PRZEGLĄD RUSYCYSTYCZNY 2022, nr 1 (177)
CC BY SA
DOI 10.31261/pr.12761

TOMASZ SZUTKOWSKI
ORCID http://orcid.org/0000-0003-3490-1707
Uniwersytet Szczeciński

W POSZUKIWANIU KOLOKACJI I ICH ODPOWIEDNIKÓW PRZEKŁADOWYCH Z WYKORZYSTANIEM NARZĘDZI ELEKTRONICZNYCH I DANYCH KORPUSOWYCH

IN SEARCH FOR COLLOCATIONS AND THEIR TRANSLATIONAL EQUIVALENTS WITH THE USE OF ELECTRONIC TOOLS AND CORPUS DATA

This paper presents a problem of stabilized word combinations (collocations) extraction and determining their translational equivalents in Polish and Russian languages. Multi-word language units have been the subject of linguistic research in various aspects of language for many years. However, still relatively few linguists and translators use in their work electronic corpora data, that in the 21st century should be considered as an important source of linguistic information, including real and potential translational equivalents. Author discusses the possibilities of using software as one of the methods of collocation extraction from text material and analyses selected units in a parallel corpus.

Przełom XX i XXI wieku naznaczony był gwałtownym rozwojem technologicznym w niemal wszystkich sferach ludzkiej działalności. Nie tylko nauki ścisłe i przyrodnicze zaczęły na wielką skalę wykorzystywać coraz większe możliwości obliczeniowe komputerów w celu gromadzenia i przetwarzania ogromnych ilości danych. Świat elektroniki nie pozostał jednak na peryferiach szeroko pojętych nauk humanistycznych, w tym także językoznawstwa, które wkroczyło w nową fazę rozwoju zwłaszcza pod szyldem lingwistyki komputerowej i korpusowej[1].

Na marginesie niniejszych rozważań pozostawiam kwestię genezy i aktualnego stanu rozwoju tej językoznawczej subdyscypliny[2].

[1] Zob. K. Church, M. Liberman, *The Future of Computational Linguistics: On Beyond Alchemy*, „Frontiers in Artificial Intelligence" 2021, vol. 4, s. 1–18.

[2] Zob. M. Świdziński, *Lingwistyka korpusowa w Polsce — źródła, stan, perspektywy*, „LingVaria" 2006, nr 1, s. 23–34.

W jej kontekście pragnę skupić się natomiast na kilku istotnych zagadnieniach, jakimi są: status wielowyrazowych jednostek języka (nazywanych niekiedy kolokacjami), możliwości ich wyodrębniania w konkretnych tekstach oraz wykorzystanie danych korpusowych w celu ustalenia optymalnych odpowiedników przekładowych. Z moich kilkunastoletnich obserwacji wynika, że w środowisku wielu językoznawców — szczególnie slawistów i rusycystów — zastosowanie narzędzi cyfrowych w badaniach nad językiem spotyka się z dość stanowczym oporem. Naturalnie, wciąż kompetencja specjalisty jest niezbędna w badaniach naukowych; bez niej choćby ogromna ilość wyekscerpowanych wartościowych danych pozostanie jedynie nieuporządkowanym i nieopracowanym materiałem. Warto jednak zwrócić uwagę na fakt, że dzięki różnorodnym programom, aplikacjom i korpusom ów materiał można nie tylko dużo szybciej pozyskać, ale przede wszystkim już na wstępnym etapie dokonać jego selekcji, segregacji czy klasyfikacji. Sposób wykorzystania technologii komputerowej w badaniach językoznawczych może mieć szeroką skalę i zróżnicowany charakter. Z pewnością nie we wszystkich tego typu działaniach niezbędna będzie profesjonalna kompetencja informatyczna, jak np. znajomość języków programowania, tworzenia złożonych baz danych, filtrów ekscerpcyjnych itp. W wielu sytuacjach wystarczy nawet podstawowa umiejętność obsługi komputera i zapoznanie się z nieskomplikowanymi narzędziami informatycznymi, aby móc skutecznie pozyskiwać oraz interpretować istotne dla badań naukowych dane. Celem niniejszego artykułu jest właśnie krótka prezentacja zautomatyzowanej metody ekscerpcji wielowyrazowych jednostek języka z dowolnych tekstów oraz ustalania ich odpowiedników na materiale polsko-rosyjskich relacji przekładowych.

Problem statusu wielowyrazowych jednostek języka nie jest w światowej lingwistyce ani nowy, ani tym bardziej zaskakujący. Można powiedzieć, że co najmniej od czasów Charlesa Bally'ego i jego koncepcji stylistyki funkcjonalnej, czyli od początku XX wieku, kwestia ustabilizowanych struktur leksykalnych na stałe wpisała się w dość wyraźnie zarysowany nurt badawczy, który później — dzięki m.in. pracom Wiktora Winogradowa — ukształtował współczesną frazeologię[3]. W skrajnej postaci przyjęła ona formę idiomatyki, jednak to nie tylko idiomy (np. „nie przebierać w słowach", „być panem sytuacji", „słomiana wdowa", „ręka rękę myje", „wyjść na światło dzienne") wy-

[3] Zob. М.А. Алексеенко, *Ещё раз о нерешённых проблемах фразеологии*, „Slavica Stetinensia" 1998, nr 8, s. 83–103.

kazują specyficzną spoistość semantyczno-strukturalną. Między nimi a tzw. swobodnymi połączeniami wyrazowymi należy dostrzec bardzo obszerny i zróżnicowany jakościowo zbiór jednostek synkretycznych, peryferyjnych, częściowo zleksykalizowanych, znajdujących się niejako w pośrednim stadium procesu idiomatyzacji[4] (np. „salon fryzjerski", „świeżo malowane", „szczyt kampanii", „szczebel wojewódzki", „wspominać nawiasowo"). Proces ten natomiast jest determinowany szczególnego rodzaju zależnościami lub inaczej — jak to ujął jeszcze w XIX wieku Jan Karłowicz — ustalonymi zwyczajem powinowactwami danej jednostki leksykalnej z innymi wyrazami[5]. Badania owych zależności czy powinowactw ma fundamentalne znaczenie dla badań językoznawczych, na co zwrócił uwagę Andrzej Bogusławski — „praktycznie zadaniem głównym", „centralną kwestią lingwistyczną", „pierwszoplanowym problemem badawczym" jest poszukiwanie i wyodrębnianie wielowyrazowych jednostek języka[6], ponieważ ich liczba przybiera paradoksalnie milionowe wartości, co w porównaniu z jednowyrazowymi jednostkami jest wielkością niewspółmiernie większą[7]. Aspekt onomazjologiczny badań nad wielowyrazowymi jednostkami podkreśla szczególnie Wojciech Chlebda:

> Jeżeli nazywanie jest aktem wieńczącym proces poznawania tego właśnie, a nie innego, wycinka, obszaru czy aspektu świata, to, chcąc mieć orientację w strukturze świata i naszych skomplikowanych z nim relacji, musimy w nieskończonym procesie poznawczym ustalać, jakie pojęcia o fragmentach tego świata jakim werbalizatorom odpowiadają. Mówiąc konkretniej: ustalmy granice werbalizatora, kształtowy zasięg i składowe danej formy słownej, a dojdziemy do ustalenia zasady odrębności i granic odpowiadającego mu pojęcia — tego właśnie, a nie innego, pojęcia o tym i tylko o tym wycinku, obszarze czy aspekcie rzeczywistości[8].

---

[4] T. Szutkowski, *Współczesna paremiografia rosyjska. Stan. Problemy. Perspektywy*, Volumina.pl, Szczecin 2015, s. 46.

[5] J. Karłowicz, *Przyczynki do projektu wielkiego słownika języka polskiego*, „Rozprawy i sprawozdania Wydziału Filologicznego Akademii Umiejętności" 1876, t. IV, s. XXVIII–XXX.

[6] A. Bogusławski, *Uwagi o pracy nad frazeologią*, w: Z. Saloni (red.), *Studia z polskiej leksykografii współczesnej*, t. 3, Wyd. Filii UW w Białymstoku, Białystok 1989, s. 18–19.

[7] Skalę i wagę tego zjawiska bardzo dobrze oddaje nowatorskie opracowanie leksykograficzne zespołu pod kierunkiem Profesora Wojciecha Chlebdy: W. Chlebda (red.), *Polsko-rosyjski słownik par przekładowych. Tom zbiorczy Podręcznego idiomatykonu polsko-rosyjskiego (z. 1–5)*, Wyd. Uniwersytetu Opolskiego, Opole 2014.

[8] W. Chlebda, *Reprodukty na warsztacie*, w: W. Chlebda (red.), *Na tropach reprodukтów. W poszukiwaniu wielowyrazowych jednostek języka*, Wyd. Uniwersytetu Opolskiego, Opole 2010, s. 9–10.

Wraz z rozwojem badań nad semantyką, strukturą i funkcją związków wyrazowych wykazujących pewien stopień leksykalizacji pojawiły się różne koncepcje dotyczące ich nazewnictwa. Trudno byłoby w granicach jednego artykułu przedstawić choćby zarys tego problemu. Można jednak wymienić co najmniej kilkanaście terminów funkcjonujących we współczesnym językoznawstwie: „frazeologizm", „frazem", „kolokacja", „odtwarzalny wielowyrazowiec"/„reprodukt"[9], „peryfraza", „analityzm", „konstrukcja opisowa", „multiwerbizm", „połączenie konwencjonalne", „połączenie frazeologiczne"[10] i in. Za każdą z tych propozycji terminologicznych stoją mniej lub bardziej przekonujące racje i stanowiska. Specyficzne rozmycie niektórych z tych pojęć wymaga od badacza niemal każdorazowego zdefiniowania używanych terminów. Taka sytuacja ma miejsce przede wszystkim w przypadku frazeologizmu, który w zależności od tzw. wąskiego lub szerokiego rozumienia frazeologii[11] może oznaczać diametralnie różne jakościowo byty językowe. Nie wdając się jednak w szczegółowe analizy, należy zauważyć, że wspólnym mianownikiem jednostek języka kryjących się za wskazanymi terminami jest przede wszystkim cecha odtwarzalności oraz ich co najmniej binarna struktura. I to właśnie konglomerat tych dwóch parametrów optymalnie oddaje termin „kolokacja" (z łac. „collocatio" — 'ułożenie, układ, umieszczenie, ustawienie'), który zaistniał w środowisku brytyjskich językoznawców w latach pięćdziesiątych XX wieku[12]. Badacze ci podkreślali głównie występowanie określonych schematów strukturalnych oraz specyficzne predyspozycje w zakresie łączliwości jednego leksemu z innymi leksemami. Istotę kolokacji trafnie oddaje definicja Tona van der Woudena: „Collocation is a term that rewers to the mechanism, or fact, that certain words are regularly found in the company of other words"[13]. Można zatem stwierdzić, że każdy leksem niejako wytwarza szczególnego rodzaju sieć wiązań z innymi jednostkami leksykalnymi, preferuje współwystępowanie jednego elementu lub kilku, a z in-

[9] W. Chlebda, *Reprodukty na warsztacie…*, s. 10.

[10] E. Białek, *Kolokacja w przekładzie. Studium rosyjsko-polskie*, UMCS, Lublin 2009, s. 8.

[11] T. Szutkowski, *Współczesna paremiografia rosyjska…*, s. 37.

[12] Zob. J.R. Firth, *Papers in Linguistics 1934–1951*, Oxford University Press, London 1957. Stan badań w zakresie kolokacji w językoznawstwie rosyjskim, polskim i zachodnioeuropejskim omawia Ewa Białek: E. Białek, *Kolokacja w przekładzie…*, s. 8–14.

[13] T. van der Wouden, *Negative Contexts. Collocation, Polarity and Multiple Negation*, Routledge, London 1997, s. 5.

nymi nie wchodzi w językową koegzystencję (np. „odnieść sukces", „ponieść porażkę", ale nie „*ponieść sukces", „*odnieść porażkę"). W zjawisku tym uczestniczy oddziaływanie — z jednej strony — uzusu, a z drugiej — normy językowej[14]. Nie zawsze jednak ich wektory są zwrócone w tym samym kierunku[15]. Kolokacjom towarzyszy również zjawisko wariantywności. Ponadto, nie wszystkie muszą odznaczać się wysoką frekwencją — niektóre z nich to okazjonalizmy, językowe efemerydy.

Ustalenie wspomnianych przez Chlebdę granic jednostek językowych jest w zasadzie prostą procedurą jedynie w odniesieniu do pojedynczych leksemów. O wiele trudniej wyodrębnić jednostki nieciągłe, wieloelementowe czy — używając nomenklatury lingwistyki komputerowej — N-gramowe[16] (dwu i więcej-gramowe: bigramy, trigramy), czyli kolokacji. Zadaniem badacza jest więc ustalenie stopnia ich spoistości semantyczno-strukturalnej, a jednym z jej głównych wskaźników będzie zjawisko reprodukowalności w mowie potwierdzone nie tylko frekwencją, ale także wartością tzw. siły kolokacji. Cel ten można osiągnąć, odwołując się do konkretnych danych tekstowych.

Językoznawca staje zatem w pierwszej kolejności przed wcale niełatwym zadaniem ekscerpcji kolokacji. Może je pozyskać na kilka sposobów. Jednym z nich jest wciąż metoda tradycyjna, „ręczna", czy też — inaczej mówiąc — fiszkowa. Z pewnością bardzo pracochłonna i czasochłonna, gdyż wymaga uważnej lektury oraz analizy mniej lub bardziej obszernego materiału tekstowego. Przyczyny, które utrudniają ekscerpcję kolokacji (wielowyrazowego reproduktu), wymienia i omawia szczegółowo Chlebda. Są to: 1) nieobecność wyraźnie zakreślonych granic początku i końca jednostki; 2) występowanie w strukturze jednostki elementów „widocznych" i „niewidocznych"; 3) wariantywność (wymienność komponentów, możliwość poszerzania lub skracania struktury wyjściowej); 4) możliwość istnienia jednostki rozrzuconej w tekście; 5) polisemia (szczególnie

[14] Zob. M. Bańko, *Słownik dobrego stylu, czyli wyrazy, które się lubią*, PWN, Warszawa 2006; E. Geller, A. Dąbrówka, *Słownik stylistyczny języka polskiego*, Świat Książki, Warszawa 2007.

[15] Wpisując w wyszukiwarkę www.google.pl niepoprawne warianty „*ponieść sukces", „*odnieść porażkę", możemy odnaleźć co najmniej kilkadziesiąt przykładów ich użycia w tekstach.

[16] Zob. D. Jurafsky, J. Martin, *Speech and Language Processing*, Prentice-Hall, New Jersey 2008.

neosemantyzacja jednostki już w języku funkcjonującej)[17]. Ten niełatwy jak widać proces można nieco przyspieszyć, wykorzystując odpowiednie narzędzia informatyczne, które jednak absolutnie nie zwalniają badacza z niezbędnej czujności, wykorzystania własnej kompetencji oraz lingwistycznej intuicji. Metoda automatycznej ekscerpcji dostarczy określonego językowego „surowca", wymagającego mimo wszystko dalszej starannej i rzetelnej obróbki w zależności od ustalonego celu badawczego.

W 2010 roku w jednym z artykułów wspólnie z Jurijem Fiedoruszkowem[18] omówiliśmy sposób ekscerpcji związków atrybutywnych z wykorzystaniem wyrażeń regularnych (*regular expressions*)[19]. Metoda ta wymaga użycia specjalnych kodów, nazywanych inaczej ankodami (ros. аношкинский код), w postaci dwuliterowych identyfikatorów odpowiadających kombinacji parametrów semantyczno-gramatycznych[20]. Procedura pozyskania zadanych modeli strukturalnych zakłada cztery etapy: 1) przygotowanie bazy ekscerpcyjnej w postaci tekstu lub tekstów (korpusu), 2) wprowadzenie tekstu do modułu analizatora morfologicznego w celu otrzymania odpowiednich danych w postaci tagów (znaczników) — do każdej jednostki leksykalnej zostaje przyłączony odpowiadający jej znacznik gramatyczny; 3) modyfikacja tekstu umożliwiająca zastosowanie filtrów ekscerpcyjnych składających się z ankodów i wyrażeń regularnych; przykładowy tekst w postaci linearnej:

> Беседа главного редактора газеты „Тамбовские известия" А.К. Клименко с одним из лидеров Тамбовского городского отделения партии пенсионеров [...]

przyjmuje postać następującego segmentu (wyraz + ankod):

> Беседа га
> главного йб

[17] W. Chlebda, *Nieautomatyczne drogi dochodzenia do reproduktów wielowyrazowych*, w: W. Chlebda (red.), *Na tropach reproduktów*..., s. 19.

[18] Ю. Федорушков, Т. Шутковски, *Лексико-грамматическая сочетаемость атрибутивных словосочетаний русского языка в контексте методов компьютерной эксцерпции*, w: R.G. Tirado, L. Sokolova, I. Votyakova (red.), *La lengua y literatura rusas en el espacio educativo internacional: estado actual y perspectivas*, Rubiños, Granada 2010, s. 1564–1569.

[19] Zob. J. Hopcroft, R. Motwani, J. Ullman, *Wprowadzenie do teorii automatów, języków i obliczeń*, przekł. B. Konikowska, PWN, Warszawa 2003.

[20] www.aot.ru (10.10.2021).

редактора аб
газеты гб
Тамбовские йт
известия еб;

4) zastosowanie zadanego ekscerptora, np. .*йб$\n.*аб*v, który wyselekcjonuje z danej próbki tekstu związki wyrazowe o strukturze ‘przymiotnik + rzeczownik’ w dopełniaczu liczby pojedynczej (np. „транспортного налога”, „прожиточного минимума”, „русского языка”)[21].

Strukturę przykładowych filtrów ekscerpcyjnych przedstawia poniższa tabela[22]:

Filtr ekscerpcyjny	**Wyekscerpowany związek wyrazowy**
^.*аа$\n.*\тйа	красивый стол
^.*аб$\n.*\тйб	красивого стола
^.*ав$\n.*\тйв	красивому столу
^.*аг$\n.*\тйг	красивый стол
^.*ад$\n.*\тйд	красивым столом
^.*ае$\n.*\тйе	красивом столе
^.*аж$\n.*\тйт	красивые столы
^.*аз$\n.*\тйу	красивых столов
^.*аи$\n.*\тйф	красивым столам
^.*ай$\n.*\тйх	красивые столы
^.*ак$\n.*\тйц	красивыми столами
^.*ал$\n.*\тйч	красивых столах

Omówiona pokrótce metoda umożliwia ekscerpowanie związków wyrazowych (w tym także kolokacji) o z góry założonej strukturze bez względu na stopień leksykalizacji ich komponentów składowych, co może być z punktu widzenia językoznawcy istotną wadą, gdyż wymaga skrupulatnej analizy całej listy otrzymanych wielowyrazowych struktur języka (selekcję danych może jednak znacząco usprawnić oprogramowanie statystyczne[23]). Jej zaletą natomiast jest z pewnością możliwość szybkiego pozyskania danych do badań nad modela-

[21] Ю. Федорушков, Т. Шутковски, *Лексико-грамматическая сочетаемость*..., s. 1567–1568.

[22] Tamże, s. 1568.

[23] Zob. P. Wierzchoń, *Pięć bardzo skutecznych (sprawdzonych) sposobów na masowe wyodrębnianie wielowyrazowych segmentów podejrzanych o frazematyczność (czyli reproduktów)*, w: W. Chlebda (red.), *Na tropach reproduktów*..., s. 106–107.

mi strukturalnymi grup wyrazowych (werbo-nominalnych, werbo--adwerbialnych, atrybutywnych itd.)[24].

Poszukiwanie kolokacji ułatwiają obecnie coraz lepiej rozwinięte narzędzia korpusowe. Dla badacza kluczową kwestią jest możliwość wyekscerpowania związków wyrazowych, które charakteryzują się już pewnym stopniem leksykalizacji, a więc nie kwalifikują się do zbioru tzw. swobodnych połączeń wyrazowych. Narodowy Korpus Języka Polskiego udostępnia użytkownikom specjalnie opracowany moduł ekscerpcyjny, nazywany kolokatorem[25]. Narzędzie to umożliwia wyszukiwanie na podstawie kryterium leksykalnego ośrodka kolokacji. Po wpisaniu w okno kolokatora leksemu „alarm" otrzymujemy następujące wyniki (poniżej zamieszczam jedynie początkowe rekordy z listy rankingowej, zob. ilustrcję na następnej stronie).

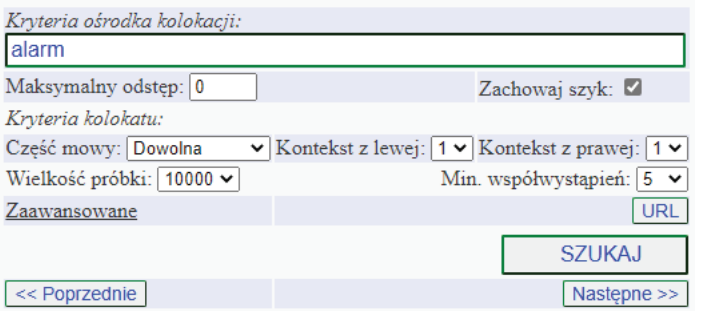
Kolokator Narodowego Korpusu Języka Polskiego podaje bardzo istotne dla badacza dane: liczbę wystąpień ośrodka kolokacji, pasujące współwystąpienia wraz z frekwencją, liczbę kolokacji spełniających zadane kryteria. Lista rankingowa jest jednak porządkowana według współczynnika Chi^2, dzięki któremu oblicza się wskaźnik siły kolokacji[26]. Współczynnik ten, w przeciwieństwie do rozkładu częstości, pozwala na wyciągnięcie poprawnych wniosków dotyczących identyfikacji rzeczywistych, a nie pozornych wielowyrazowych jednostek języka[27], gdyż informuje o statystycznej istotności korelacji między dwiema określonymi zmiennymi, którymi w tym wypadku są komponenty danej kolokacji. Naturalnie, nie można uzyskanych danych traktować bezkrytycznie. Wyekscerpowane jednostki wymagają i tym razem od badacza analizy oraz weryfikacji. Struktura powyższej przykładowej listy rankingowej daje już pewien obraz łączliwości wybranego leksemu. Mamy zatem następujące związki wyrazowe: „fałszywy alarm", „alarm powodziowy", „alarm szalupowy", „alarm przeciwpowodziowy", „alarm bombowy", „alarm pożarowy", „wszcząć alarm", „alarm wibracyjny", „alarm przeciwpowodziowy", „alarm antyterrorystyczny", „alarm przeciwlotniczy",

[24] Zob. J. Fiedoruszkow, *Metody automatyzacji ekscerpcji konstrukcji atrybutywnych języka rosyjskiego*, w: W. Chlebda (red.), *Na tropach reproduktów…*, s. 59–85.

[25] www.nkjp.uni.lodz.pl (10.10.2021).

[26] Zob. E. Babbie, *Badania społeczne w praktyce*, PWN, Warszawa 2007.

[27] M. Hebal-Jezierska, *Podstawowe zasady korzystania z korpusów przy badaniu języka*, w: W. Chlebda (red.), *Na tropach korpusów…*, s. 29.



3.	**szalupowy**	alarm___szalupowy (5),
4.	**przeciwpowodziowy**	alarm___przeciwpowodziowy (34),
5.	**bombowy**	alarm___bombowy (39), bombowy___alarm (3),
6.	**pożarowy**	alarm___pożarowy (8),
7.	**wszczął**	wszczął___alarm (21), wszczęła___alarm (16), wszczął___alarm (10), wszczęto___alarm (9), wszczęli___alarm (8), alarm___wszczęty (2), wszczęły___alarm (2), alarm___wszczął (2), alarm___wszczęły (2), wszczęłaby___alarm (1), alarm___wszczęli (1), alarm___wszczęto (1), wszczęłaby___alarm (1),
8.	**wibracyjny**	alarm___wibracyjny (8),
9.	**przeciwpożarowy**	alarm___przeciwpożarowy (11), przeciwpożarowa___alarm (1),
10.	**antyterrorystyczny**	alarm___antyterrorystyczny (3), antyterrorystyczny___alarm (2),
11.	**przeciwlotniczy**	alarm___przeciwlotniczy (10),
12.	**próbny**	próbny___alarm (14),
13.	**podnieść**	podnieśli___alarm (23), podniosła___alarm (16), podniósł___alarm (9), podniesiono___alarm (6), podnieść___alarm (6), podniosły___alarm (6), alarm___podniósł (5), podniesie___alarm (4), alarm___podniesiony (3), alarm___podniosła (2), podnieśliście___alarm (2), alarm___podnieśli (2), podniosłem___alarm (1), podniesiesz___alarm (1), podniosło___alarm (1), podniosłam___alarm (1), alarm___podniesie (1), podnieśliśmy___alarm (1), podniosłyby___alarm (1),
14.	**wszczynać**	wszczynają___alarm (4), wszczynano___alarm (3), wszczynać___alarm (3), wszczyna___alarm (3), alarm___wszczynany (1), wszczynali___alarm (1), wszczynając___alarm (1),
15.	**podniosły**	podniosła___alarm (16), podniosły___alarm (6), alarm___podniosła (2), podniosła___alarm (1),
16.	**ogłosić**	ogłoszono___alarm (28), ogłosił___alarm (10), ogłoszony___alarm (7), ogłosić___alarm (7), alarm___ogłoszono (4), ogłosili___alarm (3), ogłoszą___alarm (3), ogłosi___alarm (3), ogłosiłem___alarm (2), ogłosiła___alarm (2), ogłosiło___alarm (1), ogłosiły___alarm (1), ogłosiliśmy___alarm (1), alarm___ogłosiły (1), alarm___ogłosił (1), alarm___ogłoszony (1),
17.	**zarządzić**	zarządzono___alarm (7), zarządził___alarm (7), zarządzić___alarm (3), zarządź___alarm (2), zarządziłem___alarm (2), zarządzili___alarm (1), zarządza___alarm (1), zarządzę___alarm (1),
18.	**spłoszyć**	spłoszył___alarm (4), alarm___spłoszył (1),
19.	**ogłaszać**	ogłaszam___alarm (13), ogłasza___alarm (7), ogłaszamy___alarm (3), ogłaszając___alarm (1), ogłaszającego___alarm (1), alarm___ogłasza (1), alarm___ogłaszając (1),
20.	**włączyć**	włączyć___alarm (12), włączyła___alarm (4), włączył___alarm (4), włączy___alarm (4), alarm___włączył (4), włączony___alarm (3), włączcie___alarm (1), włącze___alarm (1), włączyliśmy___alarm (1),
21.	**bojowy**	alarm___bojowy (17), bojowej___alarm (1),
22.	**odwołać**	alarm___odwołano (10), alarm___odwołany (7), odwołano___alarm (7), odwołał___alarm (3), odwołać___alarm (2), odwołany___alarm (2), odwołali___alarm (1), odwołałem___alarm (1), odwołaliśmy___alarm (1), alarm___odwoła (1),

”

próbny alarm”, „podnieść alarm”, „wszczynać alarm” itd. Dane korpusowe odzwierciedla m.in. słownik Mirosława Bańki, który dla celów poprawnościowych odnotowuje najczęstsze kolokacje[28]. Jak widać, wiele z nich pokrywa się z wynikami próbnej ekscerpcji:

alarm (stan alarmowy) • **fałszywy** ~ • ~ **bojowy**, bombowy, lotniczy • ~ **przeciwpożarowy**, przeciwpowodziowy • **ogłosić** ~, podnieść, wszcząć, zarządzić • **bić na** ~, uderzyć na ~, narobić ~u *pot.*

alarm (urządzenie) • włączyć ~, wyłączyć

album • ~ fotograficzny, pamiątkowy, rodzinny

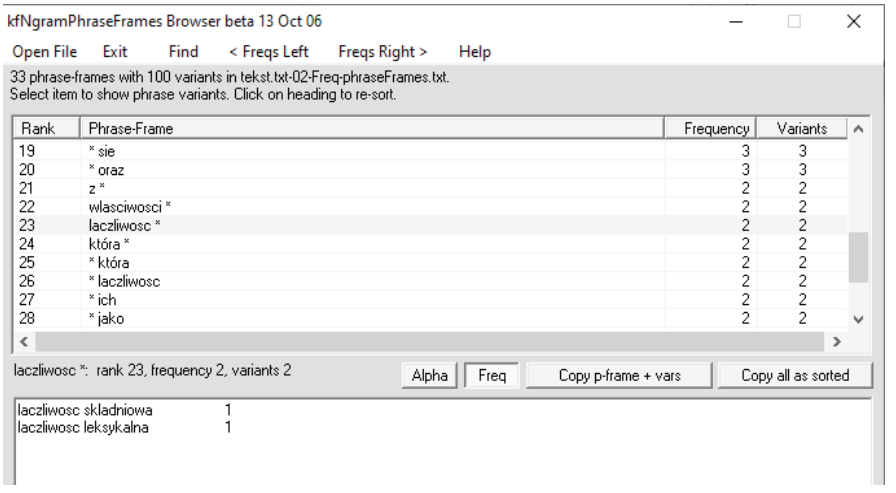
alej|a • **główna** ~a, boczna • **szeroka** ~a • **cienista** ~a • **brzozowa** ~a, kasztanowa, lipowa i in. • ~a **parkowa**, spacerowa • **iść** ~ą • ~a **prowadzi** ***dokądś***, wiedzie

[28] M. Bańko, *Słownik dobrego stylu*..., s. 3.

Poza narzędziami korpusowymi językoznawca ma także do dyspozycji inne sposoby pozyskania potrzebnego materiału badawczego, jak na przykład odrębne oprogramowanie, w którym analizie możemy poddać dowolny tekst (w korpusach zasób tekstów jest w danym momencie ograniczony i zamknięty). Mowa o programie kfNgram opracowanym przez Williama Fletchera[29] w celu generowania listy n--gramów w plikach tekstowych *txt lub HTML. Wartość „n" przybiera postać dowolnej liczby całkowitej i stanowi określoną sekwencję słów wedle zadanych parametrów. W pierwszym etapie należy sprawdzić, czy plik z przeznaczonym do badań tekstem ma nadany właściwy format (*txt lub HTML). Prostym sposobem zmiany formatu jest wklejenie tekstu do znajdującego się w akcesoriach środowiska MS Windows notatnika. Tak sformatowany materiał językowy można poddać analizie. Interfejs programu jest dość prosty w układzie i strukturze. Wybrany plik tekstowy należy wprowadzić do systemu. Przy ustawieniu minimalnego, jednokrotnego wystąpienia, wszystkie składniki tekstu (łącznie ze znakami interpunkcyjnymi) zostaną potraktowane jak oddzielne segmenty. Następnie podzielony na n-gramy plik można poddać analizie frekwencyjnej według ustalonego schematu strukturalnego, a więc np. na bigramy, trigramy, czyli związki wyrazowe 2-, 3-składnikowe. Wyniki analizy są prezentowane w formie listy rankingowej z podaną frekwencją oraz liczbą potencjalnych wariantów (zob. ilustrację na następnej stronie).

Na podstawie analizy fragmentu tekstu o tematyce językoznawczej uzyskujemy bigramy, z których część zakwalifikujemy do zbioru kolokacji, jak np. „łączliwość leksykalna", „łączliwość składniowa", „łączliwość wyrazów", „współwystępowanie wyrazów", „jednostka wielowyrazowa", „właściwości składniowe" itd. Należy oczywiście zwrócić szczególną uwagę na to, że program kfNgram podaje wyłącznie współczynnik częstości, który absolutnie nie decyduje o tym, czy dany bigram lub trigram uzyskuje status kolokacji. Na tym etapieznów konieczna jest kompetencja badacza-językoznawcy, który ów status oceni i zweryfikuje (z pewnością kolokacjami nie będą związki: „podobny wyraz", „jeden język", „typ wyrazu"). Warto zauważyć, że omawiane narzędzie informatyczne może mieć daleko szersze zastosowanie, znacznie wykraczające poza wyłączne poszukiwanie wielowyrazowych jednostek języka (np. w analizie stylometrycznej).

[29] www.kwicfinder.com/kfNgram (12.10.2021).



Kolokacje w procesie translacji odgrywają fundamentalną rolę. Dotykamy w ten sposób kluczowego zagadnienia, jakim jest pojęcie i istota jednostki tłumaczenia. Białek wyróżnia dwa główne typy tych jednostek:

> Przez leksykalną jednostkę przekładu rozumiem pojedyncze leksemy, przez ponadleksykalną zaś — związki wyrazowe, zdanie lub większe fragmenty oryginału, które tylko i wyłącznie jako całość są jednostką sensu, i tylko dla takiej całości może być dobrany odpowiednik w języku przekładu. Przy ponadleksykalnych jednostkach przekładu sens wypływa z integracji sensów wszystkich komponentów, a przekład z różnych względów na poziomie wyrazów nie jest możliwy lub nie jest wskazany[30].

W definicji tej badaczka podkreśla wagę integracji sensów komponentów wielowyrazowej jednostki języka (a tym samym jednostki tłumaczenia), której nie można zignorować w procesie przekładu, gdyż w przeciwnym razie takie tłumaczenie może wprowadzić do tekstu w języku docelowym sens niezgodny lub całkowicie sprzeczny z intencją autora oryginału. Praca tłumacza to przede wszystkim ciągłe poszukiwanie adekwatnych środków wyrazu sensu, którego nośnikami są jednostki tekstu języka wyjściowego. W kontekście omawianej problematyki kolokacji należy mieć na względzie elementarny fakt — jak podkreśla Chlebda — że mówiący zazwyczaj posługują się w swoich wypowiedziach reprodukowanymi jednostkami ponad-

[30] E. Białek, *Kolokacja w przekładzie*..., s. 46.

leksykalnymi. Jednostki jednowyrazowe w zasobach języka tworzą zdecydowanie mniej liczny zbiór[31]. To jakże ważne spostrzeżenie nakazuje tłumaczowi — z jednej strony — dostrzeganie w tekście oryginału poszczególnych kolokacji jako jednostek będących nośnikiem sensu, z drugiej natomiast — dobieranie do nich ekwiwalentnych jednostek w języku przekładu. Obecnie w epoce intensywnego rozwoju technologii komputerowych sam słownik nie jest już wystarczającym źródłem ekwiwalencji przekładowej[32]. Korzystanie z korpusów i innych elektronicznych baz danych powoli staje się w codziennej pracy tłumaczy normą i nieodłącznym elementem praktyki zawodowej. Z moich obserwacji wynika, że zwłaszcza adepci niełatwego niewątpliwie rzemiosła translatorskiego, ale także sztuki translatorskiej, niezbyt chętnie korzystają z nowoczesnych narzędzi znacząco ułatwiających i przyspieszających realizację zadania doboru ekwiwalentnych jednostek tłumaczenia (wyjątkiem jest niewątpliwie translator Google).

Bardzo praktycznym narzędziem w ekscerpowaniu kolokacji oraz poszukiwaniu ich tekstowych ekwiwalentów jest czeski korpus elektroniczny InterCorp[33], przydatny nie tylko w eksploracjach języka czeskiego, gdyż dzięki innym podkorpusom umożliwia on analizę jednostek np. w polsko-rosyjskich relacjach przekładowych. Posługując się już wcześniejszym przykładem leksemu „alarm" jako ośrodka kolokacji, można zobrazować zasób potencjalnych wielowyrazowych jednostek języka (w tym przypadku — bigramów) wyekscerpowanych w podkorpusie języka polskiego (zob. ilustrację na następnej stronie)[34].

Z tabeli tej badacz może odczytać wiele bardzo wartościowych informacji, ponieważ korpus oblicza jednocześnie kilka istotnych w analizie kolokacji wskaźników statystycznych. Oprócz frekwencji są to: wartość MI3 i MI (prawdopodobieństwo wystąpienia dwóch słów jednocześnie), T-score (tzw. miara kontrastu — im wyższy wy-

[31] W. Chlebda, *Nieautomatyczne drogi…*, s. 16.

[32] O źródłach ekwiwalentów przekładowych pisze Wojciech Chlebda: W. Chlebda, *Korpusologia użytkowa dla początkujących i zaawansowanych*, w: W. Chlebda (red.), *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*, Wyd. Uniwersytetu Opolskiego, Opole 2013, s. 12–13.

[33] https://www.korpus.cz (15.10.2021).

[34] https://www.korpus.cz/kontext/collx?maincorp=intercorp_v13_pl&viewmode=kwic&pagesize=40&attrs=word&attr_vmode=visible-kwic&base_viewattr=word&refs=%3Ddoc.id&q=~OuekoW62Iuio&cattr=word&cfromw=-5&ctow=5&cminfreq=3&cminbgr=1&cbgrfns=3&cbgrfns=m&cbgrfns=t&cbgrfns=d&cbgrfns=p&cbgrfns=r&cbgrfns=s&cbgrfns=l&csortfn=d&collpage=1 (15.10.2021).

	Filtr	word	Freq	MI3	MI	T-score	logDice ▼	MI.log_f	relative freq. [%]	min. sensitivity	log likelihood
1.	p / n	Alarm	74	25.915	13.496	8.602	10.012	58.267	20.274	0.037	1255.402
2.	p / n	fałszywy	78	25.546	12.975	8.831	9.977	56.693	14.130	0.039	1261.687
3.	p / n	Fałszywy	58	26.115	14.399	7.615	9.797	58.711	37.908	0.029	1068.886
4.	p / n	włączył	60	23.968	12.154	7.744	9.490	49.964	8.000	0.030	897.740
5.	p / n	alarm	52	22.796	11.395	7.208	9.110	45.242	4.727	0.026	721.352
6.	p / n	Czerwony	34	21.341	11.166	5.828	8.622	39.699	4.033	0.017	460.304
7.	p / n	włączy	25	22.580	13.292	5.000	8.590	43.307	17.606	0.013	415.674
8.	p / n	pożarowy	22	24.421	15.503	4.690	8.486	48.608	81.481	0.011	456.184
9.	p / n	uruchomił	19	20.728	12.232	4.358	8.139	36.644	8.444	0.010	286.030
10.	p / n	przeciwpożarowy	17	23.016	14.841	4.123	8.109	42.896	51.515	0.009	326.740
11.	p / n	wyłączyć	20	18.952	10.308	4.469	7.829	31.382	2.225	0.010	246.472
12.	p / n	biją	16	19.240	11.240	3.998	7.795	31.844	4.244	0.008	218.132
13.	p / n	czerwony	22	18.744	9.825	4.685	7.743	30.806	1.592	0.011	256.287
14.	p / n	wyłączony	15	18.964	11.150	3.871	7.702	30.915	3.989	0.008	202.596
15.	p / n	fałszywego	14	19.170	11.555	3.740	7.672	31.293	5.283	0.007	197.131
16.	p / n	bije	18	18.389	10.049	4.239	7.642	29.589	1.860	0.009	215.290
17.	p / n	Ogłaszam	13	19.632	12.232	3.605	7.638	32.280	8.442	0.007	195.656
18.	p / n	bojowy	13	19.507	12.106	3.605	7.629	31.949	7.738	0.007	193.298
19.	p / n	Ogłosić	11	21.717	14.798	3.317	7.489	36.772	50.000	0.006	210.470
20.	p / n	włączony	13	17.934	10.533	3.603	7.422	27.796	2.600	0.007	164.260

nik wskaźnika, tym większe prawdopodobieństwo, że dana kombinacja słów jest bardziej ustabilizowana, a więc może stanowić kolokację), logDice (tzw. logarytm Dice'a — pomiar częstotliwości bigramu ze średnią częstotliwością jego składowych), frekwencja względna oraz wskaźnik asocjacji (stosunek częstotliwości bigramu do częstotliwości jednego z jego składników)[35]. W zamieszczonej powyżej liście rankingowej wybrano jako wskaźnik selekcji parametr logDice, którego obliczenia siły kolokacji nie zależą od wielkości korpusu, co z kolei bardzo ułatwia porównywanie danych między różnymi bazami tekstowymi. Wskaźnik logDice przyjmuje wartość od -∞ do +14. Unifikując m.in. warianty graficzne (np. mała i duża litera) i gramatyczne, wysoki wskaźnik siły kolokacji wykazują następujące jednostki: „fałszywy alarm", „włączyć alarm", „czerwony alarm", „alarm pożarowy", „uruchomić alarm", „alarm przeciwpożarowy", „wyłączyć alarm", „bić na alarm", „ogłaszać/ogłosić alarm", „alarm bojowy", „włączony alarm". Kolokacje te przyjmują wartość logDice w przedziale (7,422, 9,977). W korpusie można zweryfikować kontekstowo każde wyekscerpowane wystąpienie analizowanej jednostki[36]:

[35] Wszystkie wskaźniki wraz ze wzorami zostały dokładnie omówione pod adresem: https://wiki.korpus.cz/doku.php/pojmy:asociacni_miry?redirect=1#miscore_a_mi3 (15.10.2021).

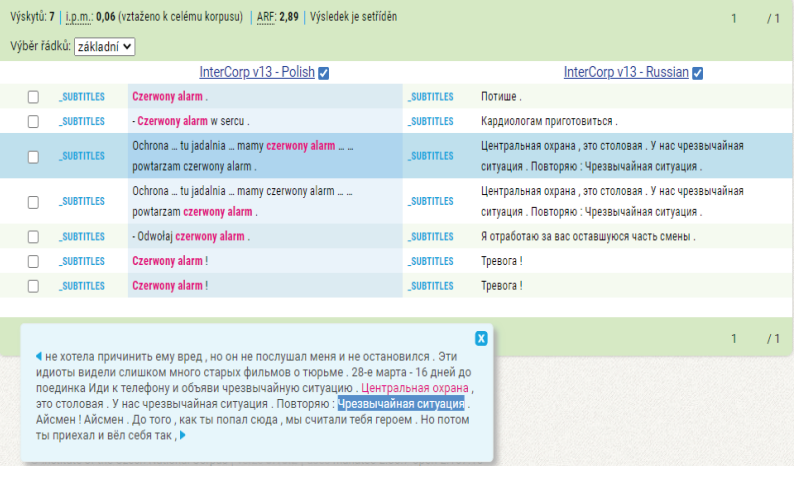
[36] https://www.korpus.cz/kontext/view?maincorp=intercorp_v13_pl&viewmode=align&pagesize=40&attrs=word&attr_vmode=visible-kwic&base_viewattr=word&refs=%3Ddoc.id&q=~asWYgci066KI (15.10.2021).

Na podstawie analizy poszczególnych przykładów i kontekstów ich użycia wyłania się kolokacja „czerwony alarm” w znaczeniu ‘stan gotowości, mobilizacja’ (synonim kolokacyjny: „czerwony alert”). Podobne badanie można przeprowadzić w innym języku. Mając na względzie polsko-rosyjskie relacje przekładowe, wykorzystamy w tym celu podkorpus języka rosyjskiego. Po wpisaniu w kolokator leksemu „тревога” uzyskujemy następujące wyniki wyszukiwania:

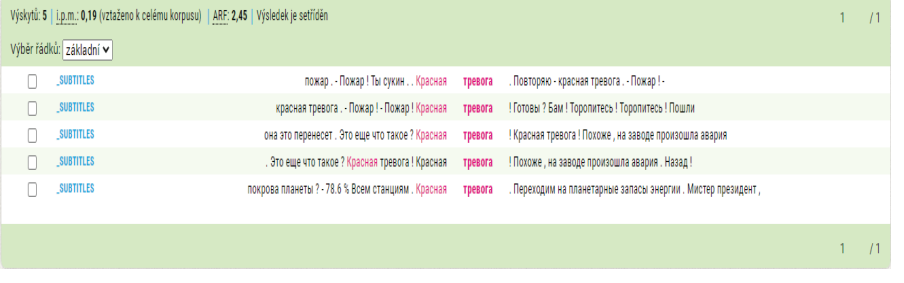
	Filtr	word	Freq	MI3	MI	T-score	logDice ▼	log likelihood	min. sensitivity	MI.log_f	relative freq. [%]
1.	p / n	вызывает	44	21.075	10.156	6.627	9.484	535.521	0.039	38.661	4.977
2.	p / n	сигнал	38	20.855	10.360	6.160	9.440	473.300	0.034	37.953	5.732
3.	p / n	Ложная	17	22.499	14.324	4.123	8.921	328.827	0.015	41.402	89.474
4.	p / n	вызывают	15	17.834	10.021	3.869	8.394	179.292	0.013	27.783	4.532
5.	p / n	Тревога	12	19.262	12.092	3.463	8.365	179.739	0.011	31.016	19.048
6.	p / n	тревога	13	17.649	10.248	3.603	8.275	159.566	0.012	27.046	5.306
7.	p / n	ложная	11	19.505	12.586	3.316	8.266	173.300	0.010	31.276	26.829
8.	p / n	поводу	32	18.240	8.240	5.638	8.204	303.086	0.013	28.812	1.319
9.	p / n	Сигнал	11	18.233	11.315	3.315	8.196	151.924	0.010	28.116	11.111
10.	p / n	воздушной	9	18.136	11.796	2.999	7.956	130.731	0.008	27.162	15.517
11.	p / n	ложной	9	17.709	11.369	2.999	7.932	125.006	0.008	26.178	11.538
12.	p / n	голосе	10	15.633	8.989	3.156	7.695	104.971	0.009	21.556	2.217
13.	p / n	объявлена	7	17.125	11.511	2.645	7.597	98.678	0.006	23.935	12.727
14.	p / n	чувство	16	15.557	7.557	3.979	7.412	136.139	0.008	21.410	0.821
15.	p / n	тепловой	6	16.917	11.748	2.449	7.393	86.697	0.005	22.860	15.000
16.	p / n	Внимание	8	14.617	8.617	2.821	7.359	79.803	0.007	18.934	1.713
17.	p / n	сигналом	6	15.748	10.578	2.448	7.332	76.431	0.005	20.583	6.667
18.	p / n	охватила	6	15.080	9.910	2.447	7.271	70.725	0.005	19.283	4.196
19.	p / n	поднять	8	14.151	8.151	2.818	7.206	74.613	0.007	17.910	1.240
20.	p / n	Повторяю	6	14.458	9.288	2.446	7.186	65.472	0.005	18.074	2.727

Przyjmując ponownie za kryterium siły kolokacji wartość wskaźnika logDice, można wyodrębnić w analizowanych tekstach następujące kolokacje: „вызывать тревогу”, „сигнал тревоги”, „ложная тревога”, „тревога по поводу”, „воздушная тревога”, „тревога в голосе”, „объявить тревогу”, „чувство тревоги”, „тревога по тепловой защите”, „тревога охватила кого-либо”, „поднять тревогу” i in.

Korpus InterCorp umożliwia również wyszukiwanie ekwiwalentów nie tylko na poziomie jednostek jednowyrazowych, ale także całych kolokacji. Analiza poszczególnych kontekstów daje tłumaczowi szerokie spektrum potencjalnych translatów, znacznie w wielu przypadkach wykraczających poza dane słowników dwujęzycznych. Po wpisaniu w okno wyszukiwarki kolokatora jednostki „czerwony alarm", korpus wyświetla następujące dane[37]:



Wyłączając z analizy tłumaczenie opisowe lub opuszczenia, na podstawie wyekscerpowanych danych można ustalić co najmniej dwie pary przekładowe: „czerwony alarm" → „чрезвычайная ситуация", „czerwony alarm" → „тревога". W podkorpusie języka rosyjskiego możemy co prawda odnaleźć jednostkę „красная тревога", jednak w korpusie paralelnym nie ma ani jednego przykładu polskiego translatu:



[37] https://www.korpus.cz/kontext/view?maincorp=intercorp_v13_pl&viewmode=align&pagesize=40&attrs=word&attr_vmode=visible-kwic&base_viewattr=word&refs=%3Ddoc.id&q=~buIQek8Sk608 (15.10.2021).

Ekwiwalencję tych dwóch kolokacji należałoby zatem zbadać na podstawie kontekstów ich użycia odrębnie w języku polskim i języku rosyjskim. Warto zwrócić uwagę na to, że słowniki polsko-rosyjskie nie odnotowują tej kolokacji. W takich sytuacjach wartość danych korpusowych jest tym bardziej wymierna i cenna.

Z powyższych rozważań wypływa kilka zasadniczych wniosków. Po pierwsze, w badaniach językoznawczych zdecydowanie większą uwagę należy poświęcić wielowyrazowym jednostkom języka w różnych aspektach — formalnym, semantycznym i funkcjonalnym (pragmatycznym). W wielu ośrodkach naukowych widać od kilkunastu lat postęp tych badań i ich konkretne rezultaty. Przykładem tego jest przede wszystkim seria idiomatykonu, nad którą intensywnie pracuje opolski zespół filologów pod kierunkiem Profesora Wojciecha Chlebdy. Powyższy postulat nie wynika absolutnie z subiektywnych preferencji lub językoznawczej „mody", lecz z obiektywnych przesłanek dotyczących ontologii języka naturalnego i jego specyfiki, o czym była mowa na wstępie niniejszego artykułu. Po drugie, mając na względzie fakt, że ilościowo kolokacje przeważają nad jednowyrazowymi jednostkami, proces ich wyodrębniania zwłaszcza dla celów leksykograficznych i przekładowych można znacząco przyspieszyć i ułatwić, stosując odpowiednie narzędzia elektroniczne. Jak starałem się to ukazać, nie są one nazbyt skomplikowane, co powinno — w moim przekonaniu — zachęcić zarówno językoznawców starszego, jak i młodszego pokolenia do korzystania z nich w codziennej pracy naukowej i dydaktycznej. Zaprezentowane dane statystyczne tym bardziej podnoszą jakość analizy, a wnioski wyprowadzone na jej podstawie stają się jeszcze bardzie obiektywne. Po trzecie, elektroniczne bazy danych w postaci korpusów (ale nie tylko) znajdują bardzo praktyczne zastosowanie w tłumaczeniach, a proces ustalania par przekładowych na podstawie realnych danych tekstowych (nie kwestionując absolutnie wartości tradycyjnych słowników dwujęzycznych) nie tylko może przyczynić się do podniesienia jakości tłumaczonych tekstów, ale także jest w stanie realnie wspomóc osiągnięcia współczesnej leksykografii przekładowej. Na koniec warto przywołać coraz bardziej aktualne i popularne w językoznawstwie pojęcie frazeotranslacji, które wydaje się swego rodzaju naukowym spoiwem powyższych konstatacji[38]. Temu zagadnieniu należałoby jednak poświęcić odrębne opracowanie.

[38] Zob. J.-P. Colson, *Computional phraseology and tranlation studiem: from theoretical hypotheses to practical tools*, w: G. Corpas Pastor, Jean-Pierre Colson (red.), *Computational Phraseology*, John Benjamins, Amsterdam-Philadelphia 2020, s. 65–81.

REFERENCES

Alekseyenko, Mikhail. "Yeshchë raz o nereshënnykh problemakh frazeologii." *Slavica Stetinensia,* 1998, no. 8: 83–103 [Алексеенко, Михаил. "Ещё раз о нерешённых проблемах фразеологии." *Slavica Stetinensia* 1998, no. 8: 83–103].

Babbie, Earl. *Badania społeczne w praktyce*. Warszawa: PWN, 2007.

Bańko, Mirosław. *Słownik dobrego stylu, czyli wyrazy, które się lubią*. Warszawa: PWN, 2006.

Białek, Ewa. *Kolokacja w przekładzie. Studium rosyjsko-polskie*. Lublin: UMCS, 2009.

Bogusławski, Andrzej. "Uwagi o pracy nad frazeologią." *Studia z polskiej leksykografii współczesnej*. Tom 3. Saloni, Zygmunt. Ed. Białystok: Wyd. Filii UW w Białymstoku, 1989: 13–30.

Chlebda, Wojciech. "Korpusologia użytkowa dla początkujących i zaawansowanych." *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*. Chlebda, Wojciech, Ed. Opole: Wyd. Uniwersytetu Opolskiego, 2013: 7–15.

Chlebda, Wojciech. "Nieautomatyczne drogi dochodzenia do reproduktów wielowyrazowych." *Na tropach reproduktów. W poszukiwaniu wielowyrazowych jednostek języka*. Chlebda, Wojciech. Ed. Opole: Wyd. Uniwersytetu Opolskiego, 2010: 15–35.

Chlebda, Wojciech. Ed. *Polsko-rosyjski słownik par przekładowych. Tom zbiorczy Podręcznego idiomatykonu polsko-rosyjskiego (z. 1–5)*. Opole: Wyd. Uniwersytetu Opolskiego, 2014.

Chlebda, Wojciech. "Reprodukty na warsztacie." *Na tropach reproduktów. W poszukiwaniu wielowyrazowych jednostek języka*. Chlebda, Wojciech. Ed. Opole: Wyd. Uniwersytetu Opolskiego, 2010: 7–13.

Church, Kenneth. Liberman, Mark. "The Future of Computational Linguistics: On Beyond Alchemy." *Frontiers in Artificial Intelligence,* 2021, vol. 4: 1–18.

Colson, Jean-Pierre. "Computional phraseology and tranlation studiem: from theoretical hypotheses to practical tools." *Computational Phraseology*. Corpas Pastor, Gloria. Colson, Jean-Pierre. Eds. Amsterdam-Philadelphia: John Benjamins Publishing Company, 2020: 65–81.

Fedorushkov, Yuriy, and Shutkovski, Tomash. "Leksiko-grammaticheskaya sochetayemost' atributivnykh slovosochetaniy russkogo yazyka v kontekste metodov komp'yuternoy ekstserptsii." *La lengua y literatura rusas en el espacio educativo internacional: estado actual y perspectivas*. Tirado, Rafael Guzmán, and Sokolova, and Larisa, and Votyakova, Irina. Eds. Granada: Rubiños, 2010. 1564–1569 [Федорушков, Юрий, and Шутковски, Томаш. "Лексико-грамматическая сочетаемость атрибутивных словосочетаний русского языка в контексте методов компьютерной эксцерпции." *La lengua y literatura rusas en el espacio educativo internacional: estado actual y perspectivas*. Tirado, Rafael Guzmán, and Sokolova, Larisa Votyakova, and Irina. Eds. Granada: Rubiños, 2010: 1564–1569].

Fiedoruszkow, Jurij. "Metody automatyzacji ekscerpcji konstrukcji atrybutywnych języka rosyjskiego." *Na tropach reproduktów. W poszukiwaniu wielowyrazowych jednostek języka*. Chlebda, Wojciech. Ed. Opole: Wyd. Uniwersytetu Opolskiego, 2010: 59–85.

Firth, John Rupert. *Papers in Linguistics 1934–1951*, London: Oxford University Press, 1957.

Geller, Ewa. Dąbrówka, Andrzej. *Słownik stylistyczny języka polskiego*. Warszawa: Świat Książki, 2007.
Hebal-Jezierska, Milena. "Podstawowe zasady korzystania z korpusów przy badaniu języka." *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*. Chlebda, Wojciech. Ed. Opole: Wyd. Uniwersytetu Opolskiego, 2013: 17–30.
Hopcroft, John. Motwani, Rajeev. Ullman, Jeffrey. *Wprowadzenie do teorii automatów, języków i obliczeń*. Warszawa: PWN, 2003.
Karłowicz, Jan. "Przyczynki do projektu wielkiego słownika języka polskiego." *Rozprawy i sprawozdania Wydziału Filologicznego Akademii Umiejętności* 1876, tom 4: 28–30.
Jurafsky, Dan. Martin, James. *Speech and Language Processing*. New Jersey: Prentice-Hall, 2008.
Szutkowski, Tomasz. "Subkompetencja frazeologiczna studentów-neofilologów jako składnik kompetencji językowej tłumacza. Badanie pilotażowe." *Roczniki Humanistyczne,* 2019, t. LXVII, z. 7: 93–109.
Szutkowski, Tomasz. *Współczesna paremiografia rosyjska. Stan. Problemy. Perspektywy*. Szczecin: Volumina.pl, 2015.
Świdziński, Marek. "Lingwistyka korpusowa w Polsce — źródła, stan, perspektywy." *LingVaria,* 2006, no. 1: 23–34.
Wierzchoń, Piotr. "Pięć bardzo skutecznych (sprawdzonych) sposobów na masowe wyodrębnianie wielowyrazowych segmentów podejrzanych o frazematyczność (czyli reproduktów)." *Na tropach reproduktów. W poszukiwaniu wielowyrazowych jednostek języka*. Chlebda, Wojciech. Ed. Opole: Wyd. Uniwersytetu Opolskiego, 2010: 87–125.
Wouden, Ton van der. *Negative Contexts. Collocation, Polarity and Multiple Negation*. London: Routledge, 1997.