowe do debat parlamentarnych) i bywają zapisem wypowiedzi ustnych – łączy to, że są tekstem właśnie.

[2] Wszystkie tłumaczenia pochodzą od autorki.

Koncentracja na tekście jako podstawowym przedmiocie analizy stała się zapewne przyczyną coraz wyraźniejszej tendencji do wykorzystywania narzędzi językoznawstwa korpusowego w analizach KAD (Baker i in. 2008). Sam korpus definiowany jest przeważnie jako zbiór tekstów, które podlegają obróbce maszynowej (ang. *machine-readable texts*) (McEnery, Wilson 2001). Zwraca się także uwagę na jego reprezentatywność i zrównoważenie, przy czym ta pierwsza rozumiana jest jako obecność wszystkich elementów analizowanej odmiany w korpusie (Gries 2009), zaś zrównoważenie jako zachowanie właściwych (uwzględniających częstotliwość i istotność) proporcji pomiędzy reprezentacją poszczególnych elementów badanej odmiany w korpusie (Gries 2009). Baker (2006) z kolei wskazuje na rozróżnienie pomiędzy korpusami ogólnymi (a więc zrównoważonymi ze względu na język jako całość) i korpusami specjalistycznymi, w którym punktem odniesienia dla reprezentatywności jest bądź jakaś odmiana języka (np. język gazet), bądź jakiś konkretny temat (np. uprzejmość). Korpusy zarówno ogólne, jak i specjalistyczne znajdują bardzo szerokie zastosowanie w językoznawstwie i dziedzinach pokrewnych: od badań historii języka (Hebda 2011), przez analizy struktur składniowych i semantycznych (Miechowicz-Mathiasen, Scheffler 2008), aż po badania nad sposobami konceptualizacji wybranych pojęć (Fabiszak, Hebda, Konat 2012).

Są one także coraz częściej używane w analizach prowadzonych w nurcie KAD.

Jak wspomniałam powyżej, celem niniejszego tekstu jest prezentacja podstawowych technik wywodzących się z językoznawstwa korpusowego, które są stosowane w badaniach KAD. Omówię więc poszczególne metody, przedstawiając krótko ich definicje, a także przykłady zastosowania w konkretnych badaniach prowadzonych w ramach KAD. Przegląd badań łączących językoznawstwo korpusowe i KAD nie rości sobie pretensji do zupełności – wybrane przykłady mają raczej ilustrować możliwości analityczne, jakie dają poszczególne techniki, a także wielość i różnorodność tematów, do badania których je wykorzystywano. Przegląd nie ma też na celu prezentacji możliwości, funkcji i ograniczeń konkretnego oprogramowania, a raczej omówienie technik korpusowych, których wykorzystanie stało się możliwe dzięki zwiększeniu mocy obliczeniowych komputerów.

Dobór korpusu

Jedno z podstawowych pytań, jakie nasuwa się w kontekście wykorzystania korpusów w KAD, dotyczy jego wielkości i sposobu doboru. Stosunkowo najrzadziej wykorzystywanym w ramach KAD podejściem jest prowadzenie badań na ogólnych korpusach językowych. I tak na przykład Hamilton i współpracownicy (2007) rekonstruują znaczenie słowa *risk* na podstawie ogólnych korpusów brytyjskiej i amerykańskiej wersji języka angielskiego. Na kwestiach o silniejszym nacechowaniu społecznym skupia swoją uwagę Maunter (2007), analizując funkcjonowanie słowa *elderly* w wybranych podkorpusach Bank of English. Z kolei Orpin (2005) porównuje dyskursywne konstrukcje związane z użyciem dwóch angielskich słów określających nieuczciwe zachowania osób publicznych: *corruption* i *sleaze*. Robi to na podstawie analizy ich wystąpień w podkorpusie prasy brytyjskiej korpusu Bank of English. Wspólną cechą opisanych badań jest koncentracja na semantyce badanych słów – szczegółowa analiza kolokacji bądź kontekstów, w których dane słowo występuje, daje bardzo dobry wgląd w jego znaczenie i funkcjonowanie w języku, pozostawia jednak pewien niedosyt w kwestii bardziej społecznie czy krytycznie nastawionych konkluzji.

Zdecydowanie popularniejszym nurtem wśród badaczy KAD jest wykorzystanie specjalistycznych korpusów. Mogą one zostać podzielone według dwóch kryteriów:

1. odmiany języka, którą zawierają: pisane *versus* mówione;
2. części badanej populacji, którą obejmują: próbkowane *versus* pełne.

Najczęściej badaniu podlegają korpusy tekstów pisanych, stanowiące określony wycinek wszystkich tekstów, które można uznać za istotne ze względu na badane zagadnienie. W wielu przypadkach badacze decydują się jednak na tak szczegółowe sformułowanie problemu badawczego, które pozwala im objąć badaniem wszystkie teksty istotne ze względu na postawiony problem.

Tylko nieliczni badacze skupiają się na analizie mowy: badaniu podlegają wtedy transkrypcje nagrań audio pochodzących bądź z naturalnych sytuacji komunikacyjnych (Herbel-Eisenmann, Wagner 2010), bądź z badań prowadzonych metodami nauk społecznych, takimi jak wywiady czy badania fokusowe (Weninger 2010). Do zalet tego typu badań należy koncentracja na najbardziej pierwotnej i spontanicznej formie języka, jaką jest mowa. Można przypuszczać, że o stosunkowo małym wykorzystaniu tego typu korpusów w KAD decydują liczne wyzwania techniczne i organizacyjne związane z ich tworzeniem[3].

Z kolei wśród przykładów badań prowadzonych na populacjach znajdziemy analizy oparte na pełnych tekstach debat parlamentarnych dotyczących określonego tematu (Bachmann 2011; Subtirelu 2013), oficjalnych dokumentach wybranych instytucji (Albakry 2004; Freake, Gentil, Sheyholislami 2010; Kamasa 2013a) czy też tekstach medialnych obejmujących wąski okres czasowy wokół wydarzenia uznawanego przez autora za istotne społecznie (Hidalgo Tenorio 2011). Wykorzystanie korpusu obejmującego całą populację tekstów zapewnia wysoki poziom trafności prezentowanych wyników. Z drugiej jednak strony wymusza koncentrację na wąskich zagadnieniach, których istotność ze względu na stawiane sobie przez KAD cele może budzić pewne wątpliwości.

Największą i najbardziej zróżnicowaną grupę analiz stanowią badania oparte na korpusach złożonych z tekstów wybranych w jakiś sposób z szerszego zbioru. Badacze kierują się tu różnorodnymi kryteriami:

- intuicyjnym przekonaniem o istotności tekstów wybieranych do korpusu: Salama (2011) analizuje dwie wybrane przez siebie książki, w których prezentowane są odmienne wizje wahhabizmu, a Fortchner i Kolvraa (2012) pracują na korpusie przemówień polityków, których uznają za istotnych aktorów na scenie politycznej;
- wysoką pozycją w wybranych rankingach: Alcaraz-Ariza i Ángeles (2002) tworzą korpus recenzji książek medycznych, wybierając te opublikowane w najlepszych czasopismach z interesującej ją dziedziny, Lischinsky (2011) wykorzystuje raporty największych szwedzkich firm, zaś Prentice (2010) prowadzi analizy na korpusie złożonym z postów opublikowanych na największym forum dotyczącym interesującego ją tematu;
- zawieraniem określonych słów bądź poruszaniem określonych tematów: Koller (2004) pracuje na dwóch subkorpusach złożonych z tekstów opisujących odpowiednio businessmanów i businesswomen, Lukac (2011) decyduje się na analizę blogów, których autorki otwarcie deklarują bycie *pro-ana*[4], a Garbielatos i Baker (2008) tworzą złożony algorytm pozwalający im wybrać artykuły prasowe dotyczące imigrantów;
- czasem powstania: Almeida (2011) i Chen (2012) konstruują swoje korpusy, wybierając artykuły opublikowane w określonym przedziale czasowym.

[3] Do wyzwań tych zaliczyć można między innymi konieczność organizacji nagrań, zapewnienia odpowiedniego poziomu technicznego nagrań czy sporządzenia transkrypcji, która to konieczność wiąże się ze znacznie większymi nakładami środków i czasu niż w przypadku prowadzenia badań na tekstach już zdygitalizowanych, jak na przykład teksty z internetowych wydań gazet.

[4] *Pro-ana* jest nazwą grupy osób uznających anoreksję za styl życia (Lukac 2011).

Wewnątrz zaproponowanych powyżej kategorii stosowane są różnorodne sposoby wyboru konkretnych tekstów, które mają znaleźć się w badanym korpusie. Stosując terminologię opisującą sposób doboru próby w badaniach społecznych (Babbie 2003), możemy tu mówić o doborze losowym (np. Almeida 2011; Chen 2012), celowym (np. Gabrielatos, Baker 2008) czy dostępnościowym (np. Don, Knowles, Fatt 2010). Jak wskazują przytoczone powyżej przykłady, podejście takie umożliwia podejmowanie bardzo różnorodnych zagadnień, tym samym pozwala badaczowi kierować się przy wyborze analizowanego tematu przede wszystkim jego istotnością społeczną (zgodnie z założeniami przyjmowanymi w KAD), nie zaś możliwościami techniczno-organizacyjnymi. W przypadku każdego korpusu powstaje jednak pytanie, na ile wybrane teksty odzwierciedlają dominujące czy też społecznie najistotniejsze dyskursy w danej kwestii.

Korpusy tworzone na potrzeby poszczególnych projektów badawczych zasadniczo różnią się też rozmiarem: zawierają zazwyczaj od kilkudziesięciu (np. O'Halloran 2009) do kilkuset tysięcy słów (np. Bachmann 2011), choć zdarzają się także badania prowadzone na znacznie większych korpusach obejmujących od kilku (np. Don 2010) do nawet kilkuset milionów (np. Gabrielatos, Baker 2008) słów. Wielkość korpusów wydaje się być determinowana zarówno dostępnością tekstów związanych z analizowanym zagadnieniem, jak i technikami, które badacz zamierza wykorzystać: niektóre z nich pozwalają na wyciąganie wniosków jedynie w przypadku bardzo dużych korpusów, podczas gdy inne wręcz przeciwnie – znajdują zastosowanie właściwie wyłącznie w przypadku małych i średnich korpusów. Na koniec warto zauważyć, że, jak stwierdza Baker (2006: 28), nie ma prostej odpowiedzi dotyczącej tego, jaki rozmiar powinien mieć korpus, a najważniejszym kryterium, którym należy się kierować, podejmując decyzję dotyczącą jego wielkości, jest cel, jakiemu korpus ten ma służyć.

Zastosowanie korpusów w badaniach nad dyskursem wymaga wykorzystania specjalistycznego oprogramowania komputerowego przede wszystkim ze względu na wielkość analizowanego materiału badawczego, a także procedury statystyczne stosowane w poszczególnych technikach. Do najpopularniejszego oprogramowania stosowanego przez badaczy należą bezpłatny AntConc (dostępny na stronie www.antlab.sci.waseda.ac.jp/software.html) i dostępny na licencji WordSmith Tool (Scott 2013b). Oba umożliwiają pracę na korpusach z polskimi znakami diakrytycznymi, obliczanie list frekwencyjnych, list słów kluczowych i list kolokacji dla modyfikowalnych ustawień, a także eksport wyników do wybranych programów. W poszukiwaniu bardziej zaawansowanych narzędzi, warto zwrócić uwagę na te dostępne w ramach Wmatrix, opracowywane przez zespół pod kierownictwem Paula Raysona (2009).

Podstawowe techniki korpusowe stosowane w krytycznej analizie dyskursu

Analiza list frekwencyjnych

Lista frekwencyjna (ang. *word list*, *frequency list*) określana jest jako „lista wszystkich słów pojawiających się w korpusie wraz z częstotliwością ich występowania i procentowym udziałem w korpusie[5]" (Baker 2006: 51). Jest ona automatycznie generowana przez programy służące do analizy korpusów, takie jak powyżej wspomniane AntConc czy WordSmith Tool. Stanowi również podstawę dla bardziej złożonych narzędzi analitycznych, takich jak na przykład opisane poniżej słowa kluczowe. Sama w sobie może zostać uznana za najbardziej podstawowe narzędzie we wspieranej korpusowo KAD.

Generując listę frekwencyjną, badacz otrzymuje informacje dotyczące wszystkich słów występujących w tekstach poddawanych analizie, jak i częstotliwości ich występowania. Takie dane zostały wykorzystane na przykład przez Chen (2012) w badaniach dotyczących wpływu zmian politycznych w Chinach na prasę. Wpływ ten jest obserwowany przez częstotliwość użycia pozytywnie i negatywnie nacechowanych oraz neutralnych słów porównujących w diachronicznym korpusie artykułów prasowych. Z kolei Mohamad i współpracownicy (2012) wykorzystali listy frekwencyjne stworzone dla dwóch subkorpusów tekstów z podręczników do matematyki do porównania reprezentacji płci w anglojęzycznych podręcznikach wydawanych w Katarze i poza jego granicami. Analiza częstotliwości występowania wybranych zaimków, nazw zawodów czy określeń pokrewieństwa umożliwiła autorom odpowiedź na pytanie dotyczące poziomu seksizmu w obu grupach podręczników. Potencjał listy frekwencyjnej jako narzędzia wspierającego porównywanie różnych dyskursów wykorzystał również Edwards (2012) w badaniach dotyczących dyskursu Brytyjskiej Partii Narodowej (ang. *British National Party*, *BNP*). Porównanie częstotliwości występowania poszczególnych słów w manifestach BNP z 2005 i 2010 roku umożliwiło mu wybranie słów-węzłów (ang. *node-words*), których częstotliwość istotnie zmieniła się w badanych manifestach i których konteksty występowania zostały następnie poddane dalszej analizie jakościowej. W przypadku analizy list frekwencyjnych oprogramowanie jest więc wykorzystywane w celu ich wygenerowania, podczas gdy wybór jednostek leksykalnych poddawanych dalszej analizie, a także interpretacja uzyskiwanych częstotliwości pozostają po stronie badacza.

Analiza słów kluczowych

Słowa kluczowe określane są jako „punkty, wokół których toczone są bitwy ideologiczne" (Stubbs 2001: 188). Z kolei Mautner (2005), cytując Williamsa, wskazuje na zasadność, a nawet konieczność zainteresowania się słowami kluczowymi, jako że słowa powinny być widziane jako element problemów. Przytoczone wypowiedzi nie dają jednak jasnych wskazówek, jak ustalić, co jest słowem kluczowym dla danego tekstu, dyskursu lub problemu. Spotyka się tu dwa podejścia: pierwsze z nich oparte jest na wiedzy dotyczącej poruszanego zagadnienia, którą posiada badacz. Na jej podstawie wybiera on słowa kluczowe, których występowanie i funkcjonowanie w tekście zostanie później poddane dalszej analizie (np. Mautner 2005; Degano 2007). Drugi sposób poszukiwania słów kluczowych jest bardziej zakorzeniony w tradycji językoznawstwa korpusowego, zgodnie z którą słowo kluczowe definiuje się jako słowo, jakie występuje w jednym korpusie znacząco

[5] Dostępność drugiej z wymienionych informacji jest zależna od używanego oprogramowania.

częściej niż w innym (określanym jako korpus referencyjny) (Scott 2013a). W celu obliczenia tak zdefiniowanych słów porównuje się za pośrednictwem wspomnianego oprogramowania listę frekwencyjną badanego korpusu z listą frekwencyjną korpusu referencyjnego. Następnie, przy pomocy statystycznych miar kluczowości (np. logarytmiczny wskaźnik wiarygodności [ang. *log-likelihood*] czy test chi[6]), generuje się listę słów występujących statystycznie znacząco częściej[7] w badanym korpusie niż w korpusie referencyjnym.

Przy takiej procedurze określania słów kluczowych szczególnie istotne staje się zagadnienie wyboru korpusu referencyjnego, bowiem to właśnie od jego składu zależeć będzie uzyskany wynik. W tym kontekście Bondi i Scott (2010) zwracają uwagę, że raczej trudno mówić o jednym zakresie tematycznym (ang. *aboutness*) tekstu, o którym informacje pozyskujemy, analizując słowa kluczowe. Za bardziej trafne uznają przyjęcie, że istnieje wiele różnych zakresów tematycznych danego tekstu, które mogą być odkrywane przez zestawienie go z różnymi korpusami referencyjnymi. I tak, chcąc uzyskać ogólne wskazówki dotyczące pojęć istotnych w danym zbiorze tekstów, badacze posługują się zazwyczaj ogólnymi korpusami dla danego języka[8]. Natomiast kiedy ze względu na analizowane zjawisko lub stawiane pytanie badawcze istotny staje się określony kontekst, wykorzystuje się korpusy referencyjne przygotowywane specjalnie na potrzeby danego badania.

Drugie z opisanych podejść zapewnia koncentrację na wyrażeniach charakterystycznych dla badanego dyskursu, nie zaś na przykład dla gatunku: wykorzystuje je Lukac (2011), porównując badane przez siebie blogi z korpusem blogów na inne tematy. Unika w ten sposób koncentracji na zjawiskach charakterystycznych dla blogu jako gatunku internetowego. Taka strategia okazuje się też bardzo pomocna w ujawnianiu różnic w dyskursach różnych grup na ten sam temat. I tak Subtirelu (2013) wskazuje na różne podejścia do osób nieposługujących się językiem angielskim wśród zwolenników i przeciwników ustawy o wielojęzycznych kartach do głosowania. Pozwala mu je ujawnić analiza słów kluczowych uzyskanych przez zestawienie korpusu wypowiedzi zwolenników ustawy z korpusem wypowiedzi jej przeciwników. Podobną strategię stosuje Baker (2006), zestawiając głosy zwolenników i przeciwników zakazu polowania na lisy z debaty parlamentarnej dotyczącej tego zagadnienia. Dzięki temu ujawnia między innymi dyskursywne powiązanie polowania na lisy z brytyjską tożsamością, obecne w debacie zwolenników polowań czy tendencję do dosłownych i dosadnych określeń efektów polowania wśród jego przeciwników. Z kolei wykorzystanie jako korpusu referencyjnego korpusu ogólnej angielszczyzny FLOB i dalsza analiza uzyskanych w ten sposób słów kluczowych umożliwia Weninger (2010) identyfikację kategorii wspomaganej podmiotowości (ang. *facilitated agency*), charakterystycznej dla amerykańskiego dyskursu dotyczącego rewitalizacji miast i zrywającej z klasycznym podziałem obrazowania kluczowych uczestników sytuacji społecznych jako pasywnych lub aktywnych.

Wspomniane powyżej przykłady wskazują, że sama lista słów kluczowych nie umożliwia odpowiedzi na pytanie dotyczące dyskursów obecnych w badanym materiale. Konieczna jest dalsza analiza, na którą składa się przyporządkowywanie słów kluczowych do grup tematycznych (np. Gabrielatos, Baker 2008) bądź szczegółowa analiza kontekstów, w których występują w poszukiwaniu wzorców (np. Weninger 2010). Zatem podobnie jak w przypadku list frekwencyjnych program komputerowy generuje jedynie listę słów, zaś jej dalsze opracowanie należy już do badacza. Słowa kluczowe stanowią jednak istotną wskazówkę kierującą uwagę badaczy na zjawiska charakterystyczne dla analizowanego przez nich dyskursu. Mogą być więc swego rodzaju punktem wejścia (ang. *point of entrance*) do zebranych danych.

Analiza kolokacji

Kolokacja jest przez badaczy KAD definiowana zgodnie z tradycją językoznawstwa korpusowego jako częste współwystępowanie (Stubbs 2001). Dla wybranego słowa określa się więc zasięg[9] (ang. *span*), a następnie przy pomocy oprogramowania wykorzystującego określone miary statystyczne (test t, wskaźnik MI i inne[10]) generuje się listę słów występujących istotnie częściej w określonym zasięgu od słowa bazowego. Po uzyskaniu takiej listy, badacze przechodzą do bardziej jakościowej, skupionej na znaczeniu analizy. Spotyka się tu zasadniczo dwa różne modele postępowania:

- koncentracja na samych kolokacjach – słowa z otrzymanej listy łączone są w grupy tematyczne, które pozwalają określić funkcjonowanie kluczowego pojęcia w dyskursie przez odtworzenie najważniejszych domen, z którymi jest kojarzone lub w otoczeniu których występuje. Taką strategię przyjmują na przykład Freake i współpracownicy (2010) w swoich badaniach dotyczących tożsamości mieszkańców Quebecu. Pozwala im to ustalić na przykład, że dla francuskojęzycznych mieszkańców tej prowincji Quebec wiąże się przede wszystkim z narodem rozumianym jako pewna wspólnota historyczna;

- analiza kontekstu zawierającego otrzymane kolokacje – dla otrzymanych słów generuje się listy ich wystąpień w tekście wraz z najbliższym kontekstem, a następnie analizuje otrzymane fragmenty w poszukiwaniu wzorców. Takie działanie prowadzi Forchtnera i Kolvraa (2012) do ustalenia, że kluczowe dla konstrukcji tożsamości europejskiej w badanym przez nich materiale są wspólnie wyznawane wartości, które są umieszczone w kontekście wspólnej trudnej przeszłości.

Drugim istotnym zagadnieniem związanym z wykorzystaniem analizy kolokacji jest wybór słów, dla których kolokacje te mają zostać ustalone. Również tutaj badacze wybierają różnorodne sposoby

[6] Więcej informacji na temat miar statystycznych stosowanych dla obliczania słów kluczowych, a także ich ograniczeń można znaleźć np. w Gabrielatos i Marchi (2011; 2012).

[7] W niektórych badaniach wykorzystuje się także tzw. negatywne słowa kluczowe, czyli takie, które występują w analizowanym korpusie znacząco rzadziej niż w korpusie referencyjnym.

[8] W przypadku badań dotyczących tekstów w języku polskim możliwe jest skorzystanie z list frekwencyjnych przygotowanych dla Narodowego Korpusu Języka Polskiego (http://nkjp.uni.lodz.pl/).

[9] Czyli liczbę słów po prawej i lewej stronie od wybranego słowa, wśród których mają być poszukiwane kolokacje.

[10] Bardziej szczegółowe informacje dotyczące wskaźników statystycznych stosowanych do obliczania kolokacji znaleźć można w Gries (2010).

postępowania: Subtirelu (2013) poszukuje kolokacji dla ustalonych wcześniej słów kluczowych. Pozwala mu to skupić uwagę na konstrukcji pojęć charakterystycznych dla badanego przez niego dyskursu[11]. Z kolei Lischinsky (2011), ze względu na swoje zainteresowanie konstrukcją kryzysu ekonomicznego w prasie, decyduje się skorzystać ze źródeł leksykograficznych (słowniki i baza FrameNet) w poszukiwaniu słów używanych, by określać kryzys. Własną intuicją kieruje się z kolei Mautner (2005)[12], określając słowa istotne w jej opinii dla koncepcji *entrepreneurial university* i skupiając swoją uwagę na analizie ich profili kolokacyjnych.

Podsumowując, profile kolokacyjne generowane przy pomocy programu komputerowego wykorzystywane są jako podstawa do identyfikacji konstrukcji dyskursywnych stanowiących przedmiot zainteresowania badacza. Najczęściej służą one uzyskaniu bardziej szczegółowych informacji na temat funkcjonowania określonych słów w badanych tekstach. Na podstawie takiej informacji badacz identyfikuje później struktury dyskursywne. Analiza kolokacji, zwłaszcza prowadzona na dużych korpusach języka ogólnego, pozwala także na ustalenie, jak badane słowo i związany z nim koncept funkcjonują społecznie (np. Mautner 2007; Marling 2010). Lista kolokacji oddaje *stricte* ilościowe relacje pomiędzy słowami występującymi w badanym tekście. Poddana dalszej jakościowej analizie, pozwala jednak zrozumieć złożone relacje pomiędzy reprezentacjami.

[11] Podobną strategię stosują również Don, Knowles, Fatt (2010), Freake, Gentil, Sheyholislami (2010) oraz Salama (2011).

[12] Podobną strategię stosują również Mautner (2005; 2007), Hamilton, Adolphs, Nerlich (2007), Forchtner, Kolvraa (2012).

Prozodia semantyczna

Jednym ze szczególnych przypadków wykorzystania list kolokacji jest analiza prozodii semantycznej. Pojęcie to zostało zaproponowane przez Louwa (1993), który zdefiniował ją jako „powtarzającą się konsekwentnie aurę znaczeniową, którą dane słowo zostaje przepojone przez swoje kolokacje" (s. 157). Prozodię semantyczną dla wybranego słowa badacz określa więc na podstawie oceny nacechowania (najczęściej na osi pozytywne–negatywne) jego najsilniejszych kolokacji. Sama koncepcja była poddawana krytyce (zobacz np.: Whitsitt 2005) dotyczącej na przykład możliwości transferu nacechowania z jednej jednostki leksykalnej na inną czy rodzaju nacechowania, jakiego powinny dotyczyć analizy[13].

Mimo tej krytyki analiza prozodii semantycznej jest wykorzystywana w badaniach z nurtu KAD. I tak na przykład Mautner (2007) wskazuje, że analizowane przez nią słowo *elderly* ma znacznie silniejszą negatywną prozodię semantyczną, kiedy jest używane jako rzeczownik (osoba w podeszłym wieku) niż gdy występuje jako przymiotnik (starszy). Z kolei Kamasa (2013a) pokazuje, jak użycie słowa *praktyka* w odniesieniu do zapłodnienia *in vitro* przyczynia się do jego negatywnej konstrukcji w dyskursie Kościoła katolickiego w Polsce. Analizę prozodii semantycznej wykorzystują również Hamilton, Adolphs i Nerlich (2007) w swoich badaniach nad znaczeniem słów *ryzyko* i *ryzykować* czy Gabrielatos i Baker (2008), którzy analizują dyskursywną konstrukcję emigrantów w brytyjskiej prasie.

[13] Nacechowanie wyłącznie pozytywne bądź negatywne lub bardziej złożone modele analizy nacechowania (Oster 2010).

Przykłady wykorzystania prozodii semantycznej w KAD wskazują, że jest ona użytecznym narzędziem pozwalającym na identyfikację ukrytych i nieoczywistych wzorców funkcjonowania wybranych pojęć w dyskursie. Z drugiej jednak strony wątpliwości dotyczące realności psychologicznej założenia, że częste występowanie określonego wyrażenia w sąsiedztwie negatywnie nacechowanych wyrażeń powoduje zmianę jego nacechowania, każą zachować daleko posuniętą ostrożność w interpretacji wyników takich analiz.

Preferencja semantyczna

Drugim ze sposobów bardziej złożonego wykorzystania list kolokacji w ramach KAD jest analiza preferencji semantycznej. Definiuje się ją jako tendencję określonej jednostki leksykalnej do częstego kolokowania z serią jednostek należących do jednego pola semantycznego (Salama 2011). Podobnie jak w przypadku prozodii semantycznej, preferencja określana jest więc na podstawie listy kolokacji. Z tą różnicą, że tym razem badacz ocenia nie nacechowanie poszczególnych kolokacji, ale ich przynależność do określonych grup tematycznych (pól semantycznych).

Wśród badań, w których analiza preferencji semantycznej prowadzi do identyfikacji społecznych konstrukcji wybranych zjawisk, wskazać można wspomnianą już pracę Hamiltona i współpracowników (2007), w której wykorzystuje się analizę preferencji do identyfikacji dominujących dyskursów związanych z ryzykiem. Prowadzi to na przykład do stwierdzenia obecnej w badanym korpusie tendencji do oceny wielkości ryzyka. Z kolei Salama (2011) ustala z wykorzystaniem takiej analizy, że sposób reprezentacji wahhabizmu w jednym z badanych przez niego źródeł związany jest z zagrożeniem, konspiracją i koncepcją państwa policyjnego.

Koncentracja na grupach tematycznych, do których należą kolokacje słów określających zjawiska, jakimi zajmuje się badacz, prowadzi więc do ujawnienia sposobów jego reprezentowania charakterystycznych dla badanego dyskursu. Za pewną słabość tej analizy można uznać oparcie klasyfikacji do poszczególnych pól semantycznych wyłączne na intuicji badacza (nie istnieje ani jedna lista pól semantycznych, którą posługują się badacze, ani algorytm pozwalający na przydzielanie słów do takich pól w sposób charakteryzujący się wysokim poziomem intersubiektywności).

Analiza konkordancji

O ile wszystkie opisane powyżej techniki raczej wspierają analizę jakościową bądź dostarczają badaczowi wskazówek, gdzie taką analizę zacząć lub czego może ona dotyczyć, to analiza konkordancji może zostać uznana za najbardziej zbliżoną do klasycznie rozumianej jakościowej analizy dyskursu[14]. Samą konkordancję definiuje się jako „listę wszystkich wystąpień poszukiwanego terminu w korpusie, zaprezentowaną wraz z kontekstem, w którym termin ten się pojawia" (Baker 2006: 71). Długość kontekstu jest określana przez badacza i mierzona przy pomocy liczby słów lub znaków (w zależności

[14] Baker i in. (2008) wskazują na przykład, że analiza konkordancji jest jedynym narzędziem językoznawstwa korpusowego, z użyciem którego analitycy dyskursu czują się swobodnie.

od używanego oprogramowania), zaś lista konkordancji jest generowana według wybranych parametrów przez używane oprogramowanie.

Jako że konkordancje dają możliwość obserwowania wybranych terminów w ich najbliższym kontekście i tym samym rekonstrukcji dyskursów związanych z tymi terminami, ich analiza wykorzystywana jest niemal we wszystkich badaniach z nurtu KAD prowadzonych z użyciem narzędzi korpusowych. Przy pomocy powyżej opisanych technik bądź kierując się intuicją, badacz ustala słowa, których kontekst występowania zostaje następnie poddany analizie właśnie na podstawie listy konkordancji. I tak na przykład Albakry (2004), analizując konkordancje w kanadyjskim i amerykańskim raporcie dotyczącym incydentu bratobójczego ognia w Kandaharze w 2002 roku, pokazuje jak grzeczność[15] wpływa na kształt tych raportów. Edwards (2012) w swojej analizie manifestów Brytyjskiej Partii Narodowej, skupiając się na konkordancjach dla słów *our* i *British*, demonstruje wzrastającą tendencję do ukrywania rasizmu i konstrukcji grupy własnej w oparciu o pozornie bardziej inkluzywną kategorię narodowości.

Przykład bardziej osadzonego teoretyczne zastosowania kolokacji znajdziemy u Mulderrig (2011), która koduje badane przez siebie kolokacje według typów akcji (ang. *action-type*), zaproponowanych przez Hallidaya i Matthiessena (2004). Pozwala jej to ustalić diachroniczne zmiany w sposobie reprezentowania rządu w dyskursie laburzystów dotyczącym edukacji w Wielkiej Brytanii. Z kolei Kamasa (2013b), opierając się na kategoriach teoretycznych zaproponowanych przez van Leeuwena (2008), określa na podstawie odpowiednio wyszukanych konkordancji dyskursywną konstrukcję *rodziny* w oficjalnych dokumentach Kościoła katolickiego w Polsce.

Do zalet analizy konkordancji należy możliwość zbadania kontekstu występowania słów istotnych ze względu na stawiane pytanie badawcze nawet w bardzo dużych korpusach tekstów. Możliwość automatycznego wygenerowania listy wszystkich wystąpień wybranego słowa wraz z jego najbliższym kontekstem znacząco skraca proces analizy, a także podnosi jego stopień trafności (istnieje pewność, że zanalizowane zostały wszystkie wystąpienia danego słowa) i powtarzalności (różni badacze dla tego samego korpusu uzyskają zawsze tę samą listę konkordancji). Poszukiwanie wzorców wśród uzyskanych konkordancji prowadzi badaczy do identyfikacji dominujących dyskursów i sposobów dyskursywnej reprezentacji analizowanych zjawisk. Wadą jest natomiast skupienie się na słowie/słowach, nie zaś problemach, o których w tekście może być mowa, bez użycia wyszukiwanego słowa (np. przy pomocy zaimków).

Ocena zastosowania metod korpusowych w KAD

Wykorzystanie metod korpusowych w badaniach z zakresu KAD wiąże się z istotnymi zmianami w tym polu badawczym. Za najistotniejsze z nich uznajemy znaczące zwiększenie liczby danych poddawanych analizie, podniesienie stopnia przejrzystości stosowanych procedur badawczych oraz koncentrację na wzorcach ilościowych. Zmiany te, choć zyskują rosnące grono zwolenników, mogą także budzić pewne obawy i wątpliwości.

Korzyści

Zwiększenie liczby analizowanych danych podnosi trafność uzyskiwanych wyników. Badania oparte na dziesiątkach czy setkach tysięcy słów pochodzących z gazet pozwalają na udzielenie pełniejszej odpowiedzi na pytanie dotyczące reprezentacji na przykład imigrantów niż analiza kilku artykułów. Ponadto możliwość pracy ze stosunkowo dużym korpusem danych daje badaczowi szanse na jego bardziej zrównoważony dobór oraz pozwala zastosować złożone algorytmy wyboru tekstów. Zmniejsza to ryzyko skupienia się na badaniu tekstów, których wyborem kierowała wyłącznie intuicja. Co więcej, jednym z punktów krytyki podnoszonej wobec KAD jest właśnie brak reprezentatywności analizowanych tekstów (Stubbs 1997) czy też kierowanie się osobistymi pobudkami w ich wyborze (Breeze 2011). Zastosowanie dużych korpusów danych stanowi częściową odpowiedź na tę krytykę.

Z kolei podniesienie poziomu przejrzystości stosowanych procedur zwiększa możliwość replikacji prowadzonych badań. Ma również znaczenie w kontekście badań porównawczych: profile kolokacyjne czy słowa kluczowe obliczone przy pomocy określonych metod dla określonych korpusów w jednym języku bądź okresie czasu mogą być porównywane z danymi uzyskanymi w analogiczny sposób dla innego języka bądź innego okresu. Ponadto jawność i przejrzystość metod prowadzących do uzyskanych wyników mogą zwiększać poziom ich wiarygodności i tym samym czynić je bardziej przekonującymi. Podobnie jak w przypadku analizowanych danych również brak jasnych metod jest jednym z problemów wskazywanych przez krytyków KAD (por. np. Breeze 2011). Zastosowanie wybranych technik korpusowych dostarcza częściowego rozwiązania tego problemu.

Wykorzystanie opartych na wzorcach ilościowych list słów kluczowych czy list kolokacji pozwala badaczom skupić się na najczęściej powtarzających się w tekstach słowach czy konceptach. Uzyskuje się w ten sposób punkt rozpoczęcia dalszych analiz, osadzony w samych danych, nie zaś intuicjach czy wiedzy badacza. Odbiorcy wyników badań zyskują w ten sposób jasność dotyczącą kryteriów wyboru słów, które poddane zostały dalszej analizie, co ułatwia zrozumienie i interpretację prezentowanych rezultatów. Może to również prowadzić do zmniejszenia poziomu stronniczości (ang. *bias*) uzyskiwanych wyników – uwaga badacza zostaje skierowana na kwestie najczęściej pojawiające się w badanych przez niego tekstach, nie zaś na te, które wydają mu się intuicyjnie najistotniejsze.

Uzyskana dzięki oprogramowaniu i technikom korpusowym możliwość obserwacji i interpretacji wzorców ilościowych w badanych tekstach może sprzyjać ujawnianiu ukrytych i nieoczywistych tendencji obecnych w analizowanym materiale. Profile kolokacyjne wybranych słów czy też powtarzalność pewnych sposobów reprezentacji obserwowana przy pomocy analiz konkordancji pozwalają skupić uwagę na regularnościach obecnych w stosunkowo dużym zbiorze danych. Przeniesienie ciężaru

[15] Rozumiana w sposób proponowany przez Brown i Levinsona (1987).

z tego, co powtarzalne w opinii badacza na to, co powtarzalne ze względu na liczbę wystąpień może stanowić również krok w kierunku zwiększenia intersubiektywności prowadzonych analiz.

Wspomniane tutaj korzyści[16] związane z zastosowaniem metod korpusowych w KAD wydają się szczególnie istotne ze względu na przyjmowane w ramach KAD założenie o społecznej roli prowadzonych analiz. Zwiększenie reprezentatywności badanego materiału, przejrzystość stosowanych metod, koncentracja na tym, co najczęstsze czy też obserwacja wzorców ilościowych prowadzą, jak się wydaje, do podniesienia wiarygodności prezentowanych wyników, szczególnie dla szerszego niż tylko specjaliści KAD grona odbiorców. Takie podnoszenie wiarygodności można z kolei uznać za szczególnie istotne ze względu na przyjmowane w ramach KAD założenie o emancypacyjnej roli prowadzonych badań.

Koszty

Założenie o analizie dużych korpusów tekstów może modyfikować zakres podejmowanych tematów badawczych. Kieruje ono bowiem uwagę badacza na teksty, które są łatwo dostępne w wersji elektronicznej i które umożliwiają kompilację stosunkowo dużych zbiorów danych. Tendencję tę widać na przykład w prowadzeniu znaczącej części omówionych tu analiz na tekstach prasowych (np. Koller 2004; Gabrielatos, Baker 2008; Almeida 2011) czy transkrypcjach debat parlamentarnych (np. Bachmann 2011; Subtirelu 2013). Pominięte w ten sposób zostają pytania o dyskursywne konstrukcje obecne w trudniej dostępnych wypowiedziach, takich jak na przykład kazania wpływowych księży, przemówienia liderów wspólnot lokalnych czy wykłady nauczycieli w szkołach i na uniwersytetach. Koncentracja na dużych zasobach tekstowych dostępnych w formie elektronicznej może więc dwojako ograniczać pole badawcze: z jednej strony ogranicza się źródła, na podstawie których szuka się odpowiedzi na pytania badawcze. Z drugiej: ogranicza się także same stawiane pytania do takich, dla których możliwe jest uzyskanie odpowiedzi wyłącznie na podstawie analizy istniejących już i dostępnych elektronicznie tekstów.

Z kolei skupienie się na wzorcach ilościowych każe zadać pytanie, czy najczęstszy oznacza najważniejszy. Wydaje się, że brak jest przekonywujących dowodów, że to właśnie częstość występowania w dyskursie stanowi czynnik najsilniej wpływający na kształt społecznych i indywidualnych reprezentacji wybranego zjawiska. Pominięta zostaje na przykład kwestia zróżnicowanej społecznie definiowanej istotności tekstów umieszczanych w korpusie. Można przypuszczać, że pewne wypowiedzi, czy to ze względu na ich autorów, czy też kształt samej wypowiedzi (np. szczególną wyrazistość), mogą mieć silniejszy wpływ na konstrukcje dyskursywne obecne wśród odbiorców.

Na uwagę zasługuje także problem wielu zmiennych, które wpływają na uzyskiwane przy pomocy technik korpusowych wyniki. Listy słów kluczowych zależą od trafności doboru zarówno głównego korpusu, jak i korpusu referencyjnego, a także od dostosowania stosowanej miary statystycznej. Uzyskiwane profile kolokacyjne są również zależne od współczynników, przy pomocy których są obliczane, a ich wiarygodność jako źródła informacji dotyczącego konstrukcji dyskursywnych zależy także od trafności wyboru słów, dla których są obliczane. Drobny błąd czy nieścisłość pojawiająca się na którymkolwiek etapie wykorzystania technik korpusowych mogą prowadzić do skrzywienia wyników, które trudno będzie dostrzec zarówno samemu badaczowi, jak i odbiorcom jego badań.

Innym problemem, który związany jest również z samymi technikami korpusowymi, jest występowanie w nich licznych, niewyrażanych wprost i często nieuwzględnianych w interpretacjach wyników założeń. I tak na przykład wykorzystanie techniki słów kluczowych prowadzi do koncentracji na tym, co różne i pominięcia tego, co podobne. Analizie poddane zostają słowa, które najbardziej różnią się częstotliwością w badanych korpusach. Pomija się na przykład te, które występują bardzo często zarówno w korpusie głównym, jak i referencyjnym. Analiza konkordancji dla określonych słów związana jest z rekonstrukcją funkcjonowania w dyskursie wybranego słowa, nie zaś konceptu, jaki reprezentuje, a który może być wrażany także przy pomocy innych słów. Dodatkowo prowadzi ona do uznania najbliższego kontekstu danego słowa za najistotniejszy i decydujący o wynikach prowadzonych badań. Wybór kolokacji jako narzędzia ułatwiającego badanie dyskursu zakłada, że istotna jest tylko częstość współwystępowania, nie zaś inne cechy połączenia między wyrazami. Każde z tych założeń może zostać podane w wątpliwość. Każde z nich wpływa także na zakres i sposób interpretacji uzyskiwanych wyników.

Na koniec warto także wspomnieć o kwestii specyficznej dla analiz prowadzonych na teksach w języku polskim. Opisane powyżej techniki wykorzystywane są przede wszystkim dla danych w języku angielskim, który charakteryzuje się inną strukturą morfosyntaktyczną niż polski: brak odmiany rzeczownika, a także występowanie zaledwie kilku form czasownika ma wpływ na wyniki ilościowych porównań prowadzonych dla słów rozumianych jako określone sekwencje znaków. To samo pojęcie[17] w języku angielskim wyrażane jest przy pomocy zaledwie kilku różnych słów, w polskim zaś kilkunastu lub kilkudziesięciu, co osłabia analityczną siłę narzędzi, takich jak listy kolokacji czy słów kluczowych, w przypadku badań prowadzonych na niezlematyzowanych[18] korpusach. Z kolei lematyzacja pociąga za sobą liczne problemy, jak na przykład możliwości przypisania określonej formy tekstowej do różnych lematów, która z kolei może prowadzić od zniekształcenia oryginalnego tekstu[19].

Większość wspomnianych tu problemów może zostać rozwiązana przez przemyślaną i precyzyjną konstrukcję korpusów, a także uwzględnienie możliwych wątpliwości w opracowaniu i interpretacji wyników. Ich pominięcie może jednak prowadzić do niebezpiecznej sytuacji: uzyskiwane przy pomocy metod korpusowych wyniki stwarzają wrażenie bardzo wiarygodnych ze względu na ilość danych, na których są oparte, a także ograniczenie roli badacza-

[16] Pełniejszy obraz korzyści wynikających z zastosowania technik korpusowych w KAD można znaleźć, sięgając na przykład do Hardt-Mautner (1995), Orpin (2005) czy Bakera (2006).

[17] Przyjmujemy tutaj uproszczony model, w którym jednemu pojęciu odpowiada jeden leksem.

[18] Lematyzacja jest to proces przypisania każdej formie wyrazowej występującej w tekście jej formy podstawowej (lematu).

[19] Bardziej szczegółowe informacje na temat lingwistyczno-technicznych problemów związanych z przetwarzaniem polszczyzny można znaleźć na przykład w Młodzki, Przepiórkowski (2009) czy Głowińska, Przepiórkowski (2010).

-interpretatora na rzecz wykorzystania miar statystycznych. Jednak przez niedociągnięcia w procesie analizowania i interpretowania danych, mogą mieć *de facto* niską wartość poznawczą.

Podsumowanie

Zaprezentowane powyżej metody korpusowe stanowią przegląd technik najczęściej stosowanych w badaniach prowadzonych w ramach KAD. Opisane przykłady badań obrazują wielość pytań badawczych, na które szuka się odpowiedzi przy ich pomocy, a także różnorodność tekstów, do których mogą być stosowane. Przedstawione korzyści mogą zachęcać do stosowania technik korpusowych, zaś koszty wskazywać punkty, w których należy zachować szczególną ostrożność przy projektowaniu badań i interpretacji ich wyników.

Zaproponowany tu przegląd nie ma w żadnym wymiarze wyczerpującego charakteru. W ramach wspieranej korpusowo KAD stosuje się także na przykład analizę wiązek leksykalnych (Herbel-Eisenmann, Wagner 2010) czy automatyczne tagowanie semantyczne wraz z wykorzystaniem opisanych technik dla kategorii semantycznych, nie zaś poszczególnych słów (Prentice 2010). Metody korpusowe stosuje się także do operacjonalizacji kategorii analitycznych obecnych w różnych szkołach w ramach KAD, takich jak krytyczno-historyczna analiza dyskursu Wodak (O'Halloran 2009) czy propozycje van Leeuwena (Kamasa 2013b; Subtirelu 2013). Lista kosztów i korzyści zależy zaś od przyjmowanych założeń filozoficznych i teoretycznych, perspektywy, z której mają być prowadzone badania, a także przyzwyczajeń i przekonań samego badacza.

Bibliografia

Albakry Mohammed (2004) *U.S. "Friendly Fire" Bombing of Canadian Troops: Analysis of the Investigative Reports*. „Critical Inquiry in Language Studies", vol. 1, no. 3, s. 163–178.

Alcaraz-Ariza, María Ángeles (2002) *Evaluation in English-Medium Medical Book Reviews*. „International Journal of English Studies", vol. 2, no. 1, s. 137–153.

Almeida Eugenie P. (2011) *Palestinian and Israeli Voices in Five Years of U.S. Newspaper Discourse*. „International Journal of Communication", vol. 5, s. 1586–1605.

Babbie Earl R. (2003) *Badania społeczne w praktyce*. Przełożyła Agnieszka Kloskowska-Dudzińska. Warszawa: Wydawnictwo Naukowe PWN.

Bachmann Ingo (2011) *Civil partnership – "gay marriage in all but name": a corpus-driven analysis of discourses of same-sex relationships in the UK Parliament*. „Corpora", vol. 6, no. 1, s. 77–105.

Baker Paul (2006) *Using corpora in discourse analysis*. London, New York: Continuum.

Baker Paul i in. (2008) *A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK Press*. „Discourse & Society", vol. 19, no. 3, s. 273–306.

Bondi Marina, Scott Mike (2010) *Keyness in texts*. Amsterdam, Philadelphia: John Benjamins Publishing.

Breeze Ruth (2011) *Critical Discourse Analysis and Its Critics*. „Pragmatics", vol. 21, no. 4, s. 493–525.

Brown Penelope, Levison Stephen C. (1987) *Politeness. Some universals in language usage*. Cambridge: Cambridge University Press.

Chen Lily (2012) *Reporting news in China: Evaluation as an indicator of change in the China Daily*. „China Information", vol. 26, no. 3, s. 303–329.

Degano Chiara (2007) *Dissociation and Presupposition in Discourse: A Corpus Study*. „Argumentation", vol. 21, no. 4, s. 361–378.

Don Zuraidah Mohd, Knowles Gerry, Fatt Choong Kwai (2010) *Nationhood and Malaysian identity: a corpus-based approach*. „Text & Talk – An Interdisciplinary Journal of Language, Discourse & Communication Studies", vol. 30, no. 3, s. 267–287.

Edwards Geraint O.. (2012) *A comparative discourse analysis of the construction of 'in-groups' in the 2005 and 2010 manifestos of the British National Party*. „Discourse & Society", vol. 23, no. 3, s. 245–258.

Fabiszak Małgorzata, Hebda Anna, Konat Barbara (2012) *Dichotomy between private and public experience: The case of Polish wierzyć 'believe'*. „Selected Papers from UK-CLA Meetings", vol. 1, s. 164–176.

Forchtner Bernhard, Kolvraa Christoffer (2012) *Narrating a 'new Europe': From 'bitter past' to self-righteousness?* „Discourse & Society", vol. 23, no. 4, s. 377–400.

Freake Rachelle, Gentil Guillaume, Sheyholislami Jaffer (2010) *A bilingual corpus-assisted discourse study of the construction of nationhood and belonging in Quebec*. „Discourse & Society", vol. 22, no. 1, s. 21–47.

Gabrielatos Costas, Baker Paul (2008) *Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996-2005*. „Journal of English Linguistics", vol. 36, no. 1, s. 5–38.

Gabrielatos Costas., Marchi Anna (2011) *Keyness: Matching metrics to definitions*. Refereat zaprezentowany podczas *Theoretical-methodological challenges in corpus approaches to discourse studies - and some ways of addressing them*, 5 listopada, Portsmouth, Anglia.

Gabrielatos Costas, Marchi Anna (2012) *Keyness: Appropriate metrics and practical issues*. Refereat zaprezentowany podczas *CADS International Conference 2012*, 13–14 września, Bologna, Włochy.

Głowińska Katarzyna, Przepiórkowski Adam (2010) *The Design of Syntactic Annotation Levels in the National Corpus of Polish* [dostęp 15 stycznia 2014 r.]. Dostępny w Internecie: ‹http://nlp.ipipan.waw.pl/~adamp/Papers/2010-lrec-kg/lrec-nkjp_0324.pdf›.

Gregory Ian N. (2008) *Different Places, Different Stories: Infant Mortality Decline in England and Wales, 1851–1911*. „Annals of the Association of American Geographers", vol. 98, no. 4, s. 773–794.

Gries Stefan Thomas (2009) *Quantitative corpus linguistics with R. A practical introduction*. New York: Routledge.

Gries Stefan Thomas (2010) *Useful statistics for corpus linguistics* [w:] Aquilino Sánchez, Moisés Almela, eds., *A mosaic of corpus linguistics: selected approaches*. Frankfurt am Main: Peter Lang, s. 269–291.

Halliday Michael, Matthiessen Christian (2004) *An introduction to functional grammar*. London, New York: Arnold.

Hamilton Craig, Adolphs Svenja, Nerlich Brigitte (2007) *The meanings of 'risk': a view from corpus linguistics*. „Discourse & Society", vol. 18, no. 2, s. 163–181.

Hardt-Mautner Gerlinde (1995) *'Only Connect.' Critical Discourse Analysis and Corpus Linguistics* [dostęp: 20 maja 2012 r.]. Dostępny w Internecie: ‹http://ucrel.lancs.ac.uk/papers/techpaper/vol6.pdf›.

Hebda Anna (2011) *Onde and envy: A diachronic cognitive approach* [w:] Jacek Fisiak, ed., *Studies in Old and Middle English*. Frankfurt am Main: Peter Lang, s. 107–126.

Herbel-Eisenmann Beth, Wagner David (2010) *Appraising lexical bundles in mathematics classroom discourse: obligation and choice*. „Educ Stud Math", vol. 75, no. 1, s. 43–63.

Hidalgo Tenorio Encarnacion (2011) *Critical Discourse Analysis, An overview*. „Nordic Journal of English Studies", vol. 10, no. 1, s. 184–210.

Kamasa Victoria (2013a) *Naming "In Vitro Fertilization": Critical Discourse Analysis of the Polish Catholic Church's Official Do-*

cuments. „Procedia – Social and Behavioral Sciences", vol. 95, s. 154–159.

Kamasa Victoria (2013b) *Rodzina w dyskursie Polskiego Kościoła Katolickiego. Badania korpusowe z perspektywy Krytycznej Analizy Dyskursu*. „Socjolingwistyka", vol. 27, s. 139–152.

Koller Veronika (2004) *Businesswomen and war metaphors: 'Possessive, jealous and pugnacious'?* „Journal of Sociolinguistics", vol. 8, no. 1, s. 3–22.

Lischinsky Alon (2011) *In times of crisis: a corpus approach to the construction of the global financial crisis in annual reports*. „Critical Discourse Studies", vol. 8, no. 3, s. 153–168.

Louw Bill (1993) *Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies* [w:] M. Baker, ed., *Text and Technology*. Amsterdam: John Benjamins, s. 157–176.

Lukac Morana (2011) *Down to the bone: A corpus-based critical discourse analysis of pro-eating disorder blogs*. „Jezikoslovlje", vol. 12.2, s. 187–209.

Marling Raili (2010) *The Intimidating Other: Feminist Critical Discourse Analysis of the Representation of Feminism in Estonian Print Media*. „NORA – Nordic Journal of Feminist and Gender Research", vol. 18, no. 1, s. 7–19.

Mautner Gerlinde (2005) *The Entrepreneurial University*. „Critical Discourse Studies", vol. 2, no. 2, s. 95–120.

------ (2007) *Mining large corpora for social information: The case of elderly*. „Language in Society", vol. 36, no. 1, s. 51–72.

McEnery Tony, Wilson Andrew (2001) *Corpus linguistics. An introduction*. Edinburgh: Edinburgh University Press.

Miechowicz-Mathiasen Katarzyna, Scheffler Paweł (2008) *A corpus-based analysis of the Polish verb podobać się* [w:] Jacek Witkoś, ed., *Elements of Slavic and Germanic grammars: a comparative view. Papers on topical issues in syntax and morphosyntax*. Frankfurt, Berlin, Bern, Brussels, New York, Oxford, Vienna: Peter Lang, s. 89–111.

Młodzki Rafał, Przepiórkowski Adam (2009) *The WSD Development Environment* [w:] Zygmunt Vetulani, ed., *Proceedings of LTC 2009*, s. 185–189 [dostęp 15 stycznia 2014 r.]. Dostępny w Internecie: ‹http://nlp.ipipan.waw.pl/~adamp/Papers/2009--ltc-wsd/ltc-034-mlodzki.pdf›.

Mohamad Subakir Mohd Yasin i in. (2012) *Linguistic Sexism In Qatari Primary Mathematics Textbooks*. „GEMA Online™ Journal of Language Studies", vol. 12, no. 1, s. 53–68.

Mulderrig Jane (2011) *Manufacturing Consent: A corpus-based critical discourse analysis of New Labour's educational governance*. „Educational Philosophy and Theory", vol. 43, no. 6, s. 562–578.

O'Halloran Kieran (2009) *Inferencing and cultural reproduction: a corpus-based critical discourse analysis*. „Text & Talk – An Interdisciplinary Journal of Language, Discourse Communication Studies", vol. 29, no. 1, s. 21–51.

Orpin Debbie (2005) *Corpus Linguistics and Critical Discourse Analysis*. *Examining the ideology of sleaze*. „International Journal of Corpus Linguistics", vol. 10:1, s. 37–61.

Oster Ulrike (2010) *Using corpus methodology for semantic and pragmatic analyses: What can corpora tell us about the linguistic expression of emotions?* „Cognitive Linguistics", vol. 21, no. 4, s. 727–763.

Prentice Sheryl (2010) *Using automated semantic tagging in Critical Discourse Analysis: A case study on Scottish independence from a Scottish nationalist perspective*. „Discourse & Society", vol. 21, no. 4, s. 405–437.

Przepiórkowski Adam i in. (2009) *Narodowy Korpus Języka Polskiego*. „Biuletyn Polskiego Towarzystwa Językoznawczego", t. 65, s. 47–56.

Rayson Paul (2009) Wmatrix: a web-based corpus processing environment, Computing Department, Lancaster University, dostępny na stronie ‹http://ucrel.lancs.ac.uk/wmatrix›.

Rogers Rebecca i in. (2005) *Critical Discourse Analysis in Education: A Review of the Literature*. „Review of Educational Research", vol. 75, no. 3, s. 365–416.

Salama Amir H.Y. (2011) *Ideological collocation and the recontexualization of Wahhabi-Saudi Islam post-9/11: A synergy of corpus linguistics and critical discourse analysis*. „Discourse & Society", vol. 22, no. 3, s. 315–342.

Schmidt Filip, Skowrońska Marta (2006) *Człowiek w sieci przedmiotów. Socjologiczna analiza roli i znaczenia przedmiotów w przestrzeni domowej* [w:] Jacek Kowalewski, Wojciech Piasek, Marek Śliwa, red., *Rzeczy i ludzie. Humanistyka wobec materialności*. Olsztyn: Colloquia Humaniorum, s. 197–222.

Scott Mike (2013a) *WordSmith Tools Help* [dostęp 20 sierpnia 2013 r.]. Dostępny w Internecie: ‹http://www.lexically.net/downloads/version6/HTML/index.html?keywords_info.htm›.

------ (2013b) *WordSmith Tools*. Liverpool: Lexical Analysis Software.

Stubbs Michael (1997) *Whorf's Children: Critical comments on Critical Discourse Analysis (CDA)* [w:] Ann Ryan, Alison Wray, eds., *Evolving models of language. Papers from the annual meeting of the British Association for Applied Linguistics held at the University of Wales, Swansea, September 1996*. Clevedon: British Association for Applied Linguistics, s. 100–116.

------ (2001)*Words and phrases. Corpus studies of lexical semantics*. Oxford, Malden: Blackwell Publishers.

Subtirelu Nicholas Close (2013) *'English… it's part of our blood': Ideologies of language and nation in United States Congressional discourse*. „Journal of Sociolinguistics", vol. 17, no. 1, s. 37–65.

van Dijk Teun A. (2001) *Multidisciplinary CDA: a plea for diversity* [w:] Ruth Wodak, Michael Meyer, eds., *Methods of critical discourse analysis*. London: Sage, s. 95–120.

van Leeuwen Theo (2008) *Discourse and practice. New tools for critical discourse analysis*. Oxford, New York: Oxford University Press.

Weninger Csilla (2010) *The lexico-grammar of partnerships: corpus patterns of facilitated agency*. „Text & Talk – An Interdisciplinary Journal of Language, Discourse & Communication Studies", vol. 30, no. 5, s. 591–613.

Whitsitt Sam (2005) *A critique of the concept of semantic prosody*. „International Journal of Corpus Linguistics", vol. 10, no. 3, s. 283–305.

Wodak Ruth (2001) *The discourse-historical approach* [w:] Ruth Wodak, Michael Meyer, eds., *Methods of critical discourse analysis*. London: Sage, s. 63–94.

Cytowanie

Kamasa Victoria (2014) *Techniki językoznawstwa korpusowego wykorzystywane w krytycznej analizie dyskursu. Przegląd*. „Przegląd Socjologii Jakościowej", t. 10, nr 2, s. 100–117 [dostęp dzień, miesiąc, rok]. Dostępny w Internecie: ‹www.przegladsocjologiijakosciowej.org›.

Corpus Linguistics Techniques Used for Critical Discourse Analysis. An overview

Abstract: The paper aims to present corpus methods most commonly used in Critical Discourse Analysis (CDA). The issues of corpus design for CDA will be discussed and methods frequently used in such analysis will be presented: frequency lists, keywords, collocations, and concordances. Moreover, examples of research using this methods will be overviewed to provide an account of the variety of subjects and conclusions the discussed methods might lead to. The paper will conclude with some remarks on benefits and costs related to the usage of corpus methods in CDA.

Keywords: Critical Discourse Analysis, corpus methods, collocation, key words, collocations, concordances