Tworzenie materiałów do nauki języka obcego specjalistycznego z wykorzystaniem korpusów tematycznych

DOI: 10.47050/jows.2024.3.111-121

Dostępność materiałów do nauki popularnych języków obcych jest obecnie bardzo szeroka. Jednakże w przypadku języków obcych specjalistycznych sytuacja wygląda zupełnie inaczej – nauczyciele często muszą samodzielnie opracowywać materiały dydaktyczne, co może być trudne, zwłaszcza jeśli nie są ekspertami w danej dziedzinie. W takich sytuacjach pomocą mogą służyć korpusy tematyczne, które doskonale wspierają proces tworzenia tychże materiałów.

AGNIESZKA LEŃKO-SZYMAŃSKA
Uniwersytet Warszawski

Na przestrzeni ostatnich 30 lat korpusy stały się popularnym źródłem danych językowych wśród autorów specjalizujących się w tworzeniu komercyjnych materiałów do nauki języków obcych, zwłaszcza języka angielskiego. Słowniki (np.: *Oxford Advanced Learner's Dictionary of Current English*; Lea, Bradbery i Hornby 2020) i gramatyki opisowe (np.: *Longman Grammar of Spoken and Written English*; Biber i in. 1999), a także podręczniki (np.: *Touchstone*; McCarthy, McCarten i Sandiford 2004) oraz zestawy ćwiczeń (np.: *English Vocabulary in Use. Advanced*; McCarthy i O'Dell 2017) oparte są na analizach zapisów autentycznego użycia języka zgromadzonych w korpusach oraz prezentują przykłady zaczerpnięte z tych zasobów. Jednocześnie w literaturze fachowej oraz podczas konferencji i warsztatów metodycznych promowane jest także bezpośrednie korzystanie z korpusów przez nauczycieli języków obcych w celu tworzenia własnych materiałów dydaktycznych, a także przez samych uczniów do samodzielnych eksploracji danych językowych (Römer 2011). To ostatnie zastosowanie określane jest terminem „uczenia się sterowanego danymi" (ang. *data-driven learning*) wprowadzonym przez Tima Johnsa (1991).

Zalety wykorzystywania korpusów przez nauczycieli i uczących się obejmują bardziej rzetelny i dokładny opis funkcjonowania języka w konkretnych kontekstach oraz sytuacjach komunikacyjnych, a także lepsze dopasowanie materiałów do określonych potrzeb i zainteresowań ucznia. Korpusy umożliwiają ponadto ekspozycję ucznia na autentyczne przykłady użycia języka, a nie na jego uporządkowaną i wygładzoną formę, jak to często ma miejsce w bardziej tradycyjnych materiałach dydaktycznych. Co więcej, bezpośrednie korzystanie z korpusu przez osoby uczące się języka obcego rozwija w nich autonomię oraz intuicję językową (Cheng i Lam 2022). O efektywności uczenia sterowanego danymi świadczą wyniki licznych badań (Boulton i Cobb 2017).

Niestety bezpośrednie wykorzystanie korpusów przez nauczycieli i uczniów ciągle nie zyskało dużej popularności. Jednym z ważnych powodów takiego stanu rzeczy jest to, że samodzielne tworzenie zadań dydaktycznych przy użyciu korpusów jest czasochłonne oraz wymaga wiedzy i szeregu umiejętności (Leńko-Szymańska 2022). Dlatego też

nauczyciele wolą korzystać z gotowych materiałów, których oferta jest niezwykle bogata, zwłaszcza dla częściej nauczanych języków obcych. Jednakże w przypadku bardziej niszowych obszarów, takich jak nauczanie mniej popularnych języków, nauczanie języków specjalistycznych dla precyzyjnie zdefiniowanych potrzeb, czy w końcu nauczanie języka przez treść (ang. *content-based language teaching*; Chodkiewicz 2011) pozyskanie odpowiednich gotowych materiałów dydaktycznych jest znacznie trudniejsze lub wręcz niemożliwe. W takich przypadkach nauczyciele są zmuszeni tworzyć własne treści, a zasoby i narzędzia korpusowe są w tym zakresie niezwykle pomocne.

Celem tego artykułu jest opis procesu samodzielnego tworzenia materiałów dydaktycznych do nauki języka specjalistycznego z wykorzystaniem korpusów tematycznych. Przedstawione są w nim sposoby tworzenia własnych korpusów tematycznych z wcześniej zgromadzonych tekstów lub poprzez automatyczne pozyskiwanie treści ze stron internetowych. Wskazane są także popularne i łatwo dostępne narzędzia do kompilacji i analizy korpusów autorskich. Na koniec omówione będą przykłady klasycznych zadań językowych przygotowanych na podstawie danych korpusowych, a także zadań wymagających analizy danych korpusowych przez ucznia (wspomniane uczenie się sterowane danymi). Dla zilustrowania opisanych praktyk zaprezentowane zostaną korpusy tematyczne oraz oparte na nich materiały dydaktyczne stworzone w ramach europejskiego projektu programu Erasmus+ o nazwie „Teaching English as a Content Subject at the Tertiary Level (TE-Con3)”[1]. Jego celem było opracowanie innowacyjnej metodologii nauczania języka angielskiego w instytucjach szkolnictwa wyższego przy użyciu modułów treści odzwierciedlających różne dyscypliny akademickie. Zadania oparte na korpusach tematycznych uzupełniają materiały edukacyjne, które powstały w wyniku projektu. Adresatami projektu były i są uczelnie w całej Europie, a zaprojektowane materiały dydaktyczne są ogólnie dostępne w internecie[2].

1 tecon3.wn.uw.edu.pl.

2 tecon3.itee.radom.pl.

Tworzenie korpusów tematycznych

Konieczność samodzielnego tworzenia korpusów tematycznych

Korpusy tematyczne są często wykorzystywane w badaniach naukowych, ale także w dziedzinach stosowanych, takich jak leksykografia i przekład, choćby do kompilacji glosariuszy terminologicznych lub kompendiów frazeologicznych, np. *Medical Academic Vocabulary List* (MAVL; Lei i Liu 2016)[3] czy *Oxford Phrasal Academic Lexicon*[4]. Są też używane do opracowywania materiałów dydaktycznych dla języków specjalistycznych. Niestety tematyczne zasoby korpusowe bywają niezwykle rzadko udostępniane szerszej społeczności akademickiej i praktykom. Jest to spowodowane przede wszystkim ograniczeniami nakładanymi przez prawa autorskie, które utrudniają dalsze rozpowszechnianie zebranych tekstów, nie tylko w przypadku korpusów specjalistycznych, lecz także zbiorów języka ogólnego. W niektórych przypadkach dodatkowe utrudnienia w dzieleniu się zebranymi danymi mogą być związane z poufnością informacji biznesowych lub technicznych zawartych w tekstach wchodzących w skład korpusu (np. w kontraktach czy dokumentacji technicznej). Innym powodem może być fakt, że udostępnianie korpusu wiąże się z dodatkowym nakładem pracy związanym z uporządkowaniem, adnotacją i opisem tekstów, które pozwolą innym użytkownikom zrozumieć, jaka jest struktura i zawartość zasobu.

3 www.eapfoundation.com/vocab/academic/other/mavl.

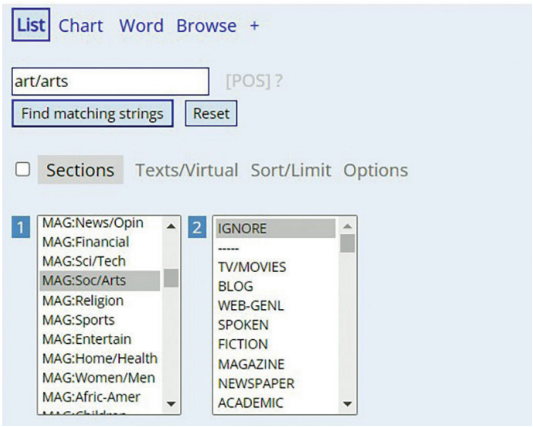
4 www.oxfordlearnersdictionaries.com/about/wordlists.

Brak dostępu do gotowych korpusów tematycznych stanowi niewątpliwie główną przeszkodę w wykorzystaniu danych korpusowych do tworzenia autorskich materiałów dydaktycznych przez nauczycieli języków specjalistycznych. Należy jednak wziąć pod uwagę, że nawet gdyby takie zbiory były łatwiej dostępne, mogłoby się okazać, że nie są odpowiednie dla szczegółowo sprecyzowanych potrzeb konkretnego nauczyciela i jego uczniów. Zebrane w nich teksty, mimo że reprezentują tę samą dziedzinę, mogą dotyczyć innych tematów, zawierać inne gatunki użytkowe bądź literackie lub okazać się za mało albo za bardzo techniczne w stosunku do celów dydaktycznych danego kursu języka specjalistycznego. Na przykład korpus języka medycznego zawierający niewielką liczbę tekstów o urologii, w którego skład

wchodzą głównie podręczniki dla studentów medycyny, może się okazać niedopasowany do potrzeb nauczyciela wspomagającego grupę urologów w przygotowaniu do egzaminu specjalizacyjnego obejmującego znajomość piśmiennictwa naukowego w języku angielskim.

Wydaje się zatem nieuchronne, że nauczyciel języka specjalistycznego pragnący skorzystać z dobrodziejstw oferowanych przez korpusy tematyczne, musi stworzyć własne zasoby, które będą dostosowane do konkretnych potrzeb jego i jego uczniów pod względem zakresu tematycznego, stopnia techniczności (specjalizacji) i różnorodności gatunków. Na szczęście dostępne są obecnie narzędzia korpusowe i informatyczne, które mogą wspomagać nauczyciela w kompilacji własnych korpusów tematycznych do celów dydaktycznych.

Rys. 1. Proces tworzenia korpusu tematycznego z części korpusu referencyjnego COCA



Źródło: COCA.

Tworzenie korpusów tematycznych z większych korpusów referencyjnych

Najprostszym, choć jednocześnie najbardziej ograniczonym sposobem na stworzenie korpusu tematycznego dla potrzeb dydaktycznych jest odwołanie się do części składowych (tzw. podkorpusów) ogromnych korpusów referencyjnych, które mają różnorodną, a jednocześnie dobrze opisaną i oznaczoną strukturę wewnętrzną obejmującą różne tematy i gatunki tekstów. Najlepszym przykładem takiego zasobu dla języka angielskiego jest liczący obecne miliard wyrazów korpus współczesnego języka angielskiego w odmianie amerykańskiej *Corpus of Contemporary American English* (*COCA*; Davies 2008–). Obejmuje on osiem głównych typów (gatunków) tekstów: język mówiony, beletrystykę, artykuły z czasopism popularnych, artykuły z gazet (dzienników), teksty naukowe, napisy filmowe i telewizyjne, blogi i inne strony internetowe. Każdy z tych gatunków podzielony jest dalej na podtypy (podgatunki) lub zakresy tematyczne. W ramach projektu Te-Con3 liczący 9,3 miliona wyrazów podkorpus artykułów pochodzących z popularnych czasopism poświęconych zagadnieniom społecznym i artystycznym posłużył jako zbiór tematyczny do stworzenia materiałów dydaktycznych na temat sztuki, których celem było uwypuklenie różnic w znaczeniu i związkach frazeologicznych dwóch pokrewnych rzeczowników: *art* i *arts* (Rys. 1).

Rys. 2. Skład korpusu wirtualnego (Media) stworzonego z części korpusu referencyjnego COCA

HELP	☐ 100	YEAR	GENRE	SOURCE	TITLE
1	☐	1990	ACAD	LatAmPopScult	Through the pantalla Uruguaya [Uruguayan screen]: The television environment for children in.....
2	☐	1991	ACAD	ArabStudies	Language and propaganda: Challenges to media interpretations of the Palestine question....
3	☐	1992	MOV	Manufacturing Consent: Noam Chomsky and the Media...	Documentary, Biography, War
	A film about the noted American linguist/political dissident and his warning about corporate media's role in modern propaganda.				
4	☐	1992	SPOK	ABC_Special	Viewpoint Politics and the Media: Reporting or Distorting?...
5	☐	1993	ACAD	IntlAffairs	Eclipse of reason: The media in the Muslim world....
6	☐	1993	ACAD	IntlAffairs	The market versus the state: The Chinese press since Tiananmen....
7	☐	1993	ACAD	IntlAffairs	The media's role in U.S. foreign policy....
8	☐	1993	ACAD	IntlAffairs	The press and power in the Russian Federation....
9	☐	1994	ACAD	ArtsEduc	Video in the classroom: A tool for reform....
10	☐	1995	ACAD	LatAmPopScult	Resistance and appropriation in Brazil: How the media and 'official culture' institutionalized.....
11	☐	1997	NEWS	AssocPress	
12	☐	1999	ACAD	SexResearch	Teenage Sexuality and Media Practice: Factoring in the Influences of Family, Friends, and School (Book)....
13	☐	2000	ACAD	PerspPolSci	Here We Go Again: Presidential Elections and the National Media....
14	☐	2001	ACAD	AmerStudies	Global Media and the Ambiguities of Resonant Americanism....
15	☐	2003	ACAD	ClearingHouse	^@What, Me Worry?^@.
16	☐	2004	NEWS	AssocPress	Developments in the news industry for May 3-10:...
17	☐	2005	ACAD	EnvironHealth	Environmental Health and the Media, Part 2: Beyond Get the Message Out/Put the Fires Out....
18	☐	2006	ACAD	SexResearch	Adolescents' Contact With Sexuality in Mainstream Media: A Selection-Based Perspective....

Źródło: COCA.

W korpusie *COCA*, a także w innych pokrewnych zbiorach dostępnych w popularnej witrynie English-corpora.org, istnieje także możliwość tworzenia tzw. korpusów wirtualnych. Oznacza to wyodrębnienie z całości korpusu grupy tekstów spełniających określone kryteria. Kryteria te mogą odnosić się do cech tekstu, takich jak jego gatunek i podgatunek, rok powstania lub konkretne źródło, z którego pochodzi (np. tytuł gazety), a także mogą obejmować określone wyrazy (formy hasłowe lub tekstowe) i wyrażenia, pojawiające się w tekście w określonej minimalnej częstotliwości. Przykładem takiej metody tworzenia zasobu tematycznego jest korpus wirtualny *Media*, skompilowany w ramach projektu Te-Con3. Składa się on z 2,3 milionów wyrazów i 100 tekstów (rys. 2) należących do prawie wszystkich głównych typów tekstów korpusu COCA, z przewagą innych stron internetowych (38 tekstów), artykułów naukowych (37 tekstów) oraz blogów (16 tekstów).

Zaletą tworzenia korpusów tematycznych z istniejących już zbiorów referencyjnych jest fakt, że użytkownik korzysta z danych tekstowych, które zostały już wcześniej zebrane według ściśle określonych kryteriów, sprawdzone pod względem zawartości, skompilowane w reprezentatywny i zrównoważony zbiór, a także zlematyzowane i otagowane, czyli opatrzone formami hasłowymi i znacznikami morfosyntaktycznymi dla każdego wyrazu. Ponadto są one już powiązane z narzędziami do ich analizy, stanowiąc tym samym zasób gotowy do użycia. Ich wadą jest brak szerokiego wyboru pośród dostępnych gatunków i tematów, ponieważ jest on ograniczony tylko do tekstów, które weszły w skład korpusu referencyjnego, z zasady niezawierającego wąsko specjalistycznych dokumentów.

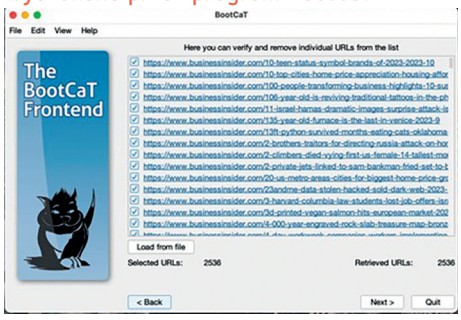
Tworzenie korpusów tematycznych z tekstów dostępnych w internecie

Najbardziej oczywistym źródłem tekstów, które mogą wejść w skład korpusu tematycznego, jest niewątpliwie internet dzięki swoim przepastnym zasobom, łatwości dostępu, a także elektronicznemu formatowi danych. Specjalistyczne blogi, czasopisma oraz książki fachowe i naukowe znajdujące się w wolnym dostępie, a także strony internetowe organizacji i przedsiębiorstw stanowią doskonały rezerwuar teksów specjalistycznych. Jednak pobieranie bezpośrednio z sieci takich tekstów – których do utworzenia korpusu potrzebnych jest bardzo wiele (setki, a nawet tysiące) – może okazać się żmudnym zadaniem. Na szczęście istnieją specjalne programy (boty), tzw. *web scrapers*, które mogą wyeliminować ręczne, powtarzalne klikanie, kopiowanie i wklejanie poprzez automatyczne pobieranie pożądanych treści ze stron i witryn internetowych. Oczyszczają one wyodrębnione teksty z niepotrzebnych kodów HTML oraz innych informacji, takich jak spisy treści, hiperłącza, przyciski nawigacji, reklamy lub zdjęcia. Dodatkowo pozyskane dane często są zapisywane w ustrukturyzowanym formacie, który pozwala potem na łatwą orientację i nawigację wśród zebranych tekstów. Przykładem takiego programu jest Web Scraper, rozszerzenie wyszukiwarki Chrome[5], które jest darmowym, stosunkowo prostym w obsłudze narzędziem, a w internecie dostępne są instrukcje i filmy przystępnie opisujące, jak z niego korzystać.

5 chrome.google.com/webstore/detail/web-scraper-free-web-scra/jnhgnonknehpejjnehehllkliplmbmhn.

Inne programy-boty, tzw. roboty indeksujące lub pająki sieciowe (ang. *web crawlers*), ułatwiają z kolei odnalezienie odpowiednich stron w internecie, zastępując ręczne poszukiwania tekstów w przeglądarce zautomatyzowanym łańcuchem działań. Użytkownik proszony jest o podanie kilku do kilkunastu słów i wyrażeń stanowiących podstawę kwerendy, a program łączy je w różne kombinacje i poszukuje stron, które je zawierają. Dodatkowo użytkownik może określić maksymalną listę stron do odszukania, a także domeny i typy dokumentów, na których program ma oprzeć indeksację (poszukiwane), lub te, które ma wykluczyć z tego procesu. W efekcie powstaje lista adresów stron internetowych (URL) spełniających kryteria określone przez użytkownika. Może je on następnie otworzyć, przejrzeć i ewentualnie wykluczyć z listy. Wiele pająków sieciowych posiada także funkcję automatycznej ekstrakcji danych z adresów zatwierdzonych przez użytkownika (czyli *web scaping* omówiony w poprzednim akapicie). Przykładem takiego programu, darmowego i prostego w obsłudze, jest BootCaT[6]. Rys. 3 przestawia wygenerowaną przez program listę stron, które spełniają określone przez użytkownika warunki.

Rys. 3. Lista stron internetowych wyszukana przez program BootCaT



Źródło: BootCaT.

6 bootcat.dipintra.it.

Zebrane w ten sposób teksty w zasadzie można już uznać za korpus tematyczny. Aby przeprowadzić jego analizę i wykorzystać dane tekstowe do tworzenia materiałów dydaktycznych, należy go połączyć z programem do analizy danych korpusowych, takim jak darmowy i łatwy w obsłudze AntConc[7]. Umożliwia on wyszukiwanie cytowań wyrazów, wyrażeń i struktur gramatycznych w korpusie (tzw. linii konkordancyjnych), generowanie list wyrazów, automatyczną ekstrakcję kolokacji i wiązek leksykalnych (n-gramów), a także generowanie listy słów kluczowych dla korpusu. Jednak znacznie lepsze efekty analizy można osiągnąć, najpierw lematyzując i tagując zebrane teksty, czyli dołączając do każdego wyrazu w tekście etykietę z informacją

7 www.laurenceanthony.net/software/antconc.

o jego formie hasłowej oraz jego cechach morfo-syntaktycznych. Etykiety te pozwalają podjąć bardziej wszechstronne wyszukiwania informacji językowych w korpusie. Lematyzację i tagowanie wykonuje się automatycznie specjalnie do tego przeznaczonym oprogramowaniem. Przykładem takiego programu jest dostępny w sieci TreeTagger[8], który został dostosowany do obsługi wielu języków. Sporo wyspecjalizowanych (choć nie zawsze przyjaznych w użyciu) narzędzi do tworzenia i analizy korpusów ogólnych oraz tematycznych w różnych językach można też znaleźć w repozytorium infrastruktury CLARIN-PL[9].

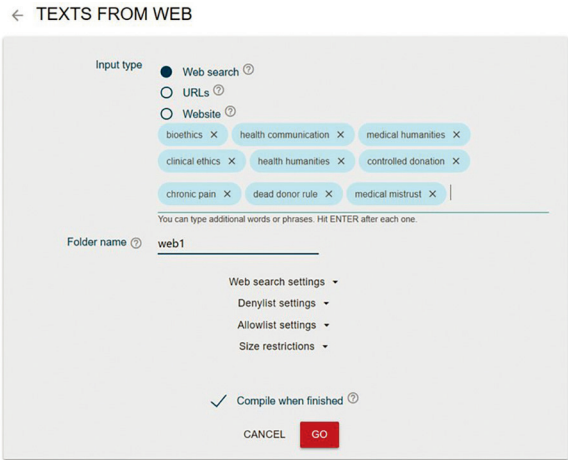
Jak widać z opisu w tym podrozdziale, proces budowania korpusu tematycznego (a w gruncie rzeczy każdego korpusu) jest wieloetapowy i – pomimo dostępności prostych w obsłudze i bezpłatnych narzędzi informatycznych – skomplikowany oraz pracochłonny. Istnieją jednakże programy, które mogą ten proces usprawnić, łącząc w sobie wszystkie wymienione wyżej funkcje. Najbardziej znanym i używanym narzędziem tego typu jest platforma internetowa Sketch Engine[10] (Kilgariff i in. 2014), z funkcjami web crawlingu, web scrapingu, lematyzacji i tagowania oraz narzędziami do analizy danych korpusowych. Dodatkowo platforma daje dostęp do ponad 800 korpusów różnego typu w ponad 100 językach, utworzonych przez badaczy i praktyków z całego świata, którzy zechcieli udostępnić swoje dane szerszej społeczności. Użytkownik może także dzielić się swoimi danymi z innymi wskazanymi użytkownikami. Ponadto platforma posiada przyjazny i rozbudowany interfejs oraz pełną dokumentację, która stanowi przystępny podręcznik obsługi programu. Co więcej, w sieci można odnaleźć wiele krótkich filmów instruktażowych, które krok po kroku pokazują działanie programu. Niestety dostęp do platformy jest nieodpłatny tylko przez miesiąc próbny, a potem wymagana jest opłata subskrypcyjna dla indywidualnych użytkowników lub instytucji.

8 www.cis.uni-muenchen.de/~schmid/tools/TreeTagger.

9 clarin-pl.eu.

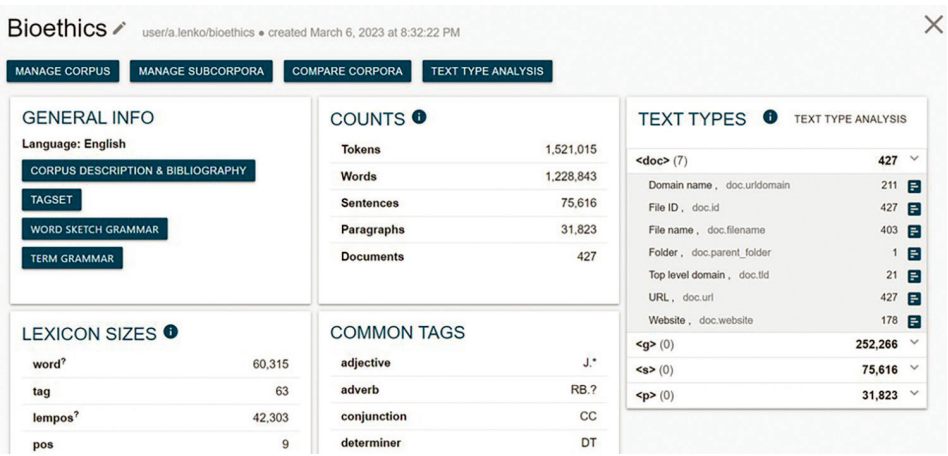
10 www.sketchengine.eu.

Rys. 4. Formularz pająka sieciowego w Sketch Engine



Platforma Sketch Engine została wykorzystana w projekcie TE-Con3 do utworzenia trzech korpusów tematycznych. Pierwszy z nich ma szerszy zakres i zawiera teksty dotyczące bioetyki. Początkowym etapem jego kompilacji było przejrzenie listy czasopism naukowych z tej dziedziny. Kwerenda wykazała, że artykuły zawarte w czasopismach są wysoce specjalistyczne i hermetyczne ze względu na naukowy język, w związku z czym nie są odpowiednie dla określonej w projekcie docelowej grupy studentów. Dlatego z tytułów zawartych w czasopismach zostało wyłonionych około 10 często powtarzających się terminów jedno- i wielowyrazowych, które następnie wpisano w formularz zbierania danych korpusowych na platformie Sketch Engine (Rys 4). Będący częścią programu pająk sieciowy zaproponował ponad 500 stron internetowych, z których kilkanaście zostało następnie ręcznie odrzuconych ze względu na nieadekwatną treść. Po zatwierdzeniu końcowej listy adresów URL w niecałe 5 minut na platformie powstał nowy korpus tematyczny liczący ponad 1,2 miliona wyrazów i 427 tekstów (Rys. 5).

Rys. 5. Zawartość korpusu Bioethics stworzonego przez pająka sieciowego w Sketch Engine



Źródło: SketchEngine.

Dwa kolejne korpusy skompilowane przy pomocy pająka sieciowego stanowiącego część Sketch Engine – Cars in cities (1,1 miliona wyrazów) oraz Automated car manufacturing (2,5 miliona wyrazów) oparte są na znacznie węziej zdefiniowanych zakresach tematycznych powiązanych z tematami lekcji zaprojektowanych w ramach projektu. W obu przypadkach słowa kluczowe pochodziły z tekstów przewidzianych jako podstawa dyskusji merytorycznych podczas zajęć.

Tworzenie korpusów tematycznych z własnych tekstów

Pomimo wydawałoby się nieskończoności zasobów internetowych, nie wszystkie typy tekstów mogą być w nim dostępne. Nauczyciel może więc wybrać budowanie korpusu z tekstów, które zgromadził z innych źródeł. W zbieranie ich można też zaangażować uczniów, prosząc o udostępnienie dokumentów lub tekstów, z którymi stykają się w swojej pracy zawodowej bądź w procesie kształcenia. Zapewni to właściwe dostosowanie zawartości merytorycznej materiałów dydaktycznych do potrzeb uczących się. Scalenie zgromadzonych samodzielnie treści w korpus i połączenie ich z programem do analizy danych korpusowych umożliwia nauczycielowi ich lepsze wykorzystanie poprzez zintegrowane przeszukiwanie i analizę. Do kompilacji korpusów tematycznych z własnych tekstów także można wykorzystać dostępne programy komputerowe. Jednym z takich narzędzi jest udostępniana bezpłatnie przez infrastrukturę CLARIN-PL platforma Korpusomat[11] (Saputa i in. 2023) obsługująca 32 języki. Wspomniana już platforma korpusowa Sketch Engine także udostępnia taką funkcję. Do platformy można przesłać teksty w wielorakich formatach, między innymi PDF, HTML i DOCX, które zostają przekształcone na zwykły tekst, a następnie – jak w przypadku dokumentów ściągniętych bezpośrednio z internetu – zlematyzowane i otagowane. Program umożliwia ręczne dopisanie metadanych do każdego dokumentu, przez co korpus może być przeszukiwany i analizowany w bardziej zniuansowany sposób.

11 korpusomat.eu.

Ta metoda budowy korpusu została wykorzystana do kompilacji zasobu tematycznego dotyczącego architektury. Łącznie 708 artykułów pochodzących z czterech prestiżowych czasopism poświęconych architekturze („Architectural Design", „Architectural Record", „Architecture Australia" i „Architect") z lat 2021 i 2020 zostało pobranych w formacie PDF z bazy czasopism pełnotekstowych udostępnianej przez Uniwersytet Warszawski. Korpus liczy 1,4 miliona wyrazów. Tabela 1 podsumowuje wszystkie korpusy tematyczne utworzone w ramach projektu TE-Con3.

Tab. 1. Korpusy tematyczne utworzone w ramach projektu TE-Con3

KORPUS	TEKSTY	ROZMIAR	METODA KOMPILACJI
Architecture	708	1,4 M	artykuły z czasopism fachowych w formacie PDF
Bioethics	427	1,2 M	web crawler ← terminologia specjalistyczna dobrana na podstawie tytułów publikacji
Cars in cities	449	1,1 M	web crawler ← słowa kluczowe dobrane ręcznie z tekstu
Automated car manufacturing	828	2,5 M	web crawler ← słowa kluczowe dobrane ręcznie z tekstu
COCA, podkorpus MAG: Soc/Arts)		9,3 M	podkorpus korpusu referencyjnego
COCA, podkorpus wirtualny Media	100	2,3 M	korpus wirtualny z korpusu referencyjnego ← wyrazy kluczowe dobrane ręcznie

Tworzenie materiałów dydaktycznych z korpusów tematycznych

Jedną z najbardziej oczywistych zalet korzystania przez nauczyciela z korpusu tematycznego jest łatwy dostęp do informacji o właściwościach danego języka specjalistycznego. Jest to szczególnie istotne, gdy nauczyciel nie jest ekspertem w dziedzinie, której języka uczy. Posługując się narzędziami do tworzenia listy wyrazów lub listy słów i wyrażeń kluczowych (ang. *keywords*) dostępnymi w programach do analizy korpusowej, nauczyciel może w automatyczny sposób wygenerować z korpusu tematycznego listę jedno- i wielowyrazowych terminów specjalistycznych, a także innych słów i wyrażeń typowych dla danej dziedziny. Dzięki innemu narzędziu – generatorowi konkordancji – nauczyciel może przyjrzeć się użyciu wybranych wyrazów, zgłębiając ich różne znaczenia w kontekście oraz studiując struktury składniowe, w jakich najczęściej występują (tzw. kolokacje gramatyczne). Można także prześledzić użycie struktur gramatycznych lub znaczników dyskursu w danej odmianie języka. Inne narzędzia umożliwiają automatyczne wyodrębnienie profili kolokacyjnych wyrazów lub ciągów powtarzających się wyrazów (tzw. wiązek leksykalnych lub n-gramów). Wyniki tych analiz stanowią doskonałą podstawę do projektowania materiałów dydaktycznych, które w unikalny sposób oddają cechy charakterystyczne nauczanej odmiany języka.

Informacje pochodzące z korpusu tematycznego mogą służyć jako podstawa do tworzenia zadań językowych dotyczących wszystkich aspektów języka: od słownictwa i terminologii poprzez frazeologię i gramatykę aż do budowy dyskursu i pragmatyki (Thompson 2022). Korpusy stanowią także niewyczerpane źródło autentycznych przykładów, które osadzone są w kontekście dziedzinowym, zatem nadają się doskonale do tworzenia materiałów do nauczania języka specjalistycznego. Nawet jeśli celem nauczania pozostają bardziej ogólne cechy języka, takie jak na przykład użycie wybranej struktury gramatycznej, cytowania wygenerowane z korpusu tematycznego i użyte do prezentacji oraz treningu sprawiają, że nauczanie osadzone jest w kontekście odpowiednim dla potrzeb ucznia. Należy jednak zaznaczyć, że zadania oparte na korpusie są najczęściej nastawione na rozwijanie raczej kompetencji językowych niż komunikacyjnych, choć dane korpusowe mogą także stanowić podstawę ćwiczeń wyrabiających sprawności językowe uczącego się, zwłaszcza sprawności pisania.

Korpusy tematyczne mogą służyć za podstawę różnego typu materiałów dydaktycznych. W najprostszy sposób nauczyciel może posłużyć się korpusem jako źródłem cytowań przydatnych do tworzenia bardziej tradycyjnych zadań językowych takich jak uzupełnianie luk (Rys 7). Nauczyciel może także stworzyć zadania wymagające od uczącego się samodzielnej analizy danych korpusowych (ang. *data-driven learning*). Dane te mogą być wstępnie wyselekcjonowane przez nauczyciela i zaprezentowane uczniom na karcie pracy (Rys 8), co eliminuje konieczność korzystania z narzędzi do analizy korpusowej przez uczących się (Boulton 2010). Gdy jednak nauczyciel uzna, że uczniowie poradzą sobie z obsługą programu, zadanie może pokierować ucznia w samodzielnym kontakcie z korpusem.

Opisane możliwości zostały wykorzystane w tworzeniu materiałów dydaktycznych opartych na korpusach tematycznych w ramach projektu TE-Con3. Łącznie opracowano 11 kart pracy dla czterech modułów tematycznych. Każda z kart zawiera od 3 do 5 zadań, które składają się na spójne sekwencje dydaktyczne. Tabela 2 zawiera zestawienie wszystkich kart wraz z informacją, jakich tematów oraz aspektów języka dotyczą.

Tab. 2. Zestawienie kart pracy stworzonych w ramach projektu TE-Con3

MODUŁ	TEMAT	ASPEKT JĘZYKA
Architecture	wood vs wooden	znaczenia, użycie
	design	łączliwość leksykalna
	aim	łączliwość syntaktyczna
	present perfect	gramatyka
	a magazine article	sprawności czytania i pisania
Biomedical sciences	case	terminologia, łączliwość leksykalna, gramatyka
	noun post-modification	gramatyka
Automotive engineering	noise	łączliwość leksykalna
	verb + by vs with	łączliwość syntaktyczna
Art and media	art vs arts	znaczenia, łączliwość leksykalna
	however	sprawność pisania

Zadanie 2 na pierwszej karcie pracy w module poświęconym architekturze (Rys 6) dotyczy subtelnej różnicy w użyciu między dwoma pokrewnymi wyrazami: *wood* i *wooden*. Leksem *wood* częściej występuje w tekstach przedstawiających szczegóły techniczne, podczas gdy *wooden* jest częściej używany w tekstach opisujących wygląd. Zadaniem uczniów jest samodzielne zaobserwowanie tej różnicy poprzez analizę trzech par zdań, z których każda zawiera oba wyrazy w dwóch różnych kontekstach, jednak jako przydawki tego samego rzeczownika. Wszystkie zdania użyte w tym zadaniu pochodzą z korpusu Architecture opisanego w sekcji 4.

Rys. 6. Fragment zadania ilustrującego różnice w użyciu wyrazów pokrewnych opracowanego na podstawie korpusu tekstów Architecture

TASK 2

Both words can be used before a noun. Study the sentences below. Can you see any differences in meaning and use between the two words?

– A brick wall features a bold mural; the **wood ceiling** and joists are painted bright yellow; furniture, floors, and walls are made of ash wood, treated with tongue oil or blackened.

– To achieve the transformation, ZGF played up the most striking aspect of the original building– its enormous clear-span floor space – leaving the vast **wooden ceiling** and its arching supports exposed, to stunning visual effect.

– ...

Z kolei w zadaniu 4 na tej samej karcie pracy (Rys 7) uczniowie mają przeanalizować i wyjaśnić różnice w znaczeniu obu wyrazów w szczególnych kontekstach językowych (*sklep z drewnem* a *sklep z drewna*), a następnie użyć ich w zdaniach z luką. Zarówno pary *wood/wooden* + rzeczownik, jak i zdania zostały wyszukane w korpusie Architecture.

Rys. 7. Fragment zadania ilustrującego różnice w znaczeniu wyrazów pokrewnych opracowanego na podstawie korpusu tekstów Architecture

TASK 4

1. In some cases, the words *wood* and *wooden* used before the same noun can have very different meanings. Compare the noun phrases below and explain what they mean.

wood shop	wooden shop
wood colour	wooden colour
wood profile	wooden profile
wood crew	wooden crew

2. Fill out the gaps in the sentences below with *wood* or *wooden*.
– Although the architect's design is nothing particularly special, this _____ **shop** was masterfully built by an expert carpenter.
– By the 1st century BC, _____ **screws** were commonly used throughout the Mediterranean world in devices such as oil and wine presses. Metal screws used as fasteners did not appear in Europe until the 1400s.
– ...

Inna karta pracy w tym samych module (Architecture) została poświęcona ćwiczeniu struktury gramatycznej, która sprawia uczniom duże problemy nawet na wyższych poziomach zaawansowania. W zadaniu 3 przedstawionych zostało dwadzieścia czasowników (wraz z ich częstotliwością), które najczęściej występują w tej strukturze (Rys 8). Uczeń proszony jest o połączenie ich w grupy o podobnym znaczeniu, a następnie ustalenie, czy wyłonione grupy wiążą się z konkretnymi relacjami czasowymi wyrażanymi przez strukturę. Jako kontynuacja tego zadania uczniowie mają wypełnić analizowanymi czasownikami siedem zdań z lukami. Zarówno informacje o częstotliwości, jak i zdania zostały zaczerpnięte z korpusu Architecture.

Rys. 8. Zadanie gramatyczne oparte na korpusie Architecture

TASK 3

1. Study the list of twenty most frequent verbs which occur in the Present Perfect tense in a selection of articles from professional architecture journals. Do you know all these verbs? If not, check their meaning in an online dictionary.

– been	495	– shown	16
– become	76	– evolved	14
– made	31	– done	12
– led	21	– taken	12
– had	21	– grown	11
– created	19	– changed	10
– seen	19	– given	10
– worked	18	– gone	10
– developed	17	– designed	10
– come	17	– increased	9

2. Can you see any similarities in meaning between some of the verbs in the list? Find the verbs in the list which are related in meaning to the verbs below.

– changed _____
– created _____
– come _____
– given _____
– seen _____

3. Which verbs in the list are likely to refer to "an action which started in the past and is still continuing"?

W niektórych zadaniach na innych kartach pracy uczniowie otrzymują instrukcje, jak samodzielnie szukać informacji w korpusie. W jednym z poleceń poświęconych kolokacjom rzeczownika *case* (moduł Biomedical sciences) uczniowie wyszukują w korpusie COCA przymiotniki, które występują w schemacie syntaktycznym *in* + przymiotnik + *cases*. Mają w ten sposób porównać wyniki z rezultatami podobnego wyszukiwania w korpusie Bioethics, które zostały zamieszczone na karcie pracy. Z kolei w zadaniu poświęconym profilowi kolokacyjnemu rzeczownika *noise* (moduł Automotive engineering) studenci porównują profil wygenerowany z ogólnego zasobu korpusowego SKELL[12] (Baisa i Suchomel 2014) z profilem, który tworzą samodzielnie, analizując zdania wybrane z korpusu specjalistycznego i zawierające analizowany rzeczownik. W karcie pracy w module Arts and media, zawierającej zadania rozwijające umiejętność pisania, uczniowie mają za zadanie wygenerować konkordancje spójnika *however* w podkorpusie COCA składającym się z tekstów z czasopism o tematyce społecznej i artystycznej (porównaj sekcja 2.2). Zadaniem uczniów jest zaobserwować, w którym miejscu w zdaniu analizowany spójnik występuje najczęściej (na początku, w środku czy na końcu). Uczniowie proszeni są także o prześledzenie interpunkcji związanej z użyciem tego łącznika.

12 skell.sketchengine.eu.

Podsumowanie

Opisany powyżej proces tworzenia materiałów dydaktycznych z wykorzystaniem korpusów tematycznych stwarza niezwykłe możliwości dla nauczycieli języków specjalistycznych. Należy jednak przyznać, że nie jest to przedsięwzięcie łatwe i wymaga nabycia szeregu kompetencji – technicznych, analitycznych oraz dydaktycznych. Obejmują one: 1) znajomość istniejących narzędzi korpusowych oraz umiejętność posługiwania się nimi, 2) biegłość w budowaniu adekwatnych zapytań i interpretacji ich wyników oraz 3) wiedzę, jak właściwie wykorzystać pozyskane informacje w procesie nauczania języka specjalistycznego, czyli jak zaplanować różnorodne zadania i jak je umiejętnie zintegrować z innymi strategiami nauczania (Leńko-Szymańska 2022). Nauczyciele języków obcych powinni zdobywać te kompetencje w ramach kształcenia zawodowego, a wytyczne w tym zakresie są już formułowane w literaturze przedmiotu (Farr i Leńko-Szymańska 2023).

Niestety oprócz licznych umiejętności opisanych powyżej tworzenie materiałów opartych na korpusach wymaga od nauczyciela dużych nakładów czasu i pracy, szczególnie na początku tego procesu, kiedy należy stworzyć własny zbiór dokumentów. Pojawia się nadzieja, że najnowsze narzędzia do generowania tekstu oparte na sztucznej inteligencji, takie jak ChatGPT[13], mogą ten proces uprościć i skrócić. Dają one użytkownikowi możliwość wygenerowania dowolnej liczby powiązanych tematycznie zdań lub fragmentów tekstu zawierających określony wyraz, wyrażenie bądź inny element językowy. Peter Crosthwaite i Vit Baisa (2023) zwracają jednak uwagę na wady takiego rozwiązania, związane przede wszystkim z brakiem informacji o źródłach wygenerowanych przez chatobota przykładów. W takiej sytuacji trudno ocenić, na ile są one reprezentatywne dla danej odmiany języka specjalistycznego i w związku z tym adekwatne dla potrzeb ucznia. W odróżnieniu od chatbotów korpusy tematyczne pozwalają także na dokładniejszy i bardziej zniuansowany opis danej odmiany języka poprzez udostępnianie informacji o częstotliwości i łączliwości, których brak w modelach GPT. Wdaje się więc, że umiejętność korzystania z korpusów tematycznych w celu samodzielnego tworzenia materiałów dydaktycznych do nauki języka specjalistycznego będzie jeszcze przez pewien czas niezwykle przydatna dla nauczycieli.

13 chat.openai.com.

BIBLIOGRAFIA

→ Baisa, V., Suchomel, V. (2014), SkELL – *Web Interface for English Language Learning*, [w:] *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Tribun EU, s. 63–70.

→ Biber, D., Finegan, E., Johannson, S., Leech, G. (1999), *Longman Grammar of Spoken and Written English*, Harlow: Pearson Education.

→ Boulton, A. (2010), *Data-Driven Learning: Taking the Computer Out of the Equation: Data-Driven Learning*, „Language Learning", nr 60(3), s. 534–572.

→ Boulton, A., Cobb, T. (2017), *Corpus Use in Language Learning: A Meta-Analysis: Meta--Analysis of Corpus Use in Language Learning*, „Language Learning", nr 67(2), s. 348–393.

→ Cheng, W., Lam, P. (2022), *What Can a Corpus Tell Us about Language Teaching?*, [w:] A. O'Keeffe, M. McCarthy (red.), *The Routledge Handbook of Corpus Linguistics*, Oxon i New York: Routledge, s. 319–332.

→ Chodkiewicz, H. (2011), *Nauczanie Języka Przez Treść: Założenia i Rozwój Koncepcji*, „Lingwistyka Stosowana", nr 4, s. 11–29.

→ Crosthwaite, P., Baisa, V. (2023), *Generative AI and the End of Corpus-Assisted Data--Driven Learning? Not so Fast!* „Applied Corpus Linguistics", nr 3(3), s. 100066.

→ Davies, M. (2008-), *The Corpus of Contemporary American English (COCA): 560 Million Words, 1990-Present*, <www.english-corpora.org/coca>, [dostęp: 1.10.2023].

→ Farr, F., Leńko-Szymańska, A. (2023), *Corpora in English Language Teacher Education: Research, Integration and Resources,* „TESOL Quarterly". Special Issue „Assessing the Effectiveness of Corpus-Based Approaches to English Language Teaching". doi.org/10.1002/tesq.3281.
→ Johns, T. (1991), *„Should You Be Persuaded": Two Samples of Data-Driven Learning Materials*, „Classroom Concordancing ELR Journal", nr 4, s. 1–16.
→ Kilgariff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014), *The Sketch Engine: Ten Years On*, "Lexicography", nr 1, s. 7–36.
→ Lea, D., Bradbery, J., Hornby, A.S. (2020), *Oxford Advanced Learner's Dictionary of Current English*, Tenth edition, Oxford: Oxford University Press.
→ Leńko-Szymańska, A. (2022), *Training Teachers and Learners to Use Corpora*, [w:] R.R. Jablonkai, E. Csomay (red.), *The Routledge Handbook of Corpora and English Language Teaching and Learning*, Oxon i New York: Routledge, s. 509–524.
→ Lei, L., Liu, D. (2016), *A New Medical Academic Word List: A Corpus-Based Study with Enhanced Methodology.* „Journal of English for Academic Purposes", nr 22, s. 42–53.
→ McCarthy, M., McCarten, J., Sandiford, H. (2004), *Touchstone (Four Levels)*, Cambridge: Cambridge University Press.
→ McCarthy, M., O'Dell, F. (2017), *English Vocabulary in Use. Advanced*, Third edition, Cambridge: Cambridge University Press.
→ Römer, U. (2011), *Corpus Research Applications in Second Language Teaching,* „Annual Review of Applied Linguistics", nr 31, s. 205–225.
→ Saputa, K., Tomaszewska, A., Zawadzka-Paluektau, N., Kieraś, W., Kobyliński, Ł. (2023), *Korpusomat.eu: A Multilingual Platform for Building and Analysing Linguistic Corpora*, [w:] J. Mikyška, C. de Mulatier, M. Paszynski, V.V. Krzhizhanovskaya, J.J. Dongarra, P.M.A. Sloot (red.), *Computational Science – ICCS 2023. 23rd International Conference, Prague, Czech Republic, July 3–5, 2023, Proceedings, Part II, number 14074 in Lecture Notes in Computer Science*, Cham: Springer Nature Switzerland, s. 230–237.
→ Thompson, P. (2022), *Corpus Analysis of Disciplinary Variation and the Teaching of ESP/EAP*, [w:] R.R. Jablonkai, E. Csomay (red.), *The Routledge Handbook of Corpora and English Language Teaching and Learning*, Oxon–New York: Routledge, s. 177–192.

Artykuł został pozytywnie zaopiniowany przez recenzenta zewnętrznego „JOwS" w procedurze *double-blind review.*

DR HAB. AGNIESZKA LEŃKO-SZYMAŃSKA Adiunkt w Instytucie Lingwistyki Stosowanej Uniwersytetu Warszawskiego. Prowadzi zajęcia z metodyki nauczania języków obcych i językoznawstwa korpusowego. Jest autorką wielu publikacji naukowych na temat wykorzystania korpusów w nauczaniu języków obcych oraz dotyczących korpusowej analizy procesów przyswajania języka drugiego.