

Korpusowe narzędzia weryfikacji frazeostylistycznej tłumaczeń

Piotr Pęzik

Uniwersytet Łódzki

<https://orcid.org/0000-0003-0019-5840>

piotr.pezik@uni.lodz.pl

Corpus tools of phraseo-stylistic verification of translation

Phraseological prefabrication can be defined as the tendency for native speakers to select recurrent word combinations, constructions and formulaic sequences to convey familiar meanings and achieve conventional discourse functions. Since such formulaicity is an essential aspect of fluency and intelligibility of naturally-occurring language, a good command of target language phraseology is also a prerequisite of successful translation. This paper provides an overview of selected corpus-based tools and resources which can be used to validate the idiomaticity of translations. In addition to discussing the role of large monitor and reference and parallel corpora and online collocation databases, the paper introduces a new tool called Treelets which enables extraction of combinatorial dictionaries from dependency-parsed corpora. This feature proves useful in enhancing translators' phraseological competence.

Keywords: corpus tools; idiomaticity; phraseme; phraseological index; translational equivalence; translation quality; translation strategies

1. Wstęp

Prefabrykacja frazeologiczna jest jednym ze stałych aspektów użycia języka. Bez setek tysięcy utrwalonych połączeń wyrazowych, tj. kolokacji, idiomów, powiedzeń, wielowyrazowych markerów dyskursu oraz częściowo utrwalonych konstrukcji

leksykalno-gramatycznych, niemożliwe byłoby utrzymanie właściwej dla różnych odmian i rejestrów komunikacji językowej poziomów płynności i jednoznaczności tekstów (Pawley, Syder, 1983). Z perspektywy odbiorcy języka, użycie jednostek frazeologicznych redukuje wysiłek kognitywny wkładany w ujednoznacznienie komunikatów językowych. Choć trudno ocenić dokładnie stosunek liczby frazemów (skonwencjonalizowanych, odtwarzanych z pamięci kombinacji wielowyrzowych) do liczby struktur czysto kompozycyjnych i doraźnie konstruowanych (syntagm), to nie ulega wątpliwości, że właściwe rozpoznanie frazemów w tekście oryginału, a także ich odpowiednie użycie w tekście przekładu jest również jednym z podstawowych parametrów jakości tłumaczenia. Zarówno tekst źródłowy, jak też wynik tłumaczenia cechuje pewien stopień idiomatyczności, który można określić mianem indeksu frazeologicznego, czyli poziomu nasycenia jednostkami frazeologicznymi. O ile dla tekstu źródłowego wartość indeksu frazeologicznego jest dana, to już poziom nasycenia tłumaczenia konstrukcjami frazeologicznymi może się istotnie wahać, w zależności od ogólnej strategii obranej przez tłumacza oraz doboru ekwiwalentów poszczególnych frazemów z oryginału.

Frazeologie analityczne dostarczają licznych typologii jednostek frazeologicznych (Chlebda, 1991)¹. W kontekście decyzji frazeostylistycznych tłumacza można przyjąć stosunkowo prosty podział frazemów ze względu na stopień wyrazistości w tekście oryginału oraz konsekwencje ich niewłaściwego przekładu. Do pierwszej grupy można zaliczyć związki frazeologiczne, czyli frazemy niekompozycyjne znaczeniowo, takie jak idiomy czyste i figuratywne. Niewłaściwe rozpoznanie ich statusu jako prefabrykatów językowych może skutkować ewidentnym brakiem ekwiwalencji znaczeniowej między tekstem przekładu a tłumaczeniem. Ilustrują to przykłady 1-4, ukazujące dwa tłumaczenia angielskiego idiomu figuratywnego *hit the nail on the head*. Oba pochodzą z amatorskich tłumaczeń napisów filmowych udostępnionych w serwisie *opensubtitles.org*, które zostały jednocześnie zindeksowane w otwartym korpusie równoległym Paralela (Pęzik, 2016). W pierwszej parze zdań polskim odpowiednikiem angielskiego idiomu jest idiomatyczne wyrażenie *trafić w sedno*. Można je uznać za udany ekwiwalent frazeologiczny zarówno pod względem stopnia prefabrykacji, jak też zakresu znaczeniowego oryginalnego idiomu.

1. O: I think you've *hit the nail on the head* (Paralela, Judas: One of Twelve, 2007).
2. T: Myślę, że *trafiła pani w sedno* (*opensubtitles.org*).

Z kolei tłumaczenie tego samego angielskiego idiomu (ukazane w przykładach 3-4) wydaje się być niepożądanym skutkiem nierozpoznania jego statusu jako jednostki frazeologicznej. Tłumacz próbuje lokalnie zinterpretować znaczenie frazemu, zamiast

¹ Terminy *jednostka frazeologiczna*, *frazem* i *prefabrykat* są tu używane wymiennie.

odwołać się do jego skonwencjonalizowanego znaczenia. W rezultacie znaczenie i funkcja dyskursywna oryginalnego tekstu ulega całkowitemu rozmyciu w tłumaczeniu.

3. O: You've *hit the nail on the head* there (Keeping Mum, 2005).

4. T: Tym *mnie dobiłaś* (Paralela, opensubtitles.org).

Osobną grupę frazemów o nieco innej charakterystyce z perspektywy przekładu stanowią częściowo lub całkowicie kompozycyjne połączenia wielowyrazowe, np. otwarte kolokacje leksykalne, a także łańcuchy kolokacyjne². Ich tłumaczenie może nastroczać nieco subtelniejszych problemów w warstwie frazeostylistycznej tekstu. Przykładem otwartej kolokacji angielskiej może być fraza *fair question*. Jest to stosunkowo kompozycyjne połączenie wyrazów autosemantycznych, tzn. występujących w swoich ogólnych (a nie właściwych dla tej kombinacji) znaczeniach. Cowie, Mackin i McCaig (1993, s. xiv) omawiając status leksykograficzny tej kolokacji, zastanawiają się nad potrzebą jej włączenia do słowników frazemów angielskich i dochodzą do wniosku, że za ostateczną decyzją o uwzględnieniu tej frazy w słowniku przemawia przede wszystkim jej częste występowanie w różnego typu debatach i dyskusjach. Analiza korpusowa tego wyrażenia potwierdza słuszność uznania frazy *fair question* za kolokację, którą cechuje przede wszystkim kontekstowa stereotypowość, a nie kryteria formalne. Niska wyrazistość otwartych kolokacji utrudnia rozpoznanie ich statusu jako jednostek skonwencjonalizowanych, które wymagają często w podobnym stopniu utrwalonych ekwiwalentów. Świadczą o tym zamieszczone poniżej przykłady amatorskiego tłumaczenia kolokacji *fair question*, której ekwiwalentami stają się odpowiednio nominalne frazy *uczciwe pytanie* (7) oraz *szczerze pytanie* (8). Pierwszy z odpowiedników wydaje się być doraźną syntagmą wynikającą z dosłownego tłumaczenia znaczenia przymiotnika *fair*.

5. *It's a fair question* (Paralela, Land of the Lost, 2009).

6. *To uczciwe pytanie* (Paralela, opensubtitles.org).

W drugim przypadku ekwiwalentem staje się regularna, choć nietrafiona w tym kontekście polska kolokacja *szczerze pytanie*. Z rezultatów wyszukiwania w korpusie Paralela wynika bowiem, że regularnym ekwiwalentem frazeologicznym angielskiej formuły (*It's a fair question*) jest dużo częściej fraza *dobrze pytanie*.

7. What's the point? *It's a fair question*; in a sense, you all have been studying for the last three weeks (Paralela, Role Play, 2012).

² Mianem łańcucha kolokacyjnego określane jest zazębianie składniowo występowanie dwóch lub więcej kolokacji binarnych (por. Tabela 6).

8. Jaki w tym cel? *To szczerze pytanie*, biorąc pod uwagę to, że wszyscy uczyliście się przez ostatnie trzy tygodnie (Paralela, opensubtitles.org).

Nawet jeżeli przetłumaczenie pojedynczej kolokacji nieskonwencjonalizowanym lub nietrafionym znaczeniowo odpowiednikiem nie powoduje wyraźnego obniżenia jakości całego tekstu tłumaczenia, to jednak nagromadzenie takich do-
rażnie konstruowanych ekwiwalentów dla utrwalonych w języku oryginału połą-
czeń wyrazowych obniża walory stylistyczne przekładu jako całości. W rezultacie interpretacja tekstu tłumaczenia wymaga większego wysiłku poznawczego i jest obciążona dodatkowym ryzykiem błędu. Istotną kompetencją tłumacza jest więc umiejętność tworzenia idiomatycznych tekstów przekładu. Brak tej kompetencji widać w cytowanych powyżej amatorskich tłumaczeniach napisów filmowych, których autorzy nie potrafili dobrać utrwalonych i jednoznacznych ekwiwalentów frazeologicznych dla występujących w oryginale frazemów.

Celem tego artykułu jest przedstawienie otwartych korpusowych zasobów i narzędzi do badań frazeologii, które mogą być przydatne zarówno w procesie kształcenia tłumaczy, jak też w codziennej pracy w warstwie frazeostylistycznej przekładu. W pierwszej jego części omówione zostają systemy wyszukiwawcze dostępne dla wybranych korpusów referencyjnych, monitorujących i równoległych, ze specjalnym uwzględnieniem zagadnień reprezentatywności tematycznej, gatunkowej i diachronicznej tych zasobów. Część druga artykułu skupia się na narzędziach do ekstrakcji i tzw. detekcji frazeologii z korpusów referencyjnych i domenowych oraz ich możliwym zastosowaniu w warsztacie tłumacza. Warto tu zaznaczyć, że celem artykułu nie jest omówienie funkcji i roli popularnych systemów komputerowego wspomagania tłumaczenia (ang. *computer aided translation*), które w roku 2019 można uznać za szeroko spopularyzowane, udokumentowane i zapewne doskonale znane ogromnej większości praktykujących tłumaczy. Pomimo intensywnego rozwoju tego typu oprogramowania, baz terminologicznych oraz systemów wyszukiwawczych dla zbiorów tekstów równoległych, wydaje się jednak, że niektóre metody, zasoby i narzędzia stanowiące osiągnięcia ogólnego językoznawstwa korpusowego odgrywają niezastąpioną rolę w weryfikacji idiomatyczności tłumaczeń.

2. Korpusowe systemy wyszukiwawcze

Systemy wyszukiwawcze dostępne dla dużych korpusów referencyjnych, takich jak Narodowy Korpus Języka Polskiego (NKJP) (Pęzik, 2012a), mogą być używane do potwierdzenia statusu frazeologicznego kombinacji wielowyrazowych w procesie tłumaczenia. Pęzik (2012b) opisuje tego rodzaju zastosowanie korpusów referencyjnych na przykładzie wyszukiwarki PELCRA dla NKJP. Ze względu na

zrównoważoną reprezentację rejestrów i typów funkcjonalnych tekstów korpus narodowy stanowi „niezastąpione źródło informacji o użyciu, rejestrze, konotacjach oraz łączyłości semantycznej i frazeologicznej wyrazów, zwrotów, form i konstrukcji gramatycznych” (Pęzik, 2012b, s. 301). Ponieważ jednak od zakończenia prac nad pierwszą edycją, NKJP stracił nieco na swej aktualności, a na kolejne jego wydanie trzeba będzie jeszcze jakiś czas zaczekać, warto w tym miejscu zilustrować zastosowanie korpusowych systemów wyszukiwawczych w pracy tłumacza na przykładzie korpusu MoncoPL (Pęzik, w druku)³, który w zawężonym stopniu pełni funkcję korpusu referencyjnego polszczyzny po roku 2010. Indeks MoncoPL zawiera próbki tekstów z serwisów informacyjnych o średniej wielkości dziennej ok. 2 mln segmentów słów. Pod koniec 2019 r. wyszukiwarka tego korpusu dawała dostęp do zbioru ponad 6 mld segmentów słów tekstów opublikowanych w latach 2010–2019. Taki rozmiar i zakres czasowy korpusu daje szansę na odnotowanie innowacji leksykalnych i frazeologicznych, które przed rokiem 2010 nie istniały lub dopiero zaczynały wchodzić do powszechnego użycia. Na przykład wielowyrazowy termin *lokowanie* produktu, którego domyślnym ekwiwalentem może być angielski termin *product placement*, w całej puli danych NKJP występuje zaledwie pięciokrotnie, w tym trzykrotnie w części zrównoważonej. W korpusie MoncoPL to samo wyrażenie występuje aż 2 059 razy (do końca października 2019 r.). Mimo iż MoncoPL ma objętość trzykrotnie większą od całej puli danych NKJP i zawiera głównie teksty informacyjne, to jednak dysproporcji tej nie da się wyjaśnić jedynie różnicami w objętości i kompozycji gatunkowej obu korpusów. Wynika ona przede wszystkim z upowszechnienia i utrwalenia się wielowyrazowego terminu *lokowanie produktu*, który zaczął występować szczególnie często w dłuższych sekwencjach formułicznych, takich jak *audycja/ tekst/ wypowiedź zawiera lokowanie produktu* dopiero po roku 2010. Potrzeba aktualizacji nawet najlepiej zrównoważonych korpusów referencyjnych uwidacznia się na tym i kilku dalej omawianych przykładach związków wielowyrazowych.

Tabela 1 Wybrane korpusy referencyjne, monitorujące i równoległe języka polskiego i angielskiego

Wyszukiwarka	Adres	Język	Wielkość korpusu w mln. słów	Typ korpusu	Zakres czasowy
PELCRA NKJP	nkjp.uni.lodz.pl	PL	1 800	Referencyjny	1918-2010
PELCRA II NKJP	pelcra.clarin-pl.eu/NKJP	j.w.	j.w.	j.w.	j.w.
MoncoPL	monco.frazeo.pl	PL	6 000 w 2019 r.	Monitorujący	Od 2010 r.
MoncoEN	monco.frazeo.com	EN	11 000 w 2019 r.	Monitorujący	Od 2015 r.
Paralela	paralela.clarin-pl.eu	PL-EN	320	Równoległy	n.d.

³ Zob. monco.frazeo.pl.

3. Wariantywność frazemów

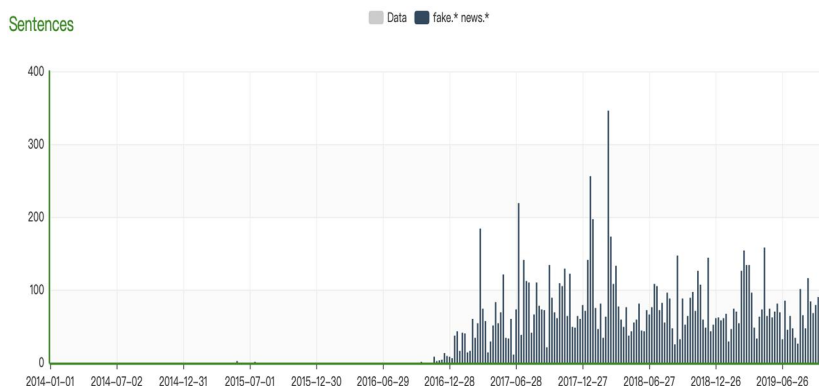
Wiele pozornie stałych związków wyrazowych cechuje wariantywność na poziomie ortograficznym, morfologicznym, leksykalnym i składniowym (Lewicki, 2003). Określenie najbardziej prototypowej postaci frazemów, które weszły do powszechnego użycia w ostatnich latach może okazać się trudne lub niemożliwe bez odwołania do aktualnych referencyjnych zbiorów danych korpusowych. Poziom idiomatyczności tłumaczenia, podobnie jak tekstów oryginalnych, może zależeć również od doboru najczęstszego wariantu danego frazemu. Istotność tej tendencji dla procesu tłumaczenia można zilustrować przykładem dwóch zbiorów kolokacji o podobnym znaczeniu, ale różnej proveniencji czasowej, zakresie znaczeniowym i stopniu skonwencjonalizowania. Pierwszy zestaw kolokacji można wydobyc z korpusów NKJP, MoncoPL i Paralela następującym zapytaniem:

falszywy**|nieprawdziwy** informacja**|wiadomość**

Aktualna wersja korpusu MoncoPL zwraca dla tego zapytania ponad 27 tys. wyników. Z kolei w całym NKJP znajdujemy niemal 4 tys. wystąpień. Różne kombinacje wyrazów pasujących do tego zapytania układają się w podobne znaczeniowo kolokacje, takie jak *nieprawdziwa informacja* czy *falszywa wiadomość*. W segmentach tłumaczeniowych zindeksowanych w korpusie Paralela są one regularnymi odpowiednikami angielskich kolokacji *inaccurate/incorrect/deceptive/false information*, również w kontekstach, gdy znamionowane przez nie fałszywe informacje dotyczą sfabrykowanych doniesień prasowych lub internetowych. Ilustruje to poniższa para segmentów z tego korpusu równoległego:

9. O: Iraq was invaded and occupied on the basis of *deliberately deceptive information* concerning the presence of weapons of mass destruction (Paralela, EP Proceedings).
10. T: Irak został zaatakowany i zajęty po celowym podaniu *nieprawdziwych informacji* o istnieniu broni masowego rażenia.

Tę komfortową dla tłumacza sytuację komplikuje pojawienie się terminu *fake news*. Jak wynika z analizy szeregów czasowych wygenerowanych z korpusu MoncoPL, ta denotatywna kolokacja mniej więcej w październiku 2016 r. zaczęła być regularnie używana w polskich tekstach prasowych (por. rys. 1).



Rysunek 1 Szereg czasowy częstości wystąpień (z granulacją tygodniową) dla wyrażenia *fake news* w korpusie MoncoPL

Co ciekawe, w całej puli danych NKJP termin ten nie występuje w ogóle, a w korpusie Paralela pojawia się na tyle sporadycznie, że trudno mówić o jego regularnych polskich odpowiednikach. Poniższe zapytanie do korpusu MoncoPL (użyte również do opracowania w/w wykresu szeregu czasowego) pozwala na analizę użycia tego zapożyczonego z angielszczyzny określenia:

fake.|fejk.*news.*|nius.*|njus.**

Intencją zapytania jest wyszukanie różnych kombinacji form wyrazowych zaczynających się od ciągu znaków *fake* lub *fejk*, po których występują różne możliwe zapisy i odmiany angielskiego wyrazu *news*, także w jego spolszczonej ortograficznie postaci *nius* lub *njus*. Analiza konkordancji wybranych spośród ponad 11 tys. wystąpień zwracanych przez wyszukiwarkę MoncoPL wskazuje, że to angielskie zapożyczenie funkcjonuje w polszczyźnie jako pseudo termin, tzn. fraza używana w funkcji terminu, ale zazwyczaj w sposób uznaniowy i bez jednoznacznego odwołania do normatywnych źródeł jego definicji, jeśli nie liczyć użyć odnoszących się wprost do nowych rozwiązań prawnych mających na celu zwalczanie dezinformacji. Mimo tej częściowej nieostrości i pewnych normatywnych oporów⁴ nietrudno sobie wyobrazić konteksty, w których tłumacz mógłby się zdecydować na oddanie angielskiej frazy *fake news* za pomocą spolszczonej formy frazy *fake news(y)*, aby jednoznacznie odwołać się do zjawiska dezinformacji

⁴ O obiekcjach wobec stworzenia hasła słownikowego dla *fake newsów* piszą między innymi autorzy Dobrego Słownika (<https://dobrysloownik.pl/slowo/fake+news/224024/#znaczenie-239461>). Wydaje się jednak, iż wykazana powyżej powszechność tej frazy oraz brak synonimicznego określenia o podobnej funkcji terminologicznej nakazują takie opory przewyżczać.

w epoce mediów społecznościowych⁵. Z punktu widzenia stylistyki tłumaczenia dodatkowych problemów może nastręczać wskazanie prototypowej postaci tej frazy. Trudno jest czysto intuicyjnie wskazać jej najczęściej używaną postać. Przymiotnik *fake* może bowiem występować w mniej lub bardziej znaturalizowanej postaci ortograficznej i morfologicznej (*fake*, *fejk*, *fejkowe*). Również rzeczownik *news* może być używany nieodmiennie, spolszczany ortograficznie (*njus*, *njusy*), a także odmieniany przez liczbę i przypadek. Tabela 2 przedstawia wyniki pięciu oddzielnych zapytań o warianty frazy *fake news* w korpusie MoncoPL. W drugiej kolumnie ukazane są warianty form głównych, których zestawienie zostało wygenerowane za pomocą funkcji *Podsumowanie*, dostępnej dla wyników dowolnego zapytania. Z zestawienia wynika, iż zdecydowanie najczęstszym wariantem bazowym jest forma zachowująca oryginalną angielską pisownię *fake news*, ale jednocześnie dopuszczająca regularną odmianę rzeczownika *news* (jeżeli przyjąć taki rozbiór składniowy polskiej wersji tej frazy).

Tabela 2 Warianty ortograficzne i morfologiczne frazy *fake news* w korpusie MoncoPL

#	Zapytanie	Przykładowe formy	Częstość
1	fake.* news.*	fake news (3806), fake newsów (2026), fake newsy (1949),...	11 167
2	fejk.* news.*	fejk news (17), fejk newsów (10)...	93
3	fejk.* njus.*	fejk njus (13), fejk njusami (10)...	37
4	fejk.* njus.*	fejk njut (5), fejk njusów (2)...	15
5	fake.* njus.*	fake njusów (3), fake njusami (2)	5

Jak widać na powyższym przykładzie, agregacja tysięcy wystąpień danej frazy w aktualnym korpusie pozwala określić prototypowe warianty stosunkowo nowych leksemów i frazemów w sposób bardziej wiarygodny niż zwykłe odwołanie się do intuicji tłumacza, czy też anegdotycznych zaświadczeń w przypadkowo znalezionych źródłach. Również użycie ogólnego przeznaczenia wyszukiwarek internetowych jako narzędzia do sporządzenia listy frekwencyjnej wariantów leksemów i frazemów jest ryzykowne ze względu na nieprzejrzysty mechanizm indeksowania danych i dostosowywanie wyników wyszukiwania do preferencji indywidualnych użytkowników. Przydatnym narzędziem okazuje się stale aktualizowany korpus monitorujący, dla którego istnieje odpowiedni system wyszukiwawczy.

⁵ Analiza korpusów diachronicznych wykazuje, że w angielszczyźnie fraza *fake news* (początkowo *false news stories*) dopiero na początku XXI wieku zaczęła wypierać nieco starsze określenie *false news* (Allan Metcalf, 2017).

4. Ekstrakcja frazeologii

Poza funkcją generowania konkordancji dla doraźnych zapytań o słowa, frazy i konstrukcje gramatyczne, wiele wyszukiwarek korpusowych posiada również moduł ekstrakcji dwuwyrazowych kolokacji pozycyjnych oraz n-wyrazowych sekwencji wyrazów, czyli tzw. n-gramów wyrazowych. Przykładem takiego rozwiązania może być dostępne w wyszukiwarce PELCRA dla NKJP narzędzie Kolokator (Pęzik, 2012b), które dla zadanego ośrodka kolokacji (tj. pojedynczego wyrazu lub frazy definiowanej korpusową składnią zapytania) wyszukuje wyrazy występujące w jego pobliżu (tzw. kolokaty), a następnie agreguje listę frekwencyjną po ich formach podstawowych. Podobną funkcję generowania tabeli typowych połączeń wyrazowych dla jedno- i wielowyrazowych ośrodków kolokacji pełni narzędzie Kolokacje dostępne w korpusie MoncoPL. Tabela 3 ukazuje listę kolokatów rzeczownikowych wygenerowanych za pomocą tego narzędzia z 541 wystąpień różnych form przymiotnika *fejkowy*. Ośrodek kolokacji został zdefiniowany zapytaniem *fejkow.**, ponieważ w słowniku używanym przez wyszukiwarkę do rozszerzeń morfologicznych nie odnotowano jeszcze takiej formy bazowej (lematu). Symbole *** definiują zero lub więcej dowolnych znaków, które mogą wystąpić w terminie *zapytania*. Na liście kolokatów pojawia się kilka powtarzalnych kombinacji, takich jak *fejkowe konto* czy też *fejkowy profil*, które pochodzą głównie z dziedziny komunikacji internetowej i mediów społecznościowych.

Tabela 3 Rzeczownikowi modyfikowane przez przymiotnik *fejkowy* w korpusie MoncoPL

#	Kolokat	Częstość
1	konto	138
2	profil	42
3	informacja	17
4	news	16
5	strona	9
6	mem	8
7	post	7
8	wiadomość	6
9	wpis	5
10	konkurs	5

Przykładem tłumaczenia, w którym użycie modułu ekstrakcji kolokacji pozwala porównać stopień utrwalenia związków wyrazowych występujących w oryginale i tłumaczeniu może być zacytowany poniżej fragment artykułu z portalu sport.pl. Ukazany w przykładzie 12 urywek tego artykułu jest w zasadzie bezpośrednim tłumaczeniem komunikatu Angielskiej Federacji Piłki Nożnej, opublikowanego na portalu społecznościowym Twitter 14 października 2019 r.

11. O: We can confirm that @England players were *subjected to abhorrent racist chanting* while playing in the #EURO2020 qualifier against Bulgaria. This is unacceptable at any level of the game and our immediate focus is supporting the players and staff involved (Twitter).
12. T: Angielska federacja wydała oświadczenie. - Możemy potwierdzić, że gracze z Anglii byli *poddani obrzydliwym rasistowskim śpiewom* podczas meczu eliminacji Euro 2020 przeciwko Bułgarii. To niedopuszczalne, a naszym bezpośrednim celem jest wsparcie graczy i personelu (sport.pl, 2019.10.14)⁶.

Jednym z problemów w warstwie stylistycznej tego tłumaczenia jest użycie konstrukcji *poddani (obrzydliwym rasistowskim) śpiewom* jako ekwiwalentu angielskiej frazy *subject(ed) to chants*. Jak widać na liście frekwencyjnej wygenerowanej narzędziem Collocations z korpusu MoncoEN (Tabela 4), rzeczownikowymi kolokacjami wyrażenia *subject(ed) to* bywają wyrazy o zdecydowanie negatywnym znaczeniu, np. *subjected to abuse/attack/violence/harassment*. Wydaje się, że zaświadczone w korpusie MoncoPL kolokacje wypełniające wzorzec *poddany/poddać się* + celownik również cechuje pewien stopień negatywnego zabarwienia, np. *poddany karze/egzekucji/torturom*. Nawet rzeczowniki o pozornie pozytywnym lub neutralnym znaczeniu, tj. *poddany terapii/leczeniu/renowacji* implikują negatywny stan podmiotu kwalifikowanego tym określeniem. Być może między innymi z tego względu niezaświadczone w dużych korpusach polszczyzny (takich jak MoncoPL czy NKJP) konstrukcja *poddani śpiewom* brzmi nieprzekonująco. Innymi słowy, rzeczownik *śpiew(y)* nie wywołuje jednoznacznie negatywnych skojarzeń i z pewnością nie tworzy regularnych kombinacji składniowych jako dopełnienie celownikowe czasownika *poddąć*. Co ciekawe, w korpusie MoncoEN dłuższa konstrukcja użyta w oryginalnym oświadczeniu Angielskiej Federacji Piłki Nożnej, tj. *subjected to chants* występuje 60 razy, zazwyczaj z różnymi przymiotnikowymi modyfikatorami rzeczownika *chants*, np. *subjected to monkey chants* (18), *subjected to chants* (14), *subjected to racist* (9)/*derisive* (5)/*distasteful* (3) *chants*. Niemal we wszystkich przypadkach jest używana jako określenie obraźliwych (w tym również rasistowskich) przyspiewek kibiców, co dodatkowo świadczy o jej skonwencjonalizowanym formalnie i kontekstowo użyciu.

⁶ Zob. <http://www.sport.pl/pilka/7,64946,25309033,rasistowski-skandal-w-sofii-sedzia-musial-dwukrotnie-przerywac.html>.

Tabela 4 Dopełnienia przymikowe wyrażenia *subjected to* zaświadczone w korpusie MoncoEN

#	Kolokat	Częstość
1	abuse	962
2	attack	487
3	scrutiny	406
4	violence	395
5	test	306
6	harassment	286
7	treatment	281
8	assault	244
9	threat	238
10	torture	206
74	chant	60

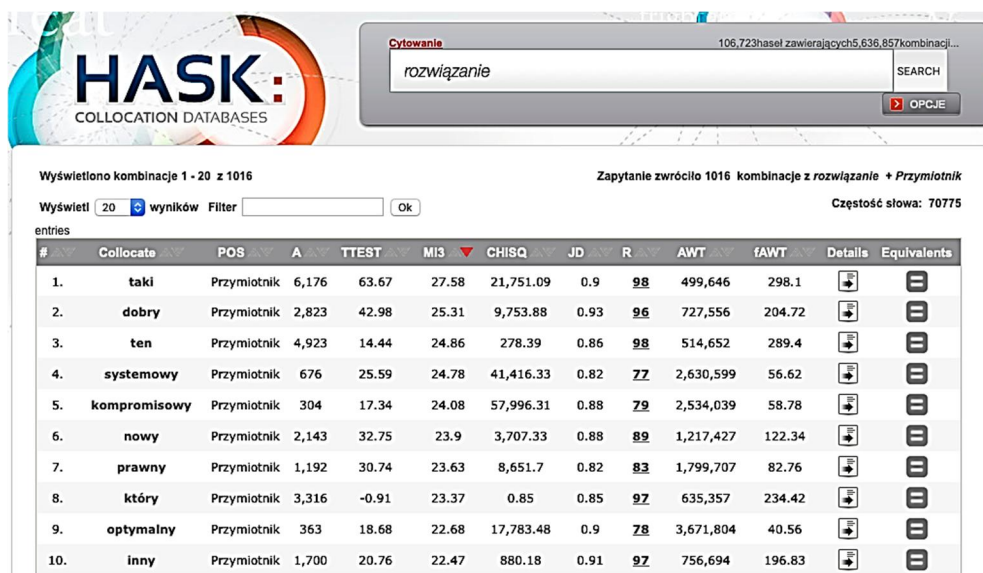
Po ustaleniu statusu konstrukcji występującej w tekście oryginału jako skonwencjonalizowanego frazemu oraz negatywnej weryfikacji utrwalenia zaproponowanego ekwiwalentu, możemy przystąpić do poszukiwań bardziej idiomatycznego odpowiednika w referencyjnym korpusie języka przekładu. O ile w korpusach równoległych, takich jak Paralela, ekwiwalentów można wyszukiwać zapytaniem o wystąpienia konstrukcji lub wyrazów z oryginału, o tyle w przypadku korzystania z korpusu referencyjnego możliwe ekwiwalenty powinna podpowiadać tłumaczowi intuicja. Możliwym ekwiwalentem rzeczownika *chants* w omawianym kontekście może być rzeczownik *przyśpiewka*, którego kolokaty przymiotnikowe, takie jak *kibicowska* (111 wystąpień MoncoPL), *stadionowa* (93), *rasistowska* (83), *antysemicka* (41) i *kibolska* (17) wskazują na jej użycie w opisach niepożądanych zachowań kibiców na stadionach. Tabela 5 ukazuje kilka kolokatów czasownikowych dla zapytania zapytania *rasistowski** przyśpiewka*** jako ośrodka kolokacji; *rasistowskie przyśpiewki* w tekstach prasowych, między innymi *intonowane* i *kierowane* pod jakimś adresem. Zaświadczenia te uzasadniają użycie w tłumaczeniu wielowyrazowego wyrażenia formułicznego, które poprawia jego idiomatyczność, np.: „Pod adresem piłkarzy reprezentacji Anglii *kierowano* (wulgarne/ohydne) *rasistowskie przyśpiewki*”.

Tabela 5 Wybrane czasowniki występujące w kontekście frazy *rasistowskie przyśpiewki* w korpusie MoncoPL

#	Kolokat	Częstość	Przykład
1	intonować	19	intonowali rasistowskie przyśpiewki (5)
2	ukarać	12	ukarany za rasistowskie przyśpiewki (1)
3	zostać	8	został ukarany za rasistowskie przyśpiewki (1)
4	mieć	7	miały miejsce rasistowskie przyśpiewki (2)
5	kierować	5	kierowali pod adresem rosyjskich kibiców obraźliwe przyśpiewki (1)

5. Słowniki kombinatoryczne

Procedurę ekstrakcji kolokacji można wykonać iteracyjnie dla wszystkich pojedynczych wyrazów w korpusie, zapisując listy zidentyfikowanych kolokatów dla każdego z zapytań. Powstaje w ten sposób baza danych, którą można określić mianem automatycznego słownika kombinatorycznego. W słowniku tym zapisane są odnotowane w korpusie połączenia wyrazowe tworzące dwuwyzrazowe kolokacje i dłuższe jednostki frazeologiczne. Przykładem takiego zasobu są bazy kolokacji polskich i angielskich HASK, wygenerowane odpowiednio z korpusów NKJP i BNC (British National Corpus; Pęzik, 2012, 2014)⁷. Ponieważ profile kolokacyjne każdego hasła zostały wcześniej wyliczone i zapisane w bazie, wyszukiwanie kolokacji w słownikach HASK jest bardzo szybkie i nie wymaga konstruowania specjalnych zapytań. Rys. 2 ukazuje tabelę wyników uzyskaną po wpisaniu podstawowej formy rzeczownika *rozwiązanie* w polu wyszukiwania. W tabeli uwzględniono, poza zwykłą częstością występowania kombinacji wyrazowych, kilka miar powiązania (np. informacja wzajemna, chi-kwadrat) oraz równomierności występowania w całym korpusie, którym w tym przypadku jest NKJP. Każdy z tych wyników może być weryfikowany na podstawie pełnych konkordancji, które znajdują się pod odsyłaczem zamieszczonym w przedostatniej kolumnie.

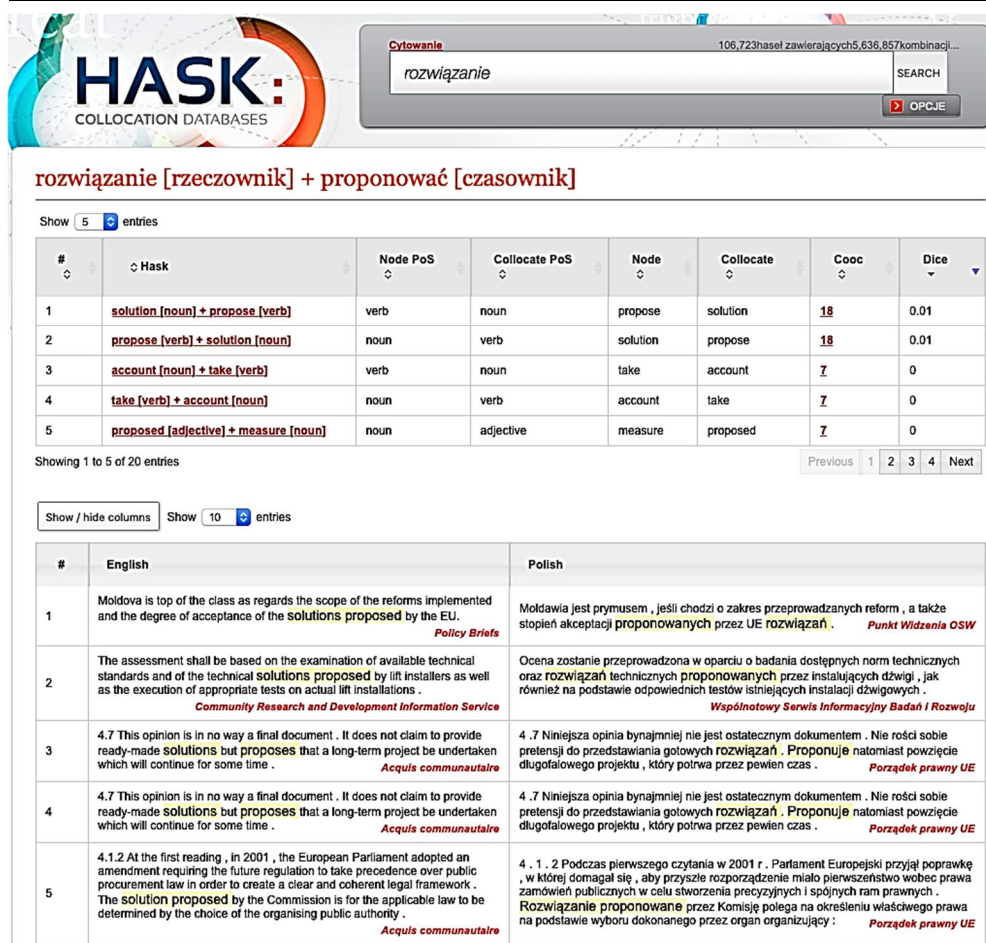


The screenshot shows the HASK Collocation Databases interface. At the top, there is a search bar with the word 'rozwiązanie' entered. Below the search bar, a table displays the results for the word 'rozwiązanie'. The table has columns for rank, collocate, part of speech (POS), and various statistical measures (A, TTEST, MIS, CHISQ, JD, R, AWT, fAWT). The results are sorted by rank, with 'taki' being the top collocate.

#	Collocate	POS	A	TTEST	MIS	CHISQ	JD	R	AWT	fAWT	Details	Equivalents
1.	taki	Przymiotnik	6,176	63.67	27.58	21,751.09	0.9	98	499,646	298.1		
2.	dobry	Przymiotnik	2,823	42.98	25.31	9,753.88	0.93	96	727,556	204.72		
3.	ten	Przymiotnik	4,923	14.44	24.86	278.39	0.86	98	514,652	289.4		
4.	systemowy	Przymiotnik	676	25.59	24.78	41,416.33	0.82	77	2,630,599	56.62		
5.	kompromisowy	Przymiotnik	304	17.34	24.08	57,996.31	0.88	79	2,534,039	58.78		
6.	nowy	Przymiotnik	2,143	32.75	23.9	3,707.33	0.88	89	1,217,427	122.34		
7.	prawny	Przymiotnik	1,192	30.74	23.63	8,651.7	0.82	83	1,799,707	82.76		
8.	który	Przymiotnik	3,316	-0.91	23.37	0.85	0.85	97	635,357	234.42		
9.	optymalny	Przymiotnik	363	18.68	22.68	17,783.48	0.9	78	3,671,804	40.56		
10.	inny	Przymiotnik	1,700	20.76	22.47	880.18	0.91	97	756,694	196.83		

Rysunek 2 Kolokacje dwuwyzrazowe rzeczownika *rozwiązanie* w słowniku HASK

⁷ Zob. pelcra.pl/hask.



HASK: COLLOCATION DATABASES

Cytowanie 106,723hasel zawierających 5,636,857kombinacji...

rozwiązanie SEARCH OPCJE

rozwiązanie [rzeczownik] + proponować [czasownik]

Show 5 entries

#	↕ Hask	Node PoS	Collocate PoS	Node	Collocate	Cooc	Dice
1	solution [noun] + propose [verb]	verb	noun	propose	solution	18	0.01
2	propose [verb] + solution [noun]	noun	verb	solution	propose	18	0.01
3	account [noun] + take [verb]	verb	noun	take	account	7	0
4	take [verb] + account [noun]	noun	verb	account	take	7	0
5	proposed [adjective] + measure [noun]	noun	adjective	measure	proposed	7	0

Showing 1 to 5 of 20 entries

Show / hide columns Show 10 entries

#	English	Polish
1	Moldova is top of the class as regards the scope of the reforms implemented and the degree of acceptance of the solutions proposed by the EU. <i>Policy Briefs</i>	Moldawia jest prymusem, jeśli chodzi o zakres przeprowadzonych reform, a także stopień akceptacji proponowanych przez UE rozwiązań. <i>Punkt Widzenia OSW</i>
2	The assessment shall be based on the examination of available technical standards and of the technical solutions proposed by lift installers as well as the execution of appropriate tests on actual lift installations. <i>Community Research and Development Information Service</i>	Ocena zostanie przeprowadzona w oparciu o badania dostępnych norm technicznych oraz rozwiązań technicznych proponowanych przez instalujących dźwigi, jak również na podstawie odpowiednich testów ścinających instalacji dźwigowych. <i>Wspólnotowy Serwis Informacyjny Badań i Rozwoju</i>
3	4.7 This opinion is in no way a final document. It does not claim to provide ready-made solutions but proposes that a long-term project be undertaken which will continue for some time. <i>Acquis communautaire</i>	4. 7 Niniejsza opinia bynajmniej nie jest ostatecznym dokumentem. Nie rości sobie pretensji do przedstawiania gotowych rozwiązań. Proponuje natomiast powzięcie długofalowego projektu, który potrwa przez pewien czas. <i>Porządek prawny UE</i>
4	4.7 This opinion is in no way a final document. It does not claim to provide ready-made solutions but proposes that a long-term project be undertaken which will continue for some time. <i>Acquis communautaire</i>	4. 7 Niniejsza opinia bynajmniej nie jest ostatecznym dokumentem. Nie rości sobie pretensji do przedstawiania gotowych rozwiązań. Proponuje natomiast powzięcie długofalowego projektu, który potrwa przez pewien czas. <i>Porządek prawny UE</i>
5	4.1.2 At the first reading, in 2001, the European Parliament adopted an amendment requiring the future regulation to take precedence over public procurement law in order to create a clear and coherent legal framework. The solution proposed by the Commission is for the applicable law to be determined by the choice of the organising public authority. <i>Acquis communautaire</i>	4. 1. 2 Podczas pierwszego czytania w 2001 r. Parlament Europejski przyjął poprawkę, w której domagał się, aby przyszłe rozporządzenie miało pierwszeństwo wobec prawa zamówień publicznych w celu stworzenia precyzyjnych i spójnych ram prawnych. Rozwiązanie proponowane przez Komisję polega na określeniu właściwego prawa na podstawie wyboru dokonanego przez organ organizujący. <i>Porządek prawny UE</i>

Rysunek 3 Widok ekwiwalentów kolokacyjnych w słowniku HASK PL

Dla tłumacza szczególnie interesująca może być opcja wyświetlania potencjalnych ekwiwalentów kolokacji wielowyrazowych, dostępna w ostatniej kolumnie każdego wiersza wyników. Pierwszym krokiem algorytmu wyszukiwania ekwiwalentów kolokacyjnych jest zbudowanie słowników kombinacji dwuwyrazowych z korpusu BNC i NKJP. Następnie, wykryte w tych korpusach kolokacje indeksowane są w segmentach korpusu Paralela. Jeżeli w danej parze segmentów równoległych występują dwie kombinacje zindeksowane wcześniej w korpusach referencyjnych, to takie wystąpienie jest doliczane do wartości kolumny współwystąpień (ang. *co-occurrences*). Rys. 3 ukazuje ekran potencjalnych ekwiwalentów znalezionych w ten sposób dla kolokacji *proponować rozwiązanie*. Wynika z niej, że w korpusie Paralela znaleziono 18 wystąpień angielskiej kolokacji *propose a solution* równoległych do polskich segmentów, które zawierały kolokację *proponować rozwiązanie*. Co ciekawe, idiomatyczność tych kombinacji potwierdza niezależne

ich występowanie w korpusach referencyjnych. W ogromnej większości przypadków nie są to więc kalki językowe, o które, jak wykazano w przykładach cytowanych w tym artykule, nietrudno w oportunistycznym korpusie równoległym, jakim jest Paralela. Poniżej w tabeli współwystąpień kolokacji wyświetlane są również konkordancje z korpusu równoległego, które pozwalają zweryfikować trafność proponowanych ekwiwalentów. Słowniki HASK stanowią cenne źródło wiedzy o setkach tysięcy utrwalonych połączeń wyrazowych występujących w referencyjnych korpusach polszczyzny i angielszczyzny oraz ich w różnym stopniu skonwencjonalizowanych ekwiwalentach występujących w korpusie Paralela.

Do ich głównych ograniczeń można zaliczyć: (1) zawężenie procesu wydobywania frazeologii do kombinacji dwuwyrazowych (2) zaledwie dwóch korpusów referencyjnych i jednego równoległego. W kolejnej sekcji artykułu zostaje przedstawiona zależnościowa metoda i oparte na niej narzędzie do ekstrakcji frazeologii, które wolne jest od obu tych ograniczeń.

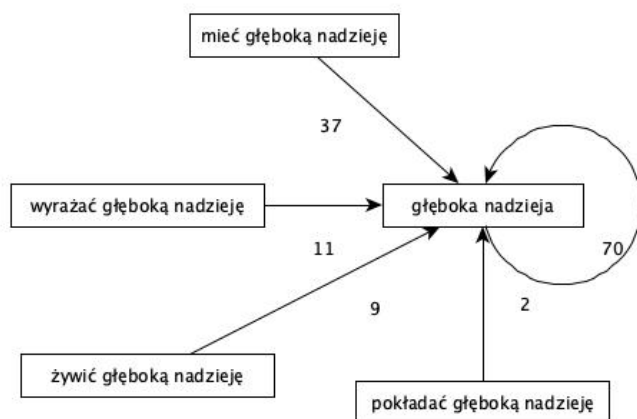
6. Frazeosyntaktyka tekstu przekładu

Prezentowane powyżej narzędzia do identyfikacji związków wyrazowych wykorzystują tzw. pozycyjne metody ekstrakcji frazeologii, które zakładają, że składniki jednostek frazeologicznych występują w liniowo definiowanym kontekście. Na przykład, w ukazanej na Rys. 2 tabeli kolokatów rzeczownika *rozwiązanie* uwzględnione są tylko takie przymiotniki, które w zrównoważonej części NKJP występowały w „odległości” maks. 2 wyrazów (przed lub po wystąpieniu wyrazu *rozwiązanie*). Alternatywą dla takiego definiowania granic frazemu mogą być podejścia „relacyjne” (Evert, 2005), w których przyjmuje się, że jednostki frazeologiczne mają przewidywalną na pewnym poziomie ogólności strukturę składniową. Na przykład w zależnościowej ekstrakcji frazeologii przyjmuje się, że frazemy tworzą tzw. kateny (Pęzik, 2017; Osborne, Putnam, Groß, 2012), czyli podgrafy drzewa składniowego-zależnościowego zdania. Zaletą takiego podejścia jest możliwość intuicyjnego definiowania frazemów złożonych z więcej niż dwóch wyrazów, a także modelowania ich struktury gramatycznej i relacji między stałymi i zmiennymi postaciami konstrukcji frazeologicznych (Pęzik, 2018).

Te nieco teoretyczne rozważania rodzą bardzo praktyczne implikacje. Pewne aspekty poprawności frazeostylistycznej tekstu tłumaczenia mają częściowo składniowy (a nie jedynie leksykalny) wymiar. Rozważmy w tym kontekście przykład kolokacji *głęboka nadzieja*, która została użyta jako ekwiwalent formulicznych wyrażań angielskich *I truly hope that*, *I sincerely hope that* oraz *I really hope that* w cytowanych poniżej tłumaczeniach wystąpień w Parlamencie Europejskim:

13. O: *I truly hope that* we can achieve this with the cooperation of the industry (Paralela, EP Proceedings).
14. T: Mam **głęboką nadzieję**, że możemy to osiągnąć przy współpracy z producentami.
15. O: *I sincerely hope that* Mrs Rajavi is right (Paralela, EP Proceedings).
16. T: Mam **głęboką nadzieję**, że pani Radżawi ma rację.
17. O: *I really hope* this report will help to solve those problems (Paralela, EP Proceedings).
18. T: **Żywię głęboką nadzieję**, że przedmiotowe sprawozdanie przyczyni się do rozwiązania tych problemów.

Poprawne stylistycznie użycie pozornie samodzielnej znaczeniowo i strukturalnie kolokacji *głęboka nadzieja* w cytowanych tłumaczeniach zależy między innymi od uwzględnienia jej dwóch subtelnych właściwości frazeosyntaktycznych. Po pierwsze, fraza *głęboka nadzieja* w ogromnej większości zaświadczonych w korpusach referencyjnych użyć jest dopełnieniem. Nie wynika to z tzw. restrykcji walencyjnych samego rzeczownika *nadzieja* (czyli zakresu jego możliwych ról syntaktycznych), ponieważ nawet w bardzo idiomatycznych frazach może on funkcjonować jako podmiot, np. *nadzieja jest matką głupich*, *nadzieja umiera ostatnia*. Dodatkowo z analizy jej użyc w dużych korpusach referencyjnych wynika, że kolokacja *głęboka nadzieja* jest powtarzalnym dopełnieniem bliższym zaledwie kilku czasowników. Ilustruje to tzw. *graf uwieżenia* (ang. *subsumption graph*) (Pęzik, 2018), ukazany na Rys. 4. Wierzchołki tego grafu reprezentują różnej długości poddrzewa zależnościowe zawierające kolokację *głęboka nadzieja*. Skierowane krawędzie oznaczają zawieranie struktur dłuższych przez krótsze z częstością wyrażaną przez wagę danej krawędzi. W grafie z Rys. 4 uwzględniono tylko takie struktury zawierające kolokację *głęboka nadzieja*, które wystąpiły w korpusie NKJP co najmniej dwukrotnie.



Rysunek 4 Graf uwieżenia kolokacji *głęboka nadzieja*

Suma wag na krawędziach wchodzących do wierzchołka kolokacji *głęboka nadzieja* z innych wierzchołków wynosi 59, co oznacza, że w 84 proc. przypadków swoich wystąpień w NKJP (59 z 79) jest ona zawarta w dłuższej, potencjalnie regularnej konstrukcji jako dopełnienie czasowników *mieć*, *wyrażać*, *żyć* lub *położyć*. Wynika z tego, że 'bezpieczne' z punktu widzenia frazeostylistyki tekstu użycia kolokacji *głęboka nadzieja* są zawarte w dłuższych strukturach zależnościowych *mieć/wyrażać/żyć głęboką nadzieję, że*. W cytowanych powyżej tłumaczeniach wyrażenia te stanowią frazeologiczne ekwiwalenty sekwencji formułicznych I *really/sincerely/frankly hope*. Przykładów kolokacji częściowo uwiecznionych w dłuższych utrwalonych strukturach można w każdym odpowiednio dużym korpusie referencyjnym znaleźć tysiące. Tabela 6 zawiera przykłady prostych fraz nominalnych złożonych z modyfikowanego przymiotnikiem rzeczownika, które w części zrównoważonej NKJP występują przeważnie lub niemal wyłącznie jako fragmenty dłuższych łańcuchów kolokacyjnych (1-5), lub idiomów (6-7).

Tabela 6 Przykłady prostych kolokacji znacząco uwiecznionych w dłuższych łańcuchach kolokacyjnych lub idiomach

#	Kolokacja	łańcuch kolokacyjny/ Idiom	Uwieszenie w proc.
1	ułatwione zadanie	mieć + ułatwione zadanie	201/212=0,95
2	wyrobione zdanie	mieć + wyrobione zdanie	99/115=0,86
3	szalona ochota	mieć + szaloną ochotę	12/13=0,92
4	należyte wnioski	wyciągnąć + należyte wnioski	13/16=0,81
5	zapewniona opieka	mieć + zapewnioną opiekę	149/161=0,93
6	szerokie echo	odbić się + szerokim echem	153/213=0,72
7	pierwsze skrzypce	grać + pierwsze skrzypce	251/284=0,88

7. Zależnościowe słowniki kombinatoryczne

Po tym nieco teoretycznym wstępie możemy powrócić do głównego tematu tego artykułu, którym są korpusowe narzędzia wspierające wyzwania frazeostylistyczne, przed którymi stoi tłumacz jako autor tekstu przekładu. Współczesne językoznawstwo korpusowe dysponuje bowiem metodami ekstrakcji frazeologii, które umożliwiają identyfikację związków wyrazowych wyższego rzędu niż proste kolokacje binarne. Przykładem takiego narzędzia jest aplikacja o nazwie Treelets⁸, opracowana m.in. przez autora tego artykułu w ramach projektu CLARIN-PL. Aplikacja ta pozwala na generowanie automatycznych zależnościowych słowników kombinatorycznych z dowolnego korpusu w dowolnym języku, o ile ma on anotację zależnościową, lub jeżeli istnieje dla niego model parsera zależnościowego w formacie UDPipe⁹. Rys. 5

⁸ Zob. <http://pelcra.pl/new/treelets>.

⁹ Zob. <http://ufal.mff.cuni.cz/udpipe> (Straka i Straková 2017).

ukazuje panel ustawień narzędzia *Treelets*, w którym zaimportowano trzy różnej wielkości korpusy: referencyjny korpus angielszczyzny (BNC), specjalistyczny korpus polszczyzny (Meteo) i korpus angielszczyzny jako drugiego języka (PLEC). W kontekście praktyki tłumaczeniowej można sobie wyobrazić zarówno potrzebę kwerend frazeologicznych w generycznych korpusach referencyjnych (takich jak NKJP), jak też specjalistycznych, które reprezentują dziedzinę i rejestr tekstu oryginału.

Corpora

Current Corpus search configuration: meteo_corpus

Corpus	Words	Sentences	Texts	Size	External Size	Action
meteo_corpus	1954208	146968	11536	1667 MB	1241 MB	Metadata Files Settings
plec_corpus	3104191	176068	4008	2441 MB	1912 MB	Metadata Files Settings
bnc_corpus	111476331	6051775	4054	64 GB	46 GB	Metadata Files Settings

+ Create Corpus

ACDs

Current ACD search configuration: meteo_acd

ACD	Corpus	Headwords	Treelets	Occurrences	Action
bnc_acd	bnc_corpus	27817	1435169	11576859	Populate Settings
plec_acd	plec_corpus	8397	56257	309175	Populate Settings
meteo_acd	meteo_corpus	5780	28223	273280	Populate Settings

+ Create ACD

Rysunek 5 Panel importu korpusów w narzędziu *Treelets*

Rys. 6 ukazuje dziewięć wzorców struktur zależnościowych, które zdefiniowano dla każdego z tych korpusów. Należy zaznaczyć, że użytkownicy aplikacji mogą samodzielnie 'projektować' struktury zależnościowe, w których wyszukiwane mają być potencjalne frazemy występujące w korpusach za pomocą interfejsu graficznego. Na podanej domyślnie liście reguł znajdują się wzorce prostych kolokacji typu *czasownik + dopełnienie* czy też *modyfikator przymiotnikowy + rzeczownik*, ale także wyższego rzędu konstrukcje, w których na przykład modyfikowane rzeczowniki funkcjonują jako dopełnienie, podmiot itp.

Makrostrukturę zależnościowego słownika kombinatorycznego wyznaczają ukazane w górnej części Rys. 6 lematy (formy bazowe wyrazów), które użytkownicy definiują samodzielnie. W podanym przykładzie hasła słownika generowane są tylko dla rzeczowników i przymiotników o minimalnej częstości, która

zależy od wielkości korpusu. Innymi słowy, dla każdego rzeczownika lub przymiotnika występującego w BNC odpowiednio co najmniej 30 i 20 razy generowana jest lista powtarzających się w tym korpusie struktur zależnościowych dla zdefiniowanych typów. Tabela 6 ukazuje podsumowanie występujących w korpusie BNC ośmiu typów składniowych struktur, w których występuje rzeczownik *point*. Łączna liczba typów leksykalnych takich struktur wynosi dla samego rzeczownika *point* 2 637 (36 735 wystąpień w tekstach).

#	POS	Count	Min freq.	Max freq.	Action
1	NOUN	18537	30	187661	 Delete
2	ADJ	9280	20	153625	 Delete

 Add

Extraction Patterns

#	Name	Desc	Generic	Node pos	Active (false/true)	Action
1	amod	Adjectival Modifiers	true	[NOUN]	☐	 Edit
2	dobj	Direct Objects	true	[NOUN]	☐	 Edit
3	amod_dobj	Adjectival Modifiers of Direct Objects	true	[NOUN]	☐	 Edit
4	amod_dobj_prep	Adjectival Modifiers of Direct Objects with Preposition	true	[NOUN]	☐	 Edit
5	prep_pobj	Prepositional Objects of Nouns	true	[NOUN]	☐	 Edit
6	amod_dobj_prep_pobj	Verbs with Direct and Prepositional Objects	true	[NOUN]	☐	 Edit
7	amod	Adv Modifiers of Adjectives	true	[ADJ]	☐	 Edit
8	advmod	Nominal subjects	true	[NOUN]	☐	 Edit
9	amod_nsubj	Nominal subjects with adj modifiers	true	[NOUN]	☐	 Edit

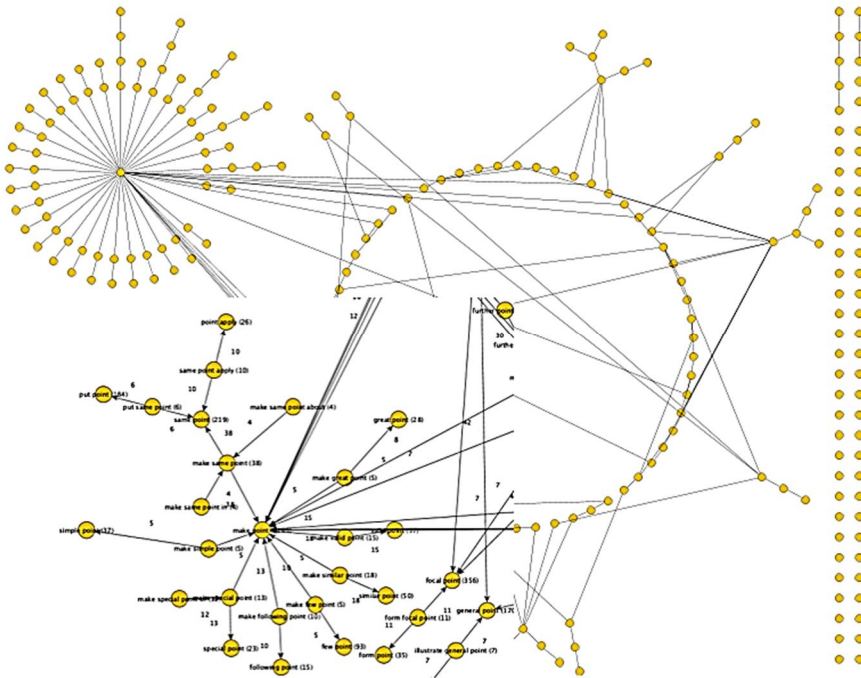
 New pattern

Rysunek 6 Reguły zależnościowej ekstrakcji frazeologii w narzędziu Treelets

Tabela 6 Typy powtarzalnych struktur zależnościowych, w których występuje rzeczownik *point* w słowniku kombinatorycznym dla korpusu BNC

#	Typ składniowy	Częstość	Przykłady
1	Nouns with Prep. Objects	1012	point of view, point of order, point of contact
2	Adjectival Modifiers	569	basic point, very point, various point
3	Direct Objects	400	separate point, bubble point, make point
4	Adjectival Modifiers as Direct Objects	301	make same point, make important point, have good point
5	Nominal Subjects	136	point find, point ensure, point offer
6	Nominal Subjects with Adj. Modifiers	125	strong point be, practical point be, striking point be
7	Adjectival Modifiers of Direct Objects with Preposition	67	provide focal point for, make special point of, have different point of
8	Direct Objects with Prep. Objects	27	put own point of view, have different point of view, maintain coherent point of view

Złożoność strukturalną różnych kolokacji rzeczownika *point* ilustruje graf uwięzienia przedstawiony poglądowo na Rys. 7. Zdolność do rozpoznania lub właściwego użycia wielu z tych struktur podczas procesu tłumaczenia jest ważnym aspektem kompetencji językowych tłumacza przekładającego z angielszczyzny, lub też tłumaczącego na ten język. Aplikacja Treelets może tłumaczyć w tym procesie wspierać pod warunkiem, że korpus, z którego wyekstrahowano zależnościowy słownik kombinatoryczny jest odpowiednio duży i reprezentatywny.



Rysunek 7 Widok grafu uwięzienia kolokacji rzeczownika *point* w aplikacji Treelets

Tłumacz poszukujący skonwencjonalizowanego odpowiednika danej frazy może skorzystać z intuicyjnych w użyciu funkcji przeszukiwania słownika, które obejmują zarówno jego makro-, jak też mikrostrukturę. Po wpisaniu zapytania o słowo lub frazę wyświetlane są zarówno podsumowania haseł dla pojedynczych leksemów podobne do tego z Tabeli 6, jak też odsyłacze do stron poszczególnych kombinacji wyrazowych. Taka przeszukiwalna baza połączeń wyrazowych może być przydatna również w tłumaczeniach specjalistycznych. Na przykład, słownik wygenerowany z korpusu komentarzy synoptyków ukazujących się w portalu *meteo.pl* może stanowić cenne źródło informacji o frazeologii i terminologii dla tłumacza przekładającego teksty z tej dziedziny. Dla zapytania o hasła i połączenia zawierające wyraz *zmiana* tego liczącego ok. dwóch milionów segmentów

zbioru tekstów użytkownik narzędzia Treelets dostaje m.in. kolokację dwuwyrazową *zachodząca zmiana*. Po kliknięciu na ten wynik, możliwe jest sprawdzenie konkordancji kolokacji, statystyk jego występowania, struktury składniowej, ale także jej stopnia uwięzienia w dłuższych powtarzalnych strukturach. Służy do tego specjalna zakładka z podsumowaniem łączliwości wewnętrznej i zewnętrznej danej kombinacji. W ten sposób od kolokacji *zachodząca zmiana* łatwo można dotrzeć do dłuższego łańcucha kolokacyjnego *zachodzące zmiany doprowadziły/doprowadzą do*, którego konkordancje ukazano na Rys. 8. Konstrukcje te są z jednej strony silnie formułiczne w badanym rejestrze języka, ale jednocześnie nie stanowią jednostek terminologicznych, których wydobyć może umożliwić użycie standardowego oprogramowania wspierającego pracę tłumacza. Warto tu zwrócić uwagę na odległość pozycyjną między składnikami omawianego wyrażenia. O ile sekwencja z przykładu 3 jest zwykłym n-gramem trójwyrazowym, to w przykładzie czwartym czasownik jest oddalony od podmiotu o ponad 20 segmentów wyrazowych. Użycie relacyjnych metod ekstrakcji frazeologii zwiększa pokrycie ekstrakcji rozproszonych formuł wyrazowych, które składają się na odpowiedni poziom idiomatyczności tłumaczenia specjalistycznego.

[zachodzący + zmiana + doprowadzić]

Concordances		Stats	Syntax	Valency
Found 4 results				
#	Sentences			
1	Zmiany zachodzące w rozkładzie ciśnienia związane z oddalaniem się niżu na wschód oraz stałym jego wypełnianiem się , doprowadziły do wyraźnego zmniejszenia się gradientu barycznego nad Polską , a tym samym osłabnięcia wiatru oraz wzrostu temperatury odczuwalnej .			
2	Zachodzące powoli zmiany w rozkładzie ciśnienia nad kontynentem europejskim , zapoczątkowane przemieszczaniem się dobrze rozbudowanego wyżu znad Morza Północnego w kierunku południowo wschodnim , już w horyzoncie czasowym naszych prognoz , doprowadzą do napływu powietrza polarnego o cechach morskich , pochodzącego z umiarkowanych szerokości geograficznych znajdujących się nad Atlantykem .			
3	Zachodzące zmiany doprowadzą do występowania typowych warunków pogodowych dla majowej wiosny .			
4	Zmiany zachodzące w rozkładzie ciśnienia przy powierzchni ziemi oraz geopotencjału w troposferze powyżej warstwy granicznej , o których szerzej pisałem w poprzednim komentarzu , doprowadziły do odbudowania się strefowej cyrkulacji atmosfery , zapoczątkowując nowy naturalny okres synoptyczny z powietrzem polarnym napływającym krótką drogą znad Atlantyku .			

Rysunek 8 Widok konkordancji pojedynczej konstrukcji w aplikacji Treelets

8. Wykrywanie frazeologii

Jeszcze jednym potencjalnym zastosowaniem systemu zależnościowej ekstrakcji frazeologii, o którym warto tu wspomnieć, jest tzw. detekcja, czyli wykrywanie frazeologii (z ang. *phraseology detection*, por. Pęzik, 2018). W odróżnieniu od opisanej powyżej ekstrakcji, która jest procesem agregacji jednostek wielowyrazowych w oparciu o odpowiednio duże korpusy, detekcja jednostek frazeologicznych

polega na wykrywaniu nawet pojedynczych ich wystąpień w tekstach na podstawie opracowanego wcześniej słownika referencyjnego. Słownikiem referencyjnym może być elektroniczna postać opracowanego tradycyjnymi metodami słownika kolokacji lub idiomów, a także automatycznie wyekstrahowana z korpusu baza kombinatoryczna. Przykładem narzędzia, które wykrywa wybrane typy syntaktyczne frazemów w tekstach angielskich jest aplikacja Phrime (Pęzik, 2017)¹⁰. Wczytując krótki fragment tekstu z przykładu 19 poniżej do modułu detekcji frazeologii tego narzędzia, otrzymujemy listę 4 kolokacji i łańcuchów kolokacyjnych wykrytych w automatycznym słowniku kombinatorycznym wygenerowanym z dużego korpusu angielszczyzny amerykańskiej.

19. O: Parliament has *condemned the use of force* and believes that conflicts in the Caucasus cannot be solved by violence; the speedy clear-up of mines will *prevent future casualties* of civilians (Paralela, EP Proceedings).
20. Parlament *potępił użycie siły* i jest przekonany, że konflikty na Kaukazie nie mogą być rozwiązywane z wykorzystaniem przemocy; szybkie odminowanie terenu *zapobieganie ofiarom* w ludziach w przyszłości.

Ukazany na Rys. 9 ekran wyników sugeruje, że w zasadzie całe wyrażenie *condemn the use of force* można uznać za stosunkowo utrwalone w korpusie referencyjnym, ponieważ poszczególne jego składniki występują w nim od kilkunastu do kilkuset razy. Również prosta kolokacja *prevent + casualties* została zaświadczona w korpusie siedmiokrotnie. Taka uzyskana za pomocą kilku kliknięć lista może dodatkowo uczulać tłumacza na obecność w tekście tłumaczenia prefabrykatów frazeologicznych, które wymagają być może równie utrwalonych i łańcuchów przyswajalnych ekwiwalentów w tekście przekładu. Użyty w polskim tłumaczeniu (20) zwrot *potępić użycie siły* jest również stosunkowo zaświadczone w korpusach referencyjnych polszczyzny łańcuchem kolokacyjnym (ponad 200 wystąpień w korpusie MoncoPL). Z kolei kombinacja czasownika *zapobiegać* z dopełnieniem celownikowym w postaci rzeczownika *ofiarom* jest dużo rzadsza, a na pewno względnie rzadsza od oryginalnego zwrotu *prevent casualties*. Oczywiście fakt ten nie wyklucza całkowicie jej użycia jako ekwiwalentu w tym kontekście, niemniej tego typu informacja zwrotna o każdej konstrukcji opisanej regułami ekstrakcji słownika kombinatorycznego może być pomocna w weryfikacji poziomu idiomatyczności tekstu oryginalnego, jak też przekładu.

¹⁰ Zob. phrime.tt.com.pl.

Parliament has **condemned** the **use of force** and believes that conflicts in the Caucasus cannot be solved by violence ; the speedy clear-up of mines will **prevent** future **casualties** of civilians .

Filter:

#		Chain	Freq.	
1	☐	condemn + use	18	    
2	☒	condemn + use + of	18	    
3	☒	use + of + force	787	    
4	☒	prevent + casualty	7	    

Rysunek 9 Kolokacje i łańcuchy kolokacyjne wykryte w jednozdaniowym fragmencie tekstu przez aplikację Phrime

Funkcjonalność wykrywania potencjalnych frazemów w oparciu o wygenerowane z korpusów referencyjnych i specjalistycznych dane jest również zaplanowana w pierwszej wersji aplikacji Treelets, której oficjalna premiera planowana jest na początek roku 2020.

8. Podsumowanie

Artykuł ten miał na celu wykazanie przydatności niektórych mniej lub bardziej znanych (NKJP, MoncoPL, Paralela), lub zupełnie nowych (Treelets) zasobów i narzędzi korpusowych, które wspierają tłumacza w procesie tworzenia i weryfikacji frazeostylityki przekładu. Przydatność wyszukiwarek korpusowych w tym zakresie zależy przede wszystkim od wielkości, reprezentatywności i aktualności zindeksowanych w nich korpusów. Wydaje się, że rozmiar korpusu liczony w miliardach słów częściowo rekompensuje braki w jego zrównoważeniu, a na pewno pomaga w poświadczaniu utrwalenia potencjalnych frazemów, które mogą służyć jako idiomatyczne ekwiwalenty tłumaczeniowe. Nieco bardziej zaawansowane narzędzia do ekstrakcji frazeologii pozwalają tworzyć słowniki kombinatoryczne z własnoręcznie kompilowanych korpusów specjalistycznych i referencyjnych. Zależnościowe metody ekstrakcji frazeologii umożliwiają weryfikację subtelnych restrykcji leksykalno-gramatycznych, od których również zależy ogólna jakość przekładu. Choć omawiane tu były tylko przykłady polskich i angielskich korpusów i baz kombinatorycznych, warto jeszcze raz podkreślić, że aplikacja Treelets może być używana potencjalnie dla kilkudziesięciu innych języków, dla których istnieją otwarte analizatory składniowe. Nadzieją autora tej aplikacji jest, że wejdzie ona do warsztatu tłumaczy niektórych z tych języków jako narzędzie wspierające ich kompetencje frazeostylizacyjne.

Bibliografia

- Chlebda, W. (2003). *Elementy frazematyki: Wprowadzenie do frazeologii nadawcy*. Łask: Leksem.
- Cowie, A., Mackin R., McCaig, I. (1975). *Oxford dictionary of current Idiomatic English*. London: Oxford University Press.
- Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations*. Unpublished PhD thesis, University of Stuttgart.
- Lewicki, A. (2003). *Studia z teorii frazeologii*. Łask: Leksem.
- Osborne, T., Putnam, M., Groß, T. (2012). *Catenae: Introducing a novel unit of syntactic analysis*. *Syntax*, 154(4), 354-96. <http://doi.org/10.1111/j.1467-9612.2012.00172.x>
- Pawley, A., Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. W: J. J. C. Richards, R. R. W. Schmidt (red.), *Language and communication* (s. 191-225). Longman: Londyn.
- Pęzik, P. (2012a). Wyszukiwarka PELCRA dla danych NKJP. W: A. Przepiórkowski, M. Bańko, R. L., Górski, B. Lewandowska-Tomaszczyk (red.), *Narodowy korpus języka polskiego* (s. 253-279). Warszawa: PWN.
- Pęzik, P. (2012b). NKJP w warsztacie tłumacza. *Narodowy korpus języka polskiego*. W: A. Przepiórkowski, M. Bańko, R. L., Górski, B. Lewandowska-Tomaszczyk (red.), *Narodowy korpus języka polskiego* (s. 301-311). Warszawa: PWN.
- Pęzik, P. (2014). Graph-based analysis of collocational profiles. W: V. Jesenšek, P. Grzybek (red.), *Phraseologie im Wörterbuch und Korpus/Phraseology in Dictionaries and Corpora*, ZORA 97 (s. 227-243). Maribor, Bielsko-Biała, Budapest, Kansas, Praha: Filozofská fakulteta.
- Pęzik, P. (2017). Experimental applications of dependency-based phraseology extraction. W: P. Pęzik, J. Waliński, J. (red.), *Language, corpora and cognition* (s. 29-55). Frankfurt am Main: Peter Lang.
- Pęzik, P. (2018). *Facets of prefabrication: Perspectives on modelling and detecting phraseological units*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.
- Pęzik, P. (w druku). Budowa i zastosowania korpusu monitorującego MoncoPL. *Forum Lingwistyczne*.
- Straka, M., Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. W: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (s. 88-99). Vancouver, Canada: Association for Computational Linguistics.