*PROBLEMY FRAZEOLOGII EUROPEJSKIEJ X*
*Oficyna Wydawnicza LEKSEM, Łask 2015*

ŁUKASZ GRABOWSKI
Uniwersytet Opolski

# O frazeologii z perspektywy językoznawstwa korpusowego Przegląd głównych nurtów badawczych ostatniego dwudziestolecia w Wielkiej Brytanii i USA

ON PHRASEOLOGY FROM A CORPUS LINGUISTIC PERSPECTIVE: AN OVERVIEW OF RESEARCH IN THE USA AND THE UK IN THE LAST TWO DECADES

This paper provides an overview of phraseological research grounded in corpus linguistics and conducted in the last two decades or so in the United States and in the United Kingdom. A number of approaches to phraseological research are presented, starting from idiom principle, pattern grammar, lexical bundles, lexical priming to skipgrams, phrase frames, concgrams to semantic sequences. Also, an attempt is made by the author to outline a definition of corpus linguistic phraseology (*frazeologia korpusowa*), a label proposed in this paper. The rationale behind this paper is an observation that, on the one hand, corpus linguistic research on phraseology in the Polish language is still scarce and, on the other, that not all Polish corpus or computational linguists are familiar with more traditional methods of phraseological research. This paper is an attempt to change this state of affairs.

## Wstęp

Celem niniejszego artykułu jest naszkicowanie zasięgu, ram i podstawowych założeń korpusowych badań nad frazeologią. Jest to w istocie także zbiór pewnych tez i rozważań nad zastosowaniem metodologii językoznawstwa korpusowego we

frazeologii przez pryzmat najważniejszych nurtów badawczych, które w ostatnich dwóch dekadach rozwinęły się w Wielkiej Brytanii i USA.[1] Przede wszystkim postaram się wyjaśnić najważniejsze – moim zdaniem – różnice między tradycyjnym paradygmatem badań nad frazeologią, ukształtowanym pod wpływem koncepcji wysuniętych przez takich badaczy, jak Wiktor Winogradow (1947/1977) czy Natalia Amosowa (1963), a podejściem korpusowym, określanym najczęściej mianem frazeologii dystrybucyjnej[2] (*distributional phraseology*), frazeologii sterowanej frekwencją (*frequency-driven phraseology*), czy też frazeologii sterowanej danymi (*data-driven phraseology*), by użyć określeń proponowanych przez m.in. Sylviane Granger i Magali Paquot (2008), czy Stefana Griesa (2008). Omówienie rozpocznę od przybliżenia klasyfikacji frazeologizmów, zaproponowanej przez Anthony'ego Cowie (1994, 1998), osadzonej bardziej w tradycji winogradowowskiej. Następnie omówię badania Rosamund Moon (1998), która przedstawiła nie tylko typologię frazeologizmów, ale także zbadała częstość ich występowania w różnych gatunkach tekstów napisanych w języku angielskim. Badanie Moon można uznać za studium na pograniczu dwóch paradygmatów. Dalej zwięźle przybliżę specyfikę językoznawstwa korpusowego, zwłaszcza w odniesieniu do metodologii prowadzenia badań nad językiem. W najważniejszej części artykułu przybliżę różne koncepcje i podejścia do badań nad frazeologią osadzone w metodologii stricte korpusowej[3], takie jak zasada idiomu (Sinclair 1987,1991), Pattern Grammar (Hunston i Francis 2000), torowanie / pryming leksykalny (Hoey 2005), zbitki wielowyrazowe (Biber i in. 1999), skipgramy (Wilks 2005), ramy frazowe (Fletcher 2007 i concgramy (Cheng i in. 2006), a także sekwencje semantyczne (Hunston 2008). W niniejszym artykule będę używał jednego terminu – frazeologia korpusowa – jako nadrzędnego określenia dla wszystkich omówionych podejść. W podsumo-

1 Niniejszy tekst został przedstawiony na posiedzeniu Sekcji Frazeologicznej Komitetu Językoznawstwa Polskiej Akademii Nauk w Warszawie w dniu 25.10.2013 r. Niektóre jego fragmenty zostały opublikowane w języku angielskim w monografii pt. *Phraseology in English Pharmaceutical Discourse: A Corpus-Driven Study of Register Variation* (Grabowski 2015b).

2 Na gruncie polskim termin „frazeologia dystrybucyjna" stosują m.in. Piotr Pęzik (2013) i Stanisław Goźdź-Roszkowski (2013).

3 Inspiracją dla niniejszego artykułu było badanie dystrybucji wybranych wzorców leksykalno-gramatycznych w korpusie tekstów akademickich napisanych w języku angielskim przez studentów o różnym stopniu kompetencji językowej, przeprowadzone przez Roemer (2009b). We wstępie teoretycznym badaczka zwięźle opisuje m.in. zasadę idiomu, Pattern Grammar, pryming leksykalny i zbitki wielowyrazowe jako jedne z najważniejszych korpusowych podejść do badań nad frazeologią, u których podstawy leży nierozerwalność leksyki i gramatyki, a także formy i znaczenia jednostek wielowyrazowych. W niniejszym artykule szerzej rozwijam niektóre koncepcje zasygnalizowane przez Roemer (2009b), a także omawiam koncepcje skipgramów, ram frazowych, congramów i sekwencji semantycznych, które również wpisują się w korpusowe badania nad frazeologią.

waniu podejmę próbę naszkicowania ram definicyjnych frazeologii korpusowej, a także wyciągnę wnioski z przedstawionych tez i rozważań.

## 2. Frazeologia w epoce przedkorpusowej

Tradycyjne podejście do badań nad frazeologią, zapoczątkowane przez takich badaczy rosyjskich, jak m.in. Winogradow, Amosowa, Szacki, czy też Archangielski, osadzone w strukturalizmie, polegało głównie na szczegółowych analizach jednostek podejrzanych o bycie frazeologizmami (najczęściej w ich postaci kanonicznej) w oparciu o szereg kryteriów formalnych, semantycznych czy też pragmatycznych (najczęściej niedokładnie sprecyzowanych i dość ogólnych), bez uwzględnienia kontekstów sytuacyjnych, w jakich owe jednostki bywają używane. Wynikiem takiego podejścia było opracowanie – głównie dla celów leksykograficznych – licznych typologii jednostek wielowyrazowych, dla których najważniejszymi kryteriami był stopień inwariantowości frazeologizmów oraz ich spójności znaczeniowej. Wszak najważniejsze w tym podejściu było opisanie frazeologizmów takimi, jakimi miałyby być przedstawione w słowniku, odzwierciedlającym zawartość leksykonu, czyli części systemu języka. Tym samym zakres frazeologii był wypadkową kryteriów stosowanych dla odróżnienia frazeologizmów od jednostek, których za takowe nie uznawano. Skutkiem takiego podejścia było zawężenie frazeologizmów do jednostek wielowyrazowych spełniających kryteria lingwistyczne określane dość arbitralnie przez badaczy. Za prototypowe frazeologizmy uważano grupy wyrazów o niezmiennej formie i nieumotywowanym znaczeniu, np. u Winogradowa (1947/1977) są to zrosty frazeologiczne (*фразеологические сращения*), u Skorupki (1962/1989) *związki stałe*, u Mielczuka (1996, 1998) pełne frazemy (*полные фраземы*). Tym samym luźne połączenia wyrazowe były albo traktowane marginalnie (tak jak u Skorupki), albo w ogóle znajdowały się poza sferą zainteresowania frazeologii (tak jak u Winogradowa). Zdaniem Wojciecha Chlebdy (2003: 20), jako opozycję do takiego analitycznego[4] podejścia można traktować bardziej syntetyczne rozumienie frazeologii, ukształtowane na gruncie koncepcji innego rosyjskiego językoznawcy teoretycznego, dialektologa i socjolingwisty – Jewgienija Poliwanowa. Jako jednostkę frazeologiczną traktował on dowolną formę językową – czy to jednowyrazową, czy wielowyrazową – systematycznie

4 Zdaniem Wojciecha Chlebdy (2003: 23), w kontekście frazeologii tradycyjnej (winogradowowskiej) przymiotnik *analityczny* oznacza tyle, co ‘rozdrabniający, dzielący, rozdzielający’ (np. idiomy od przysłów, od porzekadeł, od skrzydlatych słów itd.) w opozycji do podejścia *syntetycznego*, w którym różnice gatunkowe (tj. rozróżnianie różnych typów, czy też gatunków frazeologizmów) stają się drugoplanowe. W podejściu syntetycznym na pierwszy plan wysuwa się poszukiwanie odtwarzalnych frazemów lub reproduktów, natomiast sprawą mniejszej wagi jest ich przynależność gatunkowa.

odtwarzaną przez użytkowników języka w określonych sytuacjach. W takim podejściu ważniejsze staje się wyodrębnienie i opis frazeologizmów w powiązaniu z kontekstem ich użycia aniżeli ich szczegółowa klasyfikacja[5]. Ślady takiego syntetycznego podejścia – bardziej od tekstu do słownika, niż od słownika do tekstu – widoczne są wyraźnie w pracach m.in. Andrzeja Marii Lewickiego (1976), Andrzeja Bogusławskiego (1976, 1978, 1989), Wojciecha Chlebdy (2003, 2009, 2010), a także w wielu koncepcjach ukształtowanych na gruncie językoznawstwa korpusowego, o czym będzie jeszcze dalej mowa. To rzekłszy, należy bliżej przyjrzeć się wpływom obu przeciwstawnych nurtów badawczych na badania nad frazeologią w krajach anglojęzycznych.

Zdaniem Granger i Paquot (2008: 36), typologii jednostek wielowyrazowych opracowanych na gruncie angielskiej leksykologii i leksykografii należy szukać przede wszystkim w pracach Anthony'ego Cowie (1981, 1994, 1998), który dzieli frazeologizmy – w oparciu o ich spójność semantyczną i niezmienność formy – na dwa główne typy, tj. kompozyty (*composites*) oraz formuły (*formulas*). Pierwsze z nich, które pod względem składni obejmują jednostki mniejsze niż zdania, Cowie (1998: 17) dzieli na kolokacje ograniczone (*restricted collocations*, np. *to curry favour*), idiomy metaforyczne (*figurative idioms*, np. *to do a U-turn*), oraz idiomy klasyczne (*pure idioms*, np. *to spill the beans*). Na drugim biegunie znajdują się formuły (Cowie 1998: 4), czyli jednostki samodzielne, które z pragmatycznego punktu widzenia obejmują przysłowia, znane cytaty, skrzydlate słowa, a także różne formuły konwersacyjne. Cowie (1998: 4) dzieli je na formuły rutynowe (*routine formulae*), utożsamiane z pojedynczymi aktami mowy, np. *see you soon*, oraz formuły konwersacyjne (*speech formula*), nadające pewne ramy przekazywanym komunikatom oraz wyrażające postawy, oceny, czy też sądy użytkowników języka, np. *You don't say!* Tym samym za podstawę klasyfikacji jednostek wielowyrazowych Cowie przyjmuje kryteria formalne (składniowe), semantyczne i pragmatyczne.

Inna ważna typologia powstała na gruncie leksykografii brytyjskiej to klasyfikacja stałych połączeń wyrazowych i idiomów (*fixed expressions and idioms*), zaproponowana przez Rosamund Moon (1998: 19-25). Uwzględniając szereg kryteriów leksykalno-gramatycznych, semantycznych i pragmatycznych, Moon podzieliła wielowyrazowce na trzy szerokie kategorie, mianowicie kolokacje nietypowe (*anomalous collocations*), które stanowią problem z perspektywy leksyko-gramatyki, formuły (*formulas*), problematyczne z punktu widzenia pragmatyki i funkcji dyskursywnych, oraz metafory (*metaphors*), problematyczne ze względu

5 Jak podaje Kopotiew (2008) do jednej z najbardziej szczegółowych klasyfikacji należy zaliczyć opracowanie Mielczuka (1996), który wyodrębnił aż 54 klasy jednostek wielowyrazowych.

na stopień spójności semantycznej.[6] Zdaniem Moon (1998:19), istota tej typologii polegała na określeniu przyczyn, dla których dane wyrażenia i idiomy mogą być uważane za całostki (*holistic units*) z punktu widzenia leksykografii. Wydaje się jednak, że równie ważnym osiągnięciem Moon było wykazanie, że frazeologizmy tradycyjnie uznawane za prototypowe (czyli idiomy, powiedzenia, przysłowia) są niezwykle rzadko używane w tekstach, a ich znormalizowana frekwencja często nie przekracza progu jednego wystąpienia na milion wyrazów tekstowych (Moon 1998: 59-64). Oznacza to tyle, że ze statystycznego punktu widzenia ich występowanie w tekstach jest wynikiem przypadku. Warto nadmienić, że Moon (1998) przebadała w sumie 6 700 stałych połączeń wyrazowych i idiomów, zaczerpniętych ze słownika *Collins Cobuild English Language Dictionary* (1987), z których 45.3% (3 068) to kolokacje nietypowe, 21.3% (1 143) to formuły, a 33.4% (2 265) to metafory. Co ważne, prawie 70% z nich występuje w 18-milionowym korpusie Oxford Hector Pilot Corpus rzadziej niż raz na milion wyrazów tekstowych, a prawie 40% z nich nie występuje w korpusie w ogóle. Wyniki badania opartego na danych korpusowych pokazały zatem, że to połączenia wyrazowe przejrzyste znaczeniowo, zwłaszcza kolokacje wadliwe i frazeologiczne, formuły proste, metafory przejrzyste i półprzejrzyste są tymi jednostkami, po które użytkownicy języka angielskiego sięgają najczęściej. Tym samym to właśnie te jednostki powinny znajdować się w centrum badań nad frazeologią (a nie przysłowia, porównania, idiomy itp.), niezależnie od tego czy opis frazeologii ma posłużyć jako materiał słownikowy, czy też jako materiał dydaktyczny. Z tego punktu widzenia można jednoznacznie stwierdzić, że syntetyczna koncepcja odtwarzalnych frazemów (Chlebda 2003) oraz reproduktów (2009, 2010), spotykanych w codziennych sytuacjach użycia języka, wychodzi naprzeciw wnioskom płynącym z badania Moon (1998).

Wyniki i metodologia badania przeprowadzonego przez Moon (1998), które można uznać za studium z pogranicza frazeologii tradycyjnej i językoznawstwa korpusowego, pokazały również, że informacja o częstości występowania frazeologizmów nie tylko stanowi wartość dodaną w ich opisie, ale może stanowić istotne kryterium ich wyodrębniania, zwłaszcza kosztem tradycyjnie stosowanych, często nieprecyzyjnych, kryteriów formalnych, semantycznych czy też pragmatycznych.[7] I właśnie problem wyodrębniania jednostek wielowyrazowych (*multi-word units*) z tekstów urósł do rangi jednego z najważniejszych w korpusowym paradygmacie badań nad frazeologią, a zwłaszcza w podejściu sterowanym korpusem (*corpus-driven approach*). Zanim jednak omówię wybrane koncepcje teoretycz-

6 Warto zaznaczyć, że w swojej pracy Moon (1998: 19-25) dokonała jeszcze bardziej szczegółowych klasyfikacji frazeologizmów.

7 Opisywane problemy nabierają szczególnego znaczenia w konfrontatywnych badaniach nad frazeologią, gdzie badacze muszą zmierzyć się z nieprzystawalnością systemów frazeologicznych różnych języków.

ne wypracowane na gruncie badań korpusowych, a wspomniane już na początku niniejszego artykułu, chciałbym pokrótce wyjaśnić, jak językoznawcy korpusowi postrzegają język. Moim zdaniem, taka syntetyczna informacja ułatwi zrozumienie omawianych koncepcji.

## 3. Językoznawstwo korpusowe a badania nad frazeologią

Ogólnie rzecz ujmując, językoznawcy korpusowi badają korpusy, czyli zbiory tekstów utrwalone w postaci elektronicznej, zgromadzone według kryteriów określonych przez badacza (takich jak rozmiar, reprezentatywność, wielkość próbki itp.), których parametry są zależne od m.in. celów badania, zastosowanej metodologii, sformułowanych hipotez, wykorzystanych narzędzi itp. (McEnery i Wilson 1996; Lewandowska-Tomaszczyk 2005). Teksty zaś to duże próby pewnych policzalnych obiektów, wyrazów tekstowych. Innymi słowy, tekst to pewna struktura probabilistyczna, w której istotna jest częstość występowania jednostek i ich klas, wobec tego powinno się go badać przy użyciu takich technik, jakie stosuje np. demografia czy socjologia empiryczna: statystycznych (Piotrowski & Grabowski 2013: 60). Na pierwszy plan wysuwają się zatem zjawiska językowe które są częste i typowe, a nie rzadkie i unikatowe.

W językoznawstwie korpusowym – jak pisze Górski (2012: 292) – wyróżnia się trzy główne podejścia do badania tekstów[8]: badania ilustrowane przez korpus (*corpus-illustrated / corpus-informed approach*), oparte na korpusie (*corpus-based approach*, rzadziej nazywane *corpus-supported approach*) oraz badania sterowane korpusem (*corpus-driven approach*). W przypadku pierwszego traktujemy korpus jak informatora, rodzimego użytkownika języka, ponieważ w argumentacji posługujemy się danymi korpusowymi (Górski 2012: 292).[9] Jak pisałem wcześniej, korpusy to ograniczone zbiory tekstów. W badaniach nad frazeologią takie podejście najczęściej stosujemy do analizy kontekstów występowania zadanych przez nas jednostek wielowyrazowych czy też sprawdzenia częstości ich występowania w tekstach. W podejściu opartym na korpusie (*corpus-based approach*)

8 Lee (2008: 88) wyróżnia jeszcze podejście wywoływane/indukowane korpusem (*corpus-induced approach*), czyli wykorzystywanie korpusów w projektach prowadzonych na dużą skale, głównie polegających na automatycznej analizie danych językowych w celu pozyskania informacji dla zastosowań praktycznych, komercyjnych.

9 Zdaniem Górskiego (2012: 292) należy pamiętać jednak, że korpus nigdy do końca nie zastąpi informatora, brak informacji o występowaniu jakiejś konstrukcji językowej w korpusie nie upoważnia nas do wyciągnięcia wniosku, że takowa konstrukcja jest nieakceptowana – wszak może ona wystąpić w innym korpusie.

przeszukujemy korpus w celu potwierdzenia lub sfalsyfikowania danej hipotezy[10], osadzonej w pewnych ramach teoretycznych (Górski, 2012: 292).[11] Ostatnie podejście, sterowane korpusem (*corpus-driven*), jest najbardziej radykalne – nie stawia się tu bowiem jakichkolwiek hipotez przed rozpoczęciem badań (najczęściej korzysta się wtedy z korpusów nieadnotowanych, tj. surowych danych tekstowych). Nie jest również ważne rozróżnienie poziomów organizacji systemu języka (np. leksyka, składnia, semantyka, pragmatyka), ponieważ w praktyce wszystkie te podsystemy zazębiają się ze sobą w sytuacji przekazywania jakiegoś konkretnego znaczenia, komunikatu (McEnery i Wilson 1996: 10). Tym samym opis języka polega głównie na wyodrębnianiu z tekstów form językowych (jedno- i wielowyrazowych) oraz ich kombinacji, które – jako surowe dane – stanowią punkt wyjścia do formułowania hipotez i dalszych badań (np. polegających na określeniu funkcji dyskursywnych ciągów wyrazowych w tekstach). Takie podejście jest bardzo popularne w badaniach nad frazeologią w krajach anglojęzycznych, a najczęściej wykorzystywane jest właśnie w celu wyodrębniania – przy pomocy specjalnego oprogramowania komputerowego – jednostek wielowyrazowych (np. zbitek wielowyrazowych, ram frazowych, concgramów) w oparciu o arbitralnie określone parametry, takie jak np. długość i ciągłość form językowych[12], ich minimalna znormalizowana frekwencja w tekstach (np. 40 wystąpień na milion wyrazów), zakres dystrybucji (rozproszenia) formy językowej (np. czy dana forma występuje w pojedynczym tekście wchodzącym w skład korpusu, w różnych tekstach reprezentujących ten sam rejestr, czy w różnych typach i gatunkach tekstów).[13] Wynikiem

10 W kontekście frazeologii takim podejściem posłużyła się na przykład Moon (1998), która sfalsyfikowała hipotezę dotycząca częstego występowania jednostek wielowyrazowych o nieumotywowanym znaczeniu i niezmiennej formie (idiomy, powiedzenia itp.).

11 Najczęściej wykorzystujemy takie podejście, określając klasy obiektów poszukiwanych w korpusie (np. definiujemy określone kategorie gramatyczne, typy frazeologizmów itp.), a następnie, na podstawie danych korpusowych, weryfikujemy prawdziwość postawionej przez nas hipotezy. Ponadto w takim podejściu częstą praktyką jest korzystanie z korpusów adnotowanych, tj. opisanych w oparciu o daną teorię języka, zamiast surowych danych tekstowych.

12 Określamy czy poszukujemy sekwencji trzech, czterech lub pięciu wyrazów. Doprecyzowujemy także stopień ciągłości sekwencji wyrazów (tj. czy wewnątrz sekwencji wyrazów występują jakieś komponenty opcjonalne i zmienne (alternanty), a jeśli tak, to czy takim elementem jest jeden wyraz, czy też może więcej wyrazów.

13 W niektórych koncepcjach, a zwłaszcza w badaniach nad wyodrębnianiem kolokacji z tekstów, dodatkowo dookreśla się kryterium statystyczne, stosując różne miary zależności między zmiennymi losowymi, które w wypadku frazeologii są pojedynczymi wyrazami wchodzącymi w skład dłuższych ciągów wielowyrazowych (do takich miar należą m. in. wskaźnik informacji wzajemnej (MI-score), a także testy istotności statystycznej, np. t-test, log-likelihood test, test chi-kwadrat itp.). Więcej o zastosowaniu testów statystycznych w badaniach nad frazeologią można przeczytać w Oakes (1998), Gries (2008), Evert (2005), Biber (2009).

takiego podejścia jest często taksonomiczny opis frazeologii na podstawie danych empirycznych pozyskiwanych z korpusu.

W kontekście frazeologii badanej na sposób korpusowy warto również zwrócić uwagę na rozróżnienie między dwoma podejściami do analizy jednostek wielowyrazowych w zależności od metod ich wyodrębniania (Granger i Paquot 2008: 38). Pierwsze to badanie n-gramów lub zbitek wyrazów (*n-gram analysis / cluster analysis*), polegające na wyodrębnianiu odtwarzalnych, często powtarzających się w tekstach ciągów wielowyrazowych, niezależnie od stopnia ich idiomatyczności i struktury gramatycznej, np. *on the other hand, I guess that, in the case of, as a result of*. W zależności od długości takie jednostki nazywamy dwugramami, trzygramami itd., a ogólnie n-gramami. Można wśród nich wyróżnić nieciągle sekwencje wyrazów, wewnątrz których występują elementy opcjonalne i/lub zmienne, takie jak ramy kolokacyjne (*collocational frameworks*) (Renouf i Sinclair 1991), ramy leksykalno-gramatyczne (*lexicogrammatical frames*) (Moon 1998), skipgramy (*skipgrams*) (Wilks 2005), ramy frazowe (*phrase frames*) (Fletcher 2007), np. *in the * of, the * of the, in order to **, gdzie * oznacza komponent zmienny. Drugie podejście to badania nad współwystępowaniem pary wyrazów (*co-occurrence analysis*), polegające na wyodrębnianiu par wyrazów, których częstość występowania obok lub niedaleko siebie, najczęściej w sąsiedztwie pięciu wyrazów od lewej i prawej strony[14], jest istotna ze statystycznego punktu widzenia (np. *heavy traffic, first-hand experience*, *solid experience* itp.).[15] Typowym przykładem takich jednostek o rożnym stopniu ciągłości są kolokacje[16]. W niniejszym artykule omówię zatem najważniejsze koncepcje wpisujące się w korpusowe badania nad frazeologią.

14 Takie rozumienie kolokacji upowszechniło się od lat osiemdziesiątych ubiegłego wieku, kiedy rozpoczęto prace nad projektem leksykograficznym COBUILD w Birmingham (Wielka Brytania), dla którego potrzeb kolokaty zdefiniowano jako ‘jednostki leksykalne występujące w sąsiedztwie pięciu wyrazów od słowa kluczowego z wyższą frekwencja aniżeli można by oczekiwać na podstawie prawa średnich” (Krishnamurthy 1987). Zainteresowanie łączliwością wyrazów ma swoje źródła w pracach Johna Ruperta Firtha (np. 1957), który ukuł termin „kolokacja”, w znaczeniu tendencji niektórych wyrazów do współwystępowania w tekście, a także zauważył, że znaczenie i sposób użycia danego wyrazu w tekście jest wypadkową znaczenia i użycia wyrazów najczęściej z nim współwystępujących (czyli kolokatów). Takie podejście do badań nad frazeologią bywa obecnie, tj. w czasach, gdy w badaniach wykorzystuje się ogromne, idące już w setki milionów lub miliardy wyrazów elektroniczne korpusy językowe, nazywane podejście neofirthiańskim (np. Pęzik 2013).

15 Badania nad współwystępowaniem wyrazów koncentrują się wokół wypracowania optymalnych metod statystycznych w celu wyodrębnienia kolokacji (tj. par składających się ze słowa kluczowego i kolokatu) dla różnych zastosowań praktycznych i komercyjnych (najczęściej w leksykografii i przy opracowywaniu materiałów do nauczania języków obcych).

16 Warto nadmienić, że zarówno n-gramy, jak i kolokacje są uważane za frazeologizmy (Gries 2008: 4). Rozróżnienie to ma zatem jedynie charakter czysto techniczny.

## 3.1 Zasada idiomu i zasada otwartego wyboru

Podczas badań korpusowych nad związkami między znaczeniem przekazywanym przez jednostki języka a ich formą, Sinclair (1991: 7) zauważył, że podział na jedno i drugie ma charakter sztuczny. Wyjaśnił to w ten sposób, że używając języka w mowie i piśmie użytkownicy języka posługują się nie pojedynczymi wyrazami lub ich kombinacjami, ale gotowymi odtwarzalnymi frazami (frazeologizmami/*phraseologies*), cechującymi się jednością formy i znaczenia. Oznacza to, że to nie pojedyncze wyrazy, ale właśnie dłuższe związki wielowyrazowe są nośnikami znaczeń w tekście. Według Sinclaira (1987, 1991, 2004), język jest probabilistycznym systemem frazeologicznym (w odróżnieniu od tradycyjnego podziału systemu języka na leksykę i gramatykę) i w związku z tym niektóre frazeologizmy pojawiają się w tekstach częściej, niosą więcej znaczeń, a ich wystąpienie jest bardziej prawdopodobne w danym kontekście sytuacyjnym niż innych związków wielowyrazowych. Tym samym użytkownicy języka są bardziej ograniczeni w wyborze środków językowych niż mogłoby się to intuicyjnie wydawać. Nattinger (1980: 341) już wcześniej zauważył, że posługiwanie się językiem to proces kompozycyjny, polegający na łączeniu ze sobą prefabrykowanych fraz (*preassembled phrases*). Aby lepiej zobrazować swoją koncepcję, Sinclair (1987, 1991) zaproponował dwie zasady wyjaśniające, jak znaczenie jest przekazywane w tekście.

Pierwsza nosi nazwę „zasady otwartego wyboru" (*open-choice principle*). Zgodnie z nią w każdym momencie, kiedy użytkownik języka wybiera środki językowe (np. pojedyncze wyrazy, frazy, czy też zdania) w celu zakomunikowania określonych znaczeń, natychmiast otwiera się przed nim niezliczona ilość potencjalnych rozwiązań, jak połączyć te środki językowe z kolejnymi (co przekłada się na inkrementalne tworzenie tekstu), a jedynym ograniczeniem w tym względzie jest poprawność gramatyczna (1991: 109). Jeżeli zatem dwa lub trzy wyrazy łączą się ze sobą, to razem niosą nowe znaczenie, które jest częściowo albo zupełnie niezależne od znaczenia, jakie te same wyrazy niosą oddzielnie (np. *by all means*, *due to*). Sinclair obrazowo określa tę zasadę mianem modelu „szczeliny i wypełniacza" (*slot-and-filler model*), w którym wyrazy łączą się ze sobą w różne kombinacje, aby przekazać określone znaczenie (Sinclair 1991: 8), co w późniejszych pracach nazywa frazeologiczną tendencją języka (2004: 29).

Druga zasada, tj. „zasada idiomu" (*idiom principle*), polega na tym, że użytkownik języka ma do swojej dyspozycji ogromny wybór prefabrykowanych, odtwarzalnych frazeologizmów, które jako całość stanowią pojedyncze jednostki języka, i to pomimo tego, że można je rozbić na mniejsze segmenty, np. wyrazy, morfemy, fonemy, diakryty (Sinclair 1991:110). Takie frazeologizmy nie stanowią kompletnych struktur gramatycznych, są często nieciągłe (np. *a * of*, *in the * of, if * any*), cechują się dużym stopniem wewnętrznego zróżnicowania w zakresie lek-

syki i składni, a także łączą się z innymi wyrazami lub ciągami wyrazów. Przykładowo, w korpusie ulotek dla pacjentów, które przeanalizowałem, konstrukcja (*if * any*) najczęściej wypełniona jest ciągiem dwóch wyrazów, takich jak *you notice*, *you experience*, *you get*, *you develop*, *you suffer*, a dalej łączy się z prawej strony z semantycznie powiązanymi rzeczownikami nacechowanymi negatywnie, takimi jak *problems, symptoms, side-effects, conditions, pain*, np. *if you have any symptoms*. Tym samym wybór całej konstrukcji jest pojedynczym wyborem w celu wyrażenia konkretnego znaczenia, które przyciąga do siebie dłuższe ciągi wyrazów. Zdaniem Sinclaira (2004), kształt jednostek wielowyrazowych (frazeologizmów) jest wypadkową czterech kryteriów wyboru środków językowych, mianowicie prozodii semantycznej (tendencji wyrazów do współwystępowania z wyrazami o nacechowaniu pozytywnym lub negatywnym, np. *to cause problems / trouble*), preferencji semantycznej (tendencji wyrazów do współwystępowania z innymi wyrazami w określonym znaczeniu lub funkcji komunikacyjnej, np. *it has been revealed* / *shown / demonstrated* w artykułach naukowych zazwyczaj przekazuje znaczenie ‘wykazać coś’), koligacji (tendecji wyrazów do współwystępowania z określonymi kategoriami gramatycznymi, np. *like + -ing sth*, *like + to do sth*, *worth + -ing*) i kolokacji (tendencji wyrazów do współwystępowania z innymi wyrazami w bliskim sąsiedztwie). Suma znaczeń niesionych przez frazeologizmy jest relatywna względem sytuacji, w jakiej są one używane. Tym samym potrzeby użytkowników języka w zakresie przekazywania znaczeń (w mowie lub na piśmie) są różne w zależności od sytuacji, co determinuje zakres frazeologii należący do ich repertuaru językowego. Wyrazy mają wszak tendencję do współwystępowania ze sobą w ograniczonych kombinacjach w zależności od typu sytuacji komunikacyjnej (Sinclair 2004: 29).

## 3.2 Pattern Grammar (gramatyka wzorców)

Kolejnym podejściem do opisu regularnie powtarzających się w tekście jednostek wielowyrazowych jest gramatyka wzorców (*Pattern Grammar*) opracowana przez Susan Hunston i Gill Francis (2000), która to koncepcja czerpie dużo ze spostrzeżeń Sinclaira (1991). Zdaniem autorek, wzorzec gramatyczny (*grammar pattern*) to „zbiór wyrazów i struktur gramatycznych, regularnie powiązanych z danym wyrazem i stanowiących o jego znaczeniu” (Hunston i Francis 2000: 37). Innymi słowy, wzorce gramatyczne to frazeologizmy, które nie są pojedynczymi wyrazami ani pustymi (niewypełnionymi leksyką) strukturami gramatycznymi, ale stanowią syntezę tego pierwszego i drugiego, która umożliwia przekazanie konkretnego znaczenia w tekście. Wzorce gramatyczne są powiązane z konkretnymi znaczeniami, ponieważ często znaczenia wyrazów zmieniają się w zależno-

ści od ich występowania w rożnych wzorcach gramatycznych (np. *to start to do sth* vs. *to start doing sth*), a wyrazy występujące w takich samym wzorcach często mają podobne znaczenie, np. we wzorcu '*it* + czasownik + przymiotnik + zdanie określające zaczynające się od spójnika *that*' (*it is interesting/clear/true that* ' to ciekawe/jasne/prawda, że' lub *it is sensible/possible to* 'to sensowne / możliwe, że') przymiotniki należą do jednego pola semantycznego, wyrażając modalność, istotność czegoś, oczywistość.[17] Tym samym wzorce gramatyczne można rozpatrywać z dwóch perspektyw, tj. albo dany wyraz występuje w rożnych wzorcach gramatycznych i w każdym z nich niesie różne znaczenia, albo jeden wzorzec jest powiązany z wieloma różnymi wyrazami, które niosą podobne znaczenie (Hunston & Francis 2000: 83).

Poszukiwanie wzorców gramatycznych wymaga wykorzystania dużych adnotowanych korpusów językowych (podejście oparte na korpusie), ponieważ wzorce gramatyczne winno się wyodrębniać na podstawie wysokiej frekwencji w tekstach ilustrujących użycie języka w różnych sytuacjach komunikacyjnych (Hunston 2009: 156). Niestety ani Hunston i Francis (2000), ani Hunston (2009) nie określają żadnego minimalnego poziomu częstości występowania poszukiwanych jednostek w tekstach. Samo wyodrębnianie wzorców gramatycznych z korpusów – polegające na wnikliwej analizie konkordancji, a następnie grupowaniu potencjalnych wzorców w kategorie funkcjonalno-znaczeniowe z uwzględnieniem rodzajów informacji przekazywanych w różnych kontekstach sytuacyjnych – nie należy do zadań łatwych. Wzorce gramatyczne bywają złożone i trudno zauważalne, gdyż często są to abstrakcyjne konstrukcje, które realizują się w tekście na różne sposoby, np. wielowyrazowiec *it is ironic that* 'paradoksalne jest to, że' jest aktualizacją abstrakcyjnego wzorca '*it* + *be/seem/look* + adjective + zdanie określające zaczynające się od spójnika *that*', w którym przymiotnik wyraża ocenę zaistniałej sytuacji, np. *it is not surprising that* 'nie dziwi to, że'*, it seems very peculiar that* 'wydaje się to bardzo dziwne, że' (przykład podany za Hunston 2009: 154).

## 3.3 Torowanie leksykalne (pryming leksykalny)

Koncepcja prymingu leksykalnego zaproponowana na gruncie anglosaskim przez Micheala Hoey'ego (2005) ma swoje źródła w zjawisku torowania (zwanego również prymingiem, primingiem lub poprzedzaniem), znanego bliżej w psychologii i psycholingwistyce. W dużym skrócie torowanie polega na tym, „że dany wyraz podany podprogowo (a więc nieuświadamiany przez badanego) wpływa na pojawienie się powiązanego z nim w jakiś sposób innego wyrazu" (Kurcz 2005: 20). W praktyce, jak twierdzi Hoey (2005, 2007), wraz z każdym użyciem wyrazu lub fra-

17 Ostatni przykład podaję za Hunston i Francis (2000).

zeologizmu użytkownik języka podświadomie rejestruje jego kontekst sytuacyjny, komunikacyjny i tekstowy. Powoduje to, że wraz z upływem czasu użytkownicy języka podświadomie budują mentalny rejestr kolokacji, koligacji, a także wszelkich innych powiązań tekstowych i semantycznych wyrazów, czy też frazeologizmów. Tym samym czytając, słuchając, mówiąc czy też pisząc, używamy w różnych sytuacjach określonych wyrazów i ich kombinacji, co sprawia, że w przyszłości – używając tych samych wyrazów w podobnych sytuacjach – podświadomie odtwarzamy ich typowe kolokaty, wzorce gramatyczne oraz znaczenia. Oznacza to, że każde wywołanie wyrazu (frazeologizmu) w celu użycia go w dyskursie jest efektem kumulacji wiedzy o występowaniu tego wyrazu lub frazeologizmu, jaką użytkownik języka posiadł we wcześniejszych kontaktach z danym wyrazem lub frazeologizmem. Co za tym idzie, jeżeli tak rozumiany pryming leksykalny wynika z indywidualnego doświadczenia, to owa skumulowana wiedza nie jest własnością danych wyrazów, ale subiektywnego sposobu użycia języka, który można utożsamiać z idiolektem członków danej wspólnoty językowej. Ludzie, używając języka w różnych sytuacjach, mają zróżnicowaną[18] wiedzę o kolokacjach, koligacjach, powiązaniach tekstowych i semantycznych danego wyrazu (Hoey 2007: 9). Na przykład pacjent, lekarz, farmaceuta, analityk w laboratorium i urzędnik zajmujący się refundacją leków w wojewódzkich oddziałach NFZ-tu dysponują różną wiedzą o kolokacjach, koligacjach, powiązaniach tekstowych i semantycznych rzeczownika *lek* i używają go zupełnie inaczej wskutek indywidualnie skumulowanej wiedzy, jaką zdobyli na podstawie regularnego, powtarzalnego kontaktu z tym wyrazem w typowych (ale dla każdej z tych osób różnych) sytuacjach użycia tego wyrazu. Oznacza to, że pryming leksykalny stanowi istotny determinant profilu frazeologicznego różnych typów i gatunków tekstów, a także rejestrów języka. Jest także wypadkową doświadczenia i skumulowanej wiedzy członków wspólnoty językowej (np. profesjonalnej) o sposobie użycia języka w określonych i dobrze im znanych kontekstach sytuacyjnych.

Jak pisze Hoey (2005: 11-12), w zależności od tego, czy użytkownik języka aktywnie lub pasywnie uczestniczy w sytuacjach komunikacyjnych bądź konkretnych aktach komunikacji (przekazuje lub odbiera komunikaty wyrażone danymi aktami mowy), możemy wyróżnić pryming leksykalny produktywny (*productive*) i receptywny (*receptive*). Z tym pierwszy mamy do czynienia wtedy, gdy użytkownik języka regularnie spotyka się z danymi wyrazami w dyskursie, w którym uczestniczy, w gatunkach tekstów, którymi się posługuje, w rejestrach języka, których używa. Na przykład farmaceutka w aptece w Wielkiej Brytanii, wyjaśniając pacjentowi, jak podać dany lek lub jak go przechowywać, użyje ściśle określonych

18 Jak pisze Hoey, "speakers are primed in different ways" (2007: 9)

kolokacji i koligacji z rzeczownikiem *lek*, np. *take medicine, use medicine*[19]*, store medicine, it is important that you keep this medicine*. Receptywny pryming leksykalny wiąże się natomiast z regularnym kontaktem z danymi wyrazami w sytuacjach, w których z dużym prawdopodobieństwem użytkownik języka nigdy nie będzie aktywnie uczestniczyć, np. przeciętny pacjent przychodzący do apteki, który nie pracuje w sektorze opieki zdrowotnej, raczej nie będzie aktywnie uczestniczyć w kursokonferencjach dla lekarzy lub farmaceutów, podobnie zresztą jak autor tego artykułu. W takich sytuacjach użytkownicy języka łatwo rozpoznają, że dane wyrazy lub ich kombinacje (a także kolokacje, koligacje itd.) nie należą do typowych, spotykanych na co dzień, ale pochodzą z zupełnie innych rejestrów języka. Reasumując, istotą prymingu leksykalnego w rozumieniu Hoey'ego (2005, 2007) jest to, że użytkownicy języka podświadomie rejestrują kolokacje, koligacje i wszelkie powiązania tekstowe, semantyczne, funkcjonalne tych frazeologizmów, które występują często w typach i gatunkach tekstów, a także rejestrach języka, które są im najbardziej znane, a przy okazji kolejnego użycia danego frazeologizmu automatycznie odtwarzają znajomy kontekst sytuacyjny, w jakich jest on zazwyczaj przez nich używany.

## 3.4 Zbitki wielowyrazowe

Pojęcie zbitek wielowyrazowych (*lexical bundles*) wprowadził do literatury specjalistycznej amerykański językoznawca Douglas Biber i in. (1999) przy okazji prac nad opisem gramatyki języka angielskiego, obejmującym zarówno odmiany mówione, jak i pisane (*Longman Grammar of Spoken and Written English*). Sama koncepcja została dopracowana i zoperacjonalizowana w późniejszych publikacjach tego samego autora (np. Biber i in. 2004; Biber 2006, 2009). Ogólnie rzecz ujmując, zbitki wielowyrazowe to występujące w tekstach ciągi trzech lub więcej wyrazów, stanowiące swoisty materiał budulcowy dyskursu, niezwykle często używane i odtwarzalne w rożnych sytuacjach komunikacyjnych (Biber i in. 1999: 990-991), np. *I don't think*, *as a result*, *do you want*. Czyni je to utrwalonymi połączeniami wyrazów ściśle powiązanymi z poszczególnymi rejestrami języka.

Zbitki wielowyrazowe wyodrębnia się z korpusów językowych przy pomocy specjalistycznego oprogramowania do analizy tekstu, np. WordSmith Tools (Scott 2008), kfNgram (Fletcher 2007), AntConc (Anthony 2014). Do najważniejszych kryteriów ich wyodrębniania należą długość zbitki, jej znormalizowana frekwencja oraz zakres dystrybucji w tekstach. Tym samym zbitki są wyodrębniane z korpusów w oparciu o kryterium ortograficzne i statystyczne. Jeśli idzie o minimalną

19 Na marginesie wspomnę, że pierwsze dwie kolokacje należą do najczęstszych w korpusie ulotek dla pacjentów w języku angielskim, który przebadałem (Grabowski 2015a).

frekwencję, to przyjmuje się, że krótsze zbitki występują w tekstach częściej niż dłuższe. Dlatego w wypadku tych pierwszych stosuje się niższe progi frekwencji, np. dla zbitek czterowyrazowych Biber i in. (1999: 990) stosują próg 10 wystąpień na milion wyrazów, a dla zbitek pięciu lub sześciu wyrazów – jedynie 5 wystąpień. W badaniach tekstów szablonowych i bardziej skonwencjonalizowanych niektórzy badacze stosują wyższe progi frekwencji, np. Juknevičienė (2009), Bernardini i in. (2010), Gozdz-Roszkowski (2011a), Gaspari (2013) uznają za zbitki czterowyrazowe jedynie takie ciągi wyrazów, które występują w tekście częściej niż 40 razy na milion wyrazów tekstowych. Pozwala to zdecydowanie ograniczyć ilość materiału badawczego dla późniejszych analiz funkcji dyskursywnych zbitek. Wszak w niektórych rejestrach języka (np. akademickim, prawniczym) niektóre zbitki występują nawet częściej niż 200 razy na milion wyrazów. Aby ograniczyć specyficzne i nietypowe użycia zbitek, stosuje się kryterium dystrybucji, według którego za zbitki wielowyrazowe traktuje się tylko ciągi wyrazów występujące w co najmniej pięciu tekstach z danego gatunku lub reprezentujących ten sam rejestr języka (Biber i in. 2003: 134). Wszystko to pozwala odróżnić odtwarzalne i utrwalone zbitki, które można uznać za związki frazeologiczne typowe dla danego rejestru, od ciągów wyrazowych tworzonych bardziej doraźnie, tj. od produktów językowych.

Najczęściej zbitki wielowyrazowe stanowią część fraz rzeczownikowych i wyrażeń przyimkowych, stanowią niepełne jednostki składniowe[20], a często wchodzą w skład dwóch lub zazębiają się z dwoma lub więcej jednostkami składniowymi (np. *I don't know why*, *did you see that*). Wyjątkiem są tutaj ciągi wyrazów oddzielone znakami interpunkcyjnymi, ponieważ takie w ogóle nie są traktowane jak zbitki wielowyrazowe.

Wyniki badań przeprowadzonych przez m.in. Bibera i in. (1999, 2003, 2004), Bibera (2009), Kena Hylanda (2008, 2009), Stanisława Goździa-Roszkowskiego (2011a) pokazują znaczące różnice w zakresie dystrybucji rożnych typów strukturalnych zbitek wielowyrazowych w zależności od rejestru języka, np. 15% zbitek w konwersacjach to kompletne struktury składniowe (głównie zwroty czasownikowe lub zdania), natomiast w prozie akademickiej to tylko 5%, z których większość wchodzi w skład dłuższych fraz rzeczownikowych (nominalnych) i wyrażeń przyimkowych (Biber i in. 2003: 135). W przytłaczającej większości zbitki wielowyrazowe mają znaczenie umotywowane, które wynika ze znaczeń ich komponentów, a także są trudno dostrzegalne w tekście (Biber i in. 2003: 134), zwłaszcza w porównaniu z typowymi idiomami, porzekadłami, przysłowiami czy skrzydlatymi słowami. Zdaniem Bibera i in. (2003, 2004) wynika to z faktu, że zbitki są budul-

20 Biber i in. (1999: 995) twierdzą, że jedynie 5% zbitek wielowyrazowych w prozie akademickiej (artykuły naukowe, podręczniki akademickie) to kompletne jednostki składniowe.

cem dyskursu, występują w tekstach z wysoką frekwencją, podczas gdy idiomy, powiedzenia i podobne typy frazeologizmów o nieumotywowanym znaczeniu nie dość, że są bardzo rzadkie (co wykazała Moon we wcześniej wspomnianym badaniu), to na dodatek w tekście zazwyczaj pełnią funkcje stylistyczne (są swoistymi ornamentami tekstowymi) i dlatego bardziej rzucają się w oczy czytelnikom danego tekstu. Tym samym w badaniach nad zbitkami wielowyrazowymi w polu zainteresowania frazeologów znalazły się jednostki o umotywowanym znaczeniu, stanowiące niekompletne struktury składniowe, często występujące w tekstach, jednak słabo dostrzegalne (Goźdź-Roszkowski 2011a: 44).

Większość dotychczasowych badań nad zbitkami wielowyrazowymi wykonano na materiale tekstów napisanych w języku angielskim. Na gruncie polskim próbę zbadania częstości użycia, dystrybucji i funkcji dyskursywnych zbitek wielowyrazowych w specjalnie zebranym do tego celu korpusie ulotek dla pacjentów napisanych w języku polskim podjął Grabowski (2014), który zaproponował również szereg propozycji metodologicznych w tym zakresie. W dużym skrócie wyniki badania wykazały związek między wysoką frekwencją powtarzalnych zbitek wielowyrazowych a funkcją komunikacyjną i kontekstem sytuacyjnym, w jakim używane są zazwyczaj ulotki dla pacjentów.

## 3.5 Skipgramy, ramy frazowe i concgramy

Zdaniem Sinclaira (2001: 351-353) badania nad zbitkami wielowyrazowymi i innymi typami powtarzalnych ciągów wyrazowych (n-gramów) zapchnęły na drugi plan inne jednostki wielowyrazowe, tj. połączenia nieciągłe o zmiennej formie oraz zróżnicowanej pozycji składników względem siebie. W wyniku tych ograniczeń opracowano takie koncepcje i metody wyodrębniania jednostek wielowyrazowych, jak m.in. skipgramy (*skipgrams*) i ramy frazowe (*phrase frames*), zaproponowane odpowiednio przez Yoricka Wilksa (2005), głównie na potrzeby przetwarzania języka naturalnego, i Williama Fletchera (2007).

Zdaniem Guthriego i in. (2006), skipgramy to pewne uogólnienia n-gramów (tj. ciągów dwóch, trzech i więcej wyrazów), które oprócz sekwencji składników bezpośrednio sąsiadujących ze sobą obejmują również nieciągłe połączenia wyrazów, co pozwala na rozwiązanie problemu wariancji składników. Na przykład zdanie *I hit the tennis ball* składa się z trzech trzygramów, mianowicie *I hit the*, *hit the tennis* i *the tennis ball*, jednak więcej o frazeologii powie nam skipgram *hit the * ball*[21], który uwypukla ukryty w tym zdaniu wzorzec frazeologiczny (Guthrie 2006: 1222). Koncepcja skipgramu jest jednak mocno ograniczona, i to zwłaszcza, jeśli zastanowimy się nad jej szerszym zastosowaniem w badaniach nad fra-

21 Np. *hit the tennis/golf/cue ball*.

zeologią. Jak słusznie zauważa Wilks (2005), skipgramy ograniczają się do dwu- lub trzywyrazowych sekwencji, wewnątrz których mogą wystąpić maksymalnie cztery zmienne składniki utożsamiane z wyrazami. Oznacza to, że dwa lub trzy w jakiś sposób powiązane ze sobą wyrazy, tworzące pewien ukryty wzorzec frazeologiczny, nie zostaną wyodrębnione, jeśli znajdą się dalej niż cztery wyrazy od siebie (Warren 2010: 115).[22]

Z kolei ramy frazowe to zbiory identycznych wariantów n-gramów z wyjątkiem jednego wyrazu wchodzącego w ich skład (Fletcher 2007), wyodrębnione przy wydatnej pomocy oprogramowania komputerowego (np. aplikacji kfNgram zaprojektowanej również przez Fletchera) na podstawie analizy n-gramów będących ciągłymi sekwencjami co najmniej ośmiu wyrazów. Zdaniem Warrena (2010: 115), ramy frazowe to pewna wariancja skipgramów, ograniczona co do liczby opcjonalnych wyrazów we wzorcu frazeologicznym, który wynosi jeden (a nie maksymalnie cztery, jak w wypadku skipgramów). Program kfNgram umożliwia zarówno ręczne wpisywanie poszukiwanych ram frazowych, jak i ich automatyczne wyodrębnianie z tekstów (wraz z listą wariantów danej ramy frazowej), podając również informację na temat frekwencji danej ramy frazowej, a także frekwencji jej poszczególnych wariantów, zaktualizowanych w tekście. Przykład ramy frazowej *if you * any* zaczerpnięty z korpusu angielskich ulotek dla pacjentów przedstawiam w poniższej tabeli.[23]

Tabela 1: Rama frazowa *if you * any* i jej warianty (aktualizacje w tekście)

| **Rama frazowa** | **Warianty (aktualizacje w tekście)** | **Frekwencja** | **Liczba wariantow ramy frazowej** |
|---|---|---|---|
| *if you * any* | | 786 | 7 |
| | *if you have any* | 516 | |
| | *if you notice any* | 83 | |
| | *if you experience any* | 75 | |
| | *if you get any* | 69 | |
| | *if you develop any* | 19 | |
| | *if you think any* | 18 | |
| | *if you suffer any* | 6 | |

22 Ponadto – jak piszą Cheng i in (2006: 414) – istniejące oprogramowanie do wyodrębniania skipgramów z tekstów jest mało przyjazne dla użytkownika, ponieważ wymaga czasochłonnego wprowadzania poleceń o sformalizowanej składni.

23 Warren (2010: 115) zwraca uwagę na ograniczenie, dotyczące tylko jednego zmiennego komponentu ramy frazowej. W Tabeli 1 nie znajdziemy takich sekwencji, jak np. *if you seriously have any* albo *if you definitely have any*, gdzie zmienne są sekwencje dwóch wyrazów (np. *seriously have* albo *definitely have*).

Zdaniem Ute Roemer (2009a: 91), spis ram frazowych i ich wariantów może stanowić źródło cennych danych językowych, przydatnych w opisach profili frazeologicznych rożnych typów i gatunków tekstów, a także posłużyć za ilustrację ich zróżnicowania frazeologicznego. Ponadto Forsyth i Grabowski (2014) udowodnili, że ramy frazowe, o różnym stopniu wariantywności w tekstach, mogą również posłużyć do zmierzenia formuliczności tekstów, a tym samym do opracowania rankingu zróżnicowania frazeologicznego/formuliczności tekstów.

Niemniej jednak zarówno skipgramy, jak i ramy frazowe nie rozwiązują problemu wariancji pozycji składniowej komponentów wchodzących we wzorce frazeologiczne. Jak pisze Warren (2010: 116-121), poszukiwanie wzorców współwystępowania kolokacji *structural design* w tekstach specjalistycznych z pewnością zakończy się odnalezieniem takich sekwencji (n-gramów), jak *the structural design*, *safe structural design*, *structural framing design* itp, ale jednak sekwencje, w których słowo kluczowe i kolokat występują zamiennie w różnych pozycjach względem siebie, nie zostaną zidentyfikowane, np. *design of structural framing*, *design of structural models*. Aby temu zaradzić, opracowano koncepcję i metodę wyodrębniania tzw. concgramów (Cheng i in. 2006; Greaves 2009) przy pomocy specjalistycznego oprogramowania Concgram 1.0 (Greaves 2009). Program ten umożliwia poszukiwanie różnych wariantów frazeologizmów o dowolnej długości, uwzględniając zarówno wariancje ich składników, jak i wariancje pozycji owych składników w szyku zdania. Cheng i in. (2006: 418) definiują concgramy jako „zbiory wszystkich permutacji wariancji składników oraz ich pozycji w szyku zdania, wygenerowanych na podstawie powiązań między parami dwóch lub więcej wyrazów w tekście". Zbiory takie są dostępne pod postacią konkordancji, ilustrujących użycie danych wyrazów w pełnym kontekście tekstowym.

W odróżnieniu od terminów „słowo kluczowe" (*node*) i „kolokat" (*collocate*), przypisanych do kolokacji i podkreślających prymat tego pierwszego nad drugim, Cheng i in. (2006) używają jednego terminu „źródło" (*origin*) na oznaczenie wyrazu lub frazy współwystępującej w tekście z innym wyrazem lub frazą.[24] Mówiąc ściślej, concgramy mogą stanowić zbiory od dwóch do pięciu wyrazów współwystępujących ze sobą w sąsiedztwie dwunastu wyrazów od źródła. Tym samym dzięki concgramom jesteśmy w stanie odnaleźć w tekście wyrazy powiązane ze sobą niezależnie od ich pozycji względem siebie w szyku zdania. Zdaniem Greavesa (2009), ma to istotne znaczenie zwłaszcza w badaniach nad kolokacjami, ponieważ czasowniki i przysłówki, czy też rzeczowniki i przymiotniki, mogą występować w zdaniach w różnych pozycjach względem siebie. Na przykład przymiotniki mogą znajdować się w pozycji atrybutu albo predykatu względem rzeczowników i w obu wypadkach z frazeologicznego punktu widzenia ich współwystę-

24 Wpisując oba elementy do okienka wyszukiwania programu, oddzielamy je ukośnikiem.

powanie jest istotne. Tym samym dzięki analizie concgramów uzyskujemy wgląd we wzorce współwystępowania wyrazów w ich sekwencjach nieciągłych. Można concgramy uważać także za wyższy poziom abstrakcyjnego uogólnienia wzorców frazeologicznych względem skipgramów, ram frazowych i zbitek wielowyrazowych. Konkordancje ilustrujące aktualizacje concgramów w tekstach mogą być potencjalnie wykorzystane w celu wyodrębnienia z tekstów ram frazowych i zbitek wielowyrazowych (o ile te ostatnie spełniają określone kryteria ortograficzne i statystyczne), a także dokonania pełniejszego rozróżnienia między utrwalonymi związkami frazeologicznymi i doraźnie tworzonymi produktami językowymi.

## 3.6 Sekwencje semantyczne

Koncepcja sekwencji semantycznych zaproponowana przez Hunston (2008) jest w pewnym sensie odwróceniem i rozszerzeniem wcześniejszej koncepcji wzorców gramatycznych (Hunston i Francis 2000). Jeśli we wcześniejszym podejściu punktem wyjścia była forma, czy też kształt, wzorców gramatycznych, które wiązano ze znaczeniem przekazywanym w danym kontekście, to w wypadku sekwencji semantycznych na plan pierwszy wysuwa się znaczenie odtwarzalnych ciągów wyrazów. Oznacza to, że wzorce gramatyczne są integralną częścią sekwencji semantycznych (Hunston 2008: 272), które badaczka definiuje jako „odtwarzalne sekwencje wyrazów i fraz, które mogą znacząco różnić się formą i dlatego powinny być traktowane przede wszystkim jako sekwencje semantyczne, a nie kształtowe (formalne)". Dalej Hunston (2008: 292) pisze, że sekwencje semantyczne są produktem uwarunkowań społecznych, które wymagają od użytkowników języka przekazywania [w pierwszej kolejności] określonych komunikatów [tj. przekazywania określonych znaczeń], a nie użycia gotowych prefabrykowanych fraz. Typowa sekwencja semantyczna składa się z rdzenia (*core word*), który może być albo wyrazem leksykalnym, funkcyjnym, albo dłuższą frazą, po których znajdujemy wzorce gramatyczne w funkcji dopełnienia danego wyrazu lub frazy (np. zdanie dopełniające rozpoczynające się od spójnika *that,* tzw. *that*-clause), a także inne powiązane ze sobą znaczeniowo rodzaje fraz występujące zarówno przed, jak i po rdzeniu (Hunston 2008: 272). Na przykład znaczenie modalności, powiązane ze zobowiązaniami, koniecznością czy też możliwością zrobienia czegoś, może być powiązane z taką sekwencją semantyczną, jak np. 'to *make sure* + *that*-clause' (gdzie zwrot czasownikowy *to make sure* jest rdzeniem sekwencji), np. *You need to make sure (that) the telescope stays in correct shape* (przykład za Hunston 2008: 273-278).

Ponieważ sekwencje semantyczne to jednostki odtwarzalne w danych sytuacjach komunikacyjnych, to kluczowym kryterium przy ich wyodrębnianiu staje się wy-

soka frekwencja w tekście (Hunston 2008: 272). Jednak Hunston nie doprecyzowuje tego kryterium (pisząc na stronie 280. wyżej cytowanej pracy o 40 wystąpieniach w korpusie lub więcej). Niemniej jednak badaczka zauważa, że skumulowana frekwencja poszczególnych składników sekwencji semantycznej jest bardziej wiarygodnym kryterium niż bezwzględna frekwencja rdzenia (Hunston 2008: 272).

Schulze i Roemer (2008: 265) podkreślają, że badania nad sekwencjami semantycznymi mogą rzucić nowe światło na proces konstruowania znaczenia poprzez użycie frazeologizmów, a także mogą pomóc w zrozumieniu, jak frazeologizmy (tj. poszczególne ciągłe i nieciągłe sekwencje wyrazów) kształtują przekonania, wartości i wzorce społeczne, zwłaszcza w wypadku znaczeń przekazywanych przez jednostki wielowyrazowe typowe dla danych dziedzin specjalistycznych. Zdaniem Goździa-Roszkowskiego, który przeprowadził badanie pilotażowe polegające na analizie wybranych sekwencji semantycznych w amerykańskich tekstach prawniczych (2011b), analiza tego typu jednostek może dostarczyć cennych danych, niezbędnych do opracowania bardziej wszechstronnego profilu frazeologicznego[25] różnych typów i gatunków tekstów, typowych dla poszczególnych dyscyplin lub specjalistycznych dziedzin wiedzy, o czym wspomina również Hunston (2008: 272).

## 4. Podsumowanie

Zdaniem Wojciecha Chlebdy, „frazeologizmy są takie, jakie są kryteria frazeologiczności, przyjęte przez danego badacza, są tym, za co uzna je dany językoznawca. Powołujemy frazeologizmy do życia od nowa w analizowanych tekstach mocą tak lub inaczej sformułowanej definicji" (2003: 39). Jak zatem rozumieć frazeologię w świetle omówionych wyżej koncepcji?

Po pierwsze, wszystkie one polegają na empirycznej analizie związków wielowyrazowych i ich kombinacji (które w niniejszym artykule dla uproszczenia nazywam 'frazeologizmami') na podstawie ich obserwacji w kolekcjach tekstów elektronicznych zwanych korpusami. Wyodrębnianie frazeologizmów, rozumianych jako zjawiska tekstowe, a nie systemowe, odbywa się najczęściej automatycznie lub półautomatycznie przy pomocy specjalistycznego oprogramowania do analizy tekstu, w oparciu o mniej lub bardziej sprecyzowane kryteria ortograficzne lub

25 Terminu „profil frazeologiczny tekstu lub typu tekstu" używam w odniesieniu do modelu analitycznego zaproponowanego przez Roemer (2010). W dużym skrócie – ów model umożliwia sporządzenie wykazu jednostek frazeologicznych wyodrębnionych z tekstu lub korpusu. W praktyce wykorzystanie tego modelu sprowadza się do czterech etapów analizy tekstu, tj. wyodrębnienia frazeologizmów z tekstów, określenia stopnia wariancji wewnątrz poszczególnych frazeologizmów, określenia funkcji dyskursywnych wyodrębnionych frazeologizmów oraz zbadania dystrybucji frazeologizmów w różnych typach lub gatunkach tekstów (ibid.: 97).

statystyczne. Najczęściej są nimi długość danego ciągu wyrazów, jego frekwencja w tekście, zakres dystrybucji, czasem siła wzajemnego powiązania jego składników mierzona testami statystycznymi. Wybór takich kryteriów skutkuje nie tylko zmniejszeniem się roli intuicji badacza, ale także zepchnięciem na dalszy plan takich typów frazeologizmów, jak np. idiomy, przysłowia, metafory, czy też skrzydlate słowa, które pojawiają się w tekstach niezwykle rzadko, a które tradycyjnie znajdowały się w centrum zainteresowania frazeologów.

Wyodrębnione z tekstów na sposób korpusowy związki wielowyrazowe są najczęściej niekompletnymi strukturami składniowymi o umotywowanym znaczeniu, wynikającym ze znaczeń poszczególnych składników tych związków. Mogą one być zarówno ciągłymi, jak i nieciągłymi sekwencjami wyrazów, wykazującymi wysoki stopień wariancji w zakresie kształtu składników, a także ich pozycji w szyku zdania. Używane są one często jako całostki w różnych sytuacjach komunikacyjnych w celu przekazania konkretnego komunikatu (znaczenia). Sprawia to, że z perspektywy użytkownika języka podziały między leksyką a frazeologią są sztuczne i niepotrzebne. W tym miejscu wyraźnie słychać echo poliwanowowskiego, syntetycznego nurtu badań nad frazeologią, koncepcji Andrzeja Bogusławskiego (1976) postulującej odrzucenie przeciwstawienia leksyki i frazeologii jako różnych w jakimkolwiek nietechnicznym sensie, a także chlebdowskiej frazematyki (Chlebda 2003). Tym samym klasyfikacja frazeologizmów w oparciu o szereg kryteriów formalnych i semantycznych, rozdział jednostek leksykalnych jednowyrazowych od jednostek wielowyrazowych, a także wytyczanie granic frazeologii stają się zadaniami drugoplanowymi (Granger i Paquout 2008: 28-29). Stoi to w opozycji zwłaszcza do nurtu winogradowowskiego, w którym za frazeologizm uważano jedynie „utarte połączenie wyrazów, odtwarzane w mowie w gotowej postaci, o znaczeniu metaforycznym niewynikającym z sumy znaczeń składników, obrazowe i wzmacniające ekspresywność wypowiedzi" (Chlebda 2003: 31), a swobodne połączenia wyrazów i pojedyncze wyrazy w ogóle nie znajdowały się w polu zainteresowań frazeologów.

Zdaniem Bibera, Conrad i Cortes (2004: 372) można wyróżnić dwa typy badań empirycznych nad frazeologią na sposób korpusowy. Pierwszy z nich jest ukierunkowany na opracowanie coraz to lepszych metod i narzędzi, niezbędnych do wyodrębniania jednostek wielowyrazowych z tekstów. Drugi typ badań polega na analizie funkcji dyskursywnych jednostek wielowyrazowych w tekstach. Oba typy badań często różnią się pod względem zakresu. Jedne są próbą szczegółowego opisu wszystkich jednostek wielowyrazowych; inne koncentrują się na niewielkim zbiorze jednostek, które z jakichś powodów wydają się istotne dla badacza. Badania różnią się także pod względem wyboru i odmiennego doprecyzowania kryteriów wyodrębniania frazeologizmów z tekstów (np. długość frazeologizmów) – niektóre badania koncentrują się jedynie na ciągłych sekwencjach wyrazów, inne

na nieciągłych, a jeszcze inne na obu typach; niektóre badania koncentrują się na kolokacjach, inne na dłuższych sekwencjach wyrazów. Badacze korzystają również z korpusów o różnych rozmiarach i strukturze. Niektórzy preferują korpusy składające się z kilku tekstów, a jeszcze inni przeszukują korpusy idące w miliony wyrazów tekstowych, np. tzw. korpusy narodowe, takie jak Narodowy Korpus Języka Polskiego, czy Brytyjski Korpus Narodowy (BNC); niektórzy korzystają z pełnych tekstów, a niektórzy jedynie z ich fragmentów. Na koniec, niektórzy badają frazeologizmy w tekstach reprezentujących jeden rejestr języka, natomiast inni badacze preferują studia na korpusach zawierających różne typy i gatunki tekstów. To drugie podejście ma niewątpliwą zaletę, gdyż umożliwia prowadzenie badań porównawczych w zakresie występowania, dystrybucji i funkcji frazeologizmów w różnych rejestrach języka.[26]

Tym samym *frazeologię korpusową* można zdefiniować jako empiryczne, indukcyjne podejście do badania użycia, dystrybucji i funkcji dyskursywnych ciągłych i/lub nieciągłych odtwarzalnych związków wielowyrazowych występujących z wysoką frekwencją w tekstach, które to związki są najczęściej niepełnymi strukturami składniowymi o umotywowanym znaczeniu, wyodrębnianymi z korpusów językowych w sposób automatyczny lub półautomatyczny przy pomocy specjalistycznego oprogramowania, a także przy zastosowaniu kryteriów ortograficznych (formalnych) i statystycznych. Frazeologia korpusowa umożliwia także badaczom testowanie coraz to nowych metod wyodrębniania jednostek wielowyrazowych. W tym miejscu warto jeszcze raz wspomnieć, że w literaturze specjalistycznej funkcjonują już takie określenia, jak *frazeologia dystrybucyjna* czy *frazeologia sterowana frekwencją*. Moim zdaniem jednak *frazeologia korpusowa* ma znaczenie szersze. Dwa pierwsze określenia sugerują nacisk na analizę dystrybucji jednostek wielowyrazowych w tekstach (*frazeologia dystrybucyjna*) lub na stosowanie statystycznego kryterium frekwencji przy wyodrębnianiu jednostek wielowyrazowych z tekstów (*frazeologia sterowana frekwencją*). A przecież wyodrębnianie i badanie frazeologizmów na sposób korpusowy to zdecydowanie więcej. Obejmuje ono przecież wykorzystanie różnych metod sterowanych korpusem (*corpus-driven approach*), zwłaszcza na etapie wyodrębniania jednostek wielowyrazowych, oraz metod opartych na korpusie (*corpus-based approach*), zwłaszcza na etapie analizy konkordancji, ilustrujących użycie jednostek wielowyrazowych w pełnym kontekście tekstowym (np. aby określić ich funkcje dyskursywne). W tym drugim wypadku często korzystamy z gotowych ram teoretycznych, typologii funkcjonalnych, na których podstawie wysuwamy hipotezy, które weryfikujemy przy pomocy danych korpusowych. Wreszcie możemy wykorzystać podejście ilustro-

26 Wydaje się, że powyższa charakterystyka badań nad frazeologią na sposób korpusowy opisana przez Bibera i in. (2004) dotyczy metodologii badań korpusowych w ogóle.

wane korpusem (*corpus-illustrated approach*), aby pozyskać nową wiedzę o kontekstach występowania zadanych przez nas jednostek wielowyrazowych o określonym kształcie lub sprawdzić frekwencję danych frazeologizmów w tekstach.

Podsumowując w niniejszym artykule pozwoliłem sobie przybliżyć czytelnikom pewne tezy i rozważania nad frazeologią, na którą spojrzałem z perspektywy językoznawcy korpusowego. Wydaje się to celowe, ponieważ z jednej strony badania korpusowe nad frazeologią w Polsce są ciągle stosunkowo rzadkie, a z drugiej strony językoznawcy korpusowi i komputerowi bywają słabo zaznajomieni z bardziej tradycyjnymi nurtami badań nad frazeologią, gdzie nie było ani miejsca, ani często możliwości, aby wspomagać się w analizach komputerem. Mam nadzieję, że niniejszy artykuł jest właściwym krokiem w kierunku zmiany tej sytuacji, a także, że pomoże on zwiększyć świadomość badaczy (zarówno filologów, jak i neofilologów) w zakresie możliwości i ograniczeń, jakie wiążą się z różnymi podejściami do badań nad frazeologią.[27]

## Bibliografia

Anthony, L., 2014, AntConc (Version 3.4.3), Tokyo. http://www.laurenceanthony.net/

Bernardini, S., Ferraresi, A., Gaspari, 2010, *Institutional academic English in the European context: A web-as-corpus approach to comparing native and non-native language*, [w:] *Professional English in the European context: The EHEA challenge*, red. A. Linde Lopez, R. Crespo Jimenez, Bern, s. 27-53.

Biber, D., 2006, *University Language. A corpus-based study of spoken and written registers*, Amsterdam.

Biber, D., 2009, *A corpus-driven approach to formulaic language in English: multi-word patterns in speech and writing*.[w:] "International Journal of Corpus Linguistics" 14 (3), s. 275–311.

Biber, D., Conrad, S., Cortes, V., 2003, *Lexical bundles in speech and writing: An initial taxonomy*, [w:] *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*, red. A. Wilson, P. Rayson, T. McEnery, Frankfurt am Main, s. 71–92.

Biber, D., Conrad, S., Cortes, V., 2004, *If you look at…: Lexical bundles in university teaching and textbooks*, [w:] "Applied Linguistics" 25 (3), s. 371-405.

Biber, D., S. Johansson, G. Leech, S. Conrad, E. Finegan, 1999, *The Longman grammar of spoken and written English*, Londyn.

Bogusławski, A., 1976, *O zasadach rejestracji jednostek języka*, [w:] „Poradnik językowy" 8, s. 356-364.

27 Obecnie frazeologizmy są obiektem badań specjalistów z takich różnych dziedzin, jak np. leksykografia, przetwarzanie języka naturalnego, czy też tłumaczenie maszynowe. Dlatego też tak ważne wydaje się zwiększenie ogólnej świadomości co do możliwości i ograniczeń, jakie wiążą się z różnymi spojrzeniami na frazeologię.

Bogusławski, A., 1978. *Jednostki języka a produkty językowe. Problem tzw. orzeczeń peryfrastycznych*, [w:] *Z zagadnień słownictwa współczesnego języka polskiego*, red. M. Szymczak, Wrocław, s. 15-30.

Bogusławski, A., 1989. *Uwagi o pracy nad frazeologią*. [w:] *Studia z polskiej leksykografii współczesnej*, t. 3, red. Z. Saloni, Białystok, s. 13-30.

Cheng., W, Greaves, C., Warren, M., 2006. *From n-gram to skipgram to concgrams,* [w:] "International Journal of Corpus Linguistics" 11 (4), s. 411-433.

Chlebda, W., 2003, *Elementy frazematki: wprowadzenie do frazeologii nadawcy*. Łask.

Chlebda, W., 2009, *Idiomatykon 4: gdzie jesteśmy, dokąd zmierzamy (i parę zdań o tym, skąd przychodzimy*, [w:] *Podręczny idiomatykon polsko-rosyjski 4*, red. W. Chlebda, Opole, s. 9-38.

Chlebda, W., 2010, *Nieautomatyczne drogi dochodzenia do reproduktów wielowyrazowych*, [w:] *Na tropach reproduktów: w poszukiwaniu wielowyrazowych jednostek języka*, red. W. Chlebda, Opole, s. 15-35.

Cowie, A., 1981, *The treatment of collocations and idioms in learners' dictionaries*, [w:] "Applied Linguistics" 2 (3), s. 223–235.

Cowie, A., 1994, *Phraseology*, [w:] *The Encyclopedia of Language and Linguistics*, red. R. Asher, Oxford: s. 3168–3171.

Cowie, A. 1998, *Introduction*, [w:] *Phraseology: Theory, Analysis and Applications*, red. A. Cowie, Oxford: 1-20.

Evert, S., 2005, *The Statistics of Word Cooccurrences: Word Pairs and Collocations,* Stuttgart. Nieopublikowana praca doktorska. http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/pdf/Evert2005phd.pdf

Firth, J., 1957, *A Synopsis of Linguistic Theory 1930-1955* [w:] *Studies in Linguistic Analysis*. Oxford, s. 1-32 (cyt. w Hunston & Francis 2000: 230).

Fletcher, W., 2007, *KfNgram*, Annapolis. http://www.kwicfinder.com/kfNgram/

Forsyth, R., Grabowski, Ł., 2014. *Is there a formula for formulaic language?,* Referat przedstawiony na konferencji naukowej Formulaic Language Research Network (FLaRN 2014). Swansea, Wielka Brytania, 14-16.07.2014. [artykuł w recenzji].

Gaspari, F., 2013, *A phraseological comparison of international news agency reports published online: Lexical bundles in the English-language output of ANSA, Adnkronos, Reuters and UPI*, [w:] "Studies in Variation, Contacts and Change in English" 13. http://www.helsinki.fi/varieng/journal/volumes/13/gaspari/

Górski, R., 2012, *Zastosowanie korpusów w badaniu gramatyki,* [w:] *Narodowy Korpus Języka Polskiego*, red. A. Przepiórkowski, M. Bańko, R. Górski, B. Lewandowska-Tomaszczyk B., Warszawa, s. 291-300.

Goźdź-Roszkowski, S., 2011a, *Patterns of Linguistic Variation in American Legal English. A Corpus-Based Study*, Frankfurt am Main.

Goźdź-Roszkowski, S., 2011b, *The phrase, the whole phrase…Investigating semantic sequences in legal discourse*, Referat przedstawiony na konferencji naukowej Practical Applications of Language Corpora (PALC 2011). Łódź, 13-15.04.2011.

Goźdź-Roszkowski, S., 2013, *Frazeologia dystrybucyjna w ujęciu korpusowym. Zastosowania w analizie języka prawa,* Referat przedstawiony na konferencji naukowej

Polskiego Towarzystwa Lingwistyki Stosowanej *Języki – Teksty – Akty Komunikacyjne.* Olsztyn, 5-6.04.2013.

Grabowski, Ł., 2014, *On Lexical Bundles in Polish Patient Information Leaflets: A Corpus-Driven Study*, [w:] "Studies in Polish Linguistics" 9 (1), s. 21-43.

Grabowski, Ł., 2015a, *Keywords and lexical bundles within English pharmaceutical discourse: a corpus-driven description*, [w:] "English for Specific Purposes" 38, 23-33.

Grabowski, Ł., 2015b, *Phraseology in English Pharmaceutical Discourse: A Corpus-Driven Study of Register Variation*, Opole.

Granger, S., Meunier, F., 2008, *Introduction: The many faces of phraseology* [w:] *Phraseology: An interdisciplinary perspective*, red. S. Granger, F. Meunier, Amsterdam, s. xix-xxx.

Granger, S., Paquot, M., 2008, *Disentangling the phraseological web*, [w:] *Phraseology: An interdisciplinary perspective*, red. S. Granger, F. Meunier, Amsterdam, s. 27-50.

Greaves, C., 2009, *ConcGram 1.0: a phraseological search engine*, Amsterdam.

Gries, S., 2008, *Phraseology and linguistic theory: a brief survey*, [w:] *Phraseology: An interdisciplinary perspective*, red. S. Granger, F. Meunier, Amsterdam, s. 3-26.

Guthrie, D., Allison, B., Liu, W., Guthrie, L. and Wilks, Y., 2006, *A Closer Look at Skip--gram Modelling,* [w:] *Proceedings of Fifth international Conference on Language Resources and Evaluation (LREC)*, Genoa, s. 1222-1225.

Hoey, M., 2005, *Lexical Priming: A New Theory of Words and Language*, Londyn.

Hoey, M., 2007, *Lexical priming and literary creativity*, [w:] *Text, Discourse and Corpora*, red. M. Hoey, M. Mahlberg, M. Stubbs, W. Teubert, Londyn, s.7-30.

Hunston, S., 2008, *Starting with the small words: Patterns, lexis and semantic sequences*, [w:] "International Journal of Corpus Linguistics" 13 (1), s. 271-295.

Hunston, S., 2009, *How can a corpus be used to explore patterns?*, *The Routledge Handbook of Corpus Linguistics*, red. A. O'Keefe and M. McCarthy, Londyn, s. 152-166.

Hunston, S., Francis, G., 2000, *Pattern Grammar: a corpus-driven approach to the lexical grammar of English*, Amsterdam.

Hyland, K., 2008, *As can be seen: Lexical bundles and disciplinary variation,* [w:] "English for Specific Purposes" 27, s. 4-21.

Jukneviciene, R., 2009, *Lexical bundles in learner language: Lithuanian learners vs. native speakers,* [w:] "Kalbotyra" 61 (3), s. 61–71.

Krishnamurthy, R., 1987, *The Process of Compilation*, [w:] *Looking Up: An account of the COBUILD Project in lexical computing*, red. J. Sinclair, Londyn. http://acorn.aston.ac.uk/RK-publications/1987-looking-up-CLEAN.pdf

Kurcz, I., 2005, *Psychologia języka i komunikacji,* Wydanie drugie, Warszawa.

Lewandowska-Tomaszczyk, B., 2005, *Powstanie i rozwój językoznawstwa korpusowego*, [w:] *Podstawy językoznawstwa korpusowego*, red. B. Lewandowska-Tomaszczyk, Łódź, s. 9-26.

Lewicki, A., 1976, *Wprowadzenie do frazeologii syntaktycznej: teoria zwrotu frazeologicznego*, Katowice.

McEnery, T., Wilson, A., 1996, *Corpus Linguistics*, Edynburg.

Mielczuk, I., 1996, *Lexical functions: a tool for the description of lexical relations in a lexicon*, [w:] *Lexical Functions in Lexicography and Natural Language Processing*, red. L. Wanner, Amsterdam, s. 37-102.

Mielczuk, I., 1998, *Collocations and Lexical Functions,* [w:] *Phraseology: Theory, analysis and applications*, red. A. Cowie, Oxford, s. 21-53.

Moon, R., 1998, *Fixed Expressions and Idioms in English. A Corpus-Based Approach*, Oxford.

Nattinger, J., 1980, *A lexical phrase-grammar for ESL*, [w:] "TESOL quarterly" 14 (3), s. 337-344.

Oakes, M., 1998, *Statistics and Corpus Linguistics*, Edynburg.

Pęzik, P., 2013, *Paradygmat dystrybucyjny w badaniach frazeologicznych. Powtarzalność, reprodukcja i idiomatyzacja,* [w:] *Metodologie Językoznawstwa. Ewolucja Języka, Ewolucja Teorii Językoznawczych*, red. P. Stalmaszczyk, Łódź, s. 143–160.

Piotrowski, T., Grabowski, Ł., 2013, *Interpretacja danych frekwencyjnych z korpusów językowych: opis pewnych problemów (na kilku przykładach z życia wziętych)*, [w:] *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*, red. W. Chlebda, Opole, s. 59-71.

Renouf, A., Sinclair, J., 1991, *Collocational frameworks in English*, [w:] *English corpus linguistics*, red. K. Aijmer, B. Altenberg, Nowy Jork, s. 128-143.

Roemer, U., 2009a, *English in Academia: Does Nativeness Matter?* [w:] "Anglistik: International Journal of English Studies" 20 (2), s. 89-100.

Roemer, U., 2009b, *The inseparability of lexis and grammar. Corpus linguistic perspectives*, [w:] "Annual Review of Cognitive Linguistics" 7, s. 141-163.

Roemer, U., Schulze, R. red, 2008, *Patterns, Meaningful Units and Specialized Discourses*, [w:] "International Journal of Corpus Linguistics" 13 (3). Numer specjalny, Amsterdam.

Roemer, U., 2010, *Establishing the phraseological profile of a text type. The construction of meaning in academic discourse*, [w:] "English Text Construction" 3 (1), s. 95-119.

Scott, M., 2008, *WordSmith Tools Help*, Liverpool. www.lexically.net/wordsmith/

Sinclair, J., 1987, *The nature of the evidence*, [w:] *Looking Up: An Account of the COBUILD Project in Lexical Computing*, red. J. Sinclair, Londyn, s. 150-159.

Sinclair, J., 1991, *Corpus, Concordance, Collocation*, Oxford.

Sinclair, J., 2001, *Review* [w:] "International Journal of Corpus Linguistics" 6 (2), s. 339-359.

Sinclair, J., 2004, *Trust the text: language, corpus and discourse*, Londyn.

Skorupka, S., 1989, *Słownik frazeologiczny języka polskiego*, t. 1-2, Warszawa.

Warren, M., 2010, *Determining Aboutgrams in Engineering Texts*, [w:] *Keyness in Text*, red. M. Bondi, M. Scott, Amsterdam, s. 113-126.

Wilks, Y., 2005, *REVEAL: The notion of anomalous texts in a very large corpus*, Referat przedstawiony na konferencji Tuscan Word Centre International Workshop. Certosa di Pontignano, Włochy, 30.06-3.07. 2005.

Амосова Н.Н., 1963, *Основы английской фразеологии,* Ленинград.

Виноградов В. В., 1947, *О основных типах фразеологических единиц в русском языке*, [w:] *Сборник статей и материалов*, ред. А. Шахматов, Москва, с. 339-364.

Виноградов В. В., 1977, *Об основных типах фразеологических единиц в русском языке*, [w:] *Виноградов В. В. Избранные труды. Лексикология и лексикография*, с. 140-161. http://www.philology.ru/linguistics2/vinogradov-77d.htm

Копотев М. В., 2008, *Принципы синтаксической идиоматизации*, Helsinki. http://www.doria.fi/bitstream/handle/10024/42928/principy.pdf