Member since 2019
JM14475

Badania dyskursu wspomagane korpusowo (CADS) jako wsparcie jakościowej analizy treści. Studium przypadku wykorzystania programu SketchEngine w badaniach dyskursu

Marek Troszyński
Collegium Civitas w Warszawie

https://doi.org/10.18778/1733-8069.20.4.03

Słowa kluczowe: lingwistyka korpusowa, SketchEngine, jakościowa analiza treści, metody mieszane

Abstrakt: Artykuł przedstawia możliwość wykorzystania narzędzi lingwistyki korpusowej jako pierwszego etapu jakościowej analizy treści. W tekście omówiony jest rozwój metody badań dyskursu wspomaganych korpusem (CADS). Zasadnicza część artykułu to omówienie funkcji wybranego programu wspomagających CADS – SketchEngine. W tekście znajdziemy liczne przykłady, które objaśniają sposoby wykorzystania metod CADS i funkcjonalności SketchEngine dla analizy polskiego dyskursu prasowego. Dzięki możliwości łatwego odniesienia do tekstów źródłowych (konkordancje) SketchEngine pozwala na włączenie metod mieszanych do badań dyskursu.

Marek Troszyński
Doktor, socjolog, prowadzi Obserwatorium Cywilizacji Cyfrowej w Collegium Civitas w Warszawie. Bada dyskurs medialny dotyczący migrantów i uchodźców w Polsce oraz język komunikatów z wojny w Ukrainie. W pracy naukowej zajmuje się także zagadnieniami mowy nienawiści wobec mniejszości w Polsce. W badaniach wykorzystuje metody lingwistyki korpusowej (CL) oraz narzędzia automatycznej analizy języka naturalnego (NLP).
e-mail: mtroszynski@civitras.edu.pl

Wprowadzenie

Badacze zajmujący się analizą treści medialnych stoją przed problemem – jak zebrać i poddać analizie powiększający się nieustannie strumień komunikatów. Przy badaniu obszernych zbiorów tekstów, np. artykułów prasowych zbieranych na przestrzeni miesięcy czy systematycznie zapisywanych treści z portali internetowych, pojawia się problem doboru tekstów do analizy. Mówiąc najprościej – liczba i objętość komunikatów medialnych dawno już przekroczyły możliwości tradycyjnie wykonywanej analizy jakościowej, nawet jeśli zaangażujemy w proces dużą liczbę osób. Ponadto współcześni badacze dyskursu są zainteresowani intertekstualnością, kontekstem, w którym komunikaty są dostępne czytelnikowi (Krippendorff, 2004: XX), co dodatkowo zwiększa objętość analizowanego materiału. Coraz trudniejszym zadaniem staje się dobór tekstów do analizy, w szczególności w badaniach szeroko rozumianego dyskursu medialnego.

Dlatego aby zestawić próbę, która jest możliwa do przeanalizowania w ramach jednego projektu badawczego, stosowane są różne metody selekcji tekstów. Możemy sięgnąć po znane z metodologii sondażowej dobory: losowy, systematyczny, warstwowy. Ten sposób budowy próby jest skuteczny, jeśli wszystkie teksty uznajemy za tak samo informatywne z perspektywy pytań badawczych, podobnie jak ma to miejsce, gdy badamy wybraną populację ludzi (Krippendorff, 2004: 115).

W analizie treści dysponujemy metodami, które pozwalają na wstępne określenie istotności danego tekstu dla pytań badawczych. Możemy na przykład przeszukiwać zebrane teksty na podstawie zdefiniowanych wcześniej słów kluczy. Warto jednak posłużyć się bardziej wyrafinowanymi metodami selekcji tekstów. Do tego zadania możemy wykorzystać metody zapożyczone z lingwistyki korpusowej.

Celem tego artykułu jest pokazanie możliwości wykorzystania metod i narzędzi lingwistyki korpusowej w badaniach treści komunikatów medialnych. Sama metoda ma już długą tradycję, w szczególności w języku angielskim (Hunston, 2002), w ostatnich dwóch dekadach dodatkowo rozwinęły się możliwości wsparcia analiz poprzez użycie narzędzi komputerowych (Kilgarriff i in., 2014). Popularność wykorzystania lingwistyki korpusowej w badaniach treści wzrosła w XXI wieku (Fairclough, 2000; Piper, 2000; Baker, 2006). W proponowanym podejściu ilościowe metody lingwistyki korpusowej chcę potraktować jako narzędzia wspierające jakościową analizę treści. Przede wszystkim tak rozumiana ilościowa analiza pozwala na merytoryczny (bazujący na pytaniach badawczych) dobór tekstów do dalszych analiz. Posługując się przeznaczonymi do tego programami komputerowymi (takimi jak omawiany tu SketchEngine), możemy wyselekcjonować teksty (fragmenty tekstów), które reprezentują powtarzalne wzorce treści lub teksty specyficzne, odmienne w danym obszarze tematycznym. Tak właśnie rozumie rolę analizy korpusowej w badaniach dyskursu Paul Baker, pisząc: „pozwala [ona – przyp. M.T.] badaczom zidentyfikować mniej lub bardziej obiektywnie rozpowszechnione wzorce naturalnie występującego języka i rzadkie przypadki, z których oba mogą zostać przeoczone w analizie na małą skalę" (Baker, 2004: 346).

Lingwistyka korpusowa w badaniach dyskursu

Początki językoznawstwa korpusowego to druga połowa XIX wieku (Baker, 2006: 2), ale do drugiej połowy XX wieku rozwijało się ono w obszarze analiz leksykograficznych i badań pomocnych w nauce języków. Dopiero w latach dziewięćdziesiątych XX wieku zaczęto wykorzystywać lingwistykę korpusową do badań tekstów medialnych, np. w próbach opisania różnic pomiędzy kulturami na podstawie korpusów zawierających teksty prasowe:

> […] możemy zaproponować obraz kultury amerykańskiej z 1961 r. – męskiej do punktu machismo, militarystycznej, dynamicznej i napędzanej przez wzniosłe ideały, napędzanej przez technologię, aktywność i przedsiębiorczość – kontrastujący z kulturą brytyjską jako bardziej skłonną do grania na zwłokę i gadania, czerpanie korzyści z bogactwa, a nie jego tworzenia, a także życie rodzinne i uczuciowe, mniej zależne od spraw merytorycznych niż od względów statusu zewnętrznego (Leech, Fallon, 1992: 44).

W analizie zawartości prasy zaczęto wykorzystywać połączenie metod krytycznej analizy dyskursu (CDA) z metodami lingwistyki korpusowej (Hardt-Mautner, 1995) lub krytyki CDA z pozycji lingwistyki korpusowej, wskazując ewidentne słabości tej pierwszej (np. brak jasno określonych zasad wyboru tekstów, które są poddawana analizie) (Stubbs, 1997).

Kamieniem milowym w rozwoju metody była publikacja książki *Using corpora in discourse analysis* (Baker, 2006). Paul Baker pisze o użyciu:

> […] *korpusów* (dużych zbiorów naturalnie występujących danych językowych przechowywanych na komputerach) i procesów na korpusach (procedur obliczeniowych, które manipulują tymi danymi na różne sposoby) w celu odkrycia wzorców językowych, które mogą umożliwić nam zrozumienie sposobów, poprzez które język jest wykorzystywany w konstruowaniu dyskursów (czyli sposobów konstruowania rzeczywistości) (Baker, 2006: 1).

Książka ta jest świadomą próbą połączenia dwóch wcześniej rozłącznych dziedzin wiedzy – lingwistyki korpusowej i badań nad dyskursem. Autor daje systematyczny wykład opisujący łączenie obu metod. Zaczyna od sposobów budowania korpusów, a w kolejnych rozdziałach wprowadza pojęcia z obszaru lingwistyki korpusowej („frekwencja", „konkordancja", „kolokacje", „słowa kluczowe"), ilustrując je przykładami z własnych badań opierających się na różnych korpusach. Późniejsze jego teksty (Baker i in., 2008; Gabrielatos, Baker, 2008), bazujące na projekcie badawczym „Discourses of Refugees and Asylum Seekers in the UK Press 1996–2006", rozwijają prezentowane metody. I ponownie kwestia połączenia metod korpusowych z CDA jest wyróżniającą cechą przyjętej metodologii analizy tekstów prasowych.

Wielu innych badaczy wykorzystuje metody CADS czy szerzej lingwistyki korpusowej do analizy treści mediów. Warto tu wspomnieć prace Moniki Bednarek analizujące dyskurs prasowy w kontekście

opinii dziennikarskich (Bednarek, 2006; Potts, Bednarek, Caple, 2015) czy Ibrahima Efe (2019), który analizuje dyskurs dotyczący uchodźców z Syrii w prasie tureckiej.

W 2023 roku pojawiła się kolejna znacząca publikacja, której autorzy podsumowują dotychczasowe wykorzystanie lingwistyki korpusowej jako narzędzia analizy dyskursu publicznego. W podręczniku do analizy dyskursu, który ukazał się w serii „Cambridge Elements", autorzy posługują się terminem *Corpus-Assisted Discourse Studies* (CADS), który chcę przyjąć na potrzeby tego tekstu:

> Badania dyskursu wspomagane korpusem (*Corpus-Assisted Discourse Studies*, CADS) badają dyskurs (tj. język jako praktykę społeczną) poprzez badanie korpusów (tj. dużych cyfrowych zbiorów danych tekstowych). CADS pozwala zbadać korpus jako całość, zamiast skupiać się tylko na niektórych tekstach, które przypadkowo lub celowo mogą być tymi, które potwierdzają to, co od dawna chcieliśmy pokazać (Gillings, Mautner, Baker, 2023: 1).

Seria „Cambridge Elements" zawiera wiele pozycji książkowych opisujących w skondensowany sposób perspektywę lingwistyki korpusowej, jak choćby flagowa pozycja *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User* (Egbert, Larsson, Biber, 2020).

Podsumowując cały proces wypracowywania metody, można powiedzieć, że kluczowym momentem było postawienie pytań w ramach analiz lingwistycznych, które odnoszą się do społecznego kontekstu wypowiedzi. Inaczej mówiąc, lingwistyka korpusowa zostaje użyta do analizy problemów badawczych podejmowanych przez krytyczną analizę dyskursu. Ma ją uzupełniać, a przede wszystkim zapewniać narzędzia pozwalające na uniknięcie wybiórczego charakteru analiz treści.

Krytyczna analiza dyskursu to uznana metoda (por. np. Krzyżanowski, Forchtner, 2016), od 30 lat praktykowana przez badaczy, ma jednak swoje ograniczenia. Kluczowa z perspektywy tego tekstu jest konieczność przeczytania i zinterpretowania przez badacza, często wielokrotnie, każdego analizowanego tekstu. W zasadniczy sposób ograniczy to liczbę tekstów, jakie możemy przebadać:

> I w tym tkwi problem. Realistycznie rzecz biorąc, uważne czytanie i gęsty opis są możliwe tylko wtedy, gdy korpus jest dość mały: powiedzmy kilka artykułów prasowych lub garść transkrypcji (Gillings, Mautner, Baker, 2023: 6).

Nawet szybki rozwój oprogramowania CAQDAS (por. np. Costa i in., 2023) nie przełamuje tej granicy, choć znacząco pomaga w organizacji zbioru tekstów i przede wszystkim usprawnia kodowania i analizy wybranych fragmentów. Jednak nadal trudno mówić o reprezentatywności przeprowadzanych analiz, które ograniczają się często do kilkudziesięciu artykułów prasowych.

Specyfika CADS pozwala przeanalizować duże zbiory danych z pominięciem omówionej wcześniej kwestii doboru próby. To z kolei umożliwia wychwycenie w zbiorze tekstów prawidłowości, które mogą

stanowić punkt wyjścia do jakościowych analiz. Jeśli zatem chcemy analizować dyskurs medialny, z charakterystyczną dla niego nadprodukcją tekstów, to warto wykorzystać narzędzia CADS.

Co konkretnie CADS zmienia w procesie analizy treści? Jakie są zalety i wady wykorzystania tej metody? Podsumowania zalet opisywanej metody znajdziemy w cytowanej już książce Bakera (2006: 19):

1) zmniejszenie stronniczości badacza – oczywiście, jak w każdym badaniu szukamy odpowiedzi na pytania badawcze, przyjmując określoną perspektywę teoretyczną; jednak odwołanie się do zobiektywizowanych wzorców występowania słów poprzez porównanie ich z korpusem referencyjnym pozwala na niezależne od poglądów badacza opisanie specyfiki badanego korpusu;
2) narastający efekt dyskursu – powtarzane wzorce sprawiają, że analizowane znaczenie nie odnosi się do jednego tekstu, a staje się rzeczywistą analizą dyskursu;
3) uwzględnienie zmienności dyskursu – dyskurs jest zmienny w czasie, możemy to zobaczyć dopiero wtedy, kiedy porównujemy różne korpusy tekstów;
4) triangulacja – korpus (i metody korpusowe) stanowią punkt odniesienia dla analiz jakościowych na małą skalę.

Wszystko, co odkrywamy w procesie analizy dyskursu, bazuje na naszej interpretacji tekstów, na kategoryzacjach i nadawaniu znaczenia. Natomiast w CADS szukamy zobiektywizowanych wzorców językowych. Należy również wziąć pod uwagę, jaka jest siła oddziaływania danego tekstu. Do ilu odbiorców on dotrze, jakim autorytetem cieszy się autor, jaki ma wpływ na odbiorców (Baker, 2006: 19).

W innym tekście autorzy zwracają uwagę, że podejście korpusowe może umożliwić badaczowi oddalenie się od danych i lepsze zrozumienie znaczeń i funkcji niektórych wyborów leksykalnych dokonywanych np. w tekstach dotyczących uchodźców. Badanie dyskursu wspomagane korpusem daje nam możliwość swobodnej zmiany perspektywy od całego korpusu po pojedynczą wypowiedź (Baker, Mcenery, 2005: 223). Istotną zaletą stosowania metod lingwistyki korpusowej jest również transparentność procesu badawczego, która pozwala innym naukowcom sprawdzić przeprowadzone analizy krok po kroku (Zawadzka-Paluektau, 2023: 100).

Programy komputerowe wspomagające CADS

Zanim przejdę do omówienia sposobów wykorzystania SketchEngine w analizie treści, chcę pokazać inne programy komputerowe bazujące na metodach lingwistyki korpusowej. We wspomnianej już książce *Corpus-Assisted Discourse Studies* (Gillings, Mautner, Baker, 2023) znajdziemy przegląd oprogramowania, które pozwala na przeprowadzanie analiz tekstów opartych na metodach lingwistyki korpusowej. Ogólnie rzecz biorąc, większość programów do analizy korpusowej oferuje te same narzędzia: częstość, konkordancje, kolokacje, słowa kluczowe. Różnice występują na poziomie interfejsu oraz możliwości konfigurowania poszczególnych funkcji:

1) **WordSmith Tools** (https://www.lexically.net/wordsmith/) – program stworzony w 1996 roku przez Lexical Analysis Software i Oxford University Press; jest płatny, dostępny w wersji do zainstalowania na lokalnym komputerze;
2) **AntConc** (https://www.laurenceanthony.net/software/antconc/) – darmowy program do zainstalowania na lokalnym komputerze, umożliwia korzystanie z preinstalowanych korpusów, jak również włączenie własnych zbiorów; do dyspozycji mamy kilka dodatkowych darmowych programów wykonujących poszczególne funkcje pomocne przy analizie korpusowej – generowanie własnych korpusów (AntCorGen), tworzenie n-gramów (AntGram), analiza równoległych korpusów (AntPConc) i inne.

Trzy kolejne narzędzia powstały i są rozwijane na Uniwersytecie Lancaster – miejscu pracy wielu uznanych badaczy zajmujących się lingwistyką korpusową:

1) **#LancsBox** (http://corpora.lancs.ac.uk/lancsbox/) – darmowy program do pobrania na komputer, umożliwiający pracę na zainstalowanych lub własnych korpusach; obsługuje wiele języków, choć nie dla wszystkich oferuje oznaczanie części mowy (*POS annotation*);
2) **CQPWeb** (https://cqpweb.lancs.ac.uk/) – program w wersji usługi sieciowej; darmowy serwis, w którym po zarejestrowaniu możemy korzystać z wielu przygotowanych korpusów (w tym również korpusów historycznych wersji języka angielskiego), jak również wgrać własny korpus;
3) **Wmatrix** (https://ucrel.lancs.ac.uk/wmatrix/) – program można uruchomić jako usługę sieciową lub zainstalować go na lokalnym komputerze; darmowy tylko dla studentów i absolwentów Uniwersytetu Lancaster.

Istotną pomoc dla polskich badaczy dyskursu stanowić mogą zasoby technologii językowych tworzonych w ramach projektu CLARIN-PL (Piasecki, 2014): „celem CLARIN-PL jest dostarczenie polskim naukowcom wszelkich środków potrzebnych do prowadzenia badań wymagających wykorzystania metod przetwarzania dużych ilości tekstów" (CLARIN-PL, b.r.). Z perspektywy użytkownika mamy dostęp online do wielu narzędzi (darmowych i otwartych), które możemy swobodnie wykorzystywać w projektach naukowych. CLARIN-PL oferuje narzędzia niezbędne do rozpoczęcia pracy z każdym tekstem w języku polskim: analizator morfosyntaktyczny, parser, narzędzia do rozpoznawania nazw własnych, narzędzia do anotacji czasowników (CLARIN-PL, b.r.). Pozwala również na bardziej złożone przetwarzanie tekstu, np. analizę wydźwięku, streszczenia tekstu, wykrywanie mowy nienawiści, modelowanie tematyczne (CLARIN-PL, b.r., Login).

Szczególnie przydatne do badań dyskursu wspieranych korpusem jest narzędzie Korpusomat (b.r.). Jest to aplikacja sieciowa pozwalająca na gromadzenie i znakowanie (anotowanie) korpusów językowych zbudowanych z tekstów dostarczonych przez użytkownika (Kieraś, Kobyliński, Ogrodniczuk, 2018; Kieraś, Kobyliński, 2021). Korpusomat pozwala na automatyczne anotowanie informacji fleksyjnej (według klasyfikacji stosowanej w Narodowym Korpusie Języka Polskiego), jednostek nazewniczych, struktury składniowej, a także na podsumowania statystyczne korpusów, np. rozkład słów kluczowych czy ekstrakcję terminologii (Kieraś, Kobyliński, 2021: 54). Ważnym uzupełnieniem

dla badaczy dyskursu w perspektywie porównawczej, np. analiz treści prasy w różnych krajach, jest wersja Korpusomatu pozwalająca na automatyczne anotacje dla 29 różnych języków (Korpusomat (Beta), b.r.; Saputa i in., 2023).

SketchEngine – funkcje wykorzystywane w badaniach CADS

SketchEngine jest usługą sieciową, udostępnianą przez Lexical Computing, firmę założoną przez Adama Kilgarriffa (SketchEngine, b.r.). Aplikacja oferuje 600 korpusów i obsługę 90 języków, w tym polskiego. Jest programem płatnym, do korzystania jako usługa online. Płatna jest nie tylko możliwość korzystania z narzędzi analitycznych, musimy również wykupić miejsce na analizowane korpusy (liczone w milionach słów).

SketchEngine to nie tylko narzędzie dla lingwistów, powszechne jest jego wykorzystanie w projektach analizujących różne rodzaje dyskursu: „analiza określonego rodzaju języka pod kątem tego, co mówi nam o postawach, stosunkach władzy i perspektywach uczestników. Ten rodzaj pracy odbywa się na różnych wydziałach nauk humanistycznych i społecznych" (Kilgarriff i in., 2014: 15).

Dla ilustracji funkcji wykorzystywanych w SketchEngine prezentuję dane z badania „Obraz wojny rosyjsko-ukraińskiej w polskim dyskursie medialnym", realizowanego w Collegium Civitas. Analizie poddano artykuły z polskiej prasy (dzienników), zebrane według następujących kryteriów: daty: 24 lutego 2022 – 30 kwietnia 2022; artykuły prasowe (dzienniki), źródło: Newspoint, słowa klucze: "Ukrain*" OR "Rosj*" OR "Białoru*" OR "uchodźc*" OR "migran*" OR "imigran*" OR "Putin*" OR "Zelenski*" OR "Lukaszen*" OR "Łukaszen*".

Pozwoliło to na budowę korpusu składającego się z 3,8 mln słów:

1) „Gazeta Wyborcza", 1613 artykułów, 1075 tys. słów,
2) „Rzeczpospolita", 1550 artykułów, 1029 tys. słów,
3) „Dziennik Gazeta Prawna", 963 artykuły, 812 tys. słów,
4) „Gazeta Polska Codziennie", 874 artykuły, 344 tys. słów,
5) „Super Express", 463 artykuły + „Fakt", 423 artykuły, łącznie 267 tys. słów,
6) „Nasz Dziennik", 453 artykuły, 310 tys. słów.

Powyższe liczby podaję, aby zobrazować wielkość korpusu składającego się z artykułów prasowych. Mamy bardzo popularny w mediach temat i 6 dzienników w okresie 6 tygodni. Funkcje SketchEngine omawiam na korpusach przygotowanych dla poszczególnych gazet, zebranie ich razem umożliwia porównanie i pokazanie specyfiki tytułów.

Budowa korpusu tekstów

Korpus w znaczeniu ogólnym odnosi się do „zbiorów tekstów, które są przechowywane i dostępne elektronicznie" (Hunston, 2002: 2) i zwykle są przeznaczone do pewnych celów językowych, np. analizy kryminalistyczne, pedagogiczne lub ideologiczne.

Ważne dla przygotowania analiz językowych jest zapewnienie dla każdego języka naturalnego dużego, aktualnego, ogólnego korpusu językowego, przetworzonego za pomocą narzędzi dla tego języka, ze szkicami słów (Word Sketch). Taki korpus będziemy traktować jako korpus referencyjny, pozwalający na pokazanie specyfiki języka wykorzystywanego w analizowanych przez nas dyskursach. Warunkiem wstępnym dla analizy podstawowego zasobu dla języka jest stworzenie korpusu oraz narzędzi do segmentacji. Korpus można pobrać z sieci, korzystając z metody Corpus Factory (Kilgarriff i in., 2010) lub TenTen (Jakubíček i in., 2013).

W SketchEngine dla języka polskiego mamy domyślnie ustawiony korpus „Polish Web 2019 (plTenTen19)" zebrany metodą TenTen. Korpus ten liczy ponad 4 miliardy słów.

W większości przypadków wykorzystywane korpusy to korpusy internetowe, ponieważ sieć jest jedynym miejscem, w którym można uzyskać duże zbiory słów, obejmujące szeroki zakres typów i dziedzin tekstu. Te korpusy można aktualizować, indeksując ponownie i dodając nowy materiał (Kilgarriff i in., 2014: 23).

Oprócz wstępnie załadowanych korpusów (zarządzanych przez zespół SketchEngine) użytkownicy mogą przesyłać, budować i przetwarzać własne korpusy. Jeśli użytkownik ma już korpus, może go przesłać i zainstalować za pomocą interfejsu internetowego. Jeśli dane są już opatrzone adnotacjami (np. częściami mowy), to muszą być w formacie wejściowym SketchEngine. Użytkownik może następnie zarządzać własnymi korpusami, w tym dodawać więcej danych, usuwać i przetwarzać je, a także wykorzystywać do swoich badań za pośrednictwem podstawowych funkcji SketchEngine (Kilgarriff i in., 2014: 26).

W praktyce badań dyskursu medialnego własne korpusy to zbiory tekstów z danego medium (prasy, portali internetowych) z określonego przedziału czasowego lub o jednorodnej tematyce wyznaczonej przez słowa kluczowe wykorzystane przy kwalifikowaniu poszczególnych tekstów do tego zbioru. W tej perspektywie celowo pomijamy lingwistyczne rozumienie budowy korpusu (Kilgarriff i in., 2014: 33).

Kolejnym krokiem w przygotowaniu korpusu do analiz jest tokenizacja, czyli wyodrębnienie, w uproszczeniu, poszczególnych słów w korpusie. W prostych przypadkach można używać spacji między słowami, ale wiele języków ma wyrazy specjalne, które wymagają specyficznego traktowania językowego. Tokenizację możemy traktować jako „proces dzielenia słów według wcześniej określonej procedury" (Gillings, Mautner, Baker, 2023: 15). SketchEngine definiuje token jako najmniejszą jednostkę, z której składa się korpus. Tokeny można podzielić na słowa i niesłowa. Pierwsza kategoria

odnosi się do tokenów, które zaczynają się na literę alfabetu, drugiej kategorii używamy, gdy token zaczyna się od innego znaku (np. znak interpunkcyjny lub cyfra) (Gillings, Mautner, Baker, 2023).

Bardzo ważną dla dalszych analiz czynnością jest anotowanie korpusu, czyli tagowanie: „dodawanie do każdego słowa dodatkowych informacji, takich jak przypisywanie kategorii gramatycznej lub semantycznej" (Gillings, Mautner, Baker, 2023: 15). Korpusy języka polskiego używane w SketchEngine są anotowane za pomocą narzędzia RFTagger, wykorzystującego tagset (zbiór kategorii gramatycznych) Narodowego Korpusu Języka Polskiego (NKJP) (Przepiórkowski, 2009). Na liście tagsetu NKJP znajduje się 36 klas gramatycznych rozmieszczonych w przybliżeniu według najczęściej używanych (tradycyjnych) części mowy oraz 13 kategorii gramatycznych z ich możliwymi wartościami. Każda klasa gramatyczna ma różne kategorie gramatyczne, które mogą być określone jako obowiązkowe lub fakultatywne dla danej klasy. Właściwy znacznik zawiera kategorie gramatyczne oddzielone dwukropkiem.

Tagowanie przez oznaczenie części mowy (ang. *POS tagging*) jest powszechną praktyką stosowaną dla języka angielskiego:

> Obecnie dość powszechne jest, że korpusy poddawane są automatycznemu – i rzeczywiście bardzo niezawodnemu – gramatycznemu znakowaniu części mowy (*Part-of-Speech Tagging*). CLAWS, jeden z najczęściej używanych znaczników, ma dokładność 96–97 procent, przy czym dokładny stopień dokładności różni się w zależności od rodzaju tekstu (Gillings, Mautner, Baker, 2023: 15).

To samo działanie stanowi wyzwanie dla języków słowiańskich, w tym polskiego. Zasadniczy kłopot wynika bowiem z ich fleksyjności. Do tego dochodzi swobodny szyk wyrazów, który stanowi kolejną trudność w tagowaniu. Dlatego kluczowa dla wspomnianego wcześniej znakowania części mowy (ang. *POS tagging*) jest lematyzacja, czyli sprowadzanie wyodrębnionych tokenów do ich podstawowych form, nazywanych także formami słownikowymi. W SketchEngine użyty został TaKIPI tagger (Radziszewski, Kilgarriff, Lew, 2011: 2), wykorzystujący tagset stworzony w IPI PAN.

> Ostateczna dokładność zmodyfikowanego TaKIPI, tj. 93,53% dla wszystkich znaczników i 86,57% dla niejednoznacznych, jest wciąż znacznie poniżej wyników uzyskanych dla czeskiego (95,16% dla wszystkich tokenów) i angielskiego (ponad 97%). Oznacza to, że TaKIPI popełnia średnio jeden błąd w zdaniu, ale błąd w klasach gramatycznych jest znacznie mniejszy (Piasecki, 2007: 166).

Prawidłowo przeprowadzona lematyzacja ma bardzo duże znaczenie w analizach CADS. Przykładowo, proste zestawienie frekwencji w danym korpusie możemy dzięki temu opisać na poziomie lematów, a nie wyrazów (np. „duży", „większy", „największy" to ten sam lemat, a różne wyrazy).

Word Sketch

Word Sketch jest to „jednostronicowe podsumowanie" wybranego przez nas słowa, program zestawia kolokacje, wskazując na specyficzne dla naszego korpusu użycia wybranego wyrazu.

> Wyniki są podzielone na kategorie, zwane relacjami gramatycznymi, takie jak słowa, które służą jako dopełnienie czasownika, słowa, które służą jako podmiot czasownika, słowa, które modyfikują słowo itp. Niektóre relacje gramatyczne mogą wyświetlać statystyki użycia zamiast kolokacji (SketchEngine, b.r.).

„Szkice słów to jednostronicowe, automatyczne, korpusowe podsumowania relacji gramatycznych i kolokacyjnych słowa" (Radziszewski, Kilgarriff, Lew, 2011: 1). To szeroki zbiór informacji o tym słowie (Kilgarriff i in., 2014: 9).

> Szkic słowa może być postrzegany jako szkic słownika. System przebrnął przez korpus, aby znaleźć wszystkie powtarzające się wzorce słowa i uporządkował je, gotowe do edycji, wyjaśnienia i opublikowania przez leksykografa. Te szkice słów były używane od czasu ich pierwszego wyprodukowania (Kilgarriff i in., 2014: 10).

Dla języka polskiego mamy w SketchEngine zdefiniowane trzy typy relacji gramatycznych (Radziszewski, Kilgarriff, Lew, 2011: 2):

1) relacje symetryczne, między dwoma rzeczami o równym statusie;
2) relacje podwójne, między dwoma zależnymi przedmiotami;
3) relacje trójskładnikowe, obejmujący trzy zależne elementy.

Najczęściej spotykane są relacje podwójne. W programie zostało zdefiniowanych 14 z nich, w tym relacje rzeczownik – modyfikator, podmiot – czasownik i czasownik – dopełnienie. Word Sketch pozwala nam podać zwięzłe definicje relacji czasownik – dopełnienie, w których kolejność słów nie jest ustalona. Uzyskujemy zestawienie obejmujące dowolny czasownik następujący po oknie do sześciu elementów po prawej stronie. Relacje trójstronne wskazują relacje między trzema bytami. W gramatyce języka polskiego służą do wyodrębniania wzorców, w których rzeczowniki i czasowniki łączą się ze zwrotami przyimkowymi (Radziszewski, Kilgarriff, Lew, 2011).

Na rysunku 1 przedstawiono Word Sketch dla lematu „Ukraina", zbudowany na korpusie tekstów pochodzących z dziennika „Rzeczpospolita". Dla każdej relacji zostały wybrane dwie miary: częstotliwość (ile razy dana kolokacja występuje w korpusie) i wynik pokazujący siłę kolokacji. Ta druga wartość oparta jest na mierze statystycznej logDice (Matytcina, Grigoryanova, 2022), pozwalającej zmierzyć współwystępowanie dwóch obiektów. Miara ta jest niezależna od wielkości korpusu, co pozwala na porównania pomiędzy korpusami. Miara ta sprawdza się również do porównań siły kolokacji pomiędzy subkorpusami (Rychlý, 2008).

W prezentowanym przykładzie mamy wybrane cztery relacje: rzeczownik – modyfikator, czasownik poprzedzający, czasownik występujący po, współrzędność (współwystępowanie). W poszczególnych tabelach znajdujemy słowo wchodzące w skład kolokacji oraz „najdłuższe-najczęstsze dopasowanie" (ang. *the longest-commonest match* – LCM), czyli fragment, w którym kolokacja występuje najczęściej (Kilgarriff i in., 2015). Dopasowanie to pozwala nam intuicyjnie zrozumieć, jak zachowuje się dana kolokacja.

Rysunek 1. Fragment Word Sketch dla lematu „Ukraina" zbudowany na korpusie artykułów z dziennika „Rzeczpospolita"

Źródło: badania własne.

Patrząc na wyniki przedstawione na rysunku 1, widzimy, że w tekstach w tym korpusie znacznie częściej niż w korpusie referencyjnym pojawia się zestawienie „niepodległa Ukraina". W zestawieniu z czasownikami mamy „Rosja zaatakowała Ukrainę", „wspierać Ukrainę", „Rosja najechała Ukrainę". Najliczniejszą kolokacją jest Ukraina + mieć, jednak miara logDice jasno pokazuje nam, że siła tej kolokacji nie jest większa od innych w tym zestawieniu. Mamy tu przykład sytuacji, gdy duża liczba wystąpień kolokacji rzeczownika i czasownika nie wynika ze specyfiki naszego korpusu, jest natomiast zasadniczo charakterystyczna dla języka polskiego.

W zestawieniach dwóch rzeczowników dominują „Ukraina i Rosja", „Ukraina i Białoruś", „Polska i Ukraina". Te właśnie relacje, wygenerowane automatycznie przez SketchEngine, stanowić powinny punkt wyjścia do dalszych analiz. Nie określają nam one jednoznacznie znaczeń danych fragmentów tekstu, pozwalają jednak na wyselekcjonowanie tych właśnie fragmentów do analizy jakościowej. Dla każdej relacji mamy w interfejsie programu symbol trzech kropek, który po kliknięciu otwiera nam ekran konkordancji, czyli daje możliwość przeczytania kolokacji w kontekście, w jakim występuje w artykule, a dzięki temu poddania ich analizie jakościowej.

Konkordancje

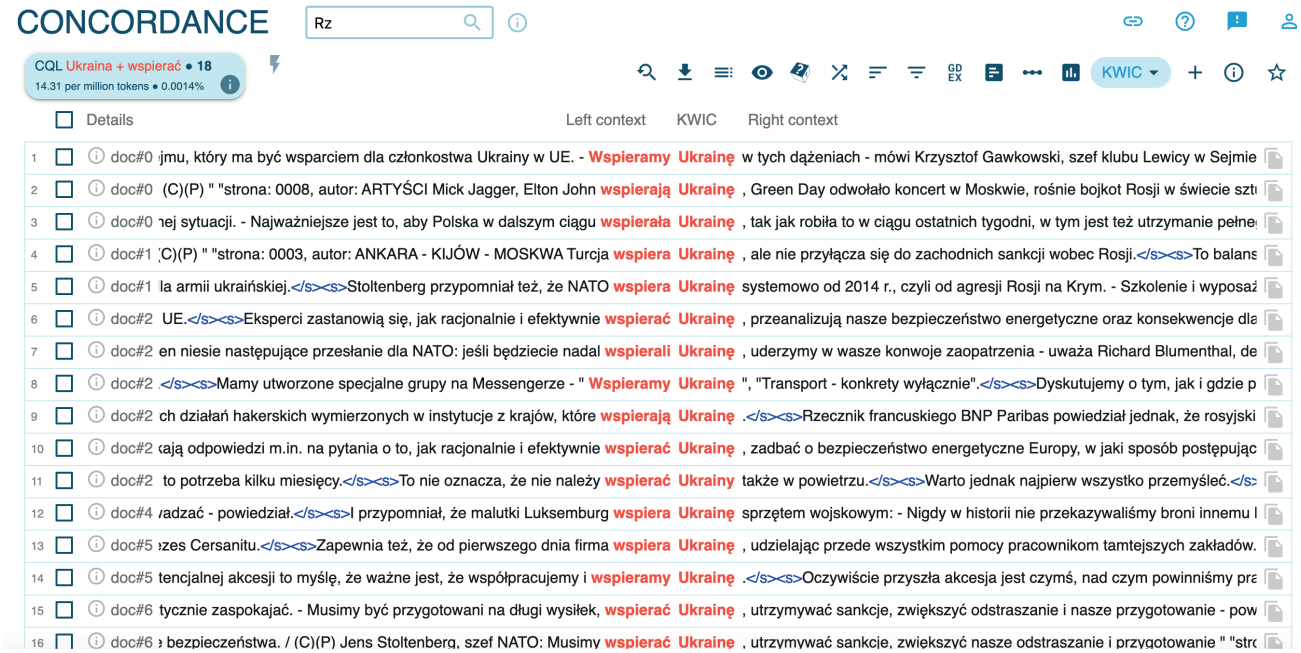
Konkordancja to podstawowe narzędzie pracy analitycznej z korpusem. Dzięki temu możemy przeczytać każdy fragment tekstu zawarty w korpusie. Jest to lista wszystkich przykładów szukanego słowa lub frazy, znalezionych w korpusie. Przenosi użytkownika do nieprzetworzonych danych, które są punktem wyjścia dalszej analizy (Kilgarriff i in., 2014: 10).

Wyszukiwane słowo prezentowane jest w formacie KWIC (ang. *Key Word In Context*) – jest ono podświetlone na środku zestawienia z możliwością przeczytania tekstu przed i po słowie kluczowym. Konkordancję można sortować, próbkować, filtrować (na przykład według kontekstu lub typu tekstu) oraz zapisywać. Dostępnych jest wiele analiz częstotliwości, w tym raporty kolokacji oraz analiza według typów tekstów (jeśli korpus ma zdefiniowane typy tekstów). Na poziomie pojedynczego wskazania użytkownik może zaznaczyć wyszukiwane hasło, aby uzyskać szerszy kontekst (aż do poziomu całego artykułu), lub zaznaczyć element w kolumnie „referencje", aby zobaczyć jego metadane (Kilgarriff i in., 2014: 13).

Istotne znaczenie dla dalszych analiz ma przyjazny interfejs użytkownika, jaki udostępnia SketchEngine. Łatwy dostęp (jedno kliknięcie) do całego tekstu, w którym występuje poszukiwany fragment, oraz możliwość swobodnego ustalania „ramek" interpretacji (rozwijania coraz to większego fragmentu tekstu) pozwalają badaczowi/badaczce na prowadzenie analizy jakościowej. Dodatkowo poszczególne fragmenty tekstu można samodzielnie kodować z poziomu interfejsu programu. Można tworzyć kategorie kodowe, filtrować zakodowane teksty, zarządzać kluczem kodowym. Dla przeprowadzenia podstawowej analizy jakościowej nie musimy eksportować fragmentów tekstu do programu CAQDAS. Możemy to zrobić, jeśli chcemy w bardziej rozbudowany sposób zarządzać kodowaniem i analizą jakościową. Eksport danych jest możliwy w podstawowych formatach – txt, csv, xlsx – lub jako pdf.

Na rysunku 2 widzimy konkordancje dla frazy „wesprzeć Ukrainę". Poszczególne fragmenty są unikalne (nie powtarzają się), są przypisane do dokumentów (w tym przypadku odpowiadających kolejnemu tygodniowi analizy), z poziomu KWIC możemy przejść do widoku zdania. Możemy je dowolnie sortować, zarówno na podstawie metadanych, jak i wskazanych słów (kolejnych tokenów) w widoku KWIC. Jeśli dany zbiór jest zbyt duży dla możliwości analizy przez badacza, łatwo można wylosować próbę o zdefiniowanej wielkości.

Rysunek 2. Fragment zestawienia konkordancji dla relacji „wspieramy Ukrainę", zbudowanej na korpusie artykułów z dziennika „Rzeczpospolita"



Źródło: badania własne.

Frekwencje

Frekwencja, czyli zestawienie występujących w tekście słów, jest chyba pierwszym skojarzeniem dla większości badaczy, gdy pada hasło „ilościowe metody analizy tekstów". Najczęstsze działania analityczne, jakie wykonujemy na podstawie informacji o częstotliwości występowania słów w danym korpusie, to sporządzenie listy słów, która porządkuje jednostki językowe (leksykalne, gramatyczne lub semantyczne) alfabetycznie lub według częstotliwości, albo wyszukanie poszczególnych jednostek językowych i porównanie ich częstości występowania w różnych korpusach lub częściach korpusów (określanych jako korpusy podrzędne).

Przy zestawieniu frekwencji warto posługiwać się stop-listą, czyli listą słów, które chcemy wykluczyć z danego zestawienia. W SketchEngine, podobnie jak w większości programów CADS, mamy możliwość utworzenia lub importu wcześniej przygotowanej własnej stop-listy.

W wielu programach możemy sporządzić listy frekwencji nie tylko pojedynczych słów, ale również tzw. n-gramów, gdzie n odnosi się do liczby słów traktowanych jako jedna jednostka (Gillings, Mautner, Baker, 2023: 17). Do porównania pomiędzy korpusami stosujemy najczęściej relatywną lub znormalizowaną frekwencję (ang. *relative or normalized frequency*) – liczbę wystąpień danego tokena na milion tokenów.

W trakcie analizy warto sprawdzić rozproszenie występowania danych tokenów (ang. *dispersion*), czyli liczbę wystąpień w poszczególnych tekstach składających się na korpus. Może się bowiem zdarzyć, że token występuje szczególnie często w konkretnych tekstach, a nie w całym korpusie. SketchEngine prezentuje dyspersje na czytelnych wykresach słupkowych jako funkcję domyślną przy sprawdzaniu frekwencji.

Rysunek 3. Zestawienie frekwencji słów dla korpusu tekstów z „Gazety Wyborczej". Wskazania dotyczą frekwencji bezwzględnej i frekwencji na milion słów w korpusie

WORDLIST GW

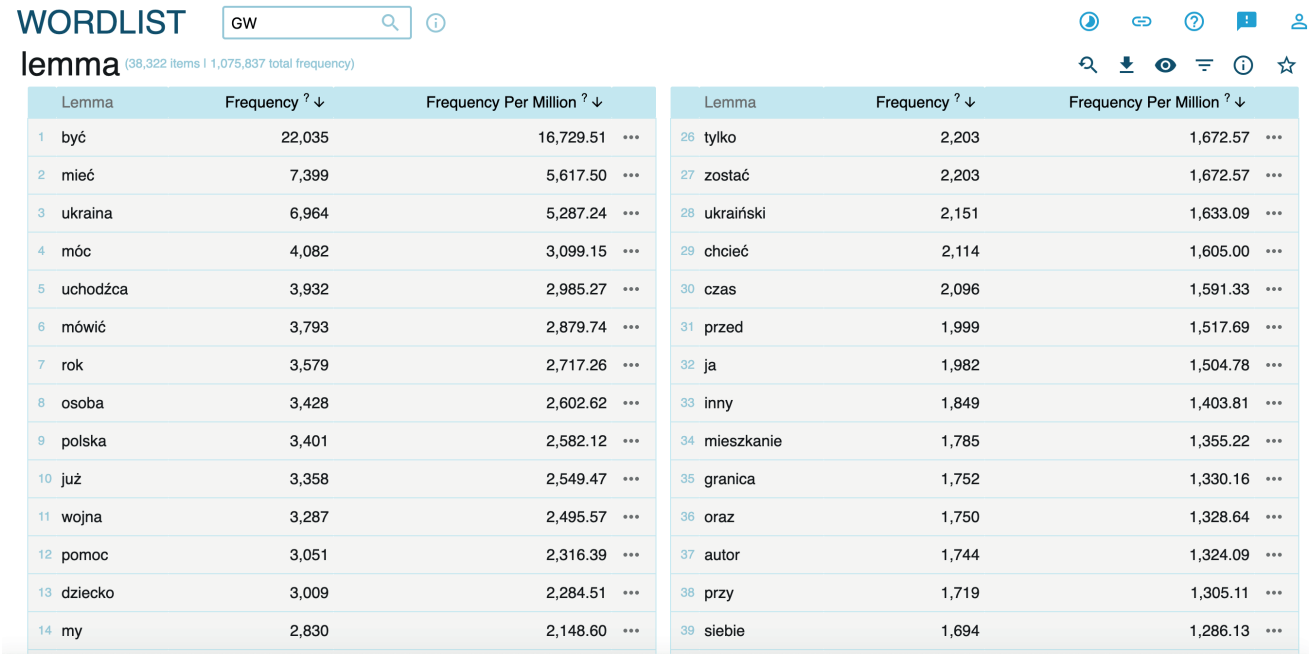
word (84,331 items I 1,075,518 total frequency)

	Word	Frequency	Frequency Per Million
1	jest	8,583	6,516.42
2	są	3,949	2,998.18
3	ukrainy	3,897	2,958.70
4	już	3,351	2,544.16
5	ma	2,667	2,024.85
6	mówi	2,450	1,860.10
7	ich	2,266	1,720.40
8	będzie	2,227	1,690.79
9	uchodźców	2,222	1,687.00
10	tylko	2,202	1,671.81
11	ze	2,023	1,535.91
12	tego	2,004	1,521.49
13	dzieci	1,974	1,498.71
14	którzy	1,949	1,479.73
26	bardzo	1,589	1,206.41
27	było	1,551	1,177.56
28	polsce	1,534	1,164.65
29	teraz	1,465	1,112.26
30	pomoc	1,453	1,103.15
31	jeszcze	1,436	1,090.25
32	nich	1,405	1,066.71
33	jej	1,402	1,064.43
34	która	1,396	1,059.88
35	lat	1,388	1,053.80
36	jednak	1,363	1,034.82
37	pracy	1,353	1,027.23
38	mają	1,345	1,021.16
39	pomocy	1,338	1,015.84

Źródło: badania własne.

Kluczową decyzją przy analizach frekwencji dla języka polskiego jest wybór, czy chcemy opisać tekst na poziomie słów (rysunek 3) czy lematów (rysunek 4). Jeśli naszym głównym celem jest analiza tematów poruszanych w zebranym korpusie, wtedy frekwencje warto analizować na poziomie lematów, szczególną uwagę zawracając na rzeczowniki. Frekwencje słów są wykorzystywane w analizach językoznawczych. Na rysunku 4 widzimy dużą liczbę wystąpień rzeczowników: „ukraina", „uchodźca", „osoba", „wojna", „polska". W praktyce badawczej taka lista lematów jest kolejnym sposobem na selekcję tekstów, do których możemy sięgnąć, korzystając z funkcji konkordancji w programie lub wyeksportować je do pogłębionej analizy jakościowej.

Rysunek 4. Zestawienie frekwencji poszczególnych lematów dla korpusu tekstów z „Gazety Wyborczej". Wskazania dotyczą frekwencji bezwzględnej i frekwencji na milion tokenów w korpusie



WORDLIST GW

lemma (38,322 items | 1,075,837 total frequency)

	Lemma	Frequency	Frequency Per Million
1	być	22,035	16,729.51
2	mieć	7,399	5,617.50
3	ukraina	6,964	5,287.24
4	móc	4,082	3,099.15
5	uchodźca	3,932	2,985.27
6	mówić	3,793	2,879.74
7	rok	3,579	2,717.26
8	osoba	3,428	2,602.62
9	polska	3,401	2,582.12
10	już	3,358	2,549.47
11	wojna	3,287	2,495.57
12	pomoc	3,051	2,316.39
13	dziecko	3,009	2,284.51
14	my	2,830	2,148.60

	Lemma	Frequency	Frequency Per Million
26	tylko	2,203	1,672.57
27	zostać	2,203	1,672.57
28	ukraiński	2,151	1,633.09
29	chcieć	2,114	1,605.00
30	czas	2,096	1,591.33
31	przed	1,999	1,517.69
32	ja	1,982	1,504.78
33	inny	1,849	1,403.81
34	mieszkanie	1,785	1,355.22
35	granica	1,752	1,330.16
36	oraz	1,750	1,328.64
37	autor	1,744	1,324.09
38	przy	1,719	1,305.11
39	siebie	1,694	1,286.13

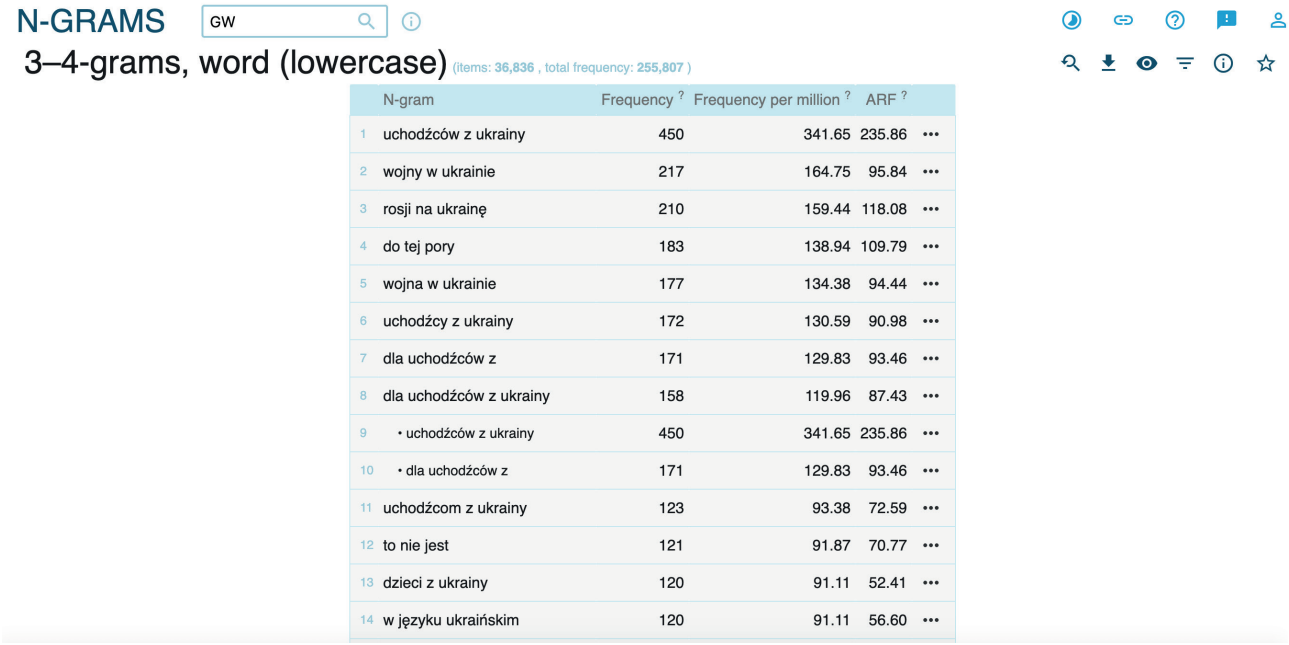
Źródło: badania własne.

N-gramy (ang. *n-grams*)

Dużo więcej informacji na temat analizowanego korpusu dostarczą nam n-gramy. Narzędzie N-gram tworzy listy frekwencji, ale dla sekwencji tokenów. N-gramy są również nazywane wyrażeniami wielowyrazowymi (MWE) lub wiązkami leksykalnymi. Użytkownik ma do wyboru opcje filtrowania, w tym wyrażenia regularne, aby szczegółowo określić, które n-gramy chce zobaczyć na liście. W skład n-gramu mogą wchodzić różne obiekty – litery, cyfry, sylaby, tokeny, słowa lub inne. N-gramy można wygenerować dla dowolnej formy, najczęściej wykorzystywane są określona forma gramatyczna lub lemat.

Wygenerowanie listy najczęściej występujących n-gramów pomoże nam dostrzec zjawiska językowe, które mogą pozostać niezauważone przy użyciu innych narzędzi. N-gramy mogą identyfikować kluczowe dla dyskursu frazy.

Rysunek 5. Zestawienie n-gramów 3–4-wyrazowych dla korpusu „Gazety Wyborczej"



N-GRAMS GW

3–4-grams, word (lowercase) (items: 36,836 , total frequency: 255,807)

	N-gram	Frequency	Frequency per million	ARF
1	uchodźców z ukrainy	450	341.65	235.86
2	wojny w ukrainie	217	164.75	95.84
3	rosji na ukrainę	210	159.44	118.08
4	do tej pory	183	138.94	109.79
5	wojna w ukrainie	177	134.38	94.44
6	uchodźcy z ukrainy	172	130.59	90.98
7	dla uchodźców z	171	129.83	93.46
8	dla uchodźców z ukrainy	158	119.96	87.43
9	• uchodźców z ukrainy	450	341.65	235.86
10	• dla uchodźców z	171	129.83	93.46
11	uchodźcom z ukrainy	123	93.38	72.59
12	to nie jest	121	91.87	70.77
13	dzieci z ukrainy	120	91.11	52.41
14	w języku ukraińskim	120	91.11	56.60

Źródło: badanie własne.

Na prezentowanym przykładzie uwzględniono następujące miary: frekwencję bezwzględną (liczbę wystąpień MWE w korpusie), frekwencję względną (liczbę wystąpień MWE na milion tokenów) i wreszcie *Average Reduced Frequency* (ARF). Ta ostatnia miara pozwala zmniejszyć znaczenie wspólnego występowania słów, jeśli występują one tylko w niewielkiej liczbie dokumentów (np. tylko w jednym) (Savicky, Hlavacova, 2002). Tu ponownie pojawia się kluczowa dla CADS kwestia dyspersji, czyli rozłożenia intensywności danego zjawiska (np. kolokacji lub słów kluczowych) w poszczególnych dokumentach korpusu (Egbert, Biber, 2018).

Możliwość wygenerowania n-gramów ułatwia nam selekcję znaczących tekstów do dalszej analizy. Łatwo odnajdujemy frazy, które najbardziej odróżniają nasz korpus od standardowego korpusu języka polskiego. Dodatkowo możemy w prosty sposób skonfrontować wyniki analizy ilościowej z naszymi oczekiwaniami co do tematów zawartych w charakterystycznych frazach. Na wskazanym na rysunku 5 przykładzie wyraźnie widać, że najbardziej charakterystyczną frazą jest n-gram „uchodźców z Ukrainy". Ma on nie tylko najwyższą liczbę wystąpień w analizowanym korpusie, ale również najwyższą miarę ARF, co wskazuje na występowanie tej frazy w dużej liczbie dokumentów.

Keywords

Jednym z najbardziej przydatnych narzędzi CADS są „słowa kluczowe" (ang. *keywords*) (por. Baker, 2004; Kilgarriff, 2009; Egbert, Biber, 2018; Egbert, Larsson, Biber, 2020)

> O słowie mówi się, że jest „kluczem", jeśli […] jego częstotliwość występowania w tekście w porównaniu z częstością występowania w zbiorze referencyjnym jest taka, że prawdopodobieństwo statystyczne obliczone za pomocą odpowiedniej procedury jest mniejsze lub równe wartości p określonej przez użytkownika (Scott, 2011).

> Słowo kluczowe można zdefiniować jako słowo, które występuje z nietypową częstotliwością w danym tekście. Nie oznacza to wysokiej częstotliwości, ale nietypową częstotliwość w porównaniu z jakimś korpusem referencyjnym (Scott, 1997: 236).

Wyznaczanie tych wyrażeń jest nazywane ekstrakcją terminów. Jest to automatyczna metoda analizy tekstu w celu identyfikacji fraz spełniających kryteria terminów (zazwyczaj są to wyrażenia zawierające rzeczownik). Ekstrakcja terminologii ma zastosowanie w analityce tekstu, gdzie jest wykorzystywana do modelowania tematów, eksploracji danych i wyszukiwania informacji z tekstu nieustrukturyzowanego. Jest używana w analizach w podobny sposób jak *topic modeling* – narzędzie z obszaru *Natural Language Processing* (por. np. Törnberg, Törnberg, 2016; Heidenreich i in., 2019; Isoaho, Gritsenko, Mäkelä, 2021; Chen i in., 2023). Rezultaty są podzielone na słowa kluczowe – składające się z jednego tokena i terminy (ang. *terms*). Termin to wyrażenie wielowyrazowe (składające się z kilku tokenów), które pojawia się częściej w jednym korpusie (korpusie głównym) w porównaniu z innym korpusem (korpusem referencyjnym).

W SketchEngine za ekstrakcję terminów odpowiedzialne jest narzędzie OneClick. W tym procesie łączone są metody statystycznej i lingwistycznej ekstrakcji terminów, która jest dodatkowo wspomagana przez porównanie języka przesłanego tekstu z językiem ogólnym. To właśnie porównanie tekstów dziedzinowych i tekstów ogólnojęzykowych pozwala określić, które pozycje leksykalne są terminologią, a które tylko często występującymi frazami bez znaczenia dla danej dziedziny.

Kluczowość nie jest prostą cechą (Gabrielatos, Marchi, 2012). Niezależnie od tego, jak obiektywnie obliczana jest wielkość efektu i istotność statystyczna, identyfikacja pozycji jako kluczowej zależy od wielu subiektywnych decyzji dotyczących:

1) progów częstości, efektu – wielkość i istotność statystyczna,
2) charakteru jednostek językowych, które są przedmiotem analizy,
3) atrybutów porównywanych korpusów.

Mówiąc najprościej, analiza ilościowa niekoniecznie pociąga za sobą obiektywizm. Dlatego ważne jest, aby decyzje te były oparte na opisanych wcześniej zasadach, tak aby można było powtórzyć analizę ilościową (Gabrielatos, 2018: 253).

W SketchEngine możemy decydować o charakterze prezentowanej listy słów (terminów) kluczowych. Służy do tego zmienna „Focus on", która w interfejsie przyjmuje wartości zmieniające się o rząd wielkości (0,1, 1, 10, 100, 1000 itd.). Dopiero przy takich zmianach tego parametru widzimy wyraźną różnicę w wynikach zestawienia. Jeśli ustawimy jej wartości na „rare" (niskie wartości), to otrzymamy listę terminów bardzo rzadkich, zdecydowanie nietypowych dla korpusu referencyjnego. Pozwala to np. wyznaczyć terminologię dla danego zbioru (rysunek 6).

Jeśli ustawimy jej wartość na „common" (wysokie wartości), to otrzymamy listę słów, które występują bardzo często w korpusie referencyjnym. Oczywiście nasza lista pokazuje te terminy, które są specyficzne (występują znacząco częściej) w naszym korpusie niż w korpusie referencyjnym. W badaniu każdego korpusu warto przetestować kilka ustawień tego parametru.

Rysunek 6. Lista terminów kluczowych dla korpusu artykułów z „Dziennika Gazety Prawnej", Focus on: rare (0,01)

KEYWORDS DGP

SINGLE-WORDS ✓ MULTI-WORD TERMS ✓

reference corpus: Polish Web 2019 (plTenTen19) (items: 83,842)

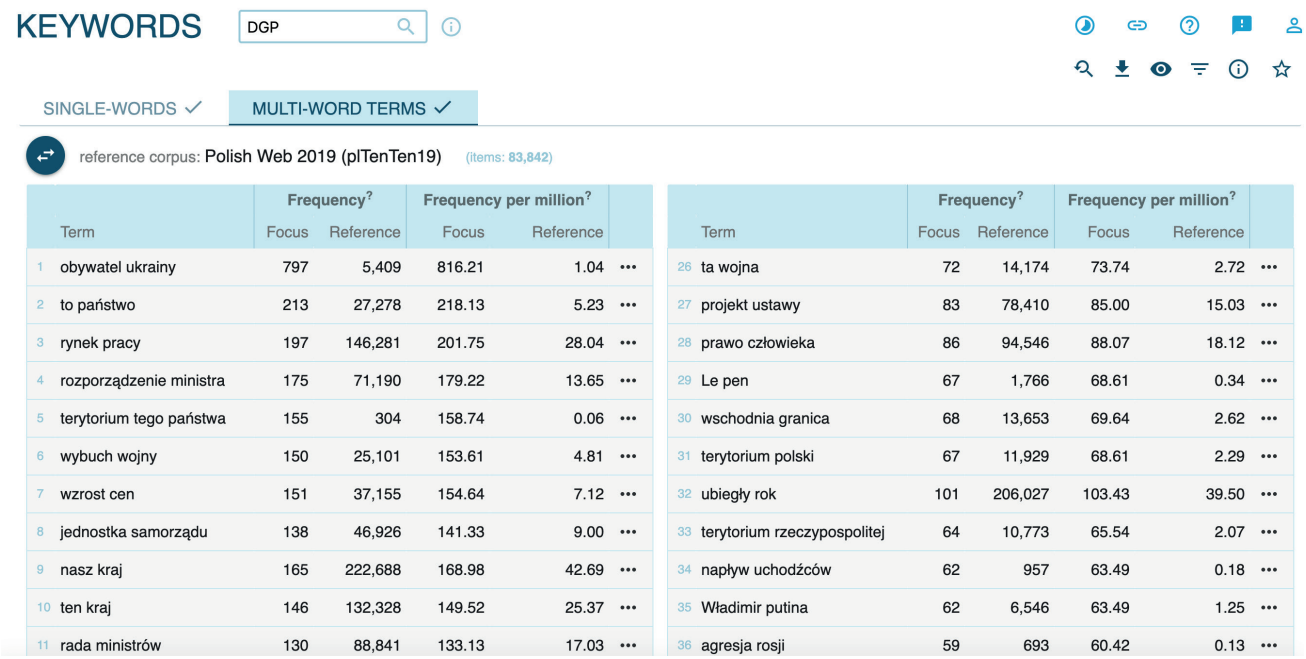
	Term	Frequency? Focus	Frequency? Reference	Frequency per million? Focus	Frequency per million? Reference
1	polski ład	74	50	75.78	< 0.01
2	Polski ład	31	0	31.75	0.00
3	terytorium tego państwa	155	304	158.74	0.06
4	krajowy plan odbudowy	22	0	22.53	0.00
5	rządowa agencja rezerw	23	7	23.55	< 0.01
6	rozporządzenie ministra klimatu	18	0	18.43	0.00
7	ukraiński uchodźca	34	55	34.82	0.01
8	odbudowa ukrainy	20	11	20.48	< 0.01
9	legislacyjny projekt	26	34	26.63	< 0.01
10	przepis specustawy	26	43	26.63	< 0.01
11	nowelizacja specustawy	30	63	30.72	0.01

Źródło: badania własne.

Na rysunku 6 widzimy, że ustawienie parametru „Focus on" na wartość 0,01 zwróciło nam listę terminów, które są bardzo specyficzne dla analizowanych tekstów. Na pierwszych miejscach mamy nazwy własne (np. „Polski Ład", „Krajowy Plan Odbudowy") lub terminy prawne (np. „przepis specustawy"). Posługując się analizą konkordancji oraz mając jasno sformułowane pytania badawcze wobec treści artykułów w korpusie, widzimy, że ten poziom funkcji keywords nie daje nam oczekiwanych odpowiedzi.

Natomiast po ustawieniu tego parametru na wartość 100 000 (rysunek 7) otrzymaliśmy listę wyrażeń, które dużo lepiej oddają oczekiwaną tematykę tekstów. Ponownie potwierdzić to możemy, przeglądając konkordancje dla poszczególnych terminów. Ustawienie parametru na konkretną wartość należy wypracować metodą prób i błędów, jest ono zależne od zawartości tematycznej konkretnego korpusu.

Rysunek 7. Lista terminów kluczowych dla korpusu artykułów z „Dziennika Gazety Prawnej", Focus on: common (100 000)



	Term	Frequency Focus	Frequency Reference	Frequency per million Focus	Frequency per million Reference
1	obywatel ukrainy	797	5,409	816.21	1.04
2	to państwo	213	27,278	218.13	5.23
3	rynek pracy	197	146,281	201.75	28.04
4	rozporządzenie ministra	175	71,190	179.22	13.65
5	terytorium tego państwa	155	304	158.74	0.06
6	wybuch wojny	150	25,101	153.61	4.81
7	wzrost cen	151	37,155	154.64	7.12
8	jednostka samorządu	138	46,926	141.33	9.00
9	nasz kraj	165	222,688	168.98	42.69
10	ten kraj	146	132,328	149.52	25.37
11	rada ministrów	130	88,841	133.13	17.03
26	ta wojna	72	14,174	73.74	2.72
27	projekt ustawy	83	78,410	85.00	15.03
28	prawo człowieka	86	94,546	88.07	18.12
29	Le pen	67	1,766	68.61	0.34
30	wschodnia granica	68	13,653	69.64	2.62
31	terytorium polski	67	11,929	68.61	2.29
32	ubiegły rok	101	206,027	103.43	39.50
33	terytorium rzeczypospolitej	64	10,773	65.54	2.07
34	napływ uchodźców	62	957	63.49	0.18
35	Władimir putina	62	6,546	63.49	1.25
36	agresja rosji	59	693	60.42	0.13

Źródło: badania własne.

Na rysunku 7 na pierwszym miejscu znajdujemy termin „obywatel ukrainy". Miara „frequency per milion" (liczba wystąpień na milion tokenów) pokazuje nam, że w naszym korpusie pojawia się on około 800 razy częściej niż w standardowym korpusie języka polskiego. Jeszcze bardziej charakterystycznym terminem jest „terytorium tego państwa" z frekwencją 2,5 tys. razy większą niż w korpusie referencyjnym. Dalej mamy „napływ uchodźców" (350 razy częściej) i „agresja rosji" (460 razy częściej). Fragmenty zwierające te frazy stają się kandydatami do uważnej analizy jakościowej.

Wnioski

Do wykorzystania CADS w praktyce badawczej przyczynia się niewątpliwie rozwój narzędzi komputerowych, takich jak omawiany w tym tekście SketchEngine. Narzędzia te wykorzystują metody lingwistyki korpusowej, uzupełnione o automatyczne przetwarzanie tekstu. Korpus, który jest poddany tokenizacji, jest również anotowany przez oznaczenie części mowy (ang. *POS tagging*), co pozwala na przeprowadzenie rzetelnych analizy dla polskich tekstów medialnych.

SketchEngine ma intuicyjny interfejs, który jest wyposażony w wiele informacji objaśniających działanie poszczególnych funkcjonalności programu. Dla użytkowników chcących zautomatyzować proces analizy danych dostępny jest również API programu (po utworzeniu konta w aplikacji).

Z perspektywy praktyki badawczej kluczowe są dwie funkcje programu: Word Sketch i Keywords. Pierwsza z nich pozwala w szybki sposób „zobaczyć" użycie kluczowych dla analizowanej tematyki słów w całym korpusie, bez względu na jego wielkość. Pozwala też na pokazanie różnic pomiędzy porównywanymi korpusami tekstów. Co ważne, możemy porównać nasz korpus nie tylko z korpusem referencyjnym języka polskiego – możemy zestawić dwa własne korpusy tekstów, np. pochodzące z różnych tytułów prasowych lub z różnych okresów. Druga ze wskazanych funkcji – Keywords – pozwala na ilościowe określenie tematyki analizowanego korpusu tekstów. Oczywiście jako wynik otrzymujemy jedynie MWE (wyrażenia wielowyrazowe), ale tu właśnie powinno nastąpić przejście do konkordancji i możliwość jakościowej interpretacji wskazanych tekstów. Ten sposób określania tematyki zbioru tekstów wydaje się bardziej intuicyjny i łatwiejszy do interpretacji przez badaczy niż narzędzia NLP, takie jak *topic modeling* (Chen i in., 2023).

Warto przy tym podkreślić, że wszystkie zadania, które zostały opisane powyżej dla programu SketchEngine, są możliwe do wykonania z wykorzystaniem narzędzi lingwistycznych dostarczanych przez CLARIN-PL. Przewagą SketchEngine jest integracja usług i przyjazny interfejs użytkownika, przewagą CLARIN-PL – otwartość technologii i darmowy dostęp. Przed wyborem narzędzia do analizy warto dokładniej przyjrzeć się obu rozwiązaniom.

Z perspektywy metodologii badań nad dyskursem najistotniejszą rolę w procesie analizy treści odgrywają konkordancje. Przyjazny dla użytkownika sposób przejścia do tekstów źródłowych to krok milowy w rozwoju CADS. Pozwala on potraktować wszystkie omawiane analizy ilościowe jako pierwszy etap w procesie badawczym, specyficzny dla danego korpusu rodzaj selekcji tekstów do analizy jakościowej. Dzięki temu omijamy trudności związane z doborem (losowym lub warstwowym) próby do badania. Metody CADS pozwalają wytypować fragmenty tekstów (lub całe artykuły), które w zależności od celu badania pokazują powtarzalne elementy korpusu lub najbardziej specyficzne zestawienia słów. Uproszczony interfejs, pozwalający na kodowanie znaczeń i zarządzanie kodami, daje możliwość dokończenia analizy jakościowej w obrębie tej jednej aplikacji. Badacze chcący wykorzystać większą pulę możliwości analizy jakościowej zakończą korzystanie z CADS na eksporcie wytypowanych tekstów do programu CAQDAS.

Ta właśnie cecha pozwala traktować CADS jako podejście z obszaru metod mieszanych (ang. *mixed methods*) (Creswell, 2009). Rozwiązuje to podstawowy, wspomniany na wstępie tekstu problem badaczy dyskursu – jak zinterpretować (w odniesieniu do sensów wypowiedzi) duże zbiory tekstu. Perspektywa lingwistyki korpusowej odwraca ten problem – im większy korpus, tym bardziej rzetelne wyniki otrzymujemy.

Bibliografia

Baker Paul (2004), *Querying Keywords: Questions of Difference, Frequency, and Sense in Keywords Analysis*, „Journal of English Linguistics", vol. 32(4), s. 346–359, https://doi.org/10.1177/0075424204269894

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

Baker Paul, Mcenery Tony (2005), *A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts*, „Journal of Language and Politics", vol. 4(2), s. 197–226.

Baker Paul, Gabrielatos Costas, Khosravinik Majid, Mcenery Tony, Wodak Ruth (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(3), s. 273–306, https://doi.org/10.1177/0957926508088962

Bednarek Monika (2006), *Evaluation in Media Discourse: Analysis of a Newspaper Corpus*, London–New York: Continuum.

Chen Yingying, Peng Zhao, Kim Sei Hill, Choi Chang Won (2023), *What We Can Do and Cannot Do with Topic Modeling: A Systematic Review*, „Communication Methods and Measures", vol. 17(2), s. 111–130, https://doi.org/10.1080/19312458.2023.2167965

CLARIN-PL (b.r.), https://ws.clarin-pl.eu/ [dostęp: 10.05.2024].

CLARIN-PL (b.r.), Login, https://services.clarin-pl.eu/services [dostęp: 10.05.2024].

Costa Antonio Pedro, Moreira Antonio, Freitas Fabio, Costa King, Bryda Grzegorz (red.) (2023), *Computer Supported Qualitative Research*, Cham: Springer International Publishing, https://doi.org/10.1007/978-3-031-31346-2

Creswell John W. (2009), *Editorial: Mapping the Field of Mixed Methods Research*, „Journal of Mixed Methods Research", vol. 3(2), s. 95–108, https://doi.org/10.1177/1558689808330883

Efe İbrahim (2019), *A corpus-driven analysis of representations of Syrian asylum seekers in the Turkish press 2011–2016*, „Discourse and Communication", vol. 13(1), s. 48–67, https://doi.org/10.1177/1750481318801624

Egbert Jesse, Biber Douglas (2018), *Incorporating text dispersion into keyword analyses*, „Corpora", vol. 14(1), s. 77–104, https://doi.org/10.3366/cor.2019.0162

Egbert Jesse, Larsson Tove, Biber Douglas (2020), *Doing Linguistics with a Corpus. Methodological Considerations for the Everyday User*, Cambridge: Cambridge University Press, https://doi.org/10.1017/9781108888790

Fairclough Norman (2000), *New Labour, New Language?*, London: Routledge.

Gabrielatos Costas (2018), *Keyness analysis: Nature, metrics and techniques*, [w:] Charlotte Taylor, Anna Marchi (red.), *Corpus Approaches To Discourse: A critical review*, Oxford: Routledge, s. 225–258.

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(1), s. 5–38, https://doi.org/10.1177/0075424207311247

Gabrielatos Costas, Marchi Anna (2012), *Keyness: Appropriate metrics and practical issues Discourse-Oriented Corpus Studies View project Conditionals and Modality View project. CADS*, https://www.researchgate.net/publication/261708842 [dostęp: 10.05.2024].

Gillings Mathew, Mautner Gerlinde, Baker Paul (2023), *Corpus-Assisted Discourse Studies*, Cambridge: Cambridge University Press, https://doi.org/10.1017/9781009168144

Hardt-Mautner Gerlinde (1995), *„Only Connect." Critical Discourse Analysis and Corpus Linguistics*, „UCREL Technical Paper", no. 6.

Heidenreich Tobias, Lind Fabienne, Eberl Jakob-Moritz, Boomgaarden Hajo G. (2019), *Media Framing Dynamics of the „European Refugee Crisis": A Comparative Topic Modelling Approach*, „Journal of Refugee Studies", vol. 32(1), s. i172–i182, https://doi.org/10.1093/jrs/fez025

Hunston Susan (2002), *Corpora in Applied Linguistics*, Cambridge: Cambridge University Press.

Isoaho Karoliina, Gritsenko Daria, Mäkelä Eetu (2021), *Topic Modeling and Text Analysis for Qualitative Policy Research*, „Policy Studies Journal", vol. 49(1), s. 300–324, https://doi.org/10.1111/psj.12343

Jakubíček Milos, Kilgarriff Adam, Kovář Vojtech, Rychlý Pavel (2013), *The TenTen Corpus Family*, [w:] *7th International Corpus Linguistics Conference CL*, s. 125–127, https://www.sketchengine.eu/wp-content/uploads/The_TenTen_Corpus_2013.pdf [dostęp 10.05.2024].

Kieraś Witold, Kobyliński Łukasz (2021), *Korpusomat – present state and the future of the project*, „Jezyk Polski", R. 101, z. 2, s. 49–58, https://doi.org/10.31286/JP.101.2.4

Kieraś Witold, Kobyliński Łukasz, Ogrodniczuk Maciej (2018), *Korpusomat – a Tool for Creating Searchable Morphosyntactically Tagged Corpora*, „Computational Methods in Science and Technology", vol. 24(1), s. 21–27, https://doi.org/10.12921/cmst.2018.0000005

Kilgarriff Adam (2009), *Simple maths for keywords*, [w:] *Proceedings of the Corpus Linguistics Conference. Liverpool, UK. 2009*, https://www.sketchengine.eu/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf [dostęp: 10.05.2024].

Kilgarriff Adam, Baisa Vit, Rychlý Pavel, Jakubíček Milos (2015), *Longest-commonest Match*, [w:] *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, s. 397–404, https://www.sketchengine.eu/wp-content/uploads/Longest-commonest_eLex2015.pdf [dostęp: 10.05.2024]

Kilgarriff Adam, Reddy Siva, Pomikálek Jan, Pvs Avinesh (2010), *A Corpus Factory for many languages*, [in:] Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, Daniel Tapias (red.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta: European Language Resources Association, s. 904–910, https://aclanthology.org/L10-1044/ [dostęp: 10.05.2024].

Kilgarriff Adam, Baisa Vit, Bušta Jan, Jakubícek Milos, Kovár Vojtech, Michelfeit Jan, Rychlý Pavel, Suchomel Vit (2014), *The Sketch Engine: Ten years on*, „Lexicography", vol. 1(1), s. 7–36, https://doi.org/10.1007/s40607-014-0009-9

Korpusomat (b.r.), https://korpusomat.pl/ [dostęp: 10.05.2024].

Korpusomat (Beta) (b.r.), https://korpusomat.eu/ [dostęp: 10.05.2024].

Krippendorff Klaus (2004), *Content analysis. An Introduction to Its Methodology*, Thousand Oaks–London–New Delhi: Sage Publications.

Krzyżanowski Michał, Forchtner Bernhard (2016), *Theories and concepts in critical discourse studies: Facing challenges, moving beyond foundations*, „Discourse & Society", vol. 27(3), s. 253–261, https://doi.org/10.1177/0957926516630900

Leech Geoffrey, Fallon Roger (1992), *Computer corpora – What do they tell us about culture?*, „ICAME Journal", vol. 16, s. 29–50.

Matytcina Marina S., Grigoryanova Tatiana (2022), *Statistical Methods for Extracting Collocations from a Text Corpus*, [w:] *2022 2nd International Conference on Technology Enhanced Learning in Higher Education (TELE)*, Lipetsk: IEEE, s. 55–57, https://doi.org/10.1109/TELE55498.2022.9801038

Piasecki Maciej (2007), *Polish Tagger TaKIPI: Rule Based Construction and Optimisation*, „Task Quarterly", vol. 11(1–2), s. 151–167, https://www.researchgate.net/publication/272685698 [dostęp 10.05.2024].

Piasecki Maciej (2014), *User-driven Language Technology Infrastructure -t he Case of CLARIN-PL*, [w:] *9th Language Technologies Conference Information Society – IS 2014*, s. 7–13, https://nl.ijs.si/isjt14/proceedings/isjt2014_01.pdf [dostęp: 10.05.2024].

Piper Alison (2000), *Some People Have Credit Cards and Others Have Giro Cheques: "Individuals" and "People" as Lifelong Learners in Late Modernity*, „Discourse and Society", vol. 11(4), s. 515–542.

Potts Amanda, Bednarek Monika, Caple Helen (2015), *How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina*, „Discourse and Communication", vol. 9(2), s. 149–172, https://doi.org/10.1177/1750481314568548

Przepiórkowski Adam (2009), *A comparison of two morphosyntactic tagsets of Polish*, [w:] *Representing Semantics in Digital Lexicography: Proceedings of MONDILEX Fourth Open Workshop*, s. 138–144, https://nlp.ipipan.waw.pl/~adamp/Papers/2009-mondilex/article.pdf [dostęp: 10.05.2024].

Radziszewski Adam, Kilgarriff Adam, Lew Robert (2011), *Polish Word Sketches*, https://www.sketchengine.eu/wp-content/uploads/Polish_Word_Sketches_2011.pdf [dostęp: 10.05.2024].

Rychlý Pavel (2008), *A Lexicographer-Friendly Association Score*, [w:] *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008*, s. 6–9, https://nlp.fi.muni.cz/raslan/2008/papers/13.pdf [dostęp: 10.05.2024]

Saputa Karol, Tomaszewska Aleksandra, Zawadzka-Paluektau Natalia, Kieraś Witold, Kobyliński Łukasz (2023), *Korpusomat.eu: A Multilingual Platform for Building and Analysing Linguistic Corpora*, [w:] Jiří Mikyška, Clélia de Mulatier, Maciej Paszynski, Valeria V. Krzhizhanovskaya, Jack J. Dongarra, Peter M.A. Sloot (red.), *Computational Science – ICCS 2023. 23rd International Conference, Prague, Czech Republic, July 3–5, 2023, Proceedings, Part II*, s. 230–237, https://nlp.ipipan.waw.pl/Bib/sap:etal:23:iccs.pdf [dostęp: 10.05.2024].

Savicky Petr, Hlavacova Jaroslava (2002), *Measures of word commonness*, „Journal of Quantitative Linguistics", vol. 9, s. 215–231.

Scott Mike (1997), *PC analysis of key words – and key key words*, „System", vol. 25(2), s. 233–245.

Scott Mike (2011), *WordSmith Tools Manual, Version 6*, Stroud: Lexical Analysis Software Ltd., https://lexically.net/downloads/version6/wordsmith6.pdf [dostęp: 10.05.2024].

SketchEngine (b.r.), https://www.sketchengine.eu/ [dostęp: 10.05.2024].

Stubbs Michael (1997), *Whorf's Children: Critical Comments on Critical Discourse Analysis (CDA)*, [w:] Ann Ryan, Alison Wray (red.), *Evolving Models of Language*, Clavendon: Multilingual Matters, s. 100–116.

Törnberg Anton, Törnberg Petter (2016), *Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum*, „Discourse & Society", vol. 27(4), s. 401–422, https://doi.org/10.1177/0957926516634546

Zawadzka-Paluektau Natalia (2023), *Ukrainian refugees in Polish press*, „Discourse and Communication", vol. 17(1), s. 96–111, https://doi.org/10.1177/17504813221111636

Cytowanie

Marek Troszyński (2024), *Badania dyskursu wspomagane korpusowo (CADS) jako wsparcie jakościowej analizy treści. Studium przypadku wykorzystania programu SketchEngine w badaniach dyskursu*, „Przegląd Socjologii Jakościowej", t. XX, nr 4, s. 44–67, https://doi.org/10.18778/1733-8069.20.4.03

Corpus-Assisted Discourse Studies (CADS) as Support for Qualitative Content Analysis: A Case Study Using SketchEngine in Discourse Research

Abstract: The article presents the potential use of corpus linguistics tools as an initial stage in qualitative content analysis. It discusses the development of Corpus-Assisted Discourse Studies (CADS). The core part of the article is a discussion of the functions of a program supporting CADS – SketchEngine. The text includes numerous examples that illustrate the ways of using CADS methods and SketchEngine functionalities for analyzing Polish press discourse. By enabling easy reference to source texts (concordances), SketchEngine facilitates the inclusion of mixed methods in discourse research.

Keywords: corpus linguistics, SketchEngine, Qualitative Content Analysis, mixed methods