Piotr Paryzek

Pozyskiwanie danych leksykalnych z tekstów elektronicznych

(na materiale czasopisma naukowego)

Promotor: dr hab. Piotr Wierzchoń, prof.

Uniwersytet im. Adama Mickiewicza w Poznaniu
Instytut Językoznawstwa

Poznań 2011

Panu Profesorowi Piotrowi Wierzchoniowi
bardzo dziękuję za cierpliwość, trafne uwagi
i opiekę nad pracą

Spis treści

1 Wstęp

Niniejsza praca dotyczy wybranych metod pozyskiwania, czyli ekscerpcji, informacji o charakterze leksykalnym z elektronicznych zbiorów tekstów.

Jej celem jest, po pierwsze, sformułowanie nowych, oryginalnych metod, które mogą być użyteczne w pozyskiwaniu materiału do analiz leksykalnych, a następnie zbadanie ich na wybranym zbiorze tekstów.

Planowano opracowanie metod niewymagających zaawansowanej znajomości programowania komputerowego, a jednocześnie umożliwiających uzyskanie wartościowych wyników, gdzie za wartościowość metody uznaje się daną wydajność ekscerpcyjną. Trzy sformułowane metody dopracowano i zoptymalizowano.

W rozdziale 2 „Korpus i językoznawstwo korpusowe" przedstawiono zarys historii badań korpusowych oraz informacje dotyczące korpusów językowych (ich cechy, typologię, przykłady oraz narzędzia przydatne w badaniach nad korpusami).

W rozdziale 3 znajdują się informacje ogólne dotyczące wyrażeń regularnych, ich składni itp., w tym szczegółowy opis wyrażeń zastosowanych w niniejszej pracy.

W rozdziale 4, będącym częścią eksperymentalną pracy, opisano zbiór tekstów wykorzystany w badaniach oraz przedstawiono dokładnie – z podziałem na poszczególne kroki – kolejno trzy metody pozyskiwania informacji leksykalnych:

1. Metoda ekscerpcji neologizmów, czyli agnonimów słownikowych, które w tekście pracy określa się jako wyrazy nowe[1].

2. Metoda ekscerpcji kolokacji w oparciu o akronimy (metoda akronimowa).

3. Metoda ekscerpcji kolokacji rzeczownikowych.

W przypadku każdej metody dokonano omówienia uzyskanych wyników, w tym odniesiono się do nich krytycznie, a także oceniono możliwości rozbudowy każdej metody i ewentualnych dalszych badań.

Istotnym elementem pracy jest punkt 5. czwartego rozdziału, w którym podsumowano przebieg prac w przypadku trzech zaproponowanych i zastosowanych metod.

W rozdziale 5 odniesiono się do spełnienia celu pracy oraz zawarto wnioski ogólne dotyczące możliwości wykorzystania wyodrębnionych metod pozyskiwania danych leksykalnych w badaniach korpusowych.

Pracę zamyka streszczenie w języku angielskim oraz bibliografia.

Praca wpisuje się w nurt badań prowadzonych w obrębie językoznawstwa stosowanego. Oznacza to, że celem dociekań naukowych jest stworzenie takiej teorii, której głównym komponentem będzie komponent ekscerpcyjny, prowadzący do uzyskania wstępnie przewidywanego wyniku ekscerpcyjnego (w postaci np. kolokacji występujących w tekstach naukowych).

[1] Pojawiły się tu dwa terminy: neologizm i agnonim słownikowy. Neologizm to wyraz nowo powstały, zaś agnonim słownikowy to wyraz nieobecny w danym słowniku. Uściślenie to jest o tyle istotne, że pojęcie neologizmu jest relewantne lingwochronologizacyjnie, a pojęcie angnonimu słownikowego jest pojęciem lingwochronologizacyjnie neutralnym. W pracy, dla uproszczenia wywodu, stosuje się określenie jednostka nowa, mając w pamięci to, że faktycznie poszukujemy po prostu agnonimów słownikowych. Dopiero specjalnie dedykowana ku temu teoria lingwochronologizacyjna byłaby w stanie orzec o neologiczności względem danej granicy datacji agnonimów słownikowych.

2 Korpus i językoznawstwo korpusowe

Według Barbary Lewandowskiej-Tomaszczyk *językoznawstwo korpusowe*:

> jest jedną z części językoznawstwa komputerowego[2] i zajmuje się analizą języka zgromadzonego w korpusach językowych, czyli komputerowych zbiorach autentycznych tekstów językowych, mówionych i pisanych (Lewandowska-Tomaszczyk 2005: 11).

Zbioru tekstów nie można zatem uznać za korpus[3], jeśli nie można go zapisać w sposób elektroniczny i w ten sam sposób odczytywać. Inaczej mówiąc – w tym ujęciu zbiorów tekstów zgromadzonych przed wprowadzeniem do językoznawstwa komputerów nie należy określać mianem korpusów.

Jednym z pierwszych podręczników językoznawstwa korpusowego był Aarts i Meijs (1984), chociaż pojęcie to zostało użyte już wcześniej, na przykład w Aarts i van den Heuvel (1982). Ponadto w literaturze angielskojęzycznej stosuje się także (por. Taylor 2008, szczegółowa analiza występowania nazw dyscypliny w publikacjach naukowych) pojęcia *corpus/corpus-based/corpus-driven/corpus assisted* + *analysis/approach/study*, a także wyodrębnia *corpus-driven linguistics* (Tognini-Bonelli 2001).

Znacznie szersze ujęcie definicji korpusu przedstawił wcześniej Kennedy (1998: 1), według którego „korpus jest zbiorem tekstów pisanych lub

[2] Na temat językoznawstwa komputerowego por. także Grishman 1986, Boguraev 1995, Clark et al. 2010.

[3] Źródłosłów słowa „korpus" jest łaciński i oznacza ciało (pierwsze znaczenie w internetowym *Słowniku języka polskiego* Wydawnictwa Naukowego PWN opracowanego na bazie *Uniwersalnego słownika języka polskiego* pod red. S. Dubisza). Znaczenie językoznawcze umieszczono na ostatniej, dziewiątej pozycji.

transkrypcji wypowiedzi mówionych; na jego podstawie prowadzi się analizy językoznawcze i dokonuje opisu języka". I kontynuuje: „wprowadzenie komputerów do badań korpusowych nie zmieniło istoty tych badań, mimo że wielu badaczy tak sądzi, a w żadnym razie komputery nie wyznaczają początku badań korpusowych, mimo że znacznie je przyspieszyły i ułatwiły". Należy ponadto pamiętać, że mimo ogromnego postępu technologicznego w wielu aspektach badań korpusowych ingerencja człowieka jest nieodzowna (na przykład w weryfikacji analizy morfologicznej i składniowej). Podobnego zdania są McEnery i Wilson (1996: 1), którzy podają niewątpliwie mniej, a nawet zupełnie nienaukową, za to szerszą i naszym zdaniem zdroworozsądkową definicję językoznawstwa korpusowego jako „badanie języka w oparciu o rzeczywiste przykłady jego użycia". Por. także elementy językoznawstwa korpusowego w Biber et al. 1998.

Który z tych poglądów jest bardziej przekonywający? Wydaje się, że nieco sztuczne wyznaczenie początku językoznawstwa korpusowego i zrównanie go z pojawieniem się komputerów nie uwzględnia tego, że badaniach tych kluczowe jest postawienie odpowiedniej hipotezy naukowej i odpowiedź na pytanie: „co badać". A. Bogusławski i M. Danielewiczowa:

> W językoznawstwie, a właściwie w całej nauce z filozofią na czele, rzeczą absolutnie podstawową jest właściwe wyodrębnienie przedmiotu opisu. Chodzi bowiem o to, żeby opisywać to, co jest, nie zaś to, czego nie ma. Studium niebytu, choćby było nad wyraz uczone i rygorystyczne, a przy tym pięknie sformalizowane, jest z naukowego punktu widzenia zupełnie nieinteresujące (Bogusławski, Danielewiczowa 2005: 8).

Komputer jest nadal jedynie narzędziem, które umożliwia badanie i opisywanie tego, co jest. Technologia zatem nie wyznacza początku badań, a jedynie – i aż – jakościowo oraz ilościowo zmienia sposób ich prowadzenia. Ponadto niewątpliwie obecnie każdy korpus *jest* zbiorem komputerowym, a nawet korpusy starsze – początkowo dostępne jedynie w wersji papierowej –

doczekały się przetworzenia do formatu elektronicznego (na przykład korpus zgromadzony w ramach projektu *Survey of the English Language*).

Należy w tym miejscu odróżnić korpus rozumiany współcześnie od zbioru tekstów, określanego niekiedy mianem *archiwum* (Kennedy 1998). Drugie pojęcie ma szerszy zakres, ponieważ – w przeciwieństwie do korpusu – nie narzuca ograniczeń w doborze tekstów, ich proporcjach i strukturze, a często jego głównym zadaniem i powodem gromadzenia tekstów są nie badania naukowe, a dostęp do tekstów (czytanie). Korpus zaś jest zbiorem tekstów charakteryzującym się systematycznością, zgromadzonym ściśle według założonego planu (patrz rozdział 1, punkt 2), nie jest „zbiorem tekstów zgromadzonych w sposób losowy" (Aston, Burnard 1998: 21). Leech (1991: 11) stwierdza, że „różnica między archiwum a korpusem polega na tym, że korpus opracowuje się z myślą o reprezentatywności", rozumianej jako celowy, świadomy i zaplanowany dobór tekstów w zbiorze (rozwinięcie pojęcia przedstawiono w rozdziale 2, punkt 2.2).

Kennedy (1998: 11) wymienia szereg zastosowań zbiorów tekstów:

opracowywanie słowników, list słów i gramatyk deskryptywnych, badania synchroniczne i diachroniczne odmian języka, zagadnienia stylu, nauczanie języka ojczystego i obcego, badania statystyczne rozkładu fonemów, badania dotyczące interpunkcji, fleksji, derywacji, analiza kolokacji i innych elementów języka.

2.1 Historia badań korpusowych

Językoznawstwo korpusowe we współczesnym ujęciu, czyli wykorzystujące technologię cyfrową, pojawiło się w latach 50. ubiegłego wieku, zaś największy postęp tej dyscypliny łączy się z wprowadzeniem tzw. mikrokomputerów w latach siedemdziesiątych XX wieku oraz – w latach osiemdziesiątych – dysków optycznych. Od tego momentu wraz ze wzrostem mocy obliczeniowej i rozwojem oprogramowania – które umożliwiały szybkie przetwarzanie danych oraz ich szczegółową analizę – rozpoczął się duży rozwój w tej dziedzinie badań językoznawczych.

2.1.1 Krytycyzm Chomsky'ego

Tak jak w historii językoznawstwa upowszechnienie komputerów przypada na lata 70. ubiegłego wieku, tak w sferze postrzegania korpusu jako źródła wiedzy o języku istotną cezurę stanowią lata 50. Wtedy to – pod znacznym wpływem poglądów N. Chomsky'ego (1957: 1965) – językoznawstwo zaczęło odwracać się od empiryzmu na rzecz racjonalizmu, czyli od badań opartych na rzeczywistych przykładach użycia języka w stronę danych wygenerowanych przez badacza i jego własnych osądów, a więc kompetencji językowej (wewnętrznych umiejętności językowych) nie zaś performancji (jej uzewnętrznienia, czyli zastosowania języka, które nie zasługuje – przeciwnie do kompetencji – na dogłębne badania); por. Chomsky 1957, 1965. Stan ten mógł wynikać z przeświadczenia niektórych wczesnych przedstawicieli językoznawstwa korpusowego (m.in. Harris 1951), że język jest tzw. *bytem skończonym*, a zatem korpusy są jego obiektywnym i pełnym obrazem. Krytyka językoznawstwa korpusowego była tym łatwiejsza, że przy słabości metod numerycznych i technologii komputerowej bardziej wszechstronna i precyzyjna analiza korpusu była zadaniem niewykonalnym.

Swój jednoznaczny krytycyzm wobec korpusu Chomsky uzasadniał tym, że „korpus nigdy nie stanie się narzędziem przydatnym językoznawcy, ponieważ zadaniem badacza jest stwarzanie modeli kompetencji językowej, a nie użycia języka", które w niewielkim stopniu odzwierciedla kompetencję (McEnery, Wilson 1996). Ponadto korpus nie stanowi obiektywnego odzwierciedlenia rzeczywistości językowej – „jest skończony i w pewnym sensie przypadkowy" (Chomsky 1957: 15) – która ma charakter nieskończony („przypuszczalnie nieskończony"; Chomsky 1957: 15), zatem nie można na jego podstawie formułować wniosków ogólnojęzykowych, ponieważ „rozważania probabilistyczne nie mają żadnego związku z gramatyką" (Chomsky 1964: 215). Ilustracją tego zagadnienia jest podawany przez Chomsky'ego przykład: zdanie *I live in New York* występuje w języku częściej niż zdanie *I live in Dayton*, *Ohio* z tej prostej przyczyny, że Nowy Jork ma więcej mieszkańców. Chomsky stwierdza ponadto, że „każdy korpus naturalny jest wypaczeniem rzeczywistości. Niektóre zdania nie występują w nim, ponieważ są oczywiste, inne, ponieważ są nieprawdziwe, zaś jeszcze inne, ponieważ byłyby nieuprzejme" (Chomsky 1957: 159). Badacz twierdzi wreszcie, że tylko 5% wypowiedzi użytkowników języka to wypowiedzi gramatyczne, reszta zaś jest gramatycznie niepoprawna, a więc badanie performancji (realnych użyć) nie ma sensu (Chomsky 1965). Zamiast niereprezentatywnego i skończonego korpusu źródłem prawdy o języku miałby być sam językoznawca, jego kompetencja językowa i introspektywne obserwacje dzięki umysłowi pozwalające ogarnąć większą liczbę wypowiedzi (McEnery, Wilson 1996: 9).

Krytycyzm wobec językoznawstwa korpusowego wyrażał ponadto Abercrombie (1965), który – w przeciwieństwie do Chomsky'ego – skupiał się na kwestiach o charakterze praktycznym, takich jak pracochłonność i czasochłonność badań oraz ich podatność na błędy związana z zaangażowaniem wielu nieraz tysięcy osób opracowujących materiał językowy.

Jednoznaczny krytycyzm Chomsky'ego i podważanie użyteczności korpusów w ogóle doprowadziły do spadku zainteresowania badaniami korpusowymi, jednak nie spowodowały ich wygaśnięcia. Nadal prowadzono analizy w dziedzinie fonetyki, w której użyteczność danych rzeczywistych trudniej było podważyć, w badaniach akwizycji języka ojczystego przez dzieci, ponieważ w tym przypadku – z uwagi na wiek użytkownika języka – informacji introspektywnych nie można było uzyskać oraz w językoznawstwie historycznym, w przypadku którego także nie było możliwości skorzystania z wiedzy o kompetencji językowej. Ponadto w latach 60. XX wieku Quirk (1960) rozpoczął prace nad projektem *Survey of English Usage*, na którym oparto dzieło *A Comprehensive Grammar of The English Language* (Quirk et al. 1985).

2.1.2 Rozwój badań korpusowych

Poglądy krytyków językoznawstwa korpusowego (por. też Fillmore 1992, Horrocks 1987, Matthews 1981) doprowadziły do zarzucenia poglądu, że korpus jest obiektywną reprezentacją języka uznawanego za byt skończony (niektórzy badacze podkreślający użyteczność korpusów już wcześniej uznawali te zastrzeżenia, por. Hockett 1948). Czy jednak absolutny obiektywizm korpusu systemu języka jest warunkiem koniecznym przydatności korpusu do badań językoznawczych? Można argumentować, że w niektórych dziedzinach – na przykład w analizie częstotliwości występowania słów – korpus jest niezastąpiony, w innych zaś, jak zauważa Leech (1992), pozwala badać język – przy wszystkich ograniczeniach – w sposób bardziej systematyczny i naukowy niż za pomocą własnych, introspektywnych uogólnień, to one bowiem są według niego z zasady nieobiektywne (por. także Partington 1998 oraz Sampson 1992, stwierdzający, że frazy, które bada językoznawstwo kierujące się introspekcją, są odmienne od fraz występujących w korpusach, a więc naturalnie występujących w języku). Na przykład badanie kolokacji i częstotliwości ich występowania

nie jest możliwe introspektywnie, intuicyjnie – współczesne metody ich analizy wykorzystują osiągnięcia językoznawstwa korpusowego (Aston, Burnard 1998: 14).

Językoznawstwo często korzystało ze źródeł, na podstawie których formułowano teorie o funkcjach, charakterze, elementach składowych i budowie języka. Wcześniej, a także w ramach podejścia Chomsky'ego, źródłem uogólnień formułowanych przez językoznawców były intuicja i introspekcja oraz doświadczenia, a także mniej lub bardziej subiektywna obserwacja języka i jego elementów (Kennedy 1998: 7). W przypadku badań korpusowych podstawą jest sam tekst, ewentualnie wzbogacony o dodatkowe adnotacje (por. rozdział 2, punkt 3.6). Jest bowiem, w ujęciu Teuberta i Krishnamurthy (2007), podstawowa różnica między językoznawstwem korpusowym opartym na *parole* (a więc na tym, co wytworzone przez użytkowników języka) a językoznawstwem, którego przedstawicielem jest Chomsky, opartym na *langue*, a więc językoznawstwem introspektywnym.

Okres osłabienia językoznawstwa korpusowego trwał stosunkowo długo (por. też Johansson 1991). Dopiero rozwój technologii i oprogramowania komputerowego pozwolił na szybki rozwój korpusologii, który trwa do dziś (McEnery 1996: 17).

Przyrost badań korpusowych w drugiej połowie XX wieku po okresie stagnacji dobrze obrazuje tabela zamieszczona przez Johanssona (1991: 312):

Tabela 1. Liczba badań korpusowych w II połowie XX w. (Johansson 1991:312)

Lata	Liczba badań
do 1965	10
1966-1970	20
1971-1975	30
1976-1980	80
1981-1985	160
1986-1991	320

Sampson (2005) szczegółowo opisuje tendencję zwiększania udziału metod empirycznych w językoznawstwie, koniecznych, ponieważ intuicja często zawodzi badaczy (por. przykłady, Sampson 2005: 18–19). Według niego, jednak, tendencja ta dotyczy głównie językoznawstwa komputerowego, a ponadto – w ostatnich latach uległa zahamowaniu, a nawet odwróceniu.

Na podstawie analizy publikacji w czasopiśmie *Language* od roku 1960 do 2002, podzielonych na oparte na danych empirycznych, oparte na intuicji oraz nieokreślone, Sampson stwierdza, że – jakkolwiek w latach 60. do końca 80. XX wieku występował niewielki wzrost liczby publikacji z dziedziny badań korpusowych, dopiero w latach 90. nastąpiło „coś w rodzaju eksplozji" (2005: 16). Jak zauważa Gries "over the past few decades, corpus linguistics has become a major methodological paradigm in applied and theoretical linguistics" (2006: 191).

Wydaje nam się, choć zważywszy na treść niniejszej rozprawy trudno o opinię bezstronną, że wraz z dalszym rozwojem technik informatycznych znaczenie badań korpusowych – opartych przecież na realnych tekstach, których zróżnicowanie jest funkcją dociekliwości i staranności badacza, a nie na subiektywnej intuicji – będzie się utrzymywało, jakkolwiek poglądy krytyków, na przykład Chomsky'ego ("My judgment, if you like, is that we

learn more about language by following the standard method of the sciences. The standard method of the sciences is not to accumulate huge masses of unanalyzed data and to try to draw some generalization from them" (2004: 97)) się nie zmieniły.

Warto ponadto zauważyć, że według Leecha „computer corpus linguistics defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject" (Leech 1992:106), a więc stanowi z perspektywy filozoficznej odmienny punkt widzenia na badania językoznawcze.

2.2 Charakterystyka korpusów językowych

Sinclair (1996) rozwinął pierwotne założenia sformułowane przez H. Kučerę i W. N. Francisa (1967) przy tworzeniu pierwszego korpusu elektronicznego (założenia konstrukcji tego korpusu były następujące: wielkość – 1 milion słów[4], jeśli to tylko możliwe, równowaga różnych typów tekstów ze źródeł pisanych, 500 tekstów po 2000 słów każdy). Obecnie, rzecz jasna, byłyby to warunki o znaczeniu czysto historycznym, niedostosowane do postępu technologii komputerowej, a także – w odniesieniu do postulowanej obecności tekstów pisanych i braku mówionych – niewystarczające.

2.2.1 Sinclaira koncepcja konstruowania korpusów (1996)

Zasady konstruowania korpusów według J. Sinclaira dotyczą *wielkości*, *jakości*, *prostoty* oraz *udokumentowania* i przedstawiają się następująco:

Wielkość

Korpus powinien być maksymalnie obszerny, ograniczony właściwie wyłącznie aktualnym stanem technologii, zatem nie sposób określić właściwej, odpowiedniej i wystarczającej wielkości korpusu. Ponadto, w przypadku korpusów monitorujących (patrz część 2.3.1) nie mówi się o wielkości korpusu, a raczej o strumieniu danych stale go zasilającym. Sinclair (1991: 9) zauważa, że nawet korpus zawierający miliard wyrazów nie będzie dawał wystarczających informacji o kontekście użycia słów rzadkich lub specjalistycznych, ponieważ do celów statystycznych konieczny jest zawsze

[4] Słowo (wyraz) jest definiowane jako „łańcuch liter (znaków alfanumerycznych), ograniczonych spacją lub znakiem interpunkcyjnym" (Kennedy 1999: 206). Definicje słowa w językoznawstwie polonistycznym por. m. in. Bańko 2002, Wierzchoń 1998, Grochowski 1982, Kurkowska 1975, Zgółkowa, Bułczyńska 1987.

więcej niż jeden przykład użycia danego słowa. Według Kennedy'ego (1998: 67) w korpusie zawierającym ponad pięćdziesiąt milionów wyrazów 40 – 50% wyrazów występuje tylko raz, są to więc tak zwane *hapax legomena* (w języku fleksyjnym, takim jak język polski, odsetek wyrazów występujących tylko jeden raz jest większy). W przypadku badania ustalonych związków wyrazowych problem ten jest jeszcze poważniejszy, ponieważ badaniu podlegają nie pojedyncze wyrazy, ale ich sekwencje (co najmniej dwuskładnikowe), przy czym kolejność składników jest znacząca.

Pożądana wielkość korpusu zależy ponadto od rodzaju prowadzonych badań językoznawczych – w przypadku badania cech fonetycznych wystarczający powinien być korpus liczący około 100 tys. słów (Kennedy 1998: 68), zaś do analizy składniowej wystarczać ma korpus o wielkości około miliona słów. Zawsze jednak należy pamiętać, że „im dana cecha podlegająca badaniu występuje rzadziej, tym większy powinien być korpus" (McEnery, Wilson 1997: 154).

Czy z kolei korpus może być zbyt duży? Teoretycznie tak, jeśli wskutek ograniczonych zasobów technologicznych jego przetwarzanie bądź analiza stają się niewykonalne, jednak przy uwzględnieniu możliwości dostępnych badaczom komputerów zagadnienia sprzętowe nie stanowią istotnego ograniczenia. W niniejszych badaniach wykorzystano komputer z procesorem 1,8 MHz, 1280 MB pamięci operacyjnej i dyskiem twardym o pojemności 30 GB.

Pojawia się tu problem nie tylko całkowitej liczby wyrazów w korpusie, ale liczby i rodzaju jego składników (tekstów bądź ich fragmentów) – a w nim echa zastrzeżeń Chomsky'ego. Sugerował on bowiem, aby rodzaje tekstów, które docierają do większej liczby odbiorców (np. gazety i audycje emitowane przez rozgłośnie ogólnokrajowe), były szerzej reprezentowane w porównaniu do tych, które docierają do stosunkowo niewielu (np. zapisy rozmów, wykładów itp.).

Istotny jest ponadto rozmiar pojedynczych tekstów (bądź ich części, czyli próbek) składających się na korpus oraz liczba tekstów z poszczególnych kategorii. Według Oostdijk (1988) teksty w korpusie powinny mieć wielkość około 20 tys. słów, zaś Biber (1999) opowiada się za mniejszym rozmiarem – około 2 do 5 tys. słów.

Jakość

Autentyzm tekstów wpływa na jakość korpusu, muszą one zatem odzwierciedlać naturalny i niezakłócony tok porozumiewania się, chyba że korpus ma celowo obejmować wypowiedzi zakłócone, na przykład wtrącenia. Wynika z tego, że ewentualna ingerencja językoznawcy (taka jak poprawianie błędów typograficznych oraz wszelkie globalne operacje wykonywane na tekście mogące naruszyć jego oryginalność, na przykład wprowadzanie spacji między słowami a oznaczeniami odnośników literaturowych, aby wyeliminować ciągi znaków literowych, po których bezpośrednio następuje cyfra) powinna być zawsze ograniczona do minimum, ponieważ może prowadzić do zarzutu manipulacji materiałem źródłowym.

W przypadku zapisów wypowiedzi ustnych jakość może dotyczyć dokładności transkrypcji, będącej etapem nieobiektywnym (zależnym od osoby dokonującej transkrypcji lub oprogramowania do transkrypcji mowy).

Jednoznaczność

Zasoby językowe powinny być dostępne w formacie tekstowym (obecnie coraz częściej stosuje się format SGML lub XML), zaś wszystkie znaczniki muszą być jawnie wyróżnione na tle tekstu zasadniczego, a także powinna istnieć możliwość ich oddzielenia, na przykład znakami przyjętymi w stosowanej konwencji. Dotyczy to w szczególności korpusów anotowanych, w których informacje dodatkowe najlepiej wyróżniać zgodnie z pewną konwencją (na przykład zgodnie ze standardem TEI, patrz rozdział 2, punkt 4.2). Ponadto, zbyt duża ilość informacji dodatkowych utrudnia wyodrębnienie

i wzrokową analizę tekstu zasadniczego (tekst ten jest ukryty wśród adnotacji), o ile oczywiście brak jest funkcji ukrywania żądanych parametrów informacji.

Udokumentowanie

Każdy element korpusu powinien być właściwie udokumentowany danymi określającymi pochodzenie tekstu, pozwalającymi go jednoznacznie zidentyfikować. Informacje takie (np. nagłówki formatu SGML rozpoczynające się znacznikiem <HEADER> mogą obejmować informacje o nazwie pliku źródłowego, autorze, roku powstania tekstu i tytule) są oddzielone od tekstu zasadniczego.

Na przykład w Korpusie Języka Polskiego Wydawnictwa Naukowego PWN zadanie w polu wyszukiwania słowa *neuron* daje wynik następujący:

> postęp w dziedzinie mikrominiaturyzacji sprzętu elektronicznego. Półprzewodnikowe obwody scalone stanowią w obecnej chwili podstawowy przedmiot zainteresowania mikroelektroniki. Przypuszcza się, że właśnie ta technika umożliwi w przyszłości wytwarzanie niezawodnych, szybko działających układów elektronicznych o gęstości upakowanie elementów zbliżonej do gęstości upakowania neuronów w mózgu ludzkim.

W nagłówku znajdują się następujące informacje:

Typ tekstu: Prasa
Tytuł: Młody Technik
Nr: 3
Miejsce wydania: Warszawa
Rok: 1971

2.2.2 Reprezentatywność

Do wspomnianych parametrów warto dodać jeszcze *reprezentatywność*. Według Lecha (1991) reprezentatywność korpusu polega na tym, że

obserwacje i wnioski wywiedzione na podstawie korpusu można rozciągnąć na język we wszystkich jego przejawach bądź na konkretny wycinek rzeczywistości językowej (tj. ze względu na parametry: socjolingwistyczne, stylistyczne, geograficzne itd.). Według Summers (1991) osiągnięcie reprezentatywności może być rozumiane na wiele sposobów, takich jak: elitarność tekstów (czyli ich wartość naukowa bądź literacka), dobór przypadkowy (a zatem reprezentatywność probabilistyczna), cyrkulacja (obieg i dostępność tekstu dla odbiorców, co faworyzuje publikacje wysokonakładowe), typowość tekstu (parametr w znacznym stopniu subiektywny), dostępność tekstów, dostosowanie do charakteru czytelnictwa w danej populacji i empiryczne dostosowanie metody doboru tekstów do ustalonych wymagań językoznawczych (w kwestii doboru tekstów por. Hunston 2002). Widać więc, że – być może z wyjątkiem doboru przypadkowego, który w teorii byłby reprezentatywny, jednak przypadkowość determinowana jest warunkami rozkładu losowego, wstępnymi założeniami dotyczącymi jakości próby itd. – pojęcie reprezentatywności jest w dużej mierze konstruktem teoretycznym i intuicyjnym, choć niewątpliwie podstawowym w konstruowaniu korpusów, na podstawie których mają być formułowane uogólnienia dotyczące całości języka. Należy jednak pamiętać, że każde takie uogólnienie ma charakter ekstrapolacji, ponieważ „każde sformułowanie dotyczące informacji występującej w korpusie odnosi się jedynie to samego korpusu, a nie języka lub rejestru, wobec którego korpus jest jedynie próbką" (Hunston 202: 23), z kolei „korpus, niezależnie od wielkości, jest reprezentatywny jedynie wobec siebie samego" (Partington 1998: 146).

Oliva i Květoň (2002) określają reprezentatywność w sposób bardziej sparametryzowany jako:

– reprezentatywność jakościową (ang. *qualitative representativity*), która oznacza, że każdy poprawny w danym języku bigram (lub szerzej – element będący przedmiotem badania) występuje w korpusie

(reprezentatywność pozytywna) oraz że w korpusie tym nie występuje żaden bigram niepoprawny w danym języku (reprezentatywność negatywna);

– reprezentatywność ilościową (ang. *quantitative representativity*), która oznacza, że stosunek liczby wystąpień danego bigramu w korpusie do liczby w „pełnym zbiorze wypowiedzi w języku" odpowiada stosunkowi liczby wystąpień wszystkich bigramów w korpusie do liczby w „pełnym zbiorze wypowiedzi w języku".

Zbiór tekstów obciążony w mniejszym lub większym stopniu przypadkowością (to znaczy taki, że kryteria doboru tekstów nie są ustalane przed rozpoczęciem gromadzenia tekstów) uwarunkowany dostępnością pewnego rodzaju tekstów określa się jako korpus oportunistyczny (Lewandowska-Tomaszczyk 2005: 29), zwany także kanibalistycznym[5]. Mimo że obserwacje poczynione na korpusie tego typu trudniej ekstrapolować na system języka, mogą one stanowić cenne źródło informacji o poszczególnych, odpowiednio wąskich (to znaczy zależnych od charakteru tekstów składających się na zbiór) jego wycinkach (na przykład, jeśli gromadzimy teksty publicystyczne z danego tytułu prasowego, uzyskujemy informacje przede wszystkim o języku używanym w publicystyce). Podstawową bowiem kwestią jest tu rozróżnienie między badaniami jakościowymi i ilościowymi (McEnery 1996: 62). W badaniach, w których zagadnienia ilościowe (np. częstotliwość występowania słów) nie mają znaczenia, uwzględniane – opisywane, analizowane – jest każde, także pojedyncze wystąpienie szukanego elementu (wyrazu, kolokacji, terminu itd.). Jeśli z kolei celem są uogólnienia (szczególnie istotne statystycznie o charakterze ilościowym, na przykład modele statystyczne wyjaśniające dane obserwacyjne) dotyczące całego języka (lub ewentualnie jego ściśle określonego wycinka), niezbędne jest dążenie do możliwie największej reprezentatywności w odniesieniu do badanego obszaru. Nie opisuje się tu pojedynczych form lub zjawisk językowych, ale formy lub

[5] Należy w tym miejscu wspomnieć o tzw. korpusie wirtualnym tworzonym w oparciu o dane dostępne w Internecie (Teubert, Čermáková 2004).

zjawiska występujące z odpowiednio dużą częstotliwością, wystarczającą na przykład do przeprowadzenia analizy statystycznej.

Schmid (1993) stwierdza, że w językoznawstwie korpusowym najlepiej stosować w jednym projekcie wiele metod, dzięki czemu można w pełni wykorzystać zalety metod ilościowych i jakościowych. W ten sposób badanie korpusu lub zbioru tekstów ma charakter możliwie najbardziej wszechstronny.

2.2.3 Koncepcja McEnery'ego i Wilsona (1996)

Według McEnery'ego i Wilsona korpus (1996: 21) może być rozumiany jako niemal dowolny zbiór tekstów; we współczesnym ujęciu, jednak, trzeba uwzględnić cztery podstawowe zagadnienia dotyczące właściwości korpusu: *gromadzenie tekstów i reprezentatywność*, *skończona wielkość*, *możliwość automatycznego przetwarzania* oraz *dostępność*.

Gromadzenie tekstów i reprezentatywność

Ponieważ celem konstruowania korpusu jest najczęściej zgromadzenie tekstów możliwie najbardziej różnorodnych (a nie na przykład tekstów jednego autora, takiego jak zbiór wszystkich dzieł Williama Shakespeare'a, patrz http://shakespeare.mit.edu) – nawet jeśli chodzi o określoną, ograniczoną dziedzinę – konieczne jest dokonanie wyboru, czyli zgromadzenie próbek tekstów. Należy przy tym dążyć do zebrania tekstów możliwie reprezentatywnych dla obszaru objętego badaniami, czyli w najlepszy możliwy sposób odzwierciedlających średnią dla hipotetycznego zbioru obejmującego wszystkie teksty. Przyjęte reguły doboru tekstów w danym określa się jako *sampling frame* (McEnery 2003: 449).

Skończona wielkość

Jakkolwiek znaczna liczba korpusów to zbiory zamknięte, czyli takie, których gromadzenie zakończyło się, istnieją korpusy monitorujące, które

powstają przez czas dłuższy (patrz rozdział 2, punkt 3.1), co umożliwia ich dostosowanie do ciągłego rozwoju języka. Zazwyczaj jednak już przed rozpoczęciem opracowywania korpusu należy określić sposób gromadzenia tekstów, liczbę tekstów oraz liczbę wyrazów w tekście (próbce tekstu).

Możliwość automatycznego przetwarzania

Mimo że automatyczne przetwarzanie nie jest warunkiem *sine qua non* dla korpusu, przy obecnym jednak stanie rozwoju korpusologii i techniki pojęcie korpusu jest równoważne *korpusowi elektronicznemu*. Jednym z ostatnich korpusów dostępnych w postaci drukowanej jest *A Corpus of English Conversation* (Svartvik, Quirk 1980), który przy obecnym stanie rozwoju językoznawstwa korpusowego stanowi swoistą ciekawostkę, mimo że wersja elektroniczna jest także dostępna (jako *London-Lund Corpus*, Svartvik 1990).

Dostępność

Ponownie jak w przypadku poprzednich kryteriów dostępność nie jest warunkiem niezbędnym dla korpusu, powinna jednak istnieć wśród osób opracowujących korpusy i nimi dysponujących tendencja do udostępniania korpusów szerszej grupie badaczy, co dzięki identyczności danych wejściowych (ten sam korpus) umożliwia porównywanie wyników osiągniętych w różnych badaniach i optymalizację stosowanych metod.

Nierzadko dostępna bezpłatnie jest jedynie część korpusu, zaś wersja pełna udostępniana jest odpłatnie (np. Korpus Języka Polskiego Wydawnictwa Naukowego PWN) lub też bezpłatnie udostępniane są jedynie narzędzia korpusowe dające ograniczone wyniki (np. *Bank of English*, który udostępnia wersję demonstracyjną narzędzia do sporządzania konkordancji i wyszukiwania kolokacji).

Ważne jest zatem, aby konstruowanie korpusu odbywało się w sposób możliwie ścisły, tj. uświadomiony teoretycznie, i zaplanowany – dostosowany

do tego, jakie badania na jego podstawie mają być prowadzone. Przede wszystkim chodzi tu o rozróżnienie badań ilościowych, które wymagają właściwych proporcji względnych między elementami korpusu (rodzaj tekstów, ich objętość, dobór), i jakościowych, we wszystkich jednak wypadkach materiał źródłowy powinien przede wszystkim mieć odpowiednią jakość (por. rozdział 2, podrozdział 2.1), aby nie fałszował wyników badań.

2.3 Typologia korpusów

Ponieważ w każdej dziedzinie badań naukowych istnieje niezmiennie potrzeba skategoryzowania elementów, które dziedzina ta obejmuje, także i w językoznawstwie korpusowym wyróżnia się szereg typów korpusów, różniących się przydatnością w konkretnych rodzajach badań. Klasyfikację (Lewandowska-Tomaszczyk 2005: 29) wprowadza się ze względu na *sposób gromadzenia materiału*, *zawartość danych korpusowych*, *reprezentatywność*, *format*, *organizację danych* oraz *cel badawczy*.

2.3.1 Korpusy referencyjne i monitorujące

Korpusy referencyjne (inaczej: *ogólne*, *statyczne*) to zbiory tekstów będące niejako „fotografią" stanu języka lub jego fragmentu w danym punkcie jego rozwoju[6]. Ponieważ są korpusami, będącymi podstawą odniesienia, powinny charakteryzować się możliwie największą reprezentatywnością, a co za tym idzie – dążeniem do obiektywności. W szczególności wymaga to ścisłego i dokładnie zaplanowanego zrównoważenia proporcji między tekstami pisanymi (które, co oczywiste, łatwiej jest gromadzić, ponieważ obecnie – dzięki rozwojowi Internetu i oprogramowania do odczytu dokumentów skanowanych – są bardziej dostępne) a mówionymi, a w ich obrębie – równowagi między poszczególnymi gatunkami oraz między tekstami o charakterze oficjalnym a prywatnymi (stąd – korpus zrównoważony). Po zgromadzeniu odpowiedniej ilości materiału obejmującego możliwie krótki okres korpus zamyka się i udostępnia do badań. Pierwsze korpusy elektroniczne miały taki właśnie charakter (patrz punkt 2.4.2).

[6] Ponieważ korpusy takie zawierają różne rodzaje tekstów, w tym zazwyczaj mówione i pisane, Aston i Burnard (1998) określają je mianem korpusów mieszanych, pojawia się także pojęcie korpusu uniwersalnego (Meyer 2002).

Korpusy monitorujące (inaczej: *dynamiczne*, *otwarte*, por. Kennedy 1998), w których dane zbiera się w sposób ciągły, stanowią więc już nie opis języka w danym punkcie czasu, jak korpusy referencyjne, a serię ujęć stanu języka. Korpusy takie pozwalają analizować tendencje zmian w języku w ciągu okresu, z którego teksty pochodzą, a więc wykorzystuje się w tym wypadku podejście diachroniczne do języka. Przy odpowiednio dużej ilości danych umożliwia to zbadanie stanu języka w dowolnie wybranym punkcie w czasie. Jedyne ograniczenie rozmiaru stanowią tu zasoby finansowe i technologiczne (Sinclair 1991). Ponadto, ponieważ gromadzenie danych do takich korpusów jest zazwyczaj mniej usystematyzowane (to znaczy w większym stopniu gromadzi się teksty w danej chwili dostępne i powstające), korpusy monitorujące są mniej reprezentatywne i mają charakter bardziej oportunistyczny. Ze względu na konieczność zaangażowania zasobów technologicznych potrzebnych do zgromadzenia danych, ich zapisania i przetworzenia, objęte są programami typowo komercyjnymi, a więc ich dostępność do badań naukowych jest ograniczona lub obciążona znacznymi kosztami (por. Kennedy 1998: 61). Ich przydatność do celów badawczych polega na ogromnej ilości zebranych danych i możliwości obserwowania nieustannie zmieniającego się języka.

2.3.2 Korpusy ogólne i specjalistyczne

Korpusy ogólne mają odzwierciedlać język niezawierający wypowiedzi o charakterze specjalistycznym bądź dialektowym, a więc stanowią zbiór tekstów o charakterze codziennym, popularnym. Służą do formułowania generalnych wniosków, na przykład o stosowanym przez użytkowników słownictwie i gramatyce.

Korpusy specjalistyczne[7] to takie, w których gromadzi się teksty stanowiące pewien wycinek systemu języka ograniczony do pewnej grupy jego użytkowników i konkretnych sytuacji (np. język dokumentów patentowych, język umów cywilno-prawnych, język rzemiosła, język blogów internetowych[8], subkultur młodzieżowych, osób uczących się danego języka[9] itd.). Przykładem takiego korpusu jest zbiór nagrań i ich transkrypcji, obrazujący rozwój kompetencji językowych u dzieci (Carterette, Jones 1974), *Michigan Corpus of Academic Spoken English* (MICASE) (Poos, Simpson 2002) oraz korpus języka angielskiego używanego w przemyśle naftowym (Zhu 1989).

Tego typu zbiorem jest ponadto korpus wykorzystywany w badaniach przedstawionych w niniejszej pracy, czyli kompletny zbiór tekstów opublikowanych w czasopiśmie naukowym zajmującym się naukami przyrodniczymi i medycznymi z określonego okresu.

2.3.3 Korpusy pełnotekstowe i próbkowane

Korpusy pełnotekstowe są rodzajem korpusów, w których gromadzi się teksty w całości, co jest podejściem typowym i najprostszym w konstruowaniu korpusów.

Korpusy próbkowane – zawierają fragmenty tekstów o określonej długości. Takie podejście może być podyktowane dążeniem do osiągnięcia maksymalnej reprezentatywności, aby odzwierciedlić system języka możliwie najpełniej, bądź ograniczeniem docelowej wielkości korpusu ze względów technologicznych. Należy pamiętać, że w badaniach stylistyki lub dyskursu ograniczony rozmiar próbek może nie wystarczyć do sformułowania jednoznacznych tez. Summers (1991: 5) stwierdza, że „główną zasadą przy

[7] Meyer (2002) wprowadza pojęcie korpusu do celów specjalnych (w odróżnieniu od korpusu uniwersalnego).
[8] Jung 2007.
[9] Xunfeng, Kawecki 2001.

ustalaniu sposobu próbkowania tekstów jest szeroka gama rodzajów tekstów zdefiniowana w sposób obiektywny". Pierwszy korpus elektroniczny Kučery i Francisa (1967) miał taki właśnie charakter.

2.3.4 Korpusy języka pisanego i mówionego

Nie ulega wątpliwości, że korpus, który ma charakteryzować się możliwie największą reprezentatywnością, musi zawierać teksty pisane i mówione. Problemem jest jednak ustalenie ich proporcji. Wydaje się oczywiste, że język mówiony jest wykorzystywany przez typowego użytkownika częściej niż pisany. Z drugiej jednak strony tekst pisany może być rozpowszechniany w ogromnym nakładzie, zaś pojedyncza rozmowa – rozgrywać się wyłącznie między dwojgiem ludzi, co oznacza, że tekst w pierwszym przypadku oddziałuje na wiele osób, w przeciwieństwie do sytuacji drugiej.

Korpusy języka pisanego – dzięki dostępności wielu rodzajów tekstów w sieci WWW (teksty publicystyczne, naukowe, literackie, blogi internetowe itd.), a także coraz większym możliwościom oprogramowania odczytującego teksty pisane (OCR[10]) zbiory tekstów pisanych gromadzi się łatwo (niekiedy łatwość gromadzenia niektórych typów tekstów może negatywnie rzutować na reprezentatywność, łatwiej bowiem na przykład zgromadzić teksty z gazet bądź czasopism niż korespondencję prywatną), choć rzecz jasna nie odzwierciedlają one systemu języka. Ponadto korpusy te były chronologicznie pierwsze – korpusy języka mówionego pojawiły się w językoznawstwie później.

Korpusy języka mówionego – stanowią niezbędny element korpusów referencyjnych. Niektóre takie korpusy (na przykład *Survey of English Dialects*) istniały przez pewien czas jedynie w postaci nagrań, co stanowi odstępstwo od wymogu zapisania korpusu. Gromadzenie korpusów języka

[10] Por. Hussmann, Deng 2005, Taghva, Stofsky 2001, Nakano 2000.

mówionego, a więc – nagrywanie użytkowników w sytuacjach najbardziej zbliżonych do naturalnych (oznacza to na przykład, że nagrywanie odczytów lub sztuk teatralnych nie jest rozwiązaniem optymalnym, ponieważ – mimo że mówione – teksty odzwierciedlają słowo pisane), a następnie transkrypcja i ewentualna adnotacja nagrań, jest procesem niezwykle kosztownym i długotrwałym. Obecnie – wraz z rozwojem technologii komputerowej – dostępne są także specjalistyczne korpusy języka mówionego (patrz Lamel et al. 1991, Svartvik, Quirk 1980, Knowles et al. 1992, Otwinowska-Kasztelanic 2000 w sprawie korpusu języka mówionego młodego pokolenia Polaków, Meyer 2002 w sprawie ograniczeń korpusów języka mówionego, Clopper et al. 2006).

2.3.5 Korpusy jednojęzyczne i wielojęzyczne, równoległe i porównywalne

Korpusy jednojęzyczne – gromadzą teksty w jednym języku (na przykład korpusy narodowe: amerykański, brytyjski, czeski, grecki, rosyjski (Sharoff 2003), słowacki, a także *Brown Corpus*, *The Helsinki Corpus of English Texts*).

Korpusy wielojęzyczne (na przykład korpus *Hansard*, korpus *EuroParl*[11], korpus *JRC-Acquis*[12]) – zawierają teksty w więcej niż jednym języku, przy czym musi istnieć między tekstami pewien związek (o typach związku por. niżej), a więc wyklucza się przypadkowość.

Korpusy wielojęzyczne równoległe (inaczej: paralelne) – zawierają teksty oryginalne i ich przekłady na język lub języki obce. Korpusy takie mogą być uporządkowane względem zdań – to znaczy określa się, które zdania w tekście oryginalnym są ekwiwalentami zdań w tłumaczeniu – lub pojedynczych wyrazów. Im większa jednak dokładność takiego

[11] http://www.statmt.org/europarl
[12] http://langtech.jrc.it/JRC-Acquis.html

uporządkowania, tym większa liczba przypadków, w których ekwiwalencja (zdania lub wyrazu) nie jest typu „podmiany 1:1" (Lewandowska-Tomaszczyk 2005: 85). Jednym z przykładów jest tu korpus CRATER (McEnery et al. 1997). Innym przykładem takiego korpusu obejmującego szereg języków europejskich jest baza aktów prawnych Wspólnot Europejskich Eurlex (http://eur-lex.europa.eu/pl/index.htm), zawierająca ogromną liczbę dyrektyw, rozporządzeń i projektów tłumaczonych na języki wszystkich, obecnie 27, Państw Członkowskich. Korpusy takie należy traktować z pewną ostrożnością, ponieważ zazwyczaj tłumaczenie wymusza zmniejszenie naturalności tekstu wynikowego. Ponadto teksty takie narażone są na wszelkie typowe błędy tłumaczeniowe wywołane na przykład interferencją. Przykładem jest tu przenoszenie swobodnego szyku wyrazów z języka polskiego do języka angielskiego, w którym szyk wyrazów ma charakter bardziej stały; z kolei przy tłumaczeniu w odwrotnym kierunku następuje stosowanie zapożyczeń w miejsce istniejących wyrazów rodzimych; na przykład *ogrzewanie pod refluksem* (ang. *under reflux*) zamiast *pod chłodnicą zwrotną* (por. Dzierżanowska 1990).

Korpusy wielojęzyczne porównywalne – zawierają teksty podobne, jednak nie identyczne treściowo, które łączyć może tematyka, data powstania, autor itd., konstruowane za pomocą tych samych reguł (ang. *sampling frame*). Porównywalność tekstów w dwóch językach może wiązać się z następującymi kategoriami (Lewandowska-Tomaszczyk 2005: 52): tematyka (identyczna lub podobna dziedzina, ewentualnie podział na poddziedziny w ramach poszczególnych dziedzin), okres powstania (podobny dla tekstów w obu językach, znaczenie ma tu rzecz jasna szybkość rozwoju danej tematyki), styl (podobna charakterystyka stylistyczna), typ publikacji (teksty książkowe, prasowe itd.), środek przekazu (jeśli na przykład w jednym języku gromadzi się teksty mówione, w języku drugim powinny także występować teksty tego rodzaju). Zaletą tego typu korpusów jest naturalność tekstów, to znaczy brak interferencji związanej z tym, że tekst powstaje jako tłumaczenie z języka obcego. Interferencja może dotyczyć wszystkich aspektów tekstu (leksyka, składnia, stylistyka itd.). Przykładem korpusów zawierających teksty napisane

w jednym języku, ale porównywalnych ze sobą, są *Brown Corpus*[13] (odmiana amerykańska języka angielskiego) oraz wzorowane na nim (to znaczy opracowywane zgodnie z tymi samymi zasadami) *LOB Corpus*[14] (odmiana brytyjska) i *Kolhapur Corpus*[15] (odmiana indyjska). Z kolei typowymi korpusami porównywalnymi są *Aarhus Corpus of Contract Law* (języki duński, angielski, francuski) oraz projekt PAROLE[16] (języki państw członkowskich Unii Europejskiej).

2.3.6 Korpusy nieanotowane i anotowane

Korpusy nieanotowane (nieindeksowane) zawierają wyłącznie tekst, co oznacza, że teksty po zgromadzeniu nie zostały wzbogacone o żadne informacje dodatkowe. Korpus nietagowany ma ograniczone zastosowanie do uzyskiwania miarodajnych list frekwencyjnych, ponieważ te same wyrazy mogą pełnić różne funkcje gramatyczne (Biber et al. 1998: 30).

Korpusy anotowane zawierające dodatkowe informacje o tekstach:

a) dane o formatowaniu zastosowanym w dokumencie, czyli użyte czcionki (ważne w analizie dyskursu, na przykład znaczenie dyskursywne kursywy i wytłuszczenia), podział na strony i akapity,

b) dane o samym tekście – autor, jego płeć (istotna przede wszystkim w przypadku korpusów języka mówionego), pochodzenie i wykształcenie, rok powstania lub nagrania, gatunek,

[13] Francis, Kucera 1964.
[14] Johansson et al. 1986.
[15] Shastri 1985, Shastri 1988.
[16] Calzolari et al. 1996.

c) dane metajęzykowe, do których zalicza się etykiety części mowy, części zdania (typ fraz), cechy prozodyczne (informacje tego rodzaju podawał już *Survey of English Usage* z lat 60. XX wieku, Kaye 1988) lub transkrypcję w przypadku tekstów mówionych, dane semantyczne i leksykalne.

Dzięki anotowaniu łatwiejsze staje się wykorzystywanie informacji zgromadzonych w korpusie, nawet przez osoby nieznające danego języka, możliwe jest prowadzenie wielokrotnych analiz korpusu, a ponadto zwiększa się ich przejrzystość (McEnery 2003: 454–455).

Wymagania dla operacji anotowania podaje Leech (1993):

a) powinna istnieć możliwość łatwego oddzielenia adnotacji od treści korpusu, co również oznacza, że należałoby unikać wstawiania adnotacji wewnątrz wyrazów korpusu (drugi warunek ma znaczenie tylko dla odczytywania korpusu, ponieważ za pomocą narzędzi komputerowych właściwie oznaczone adnotacje łatwo jest oddzielić od treści),

b) powinna istnieć możliwość wyodrębnienia adnotacji z tekstu korpusu, co pozwala na ich oddzielną analizę,

c) wszystkie symbole zastosowane w adnotacjach należy dokładnie opisać, aby użytkownik korpusu miał pełną jasność co do ich znaczenia; pomocne jest ponadto dobranie symboli w sposób ułatwiający ich interpretację (na przykład oznaczenie przyimka symbolem PREP lub podobnym zamiast kodu liczbowego), a zatem w sposób możliwie intuicyjny,

d) należy podać zastosowaną metodę anotowania korpusu oraz podać informacje o ewentualnych poprawkach wprowadzanych ręcznie,

e) konwencje zastosowane przy adnotowaniu powinny być w możliwie największym stopniu zgodne z rozpowszechnionymi w praktyce badawczej zasadami, co do których istnieje możliwie największa zgoda.

Krytycy korpusów anotowanych twierdzą, że dodatkowe informacje narzucają jedną, określoną interpretację korpusu (można jednak rzecz jasna zawsze zaproponować interpretację własną, odmienną od oryginalnej). Ponadto niedokładności procesu anotowania, a więc konieczność anotowania manualnego, wprowadzają niespójność adnotacji (por. McEnery 2003: 456–457, Sinclair 1992).

2.3.7 Korpusy synchroniczne i diachroniczne

Korpusy synchroniczne – gromadzone w możliwie krótkim przedziale czasu, przez co mogą być uznawane za obraz stanu języka w danym punkcie jego rozwoju.

Korpusy diachroniczne, określane także jako korpusy historyczne – zawierają teksty powstałe na przestrzeni dłuższego czasu. Wykorzystywane są w badaniach leksykograficznych i analizie dyskursu, a także badaniach dialektologicznych i stylistycznych (Biber et al. 1998: 204). Przykładem jest tu *The Helsinki Corpus of English Texts* (patrz rozdział 2, punkt 4.2) oraz *Complete Corpus of Old English*.

2.4 Korpusy historyczne i współczesne

Wraz z postępem w badaniach w dziedzinie językoznawstwa korpusowego zwiększa się także liczba zbiorów źródłowych, zarówno prywatnych, jak i dostępnych dla szerszej liczby badaczy poprzez Internet (zwykle darmowe są jedynie wersje demonstracyjne, wersja pełna jest płatna – na przykład *Bank of English*) lub na nośnikach optycznych. Korzystać można ze zbiorów interesujących poznawczo lub egzotycznych, np. wszystkich przemówień inauguracyjnych prezydentów Stanów Zjednoczonych, ogromnego korpusu mannheimskiego (kilkaset milionów wyrazów, http://www.ids-mannheim.de/kl/corpora.html) oraz korpusu języka staro- i średniofrancuskiego; więcej przykładów na stronie http://www.athel.com/corpus.html).

2.4.1 Korpusy ery przedkomputerowej

Przed wprowadzeniem do korpusologii komputerów, co nastąpiło w początkach lat 60. XX wieku, badania korpusowe prowadzono w szeregu obszarów (Kennedy 1998: 13):

1. **Teksty biblijne i literackie** (w tym konkordancje Biblii Crudena, patrz rozdział 2, punkt 5.6)

2. **Ortografia**

Kaeding (1897) wykorzystał ogromny jak na ówczesne standardy, złożony z około 11 milionów wyrazów korpus do określenia częstotliwości występowania liter i ciągów liter (przedrostków, tematów i przyrostków) w języku niemieckim. Studium miało służyć usprawnieniu systemu notacji

stenograficznej, stąd litery i ich ciągi odpowiadały znakom stenograficznym. W pracę tę zaangażowano około 5000 ochotników.

3. **Leksykografia**

Jedna z przełomowych prac w historii leksykografii to wydany w 1755 r. *Słownik języka angielskiego* Samuela Johnsona. W przeciwieństwie do publikacji wcześniejszych Johnson – zgodnie z założonym wcześniej planem (Johnson 1747) – zgromadził ponad 150 tysięcy cytatów z tekstów literackich i na ich przykładzie opisywał znaczenie oraz użycie 40 tysięcy wyrazów hasłowych (por. Reddick 1990, Hitchings 2005, Hanks 2005, Landau 2005). Ponadto badania korpusowe wykorzystywano do studiów nad dialektami, czego wynikiem były *The Existing Phonology of English Dialects* (Ellis 1889) oraz *English Dialect Dictionary* (Wright 1905). Podobnie *Oxford English Dictionary* z roku 1928 został opracowany na podstawie fiszek z cytatami z literatury (w 1882 r., czyli na dwa lata przed opublikowaniem pierwszego tomu, zgromadzono ich 3,5 miliona), które razem obejmowały kilkadziesiąt milionów wyrazów; patrz Mugglestone 2005 oraz praca, w której porównano podejście do tworzenia słownika zastosowane w dziele Johnsona i *Oxford English Dictionary* (Silva 2005).

4. **Gramatyka**

Korpus może być podstawowym źródłem danych, na podstawie którego opracowuje się gramatyki opisowe.

Jedną z najważniejszych późniejszych prac tego rodzaju był *Survey of English Usage* (Quirk 1968). Podstawowym celem korpusu było uzyskanie danych, na podstawie których można by opisać gramatykę wykorzystywaną przez dorosłych, wykształconych (co wpłynęło na charakter tekstów – stosunkowo oficjalny, często akademicki) użytkowników brytyjskiej odmiany języka angielskiego. Badania rozpoczęto w 1959 r., gromadząc teksty pisane (informacyjne, szkoleniowe, korespondencję, dzienniki, sztuki teatralne) oraz nagrania (monologi, dialogi, audycje radiowe, przemówienia, rozmowy

telefoniczne), które następnie transkrybowano i zapisywano na papierowych fiszkach. Niezwykle istotne jest to, że dążono do równowagi tekstów mówionych i pisanych. W ten sposób zgromadzono 200 tekstów po 5000 wyrazów każdy (zgodnie z wcześniej wspomnianymi założeniami Kučery i Francisa). Wszystkie teksty adnotowano, dodając dokładny opis gramatyczny. Obecnie korpus jest już całkowicie skomputeryzowany, będąc przykładem ewolucji zbioru tekstów powstałego przed zastosowaniem w językoznawstwie komputerów. Na podstawie badań wydano referencyjną gramatykę[17] języka angielskiego *A Comprehensive Grammar of The English Language* (Quirk et al. 1985). Informacje na ten temat znaleźć można także w: Hunston, Francis 2000, Lewandowska-Tomaszczyk et al. 2001.

2.4.2 Wybrane korpusy komputerowe

2.4.2.1 Brown Corpus

Pierwszym korpusem komputerowym był zbiór tekstów Kučery i Francisa powstały w roku 1967 (Kučera i Francis, 1967) – *Brown University Standard Corpus of Present-Day American English* (znany bardziej jako *Brown Corpus*), który – choć niewielki jak na dzisiejsze warunki – był w tych czasach dziełem przełomowym. Powstał on w okresie niełatwym dla językoznawstwa korpusowego, naznaczonym krytycyzmem Chomsky'ego (patrz rozdział 2, punkt 1.1). Składał się z tekstów powstałych w Stanach Zjednoczonych w roku 1961 (zawierał 500 próbek po 2000 wyrazów), należących do 15 gatunków literackich (teksty prasowe, w tym reportaże, artykuły redakcyjne i recenzje, religijne, hobbistyczne, popularne, literatura piękna, naukowe, beletrystyka, w tym literatura naukowa, przygodowa, romantyczna, fantastyka naukowa i humor). Co ciekawe, pierwsza wersja korpusu zapisana była na kartach perforowanych, zaś na przykład wielkie litery oznaczono gwiazdką.

[17] Typologia gramatyk, por. Chalker 1994.

Według Teuberta (2004) to właśnie Francis jako pierwszy zastosował termin „korpus” do określenia elektronicznego zbioru tekstów.

2.4.2.2 British National Corpus

Korpus ten jest zbiorem tekstów mówionych i pisanych brytyjskiej odmiany języka angielskiego z końca wieku XX. Jest to korpus monolingwalny, synchroniczny, ogólny, próbkowany (w przypadku tekstów pisanych gromadzi się fragmenty o długości do 45 tysięcy wyrazów, co umożliwia osiągnięcie większej reprezentatywności). Korpus nie jest zrównoważony pod względem tekstów pisanych i mówionych (odpowiednio 90% i 10%), przy czym teksty pisane obejmują artykuły prasowe (ogólne i specjalistyczne), podręczniki, literaturę piękną, listy, wypracowania i inne, zaś część mówiona zawiera swobodne konwersacje nagrywane przez ochotników, a także mające zdecydowanie bardziej oficjalny charakter przemówienia (debaty polityczne, kazania) oraz nagrania audycji radiowych i programów z udziałem słuchaczy.

Korpus jest całkowicie adnotowany zgodnie ze standardem TEI[18]. Zawiera informacje o częściach mowy adnotowane przez automatyczny tagger CLAWS (Garside 1987) oraz dane metatekstowe (nagłówki, podział na paragrafy itd.). Całkowita wielkość korpusu to 100 milionów wyrazów w 4054 tekstach zajmujących około 1,5 GB pamięci dyskowej. Gromadzenie rozpoczęto w 1991 r. i zakończono w roku 1994, zaś pierwsze wydanie ogólne korpusu pojawiło się rok później. W roku 2007 opublikowano wydanie trzecie *BNC XML Edition*. Wydanie to jest dostępne na DVD obok korpusu *BNC Baby*

[18] TEI (Text Encoding Initiative) to konsorcjum powołane w celu opracowania standardu kodowania tekstów w postaci cyfrowej, służącego przede wszystkim badaniom naukowym, ale także użytkownikom komercyjnym. Wytyczne TEI dotyczą wszystkich aspektów prezentowania tekstów (ustępy, interpunkcja, wytłuszczanie i cytowanie, słowa obce, nagłówki, znaki niestandardowe, transkrypcja mowy, wykresy, wzory i tabele, a także dodawanie informacji kontekstowych, które ma zastosowanie w anotowaniu korpusów). Najnowsze wydanie wytycznych TEI (*P5 Guidelines)*, które ukazało się 1 listopada 2007 r., korzysta z możliwości, jakie daje język XML.

zawierającego cztery miliony wyrazów. Interesujące jest to, że oba wydania zawierają także wersję korpusu odmiany amerykańskiej języka angielskiego *Brown Corpus* w formacie XML.

2.4.2.3 American National Corpus

Ponieważ odmiany języka angielskiego brytyjska i amerykańska znacznie – i coraz bardziej – się różnią (Kachru 1982, Algeo 2006, Peters 2004, Trudgill, Hannah 2002), konieczne było stworzenie samodzielnego korpusu angielszczyzny amerykańskiej, wzorowanego zresztą na korpusie brytyjskim BNC. W planach jest stworzenie korpusu zasadniczego zawierającego przynajmniej 100 milionów słów pisanej i mówionej odmiany języka, obejmującego różnorodne rodzaje tekstów, w tym gatunki, które pojawiły się użyciu niedawno, np. blogi, listy elektroniczne i czaty. Pierwsze wydanie korpusu liczące ponad 11 milionów słów ukazało się w roku 2003. Korpusowi zasadniczemu będzie towarzyszyć dodatkowy zbiór tekstów zawierający kilkaset milionów słów i oferujący szerokie spektrum danych (Fillmore 1998, Ide 2001, 2004).

2.4.2.4 Cobuild, Bank of English

Jest to korpus monitorujący, zawierający obecnie 524 milionów słów.

Początkowo projekt był prowadzony w ramach współpracy wydawnictwa Collins i Wydziału Języka Angielskiego uniwersytetu w Birmingham jako *Collins Birmingham University International Language Database* (*Cobuild*). Na jego podstawie miano opracować nowy słownik języka angielskiego, dlatego też korpus „miał odpowiadać potrzebom uczących się języka, nauczycieli i innych użytkowników, a także stanowić wartościowe źródło informacji dla badaczy zajmujących się współczesnym językiem

angielskim" (Renouf 1987: 3). Korpus zawierał jedynie jedną czwartą tekstów mówionych, z naciskiem na teksty o charakterze ogólnym i naturalne użycie języka, głównie brytyjskie, z mniejszym udziałem amerykańskich. W roku 1982 zawierał 7,3 miliona wyrazów (Kennedy 1998: 46), zaś obecnie – już jako *Bank of English* – mieści 524 milionów wyrazów i jako korpus monitorujący jest stale rozbudowywany. Z gromadzeniem tak ogromnej liczby danych wiązały się zresztą pewne problemy (Blackwell 1993), takie jak rozpowszechniona wówczas niekompatybilność oprogramowania i sprzętu komputerowego oraz błędy typograficzne.

Na podstawie korpusu opracowano w 1987 r. rewolucyjny, ponieważ oparty na korpusie, monolingwalny słownik uczniowski *Collins COBUILD English Language Dictionary* (nowe wydanie, Sinclair 1995).

2.4.2.5 International Corpus of English (ICE)

Zadaniem tego korpusu (Greenbaum 1996) jest gromadzenie danych do badań porównawczych odmian języka angielskiego na całym świecie, dlatego też w prace zaangażowanych jest 15 zespołów, które przygotowują oddzielne korpusy elektroniczne w krajach i regionach, takich jak: Afryka Wschodnia, Australia, Filipiny, Fidżi, Hongkong, Indie, Irlandia, Jamajka, Kanada, Malezja, Nowa Zelandia, RPA, Singapur, Sri Lanka, Stany Zjednoczone i Wielka Brytania.

Najwcześniejsze zgromadzone teksty pochodzą z roku 1990. Autorów wypowiedzi pisemnych lub ustnych dobrano, tak by język angielski był dla nich językiem ojczystym; ewentualnie dopuszczalna była także sytuacja przeprowadzenia się do kraju, w którym język angielski jest powszechnie używany, i zdobycie tam wykształcenia. Dzięki takiej strukturze korpusu będzie możliwe badanie odmian języka angielskiego, w tym różnic w ich cechach gramatycznych. Trzeba jednak pamiętać, że osiągnięcie całkowitej porównywalności korpusów składowych w obszarze takich parametrów, jak

płeć, wiek, pochodzenie społeczne lub etniczne, nie będzie możliwe. Na przykład w niektórych społeczeństwach stosujących na co dzień język angielski istnieje nierównowaga między liczbą dziewcząt i chłopców uczących się w szkołach.

Każdy korpus składowy zawiera około 1 miliona słów (500 tekstów po około 2000 wyrazów) w tekstach następujących gatunków:

- teksty mówione:
    - dialogi (prywatne i publiczne),
    - monologi (niezapisane i zapisane, czyli odczytywane),
- teksty pisane:
    - niepublikowane (wypracowania uczniów i studentów oraz listy),
    - publikowane (naukowe, popularne, reportaże, instruktażowe, artykuły redakcyjne, literackie).

Co ważne, w korpusie tym przeważają nad tekstami pisanymi teksty mówione (60%). Można to uzasadnić tym, że różnice między odmianami języka angielskiego uwidoczniają się szczególnie w mowie, zaś w piśmie są mniej wyraziste.

Korpus jest adnotowany następującymi informacjami:

- dane metajęzykowe (oznacza się: strukturę tekstu, nagłówki, układ typograficzny oraz adnotacje fonetyczne, takie jak nakładanie się wypowiedzi, pauzy, wahania itd.),
- części mowy – adnotowane za pomocą taggera TOSCA,

- parsing syntaktyczny – automatyczny przy użyciu programu opracowanego na uniwersytecie w Nijmegen.

2.4.2.6 The Helsinki Corpus of English Texts

Korpus ten (Kytö 1988, 1994), gromadzony w ramach projektu rozpoczętego w roku 1984, zawiera wyjątki z tekstów ciągłych powstałych w okresie od 750 (pierwsze zachowane teksty) do 1700 r. (koniec okresu wczesnego nowoczesnego języka angielskiego) – część diachroniczna – oraz transkrypcje współczesnych wywiadów z użytkownikami dialektów wiejskich w Wielkiej Brytanii (część dialektalna). Zawiera około 1,5 miliona słów, z czego 75% obejmuje część diachroniczna, czyli teksty historyczne.

2.4.3 Korpusy w Polsce

Jednym z pierwszych doniosłych wydarzeń w polskim językoznawstwie korpusowym było sporządzenie list frekwencyjnych języka polskiego (Kurcz et al. 1974–1977), a na ich podstawie – w latach dziewięćdziesiątych ubiegłego wieku – *Słownika frekwencyjnego polszczyzny współczesnej* (Kurcz et al. 1990).

2.4.3.1 PELCRA

Projekt ten jest wynikiem współpracy Katedry Języka Angielskiego Uniwersytetu Łódzkiego (obecnie – Katedry Języka Angielskiego i Językoznawstwa Stosowanego) i Departamentu Językoznawstwa i Współczesnego Języka Angielskiego Uniwersytetu w Lancaster. PELCRA (*Polish and English Language Corpora for Research and Application*) miał następujące zadania (Lewandowska-Tomaszczyk 2005: 19): gromadzenie

danych języka polskiego, danych o języku angielskim używanym przez jego polskich użytkowników a także tekstów oryginałów i ich przekładów w parze językowej polski-angielski. Celem drugim było zapisanie zgromadzonych danych w postaci elektronicznej (wykorzystano strukturę analogiczną do *British National Corpus*), a także udostępnianie aplikacji służących do obróbki zawartych w korpusie danych (szereg narzędzi do badania konkordancji, fleksji, częstości słów i kolokacji) (http://pelcra.ia.uni.lodz.pl/intro_pl.php).

W skład korpusów językowych opracowanych w ramach projektu wchodzą:

- *Polski Korpus Narodowy* (*Polish National Corpus*) – korpus referencyjny języka polskiego zawierający ponad 130 milionów wyrazów, z czego udostępniona do powszechnego użytku jest jego część zbalansowana licząca 93 milionów segmentów słów.

- *Polski Korpus Uczniowski Języka Angielskiego* (*Polish Learner English Corpus*) liczący obecnie około 500 000 słów, w którym zawarto dane zgromadzone od uczących się na różnych poziomach zaawansowania. Korpus ten posłuży do badań kontrastywnych (np. badania określoności, części mowy, stałych związków wyrazowych), a także analizy błędów popełnianych przez uczących się.

- *Polski Multimedialny Korpus Konwersacyjny* (*Polish Spoken Conversational Multimedia Corpus*, 500 000 słów, docelowo – 1 milion). Część korpusu referencyjnego obejmująca dane mówione. Rejestrowano wypowiedzi spontaniczne, przy czym rozmówcy nie byli informowani o nagrywaniu. Korpus jest adnotowany – zawiera informacje o płci, wieku i wykształceniu rozmówców oraz o podstawowych cechach prozodycznych. Dane dostępne są także w postaci plików dźwiękowych MP3, co umożliwia bezpośrednią analizę charakterystyki prozodycznej.

- Angielsko-polskie i polsko-angielskie korpusy paralelne (równoległe) i porównywalne. Korpusy zawierające głównie teksty oryginalne i ich tłumaczenia (przede wszystkim w kierunku z języka polskiego na angielski) czasopism i książek, służące do badań kontrastywnych i translatologicznych.

Obecnie projekt PELCRA jest realizowany w ramach Narodowego Korpusu Języka Polskiego (NKJP).

2.4.3.2 Narodowy Korpus Języka Polskiego

Korpus ten jest tworzony we współpracy Instytutu Podstaw Informatyki PAN (koordynator), Instytutu Języka Polskiego PAN, Wydawnictwa Naukowego PWN oraz Zakładu Językoznawstwa Komputerowego i Korpusowego Uniwersytetu Łódzkiego.

Wedle założeń (Przepiórkowski et al. 2009) korpus ten ma zawierać 1 miliard wyrazów, z czego subkorpus liczący co najmniej 300 milionów wyrazów ma być starannie zrównoważony. Na podstawie korpusu konwersacyjnego PELCRA (patrz rozdział 2, podrozdział 4.3.1) zostanie opracowana część (30 milionów wyrazów) zawierająca wypowiedzi mówione (przemówienia publiczne, protokoły obrad parlamentarnych, debaty telewizyjne i inne, w tym podzbiór z wypowiedziami naturalnymi użytkowników języka – 3 miliony wyrazów).

Cały korpus będzie anotowany pod względem lingwistycznym, strukturalnym i bibliograficznym, przy czym ze względu na wielkość korpusu anotacja manualna będzie obejmować subkorpus złożony z 1 miliona wyrazów.

Wersja demonstracyjna Narodowego Korpusu Języka Polskiego, obejmująca około 500 milionów wyrazów, jest dostępna na stronie internetowej http://nkjp.pl/index.php?page=6&lang=1.

2.4.3.3 Korpus Języka Polskiego PWN

Korpus opracowany przez Wydawnictwo Naukowe PWN, dostępny częściowo w Internecie.

Wersja sieciowa, dostępna bezpłatnie, jest zrównoważona tematycznie i gatunkowo. Obejmuje 661 próbek tekstu o objętości od 1 do 1,5 arkusza wydawniczego w przypadku książek i czasopism oraz 0,5 arkusza w przypadku poezji i gazet codziennych. Próbki pobierano z 75 książek beletrystycznych, 115 niebeletrystycznych i 154 numerów 72 tytułów prasowych. Ponadto w korpusie umieszczono 69 plików z transkrypcjami rozmów i 85 plików z ulotkami reklamowymi i innymi tekstami użytkowymi. Łącznie korpus ten obejmuje ponad 3,7 miliona wyrazów.

Wersja pełna z kolei obejmuje 2070 próbek tekstów o rozmiarach od 1 do 6 arkuszy wydawniczych, pobieranych z 195 książek beletrystycznych i 191 niebeletrystycznych oraz 977 numerów 185 tytułów prasowych. Ponadto 84 pliki z transkrypcjami rozmów, 272 pliki z ulotkami reklamowymi oraz 207 plików z tekstami internetowymi.

Korpusy mają charakter diachroniczny, ponieważ umieszczono w nich nie tylko teksty współczesne, ale także dawniejsze (4,5% tekstów z lat 1920–1945 oraz 10% tekstów z lat 1946–1969).

Pliki znajdujące się w obu korpusach są indeksowane. Tagi obejmują informacje o strukturze tekstu, wyrazach i konstrukcjach nietypowych bądź błędnych oraz dane o autorach lub uczestnikach rozmów.

Informacje o korpusie są dostępne na stronie http://korpus.pwn.pl/.

2.4.3.4 Korpus IPI PAN

Korpus ten, liczący 250 mln wyrazów, został utworzony przez Zespół Inżynierii Lingwistycznej w Instytucie Podstaw Informatyki PAN (IPI PAN) (Przepiórkowski 2004).

Obecnie dostępne jest 2. wydanie korpusu z marca 2006 r.: wersja pełna oraz próbka, zawierająca ponad 30 mln wyrazów, której skład jest następujący: proza współczesna (ponad 10%), proza dawna (prawie 10%), teksty książkowe niebeletrystyczne (10%), prasa (50%), stenogramy parlamentarne (15%), ustawy (5%).

Pełne informacje znajdują się na stronie internetowej www.korpus.pl.

2.4.4 Zbiory tekstów

Zbiory omówione powyżej to typowe korpusy, a więc kolekcje tekstów gromadzone przez językoznawców do badań, podporządkowane z góry ustalonym regułom. Wartościowym, także naukowo – chociaż niewątpliwie niereprezentatywnym i nieoddającym języka jako całości – rodzajem materiału, są zbiory tekstów, które coraz częściej dostępne są, za opłatą (archiwum *Gazety Wyborczej*, *Rzeczpospolitej*, *Science*) lub bez (archiwum *Die Zeit*, *Projekt Gutenberg*) – w sieci WWW w postaci elektronicznej.

Problemem może być w takim wypadku dostęp do archiwum jako całości, a więc pobranie pełnych tekstów, najdogodniej w formacie .html lub .txt. Część bowiem plików – na przykład artykułów prasowych – dostępna jest wyłącznie w formacie .pdf, który najczęściej (w przypadku plików tworzonych przez oprogramowanie OCR) nie nadaje się do badań bez wcześniejszego przekształcenia w format możliwy do odczytania, natomiast to przekształcenie wiąże się z koniecznością wykonania szeregu czynności manualnych (np. problem interpretacji znaków przeniesienia).

Poniżej podano wybrane przykłady zbiorów tekstów:

- archiwum tygodnika *Time* – liczące ponad 100 milionów wyrazów archiwum obejmujące teksty od roku 1923 (http://corpus.byu.edu/time/),

- *Projekt Gutenberg* – archiwum gromadzi teksty książek znajdujących się w domenie publicznej, a zatem takich, do których wygasły prawa autorskie (obecnie w USA wygaśnięcie praw następuje 50 lat po śmierci autora). Obejmuje literaturę popularną i piękną oraz źródła o charakterze leksykograficznym czy encyklopedycznym (słowniki, encyklopedie, tezaurusy) głównie w ramach kultury zachodniej, w języku angielskim, ale także francuskim, niemieckim, fińskim i innych. Język polski także jest reprezentowany, niestety jednak szczątkowo (http://www.gutenberg.org/wiki/Main_Page); dostępne są na przykład następujące teksty: *Pan Tadeusz* oraz *Sonety* Adama Mickiewicza, *Sklepy cynamonowe* Brunona Schulza i inne pozycje; wyszukiwarka w 2006 r. podawała 25 pozycji, zaś 12 maja 2011 r. dostępnych było 29 pozycji, co oznacza, że przyrost nie jest znaczny, z kolei w przypadu języka angielskiego było to 30178 pozycji,

- *Oxford Text Archive* – kilkaset tysięcy tekstów naukowych (w tym także korpusów językowych) w wielu językach w postaci elektronicznej (Burnard 1988, Edwards 1993) (http://ota.ahds.ac.uk/),

- *European Corpus Inititative* – projekt rozpoczęty w 1992. Zawiera teksty przede wszystkim w językach europejskich, w tym dużą liczbę wydań gazety *Frankfurter Rundschau*. Mimo obecności wielu języków Europy Środkowo-Wschodniej, tekstów polskich w zbiorze nie ma

(http://www.elsnet.org/resources/eciCorpus.html),

- archiwum *Gazety Wyborczej* – największe z ogólnodostępnych archiwum polskich tekstów prasowych, zawierające ponad 3 miliony tekstów: z wydania głównego począwszy od 1992 roku, zaś z wydań lokalnych – od roku 1999 (http://szukaj.gazetawyborcza.pl/archiwum/0,0.html),

- archiwum tygodnika *Polityka* – zawiera teksty od roku 1998 do chwili obecnej. Umożliwia wyszukiwanie zaawansowane (szukane słowo, autor, słowa w tytule, zakres dat wydania, rubryka, dział),

- archiwum dziennika *Rzeczpospolita* – zawiera teksty od roku 1993 do chwili obecnej, w tym dokumenty z serwisu prawnego i ekonomicznego oraz z dodatków.

2.5 Narzędzia komputerowe i procedury stosowane w badaniu korpusów

Komputery we współczesnym językoznawstwie korpusowym stanowią podstawowe narzędzie pracy – zapewniają ogromną szybkość, powtarzalność i odtwarzalność operacji, pozwalają wykluczyć nieuchronny w przypadku analizy manualnej element przypadkowości w studiach lingwistycznych i pozwalają pracować z bardzo rozbudowanymi zbiorami tekstów. Oferują możliwości wyszukiwania ciągów znaków, ich sortowania, a także wykonywania obliczeń statystycznych, a co za tym idzie – przejścia od językoznawczych badań jakościowych do ilościowych, umożliwiających formułowanie uogólnień (Kennedy 1998: 5, por. także Tzoukerman, Klavans, Strzalkowski 2003: 529–544).

2.5.1 Wyszukiwanie i zastępowanie informacji

Do celów tych, czyli wyszukiwania i zastępowania danych, służą programy dedykowane oraz rozbudowane edytory tekstu, a także mniej zaawansowane programy przeznaczone konkretnie do wyszukiwania określonych ciągów znaków typu *znajdź i zastąp*. Umożliwiają one najczęściej przetwarzanie wsadowe, w przypadku którego w jednej operacji przetwarzanych jest wiele plików. W badaniach korpusowych ogromne znaczenie mają wyrażenia regularne – opisane szerzej w części 3 niniejszej pracy. Edytory odpowiednie do badań językoznawczych umożliwiają posługiwanie się wyrażeniami regularnymi o zaawansowanej składni. Przykładami edytorów, które umożliwiają korzystanie z wyrażeń regularnych, są TextPad (www.textpad.com), EditPad (www.editpadpro.com), EmEditor (www.emeditor.com), UltraEdit (www.ultraedit.com) oraz AptEdit (www.aptedit.com).

2.5.2 Frekwencja

Sporządzanie list frekwencyjnych (list częstości wystąpień wyrazów, ich połączeń, konstrukcji gramatycznych) dla korpusów polega na przyporządkowaniu każdemu wyrazowi liczby jego wystąpień i posortowaniu wyrazów w porządku alfabetycznym lub według liczby wystąpień. Jeśli informację taką połączyć z kontekstem, w którym dany wyraz występuje, bądź funkcją, którą pełni w zdaniu, otrzymujemy wartościowe informacje o użyciu wyrazu w danej funkcji[19]. Podobnie, listy frekwencyjne można tworzyć na podstawie par, trójek i większej liczby współwystępujących wyrazów (n-gramów). Umożliwia to badania szeregu rodzajów stałych połączeń wyrazowych (na przykład fraz rzeczownikowych, kolokacji itd.).

Jednym z najbardziej oczywistych zastosowań list frekwencyjnych jest dydaktyka języków obcych, w której są wykorzystywane do ustalania słownictwa podstawowego, rozumianego jako wyrazy najczęściej w języku występujące oraz przy tworzeniu słowników, ponieważ kolejne tzw. znaczenia leksykograficzne danego hasła uporządkowane są zazwyczaj zgodnie z częstością ich stosowania w języku.

2.5.3 Lematyzacja

Lematyzacja (nazywana niekiedy hasłowaniem, Bień, Szafran 2001: 72 lub tematyzacją, Lewandowska-Tomaszczyk 2005: 68) to sprowadzanie form fleksyjnych wyrazu do formy podstawowej, hasłowej, czyli leksemu (lemmy). W ten sposób w języku angielskim oraz polskim, na przykład, rzeczowniki sprowadzane są do mianownika liczby pojedynczej (jeżeli forma ta oczywiście istnieje), zaś czasowniki – do bezokolicznika. Zmniejsza się wówczas ogólna liczba różnych słów w korpusie, a dzięki takiemu przetworzeniu tekstu można

[19] Por. na przykład Golding, Schabes 1996 o kontekstowym systemie korekty pisowni.

łatwo znaleźć wszystkie wystąpienia danego słowa niezależnie od formy fleksyjnej, którą w danym kontekście przyjmuje (Klein, Simmons 1963), a w rezultacie – przeprowadzić badania statystyczne występowania danego leksemu.

Lematyzację – wykonywaną za pomocą tzw. algorytmów stemmingu (por. Wierzchoń 2004) wykorzystuje się głównie w badaniu języków charakteryzujących się mniej złożonym systemem fleksyjnym, w tym w języku angielskim (por. Porter 1980, De Schryver 2003, O'Grady et al. 1997).

Współczesne programy komputerowe służące do lematyzacji korzystają z co najmniej dwóch metod hasłowania:

1. Reguły zapisane w pamięci programu (program przechowuje wszystkie formy danego wyrazu i na ich podstawie określa wyraz hasłowy). Na przykład forma dopełniacza liczby pojedynczej *wody* zostaje sprowadzona do formy mianownikowej *woda*.

2. Jeśli dana forma w pamięci lematyzatora nie istnieje, program próbuje „odcinać" od analizowanej formy prefiksy, aby uzyskać formę znajdująca się w wykazie wyrazów (na przykład z wyrazu *deinternationalisation* odcięty zostanie prefiks *de-* i pozostanie znana forma *internationalisation*).

Podstawowym problemem jest odróżnienie homografów, a więc wyrazów identycznych pod względem graficznym, jednak różniących się znaczeniem, jak w następujących przykładach:

a) *How the generalized Maxwell equations can be derived from the Einstein postulate.*

b) *Hinged closure attachment for insulated beverage can container - US Patent 4872577 from Patent Storm.*

W obu zdaniach występuje wyraz „can” – w pierwszym jako czasownik modalny, w drugim zaś – jako rzeczownik. Jedyną możliwością właściwego przyporządkowania wyrazu hasłowego jest w tym miejscu analiza całego kontekstu i przyporządkowanie części mowy każdemu wyrazowi w zdaniu. Por. Artiles et al. (2004) na temat dezambiguacji znaczenia wyrazów z użyciem korpusu treningowego w przestrzeni kontekstowej.

2.5.4 Analiza części mowy

Adnotacja w korpusie części mowy (ang. *part-of-speech annotation*), określana także jako adnotacja gramatyczna lub morfosyntaktyczna, jest procesem bardziej złożonym od lematyzacji i często prowadzonym równolegle, także z tego względu, że informacja o części mowy pozwala z większą precyzją określić formę wyrazu hasłowego, na przykład rozróżnić homografy (choć tylko, jeśli różnią się częścią mowy; różnice semantyczne pozostają poza zasięgiem tej metody).

Programy komputerowe do automatycznej analizy części mowy to analizatory morfologiczne, zaś zestaw symboli opisujących poszczególne kategorie zwany jest z języka angielskiego tagsetem.

Tagset powinien w sposób kompletny (ze względu na przyjętą gramatykę morfologiczną) opisywać wszystkie kategorie w danym języku w sposób możliwie precyzyjny, tak by dla każdego wyrazu istniała odpowiadająca mu kategoria, nawet jeśli w danym przypadku analizator dokonałby błędnego przypisania. Mimo że pierwsze analizatory komputerowe korzystały z niezłożonej listy kategorii, a ponadto każdy program opierał się na innej liście (Stolz 1965, Klein, Simmons 1963), co uniemożliwiało porównywanie wyników, tagsety współczesne charakteryzują się dekompozycyjnością i hierarchicznością (Lewandowska-Tomaszczyk 2005: 78). Dekompozycyjność tagsetów polega na tym, że każdy znak lub ich ciąg ma odrębne znaczenie i w różnych kombinacjach może wchodzić w skład

większej liczby indeksów. Z kolei hierarchiczność powoduje, że sekwencja znaków indeksu składa się ze znaków od bardziej ogólnych do szczegółowych. Dzięki temu konstrukcja indeksów jest logiczna, a liczba różnych ich składników powinna być mniejsza od całkowitej liczby indeksów, co ułatwia posługiwanie się nimi.

Praca analizatora morfologicznego składa się z etapu tokenizacji, właściwego etapu analizy oraz – w przypadku analizatorów bardziej zaawansowanych – dezambiguacji (Lewandowska-Tomaszczyk 2005: 82).

Tokenizacja

Tokenizacja to podział tekstu wejściowego na jednostki, które zostaną poddane analizie. W języku polskim za granice wyrazów można przyjąć spację lub znak interpunkcyjny (na przykład przecinek, kropka, cudzysłów zamykający, średnik, dwukropek, wykrzyknik, znak zapytania), jednak język angielski, na przykład, już na tym etapie analizy powoduje pewne trudności:

Jack's at home (≡*Jack is at home*): Jack's to dwa słowa.

Jack's home is awesome: *Jack's* to jedno słowo (dopełniacz saksoński).

Trudność ponadto sprawiają w języku angielskim wyrazy pisane z łącznikiem (np. *B-52*, *self-evaluation*, *twenty-two*), które uznaje się za pojedynczą jednostkę, i frazy typu *acid-containing*, które stanowią dwie jednostki.

Widać więc, że już na pierwszym, całkowicie prostym wydawałoby się etapie tokenizacji pojawiają się problemy. Trudniejsza jest tokenizacja w przypadku języków aglutynacyjnych, w których słowa zbudowane są z większej liczby morfemów niż w przypadku języków syntetycznych (por. na przykład Mikheev 2003: 210). Ponadto postać tekstu może utrudniać tokenizację, na przykład tekst napisany wyłącznie wielkimi literami

(rozróżnienie końca zdania od skrótu) bądź teksty wytworzone przez automatyczne systemy rozpoznawania mowy.

Analiza

Na etapie analizy wykorzystuje się dwie metody analogiczne do stosowanych przy lematyzacji, to znaczy porównywanie danych wejściowych z listą słów i ich form (uzyskaną w wyniku analizy innych korpusów lub na podstawie słownika) oraz analizę morfologiczną (ang. *affix stripping*) dokonywaną na podstawie zbioru reguł fleksyjnych danego języka. Na tym etapie analizator w wielu przypadkach przypisuje wyrazowi więcej niż jedną możliwą kategorię (np. w języku polskim *kara* jako rzeczownik lub przymiotnik, zaś w języku angielskim *water* jako rzeczownik lub czasownik), dlatego niezbędny jest etap kolejny, czyli dezambiguacja.

Dezambiguacja

Na tym etapie dane wejściowe to wszystkie wyrazy, którym przypisano więcej niż jedną kategorię, zatem konieczny jest wybór jednego, poprawnego indeksu. Ponieważ na podstawie samego wyrazu i jego formy nie jest możliwe przypisanie jednoznaczne, następuje analiza kontekstu, czyli otoczenia badanego wyrazu.

Wg T. Lewandowskiej-Tomaszczyk (2005: 85) istnieją cztery rodzaje programów do dezambiguacji, czyli dezambiguatorów:

1. Źródło reguł: znajomość gramatyki autora programu; sposób kodowania wiedzy: reguły. Oznacza to, że autor – kierując się własną wiedzą – formułuje niezmienne reguły służące do dezambiguacji. Pierwsze dezambiguatory konstruowano w ten właśnie sposób (np. CGC, Klein, Simmons 1963 oraz TAGGIT, Greene, Rubin 1971, por. też Karlsson 1995).

2. Źródło reguł: korpus tekstów; sposób wyrażenia: prawdopodobieństwo. Dezambiguator tego typu konstruuje się w ten sposób,

że na podstawie korpusu indeksowanego (anotowanego) określa się prawdopodobieństwo wystąpienia danej sekwencji indeksów. Następnie na podstawie uzyskanych wartości w procesie analizy korpusu nieanotowanego wylicza się prawdopodobieństwa wystąpienia poszczególnych kombinacji indeksów (najwyższe prawdopodobieństwo – przy założeniu odpowiedniej różnicy prawdopodobieństwa największego i drugiego z kolei – oznacza poprawne przypisanie).

3. Źródło reguł: korpus tekstów; sposób wyrażenia: reguły. Podobnie jak w pierwszym typie, formułuje się ścisłe reguły, powstają one jednak nie w oparciu o wiedzę, lecz o indeksowany korpus. Przykładem takiego dezambiguatora jest program Brilla (patrz poniżej).

4. Źródło reguł: wiedza autora programu, sposób wyrażenia: prawdopodobieństwo. W tym wariancie autor musiałby w oparciu o wiedzę własną określić częstotliwość występowania poszczególnych sekwencji indeksów. Z oczywistych względów dezambiguator taki nie może powstać, ponieważ użytkownicy języka (nawet jeśli dany język jest ojczysty) nie mogą na podstawie swojej wiedzy określić żadnych danych statystycznych obejmujących język jako całość.

Jakość wyników uzyskiwanych przez analizator morfologiczny określa się na podstawie ilości podawanych informacji oraz dokładności adnotacji (porównanie wyników z dużym, zróżnicowanym, adnotowanym korpusem odniesienia) (por. Voutilainen 2003: 223).

2.5.4.1 Przykładowe analizatory morfologiczne

Prace nad analizatorami morfologicznymi rozpoczęto już w latach 50. XX wieku (Voutilainen 2003: 223), zaś jednym z pierwszych komputerowych analizatorów morfologicznych był TAGGIT (Greene, Rubin 1971), wykorzystany do adnotacji *Brown Corpus*. W analizatorze tym zapisano szereg

reguł kontekstowych (3300 reguł) pozwalających przypisać części mowy. Takie podejście – jakkolwiek w owym czasie jedyne – miało podstawowy mankament: jeśli żadna z reguł nie odpowiadała danemu wyrazowi, analizator nie był w stanie podać żadnego przypisania. Zawierał on aż 87 kategorii części mowy i charakteryzował się skutecznością wynoszącą około 77%. Pozostałe 23% wyrazów zostało adnotowanych ręcznie (Francis i Kučera 1982: 9). Tagset zastosowany w analizatorze charakteryzował się częściową dekompozycyjnością, chociaż na przykład czasowniki posiłkowe (*be*, *do*, *have*) oznaczano odpowiednio symbolami BE, DO i HV, a więc niemającymi charakteru dekompozycyjnego (brak symbolu czasownika i jego typu).

Znacznie bardziej zaawansowanym narzędziem był system adnotacji CLAWS (*Constituent Likelihood Automatic Word-tagging System*), w którym wykorzystano reguły prawdopodobieństwa oparte na otwartym modelu Markova (por. na przykład Seymore 1999, Cutting et al. 1992), w którym kategorie gramatyczne i ich częstotliwość podane są bezpośrednio (Voutilainen 2003: 224). Postęp w porównaniu do analizatora TAGGIT był znaczny, ponieważ skuteczność adnotacji korpusu LOB (patrz punkt 2.3.5) wynosiła już 96–97% w zależności od rodzaju tekstu (Garside 1987: 9). Tagset wykorzystany w analizatorze był bardzo podobny do wykorzystanego w programie TAGGIT, aby zapewnić kompatybilność obu korpusów, dokonano jednak pewnych modyfikacji, tak że całkowita liczba kategorii wynosiła 133.

Adnotacja składa się z etapu przypisywania każdemu wyrazowi w analizowanym korpusie wszystkich możliwych kategorii gramatycznych (moduł WORDTAG analogiczny do programu TAGGIT, a więc nieuwzględniający kontekstu), poszukiwania idiomów (IDIOMTAG) oraz dezambiguacji (program CHAINPROBS), czyli badania wszystkich wyrazów, którym przypisano więcej niż jeden indeks.

Program CHAINPROBS rozpatruje wyrazy w kontekście, określając prawdopodobieństwo poprawności przypisanych przez WORDTAG kategorii.

Jeśli wskaźnik prawdopodobieństwa dla jednej kategorii jest odpowiednio wysoki, pozostałe wcześniej przypisane kategorie zostają odrzucone. Jeśli nie – dezambiguacji dokonuje człowiek.

Matryca prawdopodobieństwa została wyznaczona na podstawie *Brown Corpus*. Zawiera ona wartości prawdopodobieństwa takiego zdarzenia, że wyraz kategorii Y wystąpi po wyrazie kategorii X (prawdopodobieństwo przejścia, ang. *transition probability*). W przypadku większych sekwencji wyrazów, które muszą zostać poddane dezambiguacji, konieczne jest obliczanie prawdopodobieństwa wystąpienia ciągu więcej niż dwuskładnikowego (obliczane są prawdopodobieństwa dla każdej kombinacji możliwych przypisań; wartość najwyższa jest rozwiązaniem).

Ostatnim segmentem, który po raz pierwszy zastosowano w programie CLAWS, jest moduł IDIOMTAG. Ponieważ analizując dane wynikowe wytworzone przez dwa poprzednie moduły, stwierdzono, że do niektórych grup wyrazów kategorie zostały przypisane błędnie, wprowadzono prosty moduł regułowy, który wyszukuje grupę około 150 fraz i wprowadza poprawne adnotacje. Przykładem jest wyrażenie *as well as*, które otrzymuje identyfikator CC (spójnik).

Ostatnim etapem analizy jest edycja ręczna: analizuje się i określa jednoznaczne przypisanie dla wszystkich przypadków pozostawienia przez program więcej niż jednej adnotacji wyrazu oraz sprawdza wszystkie przypisania w całym korpusie, tak by osiągnąć całkowitą poprawność adnotacji.

Do analizy *British National Corpus* wykorzystano poprawioną wersję programu CLAWS (Leech et al. 1994: 54), w którym zmniejszono do 60 liczbę kategorii, rozszerzono do 10 tysięcy listę słów w module WORDTAG i znacznie zwiększono możliwości modułu IDIOMTAG.

Przykładem trzeciego typu dezambiguatorów jest program Brilla (1992). Według autora ma on szereg zalet w porównaniu z programami

probabilistycznymi: mniejszą ilość danych w pamięci programu, przejrzystość reguł oraz łatwiejszą wymienność między korpusami i językami, a ponadto charakteryzuje się porównywalną skutecznością.

Pierwszy moduł – analizator – wykorzystuje metodę najprostszą (lista słów), jednak zawsze przypisuje tylko jedną kategorię. W przypadku wyrazów nieznanych przypisuje wyrazowi nieznanemu kategorię tę samą, co wyraz znany kończący się tymi samymi trzema literami, zaś w przypadku wyrazów nieznanych rozpoczynających się wielką literą zakłada, że są to nazwy własne. Na tym etapie poprawność wynosi około 92,1% (Brill 1992: 113).

W następnym etapie stosowane są reguły. Powstały one na podstawie korpusu treningowego (*Brown Corpus*). W przypadku, gdy przyporządkowanie (w porównaniu z indeksowanym korpusem referencyjnym) było błędne, program formułuje „łaty" (ang. *patches*), czyli proste reguły typu:

Jeśli wyraz został oznaczony indeksem *a* i znajduje się w kontekście C (tzn. na przykład jest poprzedzany przez indeks *z* lub indeks *z* następuje po nim lub przed tym wyrazem występuje na przykład indeks *w*, zaś po wyrazie – indeks *z*, zmień indeks na b).

Następnie wyliczana jest skuteczność danej reguły, czyli liczba poprawionych indeksów, które wcześniej przypisano błędnie, zaś zweryfikowane reguły są wykorzystywane do poprawiania przypisania korpusu właściwego. W ten sposób poprawność indeksowania wzrasta do około 95,9%.

Obecnie wprowadza się nowe systemy wykorzystujące modele Markova, czyli modele stochastyczne umożliwiające wnioskowanie i obliczenia modeli, które w innym przypadku nie poddawałyby się analizie, i uczenie maszynowe (ang. Machine Learning) oraz metody łączące modele opracowywane ręcznie i generowane automatycznie, w tym funkcję optymalizacji energii (Padro 1997, por. Voutilainen 2003: 227).

Należy wreszcie zwrócić uwagę na analizatory specyficzne dla języka polskiego, przede wszystkim SAM opracowany przez K. Szafrana (por. Bień, Szafran 2001, Hajnicz, Kupść 2001) oraz PoMor autorstwa R. Wołosza (Saloni, Wołosz 2001, Wołosz 2005), AMOR (Rabiega-Wiśniewska, Rudolf 2002), a także Morfeusz (Woliński 2006).

W analizatorze PoMor zastosowano rozbudowany zestaw indeksów, zgodny zasadniczo z pracą Tokarskiego (1993). Umożliwia on ponadto rozpoznanie znaków interpunkcyjnych, niektórych skrótów oraz liczb rzymskich i arabskich, a także zawiera znaczny zasób nazw własnych.

Przeglądu analizatorów morfologicznych dla języka polskiego (Gram, PoMor, SAM, LEM, XeLDA, AMOR) dokonały Hajnicz i Kupść (2001), skupiając się na zbadaniu skuteczności działania sześciu analizatorów morfologicznych w oparciu o analizę list słów (konstruowanych w sposób zasadniczo intuicyjny) poprawnych i niepoprawnych oraz czterech tekstów gazetowych.

2.5.5 Parsing

Parsing to pełna analiza syntaktyczna tekstu – a w ogólniejszym ujęciu każdej liniowej sekwencji elementów, w której wzajemna kolejność poszczególnych elementów podlega pewnym ograniczeniom związanym z daną gramatyką, czyli pewnym zbiorem reguł (Grune, Jacobs 1990: 13) – prowadząca do przypisania poszczególnym wyrazom pełnej adnotacji zawierającej informację już nie o części mowy, ale części zdania. Przypisanie to przeprowadza się zazwyczaj po analizie morfologicznej. Z racji typowej postaci analizy zdania korpusy poddane analizie syntaktycznej określa się jako „banki drzew” (częściej stosuje się określenie angielskie *treebanks*). Aby ułatwić zapis, drzewa opisujące poszczególne zdania nie mają postaci graficznej, lecz tekstową. Na przykład w notacji stosowanej w British National Corpus adnotowane zdanie „Claudia sat on a stool” ma następującą postać:

[S[NP Claudia_NP1 NP] [VP sat_VVD [PP on_II [NP a_AT1 stool_NN1 NP] PP] VP] S][20]

(McEnery, Wilson, 1996: 44)

Ponieważ nawet dla specjalisty analiza składniowa w pewnych przypadkach nie jest zadaniem trywialnym, parsing automatyczny jest jednym z najtrudniejszych problemów komputerowej analizy tekstu, wskutek czego skuteczność programów do analizy syntaktycznej (parserów) jest niższa niż analizatorów morfologicznych. Dlatego też często parsing komputerowy wspomaga się przez adnotowanie ręczne, którego ograniczeniem jest z kolei niemożność uzyskania całkowitej powtarzalności adnotacji i znacznie mniejsza wydajność pracy.

Podobnie jak w przypadku analizy morfologicznej, istnieją dwie główne metody analizy syntaktycznej: probabilistyczna i regułowa, zaś do elementów przysparzających największych trudności należą koordynacja, braki ciągłości wypowiedzenia i elipsy (Kennedy 1998: 232). Szczególnie kłopotliwa bywa automatyczna analiza tekstów mówionych, w których różnego typu potknięcia, wahania, dopowiedzenia i opuszczenia logicznie wynikające z wypowiedzi poprzedzającej zdarzają się zdecydowanie najczęściej.

Parsing jest wykorzystywany (Grune, Jacobs 1990: 11) nie tylko w językoznawstwie (analiza tekstu, analiza korpusowa, tłumaczenie maszynowe), ale także w redagowaniu i konwersji dokumentów, informatyce (np. opracowywanie kompilatorów, interfejsy baz danych, sztuczna inteligencja), a nawet naukach biologicznych (składanie wzorów chemicznych i identyfikacja chromosomów).

[20] S – zdanie (sentence), NP – fraza rzeczownikowa (noun phrase), VVD – czas przeszły, VP – fraza czasownikowa (verb phrase), NN – rzeczownik w liczbie pojedynczej lub rzeczownik zbiorowy, PP – fraza przyimkowa (prepositional phrase).

2.5.6 Konkordancja

Konkordancja to lista słów występujących w danym tekście (określanych niekiedy jako słowa kluczowe, ang. keywords) w postaci cytatów obejmujących słowa poprzedzające dane słowo i po nim następujące (liczbę słów otaczających można zmieniać zależnie od potrzeb). W ten sposób uzyskuje się przejrzysty wykaz wszystkich użyć danego wyrazu w całym tekście lub w zbiorze tekstów (a zatem także ich znaczeń), a także częstotliwość użycia wyrazu. Otrzymaną listę można poddać sortowaniu alfabetycznemu w dowolnej konfiguracji, a dzięki temu – badać kolokacje, znaczenia słów oraz przykłady użycia słów bądź fraz[21].

Mimo że obecnie konkordancje są automatycznie generowane przez programy komputerowe (dostępne na przykład wraz z korpusem PELCRA), pojęcie konkordancji i pierwsze prace nad tym zagadnieniem sięgają wieku XVIII, kiedy to Alexander Cruden w 1737 r. wydał konkordancję Biblii Króla Jakuba.

Przykładem współczesnego programu do wyszukiwania konkordancji są narzędzia dostępne w korpusie *Bank of English* wydawnictwa Collins. Zapytania formułuje się korzystając z następujących parametrów:

- operatory logiczne (np. AND, OR, NOT),
- połączenia słów, w tym podanie maksymalnej odległości między dwoma wyrazami w formacie słowo+Nsłowo, gdzie N jest maksymalną liczbą słów występującą między szukanymi słowami.

Przykład dla light+beam:

```
year. [p] Departing from
   the established light beam system found in
                         original Optonics,
                         Magnetonics
```

[21] Zastosowanie konkordancji w badaniach leksykograficznych, por. min. Tapanainen, Järvinen 1998).

a Magnetonic for 573 years, compared with light	**beam**	Optonic which can go flat in about six weeks
pulls the train along, head on. A light	**beam**	exerts pressure. As the power of the laser was
the material glow but you can't see the light	**beam**	itself." I remember now. The other light we
precise, coherent, pinpoint, concentrated light	**beam**	, because the original light source is itself
drove out the heat of sorrow. Sometimes a light	**beam**	beam shot through the ice, a slash of raving

Przykład dla light+3beam:

to the floor, its light shooting a frail	**beam**	that fanned out eerily across the carpet.
the material glow but you can't see the light	**beam**	itself." I remember now. The other light we
Flicking on the	**beam**	I scanned the beach. Nothing moved and I relit
at the caries, using a fibre-optic light. The	**beam**	then vaporises the decay and, unlike the
see the lighthouse, so you get the light of the	**beam**	all night if you leave the curtain open, which
precise, coherent, pinpoint, concentrated light	**beam**	, because the original light source is itself
of two men. Then the light flashed upward, its	**beam**	moving over the rocks and brush below him. He
to emit very short bursts of light in a narrow	**beam**	can circumvent this problem. A 200j laser

– znak @ odpowiadający wszystkim formom fleksyjnym danego wyrazu w formacie słowo@

Przykład dla beam@

and then looked over the top. The flashlight	**beam**	shone through a blue haze of gunpowder smoke
fog, clouds, and smoke disrupted the laser	**beam**	--a deft air crew could put an LGB within ten
said. [p] It did. I made a note. [p] Then he	**beamed**	, cracked with laughter and the quite
PHOTO WITH CAPTION [/c] Sunlight in the oak-	**beamed**	dining room, a temple of excellence [p] The
s sports spokesman, Steve Tshwete, positively	**beaming**	when they did the first of two clinics in
I'm afraid it's a hopeless cause." Owen,	**beaming**	at the camera, replies: `I'm your man." Then
projection covering the intersection of ribs or	**beams**	in a ceiling [p] Corbel: Support for a beam

– znaki zastępcze (ang. *wildcards*), które obejmują składnię wyrażeń regularnych,

– etykiety części mowy (ang. *Part-of-speech tags*) w formacie słowo/etykieta (dostępnych jest 18 różnych etykiet):

Przykład dla beam/NOUN:

each in turn to a toy boat rigged with a cross-	**beam**	mast, and tried them out in his bath. Using a

sake of simplicity (and safety) this splendid	**beam**	engine is powered not by steam but by air from
Africa. [p] A fire swept through the Marconi	**beam**	squatter camp in the early hours of Monday 9th
and the frontier: Elijah Craig, LW Harper, Jim	**Beam**	George Dickell and Jack Daniels sound like a
other bits, leaving little reminders. A roof-	**beam**	props up a wall. A door leads nowhere in
second lead. [p] To trigger the flash burst, a	**beam**	of infrared light is transmitted to a small
headroom and this, combined with her narrow	**beam**	, results in the smallest level of
in the master bedroom, where the great oak	**beam**	half hiding a Venetian gilt lantern) and a
the life from me. But I still have the freeze	**beam**	and more missiles and I manage to stun the
with water to waist level, with only a narrow	**beam**	to balance and sleep on. In May a United
at each of the X-ray sources in turn. The	**beam**	weighs nothing and is controlled by magnets
keys, there sure enough, and a half-twist for	**beam**	headlights. The dashboard's fluorescence cast
figure moving round the room behind a small	**beam**	of light, probably from a torch. There seemed

– wyszukiwanie kolokacji na podstawie informacji wzajemnej (ang. *mutual information*)[22] i wskaźnika T (ang. *T-score*)[23], por. też Kilgariff 2002:

[22] Wskaźnik ten, wprowadzony do praktyki badawczej w pracy Church i Hanks (1989), mierzy stosunek prawdopodobieństwa wystąpienia bigramu do iloczynu prawdopobieństw wystąpienia jego składników niezależnie ($MI(x,y) = \log_2(P(x,y) / (P(x)\times P(y)))$). W ten sposób można stwierdzić, czy składniki bigramu są ze sobą powiązane, ponieważ wskaźnik MI przyjmuje

Przykład dla informacji wzajemnej (słowo *beam*):

Collocate	Corpus Freq	Joint Freq	Significance
splitters	8	3	12.475680
scotty	18	4	11.720717
flashlight	167	32	11.506916
laser	384	37	10.515013
marconi	61	3	9.544650
infrared	143	7	9.537908
electron	98	4	9.275688
torch	260	10	9.189950
electrons	116	4	9.032393
headlights	195	6	8.867990

Przykład dla wskaźnika T (słowo *beam*):

Collocate	Corpus Freq	Joint Freq	Significance
a	973489	155	7.297412
the	2313407	265	6.914402
laser	384	37	6.078603
flashlight	167	32	5.654909
light	8915	29	5.276078

– możliwość zawężenia wyszukiwania do podkorpusów:

  - brytyjskie książki, audycje radiowe, teksty z gazet i czasopism,
  - amerykańskie książki i audycje radiowe,
  - angielszczyzna brytyjska mówiona.

Spośród wielu historycznych i współczesnych formatów zapisu konkordancji obecnie najczęściej wykorzystuje się format KWIC (ang. *Key*

wysokie wartości w przypadku składników występujących często w połączeniu i niskie, jeśli ich występowanie w połączeniu jest jedynie przypadkowe.

[23] Wskaźnik T zastosowali w językoznawstwie Gale et al. (1991). W porównaniu do wskaźnika MI ten wskaźnik uwzględnia także większą wagę wyników uzyskanych na podstawie większej liczby wystąpień (istotność statystyczna), ponieważ im większa jest suma częstości, tym wyższy wskaźnik T.

Word in Context), w którym po numerze linii, w której występuje szukany wyraz, podawany jest kontekst lewostronny (długość ustawiana przez użytkownika), szukany wyraz oraz kontekst prawostronny (długość ustawiana przez użytkownika).

3 Wyrażenia regularne w badaniach korpusowych

Automatyzacja operacji na korpusach wymaga zastosowania narzędzia pozwalającego na ścisłą (zgodną z intencją badacza) ekscerpcję szukanego ciągu znaków zgodnie z zasadami logiki. Niezbędną pomocą są tu wyrażenia regularne (ang. *regular expressions*), dzięki którym ze zbioru tekstów łatwo wydobyć szukane frazy. W rozdziale omówione zostaną podstawy wyrażeń regularnych oraz przykłady ich wykorzystania w badaniach lingwistycznych z punktu widzenia językoznawcy, a nie informatyka, a więc bez wykorzystywania złożonych języków programowania (Friedl 1998, Good 2004, Habibi 2003, Karttunen et al. 1996).

3.1 Wyrażenia regularne

Wyrażenia regularne to ciągi znaków zastępujących inne ciągi znaków (szukanych), przy czym pewne stosowane znaki (zgodnie z konwencją umowną) zastępują całą klasę znaków szukanych, na przykład litery, cyfry, słowa. Obok znaków dosłownych (np. liter, cyfr) stosuje się więc metaznaki mające znaczenie umowne, które omówiono poniżej. Wprowadzone wyrażeniem regularnym znaki są interpretowane przez program (edytor tekstowy). Mimo że ogólne zasady pozostają podobne, poszczególne programy (edytory tekstowe oraz środowiska programowania) różnią się składnią wyrażeń i możliwościami. Dlatego przed zastosowaniem konkretnego narzędzia należy zapoznać się z zastosowaną w nim konwencją. Podane poniżej przykłady zostały zastosowane i zweryfikowane przede wszystkim w edytorze TextPad (standard POSIX P1003.2), a także – EditPro. Edytor tekstu Word w pakiecie MS Office również pozwala wyszukiwać niektóre

proste wyrażenia regularne. Są one określane jako „symbole wieloznaczne" i obejmują dowolny znak, znak w zakresie, początek wyrazu, koniec wyrazu, wyrażenie i liczbę wystąpień (nazewnictwo według polskiej wersji programu MS Word 2003).

Bardziej zaawansowane edytory (np. EditPro) pozwalają w dużym zakresie stosować wyrażenia regularne także w przypadku operacji zastępowania (polecenie Replace).

3.2 Metaznaki

Podstawowe metaznaki (znaki mające znaczenie umowne) rozpoznawane przez program TextPad są następujące:

1. Klasa znaków:

   [*znak*, *znaki lub zakres*] – dopasowuje wszystkie znaki podane w nawiasie kwadratowym, przy czym w nawiasie mogą występować znaki oddzielone myślnikiem, oznaczające przedział.

   Przykłady:

   `[gra]`

   Działanie: dopasowuje wszystkie ciągi małych liter g, r i a[24] występujących w tej kolejności.

   `[a-z]`

   Działanie: dopasowuje wszystkie wystąpienia małych liter w tekście (wyrażenie stosowane w połączeniu z innymi znakami).

   `[0-9]`

   Działanie: wyszukuje wszystkie wystąpienia cyfr.

2. Dowolny znak:

   `.` (kropka) odpowiada dowolnemu znakowi

   Przykład:

   `ga.nik`

   Działanie: wyszuka ciąg znaków *gaźnik*, ale także *gajnik* w słowie *zagajnik*.

[24] W niektórych edytorach tekstowych (np. TextPad, EditPro) istnieje możliwość określenia, czy wyszukiwanie ma uwzględniać różnice między literami małymi a wielkimi (case sensitive/case insensitive). Jeśli na przykład w analizie nie ma znaczenia, czy wyraz rozpoczyna się wielką literą (np. na początku zdania), wyszukiwanie powinno ignorować rozróżnienie między literami małymi a wielkimi. W takim wypadku wyrazy Man, man i MAN będą uznawane za identyczne.

3. Alternatywa:

| – symbol ten oznacza *lub*

Przykład:

```
woda|wody|wodzie|wodą|wodo
```

Działanie: dopasowuje wszystkie wystąpienia ciągu znaków *woda* we wszystkich przypadkach (wyszukane zostanie też np. słowo *powoda*).

Wyrażenie to można uprościć korzystając z nawiasów okrągłych:

`wod(a|y|zie|ą|o)` – ponieważ ciąg znaków *wod* jest wspólny dla wyszukiwanych ciągów.

4. Początek wiersza:

^ – wyszukuje początek wiersza (w zależności od ustawień zawijania tekstu może być to początek akapitu).

Przykład:

```
^There
```

Działanie: wskaże wszystkie wiersze (akapity) rozpoczynające się od ciągu znaków *There* (a zatem także *Therefore*).

5. Koniec wiersza:

$ – wyszukuje koniec wiersza

Przykład:

```
[1-9]$
```

Działanie: dopasowuje wszystkie cyfry z przedziału od 1 do 9 występujące na końcu wiersza.

6. Sekwencja „unikowa”:

\ – ponieważ pewne znaki interpretowane są jako metaznaki (o znaczeniu umownym), sekwencja unikowa przed takim znakiem usuwa znaczenie umowne, czyli metaznakowe.

Przykład:

```
\.$
```

Działanie: wyszukuje wszystkie końce wierszy kończące się kropką (bez zastosowania sekwencji unikowej dopasowane zostałyby wszystkie wiersze kończące się dowolnym znakiem (niedopasowane pozostałyby jedynie wiersze puste).

7. Kwantyfikatory – różna ilość wystąpień znaku lub ciągu znaków:

a) znak lub ciąg opcjonalny (zero lub jedno wystąpienie)

? – znak (lub znaki w nawiasie okrągłym), po którym następuje ten metaznak, jest traktowany opcjonalnie, tzn. nie musi wystąpić w wyszukanej frazie. Znak umożliwia skrócenie wyrażenia regularnego, ponieważ odpowiednio użyta alternatywa jest mu równoważna.

Przykład:

```
mag?ma
```

Działanie: odnajduje ciągi znaków *magma* oraz *mama* (równoważne mama|magma).

b) dowolna liczba wystąpień

* – dopasowuje wyrażenie, jeśli znak lub ciąg znaków w nawiasie okrągłym, które go poprzedzają, występuje zero lub więcej razy.

Przykład:

```
\.\.*
```

Działanie: dopasowuje wszystkie wystąpienia co najmniej dwu kropek następujących po sobie (zastosowano także sekwencję unikową).

c) co najmniej jedno wystąpienie

+ – dopasowuje wyrażenie, jeśli znak lub ciąg znaków w nawiasie okrągłym, które go poprzedzają, występuje co najmniej raz.

Przykład:

```
\.\.+
```

Działanie: dopasowuje wszystkie wystąpienia co najmniej dwu kropek następujących po sobie (zastosowano także sekwencję unikową) – identyczne z przykładem powyżej.

```
[a-z]+
```

Działanie: znajduje wszystkie wystąpienia ciągów znaków literowych (wyrażenie takie dopasuje wszystkie wyrazy rozumiane jako ciągi znaków literowych ograniczone znakami nieliterowymi bądź początkiem lub końcem wiersza).

8. Negacja:

[^znak (znaki)] – znak następujący po symbolu negacji nie może wystąpić w dopasowywanym wyrażeniu.

Przykład:

```
tyt[^(an)]
```

Działanie: wyrażenie nie dopasuje ciągów znaków typu: tytaniczny, tytan, dopasuje zaś ciąg znaków inwestytura (co ciekawe, wyrażenie to nie dopasuje ciągów znaków typu apatyt$ (ogólniej: ciągów znaków, w których po sekwencji `tyt` nie występuje inny znak).

9. Granice wyrazu:

\< oraz \> – dopasowują odpowiednio początek i koniec wyrazu (ściślej – ciągu znaków alfanumerycznych). Te sekwencje metaznaków traktują łącznik jako granicę wyrazów, zatem np. wyrazy pół-Polak, czarno-biały będą interpretowane jako dwa wyrazy.

Przykłady:

```
\<ga?nik
```

Działanie: dopasuje ciąg znaków gaźnik, gaźnikowy, ale zagajnik – nie.

```
ga?nik\>
```

Działanie: dopasuje zarówno ciąg znaków gaźnik, jak i zagajnik, ale nie gaźnikowy. Warto zauważyć, że wyrażenie takie nie dopasuje ciągu znaków gaźnik1. Ciągi tego typu mogą często występować w zbiorach tekstów, w których stosowano przypisy (w tekście oryginalnym – przed konwersją do pliku tekstowego – liczba oznaczająca przypis występowała jako indeks górny).

10. Odwołania wsteczne:

\(*ciąg znaków lub metaznaków*)\ – jest to niezwykle przydatna konstrukcja, jeśli w wyrażeniu regularnym chce się wykorzystać zmienne, czyli odwołać się do (ciągu) znaków dopasowanego wcześniej. Odwołanie do takiej sekwencji to ukośnik lewy i cyfra (od 1 do 9) oznaczająca numer kolejny odwołania (np. program TextPad obsługuje do dziewięciu odwołań w jednym wyrażeniu). Typ ten został wykorzystany w pracy do ekscerpcji różnych typów ciągów interesujących z punktu widzenia badań językoznawczych.

Przykład:

```
\(<\([a-z]\)[a-z]*_\1
```
[25]

Działanie: wyrażenie to wyszukuje znak literowy (pierwszy w ciągu), po którym następuje zero lub więcej znaków literowych (wyraz), spacja, a następnie znak literowy identyczny z pierwszym znakiem ciągu. Wyrażenie zatem dopasuje następujące ciągi: *biały b*, *kropka k* itd.

[25] Spacje w wyrażeniach regularnych będą w tekście umownie oznaczane podkreśleniem.

3.3 Zastępowanie wyrażeń

W badaniach językoznawczych podstawową kwestią nie jest jedynie wyszukanie (dopasowanie) wyrażenia, ale globalne (w całym tekście lub zbiorze tekstów) zaznaczenie w poddawanym analizie zbiorze tekstów wszystkich wystąpień znaków odpowiadających wyrażeniu regularnemu. Nie wystarczy tu jedynie zaznaczenie wiersza, w którym nastąpiło dopasowanie (funkcja programu TextPad). Bezwzględnie konieczne jest dokładne oznaczenie granic dopasowania, ponieważ w ten sposób możliwe jest późniejsze sporządzenie listy ekscerpcyjnej. Jeśli więc znaleziono skrót występujący w nawiasie, przed którym występuje jego rozwinięcie (patrz rozdział 4, punkt 3), na przykład *accelerator mass spectroscopy (AMS)*, konieczne jest oznaczenie w tekście granic dopasowanego wyrażenia.

3.3.1 Metaznaki zastępowania

1. Koniec wiersza:

\n – dopasowuje koniec wiersza. Ma podobne znaczenie, co metaznak $, jednak jest rozpoznawany także w wyrażeniach zastępujących. Korzystając z tej sekwencji metaznaków można w prosty sposób dokonywać konwersji między listami, których elementy oddzielone są znakami końca paragrafu, a wyliczeniami, w których elementy oddzielone są wybranym znakiem – np. przecinkami.

Przykłady:

Znajdź: `,_`

Zastąp: `\n`

Działanie: wyrażenie pozwala zamieniać wyliczenie, w którym elementy oddzielone są przecinkami, na listę, na przykład wyliczenie następujące:

acetylcholinesterases, *acron*, *acrosome*, *acylhomoserine*, *affinities*, *aminoglycosides*, *amphipaths*, *anaphylatoxin*, *aneuploidy*, *angiogenesis*, *angiotensin*, *anomalously*

zostanie zamienione na listę:

acetylcholinesterases
acron
acrosome
acylhomoserine
affinities
aminoglycosides
amphipaths
anaphylatoxin
aneuploidy
angiogenesis
angiotensin
anomalously

2. Całe dopasowane wyrażenie

& – wstawia cały ciąg znaków dopasowany przy użyciu wyrażenia regularnego.

Przykłady:

Znajdź: `<\wod(a|y|zie|ą|o)\>`

Zastąp: `%%&%%`

Działanie: znajduje wszystkie wystąpienia wyrazu *woda* we wszystkich przypadkach liczby pojedynczej i oznacza te wystąpienia (oznaczenie powinno zawierać znaki lub sekwencję znaków, która w żadnym miejscu badanego zbioru tekstów nie została użyta, po

zastosowaniu operacji zastąpienia będzie więc jednoznacznie określać wszystkie wystąpienia).

3. Odwołania wsteczne

Odwołania wsteczne, których składnia została opisana powyżej, pozwalają na wykonywanie operacji na tekście z użyciem zmiennych, tzn. na przykład zmiany kolejności ciągów znaków.

Przykład:

Znajdź: `\(\<[a-z]+\>\) \(\<[a-z]+\>\)`

Zamień: `\2 \1`

Działanie: wyrażenie to zastosowane w przypadku listy złożonej z sekwencji dwuwyrazowych (np. listy kolokacji) zamieni miejscami wyrazy we frazie.

3.4 Przykłady wyrażeń regularnych stosowanych w pracy

W tym punkcie rozdziału zamieszczono wyrażenia regularne własnego pomysłu, umożliwiające automatyzację procedur ekscerpcyjnych.

1. Wyszukiwanie na liście (ściślej: wyliczeniu, w którym elementy oddzielone są przecinkami) bigramów następujących po sobie wyrażeń różniących się występującymi na końcu bigramu literami s, es lub ‘s (odpowiadającymi w języku angielskim liczbie mnogiej lub dopełniaczowi saksońskiemu).

Znajdź: `\([a-z]+\)_\([a-z]+\),_\1_\2(s|es|'s|s')`

Działanie: na liście bigramów (wyliczeniu oddzielonym przecinkami) znajduje wszystkie bigramy, po których występuje liczba mnoga tegoż bigramu bądź dopełniacz saksoński (dotyczy to wyrazów języka angielskiego, oprócz wyjątków).

Zamień: `<\1 \2>`

Działanie: po zastosowaniu wyrażenia bigram w liczbie mnogiej jest usuwany, zaś bigram w liczbie pojedynczej oznaczany znakami mniejszości i większości, dzięki czemu łatwo taki element oddzielić od reszty tekstu.

2. Wyszukiwanie trigramów, po których w nawiasie występuje ich akronim (dokładniej – wyszukiwanie ciągów trzech wyrazów, po których w nawiasie okrągłym następują pierwsze litery poszczególnych wyrazów).

Znajdź:

`\([A-Z]\)[a-z]+[-_]\([A-Z]\)[a-z]+[-_]\([A-Z]\)[a-z]+_\(\1\2\3\)`

Działanie: wyrażenie dopasowuje ciągi trzech wyrazów oddzielonych spacjami lub łącznikami[26], po których następuje spacja i skrót w nawiasie okrągłym. Kolejne odwołania wsteczne dotyczą pierwszych liter poszczególnych wyrazów.

Zamień: `@@&}}`

Działanie: oznacza wyszukane frazy w sposób jednoznaczny, ułatwiając ich późniejsze wyodrębnienie.

Na przykład zdanie:

Documents relating to the safety of the proposed nuclear-waste repository at Yucca Mountain in Nevada may have been falsified, the US Department of Energy (DOE) revealed last week.

zostanie zastąpione następującym:

Documents relating to the safety of the proposed nuclear-waste repository at Yucca Mountain in Nevada may have been falsified, the US @@Department of Energy (DOE)}} revealed last week.

[26] Konieczne jest w wyrażeniu dopuszczenie występowania łączników, aby objąć wyszukiwaniem wyrażenia następujące: *arabinose-binding protein*, *carbide-derived carbon*, *caspase-activated deoxyribonuclease*.

4 Metody komputerowej ekscerpcji informacji językowej ze zbiorów tekstów

4.1 Podstawa materiałowa pracy

Dane wejściowe wykorzystane w badaniach to zbiór pełnych artykułów opublikowanych w czasopiśmie naukowym *Nature* w latach 1997–2005 (w sumie niemal 450 wydań) pobranych ze strony internetowej czasopisma w postaci plików .html (działy *Editorials*, *Research Highlights*, *News*, *News Features*, *Correspondence*, *News and Views*, *Brief Communications*, *Articles* i *Letters*). W tabeli 2 poniżej przedstawiono szczegółowe dane dotyczące poszczególnych roczników czasopisma.

To właśnie czasopismo wybrano, kierując się następującymi przesłankami:

1. *Nature* opisuje najnowsze dokonania w naukach przyrodniczych i medycynie ze szczególnym uwzględnieniem dyscyplin rozwijających się najszybciej, w których pojawia się wiele nowych koncepcji i pojęć interesujących z perspektywy językoznawczej (jednostki nowe nienotowane leksykograficznie, kolokacje)[27].

2. *Nature* jest jednym z cieszących się największym zaufaniem czasopism naukowych, mającym jedną z najwyższych miarę oddziaływania (ang. *impact factor*, IF; wartość dla tygodnika za rok 2004 wynosi 32,2);

[27] Analizę korpusu tekstów naukowych (86 tekstów, około 1 miliona słów) w dziedzinie językoznawstwa stosowanego podaje Coniam (2004).

publikuje artykuły poddawane surowemu procesowi recenzowania, co powinno gwarantować poprawność merytoryczną i językową treści.

3. Czasopismo ukazuje się co tydzień, co oznacza, że można zgromadzić znaczną ilość danych. Dane dostępne są w formacie .html; format ten stosunkowo łatwo przekształcić w format .txt, który doskonale nadaje się do dalszej analizy.

   Tabela 2. Charakterystyka badanego korpusu.

Rok	**Wielkość pliku [MB]**[28]	**Liczba słów, w tysiącach**[29]
1997[30]	17,89	2737
1998	30,59	4703
1999	29,47	4511
2000	35,28	5412
2001	37,40	5736
2002	33,00	5042
2003	32,57	4985
2004	36,56	5611
2005	39,39	6003
Suma	**292,15**	**44740**

4.1.1 Opracowywanie danych wejściowych

W celu przygotowania danych wejściowych przeprowadzono następujące operacje:

1. Łączenie plików .html (w pierwotnym zbiorze jeden plik .html odpowiadał pojedynczemu tekstowi) i utworzenie plików zbiorczych (wynikowych) obejmujących cały rocznik.

[28] W plikach tekstowych.
[29] W plikach tekstowych.
[30] Znacznie mniejsza objętość rocznika 1997 wynika z tego, że pewna liczba wydań tygodnika nie była dostępna w formacie .html, a zatem te wydania i teksty nie zostały włączone do zbioru.

2. Usuwanie z plików wynikowych wszystkich elementów powtarzających się, to znaczy innych niż treść artykułu (np. informacje o prawach autorskich, odnośniki znajdujące się poza treścią artykułu, struktura strony internetowej). Bibliografię pozostawiono bez zmian.

3. Usuwanie tagów html.

4. Przekształcenie dużych plików .html (jeden plik to jeden pełny rocznik) w pliki tekstowe (w sumie dziewięć plików).

4.1.2 Uzyskany korpus w świetle zaprezentowanej typologii korpusów

Korpus, który przygotowano za pomocą powyższej procedury, można scharakteryzować w sposób następujący w oparciu o typologię korpusów (por. rozdział 2, podrozdział 3):

1. Niereferencyjny i niemonitorujący; jakkolwiek rozpatrywany korpus ma charakter statyczny, to znaczy jego konstruowanie zostało zakończone, nie jest referencyjny, ponieważ nie jest reprezentatywny wobec całości języka, a nawet jednego z jego rejestrów (teksty naukowe); zawiera jedynie prasowe teksty naukowe z jednego źródła; nie jest to ponadto korpus monitorujący, chociaż dzięki podziałowi materiału na roczniki (8 roczników pełnych) można by prowadzić badania o charakterze diachronicznym.

2. Specjalistyczny – ponieważ jest zbiorem tekstów naukowych, niniejszy korpus zawiera niewątpliwie treści o charakterze specjalistycznym, a więc odzwierciedlającym pewien jedynie wycinek rzeczywistości językowej.

3. Pełnotekstowy – przy konstruowaniu korpusu nie zastosowano próbkowania. Zawiera on pełne teksty artykułów.

4. Języka pisanego – otrzymany korpus nie zawiera transkrypcji języka mówionego.

5. Jednojęzyczny – korpus jest zbiorem tekstów z czasopisma angielskojęzycznego (wydawanego w Wielkiej Brytanii).

6. Nieanotowany – uzyskany korpus zawiera wyłącznie tekst, bez dodatkowych informacji. Teksty wejściowe w formacie html zawierają informacje dotyczące położenia ciągu znaków (nagłówek, stopka, tytuł itd.), jednak w procesie obróbki korpusu są one usuwane i powstaje plik w formacie txt.

7. Synchroniczny – badany korpus obejmuje teksty ze stosunkowo krótkiego przedziału czasu, chociaż dzięki podziałowi na poszczególne roczniki mógłby zostać wykorzystany do badań o charakterze diachronicznym, na przykład frekwencji określonych wyrazów nienotowanych leksykograficznie.

4.2 Porównanie wybranych metod ekscerpcji jednostek nowych

W tej części rozprawy omówionych i porównanych zostanie kilka metod półautomatycznych, służących do wyszukiwania neologizmów w korpusach językowych (przy zastrzeżeniu, że pojęcie jednostki nowej (neologizmu) traktujemy nie lingwochronologizacyjnie, lecz agnonimicznie; jednostką nową jest wyraz nowy względem posiadanej bazy materiałowej). Zawierają one procedury automatyczne, ale jednocześnie wymagają udziału człowieka. Wszystkie te metody łączy koncepcja wyróżników, czyli cech tekstowych – związanych zarówno z leksykonem, jak i graficznych (interpunkcja), które poprzedzają (wyróżniki leksykalne) lub otwierają i zamykają[31] (wyróżniki interpunkcyjne)[32] wyodrębnianą frazę – których obecność wiąże się ze zwiększonym w porównaniu z innymi fragmentami tekstu występowaniem jednostek nowych. Ekscerpcję i porównanie przeprowadzono, korzystając z liczącego około 45 milionów wyrazów korpusu artykułów czasopisma naukowego *Nature*. Postulowane jednostki nowe wyodrębniono przy użyciu analizy morfologicznej oraz określono częstość ich występowania za pomocą wyszukiwarki internetowej Google (por. Brin, Page 1998). Rezultatem pracy jest lista 1000 jednostek nowych. Na podstawie wyników określono efektywność poszczególnych metod.

[31] Chafe (1998) określa odcinki wypowiedzi zawarte między znakami interpunkcyjnymi mianem „punctuation units" (jednostek interpunkcyjnych).

[32] Wykorzystanie znaków interpunkcyjnych w przetwarzaniu języka naturalnego omawiają Say i Akman (1997). W swojej analizie uwzględniają nie tylko znaki interpunkcyjne typu przecinek, średnik, dwukropek, kropka, kreska itd., ale także znaki końca akapitu, tabele, listy oraz zmianę czcionki (na przykład kursywa). Zwracają uwagę na znaczenie dyskursywne stosowanej interpunkcji.

4.2.1 Wprowadzenie

Obecnie, wraz z coraz szybszym rozwojem nauki i techniki pojawia się więcej nowych rzeczy i zjawisk wymagających nazwania, wskutek czego zasób słownictwa wzbogaca się o znaczną liczbę jednostek nowych. Pojawia się wobec tego konieczność ich odnotowania przez leksykografię. Ponieważ to język angielski stał się dominującym językiem nauki i techniki, co wyraża się także obfitością wprowadzanych do języka neologizmów, ten właśnie język został objęty badaniami.

Tradycyjna[33] metoda ekscerpcji jednostek nowych[34] (mówiąc ściślej: agnonimów, czyli jednostek nieznanych filtrowi, względem którego poszukuje się jednostek nowych) polega na ich wyszukiwaniu ręcznym. Odpowiednio wykwalifikowana osoba – najkorzystniej leksykograf – czyta tekst i wynotowuje nieznane jednostki leksykalne. Następnie weryfikuje się ich obecność w wybranych słownikach, sprawdza pisownię i znaczenie (ten etap charakteryzuje się stosunkowo dużą subiektywnością) i opracowuje listę jednostek nowych. Mimo że metoda ta w założeniu jest niezwykle precyzyjna[35] (w teorii wyszukane powinny zostać wszystkie nowe, tj. nieznane względem listy będącej podstawą filtrowania, jednostki leksykalne w tekście lub zbiorze tekstów), obciążona jest równocześnie wadami, ponieważ jest mało ekonomiczna, tj. mało wydajna – ilość publikowanych współcześnie tekstów jest tak duża, że ich dokładna analiza, a nawet jakakolwiek analiza manualna staje się niemożliwa.

Dlatego właśnie w badaniach ekscerpcyjnych pojawia się konieczność automatyzacji, czyli wprowadzenia metod komputerowych, które ułatwiłyby wyszukiwanie jednostek nowych, skróciłyby czas i ograniczyły udział człowieka do minimum.

[33] Por. *Wstęp* w *Słowniku języka polskiego* red. W. Doroszewski (1958 – 1969), por. Bańko 2001, Wierzchoń 2004.
[34] Por. Buttler 1962, 1993, Wawrzyńczyk 1994, 1999, Smółkowa 2001, Stoberski 1976.
[35] Por. wyniki ekscerpcji manualnej dla języka polskiego: Wawrzyńczyk 2000.

Automatyzacja badań leksykograficznych jest częściej dziełem inżynierów niż leksykografów. Najbardziej znaczne postępy w tej dziedzinie odnotowano w obszarach wyszukiwania kolokacji i korekty pisowni (Dias 2000, Golding, Schabes 1996[36]). Gries i Stefanowitsch (2004) przedstawili metodę, którą nazywają *analizą kolostrukturalną*, analizy dystrybucji alternacji, czyli par konstrukcji gramatycznych o podobnym, ale nieidentycznym znaczeniu, np. dwie formy czasu przyszłego w języku angielskim (*will* i *be going to*), użycie strony biernej i czynnej, użycie dopełniacza saksońskiego lub frazy przyimkowej, położenie dopełnień w zdaniu, konstrukcja *try to* wobec *try and*. Można ją wykorzystać do badania niewielkich różnic między konstrukcjami pozornie synonimicznymi oraz do analizy samego pojęcia alternacji.

Obecnie stosowane metody wykorzystują niezwykle złożony aparat matematyczny, który pozwala uzyskać listy kolokacji[37]. Te z kolei powinny zostać następnie przekazane leksykografom do dokładniejszej analizy, ale najczęściej ten etap jest niestety pomijany.

Jeśli jednak na uwadze mieć ekscerpcję neologizmów, analizy statystyczne są zdecydowanie mniej przydatne, ponieważ neologizmy są pojedynczymi i unikatowymi jednostkami leksykalnymi (w innym ujęciu – pojedynczymi słowami – *hapax legomenon*), zaś wszelkie dane dotyczące częstości ich występowania nie mają zastosowania, ponieważ neologizm interesujący z punktu widzenia leksykografa to taki, który może wystąpić zaledwie raz w tekście lub w zbiorze tekstów (korpusie), zaś jego istnienia i położenia nie sposób przewidzieć metodami matematycznymi.

W tym miejscu zamiast metod o charakterze matematycznym można wykorzystać metody językoznawcze. Cenne wskazówki i podstawy koncepcji wykorzystywanej w ekscerpcji sformułował na gruncie języka polskiego

[36] Opracowano metodę korekty błędów w pisowni, w wyniku których powstają wyrazy poprawne ortograficznie, jednak niepasujące do kontekstu (np. *peace* i *piece*, *quiet* i *quite*), a także niepoprawne ze względów gramatycznych (np. rozróżnienie między *among* i *between* oraz *amount* i *number*), por. Mays et al. 1991, Gale et al. 1993, Golding 1995.
[37] Por. Buczyński 2004, Moszczyński 2005

Chlebda (1991), który zauważył, że frazemy[38] występują wewnątrz cudzysłowów lub po pewnych zwrotach; przykłady dla języka polskiego to *tak zwany*, *jak to się mówi*. Założenia te rozwinął P. Wierzchoń i przełożył na zautomatyzowaną lub półautomatyczną (konieczne jest sprawdzenie listy wynikowej przez językoznawcę) metodę ekscerpcji jednostek nowych. Podstawowe założenie metody to badanie sąsiedztwa jednostki leksykalnej oznaczonej uprzednio jako jednostka nowa. Według jednego założenia jednostki nowe występują (lub też: występują z większą częstością) wewnątrz cudzysłowów. Oznacza to, że frazy ograniczone znakami cudzysłowu mogą zostać wykorzystane jako dane wejściowe w późniejszej analizie (metodę przestawiono szczegółowo w pracach Wierzchoń 2003, 2005a). Koncepcja druga opiera się na jednostkach leksykalnych często poprzedzających jednostki nowe, które są specyficzne dla danego języka; przykład dla języka angielskiego to zwrot *so-called*. Podstawy teoretyczne i wskazówki praktyczne dotyczące wyodrębniania takich fraz – ze szczególnym uwzględnieniem języków fleksyjnych[39], na przykład polskiego, co wymaga obszernego zastosowania wyrażeń regularnych – przedstawiono w pracy Wierzchoń 2002.

Oba typy metod zakładają wykorzystanie analizy morfologicznej[40], która ponadto może być zastosowana jako samodzielny sposób ekscerpcji jednostek nowych. Można bowiem poddać analizie morfologicznej cały tekst (zbiór tekstów), takie jednak podejście wymaga operowania dużym, w żaden sposób niezredukowanym plikiem.

Celem przedstawionych badań jest weryfikacja przydatności powyższych metod do analizy angielskich tekstów naukowych opublikowanych na przestrzeni kilku minionych lat, które – dzięki rodzajowi tekstów (nauka i technika) – powinny (spodziewamy się tego intuicyjnie, tj.

[38] Por. definicję *frazemu*: „każdy znak językowy niezależnie od statusu semantyczngo i struktury formalnej, który stanowi nazwę potencjału treściowego (pojęcia), do którego odnosi się (wypowiada) mówca jako do symbolu względnie stałego" (Chlebda 1991: 27).

[39] Języki fleksyjne, czyli języki wykorzystujące formy fleksyjne zawierające afiksy (morfemy określające relacje gramatyczne, np. rodzaj, liczbę przypadek itd.; jeden afiks nierzadko reprezentuje więcej niż jedną kategorię gramatyczną).

[40] Analiza morfologiczna umożliwia automatyczne odrzucenie form znanych analizatorowi i w związku z tym niemających znaczenia dla ekscerpcji neologizmów.

jest to nasze założenie wstępne, czyli hipoteza ekscerpcyjna) zawierać znaczną liczbę jednostek nowych, porównanie poszczególnych metod i przedstawienie danych statystycznych dotyczących efektywności, a także automatyczna redukcja liczby słów w wykazie ostatecznym dzięki sprawdzeniu częstości ich występowania w Internecie przy użyciu wyszukiwarki. W ostatnim etapie możliwe jest wyodrębnienie słów występujących najrzadziej przez określenie frekwencji występowania każdego słowa poddawanego analizie w zbiorze stron zaindeksowanych przez wyszukiwarkę.

4.2.2 Założenia

4.2.2.1 Korpus

Korpus[41], czyli zbiór tekstów wykorzystywany jako dane wejściowe dla badanych metod, należy dobrać tak, by zmaksymalizować potencjalną liczbę podlegających ekscerpcji jednostek nowych, definiowanych jako jednostki leksykalne najprawdopodobniej niezarejestrowane dotąd przez leksykografów oraz nieodnotowane w słownikach, takich jak Oxford English Dictionary (http://www.oed.com) bądź Merriam Webster (http://www.merriam-webster.com/dictionary). Pojawiają się tutaj dwie użyteczne badawczo perspektywy językoznawcze: diachroniczna i synchroniczna. Można bowiem poddawać oglądowi teksty archaiczne – nawet jeśli wobec współczesnego języka byłyby one archaiczne – i najnowsze w związku z badaniem, rejestracją i analizą podanymi metodami występowania form nowych dla leksykografii. Z dwóch możliwych perspektyw strategia druga wydaje się zdecydowanie ciekawsza ze względów językoznawczych i ogólnych, ponieważ umożliwia obserwowanie najnowszych tendencji w rozwoju języka i rejestrowanie form nowych niemalże *in statu nascendi*. Ponadto analiza morfologiczna tekstów dawnych jest niezwykle trudna bez zastosowania analizatora morfologicznego dostosowanego do dawnej odmiany języka (analizator języka współczesnego

[41] Przyjęto w tym miejscu intuicyjną definicję korpusu jako zbioru tekstów.

oznaczyłby wiele, a nawet w przypadku tekstów wcześniejszych prawdopodobnie większość, wyrazów jako nierozpoznane).

Dlatego też zasoby poddawane ekscerpcji powinny spełniać następujące warunki:

1) obecność tekstów najnowszych, czyli niedawno opublikowanych, co zwiększa prawdopodobieństwo znalezienia wyrazów nieznanych leksykografii;

2) teksty, co do których można przypuszczać, że będą zawierały znaczną liczbę jednostek nowych (ściślej rzecz biorąc – agnonimów słownikowych); jednym z najbogatszych źródeł neologizmów wydają się czasopisma naukowe (z wyjątkiem być może tekstów literackich napisanych przez twórców posługujących się neologizmami; formy takie, jednak – przy zachowaniu wartości poznawczej – rzadko wchodzą do powszechnego użycia, a w związku z tym mają mniejszą wartość dla leksykografa, tj. oczywiście w kontekście konstruowania słowników typu: narodowych, referencyjnych, wielkich itp.), w szczególności zaś dotyczące nauk przyrodniczych i medycyny, czyli dyscyplin, w których obserwuje się największy i najszybszy postęp.

4.2.2.2 Wyszukiwanie fraz

Zagadnienie o znaczeniu podstawowym jest następujące: czy możliwe jest określenie jednostek graficznych (nieleksykalnych, czyli na przykład interpunkcyjnych), leksykalnych, czyli na przykład określonych zwrotów, lub innych konkretnych jednostek łatwych do wyodrębnienia z tekstu, które poprzedzają lub ograniczają frazy, w których jednostki nowe występują z częstością większą niż w innych partiach tekstu? Czy możliwa jest automatyzacja tego procesu? Metoda zaproponowana w tej części pracy opiera się na koncepcji wyróżników, w sąsiedztwie których jednostki nowe występują

z większą częstością. W znacznej części wyróżniki można uznać za niezależne od języka, to znaczy, nawet jeśli formy graficzne, czyli słowa, są charakterystyczne dla danego języka, swoiste formy powinny występować w każdym (lub w prawie każdym) języku. Oznacza to, że metodę można rozszerzyć na języki inne niż angielski, na podstawie którego przeprowadzono niniejsze badania.

Proponujemy wyróżnić dwa podstawowe typy wyróżników: leksykalne i interpunkcyjne. Wyróżniki leksykalne to frazy, które stosunkowo częściej lub – precyzyjniej – z większym prawdopodobieństwem poprzedzają jednostki nowe. W pewnym sensie zapowiadają więc one neologizmy albo inaczej – zapowiadają jednostki nowe i nieznane.

Ponieważ językiem, na podstawie którego prowadzono badania, jest angielski, wybrano następujące zwroty, dzięki którym możliwe będzie sprawdzenie tego, czy mają przewagę nad próbką pobraną losowo ze zbioru tekstów:

- *termed*,
- *called* (wraz z *so*(-)*called*),
- *known as*,
- *defined as*.

Wybór ten ma niewątpliwie charakter subiektywny. Opiera się głównie na naszej intuicji i znajomości języka angielskiego. Można by zapewne dodać szereg innych wyróżników, wydaje się jednak, że zasadnicze definiujące wyróżniki leksykalne zostały uwzględnione.

Drugi typ stanowią wyróżniki interpunkcyjne. Można się spodziewać, że znaki te będą ograniczać frazy, w których jednostki nowe występują z większą częstością. Wybrano dwa rodzaje znaków interpunkcyjnych: cudzysłowy pojedyncze (występujące w tekstach angielskich jako znaki ‘ oraz ’ i podwójne (występujące w tekstach angielskich jako znaki “ oraz ”).

W języku angielskim nie istnieją jednoznaczne reguły użycia obu rodzajów cudzysłowów (por. Bringhurst 2002). Sprawdzeniu podlega to, czy znaki te rzeczywiście zapowiadają jednostki nowe oraz który z nich pozwala wyodrębnić więcej jednostek nowych (jeśli tak, będzie możliwe pokazanie różnic i prawidłowości w ich użyciu).

4.2.2.3 Zastosowanie analizy morfologicznej

W badaniach językoznawczych analizę morfologiczną stosuje się zazwyczaj do lematyzacji słów występujących w tekście i przypisywania indeksów opisujących formę. Wykorzystuje się w tym celu analizatory morfologiczne[42] (por. też rozdział 1, punkt 5.4). W badaniach niniejszych skupiamy się jednak nie na tym, jaka konkretna forma została użyta w tekście, ale jedynie na tym, jakich form analizator nie rozpoznaje[43]. Jakkolwiek do grupy jednostek nierozpoznanych powinny należeć wszystkie jednostki nowe, to jednak nie wszystkie wyrazy nierozpoznane będą interesujące dla celów językoznawczych. Wynika to z tego, że możliwości analizatora są ograniczone (choćby ciągłym rozwojem języka) i najprawdopodobniej wiele wyrazów stosowanych często, od dłuższego czasu zarejestrowanych zostanie oznaczonych jako nierozpoznane. Stan ten wynika, oczywiście, z faktu etapowego rozwoju analizatorów, np. w odstępach kilkuletnich.

4.2.2.4 Wyodrębnianie jednostek występujących w języku z niewielką częstością

Zawartość listy słów wygenerowanej na etapie analizy morfologicznej (czyli słów nierozpoznanych) zależy nie tylko od danych wejściowych, ale

[42] Por. Bień, Szafran 2001.

[43] W badaniach wykorzystano analizator AOT dostępny na stronie internetowej www.aot.ru. Jest to program opracowany w ramach projektu Dialing, poświęconego automatycznej analizie języków rosyjskiego, angielskiego i niemieckiego. Program składa się z bibliotek (słowników) odpowiednich języków oraz modułu analizy tekstu.

także od zastosowanego analizatora morfologicznego. Dlatego też wiele słów oznaczonych jako nierozpoznane może nie przedstawiać większej wartości z punktu widzenia poszukiwania jednostek nowych; nawet bowiem jeśli nie są rozpoznane, nie mogą w żadnym razie zostać uznane za jednostki nowe (na przykład *baseline*, *assistive*, *apoptosis*, *adversely*, *luminance*, *decreasingly*, *cutoff*, *adenosine*, *artificially*, *angiogenesis*)[44]. Istnieje zatem konieczność znalezienia metody automatycznej, pozwalającej wybrać wyrazy o największej wartości dla badacza języka. Najlepiej, jeśli lista wynikowa będzie zawierała słowa uszeregowane według częstości występowania. W jaki sposób listę taką uzyskać? Wydaje się, że jedną z najbardziej rzetelnych metod – pomijając użycie korpusów zrównoważonych, na przykład *British National Corpus* – jest oszacowanie częstości występowania słowa z wykorzystaniem sieci Internet, a dokładniej – systemu indeksowania stron. Najbardziej przydatną w niniejszych badaniach funkcją takich systemów jest podawanie liczby wystąpień danego słowa (ogólnie: szukanej frazy) na zaindeksowanych stronach (dla każdego zapytania zwracana jest liczba jego wystąpień). Na przykład w wyszukiwarce Google z językiem angielskim (www.google.com) dla zapytania *angiogenesis* podawana jest informacja:

> Results 1-10 of about 1,840,000 for angiogenesis[45]

Jeśli językiem interfejsu jest język polski (www.google.pl):

> Wyniki 1-10 spośród około 1,840,000 dla zapytania angiogenesis[46]

[44] Przykład analizy dla słów baseline i apoptosis:
baseline 0 8 LLE aa SENT1 -?? BASELINE na -1 0
baseline 0 8 LLE aa SENT1 -?? BASELINE va -1 0
« 8 2 DEL EOLN
apoptosis 10 9 LLE aa CS? SENT2 -?? APOPTOSIS na -1 0

[45] Dla pierwszej strony z wynikami wyszukiwania.
[46] Dla pierwszej strony z wynikami wyszukiwania.

4.2.2.5 Plik *random*

Aby określić bezwzględną skuteczność każdego z badanych w niniejszej części pracy wyróżników – oraz sprawdzić, czy jego wykorzystanie daje korzyści w porównaniu z losowym doborem słów do celów ekscerpcyjnych (to znaczy analizą dobranej losowo części całego zbioru tekstów) – i odpowiadających jej wyróżników, przygotowano plik *random*, który stanowi próbkę całego zbioru tekstów. Stanowi on odniesienie, z którym porównywane są pozostałe pliki.

4.2.3 Metoda

4.2.3.1 Wyszukiwanie fraz

W przypadku poddawanych analizie leksykalnych wyróżników jednostek nowych (*termed*, *called*, *defined as* i *known as*) wyszukiwano i wyodrębniano frazy od wyróżnika do najbliższej kropki.

W ten sposób uzyskiwano listę postaci:

termed acclimatization to suggest a gradual biological adaptation of soil organisms to a higher temperature.

called siderophores, *which are used by bacteria to absorb iron*, *a nutrient they need to produce essential enzymes.*

defined as tannins, *also precipitate proteins and are perceived as astringent.*

known as telomeres, *typically found at chromosome ends*, *help normal cells keep tabs on how many times they have divided.*[47]

W przypadku z kolei wyróżników interpunkcyjnych (cudzysłowy pojedyncze i podwójne) wyodrębniano całą frazę ograniczoną znakami.

[47] Przykłady z rocznika 2005.

W ten sposób uzyskiwano listę postaci:

"There are a large number of bodies already doing this. We need to pull things together under a single umbrella,"

'instruments'

Przetwarzanie danych wykonano oddzielnie dla każdego rocznika, w wyniku czego uzyskano w sumie 55 plików (sześć różnych wyróżników dla dziewięciu roczników oraz plik *random*).

Pliki źródłowe poddano modyfikacji: między słowami a odnośnikami bibliograficznymi – to znaczy między ostatni znak ciągu znaków literowych a następującą po nim sekwencję znaków numerycznych – wstawiono jedną spację.

Wynik tej operacji był następujący (przykład):

Slow group velocity was measured recently in photonic crystal structures with ultrafast pulse propagation techniques 15, *16.*

W ten sposób spacje zostały umieszczone także we frazach typu *H5N1* (po wstawieniu spacji ciąg wygląda następująco: *H 5 N 1*), jednak ponieważ badanie swoim zakresem obejmuje jedynie jednostki leksykalne, zaś cyfry nie mają znaczenia, operacja ta nie wpłynęła negatywnie na jakość danych wejściowych.

Operacje wyszukiwania były w większości przypadków proste i realizowane za pomocą nieskomplikowanych wyrażeń regularnych. Jednym i istotnym wyjątkiem okazały się cudzysłowy pojedyncze, które są trudno rozróżnialne od apostrofów – w szczególności apostrofu występującego w dopełniaczu saksońskim[48].

[48] W przypadku rozróżnienia cudzysłowów pojedynczych od apostrofów:
1) wyróżniono w tekście znaki będące apostrofami ograniczone z obu stron znakami literowymi (apostrofy), na przykład: *isn't*, *won't*, *can't*, *master's*, tak by odróżnić je od cudzysłowów pojedynczych.
2) w przypadku dopełniacza saksońskiego liczby mnogiej (na przykład *masters'*) nie jest możliwe odróżnienie go od cudzysłowu pojedynczego.

Plik *random* utworzono w sposób następujący:

– dane wejściowe: rocznik 2005;

– wszystkie wiersze puste usunięto;

– wynikowa liczba wierszy: 94.000;

– co 1250. wiersz wyodrębniono do dalszej analizy[49]. W ten sposób otrzymano niesprawiającą trudności analitycznych próbkę (5461 słów) reprezentującą cały zbiór tekstów. Plik wynikowy przetwarzano tak, jak pozostałe pliki.

W tym etapie uzyskano następujące wyniki (przykłady z rocznika 2005):

1. termed:

termed extrinsic noise), *and the results suggested that some components of extrinsic noise affect gene expression in general 1 1.*
termed duplication shadowing, *suggests that loci near clusters of segmental duplication may be more susceptible to duplication deletion*, *probably due to an increased frequency of non allelic homologous recombination 1 8.*
termed lipoproteins are a major constituent of these extracellular compartments 3, *yet their role in CD 1 antigen presentation has not been investigated.*
termed slow dynamics 2 3, *meaning that the modulus slowly returns to equilibrium over several hours or even days after the wave energy has disappeared.*
termed the BHC or BRAF HDAC complex, *which is required for the repression of neuronal specific genes* 1.

2. Known as:

known as the particle.

known as ACE D, *makes more ACE than the other common version*, *ACE I.*

[49] Należy przyznać, że taki dobór wielkości próbki jest całkowicie intuicyjny i nie wynika z dążenia do uzyskania jej ścisłej statystycznie reprezentatywności wobec zbioru tekstów.

known as myogenesis, begins in transient blocks of tissue called somites.

known as British India may be a country for political purposes, but in no proper sense of the word do they constitute a nation.

known as brownian motion – also heralded a revolution in physical thought.

3. Defined as:

defined as being connected if any of their constituent nodes are linked.

defined as 2 metres or more in length) were discovered, starting in 1828 and ending in 1996.

defined as that of the normal yellow gene in D.

defined as tannins, also precipitate proteins and are perceived as astringent.

defined as being due to relatively stable changes in gene expression without changes in the DNA sequence of the gene.

4. Called:

called the Cubiculum of the Ocean.

called siderophores, which are used by bacteria to absorb iron, a nutrient they need to produce essential enzymes.

called AHLs, or N-acylhomoserine lactones.

called circumstellar disks, and to look at disks ranging from a million years old up to the age of the Sun is to look at the planetary construction process.

called the triple process.

5. Cudzysłowy pojedyncze:

'instruments'

'Lisbon objectives'

‘*junior Nobel prize*’

‘*biospherians*’

‘*Sistine Chapel of its time*’

6. Cudzysłowy podwójne:

“*With world attention focused on natural disasters*, *it is an idea that many people feel is ripe*,”

“*This needs to be put together now*,”

“*There are a large number of bodies already doing this. We need to pull things together under a single umbrella*,”

“*When scientific bodies tell governments what to do they get rejected*,”

“*Unless there is a supplemental appropriation*, *then the dollars pledged will definitely have to come out of current budgets and thus will compete with other needs*,”

Już „wzrokowa” analiza uzyskanych fraz ujawnia obecność interesujących, np. pod względem morfologicznym, jednostek leksykalnych, na przykład *biospherians* (dla wyróżnika cudzysłowy pojedyncze), *somites* (dla wyróżnika *known as*) i *acylhomoserine* (dla wyróżnika *called*).

W tabeli 3 poniżej przestawiono wyniki wyszukiwania fraz (suma dla wszystkich roczników).

Tabela 3. Wyniki sumaryczne dla wyszukiwania fraz

Wyróżnik	**Liczba wystąpień**	**Całkowita liczba słów**	**Średnia liczba słów na wystąpienie wyróżnika**
termed	1014	13685	13,50
called	8067	111977	13,88
defined as	1356	24649	18,18
known as	4178	59796	14,31
cudzysłowy pojedyncze	53949	95506	1,77
cudzysłowy podwójne	45457	610412	13,43

W przypadku cudzysłowów podwójnych uzyskano największą ilość danych (wyrazów) wynikowych, zaś w przypadku cudzysłowów pojedynczych – największą liczbę wystąpień oraz zdecydowanie najmniejszą długość frazy (1,77). Następstwa tego stanu rzeczy i jego wpływu na wyniki ekscerpcji zostaną omówione poniżej.

4.2.3.2 Analiza morfologiczna

Analiza morfologiczna jest jednym z dwóch zasadniczych etapów ekscerpcji jednostek nowych z plików otrzymanych w wyniku wyodrębniania fraz. Jej celem jest wyszukanie słów nierozpoznanych przez analizator morfologiczny, czyli oznaczonych jako nieznane. Ten etap powinien bardzo znacznie zmniejszyć liczbę analizowanych słów, ponieważ znaczna większość słów zostanie rozpoznana, a w związku z tym – zostanie wyłączona z dalszej analizy. Dane uzyskane w tym etapie stanowią ponadto materiał wykorzystywany w etapie kolejnym.

Pliki poddano niezbędnemu przetworzeniu (na przykład ukośniki, apostrofy i łączniki zamieniono na spacje)[50].

Przykład (dla wyróżnika *called*, rocznik 2005):

1) *called ACE*, *or angiotensin-converting enzyme.*

zamieniono na:

called ACE, *or angiotensin converting enzyme.*

2) *called a free-mass interferometer*, *has pretty much wiped the competition off the face of the Earth (the second half of the book).*

zamieniono na:

called a free mass interferometer, *has pretty much wiped the competition off the face of the Earth (the second half of the book).*

3) *called WSN/33*, *which was created in 1940 in a London lab.*

zamieniono na:

called WSN 33, *which was created in 1940 in a London lab.*

Tak przygotowane pliki poddano analizie morfologicznej. W oparciu o pliki wynikowe uzyskano listy słów nierozpoznanych przez analizator. Zastosowano analizator AOT[51].

[50] Wykorzystywany analizator morfologiczny traktuje frazy zawierające łączniki i ukośniki (na przykład *and/or* lub *much-disputed*) jako całe słowa i często oznacza je jako nieznane (przykłady: *risk-reduction* (*schemes*) lub *earthquake-prone* (*areas*)).

[51] Analizator można pobrać na stronie www.aot.ru, por. Sokirko 2004 (data pobrania: 12 listopada 2005).

Przykładowy wynik analizy morfologicznej dla frazy rozpoczynającej się wyróżnikiem *so-called*:

> so-called cardiospheres contain a mixture of cardiac cell types, and, when transplanted into mice after an acute heart attack, they formed vascular cells and heart muscle cells, albeit at a rather low frequency.

Wynik analizy morfologicznej[52]:

```
These                    0 5 LLE Aa NAM? SENT1 +?? THIS ed 104627 0
□                        5 1 DEL SPC
so-called                6 9 LLE aa +?? SO-CALLED aa 21451 0
□                        15 1 DEL SPC
cardiospheres            16 13 LLE aa -?? CARDIOSPHERE nb -1 0
 cardiospheres           16 13 LLE aa -?? CARDIOSPHERE vb -1 0
□                        29 1 DEL SPC
contain                  30 7 LLE aa +?? CONTAIN va 74920 0
□                        37 1 DEL SPC
a                        38 1 LLE aa +?? A xc 397 0
 a                       38 1 LLE aa +?? A xf 537 0
□                        39 1 DEL SPC
mixture                  40 7 LLE aa +?? MIXTURE na 208 0
□                        47 1 DEL SPC
of                       48 2 LLE aa +?? OF xc 486 0
□                        50 1 DEL SPC
cardiac                  51 7 LLE aa +?? CARDIAC aa 6485 0
 cardiac                 51 7 LLE aa +?? CARDIAC na 33226 0
□                        58 1 DEL SPC
cell                     59 4 LLE aa +?? CELL na 33696 0
 cell                    59 4 LLE aa +?? CELL va 74707 0
□                        63 1 DEL SPC
types                    64 5 LLE aa +?? TYPE nb 71106 0
 types                   64 5 LLE aa +?? TYPE vb 87209 0
,                        69 1 PUN
□                        70 1 DEL SPC
and                      71 3 LLE aa +?? AND xa 101456 0
,                        74 1 PUN
□                        75 1 DEL SPC
when                     76 4 LLE aa +?? WHEN xa 101505 0
 when                    76 4 LLE aa +?? WHEN ba 3495 0
 when                    76 4 LLE aa +?? WHEN na 72786 0
□                        80 1 DEL SPC
transplanted             81 12 LLE aa +?? TRANSPLANT vcvd 77987 0
□                        93 1 DEL SPC
into                     94 4 LLE aa +?? INTO xc 472 0
□                        98 1 DEL SPC
mice                     99 4 LLE aa +?? MOUSE nb 104047 0
□                        103 1 DEL SPC
after                    104 5 LLE aa +?? AFTER xa 101451 0
 after                   104 5 LLE aa +?? AFTER xc 405 0
```

[52] Zastosowano analizator AOT, por. przypis 38.

```
 after                    104 5 LLE aa +?? AFTER ba 619 0
 after                    104 5 LLE aa +?? AFTER aa 4028 0
 after                    104 5 LLE aa +?? AFTER na 27355 0
                          109 1 DEL SPC
an                        110 2 LLE aa +?? AN xa 101455 0
 an                       110 2 LLE aa +?? AN xf 538 0
                          112 1 DEL SPC
acute                     113 5 LLE aa +?? ACUTE na 27077 0
 acute                    113 5 LLE aa +?? ACUTE aa 96868 0
                          118 1 DEL SPC
heart                     119 5 LLE aa +?? HEART na 46306 0
 heart                    119 5 LLE aa +?? HEART va 75933 0
                          124 1 DEL SPC
attack                    125 6 LLE aa +?? ATTACK na 29319 0
 attack                   125 6 LLE aa +?? ATTACK va 74300 0
,                         131 1 PUN
                          132 1 DEL SPC
they                      133 4 LLE aa +?? THEY pl 104626 0
                          137 1 DEL SPC
formed                    138 6 LLE aa +?? FORMED aa 10968 0
 formed                   138 6 LLE aa +?? FORM vcvd 75696 0
                          144 1 DEL SPC
vascular                  145 8 LLE aa +?? VASCULAR aa 25792 0
                          153 1 DEL SPC
cells                     154 5 LLE aa +?? CELL nb 33696 0
 cells                    154 5 LLE aa +?? CELL vb 74707 0
                          159 1 DEL SPC
and                       160 3 LLE aa +?? AND xa 101456 0
                          163 1 DEL SPC
heart                     164 5 LLE aa +?? HEART na 46306 0
 heart                    164 5 LLE aa +?? HEART va 75933 0
                          169 1 DEL SPC
muscle                    170 6 LLE aa +?? MUSCLE na 53954 0
                          176 1 DEL SPC
cells                     177 5 LLE aa +?? CELL nb 33696 0
 cells                    177 5 LLE aa +?? CELL vb 74707 0
,                         182 1 PUN
                          183 1 DEL SPC
albeit                    184 6 LLE aa +?? ALBEIT xa 101452 0
                          190 1 DEL SPC
at                        191 2 LLE aa +?? AT xc 423 0
                          193 1 DEL SPC
a                         194 1 LLE aa +?? A xc 397 0
 a                        194 1 LLE aa +?? A xf 537 0
                          195 1 DEL SPC
rather                    196 6 LLE aa +?? RATHER ba 2751 0
 rather                   196 6 LLE aa +?? RATHE ab 96981 0
                          202 1 DEL SPC
low                       203 3 LLE aa +?? LOW ba 2165 0
 low                      203 3 LLE aa +?? LOW na 51459 0
 low                      203 3 LLE aa +?? LOW va 76358 0
 low                      203 3 LLE aa +?? LOW aa 102318 0
                          206 1 DEL SPC
frequency                 207 9 LLE aa +?? FREQUENCY na 92350 0

.                         216 1 PUN CS? SENT2
```

Analizator nie rozpoznał wyrazu *cardiosphere* (znacznik -??), chociaż prawidłowo oznaczył część mowy (rzeczownik).

Z plików wytworzonych przez analizator wyodrębniono wszystkie wyrazy oznaczone jako nierozpoznane (znacznik -??) i utworzono listy, które poddano dalszym operacjom:

1. Usunięto wszystkie słowa zawierające przynajmniej jedną wielką literę, aby w plikach nie występowały nazwiska, nazwy własne, akronimy itd.

   Przykłady usuniętych słów (wyróżnik *called*, rocznik 2005): Schrödinger, Sierpinski, ACE, Marangoni, Josephson, MAGE, microRNA, BH3).

   W tym etapie utracone zostały także wszystkie potencjalne jednostki nowe zawierające wielkie litery (na przykład występujące na początku zdania). Ponieważ jednak celem podstawowym jest automatyzacja analizy (zorientowana na osiągnięcie wysokiej wydajności), operacja ta wydaje się uzasadniona.

2. Wszystkie słowa składające się z jednej lub dwóch liter usunięto. W ten sposób w plikach wynikowych nie występują pojedyncze litery *s* pozostałe po zastąpieniu apostrofów spacjami. Operacja ta nie może w żaden sposób wpływać negatywnie na ekscerpcję jednostek nowych.

   Przykłady usuniętych słów: n (we frazie n-channel zamienionej na n channel), k (we frazie k-SAT zamienionej na k SAT).

3. Słowa na listach uszeregowano w kolejności alfabetycznej i usunięto powtórzenia.

Poniżej podano przykładowe wyniki uzyskane w tym etapie (przykłady z rocznika 2005):

1. Termed:

acetoxonium, *adenomatous*, *allelic*, *allodynia*, *anammox*, *antagomirs*, *antigenic*, *arteriogenesis*, *autoimmunity*, *biodiversity*, *blastospores*, *calmodulin*, *cannabinoid*, *catenin*, *chemicurrent*, *conpats*, *copaxone*, *cyclin*, *deimination*, *doxycycline*, *ecogenomics*, *enteropneusts*, *epistasis*, *et*, *eukaryotic*, *exchanger*, *extracellular*, *fru*, *genomic*, *genomics*, *glatiramer*, *haplotype*, *hemichordates*, *hemifusion*, *idiomorphs*, *inducible*, *interswitch*, *kinase*, *lipoproteins*, *lophophore*, *lymphoid*, *magnetoelectrics*, *megakaryocytes*, *metabolator*, *metagenomics*, *micromanipulation*, *minisleep*, *mitochondrial*, *mns*, *molecularly…*

2. Called:

acetylcholinesterases, *acron*, *acrosome*, *acylhomoserine*, *affinities*, *aminoglycosides*, *amphipaths*, *anaphylatoxin*, *aneuploidy*, *angiogenesis*, *angiotensin*, *anomalously*, *antiporters*, *apolipoprotein*, *appressorium*, *aquaporin*, *arbuscular*, *aren*, *argosomes*, *artesunate*, *astrocytes*, *autoimmunity*, *autoionization*, *barcoding*, *bedforms*, *benztropine*, *bevacizumab*, *biotech*, *biseparable*, *blastocyst*, *blastocysts*, *boreoeutherian*, *brainstem*, *branes*, *bromodomain*, *businesses*, *calmodulin*, *cardiospheres*, *cathodoluminescence*, *ceftazidime*, *cephala*, *chamosite…*

3. Defined as:

blastocoel, *cutoff*, *cyclin*, *decreasingly*, *decribed*, *defensin*, *desmethyl*, *distally*, *euthanization*, *fluence*, *hindcasts*, *hopane*, *ischaemic*, *kilobase*, *luminance*, *mainshock*, *mers*, *min*, *myocyte*, *nanotech*, *non*, *normally*, *orthologue*, *pixel*, *positionstart*, *positionsteady*, *postsynaptic*, *pre*, *predominantly*, *qualitatively*,

seroconversion, *shuttings*, *tastant*, *tetramer*, *timestart*, *timesteady*, *transcriptional*, *utc*, *viraemia.*

4. Known as:

achiral, *adversely*, *allorecognition*, *anammox*, *androdioecy*, *angiogenesis*, *anthracotheres*, *antiepileptic*, *antigenic*, *apoptosis*, *apoptotic*, *arbuscular*, *archaea*, *arguably*, *arogyapacha*, *aspergillus*, *assistive*, *astrocyte*, *backarc*, *betalains*, *bevacizumab*, *biodiversity*, *biomolecular*, *biosynthetically*, *blastocyst*, *bonannione*, *bonobo*, *brevetoxins*, *brevis*, *buffelgrass*, *cardioprotective…*

5. Cudzysłowy pojedyncze:

acetoxonium, *achiral*, *acron*, *actin*, *adakite*, *adaptationism*, *addback*, *adenosine*, *adipocrine*, *adipokines*, *aflatoxin*, *afterglows*, *aggrecanase*, *aggrecanases*, *aldol*, *allostatic*, *analyte*, *anammox*, *angiogenesis*, *angiogenic*, *anomalously*, *antagomirs*, *anthropodenial*, *antibias*, *antigenic*, *apo*, *apoptosis*, *apoptotic*, *apsidal*, *archived*, *artisanal*, *aubotsy*, *autoantigens*, *autotoxicus*, *barcodes*, *baseline*, *bedform*, *bedrest*, *biconical…*

6. Cudzysłowy podwójne:

abstr, *abundantly*, *accretionary*, *actin*, *adenosine*, *administratively*, *aediculatus*, *afferents*, *airsurfers*, *albicans*, *alloimmunity*, *aminobutyric*, *amplicons*, *amu*, *anarcho*, *ands*, *angiogenesis*, *angiogenic*, *angiopoietin*, *antagomirs*, *anticancer*, *anticontributive*, *antiferromagnet*, *antivivisectionists*, *antonin*, *aoml*, *apoptosis*, *apoptotic*, *apos*, *aquaculture*, *archaeal*, *archaeon*, *archivable*, *arcsec*, *artesunate*, *artificially*, *arxiv*, *astrocyte*, *astrocytes*, *astrocytogenesis*, *astrometric*, *atonia*,

auflosung, *autosomal*, *axonal*, *bacterioplankton*, *bandgap*, *barcode*, *barcoding*, *barite*, *baroclinic…*

Zestawienie liczbowe wyników tego etapu podano w tabeli 4 (dane sumaryczne dla wszystkich roczników).

Tabela 4. Wyniki analizy morfologicznej

Wyróżnik	**Liczba unikatowych słów oznaczonych jako nierozpoznane**
termed	528
called	2614
defined as	364
known as	1614
cudzysłowy pojedyncze	3915
cudzysłowy podwójne	3530
Suma	**12565**

Z tabeli wynika, że liczba słów została znacząco zredukowana w porównaniu do etapu poprzedniego – w szczególności w przypadku cudzysłowów podwójnych.

4.2.3.3 Wyodrębnianie jednostek leksykalnych o niskiej frekwencji

Ponieważ celem badań nie jest samo wyodrębnienie jednostek nowych, rozumianych jako jednostki nowe względem danej granicy datacji, ale – najkorzystniej – jednostek nieodnotowanych przez wybrane leksykony, do oszacowania, co do których jednostek można się spodziewać, że są nowe dla języka i leksykografii, wykorzystano zasoby internetowe. Ocenę oparto na

częstości występowania słowa (im mniejsza częstość, tym większe jest prawdopodobieństwo, że danej jednostki dotąd nie zarejestrowano). W tym celu wykorzystano wyszukiwarkę internetową Google[53], która przypisuje liczbę wystąpień danej jednostki na zaindeksowanych stronach internetowych. Uzyskano dane w następującym formacie: [liczba wystąpień] [słowo].

Listę wynikową poddano następnie sortowaniu według liczby wystąpień. Analizę przy użyciu wyszukiwarki Google przeprowadzono w marcu 2006 r. Indeksowanie stron internetowych jest procesem ciągłym, dlatego też dokładna liczba podanych wystąpień zależy od daty poszukiwania (a także miejsca, skąd wywołano wyszukiwanie).

Przykład dla wyróżnika *known as* z rocznika 1997 (liczba wystąpień na dzień 15 marca 2006 r.):

[52] [*asioryctitheres*]

[210] [*zalambdalestid*]

[385] [*adenotin*]

[644] [*epicathechin*]

[512] [*lucibufagins*]

[870] [*epigallocathechin*]

[984] [*cathechins*]

[856] [*epipubic*]

[53] Wyszukiwarkę tę wykorzystali także na przykład Keller i Lapata (2003) do badania częstotliwości występowania bigramów (przymiotnik-rzeczownik, rzeczownik-rzeczownik i czasownik-dopełnienie) niewystępujących w dostępnym korpusie. Stwierdzili wysoką korelację między częstotliwością określoną na podstawie wyszukiwarki i korpusu.

Jeśli chodzi o liczbę wystąpień, którą przyjęto za wartość progową pozwalającą uznać jednostkę za interesującą, a w związku z niewielką frekwencją – prawdopodobny neologizm – założono wartość 1000. Umożliwia to znaczne ograniczenie liczby jednostek na listach, a zarazem wyodrębnienie jednostek pojawiających się najrzadziej. Wartość ta została ustalona arbitralnie, jednak w sposób ostrożny (zamierzeniem było uniknięcie zawyżania liczby wyodrębnionych jednostek), a powstała lista jest niezwykle interesująca dla dalszych badań morfologicznych, w wielu bowiem przypadkach wyodrębnione jednostki są formami pochodnymi (derywatami).

Przykłady: *epipubic*, *palaeoceanographers*, *homeodynamic*, *pseudomedical*, *presqualene*, *governmentalist*, *megabillion*, *arthropodization*, *oligomolecular*, *supervolatile*, *superministry.*

Zestawienie wyników uzyskanych w tym etapie podano w tabeli 4; por. także wyniki ostateczne otrzymane po kolejnym etapie.

Przykładowa liczba wystąpień jednostek, które w marcu 2006 r. występowały z frekwencją mniejszą niż 1000 – dla wyróżnika *known as* z rocznika 1997 (w nawiasie podano aktualną liczbę wystąpień na dzień 8 stycznia 2008 r.):

asioryctitheres (55), *zalambdalestid* (227), *adenotin* (469), *epicathechin* (754), *lucibufagins* (625), *epigallocathechin* (1550), *cathechins* (1200), *epipubic* (2640).

Z przykładu tego wynika, że w ciągu niecałych dwóch lat między pierwotnym i ponownym badaniem frekwencji wybranych jednostek za pomocą wyszukiwarki nastąpiły pewne zmiany frekwencji.

4.2.3.4 Analiza manualna list słów

Mimo że wszystkie poprzednie etapy charakteryzował znaczny stopień automatyzacji, ostateczna analiza list słów musi zostać przeprowadzona przez wykwalifikowanego badacza lub rodzimego użytkownika języka. Tak więc przedstawiona metoda optymalizacji pozyskiwania materiału empirycznego zawiera w sobie oczywiście komponent analizy manualnej. Konieczne jest bowiem wyodrębnienie wszystkich przypadków błędów typograficznych, wyrazów jednoznacznie obcych (czyli nieuznanych za włączone do danego języka) oraz form jawnie okazjonalnych (tj. np. takich, które stosowane są jednorazowo, na przykład jako transkrypcja błędnej wymowy lub przejęzyczenia). Poniżej podano przykłady jednostek, co do których uznano, że należą do powyższej kategorii, i dlatego usunięto je z list ostatecznych:

- błędy typograficzne: *theoreticalpractice*, *expertiseof*, *dignityof*, *appealedto*, *humandignity*, *believesit*, *organizedaround*, *whichgovern*, *marginalenvironmental*, *alwaysalso*, *preposterousconclusions*, *beethically*, *connectivetissue*, *findingsis*, *directedtothe*, *supranationalinstitute*, *ofcompounds*,

- informacje objaśniające struktury gramatyczne występujące w innych językach: *mouseeats*, *mousegoes;*

- symbole (często pisane z użyciem indeksów górnych lub dolnych): *uvcalc*, *fstim*, *sqtz*, *ndisl*,

- wyrazy obce: *shuvuu*, *sötted*, *génopôles*, *entwicklungsbiologischer*, *betsika*, *augebitur*, *pertransibunt*, *yanzigou*, *zerstückte*, *subitaneis*, *iigiracóobitooree*, *iigiracóobiwareec*, *sorokinii*, *maúudabi*, *semicelatum*, *biostratigraphische*, *maagaríshdawacee*, *mosbachensis*, *carnegii*, *wiáha*, *macée*,

- formy archaiczne: *burlesqt*, *heareing*, *equalle*,

- formy efemeryczne: *kvestion*, *vhatever*, *physicizts*, *discuzzed*, *failurez*.

4.2.4 Wynik badań

Jednostki leksykalne wymienione poniżej to ostateczne wyniki zastosowania zaproponowanych metod z podziałem na wyróżniki i roczniki czasopisma.

1997

called

trochleated, *rosettins*, *tagamites*, *mertensian*, *palaeoceanographers*, *magnetostrophic*, *lepidotrichia*, *orviétan*, *prosaccades*, *australopiths;*

cudzysłowy podwójne

prowbley, *naturify*, *anandamidergic*, *polyphyrin*, *commaform*, *superpenumbral*, *uneconomy*, *recorrecting*, *cathechins*, *homeodynamic*, *pseudomedical*, *thatled*, *megabillion*, *radioastronomers*, *presqualene*, *governmentalist;*

cudzysłowy pojedyncze

holotheres, *synstorm*, *klinorhynchy*, *arthropodization*, *superdimer*, *oligomolecular*, *mertensian*, *eupantotheres*, *sigmage*, *supervolatile*, *megapore*, *paranotal*, *parareptiles*, *neurophilosophers*, *autapse*, *transducisome*, *copulators*, *superministry*, *medusoids*, *balloncraft*, *framboid*, *connectoplasm*, *bonebeds*, *micromoulding*, *monophagy*, *nanofossils;*

defined as

vaverage;

known as

asioryctitheres, *zalambdalestid*, *adenotin*, *epicathechin*, *lucibufagins*, *epigallocathechin*, *cathechins*, *epipubic;*

termed

presqualene;

1998

called

cyproase, *bifurcationists*, *aplanktonic*, *parapsid*, *hexabrachions*, *waiverers*, *haemoglobinase*, *osteolepiforms*, *micropolygyria*, *rhipidistians*, *lexitropsins*, *polyamorphism*, *aseismically;*

cudzysłowy podwójne

palaeopenetrometers, *axoniform*, *dehomologation*, *diskoseismic*, *biotolerance*, *megamullions*, *megamullion*, *systematicists*, *spiritdom*, *outreproduce*, *preformationists*, *radioastronomers*, *chromatosome*, *misappliance*, *scapulocoracoid;*

cudzysłowy pojedyncze

macrosyllable, *piezonail*, *altoradiometer*, *ennobelled*, *diplosyllables*, *lepiform*, *platigem*, *preprismatic*, *tachopause*, *mutasomal*, *pseudodating*, *bifurcationists*, *sapromyiophily*, *aplanktonic*, *necrolab*, *cuspier*, *parapsid*, *polynail*, *precompensates*, *cephalobid*, *sphingophily*, *neontologist*, *cavicapture*, *neoglycopolymers*, *megabeds*, *gastrophysics*, *pseudodate*, *regularist*, *megaturbidite*, *conciliance*, *megaturbidites*, *lettersound*, *synneusis*, *baryometer*, *osteolepiforms*, *tetherable*, *lepospondyl*, *palaeothermometer*, *rhamphorhynchoid*, *osteolepiform*, *pyracylene*, *vendobiont*, *tunnelized*, *superplanets*, *megachannels*, *polyamorphic*, *osteolepiformes*, *downwelled*, *transducisome*, *bakedness*, *pinscape*, *anthracosaurs*, *segnosaurs*,

actinosporean, *biofluiddynamics*, *etchability*, *echeme*, *ornithophily*, *rothomagensis*, *intrasteric*, *exflagellation*, *kleptoparasites;*

defined as

known as

loxommatids, *baphetids*, *friarbirds;*

termed

megabeds, *megaturbidites*, *megaplumes*, *exflagellation;*

1999

called

dygments, *pseudocopulations*, *isochelae*, *antishocks*, *dimycocerosate*, *gabbronorites*, *protohaem*, *formatrix*, *pyrobitumen*, *haemangioblastomas;*

cudzysłowy podwójne

scattercirrus, *fallnimbus*, *ascendstratus*, *foldedcumulus*, *reindigenizing*, *upcruitment*, *tribosphenid*, *gelbrain*, *mimeomorphic*, *photofootprint*, *zooblot*, *mesdemet*, *androgynization*, *ultracivilized*, *palaeohydrogeology*, *adamantoid*, *selfplex*, *subitize*, *decruitment*, *transgenomics*, *formatrix*, *volvelles*, *unmixedness*, *pasteurien;*

cudzysłowy pojedyncze

dygments, *pongidized*, *palaeolandslide*, *pathophage*, *neurotrophinergic*, *reindigenizing*, *slabology*, *geohopanes*, *yellowfix*, *polyplocodont*, *microdermic*, *tidalists*, *micromastic*, *untransfectable*, *asthenoliths*, *hypobradytely*, *protolarva*, *regressins*, *chaincloth*, *palaeoamericans*, *rollertube*, *pellatron*, *pseudocopulations*, *petrophage*, *highconic*, *embryonization*, *superwells*, *operomics*, *pseudized*, *rhamphorhynchoids*, *legness*, *antishocks*, *crosspriming*, *comodulated*, *mudsplashes*, *segnosaurs*, *demultiply*, *superswells*, *anthophyte*,

beerstone, *prosegments*, *inchworming*, *mudsplash*, *triconodont*, *cornealis*, *cratonization*, *deoxygenase*, *uninodal*, *miniwar*, *superalliance*, *gabbronorites*, *pseudosubstrates*, *polyamorphism*, *eupyrene*, *pinchase*, *ecdysozoan*, *pasteurien;*

defined as
trenchward;

known as
tidalists, *porolepiforms*, *gongylidia*, *aspartases*, *clathrasils*, *eutely*, *syntrophs*, *mycetocytes;*

termed
methyldiamantane, *prosegments*, *pseudolysogeny;*

2000

called
nitroimidazofurans, *chemotrophism*, *bisporphyrinate*, *mnemiopsin*, *berovin*, *photoresolution*, *mitrocomin*, *nitroimidazopyrans*, *phialidin*, *solardomes*, *ventists*, *haemangioblasts*, *palaetiological*, *presenilinase*, *fimbral*, *enterohaemolysin*, *unicolonial*, *magnetochiral*, *specillum;*

cudzysłowy podwójne
dyschromatopsics, *superannalist*, *malaricissima*, *intrabranally*, *nanostencilled*, *telanthropoids*, *dysmenorrhaea*, *arttaste*, *astrometers*, *pocilloporins*, *pagodane*, *ventists*, *chailer*, *supermeme*, *palaeobiologists*, *chailing*, *perflation*, *radiocollar*, *foistered*, *hydrinos*, *unrenounceable*, *macroelectronics;*

cudzysłowy pojedyncze
supersupershifted, *micrometozoan*, *staygreens*, *membrasome*, *intrabranally*, *physiopole*, *cosmuck*, *interdimers*, *antigeroid*, *scientaria*, *unfaults*, *postfus*,

electrofoil, *secretasome*, *telonomic*, *lymphapophyses*, *amphidromies*, *rabbitized*, *uncopying*, *secretosome*, *peristriatal*, *abgerminal*, *attolitre*, *morphodynamically*, *microbiography*, *superconvection*, *cyclosynchrotron*, *unfoldases*, *misprojections*, *supergreenhouse*, *commodifaction*, *nanoprojects*, *antitestis*, *antimimetic*, *neurorobotic*, *garbenschiefer*, *genocopy*, *rhamphorhynchoid*, *solardomes*, *transcriptosomes*, *osteoimmunology*, *chailed*, *maxizyme*, *retrohoming*, *magnetochiral*, *paralemniscal*, *tripledecker*, *equilibriation*, *garnetite*, *centauron*, *nanorover*, *quarktet*, *unfoldase*, *syntrophy*, *chaostory*, *axonopathies;*

defined as

backlabelled, *interprotomer;*

known as

malaricissima, *lamillipodia*, *haemangioblasts*, *retrohoming*, *specillum*, *equilibriation;*

termed

euagaric, *thelephoroid*, *pseudopupil*, *repressilator*, *allospecificity*, *syntrophy;*

2001

called

vermilarva, *oomicides*, *archidynamic*, *exterlibral*, *coordinometer*, *trichromator*, *gonialblasts*, *cerebrotype*, *hippopotamid*, *ornithurines*, *mediatophore*, *ferropericlase*, *paranematic*, *mitosome*, *abembryonic*, *chargons*, *elaiosomes*, *chargon*, *volicitin*, *paramagnons;*

cudzysłowy podwójne

mathematicability, *vermilarva*, *tribosphenidans*, *unrecognizedly*, *nitrosohydroxylamines*, *eupantotheres*, *supersog*, *diskoseismic*, *nonhumanistic*, *energeticists*, *canaliform*, *cichlidiots*, *symmetrodonts*, *anthropodenial*,

symmetrodont, *phytoanticipins*, *phenologists*, *turnipy*, *ornithurine*, *unsealable*, *superprotonic*, *unakin*, *palaeodata*, *fossilists*, *alkaptonuric*, *formatrix;*

cudzysłowy pojedyncze

palaeosterilization, *oomicides*, *ovasomagenesis*, *tribotheres*, *prefacilitate*, *ascohymenials*, *extracleithrum*, *ormiaphones*, *mutasomal*, *exosemiconductors*, *heteropolyblues*, *metastatistic*, *electrosomatic*, *baritization*, *nitrosoreductase*, *superphyletic*, *ovumsum*, *megabasins*, *leafpoints*, *transducisomes*, *eupantotheres*, *megabasin*, *supersog*, *cryptidin*, *cerebrotypes*, *piscidins*, *zhelestids*, *metasaccharinic*, *gliriform*, *guardees*, *quasicondensates*, *dealloyed*, *shrimpoluminescence*, *crownward*, *optipulse*, *crystallomics*, *polypodiaceous*, *chronotherapies*, *tetrahedrality*, *argosomes*, *electrolamp*, *orbitons*, *aurophilicity*, *endovanilloids*, *ovasome*, *rumbleometer*, *crosspriming*, *mudsplashes*, *signalplex*, *transducisome*, *glucolipotoxicity*, *mitosome*, *automone*, *superministry*, *polyplacophoran*, *microislands*, *abembryonic*, *guardee*, *deoxygenase*, *superhybrid*, *resublime*, *extremozymes*, *elaiosomes*, *chargons*, *cytonemes*, *calfuse*, *saccharinic*, *chargon*, *pinchase*, *slushball*, *preformationist*, *condylarths*, *tectosphere*, *immunoediting;*

defined as

ultradivided, *zonasulcate;*

known as

homomeroquinene, *tribosphenidans*, *asomatopagnosia*, *superbrownian*, *flexoelectricity;*

termed

cataflexi, *hyperflexistyled*, *exosemiconductors*, *cytonemes;*

2002

called

ethesiometer, *controlniks*, *undecouplable*, *complexomics*, *prodiginine*, *ladderanes*, *trimethylmethoxysilane*, *gamergates*, *oligopyrrole*, *photoheterotrophy*, *mitosome*, *nanowhisker*, *thermochromatography*, *lipochitin;*

cudzysłowy podwójne

zeppelinoids, *unchemist*, *chemoinformatician*, *cerebrotype*, *paramilitia*, *autocreative*, *chromicized*, *obsessionism*, *microlepton*, *baryometer*, *brainspinning*, *aerosolizable*, *photoheterotrophy*, *anthropozoic*, *brainedness*, *anammoxidans*, *electrions*, *biofraud*;

cudzysłowy pojedyncze

protoanthracosaurs, *muonionization*, *palaeophosphatometry*, *bradyfauna*, *equivalogue*, *phosphatometer*, *transloxer*, *pseudohaemolysin*, *controlniks*, *avnosmia*, *undecouplable*, *coelibactin*, *spexels*, *supercorrelation*, *hoxology*, *unifolds*, *complexomics*, *stromatoloids*, *coelichelin*, *printspeak*, *peytoia*, *microchimaerism*, *antiuricosuric*, *microleptons*, *nanothermometers*, *resulphurized*, *panchabhoota*, *vaccinomics*, *pseudodimer*, *baryometer*, *hyperscanning*, *distalized*, *lepospondyl*, *pseudomagnetic*, *dihapto*, *photoaptamer*, *osteodontokeratic*, *superchemistry*, *ladderane*, *divisome*, *anammoxidans*, *anthracosaurs*, *timesome*, *triconodonts*, *unclumped*, *inchworming*, *aftercontractions*, *nanothermometer*, *biomotors*, *dispersalist*, *megaregolith*, *repressilator*, *metacarbonate*, *transdifferentiating*, *paramorphs*, *slushball*, *tomographer;*

defined as

ladderane, *tailbeat;*

known as

flexibilatis, *somatoaxon*, *cornutes*, *mitrates*, *lamellipods*, *minifilaments*, *muscivorus*, *pericentromere;*

termed

pseudohaemolysin, *epiparasites*, *fluorophosphine;*

2003

called

homodisciplinary, *aplanktonic*, *osteochondroprogenitor*, *lorisiforms*, *missionnum*, *overdeepenings*, *phaseonium*, *routinizable*, *nucleofilaments*, *neurocrystalline*, *microsyntenic*, *interglomerular*, *cytonemes*, *mitosomes*, *ladderane*, *enteropneusts*, *strepsirrhines*, *spectrosome*, *nanoprisms;*

cudzysłowy podwójne

nanolecture, *tracheals*, *retrovaccinology*, *exosymbiotic*, *eutelic*, *oresmen*, *vegetalization*, *supernebula*, *superswells*, *neurocrystalline*, *superinvar*, *intramers*, *immolative*, *chronophotograph*, *heliocentricism*, *nanolitres;*

cudzysłowy pojedyncze

hibbenisms, *colliculoreticulospinal*, *palaeopasteurization*, *chaperonology*, *heterodisciplinary*, *neohubbertarians*, *nanoharvesting*, *glottoclock*, *protoconoid*, *nanolouse*, *incommensuracy*, *scoopophobia*, *poppase*, *aplanktonic*, *rostralia*, *tauidion*, *eucentricity*, *isoindene*, *toothcombed*, *missionnum*, *duails*, *fluorobodies*, *dephosphins*, *superwetting*, *functionation*, *polyaxonal*, *overdeepenings*, *brodae*, *antiglass*, *supersegment*, *baroplastics*, *synaptotoxic*, *phaseonium*, *routinizable*, *supernebula*, *sonocytology*, *posteriorized*, *ubiquityl*, *subjunctional*, *paraspeckles*, *ladderane*, *xenoscience*, *preferers*, *phytometer*, *gammation*, *cuckolders*, *biobugs*, *phluorin*, *plantibody*, *dedifferentiates*, *lovespot*, *propiece*, *attophysics*, *degradome*, *microacoustics*, *cuckolder*, *retrotranslocated*, *gerontogenes*, *nanopod*, *nanocage*, *toplighting*, *meltback*, *pinchase*, *sulphenyl*, *bonebeds*, *regassed*, *probablys*, *microstreaming*, *tectosphere*, *systemicity;*

defined as

vinculinaggresomes, *opistodontians*, *sphenodonts*, *doliolaria;*

known as

anthracotheriid, *afrotheres*, *ambulacraria*, *vegetalization*, *glaciohydraulic*, *ladderanes*, *hypobranchials*, *hyperstriatal*, *doliolaria*, *lysenin*, *enteropneusts*, *malariotherapy*, *ceratobranchials;*

termed

polyaxonal, *baroplastics*, *cryptospores*, *paraspeckles*, *plastochrons*, *immolative*, *cuckolder;*

2004

called

polymorula, *quduties*, *thaxtomins*, *paranotal*, *chordamesodermal*, *palmitoylputrescine*, *lasetron*, *polyhook*, *undruggable*, *oosome*, *telepreventive*, *phosphoroamidite*, *heterotachy*, *magnetoelectrics*, *geopoetry*, *sageing*, *mitosomes*, *destabilase*, *intramers*, *haemangioblast*, *superlubricity*, *meristemoid*, *hanatoxin*, *batrachotoxins*, *monophagy;*

cudzysłowy podwójne

mouthwashology, *thermodramatics*, *gerontocractic*, *dodaersen*, *yeastification*, *tensegral*, *antivaccinationism*, *progressible*, *uninvitation*, *sunklands*, *concestor*, *geopoetry*, *coethnic*, *cryptofauna*, *multicontinental;*

cudzysłowy pojedyncze

innateosome, *opportunitroph*, *nanosalt*, *nanocrimelab*, *techniquities*, *apternodontids*, *brachylabic*, *nanosociology*, *superpostdocs*, *merotely*, *macrolabic*, *morphogeologic*, *hyperlethality*, *minifacets*, *degelled*, *diftox*, *multirhythm*, *yeastification*, *alcosols*, *rheoreversible*, *nonsolvers*, *metaethnic*, *foldhunter*, *camgaroos*, *hypolithon*, *hypoliths*, *countergradients*, *helixhunter*, *anoxicity*, *intervality*, *norhipposudoric*, *monoreceptor*, *nanaerobes*,

microdiverse, *overdifferentiated*, *hipposudoric*, *receptorology*, *boxological*, *mussid*, *photocaged*, *megaturbidite*, *amphitelic*, *anthropodenial*, *calcichordates*, *lasetron*, *hyperspecificity*, *demethylimination*, *ignorosphere*, *osteolepiforms*, *palaeothermometer*, *telepreventive*, *undruggable*, *sphenosuchian*, *macroelectronic*, *tilepath*, *uninvitation*, *softwiring*, *intramers*, *syntelic*, *sideground*, *sageing*, *prehairpin*, *antagomirs*, *altriciality*, *glassformers*, *faviids*, *cryptofauna*, *arctometatarsalian*, *fermionized*, *preorganize*, *polyheads*, *superlubricity*, *faviid*, *cainism*, *thunniform*, *ultracontigs*, *erectines*, *megamouse*, *nanisani*, *polyamorphism*, *exteins*, *deiminated*, *instructionist*, *myoseptum*, *syntrophy;*

defined as

polygerm, *hypolithon*, *hypoliths;*

known as

cantharidiphilous, *unhedgehog*, *cinctans*, *ctenocystoids*, *pintronic*, *tarsioids*, *homalozoans*, *trioxolanes*, *siderocalin*, *uterocalin*, *stylophorans*, *deuterostomy*, *deubiquitinase*, *coenocytes*, *perikymata*, *rifins;*

termed

merotely, *bradyopsia*, *transportins;*

2005

called

crenaters, *hydrosheds*, *infrabiological*, *thencas*, *marshballs*, *negadex*, *lophenteropneusts*, *boreoeutherian*, *cardiospheres*, *supersusceptibility*, *chronobiotics*, *argosomes*, *metalloligands*, *interchromosome*, *fruitcases*, *biseparable*, *cytonemes*, *lipopolymer*, *silicatein*, *amphipaths*, *pharyngobasilar*, *lysobisphosphatidic*, *rhopalium*, *pteroid*, *immunoediting*, *orexinergic*, *exoculata*, *trichoblast*, *geoneutrinos*, *transdetermination;*

cudzysłowy podwójne

anticontributive, *methalogical*, *fluxclimatology*, *airsurfers*, *plesiometacarpaler*, *immunobots*, *negadex*, *oxifiers*, *verticornis*, *urmetazoan*, *antagomirs*, *astrocytogenesis*, *quantitativity*, *nonconvecting*, *hemangiogenesis*, *proteorhodopsins*, *biocleaning*, *picobot*, *nontronites*, *cytonemes*, *femtotechnology*, *wiregrid*, *palagonitization*, *nepotistically*, *haemangioblast*, *transactivity*, *galaninergic*, *enteropneusts*, *gigaxonin*, *hydrinos*, *microcephalics*, *finized*, *palaeoanthropologists;*

cudzysłowy pojedyncze

aubotsy, *parasuperconducting*, *mutalecimes*, *nongouge*, *adipocrine*, *nutristad*, *hydrometropole*, *nanobaubles*, *hydrosheds*, *oscillophor*, *cotransin*, *infrabiological*, *pipmodulins*, *interologues*, *composome*, *neophrenological*, *pardoides*, *kilogirl*, *sequelog*, *premammilary*, *metabolator*, *negadex*, *lophenteropneusts*, *edentus*, *hyopsodontids*, *hyopsodontid*, *lithoheterotrophic*, *acetoxonium*, *sequelogue*, *superpigment*, *lophenteropneust*, *chemicurrent*, *pseudodimer*, *energeticists*, *nanogratings*, *antagomirs*, *supersusceptibility*, *anthropodenial*, *omovertebral*, *megamullion*, *clusteredness*, *brainprints*, *calcichordates*, *pseudoproxy*, *ladderane*, *superphyla*, *chromathography*, *magnetoelectrics*, *finitists*, *metanodes*, *prepatterns*, *cryovolcano*, *halleyan*, *blattellaquinone*, *repressilator*, *inchworming*, *downblended*, *dewetted*, *biospherians*, *destratified*, *postselected*, *aggrecanases*, *interologs*, *superspreading*, *condylarth*, *subfilaments*, *pterobranchs*, *lymphohaematopoietic*, *adakite*, *condylarths*, *superspreaders*, *pangenes*, *multiferroics*, *sequenceable*, *dehydrative*, *nanoreactor*, *tsunameters*, *triallelic*, *superstrains*, *stressmeter*, *segnosaur*, *drugable*, *superrotation*, *pseudospins;*

defined as

shuttings;

known as

bonannione, *stylocone*, *chelifores*, *strigolactones*, *prosensory*, *urmetazoan*, *transresistivity*, *arogyapacha*, *retrotranspose*, *unsynapsed*, *rhizophore*, *diacylglycerides*, *gigaxonin*, *anthracotheres*, *androdioecy;*

termed

wingbia, *metabolator*, *acetoxonium*, *chemicurrent*, *conpats*, *antagomirs*, *minisleep*, *magnetoelectrics*, *enteropneusts*, *idiomorphs*, *deimination*, *multiferroics*, *tracheoles.*

4.2.5 Analiza wyników

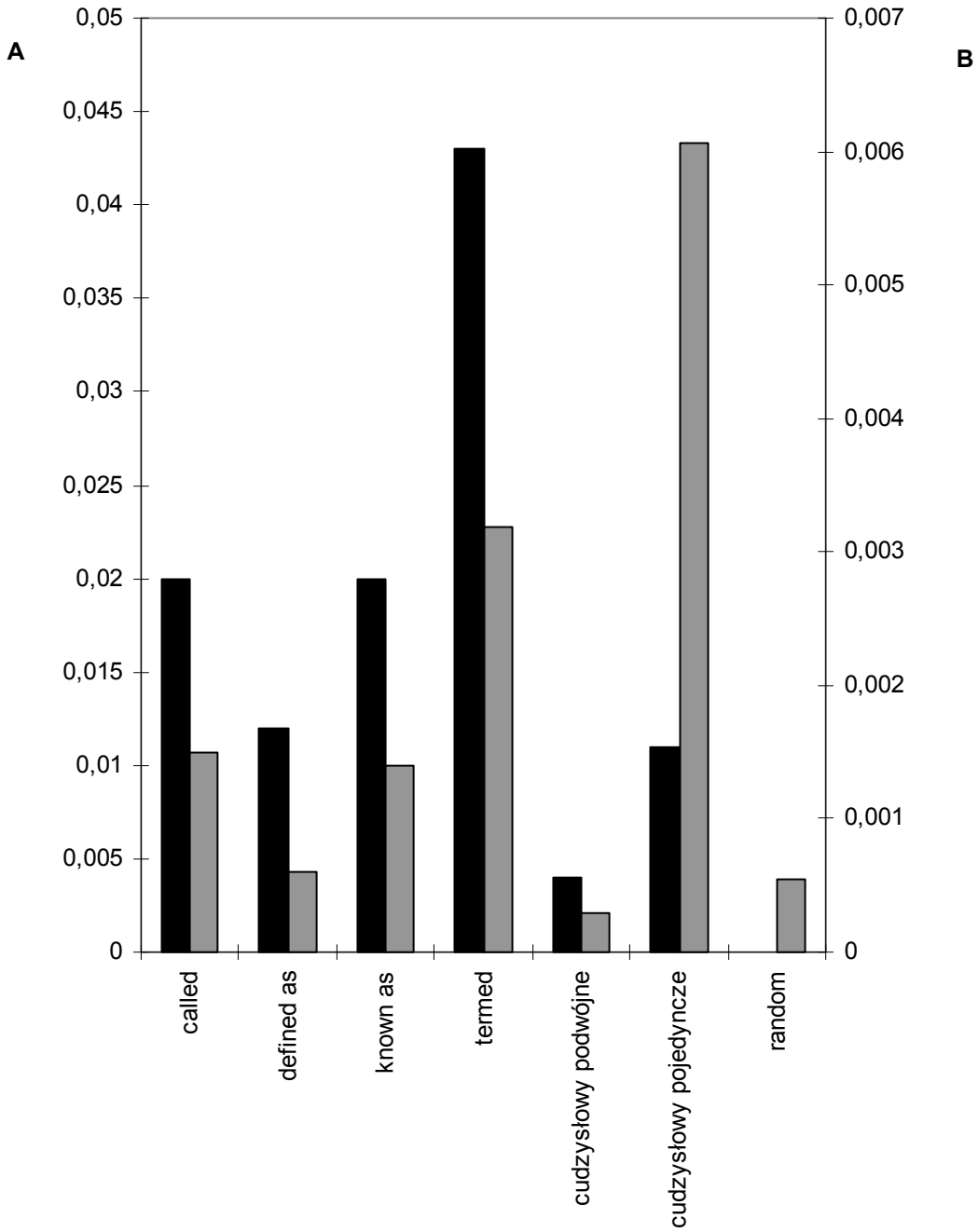
W tabeli 5 podano dane statystyczne opisujące poszczególne wyróżniki oraz plik *random* jako odniesienie (dane sumaryczne dla wszystkich roczników), zaś wykres 1 przedstawia graficzne porównanie uzyskanych wyników.

Tabela 5. Dane statystyczne wyróżników.

Wyróżnik	**Liczba jednostek nowych**	**Liczba jednostek nowych/ liczba wystąpień wyróżnika**	**Liczba jednostek nowych/ całkowita liczba słów poddanych analizie**
called	163	0,020	0,0015
defined as	16	0,012	0,0006
known as	84	0,020	0,0014
termed	44	0,043	0,0032
cudzysłowy podwójne	191	0,004	0,0003
cudzysłowy pojedyncze	581	0,011	0,0061

Wyróżnik	**Liczba jednostek nowych**	**Liczba jednostek nowych/ liczba wystąpień wyróżnika**	**Liczba jednostek nowych/ całkowita liczba słów poddanych analizie**
plik *random*	3	-	0,00055
Suma	1082	-	-

Wykres 1. Stosunek liczby jednostek nowych do liczby wystąpień (A, czarny) oraz stosunek liczby jednostek nowych do całkowitej liczby słów poddanych analizie (B, szary).



4.2.6 Możliwości rozszerzenia metody

Zaproponowaneą metodę można bez znaczniejszych trudności zastosować w badaniu innych języków po wprowadzeniu odpowiednich poprawek – zarówno do wyróżników graficznych, jak i leksykalnych.

W kwestii wyróżników leksykalnych, język angielski – jako język w dużym stopniu analityczny (Whaley 1997, Comrie 1989, Mallinson, Blake 1981, Greenberg 1974), nie stwarza poważniejszych problemów w wyodrębnianiu wyróżników, ponieważ występują one w jednej formie, bez form fleksyjnych (*termed*, *known as*, *called*, *defined as*), czyli w procesie ekscerpcji nie występują dodatkowe trudności związane z fleksją, można zatem wyróżniki z łatwością wyszukać w tekście.

Z kolei języki fleksyjne wykorzystują znacznie więcej różnych form tego samego wyrazu (liczby, przypadki, osoby itd.) (Van Valin 2001), przez co wyodrębnianie fraz jest procesem nieco trudniejszym, możliwym jednak do wykonania przy użyciu odpowiednich wyrażeń regularnych. Poniższe przykłady dotyczą języka polskiego:

Jako wyróżniki leksykalne wykorzystać można następujące wyrażenia regularne:

1. *tzw*.

2. *określan*[a-z]+ *jako* (defined as).

3. *definiowan*[a-z]+ *jako* lub *definiuje się jako*.

4. *zwan*[a-z]+ lub *nazywan*[a-z]+.

(Metodę opisano szczegółowo w odniesieniu do większości z podanych powyżej wyróżników w pracy Wierzchoń 2002.)

Podobnie należałoby wprowadzić pewne modyfikacje w analizie wyróżników graficznych. Język polski wykorzystuje cudzysłowy lewe dolne i prawe górne, które należy zamienić na znaki ASCII (") występujące w plikach tekstowych. Ponadto, cudzysłowy pojedyncze nie są stosowane (a co najmniej nie jest to zgodne z zasadami pisowni), wskutek czego cenne dla leksykografa rozróżnienie obserwowane w języku angielskim nie występuje. W języku polskim bowiem zarówno cytowania (w języku angielskim częściej oznaczane cudzysłowami podwójnymi), jak i wyrażenie niepewności, zaznaczenie ironii, neologizmu lub neosemantyzmu (w języku angielskim oznaczane częściej cudzysłowami pojedynczymi) przedstawiają cudzysłowy podwójne; wskutek tego można przypuszczać, że skuteczność wyróżników graficznych w ekscerpcji neologizmów (stosunek liczby jednostek nowych do całkowitej liczby słów w wyodrębnionych frazach) może być niższa niż w przypadku języka angielskiego.

Interesujące byłoby ponadto zbadanie zagadnień w perspektywie chronologii lingwistycznej, tj. po prostu z uwzględnieniem parametrów chronologicznych. Na podstawie posiadanego zbioru tekstów można by prześledzić zmiany frekwencji danej jednostki nowej w czasie i na tej podstawie sformułować wnioski dotyczące tego, czy jednostka nowa jest używana częściej i staje się powszechna, czy też jej wystąpienie miało charakter okazjonalny i było przypadkiem jednokrotnym. Do takich badań można by wykorzystać wyszukiwarkę, jednak wtedy należałoby analizę prowadzić systematycznie, na przykład co rok (nie jest możliwe uzyskanie z wyszukiwarki internetowej informacji o frekwencji danego słowa czy frazy z przeszłości). Ściślej rzecz biorąc, należałoby w pierwszym kroku operacyjnym przyjąć granicę datacji, względem której chce się określić neologiczność badanej jednostki, a następnie, przy zastosowaniu stosownej teorii lingwochronologizacyjnej, badać neologiczność danej jednostki względem granicy datacji.

4.2.7 Możliwości udoskonalenia metody

Mimo że przedstawiona metoda jest w zasadzie w znacznym stopniu zautomatyzowana, jej realizacja odbywała się etapami z użyciem programów komputerowych, w tym makropoleceń, rejestrowanych jednak bez programowania[54]. Bez większych trudności mogłaby jednak zostać przekształcona w kompletny pakiet umożliwiający wysoce zautomatyzowaną ekscerpcję jednostek nowych. Użytkownik określałby plik z danymi wejściowymi, definiował szukane wyróżniki – frazy lub znaki interpunkcyjne – i podawał wartość progową stosowaną w etapie określania częstości wystąpień w Internecie. Wynikiem pracy pakietu byłaby lista słów uszeregowanych wedle liczby wystąpień w zasobach wyszukiwarki internetowej. Jedynym etapem, który należałoby przeprowadzić ręcznie, byłaby ostateczna analiza i odrzucenie błędów typograficznych, wyrazów obcych, form efemerycznych itd. Mówiąc o odrzuceniu form efemerycznych, mamy na uwadze stan, w którym świadomie formy te są odrzucone, ponieważ może zajść taka ewentualność, że formy te – jako szczególnie cenne pod względem np. morfologicznym – będą poddawane osobnej analizie. W zależności od zastosowanych w pakiecie analizatorów morfologicznych mógłby on nawet pracować na więcej niż jednym języku.

Byłby to kierunek niewątpliwie bardzo pożądany z punktu widzenia zastosowania i popularyzacji metod, wymagałby jednak ścisłej, skoordynowanej współpracy zespołu złożonego z językoznawców formułujących wskazówki dotyczące metod i informatyków wykorzystujących je do stworzenia całościowego pakietu. Chociaż autor nie jest informatykiem, wydaje się, że skoro poszczególne elementy takiego pakietu są dostępne (analizator morfologiczny, wyszukiwarka internetowa, makropolecenia), stworzenie takiego pakietu nie wymagałoby zaangażowania sił przekraczających możliwości niewielkiej grupy osób zorientowanych na postawiony sobie cel.

[54] Więcej informacji dotyczących makropoleceń, patrz Pluciński 2000, Snarska 2007.

4.2.8 Dyskusja wyników

Mimo że w wyniku zastosowania metody otrzymano znaczną liczbę (1082) interesujących jednostek nowych o charakterze agnonimów leksykograficznych, obciążona jest ona pewnymi wadami:

a) nie wszystkie wyrazy w tekście poddawane są analizie; wyszukiwanie dotyczy jedynie fraz ograniczonych wyróżnikami graficznymi lub następującymi po wyróżnikach leksykalnych; to jednak właśnie dzięki zastosowaniu wyróżników prawdopodobieństwo znalezienia jednostki nowej jest znacznie większe niż w przypadku przeszukiwania losowego i znacznie szybsze niż metody manualne;

b) w związku z przyjętą procedurą niektóre formy są usuwane; dotyczy to przede wszystkim słów zawierających wielkie litery (a więc także wszystkich wyrazów występujących na początku zdania; nie obejmuje to jednak – co oczywiste – słów występujących po wyróżnikach leksykalnych, ponieważ te nie mogą znaleźć się na początku zdania); wydaje się, że próby pokonania tego problemu byłyby niewspółmierne do efektów, a przede wszystkim spowodowałyby konieczność rozróżnienia nazw własnych (na przykład nazwisk i nazw miejscowych), których usunięcie nastręczałoby wielu trudności, a które w opisywanej metodzie są w prosty sposób eliminowane;

c) metoda nie jest w pełni automatyczna; nie dotyczy to jedynie ostatniego etapu analizy manualnej, ale także poszczególnych etapów, które wykonuje się krok po kroku, choć w sposób globalny na całej zawartości pliku (programista połączyłby te etapy w całość, czyli w pełni zautomatyzował); wydaje się jednak, że w porównaniu do poszukiwania całkowicie ręcznego dokonano znacznych postępów.

Z drugiej jednak strony zastosowanie wyszukiwarki internetowej umożliwiło automatyczną eliminację znacznej liczby słów nieznanych

analizatorowi morfologicznemu, ale niebędących jednostkami nowymi, takich jak (przykłady z rocznika 1997):

a) błędy typograficzne: *thedynamics*, *contraints*, *searchfor*, *comittee*, *thepower*, *c oncerns*, *worldof*, *recurr*, *systemwith*, *necesssarily*, *itright*,

b) słowa obce: *oeconomicus*, *stephensi*, *voor*, *laboratoire*, *universités*, *burgdorferi*, *conventionné*, *elegans*,

c) formy archaiczne: *freind*, *helpe*, *onely*, *yeares*, *finisht*, *joineth*, *maketh.*

Jeśli chodzi o skuteczność ekscerpcyjną poszczególnych wyróżników – można jednoznacznie wskazać wyróżnik dający najlepsze rezultaty: są to cudzysłowy pojedyncze, które pozwalają wyszukać najwięcej ekscerptów zarówno w odniesieniu do wartości bezwzględnej (581 słów), jak i stosunku liczby ekscerptów do całkowitej liczby wyrazów poddanych analizie (0,0061). Ponadto dzięki wynikom można zaobserwować różnicę w użyciu między cudzysłowami pojedynczymi a podwójnymi: cudzysłowy podwójne wykorzystuje się do cytowania wypowiedzi, stąd wynika znaczna średnia liczba słów na jedno wystąpienie cudzysłowów (13,4) w porównaniu do 1,8 dla cudzysłowów pojedynczych – i skuteczność niższa nawet niż w przypadku pliku *random*. (Jest to zgodne z intuicją: liczba jednostek nowych w cytowaniach – obejmujących także wypowiedzi ustne – powinna być istotnie niższa niż w całości tekstu naukowego.)

Z kolei cudzysłowy pojedyncze są jednoznacznie wykorzystywane do oznaczania nowych użyć, pojęć (‘RNA World’, ‘pocket universes’), słów (‘biospherians’), niepewności piszącego lub użyć ironicznych (‘very’ clever). Dlatego też ich wykorzystanie w ekscerpcji jednostek nowych języka angielskiego wydaje się najbardziej uzasadnione.

Wyróżniki leksykalne także odznaczają się skutecznością (w szczególności *termed* i *called*, mniejszą skutecznością charakteryzują się

known as i *defined as*), w szczególności w odniesieniu do stosunku liczby jednostek nowych do liczby wystąpień wyróżnika. W ewentualnych przyszłych badaniach należałoby sprawdzić istnienie innych skutecznych wyróżników w języku angielskim oraz rozszerzyć analizy na inne języki. Porównanie poszczególnych wyróżników najłatwiej jest prześledzić na podstawie wykresu 1.

4.2.9 Wykorzystanie wyników do dalszych badań – badanie potencjału słowotwórczego wybranych morfemów

Celem niniejszej pracy jest przedstawienie komputerowych metod ekscerpcyjnych oraz ich wyników, nie zaś wykorzystanie uzyskanych rezultatów. Aby jednak pokazać na przykładzie, jak ciekawy materiał do dalszych badań stanowią zgromadzone jednostki nowe, zdecydowano się na sprawdzenie potencjału (tzw. produktywność) słowotwórczego wybranych morfemów. Zbadano częstość występowania następujących morfemów w jednostkach nowych z listy wynikowej: *pseudo-*, *super-*, *nano-*, *trans-*, *anthropo-*, *osteo-*, *mega-*, *micro-*, *poly-* i *meta-*.

Uzyskano następujące wyniki (tabela 6):

Tabela 6. Wyniki analizy morfemów w zgromadzonych jednostkach nowych.

Morfem	Jednostki nowe z morfemem	Liczba jednostek nowych
pseudo-	pseudocopulations pseudodate pseudodating pseudodimer pseudohaemolysin pseudolysogeny pseudomagnetic pseudomedical pseudoproxy pseudopupil pseudospins pseudosubstrates	12

Morfem	Jednostki nowe z morfemem	Liczba jednostek nowych
super-	superalliance superannalist superbrownian superchemistry superconvection supercorrelation superdimer supergreenhouse superhybrid superinvar superlubricity supermeme superministry supernebula superpenumbral superphyla superphyletic superpigment superplanets superpostdocs superprotonic superrotation supersegment supersog superspreaders superspreading superstrains supersupershifted supersusceptibility superswells supervolatile superwells superwetting	33
nano-	nanobaubles nanocage nanocrimelab nanofossils; nanogratings nanoharvesting nanolecture nanolitres; nanolouse nanopod nanoprisms nanoprojects nanoreactor nanorover nanosalt nanosociology nanostencilled	20

Morfem	Jednostki nowe z morfemem	Liczba jednostek nowych
	nanothermometer nanothermometers nanowhisker	
trans-	transactivity transcriptosomes transdetermination; transdifferentiating transducisome transgenomics transloxer transportins transresistivity	9
anthropo-	anthropodenial anthropozoic	2
osteo-	osteochondroprogenitor osteodontokeratic osteoimmunology osteolepiform	4
mega-	megabasin megabeds megabillion megachannels megamouse megamullion megaplumes megapore megaregolith megaturbidite	10
micro-	microacoustics microbiography microcephalics microchimaerism microdermic microdiverse microislands microlepton micromastic micrometozoan micromoulding micropolygyria microstreaming microsyntenic	14
poly-	polyamorphic polyamorphism polyaxonal polygerm polyheads polyhook polymorula polynail	12

Morfem	Jednostki nowe z morfemem	Liczba jednostek nowych
	polyphyrin polyplacophoran polyplocodont polypodiaceous	
meta-	metacarbonate metaethnic metalloligands metanodes metasaccharinic metastatistic	6

Okazuje się, że – przynajmniej w tej, ograniczonej próbce jednostek nowych – zdecydowanie najbardziej produktywnym morfemem jest morfem *super-* (33 słowa z tym morfemem), zaś drugi w kolejności morfem to *nano-* (20 słowa), co nie dziwi, zważywszy na rozpowszechnienie badań nad nanotechnologią. Specyfika podjętego korpusu badań jest interesująca także z tego punktu widzenia, że możliwa byłaby konfrontacja wyników uzyskanych powyżej z wynikami uzyskanymi dla np. korpusu tekstów nienaukowych. Konfrontacja ta pozwoliłaby postawić hipotezy dotyczące specyfiki morfologicznej tekstu naukowego. Na gruncie języka polskiego uzyskane dane można zestawić np. z danymi zaprezentowanymi w pracy K. Waszakowej (2005). Zestawienie to pozwoliłoby być może wykryć korelacje pomiędzy strukturą morfemową (morfologiczną) języka angielskiego w odmianie naukowej a angielszczyzną ogólną.

4.2.10 Próba przekładu uzyskanych jednostek nowych

Ponieważ na początku badań nad listą jednostek nowych interesujące wydawało się zweryfikowanie możliwości, a także celowości przeniesienia wyodrębnionych słów do języka polskiego, planowano zająć się kwestią tłumaczenia elementów uzyskanej listy na język polski.

Okazało się jednak, że – w naszej przynajmniej opinii – praca taka nie wnosiłaby nowych elementów w naukę o przekładzie, trudno zaś przypuszczać, że przyczyniłaby się do wzbogacenia słownictwa języka polskiego. Teoretycznie bowiem – tak, jednak z faktu zamieszczenia tłumaczeń (a raczej ich propozycji) w jednej publikacji (rozprawie doktorskiej), a co więcej – w postaci listy obejmującej szereg dziedzin (nauki przyrodnicze, matematyczne), nie wynika, że istniałaby jakakolwiek szansa ich zastosowania czy rozpowszechnienia w polszczyźnie. Pojawiła się ponadto następująca trudność – jak tłumaczyć podane wyrazy? – a należy pamiętać, że w języku oryginału – angielskim – pojawiają się one rzadko (przynajmniej w wyszukiwarce internetowej), zaś niekiedy ich występowanie ogranicza się wyłącznie do jednego artykułu z czasopisma *Nature*, co oznacza, że dysponujemy ubogą bazą tekstów, w których dane słowo występuje. Tak więc w tym miejscu stawia się problem potencjalnie możliwych strategii tłumaczenia wyekscerpowanych jednostek. Podjęcie tego wysiłku (w zakresie klasyfikacji) uznajemy za uzasadnione, ponieważ analizie translacyjnej podlegać będzie bardzo specyficznie wyekscerpowana grupa jednostek.

W wyniku przeprowadzonej analizy wyodrębniono jednak kilka strategii tłumaczenia, które można zastosować do części przynajmniej wyodrębnionych słów:

1) W słowie oryginalnym można wydzielić podstawę słowotwórczą (znany, istniejący w języku angielskim, tj. np. rejestrowany przez słowniki ogólne języka angielskiego, dla której istnieje odpowiednik w języku polskim) oraz przedrostek lub przyrostek, dla którego także istnieje polski odpowiednik.

Przykład: *femtotechnology* (proponowane tłumaczenie: *femtotechnologia*) – w tym wypadku wyraz składa się z rozpoznawalnego morfemu *femto-* występującego w języku polskim w jednostkach miar (na przykład *femtometr*, *femtosekunda*) i wyrazu podstawowego *technology*, który także ma odpowiednik w języku polskim.

Inne przykłady:

- *superpigment* (*super* + *pigment*), czyli *superpigment*,
- *pseudodimer* (*pseudo* + *dimer*), czyli *pseudodimer*,
- *palaeoanthropologist* (*palaeo* + *anthropologist*), czyli *paleoantropolog*,
- *nanosociology* (*nano* + *sociology*), czyli *nanosocjologia*,
- *palmitoylputrescine* (*palmitoyl* + *putrescine*), czyli *palmitoiloputrescyna*,
- *polyamorphism* (*poly* + *amorphism*), czyli *poliamorfizm*.

2) Znaczenie słowa angielskiego wydaje się znane – na podstawie definicji w tekście lub jednoznacznego kontekstu, jednak trudno jest (co nie znaczy, że jest to bezwzględnie niemożliwe, ponieważ zależy to od umiejętności autora) utworzyć pojedyncze nowe słowo w języku polskim odpowiadające oryginałowi, co oznacza, że można zastosować tłumaczenie opisowe.

Przykład: *undruggable* – proponowane tłumaczenie: „cząsteczka, w przypadku której nie jest możliwe zaprojektowanie leku ukierunkowanego na tę cząsteczkę".

Inne przykłady:

- *legness*: posiadanie nóg (ewentualnie odnóży),
- *etchability*: zdolność do poddawania się wytrawianiu, wytrawialność.

3) Jednostka angielska stanowi nazwę chemiczną, biochemiczną lub biologiczną substancji chemicznej bądź części komórki lub organizmu, przy czym dzięki obecności w języku polskim podobnych słów łatwe jest

spolszczenie słowa angielskiego. Współcześnie opracowuje lub odkrywa ogromną liczbę substancji, dla których przyjmuje się nazwy zwyczajowe. Najczęściej są one przenoszone na grunt języka polskiego z wykorzystaniem mniej lub bardziej intuicyjnych i często niejednoznacznych reguł (dlatego na przykład substancja czynna znana w języku angielskim jako *sildenafil* jest w polskich dokumentach oficjalnych tłumaczona jako *sildenafil* lub *syldenafil*)[55].

Przykład: *destabilase*; ponieważ angielskie nazwy enzymów z końcówką *-ase* przybierają w języku polskim końcówkę *-aza*, dlatego proponowanym tłumaczeniem jest destabilaza.

Inne przykłady:

- *chromatosom*: proponowane tłumaczenie – *chromatosom*,
- *neoglycopolymer*: proponowane tłumaczenie – *neoglikopolimer*.

W przypadku wielu terminów podanie przekonywająco uzasadnionych propozycji tłumaczenia byłoby – jak sądzimy – niezwykle trudne, a być może nawet niewykonalne, na przykład: *cavicapture*, *selfplex*, *gongylidia*, *superannalist*.

Ponieważ na podstawie przytoczonych przykładów można stwierdzić, że tłumaczenie uzyskanych jednostek nowych angielskich byłoby dla pewnej grupy czynnością niemal automatyczną, zaś dla innych byłoby niemożliwe, podjęto decyzję o niekontynuowaniu prac w ramach tego wątku badawczego.

[55] Warto w tym miejscu dodać, że jakkolwiek wydawałoby się, iż w nazwach substancji czynnych, a ogólnie – związków chemicznych – przyjęty szyk jest następujący: przymiotnik-rzeczownik (np. *chlorek sodu*), instytucja zajmująca się nadzorem nad preparatami farmaceutycznymi (Urząd Rejestracji Produktów Leczniczych, Wyrobów Medycznych i Produktów Biobójczych) wymaga stosowania odmiennego szyku (tzn. *sodu chlorek*). Wynika to prawdopodobnie z chęci zachowania szyku angielskiego (*sodium chloride*), chociaż nie bardzo wiadomo, czemu miałoby to służyć.

4.3 Metoda ekscerpcji kolokacji w oparciu o akronimy

4.3.1 Kolokacje, zwroty stałe i jednostki wielowyrazowe

Należy zauważyć, że w literaturze przedmiotu nie przyjęła się uniwersalna terminologia dotycząca – najogólniej rzecz ujmując – fraz składających się z dwóch lub więcej słów, które wobec siebie wykazują powinowactwo, czyli współwystępują. Frazy takie określa się mianem kolokacji (ang. *collocations*), zwrotów stałych (ang. *fixed expressions*) bądź jednostek wielowyrazowych (ang. *multi-word units*, MWU), ewentualnie leksemów wielowyrazowych (ang. *multi-word lexemes*, MWL). Z analizy literatury zamieszczonej poniżej wynika, że przez niektórych autorów terminy te stosowane są synonimicznie. W dalszej części tego typu ciągi wyrazowe, jeżeli opis nie będzie odnosił się do referowania badań w literaturze przedmiotu, będziemy nazywać po prostu jednostkami (będą to *de facto* jednostki ekscerpcji).

Niezwykle obszerne i rzetelne podsumowanie obecnego stanu badań nad teorią kolokacji[56] – z pespektywy leksykologicznej i leksykograficznej – przedstawia D. Siepmann (2005, 2006) zastrzegając się, że żadne z kryteriów definiujących kolokację nie znajduje zastosowania we wszystkich przypadkach (Siepmann 2005: 410).

Analizę cech jednostek wielosegmentowych o funkcji rzeczownikowej przedstawia Kosek (2008).

Kolokacje definiuje się i bada się z trzech odrębnych perspektyw. Wyróżniamy następujące perspektywy:

[56] Por. także wcześniejsze opracowanie (Moon 1998).

1) semantyczną (por. Benson 1986, Mel'čuk 1998, González-Rey 2002, Hausmann 2003, Grosmann, Tutin 2003) wykorzystującą wzajemną relację znaczeniową między składnikami kolokacji, która ma charakter bardziej intuicyjny;

2) frekwencyjną, bliższą językoznawstwu komputerowemu, która bada współwystępowanie dwóch lub większej liczby słów za pomocą metod statystycznych (por. Sinclair 1991, 2004, Kjellmer 1994);

3) pragmatyczną, która zakłada, że niekompozycyjność frazemów i kolokacji wynika z regularnych właściwości pragmatycznych, które określają „zależność między kontekstem sytuacyjnym a formą językową" (Siepmann 2005: 411), por. Feilke 1996, Feilke 2003).

Jakkolwiek liczbę składników kolokacji określa się jako dwa lub więcej, Siepmann wykazuje (2005: 415), że wiele kolokacji trójskładnikowych daje się zredukować do postaci binarnej (por. Schafroth 2003). W takiej postaci, według Hausmanna (1999) jeden ze składników jest semantycznie zależny (kolokat), zaś drugi – niezależny (podstawa kolokacji), co pozwala odróżnić kolokację od słów występujących obok siebie i nietworzących kolokacji.

Badając kolokacje, Siepmann stwierdza (2005: 431), że „każdy schemat koligacyjny[57] może stanowić podstawę kolokacji" (por. Hausmann 1985), zaś niektóre z nich tworzą kolokację szczególnie często.

Wynikiem tych rozważań jest następująca definicja kolokacji:

[57] Hoey (1998) definiuje koligację jako otoczenie gramatyczne, w którym dane słowo występuje lub w którym nie występuje, funkcje gramatyczne słowa w grupie oraz miejsce w ciągu wyrazów, które zajmuje. W sformułowaniu tym znaleźć można odniesienie do kontekstowej i stosunkowo szerokiej definicji kolokacji Firtha (1957: 181) „Collocations of a given word are statements of the habitual or customary places of that word... The collocation of a word or a 'piece' is not to be regarded as mere juxtaposition, it is an order of mutual expectancy".

> "a collocation is any holistic lexical, lexico-grammatical or semantic unit normally composed of two or more words which exhibits minimal recurrence[58] within a particular discourse community"
>
> (Siepmann 2005: 438)

Kolokacje były przedmiotem wnikliwych badań wielu autorów (patrz Allerton 1984, Cruse 1986, Choueka 1988). Smadja (1993: 148) uważa kolokacje za „spójne grupy leksykalne" (obecność jednego składnika kolokacji wyznacza występowanie drugiego), które powtarzają się (to znaczy występują w danym kontekście w różnych wypowiedziach) i są charakterystyczne dla danej dziedziny czy typu tekstu (często mają zastosowanie w ograniczonych obszarach). Ponadto ogólną cechą kolokacji jest to, że podstawienie w niej synonimu zamiast jednego z członów daje błędną kolokację (czy też jej brak) (Manning, Schütze 1999). Można mówić na przykład o *bagażu emocjonalnym*, trudno jednak o *pakunku emocjonalnym*[59].

Kwestii substytucji w ramach kolokacji rozumianej jako jednostka języka poświęca wiele uwagi A. Bogusławski w cyklu swoich prac (por. z ostatnich: Bogusławski 2010), w których szczególny nacisk kładzie się właśnie na ograniczenie (zamknięcie) klasy elementów, które mogą być substytuowane (por. także Bogusławski, Danielewiczowa 2005). Klasa taka winna być określona przez wyliczenie jej elementów (dla powyższego połączenia *bagaż emocjonalny* tym elementem będzie właśnie segment *bagaż*, ale nie *pakunek*). Przywołując pracę A. Bogusławskiego, należy tu wspomnieć także o licznych na ten temat opracowaniach leksykologicznych

[58] Pojęcie *minimal recurrence* (minimalnej powtarzalności), por. Kocourek 1991.
[59] Badania takie mają w związku z tym duże znaczenie w nauczaniu języków obcych, ponieważ uczący się często stosują błędne kolokacje. Ich przykłady dla języka angielskiego podaje na przykład Gitsaki et al. (2000): *many thanks*/**several thanks*, *strong coffee*/**powerful coffee*, *tap water*/**pipe water*.

M. Grochowskiego czy W. Chlebdy (seria *Idiomatykonów*), por. także Kosek 2008.

Merkel i Andersson (2000: 737) uważają, że tożsame z kolokacjami nieprzerwanymi (por. kolokacje dalekiego zasięgu, Siepmann 2005) są jednostki wielowyrazowe czy też leksemy wielowyrazowe. Próby zdefiniowania tego terminu sięgają definicji o charakterze mniej naukowym, bardziej zaś – intuicyjnym. Smith (1943) stwierdza, że leksemy takie to zwroty „peculiar to a people or nation" oraz „phrases which are verbal anomalies, which transgress … either the laws of grammar or the laws of logic". Zwraca się więc w tej definicji uwagę na idiomatyczność takich leksemów, a nawet łamanie przez nie zasad gramatycznych i logicznych. Z kolei definicja Reagana (1987: 417) – także o charakterze intuicyjnym – dotyczy niemożności wnioskowania o znaczeniu leksemu wielowyrazowego na podstawie znaczeń jego składników: „a collection of words whose meaning as a whole cannot be derived from the meanings of the individual words".

Wydaje się, jednak, że najpełniejszą definicją, która odwołuje się do tego, iż leksem wielowyrazowy – właśnie jako leksem[60] (w naszym ujęciu: jednostka ekscerpcji) – stanowi pełną integralną całość, jest określenie sformułowane przez Brundage'a et al. (1992: 4):

> „The first important property of an MWL is its lexeme status. This is the essential property in which MWLs differ from free syntagmatic constructions that are produced according to syntactic structure models every single time they are stated. On the contrary, MWLs are lexicalized and therefore reproduced as lexical units of the language system in question".

Automatyczne metody ekscerpcji kolokacji mają wielkie znaczenie w leksykografii i analizie językoznawczej (omówienie, patrz Jacquemin, Bourigault 2003: 599 – 615), ponieważ pozwalają zgromadzić duże zbiory

[60] Definiowany w sposób następujący: „A lexeme is the basic abstract unit of the lexicon" (Bussmann 1999).

wartościowych poznawczo danych. W leksykografii istnieje potrzeba ekscerpcji kolokacji, w szczególności kolokacji o charakterze terminów do celów rejestracyjnych[61]. W językoznawstwie ogólnym listy kolokacji wykorzystywane są między innymi do badania wzorców gramatycznych i analizy zmian zachodzących w języku na skalę, którą ogranicza jedynie rozmiar korpusu lub zbioru tekstów i moc obliczeniowa komputerów. Wcześniejsze metody opierały się na obliczeniach czysto statystycznych. Aparat statystyczny narzuca w takich badaniach szeroką definicję kolokacji Firtha (1957). Badania kolokacji w oparciu o metody statystyczne (por. Zhou, Dapkus 1995, Johansson 1994, Damerau 1993, Dunning 1993, Pedersen 1996, Church, Hanks 1990) polegają na podziale tekstu (tekstów) na n-gramy. Por. koncepcje użycia list słów[62] (ang. *seed word lists*), Weeber et al. 2000, oraz list stop[63] (ang. *stop lists*) (zob. Damerau 1993, Merkel et al. 1994), dzięki którym ogranicza się ilość przetwarzanych danych lub na wczesnym etapie badań wyklucza ich część. Stosując ściśle statystyczne podejście, jednak, trudno badać kolokacje dłuższe niż dwuskładnikowe (por. Blaheta, Johnson 2001), w danych wynikowych pojawiają się niepożądane „frazy przysłówkowe, przymiotnikowe, przyimkowe i spójnikowe" (Dias et al. 1999), trudności przysparza ponadto wyodrębnienie kolokacji występujących z niską częstotliwością w badanym zbiorze tekstów (por. Shimohata et al. 1997, Yamamoto, Church 1998).

Wraz z rozwojem adnotacji części mowy i parsingu rozpoczęto stosować strategie hybrydowe uwzględniające oprócz danych statystycznych także adnotacje morfologiczne i syntaktyczne (por. Krenn 2000, Lin 1998, Boguraev, Kennedy 1999, Birn 1997, Pearce 2001), dzięki którym kolokacja

[61] Okazuje się jednak, że badania kolokacji prowadzi się także na podstawie słowników, por. Pedersen 1995; w pracy tej wyodrębnia się kolokacje ze słowników technicznych, a jednocześnie wyraźnie rozgranicza pojęcie terminu zawierającego więcej niż jedno słowo od właściwej kolokacji (Pedersen 1995: 62).

[62] Lista taka obejmuje wszystkie słowa, które mają być wyszukiwane w korpusie (na przykład szukane podstawy kolokacji).

[63] Lista taka obejmuje wszystkie słowa, które jeśli stanowią składnik n-gramu, wtedy n-gram taki zostaje wykluczony z dalszych badań, czyli z analizy obliczeniowej (Scott 1998: 49).

definiowana jest jako relacja syntagmatyczna (Fontenelle 1992: 222)[64]. Metody takie dostarczają danych wyjściowych o większej jednorodności (por. Evert, Krenn 2001). Na przykład Kermes i Heid (2003) stosują system rekursywny służący do klasyfikacji kolokacji złożonych z przymiotnika i czasownika.

Kis et al. (2004) wykorzystują angielsko-węgierski korpus równoległy tekstów technicznych SZAK[65] (Kis, Kis 2003) do ekscerpcji jednostek wielowyrazowych języka węgierskiego metodami statystycznymi. Celem projektu była analiza językoznawcza języka węgierskiego, poszukiwanie wydajniejszej metody rozszerzania zasobów leksykograficznych oraz udoskonalenie programów komputerowych do badań językoznawczych. Analizowano kolokacje konkretnego typu (czasownik, rzeczownik, znacznik przypadka) wykorzystując pakiet *ngram* (Pedersen, Banerjee 2003) oraz parser HumorESK. Listy kolokacji-kandydatów poddawano analizie statystycznej, korzystając z wartości logarytmu prawdopodobieństwa (ang. *log-likelihood*, Dunning 1993) i istotności (ang. *salience*, por. Kilgarrif, Tugwell 2001).

Daudaravičius i Marcinkevičienė (2004) porównują metody wyodrębniania kolokacji na podstawie wskaźników Mutual Information, T-score i Dice'a[66] oraz proponują nową metodę, którą określają mianem *Gravity Mount* (wskaźnika ciężaru), bardziej złożonego od poprzednich wskaźników, ponieważ uwzględnia nie tylko częstość występowania składników razem i oddzielnie, ale także różnorodność słów występujących na prawo od pierwszego składnika i na lewo od drugiego składnika. Im mniejsza ta różnorodność (czyli składniki częściej występują razem), tym wyższa jest wartość wskaźnika. Badacze korzystają przy tym z korpusu dziennika „The

[64] „The term collocation refers to the idiosyncratic syntagmatic combination of lexical items and is independent of word class or syntactic structure".

[65] Korpus utworzony przez wydawnictwo, obejmujący między innymi monolingwalny subkorpus oryginalnych publikacji z dziedziny informatyki w języku węgierskim o wielkości około 500.000 wyrazów), a także bilingwalny korpus równoległy przetłumaczonych publikacji z tej samej dziedziny (około 1 mln wyrazów). Głównym celem korpusu są badania terminologiczne oraz przekładoznawcze.

[66] Wskaźnik Dice'a jest wykorzystywany do obliczania współwystępowania wyrazów lub ich grup i zależy od częstości występowania wspólnie i oddzielnie. Do wzoru obliczeniowego wprowadzono logarytm, aby łatwiej było obserwować niewielkie zmiany wartości.

Times" z roku 1995. Należy jednak zauważyć, że łańcuchy kolokacji „can be exploited for the subsequent non-automatic extraction and/or generation of collocations" (Daudaravičius, Marcinkevičienė 2004: 343), a więc metoda ma ograniczenie związane z wymogiem manualnego wyodrębniania właściwych kolokacji ze zgromadzonego materiału.

Widdows i Dorow (2005) stosują analizę grafów (Widdows, Dorow 2002) do wyodrębniania sekwencji o charakterze idiomów[67] typu „A and B", które w kolejności odwrotnej „B and A" nie występują (dokładniej: występują znacznie rzadziej), na przykład „chalk and cheese" lub „day and night". Sekwencje wynikowe obejmują idiomy, charakteryzujące się brakiem możliwości dekompozycji, a także zwroty wykazujące kierunkowość związaną z ich charakterem semantycznym. Podstawą ich badań (ważną także dla rozważań zamieszczonych w niniejszej pracy) jet to, że „language resources such as dictionaries and lexical knowledge bases give at best poor coverage of such phenomena".

Seretan et al. (2004) wykorzystują strategię hybrydową polegającą na ekscerpcji prawdopodobnych kolokacji z tekstu dzięki wykorzystaniu adnotacji syntaktycznej, przypisaniu każdej wyszukanej jednostce liczby będącej prawdopodobieństwem, że stanowi ona kolokację oraz uszeregowaniu jednostek według malejącego prawdopodobieństwa. Zastosowana strategia umożliwia ekscerpcję kolokacji składających się więcej niż z dwóch składników i nadaje się do analizy tekstów w różnych językach.

Metody te zazwyczaj korzystają z analizy morfologicznej (patrz Brill 1992, Brants 2000, Charniak 1993) i złożonego aparatu statystycznego pozwalającego wyodrębnić sekwencje wyrazów, które stanowią językową całość, czyli kolokację. Wcześniejsze metody wykorzystywały jedynie narzędzia statystyczne. Wszystkie metody statystyczne, jednak, wymagają

[67] Interesujące omówienie zagadnień związanych z idiomami i idiomatycznością, por. Čermák 2001 oraz Nunberg et al. 1994.

dużych korpusów językowych umożliwiających przeprowadzenie istotnych statystycznie obliczeń.

Innym podejściem jest wybór określonego wzorca (oznacza to, że korpusu nie przeszukuje się w całości, a jedynie analizuje struktury lub ich elementy), ekscerpcja danych z korpusu i analiza szczegółowa. Wzorzec taki może stanowić słowo o określonych właściwościach gramatyczno-składniowych – którego wybór jest mniej lub bardziej arbitralny – po lub przed którym występuje inne słowo, na przykład rzeczownik poprzedzany przez przymiotnik (por. Zinsmeister, Heid 2003, patrz także Manning 1993). Wszystkie wystąpienia takich fraz – n-gramów – zostają wyodrębnione, zaś na podstawie wyników formułuje się wnioski.

Metody analizy jednostek wielowyrazowych są w dużej mierze zbieżne z metodami wykorzystywanymi w badaniu kolokacji. Metoda przyjęta przez Merkela i Anderssona (2000), na przykład, wykorzystuje filtry językowe do ekscerpcji leksemów wielowyrazowych (por. Merkel et al. 1994). Z listy wyodrębnionych n-gramów wykluczono n-gramy rozpoczynające się od określonych słów (np. *if*, *when*, *how*, *your*, *his*) i kończące się pewnymi słowami (*the*, *a*, *an*, *for*, *in*). Tę metodę porównano z metodą inną wykorzystującą entropię[68] (rozumianą jako miarę nieuporządkowania tekstu poprzedzającego leksem wielowyrazowy, który jest elementem stałym, czyli charakteryzującym się niską entropią, i po nim następującego). W kolejnym etapie wykorzystuje się filtr językowy zastosowany w metodzie pierwszej. Okazuje się, że metoda druga pozwala wyodrębniać leksemy wielowyrazowe ze stuprocentową dokładnością (Merkel, Andersson 2000: 744), chociaż liczba wyodrębnionych leksemów jest mniejsza niż w przypadku metody pierwszej. Zastosowaną przez nich metodę można zastosować do badania różnych języków (została zbadana na przykładzie języków szwedzkiego, angielskiego, niemieckiego i francuskiego).

[68] Por. Shimohata et al. 1997.

Dias (2000) posługuje się złożonym aparatem statystycznym do ekscerpcji jednostek wielowyrazowych (przykłady: *Council of Ministers*, *to comply with*, *as soon as possible*).

Hoover (2002) bada względną skuteczność i dokładność analizy wielu zmiennych częstotliwości występowania wyrazów oraz sekwencji wyrazów do identyfikacji autorów tekstów oraz grupowania tekstów jednego autora. Badanie takie służą do analizy zagadnień stylistycznych oraz autorstwa tekstów.

Metoda zaproponowana w tej części pracy pozwala abstrahować od obliczeń statystycznych. Umożliwia wyodrębnianie integralnych jednostek języka (w literaturze określanych często jako leksemy wielowyrazowe) z wykorzystaniem ich wyróżników, czyli cech leksykalnych lub graficznych, dzięki którym możliwe jest określenie granic jednostki. Mimo że zastosowanie metody jest stosunkowo proste, uzyskane dzięki niej wyniki są wysoce interesujące dla językoznawcy; metoda nie wymaga ponadto istotniejszego udziału człowieka, a w zasadzie może być w pełni zautomatyzowana – jedynym etapem prowadzonym całkowicie przez człowieka jest ostateczne sprawdzenie wynikowej listy wyekscerpowanych jednostek.

4.3.2 Opis metody

4.3.2.1 Sformułowanie żądania

Pytaniem podstawowym dla dyskutowanej ekscerpcji jest znalezienie wyróżnika jednostki wielowyrazowej.

Można przypuszczać, że niemożliwe jest ustalenie prostej (w sensie jej automatycznego wyodrębnienia) cechy wyróżniającej w tekście wszystkie grupy słów będących jednostkami wielowyrazowymi. Jeśli tak, należy

spróbować określić cechę, która – nawet jeśli charakteryzuje jedynie niektóre jednostki wielowyrazowe (lub nawet ich zdecydowaną mniejszość) – nie występuje przy połączeniach wyrazów niebędących jednostkami wielowyrazowymi.

Można przyjąć, że akronimy ujęte w nawiasy i następujące po ciągach słów, których akronim dotyczy, spełniają ten warunek graniczny i umożliwiają automatyczną ekscerpcję kompletnych jednostek wielowyrazowych.

W metodzie wyodrębnia się zatem wszystkie wystąpienia sekwencji wielowyrazowych, po których w nawiasie występuje ich akronim[69]:

$w_1\ w_2\ \ldots\ w_n\ (a_1a_2\ldots a_n)$

gdzie:

w_n itd. – słowa (sekwencje znaków literowych oddzielonych spacjami),

a_n – pierwsza litera słowa w_n.

Aby zwiększyć efektywność metody, czyli liczbę wyników na liście końcowej, w poszukiwaniu uwzględniono jako znaki rozdzielające słowa nie tylko spacje, ale także łączniki, na przykład w jednostce wielowyrazowej *Brain-Derived Neurotrophic Factor*.

Wstępne próby zastosowania metody dowiodły, że w wykorzystywanym zbiorze tekstów nie występują akronimy dłuższe niż sześcioliterowe – odpowiadające jednostkom sześciowyrazowym. W związku z tym poszukiwanie objęło jednostki o długości dwóch, trzech, czterech, pięciu i sześciu wyrazów – wraz z określającymi je akronimami.

[69] Ściślej – akronim (skrótowiec) literowy, czyli „skrótowiec utworzony z nazw pierwszych liter wyrazów składowych zestawienia, np. PKS" (Dubisz 2003).

4.2.2.2 Założenie

Założenie leżące u podstaw metody i niezbędny warunek jej przydatności do badań to stwierdzenie, że akronim zawsze oznacza jednostkę wielowyrazową. Jeśli tak, sekwencje słów, wobec których zastosowano akronim, muszą stanowić językową całość i być wzajemnie powiązane, tak że nazwanie ich jednostkami wielowyrazowymi jest uprawnione.

Ze zdroworozsądkowego punktu widzenia jest to nieuniknione. Sekwencje słów skraca się, stosując akronimy, ponieważ wymaga tego oszczędność miejsca w tekście, a to oznacza, że wykorzystywane są one częściej niż jednokrotnie; w każdym wystąpieniu pojawiają się w identycznej formie, co świadczy o tym, że składniki sekwencji zastąpionej akronimem stanowią integralną całość. Nie jest możliwe zatem, aby elementy takiego ciągu nie były ze sobą powiązane, a nawet jedynie były powiązane w niewielkim stopniu. Co więcej, można argumentować, że w większości przypadków takie sekwencje wyrazów stanowią jednostki wielowyrazowe.

4.3.2.3 Dane wykorzystane w opisywanej metodzie

W celu zwiększenia skuteczności metody, czyli liczby wyszukanych kolokacji, wykorzystano zbiór tekstów naukowych z czasopisma *Nature*, który szerzej omówiono w rozdziale 4, punkt 1. Można bowiem twierdzić, że to właśnie w tekstach naukowych i technicznych akronimy pojawiają się najczęściej.

4.3.2.4 Procedura zastosowana w badaniach

1. Ekscerpcja fraz ze zbioru tekstów przy użyciu następującego wyrażenia regularnego (przykład dla jednostek składających się z trzech wyrazów):

```
([a-z])[a-z]+[- ]([a-z])[a-z]+[- ]([a-z])[a-z]+
[- ]\(\1\2\3\)
```

Wyrażenie to obejmuje odwołania wsteczne i umożliwia wyodrębnienie wszystkich sekwencji trzech segmentów, po których występuje ich akronim (segmenty-wyrazy w sekwencji zaczynają się od liter występujących w akronimie w odpowiedniej kolejności).

Wynik operacji

Abdur Reef Limestone (ARL)

aberrant crypt foci (ACF)

Above-Budget Foundation (ABF)

above-threshold ionization (ATI)

above-threshold-ionization (ATI)

accelerated mass spectrometry (AMS)

accelerator mass spectrometer (AMS)

accelerator mass spectrometric (AMS)

accelerator mass spectrometry (AMS)

accelerator mass spectroscopy (AMS)

accelerator-mass-spectrometry (AMS)…

2. Akronimy zostają usunięte, ponieważ w dalszym postępowaniu są nieprzydatne i neutralne względem kolejnych kroków badawczych.

3. Otrzymane ciągi zostają poddane sortowaniu (sortowanie bez uwzględnienia wielkości liter), zaś wiersze powtarzające się są usuwane.

4. Wiersze różniące się jedynie łącznikiem zostają usunięte, na przykład:

death effector domains

death-effector domains lub

downstream regulatory element lub

downstream-regulatory-element

Mimo że jednostki takie nie są identyczne, można założyć, że oznaczają to samo pojęcie (reguły stosowania łączników nie są w języku angielskim jednorodnie ustalone), a więc traktowanie ich jako różnych wyników nie byłoby właściwe i sztucznie zawyżałoby ostateczną liczbę znalezionych jednostek.

5. Manualne sprawdzenie uzyskanej listy w celu usunięcia jednostek różniących się jedynie kategorią liczby ostatniego wyrazu ciągu, na przykład:

peripheral blood leukocyte

peripheral blood leukocytes

Ten etap można zautomatyzować, stosując odpowiednio dobrane wyrażenie regularne, na przykład (podano wyrażenie dla listy jednostek trójwyrazowych oddzielonych przecinkami):

```
([a-z]+)[- ]([a-z]+)[- ]([a-z]+), \1[- ]\2
[- ]\3(s|es),
```

6. Manualne sprawdzenie listy w celu usunięcia ewentualnych błędów ortograficznych i ewentualnie typograficznych (na przykład: *circular dichrosim*) oraz fraz względem języka angielskiego obcych (na przykład: *don juan*, *Agenzia Spaziale Italiana*, *Biochemie-Zentrum Heidelberg*).

4.3.3 Wyniki

Liczbę jednostek wielowyrazowych wyodrębnionych za pomocą przedstawionej w tej części pracy metody podano w poniższej tabeli 7. Liczby określają jedynie unikatowe jednostki wielowyrazowe (przy założeniu identyczności ciągów różniących się jedynie łącznikiem).

Tabela 7. Ekscerpcja jednostek wielowyrazowych, wynik ostateczny

Liczba słów w jednostce wielowyrazowej	**Liczba wyodrębnionych jednostek**
2	1450
3	3298
4	1320
5	198
6	19
Suma	**6285**

Poniżej przedstawiono wyodrębnione jednostki. Podano wszystkie jednostki rozpoczynające się na literę a, b, c, d, e, f oraz g.

4.3.3.1 Jednostki dwuwyrazowe

Abdur Central
Abdur North
Abdur South
abscission zone
absolute humidity
absorbance units
abyssal peridotites
accumulation rates
acetylation conditions
acidic region

actinic light
action potential
activating domain
activation domain
activator protein
active site
acyl transferase
ad libitum
adaptive optics
adenylyl cyclase
adherens junction
adhesive pit
adrenergic receptor
Advanced Quantifier
aggressive polygyny
Aleutian Arc
alkaline phosphatase
alternating field
amacrine cell
amino acid
amorphous material
amplitude modulated
androgen receptor
Angerman River
animal caps
anion exchanger
annulate lamellae
anoxic depolarization
Antarctic Divergence
antennal lobe
anterior blastomere
anterior cingulate
anterior field
anterior horns
anterior prostate
anterior snout
anterior-posterior
antero-posterior
anthracotheriid unit
anti-parallel
anti-sense
antidote oligonucleotides
Antiproton Decelerator
aortic sac
apical cell
apoptotic thymocytes
Arabidopsis thaliana
arachidonic acid
arbitrary units
arbuscular mycorrhizae
arbuscular mycorrhizal
archaeological horizons
Arctic Oscillation
Area striata
artificial intelligence
Ascaris haemoglobin
assistant referee
Associated Press
astronomical unit
asymmetrical body
ataxia telangiectasia
atomic absorption
atrial siphon
atrial ventricular

atrio-ventricular
auditory field
Axon terminals
Bacillus thuringiensis
Baldwin Hills
band pass
barb ridges
barley lectin
basal cell
basal forebrain
base pair
basic task
basilar membrane
Bayes factors
beam splitter
Becklin-Neugebauer
before present
Belousov-Zhabotinsky
benzoic acid
best frequencies
best frequency
bifunctional rhodamine
Big Bear
binding energy
binding potential
binding protein
binding sites
biologically effective
Birch-Murnaghan
black holes
black-carbon
bloodstream forms
bloodstream-form
body depth
Bolwig organ
bone collar
bone marrow
bone resorption
bone-marrow
bongkrekic acid
bootstrap proportion
boron nitride
Boston University
Brain lysates
branch point
Branchial arch
Branching factor
bridging oxygen
bright flare
bright-field
Brillouin zone
British Columbia
British Petroleum
Brodmann area
Brown Norway
Bruce passage
bucco-lingual
butyl-sepharose
calf thymus
calorie restriction
calponin homology
Camp Rock
cancer-testis
Candida albicans

Canton-Special
capillary electrophoresis
capping buffer
carbonic anhydrase
cardiac actin
cardinal vein
carrier envelope
catalytic fragment
catalytic site
catch trial
cauline leaf
Cauline leaves
cellular automata
central zone
centrifugality index
Centro Eccelenza
cephalic furrow
cerebellar vermis
Cerro Paranal
Chaetodipus baileyi
characteristic frequency
charge injection
charge recombination
charge separation
charge transfer
charge-ordered
charge-ordering
charge-parity
charge-transfer
cheek depth
cheliforal ganglia
chemically-defined
cholera toxin
chondroitin sulphate
chorionic gonadotropin
Choristoneura fumiferana
choroid plexus
chromatic dispersion
chromo domain
ciliary bands
cinnamic acid
circadian time
circular dichroism
Cladh Hallan
claviger-macra
clayey mud
closest point
cluster diffusion
coat protein
cochlear microphonic
coiled coil
coincidence detector
cold work
colloidal particles
combination indices
companion cells
competitive indices
complementing factor
complex-spike
composite fermion
compound nucleus
computed tomographic
computed tomography
computer tomography

computer-generated
Computerised tomography
computerized tomography
concatemer formation
condensation nuclei
conditioned media
conditioned medium
conditioned responses
conditioned stimuli
conditioned stimulus
conduction band
confidence interval
confidence level
configuration interaction
Connecting cilium
Conservation International
consistency index
constant alkalinity
constant-frequency
constitutively active
constraint index
Consumers International
contiguous premises
continental crust
continuous wave
contrast discrimination
contribution index
cooperative polygyny
corn oil
corpora cardiaca
corpora lutea
corpus callosum
correlation coefficient
cortical plate
Cotton effect
Coulomb blockade
coupled conservation
courtship index
creatine kinase
credibility interval
crista ampullaris
critical point
cross polarization
crown completion
cryptic prophage
cubitus interruptus
Cubitus interuptus
culture medium
cycle threshold
cysteine-rich
cystic fibrosis
cytoplasmic extract
cytoplasmic incompatibility
cytoplasmic membrane
cytoplasmic receptor
dangerous contact
Danon disease
dark chocolate
dark flash
dark matter
dbl homology
death receptor
decyl maltoside
deep minimum

dehiscence zones
delay lines
dendritic caps
dendritic cell
dendritic tips
Denmark Strait
dense-core
densitometric units
density-independent
dentate gyrus
dermal papilla
detector response
diagenetic silicification
dielectric spectroscopy
Differential Display
differentiation medium
differentiation zone
dilute-lethal
diptheria toxin
direct current
Direction selectivity
Discovery passage
dislocation creep
dislocation dynamics
Disporum hookeri
dissociative recombination
dissolved oxygen
distal projection
distribution analysis
dobson units
domain wall
dominant negative
domoic acid
don juan
dorsal arms
dorsal fin
dorsal organ
dorsal projection
dorsal ramus
dorsal region
dorsal-ventral
double negative
double positive
double-negative
double-positive
double-stranded
Drosophila melanogasterr
dry weight
ductus arteriosus
dynamic range
early adolescence
early embryogenesis
early endosomes
early palindrome
Earth wind
ectoplacental cone
egg jelly
egg length
ejection time
electrical stimulation
electro-reflectance
electron diffraction
electron impact
electron ionization

electron microscope
electron microscopy
electron volt
elongation complex
elongation factor
elutriated fractions
embryo length
embryo proper
embryoid body
embryonal carcinoma
embryonic germ
embryonic length
embryonic stem
Emiliania huxleyi
empty vector
enantiomeric excess
encephalization quotient
encounter complex
endolymphatic duct
endoplasmic reticulation
endoplasmic reticulum
endoplastic reticulum
endothelial cell
endotoxin unit
endurance running
energy expenditure
energy gaps
energy transfer
enhancer-promoter
Enlow Fork
enteric sac
entorhinal cortex
entropy variance
Enzyme Commission
equal opportunities
equilibrium midpoint
estrogen receptor
ethidium bromide
ethmoid plate
ethylene glycol
ethylene oxide
European Commission
European Community
European Union
evaporation residue
evening element
event model
evolutionarily stable
evolutionary strategy
exclusion limit
expanded-pupil
Expectation Maximization
export production
expression site
external tufted
extracellular polysaccharide
extraembryonic ectoderm
Eya domain
eye height
eye separation Faeroe platform
familial hypercholesterolaemia
familiar instrumentals
far-field
far-red

Faraday mirrors
Faraday rotation
farm fallow
fast-spiking
fatty acids
fear-conditioned
feedback-based
female pronucleus
Feni drift
Fermi surface
Fermi-liquid
fetal liver
fibrous sheath
field cooling
field divergence
field emission
Field potentials
field-cooled
field-cooling
fill factor
filter analyser
fixed interval
fixed ratio
flip angle
floor plate
flow regime
flow-through
fluctuating asymmetry
fluctuation index
fluctuation-dissipation
fluorescein maleimide
focal adhesion
follicle cell
follicular lymphoma
follicular mantle
food-deprived
food-sated
foramen magnum
fork length
formin-homology
forward scatter
forward-strand
Four-jointed
Fourier transform
Fourier transformation
Fourier transforms
Fourier-filtered
Fourier-transformed
fraction modern
fractionation chamber
fracture zone
frequency modulated
frequency modulation
frequency-modulated
fronto-orbital
full length
fully relaxed
fuzzy-spreader
Galactic Centre
ganglion cell
ganglionic eminence
gas chromatograph
gas chromatographic
gas chromatography

gastrulation defective
gene conversions
gene ontology
gene product
General Atomics
General Electric
General Motors
genetic modification
genetically modified
Genome Ontology
genomic alignment
Georgetown University
Georgia passage
Geoscience Australia
germ cells
germ tubes
germinal centre
germinal vesicle
germination medium
gestational week
Giant interneurons
gibberellic acid
glucocorticoid receptor
glucose-responsive
Gly-Ser
Golgi apparatus
Golgi complex
Goodenough basin
graafian follicles
grain size
granular zone
granule cell
granulocyte-macrophage
granulosa cells
Greenland Sea
Greenland stadials
groove width
ground state
growth cone
growth factors
growth hormone
growth medium
guanylate kinase
guard cells
Gulf Coast
Gutenberg-Richter

4.3.3.2 Jednostki trójwyrazowe

Abdur Reef Limestone
aberrant crypt foci
Above-Budget Foundation
above-threshold ionization
accelerated mass spectrometry
accelerator mass spectrometer

accelerator mass spectrometric
accelerator mass spectrometry
accelerator mass spectroscopy
accelerator-mass-spectrometry
accessible surface area
accessory olfactory bulb
accessory optic system
accretion-induced collapse
acetyl CoA carboxylase
acheate-scute homologue
achromatic Fresnel optic
acid citrate dextrose
acid tolerance response
acidic activation domain
acoustic doppler velocimeter
acousto-optic modulator
actin depolymerizing factor
action potential duration
action potential waveforms
Activation-Induced Deaminase
activator-recruited cofactor
Active galactic nuclei
active galactic nucleus
Active Server Pages
active zone material
activin-response element
activin-responsive factor
acute lymphoblastic leukaemia
acute lymphocytic leukaemia
acute myelogenous leukaemia
acute myeloid leukaemia
acute myeloid leukemia

acute promyelocytic leukaemia
acute promyelocytic leukemia
acyl carrier protein
acyl-CoA synthetase
acyl-homoserine lactone
acylated homoserine lactone
adaptive cruise control
Adaptive Sequence Kontrol
address-event-representation
adductor digiti minimi
adenine nucleotide translocator
adeno-associated viral
adeno-associated virus
Adenomatous Polyposis Coli
adenovirus-associated virus
ADP ribosylation factor
adrenal hypoplasia congenita
Advance Technology Upgrading
Advanced Bladder Cancer
Advanced Cell Technology
Advanced Composition Explorer
advanced diamond composite
Advanced Expert system
Advanced Granulation Technology
Advanced Light Source
Advanced Micro Devices
Advanced Neutron Source
Advanced Photon Source
Advanced Stokes Polarimeter
Advanced Technology Program
Advanced Tissue Sciences
Advertising Standards Authority

aerobic dive limit
Aerosol Characterization Experiments
aerosol optical thickness
African National Congress
after-hyperpolarizing potential
Agenzia Spaziale Italiana
Agricultural Research Service
Agulhas leakage fauna
Air Commando Squadron
air mass factors
air-saturated water
Airway hyper-responsiveness
airway surface fluid
airway surface liquid
Akaike Information Criteria
Akaike information criterion
Alamogordo Primate Facility
Alfred Wegener Institute
alkaline fuel cell
All-Sky-Monitor
allele-discriminating signal
allele-specific oligonucleotide
Allen Telescope Array
Alpha Magnetic Spectrometer
altered peptide ligand
alternating electric field
Alternating Gradient Synchrotron
alternative mating strategy
American Anthropological Association
American Astronomical Society
American Cancer Society
American Chemical Society

American Chemistry Council
American Film Institute
American Geological Institute
American Geophysical Union
American Health Foundation
American Heart Association
American Mathematical Society
American Petroleum Institute
American Physical Society
American Psychological Association
American Sign Language
American Statistical Association
amino trifluoromethyl coumarin
amorphous solid water
Amplified spontaneous emission
Amyloid precursor protein
amyloid protein precursor
amyotrophic lateral sclerosis
Analytical Information Systems
analytical quality control
anaphase promoting complex
Anaplastic lymphoma kinase
androgen-binding protein
angiotensin converting enzyme
angle of attack
Anglo-Australian Observatory
anhysteretic remanent magnetization
Animal Health Institute
Animal Liberation Front
Animal Rights Center
Annular dark field
anomalous cosmic ray

Ant Colony Optimization
Ant Colony Routing
Antarctic Circumpolar Current
Antarctic cold reversal
Antarctic Polar Front
anterior auditory field
anterior cingulate cortex
anterior definitive endoderm
anterior ectosylvian visual
anterior forebrain pathway
Anterior primitive streak
anterior sphenoid length
anterior visceral endoderm
Anterior visceral endodermal
Anthrax Vaccine Absorbed
Anti-Ballistic Missile
anti-Mullerian hormone
anti-Zeno effect
anticodon stem-loop
antigen presenting cell
antisocial personality disorder
Apache Point Observatory
apical ectodermal ridge
Apoptosis inducing factor
Appalachian State University
apparent column density
apparent oxygen utilization
Application Control Module
Applications Development Language
Applied Physics Laboratory
Arabidopsis Genome Initiative
Arabidopsis response-regulator

arabinose-binding protein
arachidonoyl methyl ester
arbuscular mycorrhizal fungi
Archer Daniels Midland
Arctic Intermediate Waters
area of interest
area of occupancy
area specific resistance
area specific resistivity
area-restricted searching
area-specific resistivity
Argonne National Laboratory
Arizona State University
Armadillo repeat domain
Armour Research Foundation
Army Research Office
aromatic infrared bands
Artemis Comparison Tool
artificial neural network
artificial sea water
aryl hydrocarbon receptor
Associated Universities Incorporated
Astronomy Diagnostic Test
Astrophysical Institute Potsdam
Astrophysical Research Consortium
asymptotic giant branch
asynchronous transfer mode
ataxia telangiectasia mutated
ataxia-telangiectasia-mutated
Atlantic equatorial discordance
Atlantic transform fault
Atlantis transform fault

Atmospheric Dynamics Mission
atmospheric pressure ionization
Atmospheric Radiation Measurement
atomic absorption spectroscopy
Atomic Energy Commission
Atomic Energy Council
Atomic force micrographs
Atomic force microscope
Atomic force microscopy
atomic transient recorder
atomic-force microscope
atomic-force microscopy
atomic-force-microscope
ATP binding cassette
ATP citrate lyase
ATP-binding cassette
atrial natriuretic factor
atrial natriuretic peptide
attenuated total reflectance
Attenuated total reflection
auditory brainstem response
Auger electron spectroscopy
austral volcanic zone
Australian Broadcasting Corporation
Australian National University
Australian Patent Office
Australian Research Council
Automated Image Registration
Autonomous Benthic Explorer
autonomous underwater vehicle
autonomously replicating sequences
auxiliary power unit

Available Chemicals Directory
avalanche photo diode
average squared difference
Average telomere length
avian mycloblastosis virus
avian myeloblastosis virus
avidin-biotin complex
axial magma chamber
axon initial segment
baby hamster kidney
BAC-end sequences
Bacillus Calmette-Guerin
Back-scattered electron
Bacterial artificial chromosome
bacterial artificial clone
bacterial carbon demand
bacterial growth efficiency
bacterial sulphate reduction
baculovirus iap repeat
Baja California series
banded iron formation
Barberton Greenstone Belt
Barro Colorado Island
basal cell carcinoma
basal growth medium
basal metabolic rate
basal transition thickness
base excision repair
basic calcium phosphate
bayesian posterior probability
Bcd response element
bean abscission cellulase

Beckwith-Wiedemann syndrome
Beijing Genomics Institute
Beipiao Paleontological Museum
Bell-state measurement
benign prostatic hyperlasia
Bernard Price Institute
Bharatiya Janata Party
biased saccade task
Big Bang nucleosynthesis
Bight fracture zone
Binational Science Foundation
Biochemie-Zentrum Heidelberg
Biodiversity Observatory Network
Biological Information System
biological species concept
Biological Weapons Convention
biological weighting function
Biology Concept Inventories
biomembrane force probe
Biotech Industry Organization
Biotechnology Industry Organization
Biotechnology Research Institute
biotic resistance hypothesis
biotin dextran amine
biotinylated dextran amine
Bis-Tris propane
bis-tris-propa
bit error rate
Black Hole Quenchers
Black Mexican Sweetcorn
black-hole candidates
Blue native electrophoresis

body mass index
body-centred cubic
body-centred-tetragonal
body-mass index
bond-length alternation
bond-order alternation
bone marrow transplant
bone marrow transplantation
bone mineral content
bone mineral density
Bone morphogenetic protein
bone morphogenic protein
bone-morphogenetic-protein
Bonneville Power Administration
Booster Applications Facility
Bordetella pertussis toxin
Born-Oppenheimer approximation
Bose Chaudhuri Hocquenghem
Bose-Einstein condensate
Bose-Einstein condensation
bottom simulating reflector
bottom-simulating reflection
bottom-simulating reflector
bovine brain capillary
Bovine papilloma virus
Bovine serum albumin
bovine spongiform encephalopathy
brain heart infusion
Brain Science Institute
brain-specific repetitive
branchio-oto-renal
Brewster angle microscopy

Brief Communications Arising
Bristol-Myers Squibb
British American Tobacco
British Geological Survey
British Ice Sheet
British Medical Association
British Medical Journal
British Oxygen Company
British Technology Group
broad absorption line
bronchial smooth muscle
Brookhaven National Laboratory
Brookhaven Science Associates
brown adipose tissue
Brunauer-Emmett-Teller
Buck Reef Chert
Building Research Establishment
bulk inversion asymmetry
bulk silicate Earth
bundle-forming pilus
Burroughs Wellcome Foundation
Burroughs Wellcome Fund
business-as-usual
Caenorhabditis Genetics Center
calcareous clayey mud
calcite compensation depth
calcium carbonate distribution
calcium-dependent antibiotic
Calderbank Shor Steane
calf intestinal phosphatase
calf intestine phosphatase
California Current System

California Healthcare Institute
California State University
CaM-binding domain
Cambridge Antibody Technology
Cambridge Structural Database
Cambridge Structure Database
Cambridge-MIT Institute
cAMP-responsive element
Canadian Light Source
Canadian Neutron Facility
Cancer Research Campaign
Cancer Research Institute
cancer stem cell
canine distemper virus
canonical correlation analysis
canonical discriminant analysis
canonical variates analysis
Cap-binding protein
Cape Basin record
Cape Roberts Project
Cape Verde Island
carbide-derived carbon
carbohydrate-recognition domains
Carbon Nanotechnologies Incorporated
Carbon Sequestration Initiative
carbon-concentrating mechanism
carbon-fibre electrode
carbonate compensation depth
carbonate-associated sulphate
carboxy-terminal domain
cardio-vascular accident
Career Development Scheme

carrier envelope offset
case fatality proportion
caspase-activated deoxyribonuclease
Caspase-activated DNase
catabolite-activating protein
catabolite-activator protein
catalysed signal amplification
category correlation score
cathode ray tube
caudal fastigial nucleus
cauliflower mosaic virus
Cbl-associated protein
CDK-activating kinase
Celera Discovery system
Celera Drosophila genome
cell division cycle
Central Awash Complex
central conduction time
Central England Temperature
Central Indian ridge
Central Intelligence Agency
central meridian longitude
central nervous system
central North Pacific
central pattern generating
central pattern generator
central processing unit
Central Science Laboratory
Central Veterinary Laboratory
centrally nucleated fibres
centre of excellence
centre-of-mass

Centroid Moment Tensor
Ceramic Mound Period
cerebral amyloid angiopathy
cerebral spinal fluid
cervical intraepithelial neoplasia
chain length distributions
chain-length factor
Chamber of Commerce
charge coupled device
charge density wave
Charge modulation spectroscopy
charge switch technology
charge-coupled device
charge-density wave
charge-parity-time
charge-transfer efficiency
charged coupled device
charged-couple-device
charged-coupled device
checkpoint-sliding clamp
Chemical Abstract Service
Chemical Diversity Labs
chemical force microscopy
Chemical Inventory System
chemical remanent magnetization
Chemical Storage Cabinet
Chemical Strategies Partnership
Chemical Technological University
chemical vapour deposited
chemical vapour deposition
Chemical Warfare Service
Chemical Weapons Convention

Chemical-vapour deposition
chemical-vapour-deposited
chemical-vapour-deposition
Chicken embryonic fibroblast
chief executive officer
Chief Medical Officer
chief scientific officer
China Bridges International
Chinese Chemical Society
Chinese hamster ovarian
Chinese hamster ovary
chloramphenicol acetyl transferase
Christian Social Union
chromated copper arsenate
chromatin-dependent coactivation
chromatography data system
chromo-shadow domain
Chronic granulomatous disease
chronic lymphocytic leukaemia
chronic lymphocytic leukemia
chronic myelogenous leukaemia
Chronic myeloid leukaemia
chronic wasting disease
ciliary beating frequency
circular standard deviation
circular variable filter
circularly polarized light
Circumpolar Deep Water
class switch recombination
classical nucleation theory
classical receptive field
clathrin heavy chains

clathrin-mediated endocytosis
Clay Mathematics Institute
Clean Development Mechanism
clear air turbulence
cleaved-edge overgrowth
Cleveland Clinic Foundation
Climate Action Network
Climate Prediction Index
clinical investigation centre
closed head injury
cloud condensation nuclei
Cloud droplet nuclei
cloud longwave forcing
codon adaptation index
Codon enrichment correlation
coefficient of performance
cofilin homology region
cognitive behaviour therapy
coherent spin state
Cold Dark Matter
collision-induced dissociation
colony forming units
colony-forming cells
colony-forming unit
Colony-forming-unit
Colorado State University
Commercial Space Centers
Common Agricultural Policy
common carotid artery
Common Cold Unit
common lymphoid progenitor
common mid-point

Common principal components
commonly deleted region
Community of Science
Comodulation masking release
Compact Muon Solenoid
Comparative genome hybridization
comparative genomic hybridization
Competitive Enterprise Institute
competitive ligand equilibration
complementarity-determining region
complete blood count
complete Freunds adjuvant
complete hydatidiform mole
Component Object Model
compound action potential
compound refractive lens
computable general equilibrium
computational fluid dynamics
computer-aided design
Computerized axial tomography
concentrating solar power
conditioned place preference
conductivity temperature depth
congenital heart disease
congenital myasthenic syndrome
congestive heart failure
Congressional Budget Office
Conservation Research Center
Conserved Domain Database
constitutive transport element
constitutively active receptor
contact potential difference

contingent negative variation
continuous composition spread
continuous phase diagram
Continuous Plankton Recorder
continuous wavelet transform
contrast transfer function
convergent close-coupling
cooled injection system
Coomassie brilliant blue
Cooper pair box
Cooperative Research Centre
cooperatively rearranging region
Copyright Clearance Center
core encapsidation signal
core-binding factor
core-mantle boundary
Coriell Cell Repository
Corn Refiners Association
cornmeal-sugar-yeast
coronal mass ejection
coronary artery disease
coronary heart disease
corotation eccentricity resonance
corotation inclination resonance
Corruption Perception Index
cortical collecting duct
corticotrophin-releasing-hormone
corticotropin-releasing factor
Corticotropin-releasing hormone
cosmic background imager
cosmic microwave background
cosmic ray subsystems

cosmic-ray exposure
cost of transport
Coulomb failure function
Coulomb failure stress
county-parish-holding
crack opening displacement
crassulacean acid metabolism
CREB binding protein
CREB-responsive elements
Crescent Island crater
Cretaceous normal superchron
Creutzfeldt-Jakob disease
critical community size
Crk-associated substrate
cross-phase modulation
cross-sectional area
cryo-thermochromatographic separator
cryo-thermochromatography separator
crystal preferred orientation
crystalline colloidal array
crystalline electric field
Csk-binding protein
CTD-interacting domain
cucumber bulgarian virus
cucumber mosaic virus
cued saccade task
cumulative volcano amplitude
current-source density
cutaneous lymphocyte antigen
cyan fluorescent protein
cyclic nucleotide-gated
cyclin-dependent kinase

Cyclobutane pyrimidine dimer
cysteine-rich domain
cytidine deaminase activity
cytokine receptor homologous
Cytolethal distending toxin
cytoplasmic localization domain
cytoplasmic polyadenylation element
cytotoxin-associated gene
daily distances travelled
Daily energy expenditure
Data Coordination Center
Data Coordination Centre
Daughters against decapentaplegic
daughters against dpp
day in vitro
day of year
days after pollination
days in vitro
days post-coitum
Dead Sea transform
dead-end elimination
death effector domains
death-effector domain
decay-accelerating factor
DED-recruiting domain
deep copper zone
deep low-frequency
Deep Space Network
deep-vein thrombosis
Defense Science Board
degree of pyritization
degrees of freedom

delayed sequence recall
delayed-memory-saccade
delayed-type hypersensitivity
Democratic Progressive Party
dengue haemorrhagic fever
dense rock equivalent
densities of states
density functional theory
Density of states
density-functional theory
density-of-states
Department of Defense
Department of Energy
Department of Health
Department of Justice
Departments of Defense
depleted MORB mantle
depth of focus
Desert Research Institute
detergent-insoluble membrane
detrended correspondence analysis
Deutsche Forschungs Gemeinschaft
developmental systems theory
deviance information criterion
Devon Great Consols
diamond anvil cell
diamond-like carbon
Diet-induced obese
diet-induced obesity
Difference frequency generation
difference-of-gaussian
Differential cross-sections

differential gene expression
Differential interference contrast
differential scanning calorimeter
Differential scanning calorimetry
Differential thermal analysis
differential-interference contrast
differentially methylated domain
differentially methylated region
diffractive optical element
diffusion-limited aggregation
digital elevation model
Digital Library Federation
Digital Object Identifier
digital signal processing
digital terrain model
direct numerical simulation
direct site factor
direction selectivity index
directional drying technique
directly observed therapy
discrete combinatorial synthesis
discrete dipole approximation
disjunctive normal formula
disparity-tuning index
dissimilatory sulphite reductase
dissipative phase transition
Dissolved inorganic carbon
dissolved inorganic nitrogen
dissolved inorganic phosphorus
dissolved organic carbon
dissolved organic material
dissolved organic matter

dissolved organic nitrogen
dissolved organic phosphorus
dissolved reactive phosphorus
distal primitive streak
distal projection index
Distal tip cell
distal visceral endoderm
distributed adaptive control
distributed annotation system
distributed-Bragg-reflector
dithiobis-succinimidyl propionate
diurnal temperature range
diversity-oriented synthesis
DNA damage response
DNA-binding domain
DNA-binding protein
DNAX-activating protein
domain wall resistance
Doppler Wind Experiment
dorsal arm plate
dorsal marginal zone
dorsal root ganglia
dorsal root ganglion
dorsal spinal cord
dorsal ventricular nerve
dorsal ventricular ridge
dorsal-root ganglion
dorso-anterior index
dosage compensation complex
double-strand break
double-stranded break
doubly conserved synteny

doubly labelled water
downdragged hydrous peridotite
downregulation targeting signal
downstream promoter element
downstream regulatory element
Draining lymph node
Drosophila odour receptor
Drosophila stress odorant
Drug Enforcement Administration
Dublin City University
Duchenne muscular dystrophy
Duffy-binding-like
Dust accumulation rate
Dust Detection System
dust detector system
dust veil index
DYAD symmetry element
Dynamic light scattering
dynamic random dot
Dynamic Reaction Cell
dynamic resonance units
dynein intermediate chain
early after-depolarization
early endosomal antigen
early somite stages
Earth Liberation Front
Earth Observing System
East Pacific Rise
eastern equatorial Pacific
eastern North Pacific
eastern volcanic zone
echo planar imaging

ecological-footprint analysis
eddy kinetic energy
Educational Testing Service
effective elastic thickness
effective maximally entangled
effective radiated power
Egyptian Islamic Jihad
Eidgenossche Technische Hochschule
Einstein-Podolsky-Rosen
Elastic Brownian Ratchet
elastin Van Gieson
electric organ discharge
Electrical conductivity method
electro-optic modulator
electro-optical modulator
electrochemical vapour-deposition
electromagnetically induced transparency
electron capture detector
electron cyclotron resonance
electron dispersive spectroscopy
electron localization function
electron microscope tomography
Electron paramagnetic resonance
Electron spectroscopic imaging
electron spin resonance
electron transfer chain
electron transport layer
electron transport rate
electron-capture detectors
Electron-spin resonance
electron-stimulated desorption
electron-transport layer

Electronic Frontier Foundation
electronic road pricing
Electronic Visualization Laboratory
electrostatic force microscopy
embedded atom models
Emissions Trading Scheme
empirical mode decomposition
empirical orthogonal function
empirical risk minimization
end-diastolic diameter
Endangered Species Act
endogenous reverse transcription
endothelial progenitor cells
enemy release hypothesis
Energetic neutral atom
energy balance model
energy dispersive spectrometry
energy dispersive spectroscopy
Energy Information Administration
Energy Information Agency
energy-dispersive spectrometry
Energy-dispersive spectroscopy
energy-recovered linac
engineering design agreement
enhanced disease resistance
enhanced disease susceptibility
Ensembl Genome Browser
enteric muscle contraction
entomological inoculation rate
Environmental Defense Fund
Environmental Genome Project
Environmental Impact Statement

Environmental Protection Agency
environmental scanning electron
Environmental Sustainability Index
Environmental Working Group
enzymatic mutation detection
eosinophil cationic protein
eosinophil-derived neurotoxin
Epidermal growth factor
epithelial growth factor
epithelial-mesenchymal transition
Epstein Barr Virus
equation of state
equatorial Pacific Ocean
equatorial upwelling zone
equilibrium line altitude
equine chorionic gonadotrophin
equivalent current dipole
erasable electrostatic lithography
error-related negativity
Erythrocyte-binding antigens
essential light chain
estimated clutch frequency
Etch pit density
ethyl methane sulphonate
eucrite parent body
eukaryotic initiation factor
European Bioinformatics Institute
European Food Authority
European Neuroscience Association
European Neuroscience Institute
European Patent Convention
European Patent Office

European Phenology Network
European Physical Society
European Remote Sensing
European Research Area
European Research Council
European Science Foundation
European Social Fund
European Southern Observatory
European Space Agency
European Spallation Source
European VLBI Network
event-related potential
evolutionarily stable strategies
evolutionarily stable strategy
excitation-contraction coupling
excitatory amino acid
exclusive economic zone
exercise treadmill time
exonic splicing silencer
experimental allergic encephalomyelitis
Experimental autoimmune encephalomyelitis
Expressed sequence tag
expressed-sequence-tag
Extended Glaciological Timescale
extended haplotype homozygosity
extensor digitorum longus
external elastic laminae
external germinal layer
external granular layer
external granule layer
external plexiform layer
external quantum efficiency

external transcribed sequence
extra-pair young
extracellular enzyme activity
extracellular polymeric secretions
extracellular regulated kinase
Extreme Energy Events
Extremely Large Telescope
extremely low frequency
extremely metal-poor
extremely-low-frequency
extrinsic hyperpolarizing potentials
face-centred-cubic
false discovery rate
familial hypertrophic cardiomyopathy
far-infrared background
Faraday fracture zone
Faraday-rotation feature
Farm-Scale Evaluations
farnesyl pyrophosphate synthetase
Faroe Bank channel
Fast atom bombardment
fast Fourier transform
fat pad mass
fatal familial insomnia
fatty acid synthase
fear of intimacy
Federal Aviation Administration
Federal Communications Commission
Federal Security Service
Federal Trade Commission
feline immunodeficiency virus
feline spongiform encephalopathy

Fennoscandian ice sheet
ferric uptake regulator
fertilization-independent seed
fetal bovine serum
Fetal calf serum
Fibroblast growth factor
fibroblast-like synoviocytes
Field effect transistor
field of view
field-effect transistor
field-emission microscopy
figure of merit
figures of merit
filamentary sublimate residues
Final States Effect
Fine Guidance Sensors
finite element method
finite element model
finite-element analysis
first dorsal interosseous
first ionization potential
first mass function
flame ionization detection
flame ionization detector
flanking sequence tag
flavin adenine dinucleotide
flexor pollicis brevis
Flight Operations Segment
Florida State University
fluorescent intensity units
fluorescent speckle microscopy
fluoro orotic acid

Flux balance analysis
Flux Gate Magnetometer
flux-line lattice
focal adhesion kinase
focal adhesion targeting
focal plane array
focused ion beam
focused library synthesis
focused-ion-beam
foetal bovine serum
folded longitudinal acoustic
follicle stimulating hormone
follicular dendritic cell
Food Standards Agency
Force Concept Inventory
Forensic Science Service
forest products sector
fork-head-associated
former Soviet Union
Fos-related antigens
four-wave mixing
Fourier shell correlation
Fraction attributable risk
fractional quantum Hall
Fractional solvent accessibility
Frankfurt Biosphere Model
free energy perturbation
free fatty acid
free running laser
free spectral range
free-electron laser
free-energy perturbation

free-flow electrophoresis
free-induction decay
Free-space optics
friction force microscopy
Friedrich Ebert Stiftung
frizzled-related protein
frontal eye field
frontonasal ectodermal zone
FUSE-binding protein
fusiform face area
gain-assisted superluminality
gain-of-function
galactic cosmic rays
Galileo Europa Mission
gametocyte activating factor
gamma activation sequence
Gamma Ray Spectrometer
gamma-ray burst
ganglion mother cell
gap excess ratio
Garnier-Osguthorpe-Robson
gas electron multiplier
gas imaging spectrometers
gastric cancer relatives
gastric inhibitory peptide
gastric inhibitory polypeptide
gated superconducting strip
GDP dissociation inhibitor
Geikie Plateau Formation
gel permeation chromatography
Gene Expression Omnibus
Gene Network Sciences

gene-specific primer
General Accounting Office
general additive models
general circulation Model
general import pore
General linear model
general linearized model
General Medical Council
General Services Administration
general time-reversible
generalized extreme value
generalized gradient approximation
generalized least squares
generalized linear modelling
generalized Pareto distribution
generalized phase contrast
generalized valence bond
generation of diversity
Genetic Data Environment
genetic interaction density
genetic suppressor element
Genetics Computer Group
Genome Sciences Center
Genomic Science Centre
Genomic Sciences Center
genotype-relative risk
geocentric axial dipole
geocentric solar magnetospheric
Geographic Information System
Geographical Information System
geometric morphometric methods
George Washington University

Georgia Research Alliance
Geospace Environmental Modeling
geosynchronous transfer orbit
geranyl-geranyl-transferase
Germ Cell-less
germ-band elongation
German Democratic Republic
German Israeli Foundation
German-Israel Foundation
germinal matrix haemorrhage
giant molecular cloud
giga-electron volts
glacial isostatic adjustment
glass bottom boat
glass-forming ability
GlcNAc-terminated fetuin
glial fibrillary acidic
Global Biodiversity Assessment
global circulation model
Global Climate Coalition
global climate model
global dichotomy boundary
Global Environment Facility
Global Positioning System
Global Seismic Network
Global Seismographic Network
global spectral model
Global Taxonomy Initiative
global-positioning-system
globular calcified cartilage
glucocorticoid-response element
glucose phosphate isomerase

glucose uptake rate
glucose-binding protein
glutamic acid decarboxylase
Gly-Ser-Gly
glycogen synthase kinase
glycosyl-phosphatidyl inositol
GNU Public License
Good Laboratory Practice
good manufacturing practice
goodness of fit
Government Accountability Office
Government Accounting Office
Government Pharmaceutical Organization
gradient gel electrophoresis
Graduate Record Examination
graft-versus-tumour
grain boundary sliding
Gran Telescopio Canarias
Grand Unified Theory
graphitic black carbon
GRB Coordinates Network
Great Barrier Reef
Great Red Spot
Green Bank Telescope
green flourescent protein
green fluorescence protein
Green Fluorescent Protein
green fluorescent proteins
green-fluorescent-protein
Greenberger-Horne-Zeilinger
Greenland Ice Sheet
Greenwich Electronic Time

Gross Domestic Product
gross national product
Gross primary production
ground penetrating radar
ground reaction force
ground surface temperature
ground-penetrating radar
growth hormone secretagogue
GTPase accelerating proteins
GTPase activating protein
GTPase effector domain
GTPase-activating protein
GTPase-binding domain
guard mother cell

4.3.3.3 Jednostki czterowyrazowe

AIDS Vaccine Advocacy Coalition
ab initio molecular dynamics
aboveground net primary production
Academic Strategic Alliances Program
Accelerated Climate Prediction Initiative
Accelerated Strategic Computing Initiative
acid-sensing ion channel
acid-sensing ionic channel
acidic basic repeat antigen
acidic seminal fluid protein
acoustic Doppler current profiler
acquired immune deficiency syndrome
acridine orange direct counting
activated partial thromboplastin time
activation-induced cell death
activity-dependent neuroprotective protein
acyl-CoA binding protein
adsorptive cathodic stripping voltammetry
adult respiratory distress syndrome
Advanced CCD Imaging Spectrometer
Advanced Fibre-Optic Echelle
Advanced Land Observation Satellite
Advanced National Seismic System
advanced sleep-phase syndrome
Advanced Strategic Computing Initiative
Advanced Technology Solar Telescope
advection-dominated accretion flow
African Agricultural Technology Foundation
African American Vernacular English

African horse sickness virus
aldosterone secretion inhibiting factor
All India Biotech Association
all-taxa biodiversity inventory
all-taxa biological inventory
allele-specific transcript quantification
Alliance for Cell Signaling
Alliance for Cellular Signaling
Alliance for Cellular Signalling
Amboseli Elephant Research Project
American Anti-Vivisection Society
American Corn Growers Association
American Crop Protection Association
American Spinal Injury Association
American Type Culture Collection
American Type Culture Corporation
among-site rate variation
amorphous tin composite oxide
amplified fragment length polymorphism
analytical transmission electron microscopy
anatomically modern Homo sapiens
Andean low-level jet
Andean low-velocity zone
anomalous low-frequency dispersion
anorthosite-mangerite-chanockite-granite
Antarctic Marine Living Resources
antibody dependent cellular cytotoxicity
application-specific integrated circuit
Arabidopsis Biological Resource Center
Archeological Resources Protection Act
Arctic National Wildlife Refuge
Arctic Ocean Deep Water

aromatic amino-acid hydroxylase
artificial cerebral spinal fluid
arylhydrocarbon receptor nuclear translocator
as Antarctic Intermediate Water
ash-free dry mass
Asia Pacific Economic Cooperation
Atacama Large Millimeter Array
Atacama Large Millimetre Array
atmosphere general circulation model
atmospheric general circulation model
atmospheric pressure chemical ionization
atom transfer radical polymerization
Atomic Energy Control Board
Atomic Energy Regulatory Board
Australia Telescope Compact Array
Australia Telescope National Facility
Australian Geological Survey Organisation
average well colour development
back-scattered electron imagery
backward elimination multiple regression
Ballistic Missile Defense Organization
Base Excision Sequence Scanning
basic fibroblast growth factor
basic helix-loop-helix
benign familial neonatal convulsions
benign gonial cell neoplasm
Bent Hill massive sulphide
Berkeley Digital Seismic Network
Berkeley Drosophila Genome Project
Bermuda Atlantic Time Series
best linear unbiased prediction
Best linear unbiased predictors

Biodiversity Conservation Information System
Biogeochemical Ocean Flux Study
bioluminescence resonance energy transfer
Biomedical Informatics Research Network
Biomedical Primate Research Centre
Biomolecular Engineering Research Institute
Biomolecular Interaction Network Database
black hole accretion rate
blood oxygen level-dependent
blood oxygenation level-dependent
Board of Scientific Counselors
body-cavity-based lymphoma
Bombay Natural History Society
Bombyx cysteine proteinase inhibitor
bone marrow-derived macrophages
bovine aortic endothelial cells
bovine pancreatic trypsin inhibitor
bovine viral diarrhoea virus
Brain lipid-binding protein
Brain Molecular Anatomy Project
brain tumour initiating cells
brain-derived neurotrophic factor
branched-chain amino acids
Breast Cancer Research Program
British National Space Centre
British North Greenland Expedition
British Nuclear Fuels Limited
brush-border membrane vesicles
buffered saline-glucose-citrate
Bulk Solar System Initial
calcitonin gene related peptide
calcium-dependent protein kinase

calcium-induced calcium release
calcium-release-activated calcium
calf intestinal alkaline phosphatase
Caltech Parkes Swinburne Recorder
Canada-France-Hawaii Telescope
Canadian Galactic Plane Survey
Canadian national seismograph network
Cancer Chromosome Aberration Project
Cancer Chromosome Aberrations Project
cancer genome anatomy project
Cancer Research Campaign Technology
caprine arthritis-encephalitis virus
carbonaceous chondrite anhydrous mineral
carboxy-terminal cysteine-knot
Cardiac Arrhythmia Suppression Trials
Carica papaya BAC ends
Carnegie Ames Stanford Approach
cascade-induced source hardening
Case Western Reserve University
catch per unit effort
Cavaleri Ottolenghi Scientific Institut
Census of Marine Life
Central Atlantic Magmatic Province
Central Public Health Laboratory
Central Rice Research Institute
Cerro Tololo Interamerican Observatory
Certified Reference Materials Program
Charles Darwin Research Station
Charlie-Gibbs fracture zone
Chinese Economic Benefit Federation
Cholesterol and Recurrent Events
cholesterol-efflux regulatory protein

cholesteryl ester transfer protein
chromophore-assisted laser inactivation
chromophore-assisted light inactivation
clamped homogeneous electric field
classical swine fever virus
clathrin heavy-chain repeat
Clauser-Horne-Shimony-Holt
Clemson University Genomics Institute
Climate Change Science Program
cloud droplet number concentration
co-structure-directing agent
Coastal Zone Color Scanner
coastal zone colour scanner
cocaine-amphetamine-regulated transcript
code division multiple access
cold month mean temperature
Cold Spring Harbor Laboratory
College Entrance Examination Board
Columbia Accident Investigation Board
Committee on Publication Ethics
Community Multiscale Air Quality
community-level physiological profiles
Compact Airborne Spectrographic Imager
Compassion in World Farming
complementary metal oxide semiconductor
complementary metal oxide silicon
Comprehensive Test Ban Treaty
Comprehensive Yeast Genome Database
Compton Gamma Ray Observatory
Confocal laser scanning micrograph
confocal laser scanning microscope
confocal laser scanning microscopy

confocal scanning optical microscope
connective tissue growth factor
conserved non-genic sequences
constant carbon accumulation model
Constant denaturant capillary electrophoresis
conventional transmission electron microscope
convergent-beam electron diffraction
Cornell Electron Storage Ring
corrected integrated path length
Correlated Electron Research Center
cosmic microwave background radiation
Council on Governmental Relations
coupled general circulation model
Coupled Model Intercomparison Project
cowpea chlorotic mottle virus
critical resolved shear stress
cross linked enzyme crystals
crown-of-thorns starfish
cryogenic transmission electron microscopy
CSP-TRAP-related protein
cyclic nucleotide-binding domain
cytoadherence-linked asexual gene
cytoplasmic dynein intermediate chain
cytoplasmic transductional-transcriptional processor
cytosolic factor-associating domains
Dana-Farber Cancer Institute
death-inducing signaling complex
death-inducing signalling complex
deceased donor liver transplantation
Deep Sea Drilling Program
Deep Sea Drilling Project
deep western boundary current

deep-hypersaline anoxic basin
Defence Evaluation Research Agency
Defence Research Development Organization
degenerate four-wave mixing
Degree Angular Scale Interferometer
delayed sleep phase syndrome
delta sleep inducing peptide
denaturing gradient gel electrophoresis
Denmark Strait Overflow Water
Dense wavelength-division-multiplexed
dense wavelength-division-multiplexing
Department for International Development
descending contralateral movement detector
Descent Imaging Spectral Radiometer
Descent Trajectory Working Group
Detrended canonical correspondence analysis
Developmental Studies Hybridoma Bank
differential in-gel electrophoresis
differential optical absorption spectroscopy
Digital Millennium Copyright Act
digital particle image velocimetry
digital particle imaging velocimetry
direct methanol fuel cell
Disease Control Priorities Project
Donostia International Physics Center
dorsal longitudinal anastomosing vessel
dorsal root entry zone
dorsocaudal medial entorhinal cortex
double-barrier tunnel junction
double-strand break repair
Downward-Looking Visible Spectrometer
Drosophila CREB-binding protein

Drosophila insulin-like peptides
ductal carcinoma in situ
Dulbecco-modified Eagle medium
dynamic random access memory
dynamical mean-field theory
dysplastic aberrant crypt foci
dystrophia myotonia protein kinase
East Antarctic Ice Sheet
Eifel Laminated Sediment Archive
Electric Power Research Institute
electrochemical quartz crystal microbalance
electron back-scattered diffraction
electron energy loss spectroscopy
electron energy-loss spectra
electronic polymerase chain reaction
electrophoresis mobility shift assay
electrophoretic mobility shift assay
electrophoretic mobility-shift analysis
embryonic fibroblast growth factor
embryonic myosin heavy chain
encoding green fluorescent protein
endoplasmic-reticulum-associated degradation
endothelial nitric oxide synthase
endothelial nitric oxide synthetase
endothelial-leukocyte adhesion molecule
endothelium-derived hyperpolarizing factor
endothelium-derived relaxing factor
Energy Systems Research Unit
enhanced cyan fluorescent protein
enhanced green fluorescence protein
enhanced green fluorescent protein
enhanced yellow fluorescent protein

Ente Nazionale Energie Alternative
Environmental Risk Management Authority
environmental scanning electron microscopy
Environmental Systems Research Institute
Epidemic Type Aftershock Sequence
epidermal growth factor receptor
epithelial growth factor receptor
epizootic haemorrhagic disease virus
equal channel angular pressing
Equal Employment Opportunity Commission
equine infectious anaemia virus
equivalent atmospheric optical thicknesses
erbium-doped fibre amplifier
erythroid Kruppel-like factor
ethylene diamine tetraacetic acid
Eurasian basin bottom waters
Eurasian basin deep waters
European Brain Research Institute
European Business Angel Network
European Cell Biology Organization
European Life Science Forum
European Life Sciences Forum
European Life Sciences Organization
European Life Scientist Organization
European Marine Biology Symposium
European Medical Evaluation Agency
European Medical Research Councils
European Medicines Evaluation Agency
European Molecular Biology Conference
European Molecular Biology Laboratory
European Molecular Biology Organisation
European Molecular Biology Organization

European Mouse Mutant Archive
European Particle Accelerator Conference
European Photon Imaging Cameras
European Space Operations Centre
European Space Research Organization
European Space Science Committee
European Synchrotron Radiation Facility
European Technology Assessment Network
evaporation-induced self-assembly
expanded simple-tandem repeat
expression quantitative trait loci
expression site associated gene
extended cavity diode laser
external cavity diode laser
External RNA Controls Consortium
Extra-Solar Planet Imager
extra-tropical southern hemisphere
FAK related non-kinase
familial amyotrophic lateral sclerosis
family-wise error rate
Famine Early Warning System
Far Ultraviolet Spectrometer Explorer
Far Ultraviolet Spectroscopic Explorer
far-off-resonance trap
Fas-associated death domain
fast centrifugal partition chromatography
fast performance liquid chromatography
fast protein liquid chromatography
fast-repetition-rate fluorometry
fatty acid amide hydrolase
fatty acid methyl esters
fatty acid methyl esther

fatty acid transport protein
fatty acyl-CoA ligase
fatty-acid amide hydrolase
Federal Advisory Committee Act
Federal Emergency Management Agency
feline infectious peritonitis virus
ferric-reducing antioxidant potential
fibroblast growth factor receptor
Field Programmable Gate Array
field-programmable gate array
finite difference time domain
finitely extensible nonlinear elastic
first sharp diffraction peak
floor plate-conditioned medium
flow-assisted cell sorting
flowing-afterglow Langmuir probe
fluorescence activated cell sorter
fluorescence activated cell sorting
fluorescence in situ hybridization
fluorescence in-situ hybridization
fluorescence loss in photobleaching
fluorescence recovery after photobleaching
fluorescence resonance energy transfer
fluorescence resonance excitation transfer
fluorescence-activated cell sort
fluorescence-activated cell sorting
fluorescence-in-situ-hybridization
fluorescent analytical cell sorter
fluorescent resonance emission transfer
fluorescent resonance energy transfer
fluorescent-activated cell sorting
fluorophore-assisted light inactivation

Fms intronic regulatory element
focal laser-induced photolysis
Foothills Wildlife Research Facility
forward scattering spectrometer probe
forward-angle light scatter
Foundations of Computer Science
Fourier transform mass spectrometry
fractional quantum Hall effect
free left Alu monomer
Freedom of Information Act
frequency resolved optical gating
Frontier Collaborative Research Center
fronto-polar prefrontal cortex
Fukuyama congenital muscular dystrophy
full-width half-maximum
full-width zero-intensity
fully differential cross-sections
function magnetic resonance imaging
functional magnetic resonance imaging
Fungal Genetics Stock Center
Future Large Hadron Collider
Gamma-Ray Burst Monitor
Gas Chromatograph Mass Spectrometer
gas immersion laser doping
Gemini Deep Deep Survey
Gemini Multi-Object Spectrograph
Gene Set Enrichment Analysis
generalized linear mixed model
Genetic Engineering Approval Committee
genetically modified herbicide-tolerant
Genetically modified nematode-resistant
Geochemical Earth Reference Model

Geomagnetic Polarity Time Scale
Geophysical Fluid Dynamics Laboratory
Geostationary Operational Environmental Satellite
Geosynchronous Satellite Launch Vehicle
Germ cell nuclear factor
German Regional Seismic Network
germinal center kinase-related
Giant Segmented Mirror Telescope
glia fibrillary acidic protein
glial fibrillary acid protein
glial fibrillary acidic protein
glial-derived neurotrophic factor
Global Biodiversity Information Facility
Global Change Research Program
Global Climate Observing System
global mean sea level
global ocean heat content
Global Ocean Observing System
Global Ozone Monitoring Experiment
Global Population Dynamics Database
global precipitation climatology project
global sea level rise
Global Terrestrial Observing System
glucose-stimulated insulin secretion
glutamate acid-rich protein
Goddard High Resolution Spectrograph
Gona Palaeoanthropological Research Project
Goose egg-white lysozyme
graft-versus-host disease
granulocyte colony stimulation factor
granulocyte colony-stimulating factor
Greek Atomic Energy Commission

Greenland Arctic Intermediate Water
Growth hormone-releasing peptide
growth-hormone-releasing hormone
guanine-nucleotide-binding protein
guanylate cyclase activating protein
gut-associated lymphatic tissue
gut-associated lymphoid tissue

4.3.3.4 Jednostki pięciowyrazowe

Academia Sinica Plant Genome Center
Advanced Very High Resolution Radiometer
African Regional Industrial Property Organization
Akeno giant air shower array
Allele specific in situ hybridization
American Lebanese Syrian Associated Charities
amplified ribosomal DNA restriction analysis
Arctic mid ocean ridge experiment
autosomal dominant polycystic kidney disease
Basic Energy Sciences Advisory Committee
Basic local alignment search tool
bayesian Markov chain Monte Carlo
Biological Response Modifiers Advisory Committee
British American Security Information Council
British Institutions Reflection Profiling Syndicate
Burst and Transient Source Experiment
canalicular multi organic anion transporter
Canopy Operation Permanent Access System
Carbon Cycle Model Linkage Project
Carbon Dioxide Information Analysis Center
cellular retinoic acid binding protein
chemically assisted ion beam etching
Chinese Field Epidemiology Training Program

circular restricted three body problem
Clean Air Scientific Advisory Committee
climate leaf analysis multivariate program
Comprehensive Ocean Atmosphere Data Set
Continuous Electron Beam Accelerator Facility
Cooperative Research and Development Agreement
Cornell High Energy Synchrotron Source
cross polarization magic angle spinning
cross validation leave one out
Defence Advanced Research Projects Agency
Defense Advanced Research Program Agency
Defense Advanced Research Project Agency
Defense Advanced Research Projects Agency
denaturing high performance liquid chromatography
Department of Defense Polygraph Institute
double stranded RNA binding domain
double stripped fetal calf serum
edge emitting light emitting diode
Electronic Delay Storage Automatic Calculator
Energetic Gamma Ray Experiment Telescope
Energy Technical Scientific Research Institute
erasable programmable read only memory
extended dynamical mean field theory
Faint Object Camera and Spectograph
field emission scanning electron microscope
Fourier transform ion cyclotron resonance
Fourier transformed ion cyclotron resonance
Fred Hutchinson Cancer Research Center
Fusion Energy Science Advisory Committee
Fusion Energy Sciences Advisory Committee
Fusion Energy Sciences Advisory Council
Glacial North Atlantic Intermediate Water

global mean land surface temperature
Global Rural Urban Mapping Project
granulocyte macrophage colony stimulating factor
graphite furnace atomic absorption spectrometry
Gravity Recovery and Climate Experiment
Great Observatories Origins Deep Survey

4.3.3.5 Jednostki sześciowyrazowe

Antarctic muon and neutrino detector array
autosomal dominant non syndromic hearing loss
complete active space self consistent field
dorsal root acid sensing ion channel
Early Post Earthquake Damage Assessment Tool

4.3.4 Możliwości rozbudowy metody i dalszych badań

Co do zasady opisana metoda jest niezależna od konkretnego języka, ponieważ może być zastosowana do badania każdego języka, w którym występują akronimy będące pierwszymi literami wyrazów (dokładnie – ograniczonych spacją lub łącznikiem ciągów znaków literowych). Ponadto, do prowadzenia ekscerpcji nie jest wymagana znajomość danego języka (jedynie etap ostatni, szósty, analizy, który nie jest absolutnie konieczny, wymaga kompetencji językowych, a także językoznawczych). Ograniczeniem metody jest jednak to, że akronim musi następować bezpośrednio po jednostce wielowyrazowej (to znaczy po jednostce występuje spacja, następnie nawias otwierający i akronim), aby było możliwe jego wyodrębnienie. W języku polskim, na przykład, w szczególności w tekstach nienaukowych, akronimy często pojawiają się po odpowiadających im ciągach słów, jednak nie bezpośrednio. Odstęp między jednostką wielowyrazową a akronimem może obejmować kilka słów, zdanie lub nawet kilka zdań, co wyklucza lub co

najmniej znacznie utrudnia możliwość automatycznego jego wyodrębnienia. Poniżej podano przykłady z dziennika „Gazeta Wyborcza”:

> (…) szefowa Urzędu Komunikacji Elektronicznej Anna Streżyńska naciska na obniżki opłat (…).
>
> Szefowa UKE bez skrępowania szafuje przy tym wysokimi karami (…).
>
> Gazeta Wyborcza, 17 października 2006

> (…) Jerzy Polaczek postanowił, że poborem opłat (…) zajmie się Generalna Dyrekcja Dróg Krajowych i Autostrad. (…) W budżecie GDDKiA (…) przewidziano tylko 20 mln zł na prace związane z dostosowaniem do standardów autostrady (…).
>
> Gazeta Wyborcza, 17 października 2006

Ponieważ za pomocą metod w rodzaju przedstawionych w niniejszej pracy (wyrażenia regularne) rozwiązanie tego problemu zapewne nie byłoby możliwe, wydaje się, że w zastosowaniu opisanej tu metody istotnie ważny jest dobór tekstów do badań, który stanowi częściowe rozwiązanie problemu. Teksty o charakterze naukowym i technicznym zawierają znaczną liczbę jednostek wielowyrazowych, co wymusza pewną dyscyplinę w pisaniu tekstu (podawanie akronimu bezpośrednio po formie pełnej). W tekstach prasowych dyscyplina taka nie jest konieczna. Mimo że akronimy są stosowane we wszystkich niemal typach tekstów (prawdopodobnie stosunkowo najrzadziej w tekstach literackich, szczególnie poezji), ich liczba i różnorodność w tekstach nienaukowych jest bardzo ograniczona, co zmniejsza użyteczność metody, obniżając bezwzględną liczbę wyodrębnionych jednostek lub

wymusza zastosowanie odpowiednio większych korpusów bądź zbiorów tekstów.

Jeśli na uwadze mieć możliwości udoskonalenia metody, do zwiększenia jej wydajności, czyli liczby wyszukanych jednostek, można by zastosować wyrażenie regularne lub grupę wyrażeń obejmujących więcej możliwości wyszukiwania. Na przykład można by uwzględnić występowanie pewnych części mowy, na przykład przyimków, których często w akronimach się nie uwzględnia, na przykład *of* w *United States of America (USA)*. Nie miałoby to jednak wpływu na podstawowe założenia metody, której przedstawienie jest celem niniejszej części pracy.

4.3.5 Dyskusja wyników

Metoda zaproponowana w tej części pracy jest stosunkowo prosta w zastosowaniu, pozwala w sposób sprawny uzyskać interesujące wyniki i nie wymaga specjalistycznego oprogramowania, umiejętności programistycznych bądź dużej mocy obliczeniowej komputera. Może być wykorzystana do badania każdego korpusu lub zbioru tekstów z uwzględnieniem oczywistych ograniczeń, które pokrótce omówiono poniżej.

Ponieważ opisywana metoda nie wykorzystuje obliczeń statystycznych, nie jest wymagana duża wielkość korpusu, chociaż – co oczywiste – im większy korpus, tym prawdopodobnie większa liczba wyodrębnionych n-gramów i znalezionych jednostek wielowyrazowych.

Metodę można optymalizować i dostosowywać do charakterystyki poddawanego analizie korpusu i języka przez wprowadzanie ewentualnych zmian do wyrażenia regularnego stosowanego na etapie ekscerpcji. W ten sposób można by prawdopodobnie zwiększyć skuteczność metody, czyli liczbę wyodrębnionych unikatowych jednostek wielowyrazowych.

Ponieważ metoda nie opiera się na obliczeniach statystycznych, w przeciwieństwie do wielu innych metod wykorzystywanych przez badaczy, ale wykorzystuje wyróżniki (w postaci akronimów) pozwalające wyodrębnić jednostki wielowyrazowe w tekście, wynikiem jej zastosowania jest automatycznie wygenerowana lista jednostek, którą można bezpośrednio wykorzystać do innych badań językoznawczych, w szczególności – leksykograficznych.

Dzięki przyjętej metodzie wyszukiwania możliwe jest wyodrębnienie leksemów wielowyrazowych występujących w tekście lub ich zbiorze rzadko, także, jeśli forma rozwinięta została użyta jednokrotnie, przy czym po niej występował jej akronim, zaś w dalszej części tekstu posługiwano się wyłącznie akronimem. Takie pojedyncze lub rzadkie wystąpienia jednostek wielowyrazowych trudno wyodrębnić metodami statystycznymi. Możliwość ich wynotowywania przy użyciu zaproponowanej metody stanowi jej największą zaletę (tak więc jest to metoda stosunkowo skuteczna w ekscerpcji jednostek tego typu).

Przyjęte założenie i prostota metody – mimo że w znacznym stopniu korzystne w odniesieniu do realizacji metody oraz otrzymywanych wyników – determinują jednak jej ograniczenia.

Nie jest to metoda umożliwiająca ekscerpcję wszystkich lub nawet większości ciągów słów, które można uznać za jednostki wielowyrazowe, z danego tekstu lub zbioru tekstów. Jest nierozerwalnie związana z akronimami, których niewątpliwie nie stosuje się w przypadku wszystkich ciągów słów tworzących jednostki wielorazowe. Można jednak dowodzić, że im bardziej dana fraza ma charakter stały (większe prawdopodobieństwo współwystępowania jej składników), i im częściej w tekście występuje, tym większe jest prawdopodobieństwo, że zostanie użyta następująca po niej forma skrócona. Zwiększa to przydatność wynikowej listy jednostek w dalszych badaniach, ponieważ wyodrębnia się jedynie te jednostki wielowyrazowe, które mają charakter najbardziej stały.

Największą skuteczność metody osiąga się w przypadku tekstów, w których akronimy występują często, czyli tekstów naukowych i technicznych. W innych typach tekstów także wykorzystuje się akronimy, w mniejszym jednak stopniu, co może znacząco zmniejszyć liczbę wyodrębnionych leksemów.

4.4 Metoda ekscerpcji kolokacji rzeczownikowych na podstawie obserwacji parametru końcówki liczby mnogiej

4.4.1 Podstawowe założenie

Zagadnieniem będącym punktem wyjścia dla prac, których wyniki przedstawiono w niniejszej części rozprawy, było sformułowanie prostej i skutecznej metody ekscerpcji kolokacji rzeczownikowych ze zbioru tekstów. *Prosta* oznacza, że metodę mogą stosować językoznawcy nieoperujący aparatem programistycznym. Wykorzystuje się jedynie edytor tekstowy, oprogramowanie do podziału tekstu na n-gramy (patrz Manning, Schütze 1999, Banerjee, Pedersen 2003) i analizator morfologiczny[70]. Metoda, jednak, powinna być odpowiednio skuteczna i wydajna, czyli dostarczać możliwie obszernej listy kolokacji. Ponadto należałoby w jak największym stopniu ograniczyć konieczność analizy manualnej wyników.

Jak dowodzą wyniki zastosowanej metody, istnieje cecha, czyli wyróżnik, który umożliwia ekscerpcję kolokacji rzeczownikowych z listy n-gramów (w przypadku opisywanej metody są to bigramy i trigramy[71]), wygenerowanych ze zbioru tekstów napisanych w języku angielskim. Cechą wykorzystaną w przedstawianej metodzie jest końcówka liczby mnogiej i dopełniacza saksońskiego (na przykład dla rzeczownika *acid* odpowiednie formy to *acids* oraz *acid's*) występujące na końcu kolokacji (na przykład *acetic acid*). Jeśli n-gramy wygenerowane za pomocą odpowiedniego

[70] Przykładem stosunkowo wczesnych badań korpusowych, w których wyodrębniano z tekstów frazy rzeczownikowe, jest Godby (1994).

[71] Tekst dzielony jest na segmenty złożone z n (w tym wypadku 2 lub 3) słów, przy czym kolejne segmenty się nakładają (na przykład ze zdania *Slow group velocity was measured recently in photonic crystal structures with ultrafast pulse propagation techniques* powstaną następujące trigramy: *Slow group velocity*, *group velocity was*, *velocity was measured*, *was measured recently*, *measured recently in*, *recently in photonic*, *in photonic crystal*, *photonic crystal structures*, *crystal structures with*, *structures with ultrafast*, *with ultrafast pulse*, *ultrafast pulse propagation*, *pulse propagation techniques*).

oprogramowania zostaną uporządkowane w kolejności alfabetycznej, można w łatwy sposób wyodrębnić wszystkie wystąpienia danego n-gramu, po którym następuje jego forma mnoga. W metodzie nie wykorzystuje się więc danych statystycznych, na przykład częstotliwości występowania danego n-gramu w zbiorze tekstów (co nie byłoby powiązane ściśle z kolokacjami rzeczownikowymi), ale formę fleksyjną formy podstawowej (n-gram w liczbie pojedynczej). Jedynymi końcówkami fleksyjnymi występującymi w rzeczownikach w języku angielskim są końcówki liczby mnogiej i dopełniacza saksońskiego (por. Quirk et al. 1985, Loberger, Welsh 2001). Posługując się odpowiednio sformułowanym wyrażeniem regularnym, można wygenerować listę form podstawowych (w liczbie pojedynczej). Listę poddaje się następnie analizie morfologicznej w celu usunięcia wszystkich rekordów w liście, zawierających niepożądane części mowy (na przykład przedimki, przyimki i czasowniki). W ostatnim etapie uzyskaną listę poddaje się analizie manualnej i usuwa wszystkie wpisy, które nie zostaną uznane za kolokacje.

4.4.2 Dane

W badaniach wykorzystano opisany w rozdziale 4, punkt 1 zbiór tekstów opublikowanych w latach 1997–2005 w czasopiśmie naukowym *Nature*. Można bowiem uznać, że teksty naukowe zawierają znaczną liczbę trwałych kolokacji (tzn. takich, których elementy są ze sobą ściśle powiązane), które często można uznać za jednostki wielowyrazowe.

4.4.3 Przetwarzanie danych i ekscerpcja kolokacji

1. Segmentacja na n-gramy:

a. Teksty zostają podzielone na n-gramy przy użyciu odpowiedniego oprogramowania[72]. Nie wprowadza się ograniczenia częstości występowania n-gramów (w przedstawianej metodzie generuje się jedynie bigramy i trigramy), co oznacza, że lista obejmuje także n-gramy pojawiające się w zbiorze jednokrotnie. Znaki interpunkcyjne traktowane są jak słowa. Zatem sekwencja słów *particle*, *protein* nie zostanie uznana za bigram i nie zostanie włączona do listy bigramów, ponieważ oddzielenie słów przecinkiem oznacza, że nie tworzą kolokacji. Operację przeprowadza się bez uwzględnienia wielkości liter, aby nie rejestrować n-gramów różniących się jedynie wielkością liter jako odrębnych jednostek (na przykład *Solar System analogue* oraz *solar system analogue*, *Iron deficiency anaemia* oraz *iron deficiency anaemia*, *Large mode area* oraz *large mode area*, *Electromobility shift assay* oraz *electromobility shift assay*, *Breeding unit area* oraz *breeding unit area*[73]).

b. Lista wynikowa zostaje uporządkowana. Usunięte zostają wszystkie wiersze niemające postaci [a-z]+ [a-z]+ oraz wszystkie znaki interpunkcyjne.

c. Lista zostaje poddana sortowaniu (sortowanie bez uwzględnienia wielkości liter), zaś powtarzające się wiersze są usuwane.

Wynik:

...

abandoning worthy

abandonment in

abandonment is

[72] Do generowania list n-gramów wykorzystano program kfngram. Informacje o programie oraz najnowsza wersja znajdują się na stronie www.kwicfinder.com/kfNgram/

[73] Przykłady z wykorzystywanego zbioru tekstów.

abandonment of

abandons plan

Abangiophyte red

abasic lesion

Abasic local

abasic sites

Abasking shark

abatement cost

abatement costs

abatement credits

abatement exceeds

abatement on

Abathymetric survey

abating and

...

d. Wszystkie znaczniki końca wiersza zamieniane są na przecinek i spację w celu otrzymania ciągłej listy n-gramów.

Wynik:

absolute bearing, *absolute difference*, *absolute threshold*, *absolute truth*, *absolute value*, *absorption band*, *absorption*

coefficient, *absorption feature*, *absorption line*, *absorption loss*, *absorption peak*, *absorption photometer*, *absorption system*, *Abstract Mutation*, *Abstract Oestrogen*, *Abstract Protein...*

2. Ekscerpcja bigramów (trigramów), które stanowią prawdopodobne kolokacje:

a. Wyrażenie regularne wykorzystane w ekscerpcji:

Zamiana: `([a-z]+) ([a-z]+), \1 \2(s|es|'s|s')` na `<$1 $2>` dla bigramów.

Zamiana: `([a-z]+) ([a-z]+) ([a-z]+), \1 \2 \3(s|es|'s|s')` na `<$1 $2 $3>` dla trigramów.

Pierwsze wyrażenie regularne pozwala wyodrębnić wszystkie następujące ciągi bigramów lub trigramów:

AB lattice, *AB lattices*,

abatement cost, *abatement costs*,

ABC gene, *ABC gene's*

W wyniku wykonania tej operacji usunięte zostają formy mnogie i formy dopełniacza saksońskiego, zaś formy pojedyncze zostają ujęte w znaki < >.

b. Lista wynikowa zostaje uporządkowana. Usunięte zostają wszystkie ciągi nieujęte w znaki < >.

3. Analiza morfologiczna:

a. Przetwarzanie. Listę poddaje się analizie morfologicznej przy użyciu analizatora morfologicznego AOT.

b. Porządkowanie listy wynikowej. Usunięte zostają wszystkie niepotrzebne informacje[74]; pozostawia się jedynie oryginalne bigramy (lub trigramy) i indeksy przypisane przez analizator. Lista wynikowa ma następujący format:

w1 (tag1) w2 (tag2):

acetylene (na) molecule (na)

...

achieving (ve) it (na)

achieve (va) resolution (na)

...

achromatic (aa) lens (na)

...

them (pm) look (na)

...

these (ed) depth (na)

4. Ostateczne porządkowanie listy wynikowej:

a. Usunięcie wszystkich wierszy zawierających co najmniej jeden czasownik (indeksy va, vb), przedimek (indeksy xc, xf), zaimek (indeksy, xb, pb), przyimek (indeksy xa, xc), przysłówek (indeks ba).

[74] W przypadku wiersza *acetylene molecule* analizator generuje następujące dane wyjściowe:

0 0 BEG DOC
acetylene 0 9 LLE aa SENT1 +?? ACETYLENE na 26985 0
 9 1 DEL SPC
molecule 10 8 LLE aa CS? SENT2 +?? MOLECULE na 53411 0

b. Usunięcie wszystkich indeksów.

c. Ostateczny format prezentacji wyników jest następujący:

male, *posterior*, *pupal* ***abdomen***.

5. Analiza manualna listy wynikowej n-gramów i usunięcie wszystkich n-gramów, które nie zostaną uznane za kolokacje. Przykłady: *males aggregate* (człon *aggregate* został uznany za rzeczownik; w tym kontekście jest to czasownik), *reach agreement* (człon *reach* został uznany za rzeczownik; w tym kontekście jest to czasownik), *Species aim* (człon *aim* został uznany za rzeczownik; w tym kontekście jest to czasownik), *find alien* (człon *find* został uznany za rzeczownik; w tym kontekście jest to czasownik), *all alternative* (człon *alternative* został uznany za rzeczownik; w tym kontekście jest to przymiotnik), *Suzanne Anker* (imię i nazwisko, nie jest to przedmiotem zainteresowania w niniejszych badaniach).

4.4.4 Wyniki

W poniższej tabeli 8 przedstawiono liczbę kolokacji dwuwyrazowych i trójwyrazowych na liście wynikowej. Następnie podano część pełnej listy kolokacji dwuwyrazowych i trójwyrazowych (jedynie te kolokacje, których ostatni człon zaczyna się na literę A).

Tabela 8. Liczba wyodrębnionych kolokacji (n – liczba elementów w n-gramie).

n	**Suma**
2	79191
3	32853

4.4.4.1 Lista kolokacji dwuwyrazowych

male, *posterior*, *pupal* ***abdomen***,

chromatic, *chromosomal*, *chromosome*, *optical*, *spherical*, *wave* ***aberration***,

bilateral, *centrosome*, *cortical*, *laser*, *unilateral* ***ablation***,

notable, *probable* ***absence***,

saturable, *shock* ***absorber***,

atmospheric, *CD*, *gas*, *infrared*, *iron*, *methane*, *optical*, *photon*, *spectral*, *stretching* ***absorption***,

mathematical ***abstraction***,

Al, *average*, *bacterial*, *Be*, *beryllium*, *bulk*, *carbon*, *cell*, *cellular*, *charcoal*, *chemical*, *CO*, *coronal*, *cosmic*, *diatom*, *element*, *elemental*, *equilibrium*, *faunal*, *final*, *fish*, *fractional*, *global*, *greater*, *higher*, *highest*, *host*, *increased*, *increasing*, *initial*, *iron*, *isotope*, *isotopic*, *Li*, *lithium*, *local*, *lower*, *mantle*, *maximum*, *mean*, *measured*, *metal*, *methane*, *mineral*, *modal*, *molecular*, *natural*, *Ne*, *neon*, *nickel*, *nitrogen*, *nuclear*, *observed*, *olivine*, *oxygen*, *parasitoid*, *population*, *predicted*, *prey*, *primordial*, *Prochlorococcus*, *proportional*, *protein*, *relative*, *rock*, *solar*, *source*, *surface*, *system*, *taxon*, *total*, *transcript*, *viral*, *water*, *weed* ***abundance***,

drug ***abuser***,

myrmecophytic, *thorn* ***Acacia***,

Royal ***Academy***,

angular, *dimensionless*, *gravitational*, *greater*, *inertial*, *linear*, *maximum*, *normalized*, *particle*, *peak*, *radial*, *rapid*, *secular* ***acceleration***,

compact, *conventional*, *electron*, *laser*, *linear*, *particle*, *plasma*, *wakefield* ***accelerator***,

piezoresistive ***accelerometer***,

aromatic, *bond*, *electron*, *fluorescent*, *hydrogen*, *proton*, *quinone*, *splice* ***acceptor***,

lysine ***access***,

Arabidopsis, *human*, *parental*, *sequence* ***accession***,

Challenger, *nuclear*, *terrible*, *traffic* ***accident***,

multilateral ***accord***,

anecdotal, *bank*, *best*, *complete*, *detailed*, *extended*, *eyewitness*, *fascinating*, *historic*, *historical*, *interesting*, *intriguing*, *objective*, *personal*, *popular*, *published*, *readable*, *scientific*, *technical*, *thorough*, *vivid* ***account***,

carbon, *ice* ***accretion***,

abnormal, *carbon*, *cytoplasmic*, *gas*, *ice*, *local*, *melt*, *neurofilament*, *protein*, *random*, *snow* ***accumulation***,

linear ***acene***,

histone ***acetylase***,

lysine ***acetylation***,

histone ***acetyltransferase***,

considerable, *greatest*, *human*, *intellectual*, *major*, *remarkable*, *scientific*, *significant*, *technical*, *technological* ***achievement***,

acetic, *acylamino*, *amino*, *aqueous*, *arachidonic*, *aspartic*, *benzoic*, *bile*, *carbonic*, *carboxylic*, *deoxyribonucleic*, *dicarboxylic*, *dihydroxyeicosatrienoic*, *epoxyeicosatrienoic*, *fatty*, *general*, *glutamic*, *haloacetic*, *humic*, *lactic*, *Lewis*, *linoleic*, *linolenic*, *liquid*, *methacrylic*, *mugineic*, *mycocerosic*, *mycolic*, *nitric*, *nucleic*, *oleanonic*, *organic*, *oxalic*, *phosphatidic*, *phosphinic*, *phosphoamino*, *phosphonic*, *pinonic*, *ribonucleic*, *sialic*, *sulphenic*, *sulphinic*, *sulphonic*, *sulphuric*, *threonic*, *uronic*, *weak* ***acid***,

recombinant ***Acinus***,

mutual ***acquaintance***,

horizontal, *image*, *possible*, *presequence*, *recent* ***acquisition***,

substituted ***acridine***,

spheroidal ***acritarch***,

activity, *Air*, *altruistic*, *balancing*, *cognitive*, *concentration*, *juggling*, *opening* ***act***,

bacterial, *cytoplasmic*, *muscle*, *terminal*, ***actin***,

heavy, *lighter* ***actinide***,

antagonistic, *antiviral*, *anxiolytic*, *apoptotic*, *appropriate*, *behavioural*, *benzodiazepine*, *biological*, *bronchomotor*, *catalytic*, *combined*, *concerted*, *conservation*, *cooperative*, *coordinated*, *corrective*, *cytotoxic*, *depolarizing*, *directed*, *disciplinary*, *drastic*, *drug*, *dual*, *effector*, *enzyme*, *Extracellular*, *familiar*, *Future*, *gluconeogenic*, *government*, *guided*, *hand*, *human*, *hunting*, *hyperpolarizing*, *independent*, *indirect*, *individual*, *inhibitory*, *instructive*, *insulin*, *international*, *legal*, *leptin*, *local*, *manual*, *military*, *modulatory*, *motor*, *natural*, *oestrogenic*, *opposing*, *pathogenic*, *physiological*, *polymorphic*, *possible*, *postsynaptic*, *protective*, *recent*, *regulatory*, *remedial*, *repulsive*, *retaliatory*, *sequential*, *signalling*, *specific*, *steroid*, *strategic*, *synaptic*, *synergistic*, *US*, *Voluntary* ***Action***,

brain, *channel*, *common*, *cortical*, *differential*, *dopamine*, *greater*, *higher*, *induced*, *Maximal*, *neural*, *neuronal*, *parallel*, *peak*, *PET*, *phasic*, *piriform*, *rapid*, *repeated*, *significant*, *sustained*, *thermal* ***activation***,

angiogenesis, *efficient*, *endogenous*, *epigenetic*, *gene*, *PA*, *plasminogen*, *platelet*, *potent*, *proteasome*, *receptor*, *transcription*, *transcriptional* ***activator***,

animal, *environmental*, *peace*, *political* ***activist***,

film, *human*, *transitive* ***actor***,

piezoelectric, *rotational* ***actuator***,

penicillin ***acylase***,

Aleutian ***adakite***,

appropriate, *cellular*, *complex*, *defensive*, *Developmental*, *dietary*, *ecological*, *enzyme*, *evolutionary*, *feeding*, *floral*, *functional*, *genetic*, *important*, *local*,

marine, *masticatory*, *metabolic*, *molecular*, *morphological*, *physiological*, *postcranial*, *sensory*, *Social*, *specific*, *structural*, *unique* ***adaptation***,

domain, *molecular*, *scaffolding* ***adapter***,

cargo, *clathrin*, *generic*, *molecular*, *separate*, *specific*, *transcriptional* ***adaptor***,

carbon, *silicon* ***adatom***,

branch, *chronic*, *dust*, *fertilizer*, *iron*, *latest*, *nitrogen*, *nucleophilic*, *nucleotide*, *nutrient*, *oxidative*, *protein*, *random*, *reagent*, *recent*, *sequential*, *species*, *successive*, *telomere*, *valuable* ***addition***,

food, *polymer* ***additive***,

cisplatin, *covalent*, *major*, *serotonin* ***adduct***,

methylated ***adenine***,

breast, *colorectal*, *ductal*, *invasive*, *lung*, *mammary*, *renal* ***adenocarcinoma***,

alveolar, *colon*, *colorectal*, *intestinal*, *lung*, *Papillary*, *pituitary*, *villous* ***adenoma***,

human, *recombinant* ***adenovirus***,

cell, *focal*, *homophilic*, *membrane* ***adhesion***,

organic, *silicon* ***adhesive***,

brown, *cultured*, *white* ***adipocyte***,

integrity ***adjudication***,

adaptive, *age*, *flux*, *geostrophic*, *isostatic*, *manual*, *minor*, *Structural* ***adjustment***,

alum, *endogenous*, *vaccine* ***adjuvant***,

Atmospheric, *Bush*, *Clinton*, *drug*, *previous*, *repeated*, *Republican*, *US* ***administration***,

research, *science*, *senior*, *system*, *university* ***administrator***,

red ***admiral***,

depressed ***adolescent***,

sculptural ***adornment***,

molecular, *surface* ***adsorbate***,

asexual, *benthic*, *deprived*, *Drosophila*, *female*, *fertile*, *gravid*, *healthy*, *human*, *individual*, *male*, *mature*, *modern*, *mutant*, *normal*, *surviving*, *US*, *viable*, *young* ***adult***,

clinical, *cone*, *considerable*, *continual*, *exciting*, *glacial*, *glacier*, *greatest*, *ice*, *important*, *impressive*, *key*, *latest*, *major*, *maximum*, *mechanical*, *medical*, *notable*, *phase*, *potential*, *rapid*, *recent*, *Science*, *scientific*, *significant*, *subsequent*, *technical*, *technological*, *theoretical*, *therapeutic*, *zone* ***advance***,

adaptive, *additional*, *chief*, *competitive*, *considerable*, *cost*, *demographic*, *distinct*, *economic*, *evolutionary*, *fitness*, *functional*, *growth*, *important*, *inherent*, *key*, *major*, *mechanical*, *obvious*, *particular*, *possible*, *potential*, *primary*, *relative*, *selection*, *selective*, *significant*, *specific*, *survival*, *technical*, *unique* ***advantage***,

scientific ***adventure***,

job ***advertisement***,

chief, *financial*, *former*, *government*, *graduate*, *legal*, *science*, *scientific*, *technical*, *thesis* ***adviser***,

research, *science*, *scientific* ***advisor***,

environmental, *passionate*, *research* ***advocate***,

silica ***aerogel***,

paper ***aeroplane***,

absorbing, *Abstract*, *aliphatic*, *Antarctic*, *anthropogenic*, *Arctic*, *atmospheric*, *carbonaceous*, *coarse*, *detectable*, *liquid*, *marine*, *mixed*, *natural*, *organic*, *Saharan*, *secondary*, *soot*, *stratospheric*, *sulfate*, *sulphate*, *tropical*, *tropospheric*, *volcanic* ***aerosol***,

love, *nuclear* ***affair***,

auditory, *motion*, *visual* ***aftereffect***,

burst, *fading*, *faint*, *optical*, *radio* ***afterglow***,

absolute, *accretion*, *accurate*, *advanced*, *apparent*, *Ar*, *astrochronological*, *astronomical*, *average*, *boundary*, *calendar*, *calibrated*, *characteristic*, *chronological*, *clade*, *cluster*, *comparable*, *concordant*, *conventional*, *cooling*, *cosmic*, *crustal*, *crystallization*, *dark*, *dental*, *depletion*, *depositional*, *derived*, *developmental*, *distinct*, *divergence*, *dynamical*, *earlier*, *earliest*, *eruption*, *estimated*, *evaporation*, *evolutionary*, *expansion*, *exposure*, *formation*, *gas*, *geological*, *gestational*, *glacial*, *groundwater*, *house*, *ice*, *individual*, *intracellular*, *island*, *isochron*, *isotopic*, *known*, *late*, *limiting*, *lineage*, *Mammal*, *maximum*, *mean*, *median*, *metamorphic*, *meteorite*, *middle*, *mineral*, *minimum*, *mixed*, *model*, *Ne*, *Neolithic*, *numerical*, *old*, *older*, *plateau*, *Pleistocene*, *Pliocene*, *postnatal*, *precise*, *radiocarbon*, *radiometric*, *reconstructed*, *relative*, *reservoir*, *resetting*, *resulting*, *retention*, *retirement*, *source*, *specific*, *stellar*, *Stone*, *surface*, *System*, *terrace*, *thermal*, *Thermoluminescence*, *tree*, *uncertain*, *upper*, *varying*, *ventilation*, *working*, *young*, *younger*, *youngest*, *zircon* ***age***,

biomedical, *Energy*, *European*, *federal*, *fisheries*, *pollution*, *Project*, *protection*, *research*, *space*, *US* ***agency***,

commercial, *hidden*, *political*, *research*, *scientific* ***agenda***,

active, *addition*, *aetiological*, *alkylating*, *antiangiogenic*, *antibacterial*, *anticancer*, *antidiabetic*, *antifungal*, *antimalarial*, *antimicrobial*, *antiobesity*, *antiplatelet*, *antithrombotic*, *antitumor*, *antiviral*, *antiworming*, *bacterial*, *biocontrol*, *biological*, *bioterror*, *bioterrorism*, *biowarfare*, *bioweapons*, *blocking*, *causal*, *causative*, *central*, *chelating*, *chemical*, *chemotherapeutic*, *chemotherapy*, *complexing*, *contrast*, *control*, *cooling*, *corrosive*, *coupling*, *crosslinking*, *cytotoxic*, *damaging*, *destructive*, *disease*, *dispersal*, *diversifying*, *environmental*, *eroding*, *erosive*, *external*, *extraterrestrial*, *forcing*, *genotoxic*, *imaging*, *immunosuppressive*, *important*, *infectious*, *infective*, *inotropic*, *known*, *metasomatic*, *methylating*, *microbial*, *mutagenic*, *nerve*, *neuroprotective*, *oxidizing*, *patent*, *pathogenic*, *pharmacological*, *potent*,

primary, *radiosensitizing*, *reducing*, *selective*, *single*, *therapeutic*, *transport*, *triggering*, *vectoring*, *viral* ***agent***,

bubble, *cell*, *cellular*, *clinopyroxene*, *fibrillar*, *follicular*, *fractal*, *gravitational*, *insoluble*, *isotropic*, *larger*, *marine*, *metal*, *oligomeric*, *olivine*, *protein*, *PrPres*, *random*, *soil*, *translation* ***aggregate***,

bacterial, *molecular*, *sperm*, *supramolecular* ***aggregation***,

Cambrian, *extinct* ***agnathan***,

AR, *benzodiazepine*, *cAMP*, *cannabinoid*, *CB*, *channel*, *cholinergic*, *death*, *dopamine*, *ecdysone*, *endogenous*, *inverse*, *muscarinic*, *nicotinic*, *octopamine*, *partial*, *peptide*, *platelet*, *potent*, *receptor*, *selective*, *self*, *synthetic* ***agonist***,

bilateral, *collaboration*, *collaborative*, *cooperative*, *excellent*, *global*, *good*, *implicit*, *international*, *legal*, *licence*, *licensing*, *management*, *model*, *multilateral*, *negotiated*, *original*, *proliferation*, *quantitative*, *research*, *service*, *Trade*, *transfer* ***agreement***,

African, *effective*, *entire*, *human*, *international*, *US*, *World* ***aid***,

congressional ***aide***,

common ***ailment***,

central, *common*, *key*, *stated*, *ultimate*, ***aim***,

seismic ***airgun***,

commercial ***airliner***,

Arctic ***airmass***,

International ***Airport***,

allergic, *constricted*, *human*, *terminal* ***airway***,

Los ***Alamos***,

ultrasonic ***alarm***,

wandering ***albatross***,

geometric, *mean*, *surface* ***albedo***,

amino, *aryl*, *blood*, *chiral*, *secondary*, *sugar* ***alcohol***,

peptide ***aldehyde***,

Damage, *email*, *global*, *tsunami* ***alert***,

alignment, *analysed*, *BLAST*, *classical*, *clustering*, *complex*, *computational*, *computer*, *decentralized*, *design*, *detection*, *efficient*, *encryption*, *genetic*, *heuristic*, *learning*, *mathematical*, *optimization*, *prediction*, *quantum*, *reconstruction*, *robust*, *routing*, *satisfiability*, *scaling*, *scoring*, *search*, *simple*, *sophisticated*, *tree* ***algorithm***,

acid, *chance*, *chromosome*, *complete*, *duplication*, *erroneous*, *Figure*, *genome*, *genomic*, *induced*, *larger*, *local*, *mammalian*, *molecular*, *mouse*, *multiple*, *Observed*, *perpendicular*, *protein*, *pyramid*, *relative*, *sequence*, *side*, *spatial*, *spin*, *spindle*, *structural*, *structure*, *vector*, *vertical* ***alignment***,

Cinchona, *plant*, *pyrrolizidine*, *steroidal*, *trace* ***alkaloid***,

functionalized, *gaseous*, *linear* ***alkane***,

silicon ***alkoxide***,

stable ***alkylidene***,

terminal ***alkyne***,

misconduct, *specific* ***allegation***,

ADAR, *additional*, *advantageous*, *ancestral*, *AR*, *associated*, *beneficial*, *benthic*, *causal*, *CD*, *chromosomal*, *class*, *common*, *conditional*, *control*, *cut*, *deleterious*, *deletion*, *derived*, *disease*, *disrupted*, *dominant*, *duplication*, *endogenous*, *epigenetic*, *expressed*, *floxed*, *functional*, *gene*, *germline*, *host*, *human*, *hybrid*, *hypomorphic*, *I*, *Id*, *IgH*, *inherited*, *insertion*, *J*, *knockout*, *lethal*, *limnetic*, *line*, *linked*, *lox*, *maize*, *major*, *marker*, *MAT*, *maternal*, *minor*, *missense*, *modified*, *mosaic*, *mutant*, *mutated*, *negative*, *neutral*, *normal*, *Notch*, *null*, *parental*, *particular*, *paternal*, *polymorphic*, *predisposing*, *rearranged*, *recessive*, *recombinant*, *relevant*, *resistance*, *rhodopsin*, *risk*, *SER*, *Shaker*, *SHIP*, *single*, *susceptibility*, *susceptible*, *targeted*, *taster*, *unique*, *variant*, *weak*, *weaker* ***allele***,

pollen ***allergen***,

blind, *swim* ***alley***,

Global, *unholy* ***Alliance***,

platelet ***alloantigen***,

budget, *emissions*, *funding*, *national*, *solitary* ***allocation***,

cardiac ***allograft***,

error, *housing* ***allowance***,

Al, *aluminium*, *aluminum*, *amorphous*, *binary*, *bulk*, *chromium*, *cubic*, *decagonal*, *granular*, *intermetallic*, *Invar*, *iron*, *metal*, *metallic*, *nickel*, *ordered*, *semiconductor*, *silicon*, *titanium*, ***alloy***,

polymorphic ***allozyme***,

acid, *badge*, *chromatin*, *chromosome*, *conformational*, *consistent*, *diagenetic*, *epigenetic*, *Field*, *genetic*, *genomic*, *habitat*, *Human*, *irreversible*, *major*, *marked*, *metabolic*, *Minor*, *morphological*, *nucleotide*, *Phenotypic*, *receptor*, *Reversible*, *sequence*, *significant*, *specific*, *structural* ***alteration***,

bond, *periodic*, *rivalry*, *stimulus* ***alternation***,

accessible, *attractive*, *available*, *better*, *conventional*, *developing*, *plausible*, *possible*, *potential*, *practical*, *safer*, *suitable*, *test*, *viable*, ***alternative***,

laser, *radar*, *satellite*, *sonar* ***altimeter***,

average, *forest*, *higher*, *highest*, *ionopause*, *lower*, *site* ***altitude***,

undiscriminating ***altruist***,

nickel ***aluminide***,

anionic ***aluminophosphate***,

crystalline ***aluminosilicate***,

marine ***alveolate***,

sinusoidal, *standard* ***AM***,

proposed ***amendment***,

African, *Asian*, *black*, *European*, *indigenous*, *native*, *Pacific*, *young* ***American***,

modern ***Amerindian***,

acid, *backbone*, *terminal*, *Tertiary*, *xylem* ***amide***,

active, *biogenic*, *primary*, *quaternary* ***amine***,

stem ***amniote***,

absolute, *adequate*, *appreciable*, *appropriate*, *average*, *catalytic*, *cloud*, *column*, *comparable*, *considerable*, *constant*, *corresponding*, *decreased*, *decreasing*, *detectable*, *disproportionate*, *enormous*, *equal*, *equimolar*, *equivalent*, *exact*, *excess*, *excessive*, *fixed*, *greater*, *higher*, *highest*, *huge*, *identical*, *immense*, *increased*, *increasing*, *initial*, *known*, *larger*, *largest*, *lesser*, *limited*, *limiting*, *lower*, *lowest*, *macroscopic*, *massive*, *maximal*, *maximum*, *measurable*, *measured*, *microscopic*, *minimal*, *minor*, *minute*, *mismatch*, *moderate*, *modest*, *molar*, *negligible*, *normal*, *observed*, *path*, *precipitation*, *predicted*, *present*, *prodigious*, *protein*, *rainfall*, *reasonable*, *reduced*, *relative*, *required*, *residual*, *saturating*, *significant*, *sleep*, *smaller*, *stoichiometric*, *substantial*, *substoichiometric*, *sufficient*, *surprising*, *tiny*, *total*, *trace*, *tremendous*, *uniform*, *variable*, *varying*, *vast*, ***amount***,

anuran, *living*, *many*, *modern*, ***amphibian***,

control, *gene*, *genomic*, *independent*, *Multiplex*, *Serial*, *signal*, *specific*, ***amplification***,

clamp, *fibre*, *linear*, *mechanical*, *operational*, *optical*, *parametric*, *power*, *selection*, *voltage* ***amplifier***,

absolute, *annual*, *average*, *background*, *burst*, *calculated*, *channel*, *coherence*, *complex*, *continuum*, *control*, *correlation*, *current*, *displacement*, *driving*, *emission*, *equal*, *ERG*, *excitation*, *expected*, *experimental*, *field*, *focal*, *force*, *Fourier*, *freshwater*, *image*, *increased*, *increasing*, *initial*, *larger*, *lever*, *lower*, *luminosity*, *maximal*, *maximum*, *mean*, *measured*, *mini*, *mixing*, *modulation*, *noise*, *normalized*, *obliquity*, *observed*, *oscillation*, *peak*, *polarization*,

poststimulation, *potential*, *pressure*, *prestimulation*, *probability*, *pulse*, *quantal*, *reaction*, *reduced*, *reflection*, *relative*, *response*, *saccade*, *scattering*, *seasonal*, *signal*, *significant*, *smaller*, *spark*, *spike*, *stack*, *stimulus*, *strain*, *stroke*, *transition*, *tunnelling*, *typical*, *unitary*, *unwinding*, *using*, *variable*, *varying*, *velocity*, *vibration*, *visibility*, *wave*, *zeitgeber* ***amplitude***,

glass, *silica* ***ampoule***,

limb ***amputation***,

strict ***anaerobe***,

general, *local* ***anaesthetic***,

potent ***analgesic***,

electrostatic, *phosphate*, *synthetic* ***analog***,

acid, *ancient*, *artificial*, *benzoate*, *better*, *cAMP*, *cap*, *dust*, *exact*, *functional*, *hexadeoxynucleotide*, *I*, *ice*, *modern*, *molecular*, *natural*, *nucleoside*, *nucleotide*, *peptide*, *perovskite*, *phosphate*, *possible*, *potent*, *proline*, *quantum*, *stable*, *structural*, *substrate*, *sugar*, *synthetic*, *terrestrial*, *ubiquinone*, *zeolite* ***analogue***,

gas, *genetic*, *mass*, *mobility*, *polarization* ***analyser***,

biotech, *defence*, *financial*, *industry*, *nuclear*, *OMB*, *policy*, *research*, *stock*, *structural* ***analyst***,

liquid ***analyte***,

Genetic ***Analyzer***,

enriched ***anammoxosome***,

ape, *burrowing*, *common*, *crocodilian*, *cyanobacterial*, *extinct*, *genealogical*, *hominid*, *human*, *immediate*, *lichenized*, *mammalian*, *matrilineal*, *plant*, *possible*, *potential*, *predatory*, *proximate*, *recent*, *universal*, *vertebrate* ***ancestor***,

extracellular, *glycosylphosphatidylinositol*, *internal*, *protein* ***anchor***,

Magnesian, *Orogenic* ***andesite***,

fetal, *prenatal* ***androgen***,

personal, *telling* ***anecdote***,

sea ***anemone***,

aortic ***aneurysm***,

basal, *primitive* ***angiosperm***,

acute, *aerodynamic*, *aspect*, *average*, *axial*, *azimuth*, *azimuthal*, *backbone*, *beaming*, *bending*, *bifurcation*, *body*, *bond*, *chiral*, *chute*, *collision*, *cone*, *corresponding*, *critical*, *crossing*, *detection*, *diffraction*, *dihedral*, *disparity*, *divergence*, *elbow*, *emission*, *Euler*, *external*, *extinction*, *facial*, *fixed*, *flip*, *fracture*, *fresh*, *glide*, *grazing*, *Hall*, *higher*, *hinge*, *horizontal*, *impact*, *incidence*, *incident*, *inclination*, *interfacial*, *internal*, *joint*, *larger*, *lower*, *mismatch*, *nadir*, *normal*, *observation*, *obtuse*, *opening*, *orientation*, *orthogonal*, *packing*, *parallactic*, *phase*, *polar*, *polarization*, *position*, *projection*, *rake*, *random*, *ray*, *roll*, *rotation*, *scatter*, *scattering*, *shallow*, *smaller*, *smallest*, *speaker*, *specified*, *steep*, *steeper*, *tilt*, *tip*, *torsion*, *torsional*, *track*, *turn*, *turning*, *twist*, *unexpected*, *vertical*, *viewing*, *wave*, *zenith* ***angle***,

bulk ***angrite***,

carbonic ***anhydrase***,

adult, *affected*, *agricultural*, *alert*, *anaesthetized*, *ancestral*, *appropriate*, *aquatic*, *awake*, *behaving*, *book*, *certain*, *chimaeric*, *cloned*, *common*, *competing*, *complex*, *conscious*, *control*, *day*, *Defending*, *deprived*, *diploblastic*, *distinct*, *domestic*, *donor*, *Dorset*, *earlier*, *earliest*, *eight*, *endangered*, *engineered*, *exotic*, *experimental*, *extant*, *extinct*, *famous*, *farm*, *feeding*, *game*, *heterozygous*, *hibernating*, *higher*, *host*, *hovering*, *immunized*, *individual*, *infected*, *intact*, *knockout*, *lab*, *laboratory*, *land*, *larger*, *living*, *lower*, *marine*, *market*, *mature*, *microscopic*, *model*, *modern*, *modified*, *monitor*, *multicellular*, *mutant*, *naive*, *normal*, *old*, *older*, *paired*, *particular*, *prey*, *remaining*, *representative*, *resulting*, *sacred*, *simple*, *single*, *soil*, *stage*, *standardized*, *stimulus*, *strabismic*, *studied*, *study*, *susceptible*, *swimming*,

symmetrical, *terrestrial*, *test*, *total*, *trained*, *transgenic*, *treated*, *unimmunized*, *unstimulated*, *untrained*, *vaccinated*, *walking*, *whole*, *X*, *young* ***animal***,

Information ***Animation***,

acid, *anthracene*, *borate*, *carboxylate*, *charged*, *extracellular*, *gluconate*, *heteropolytungstate*, *hydrophobic*, *intracellular*, *magnetic*, *organic*, *superoxide* ***anion***,

crossed ***ankle***,

antihydrogen, *antiproton*, *neutralino* ***annihilation***,

additional, *chromosome*, *curated*, *current*, *functional*, *gene*, *genome*, *genomic*, *GO*, *preliminary*, *sequence* ***annotation***,

public, *recent* ***announcement***,

alternative, *composite*, *membrane*, *polymeric* ***anode***,

Laramie, *Proterozoic* ***Anorthosite***,

benzoquinone ***ansamycin***,

basal ***anseriform***,

automatic, *better*, *convincing*, *correct*, *definitive*, *final*, *general*, *one*, *partial*, *plausible*, *positive*, *possible*, *precise*, *proposed*, *quantitative*, *scientific*, *simple*, *ultimate* ***answer***,

acacia, *arboreal*, *Argentine*, *desert*, *extant*, *fire*, *Foraging*, *formicine*, *harvester*, *leafcutter*, *Myrmica*, *plague*, *ponerine*, *trained*, *wood* ***ant***,

AR, *behavioural*, *calmodulin*, *cannabinoid*, *CB*, *chemokine*, *competitive*, *functional*, *inducible*, *neutral*, *nicotinic*, *NMDAR*, *nodal*, *pathway*, *potent*, *protein*, *receptor*, *secreted*, *selective*, *site*, *specific* ***antagonist***,

giant ***anteater***,

characteristic, ***antecedent***,

saiga ***antelope***,

integrated, *wave* ***antenna***,

American, *forensic*, *physical* ***anthropologist***,

naive ***anthropomorphism***,

trapped ***antiatom***,

aminoglycoside, *ansamycin*, *antitumor*, *appropriate*, *existing*, *glycopeptide*, *important*, *lactone*, *macrolide*, *natural*, *particular*, *peptide*, *polyketide*, *potent*, *specific* ***antibiotic***,

natural, *potent* ***anticoagulant***,

tricyclic ***antidepressant***,

doped, *frustrated* ***antiferromagnet***,

conventional ***antifolate***,

acquired, *capture*, *CD*, *cell*, *cellular*, *common*, *differentiation*, *Ebola*, *encoded*, *endogenous*, *exogenous*, *foreign*, *Gal*, *glycolipid*, *glycoprotein*, *group*, *hemidesmosomal*, *histocompatibility*, *I*, *intracellular*, *islet*, *leukocyte*, *lipid*, *malarial*, *melanoma*, *membrane*, *microbial*, *model*, *multivalent*, *mycobacterial*, *myelin*, *neuronal*, *nuclear*, *particular*, *peptide*, *phospholipid*, *protective*, *protein*, *recombinant*, *residual*, *self*, *soluble*, *specific*, *Sphingomonas*, *surface*, *target*, *Targeting*, *tumor*, *tumour*, *ubiquitous*, *vaccine*, *variant*, *viral*, *virus* ***antigen***,

electron ***antineutrino***,

cellular, *plasma* ***antioxidant***,

corresponding ***antiparticle***,

hexagonal ***antiprism***,

diamond ***anvil***,

abdominal, *dorsal*, *systemic* ***aorta***,

African, *captive*, *extant*, *fossil*, *modern* ***ape***,

central ***Apennine***,

circular, *elliptical*, *numerical*, *rectangular*, *reference*, *stomatal*, *subwavelength*, *telescope* ***aperture***,

mummified, *parasitized*, *pea*, *wheat* ***aphid***,

phytochrome ***apoprotein***,

southern ***Appalachian***,

federal, *legal* ***appeal***,

histological, *initial*, *similar* ***appearance***,

antenniform, *Crustacean*, *dorsal*, *Drosophila*, *frontal*, *head*, *integumental*, *larval*, *pectoral*, *posterior*, *protocerebral*, *skin*, *surface*, *ventral* ***appendage***,

voracious ***appetite***,

conventional, *Delicious*, *golden*, *integrated*, *Seedling* ***apple***,

Java ***applet***,

electronic ***appliance***,

grant, *unsuccessful*, *visa* ***applicant***,

agonist, *AM*, *attractive*, *biological*, *biomedical*, *biotech*, *biotechnological*, *business*, *clinical*, *commercial*, *computer*, *control*, *current*, *database*, *drug*, *ecological*, *emerging*, *engineering*, *fertilizer*, *find*, *fungicide*, *future*, *general*, *geophysical*, *glutamate*, *grant*, *herbicide*, *immediate*, *important*, *individual*, *industrial*, *initial*, *insecticide*, *interesting*, *job*, *licence*, *local*, *marketable*, *medical*, *military*, *ms*, *MTSET*, *online*, *particular*, *patent*, *pesticide*, *possible*, *potential*, *practical*, *pressure*, *previous*, *profitable*, *promising*, *recent*, *related*, *repeated*, *routine*, *screening*, *sequencing*, *significant*, *simple*, *software*, *space*, *specific*, *subsequent*, *successful*, *sucrose*, *suggested*, *synthetic*, *technological*, *therapeutic*, *topical*, *unacceptable*, *unexpected*, *useful*, *visa*, *vision*, *wider*, *widespread* ***application***,

sample ***applicator***,

political ***appointee***,

academic, *dual*, *faculty*, *joint*, *postdoctoral*, *research* ***appointment***,

abstract, *accurate*, *alternative*, *analytic*, *analytical*, *annotation*, *antisense*, *area*, *array*, *basic*, *behavioural*, *best*, *biased*, *biochemical*, *bioinformatics*, *biological*, *biomimetic*, *chemical*, *classical*, *clinical*, *cloning*, *closest*, *collaborative*, *combination*, *combinatorial*, *combined*, *common*, *comparative*, *competitive*, *complementary*, *complementation*, *comprehensive*, *computational*, *conceptual*, *Conservation*, *conservative*, *conventional*, *creative*, *current*, *developing*, *domain*, *ecological*, *effective*, *efficient*, *element*, *empirical*, *engineering*, *European*, *evolutionary*, *existing*, *experimental*, *feasible*, *flexible*, *focused*, *force*, *fresh*, *functional*, *gene*, *general*, *generic*, *genetic*, *genomic*, *genomics*, *global*, *hybrid*, *I*, *imaging*, *immunological*, *immunotherapeutic*, *independent*, *indirect*, *individualized*, *informative*, *informed*, *innovative*, *integrated*, *integrative*, *interdisciplinary*, *international*, *knockout*, *latter*, *library*, *mapping*, *mathematical*, *mechanistic*, *model*, *modeling*, *modelling*, *modern*, *molecular*, *morphological*, *multidisciplinary*, *mutational*, *national*, *novel*, *numerical*, *obvious*, *optical*, *organic*, *Palaeontological*, *parallel*, *parsimony*, *pharmacological*, *phenomenological*, *philosophical*, *phylogenetic*, *planning*, *possible*, *potential*, *practical*, *precautionary*, *previous*, *productive*, *promising*, *proposed*, *proteomic*, *proteomics*, *quantitative*, *rational*, *recent*, *reductionist*, *regression*, *related*, *reported*, *research*, *rigorous*, *rival*, *scanning*, *scientific*, *screening*, *sequencing*, *simpler*, *sophisticated*, *standard*, *statistical*, *stochastic*, *structural*, *successful*, *system*, *systematic*, *systems*, *targeted*, *theoretical*, *therapeutic*, *therapy*, *traditional*, *transfer*, *transgenic*, *unbiased*, *usual*, *vaccine*, *viable*, *wavelet* ***approach***,

Senate ***appropriation***,

drug, *ethical*, *regulatory*, *such* ***approval***,

crystalline ***approximant***,

Born, *continuum*, *crude*, *density*, *good*, *gradient*, *linear*, *local*, *model*, *numerical*, *pair*, *reasonable*, *useful* ***approximation***,

debris ***apron***,

anticoagulant, *synthetic* ***aptamer***,

mammalian, *membrane*, *plant* ***aquaporin***,

Seattle ***Aquarium***,

basalt, *coastal*, *contaminated*, *Ganges*, *sedimentary*, *shallow*, *Trias* ***aquifer***,

Israeli ***Arab***,

axon, *Dendritic*, *retinotectal* ***arbor***,

axonal, *dendritic*, *terminal* ***arborization***,

Bergen, *canal*, *Cascade*, *circular*, *continental*, *dark*, *echo*, *Fermi*, *individual*, *island*, *Kamchatka*, *magmatic*, *partial*, *ring*, *unit*, *volcanic* ***arc***,

aortic, *branchial*, *caudal*, *gill*, *neural*, *palatal*, *pharyngeal* ***arch***,

US ***archaeologist***,

Canadian, *Pacific* ***archipelago***,

principal ***architect***,

brain, *chromosome*, *common*, *Comparative*, *complex*, *composite*, *computer*, *coupling*, *device*, *domain*, *energetic*, *genetic*, *learning*, *molecular*, *network*, *parallel*, *plant*, *promoter*, *protein*, *supramolecular*, *Surface* ***architecture***,

Alamos, *Cape*, *central*, *climate*, *data*, *electronic*, *FTP*, *institutional*, *Internet*, *Mutant*, *open*, *palaeoclimate*, *permanent*, *public*, *scientific*, *VISUAL* ***archive***,

crocodylomorph ***archosaur***,

accumulation, *active*, *adapted*, *adjacent*, *affected*, *aftershock*, *agricultural*, *ancestral*, *application*, *arm*, *association*, *auditory*, *autopodal*, *average*, *base*, *belt*, *black*, *blue*, *box*, *boxed*, *brain*, *breeding*, *broader*, *burned*, *burnt*, *catchment*, *cell*, *central*, *certain*, *circular*, *civilian*, *coastal*, *collecting*, *coloured*, *common*, *competitive*, *conservation*, *contentious*, *continental*, *control*, *controversial*, *core*, *corresponding*, *cortical*, *critical*, *culled*, *damaged*, *dark*, *darker*, *debated*, *defined*, *density*, *difficult*, *disaster*, *disease*, *dispersal*, *display*, *disrupted*, *distal*, *distribution*, *distributional*, *drainage*, *dry*, *emerging*, *enclosed*, *endemic*, *enlarged*, *epicentral*, *excavated*, *exciting*, *existing*, *experimental*, *exposed*, *extended*, *extensive*, *extrastriate*, *facial*, *fascinating*,

feeding, *fertile*, *field*, *fished*, *focus*, *foraging*, *forest*, *forested*, *fossiliferous*, *gape*, *general*, *genomic*, *geographic*, *geographical*, *geothermal*, *greater*, *green*, *grey*, *growing*, *growth*, *habitat*, *hatched*, *higher*, *highest*, *hippocampal*, *holoendemic*, *hotspot*, *huge*, *humid*, *hydrophobic*, *hypometabolic*, *hypothalamic*, *illuminated*, *image*, *important*, *incisor*, *indicated*, *infected*, *input*, *interdisciplinary*, *interesting*, *interfacial*, *intraparietal*, *introduced*, *irrigated*, *island*, *islet*, *isolated*, *junction*, *key*, *kitchen*, *lake*, *land*, *language*, *larger*, *largest*, *leaf*, *lens*, *lesion*, *lighter*, *limited*, *lipid*, *LIS*, *local*, *localized*, *lowland*, *lumen*, *macaque*, *major*, *mangrove*, *marginal*, *Mean*, *Mediterranean*, *medullary*, *membrane*, *metropolitan*, *mode*, *motion*, *motor*, *naked*, *narrow*, *natural*, *necrotic*, *neglected*, *neighbouring*, *nuclear*, *nursery*, *ocean*, *oceanic*, *olfactory*, *orange*, *outbreak*, *outcrop*, *overlap*, *parietal*, *particular*, *patch*, *peak*, *perivascular*, *polar*, *polluted*, *populated*, *pore*, *positive*, *potential*, *predicted*, *prefrontal*, *premotor*, *priority*, *problem*, *processing*, *projected*, *projection*, *promising*, *protected*, *range*, *recharge*, *rectangular*, *red*, *reef*, *reference*, *refugial*, *relative*, *remaining*, *remote*, *removal*, *representation*, *research*, *residential*, *restricted*, *ROC*, *rupture*, *rural*, *sample*, *sampling*, *scientific*, *Sea*, *secondary*, *see*, *selected*, *sensitive*, *separate*, *shaded*, *shadowed*, *sheet*, *shelf*, *show*, *showing*, *significant*, *skin*, *smaller*, *source*, *spatial*, *spawning*, *specialist*, *specialized*, *specific*, *spectral*, *spiral*, *stable*, *stained*, *stigmatic*, *stippled*, *study*, *subject*, *suitable*, *sunspot*, *surface*, *surrounding*, *survey*, *target*, *targeted*, *temporal*, *therapeutic*, *traditional*, *treated*, *trial*, *tumour*, *unadapted*, *unfamiliar*, *unit*, *unselected*, *upland*, *upwelling*, *urban*, *vast*, *vegetated*, *visual*, *vital*, *volcanic*, *watershed*, *western*, *wheat*, *white*, *whole*, *wider*, *wintering*, *wooded*, *work*, *yellow*, *young* ***area***,

international, *mating*, *open*, *scientific* ***arena***,

conserved, *equivalent*, *histone*, *methylated*, *unmethylated* ***arginine***,

abstract, *analogous*, *basic*, *central*, *common*, *compelling*, *convincing*, *dimensional*, *ethical*, *general*, *geometrical*, *good*, *heuristic*, *key*, *logical*, *persuasive*, *physical*, *previous*, *rational*, *revisionist*, *rigorous*, *scaling*, *scientific*, *simple*, *standard*, *statistical*, *supporting*, *theoretical* ***argument***,

activated, *antibody*, *bottom*, *chromosome*, *CO*, *dynein*, *effector*, *flexible*, *gonad*, *gonadal*, *horizontal*, *inner*, *lever*, *linker*, *open*, *outer*, *radial*, *reference*, *research*, *robot*, *robotic*, *spiral*, *upper*, *US*, *vertical* ***arm***,

Israeli, *US* ***army***,

alternative, *asymmetric*, *atomic*, *comparable*, *complex*, *contractual*, *Crystallographic*, *current*, *cyclic*, *domain*, *dot*, *experimental*, *funding*, *gene*, *geometric*, *helical*, *hexagonal*, *lattice*, *linear*, *local*, *looped*, *molecular*, *novel*, *nuclear*, *ordered*, *packing*, *planar*, *polypeptide*, *possible*, *random*, *spatial*, *specific*, *spin*, *splitting*, *structural*, *subunit*, *symmetric*, *tetrahedral* ***arrangement***,

Affymetrix, *all*, *antibody*, *aperture*, *Atlas*, *bipolar*, *branched*, *Broadband*, *channel*, *circular*, *colloidal*, *complex*, *condensed*, *corresponding*, *crystalline*, *data*, *dense*, *detector*, *diode*, *diverse*, *domain*, *dopant*, *dot*, *electrode*, *electroreceptor*, *exon*, *experimental*, *expression*, *extensive*, *extrachromosomal*, *filter*, *formidable*, *frequency*, *gate*, *gene*, *genome*, *glass*, *helical*, *hexagonal*, *hole*, *huge*, *hydrophone*, *individual*, *integrated*, *isotope*, *junction*, *larger*, *laser*, *LED*, *lens*, *linear*, *memory*, *microchip*, *microelectrode*, *microlens*, *microtubule*, *millimetre*, *mixing*, *nanoparticle*, *nanotube*, *nanowire*, *nipple*, *nonlinear*, *nucleosomal*, *nucleosome*, *oligonucleotide*, *optical*, *ordered*, *packed*, *parallel*, *patterned*, *peptide*, *periodic*, *phase*, *phased*, *phenotype*, *phenotyping*, *photoreceptor*, *planar*, *polarized*, *polycrystalline*, *polymer*, *probe*, *protein*, *ProteinChip*, *proteome*, *radial*, *radio*, *random*, *recording*, *regular*, *repeat*, *sample*, *search*, *seismic*, *sensor*, *solar*, *staggered*, *stimulus*, *submillimetre*, *suffix*, *supramolecular*, *surface*, *symmetrical*, *telescope*, *test*, *tiling*, *transponder*, *transposase*, *trap*, *triangular*, *vast*, *vortex*, *waveguide*, *wire* ***array***,

robotic ***arrayer***,

cycle, *fork*, *growth*, *metaphase*, *mitotic*, *transient* ***arrest***,

atrial, *cardiac*, *cause*, *heart*, *ventricular* ***arrhythmia***,

expected, *human*, *observed*, *photon*, *predicted*, *refracted*, *shear* ***arrival***,

bent, *blue*, *bold*, *bottom*, *broken*, *brown*, *curved*, *dark*, *dotted*, *green*, *grey*, *horizontal*, *induction*, *larger*, *left*, *lower*, *numbered*, *orange*, *purple*, *red*, *see*, *single*, *unlabelled*, *vertical*, *yellow* ***arrow***,

black, *blue*, *filled*, *open*, *red*, *Single*, *white*, *yellow* ***arrowhead***,

nuclear ***arsenal***,

beryllium ***arsenate***,

analytical, *cloning*, *cultural*, *experimental*, *historical*, *instrumental*, *laboratory*, *motion*, *movement*, *natural*, *oxidation*, *processing*, *sampling*, *sequencing*, *significant*, *stone* ***artefact***,

mesenteric ***arteriole***,

accompanying, *Insight*, *interesting*, *journal*, *magazine*, *manufactured*, *nature*, *Naturejobs*, *news*, *newspaper*, *opinion*, *original*, *published*, *recent*, *research*, *review*, *scientific*, *see*, *Views*, *Wikipedia* ***article***,

ginglymoid ***articulation***,

stone ***artifact***,

known, ***artiodactyl***,

British, *contemporary*, *Dutch*, *living*, *young* ***artist***,

didemnid, *extant*, *solitary* ***ascidian***,

filamentous ***ascomycete***,

vitric, *volcanic* ***ash***,

conserved, *transmembrane*, ***aspartate***,

additional, *anterior*, *appealing*, *attractive*, *Basal*, *beneficial*, *challenging*, *commercial*, *common*, *complementary*, *controversial*, *critical*, *crucial*, *distinctive*, *economic*, *essential*, *evolutionary*, *exciting*, *functional*, *fundamental*, *human*, *important*, *interesting*, *intriguing*, *key*, *lateral*, *lingual*, *major*, *molecular*, *neglected*, *novel*, *particular*, *posterior*, *Practical*, *puzzling*, *remarkable*, *security*, *sensory*, *significant*, *specific*, *striking*, *structural*, *temporal*, *unique*, *unusual*, *ventral*, *wave* ***aspect***,

individual, *petroleum* ***asphaltene***,

tracheal ***aspirate***,

stealth ***assault***,

accessibility, *aconitase*, *activity*, *adhesion*, *aerotaxis*, *agar*, *aminoacylation*, *angiogenesis*, *apoptosis*, *apoptotic*, *arginylation*, *ATPase*, *attachment*, *Bandshift*, *behavioural*, *binding*, *biochemical*, *biofilm*, *biological*, *blot*, *Bradford*, *BRET*, *CAM*, *cAMP*, *cap*, *CAT*, *cellular*, *chemotaxis*, *chip*, *clonogenic*, *clotting*, *collapse*, *colony*, *colorimetric*, *competition*, *competitive*, *complementation*, *conjugation*, *conversion*, *crystallography*, *cyclase*, *cytokine*, *cytotoxicity*, *deacetylase*, *deacetylation*, *deaminase*, *deamination*, *degradation*, *diffusion*, *dilution*, *downregulation*, *drug*, *Electrochemiluminescence*, *electrophysiological*, *ELISA*, *Elispot*, *enzymatic*, *enzyme*, *ERK*, *exchange*, *experimental*, *expression*, *extension*, *feeding*, *formation*, *functional*, *fusion*, *gene*, *gliding*, *glutathione*, *glycosylase*, *growth*, *GTPase*, *HAT*, *helicase*, *HMTase*, *immune*, *immunoabsorbent*, *immunoprecipitation*, *immunosorbent*, *import*, *incorporation*, *infectivity*, *ingestion*, *inhibition*, *injection*, *invasion*, *killing*, *kinase*, *ligase*, *liquid*, *loading*, *Lowry*, *luciferase*, *mating*, *maze*, *metastasis*, *methyltransferase*, *microinjection*, *migration*, *Molecular*, *motility*, *multiplex*, *mutation*, *Neurosphere*, *neutralization*, *nuclease*, *outgrowth*, *overlay*, *perifusion*, *phagocytosis*, *phenotypic*, *phosphatase*, *phosphorylation*, *plaque*, *plate*, *polymerization*, *precipitation*, *preference*, *Proliferation*, *protection*, *protein*, *proximity*, *pulldown*, *purification*, *quantitative*, *recognitionphagocytosis*, *reconstitution*, *Relaxation*, *repair*, *repopulation*, *reporter*, *reproducible*, *retardation*, *retrotransposition*, *screening*, *secretion*, *sedimentation*, *shift*, *specific*, *splicing*, *sprouting*, *standard*, *supercoiling*, *supershift*, *Survival*, *swarm*, *swelling*, *switch*, *synthase*, *synthesis*, *Taqman*, *telomerase*, *Thiosulphate*, *toeprinting*, *transactivation*, *transcription*, *transfection*, *transfilter*, *transformation*, *transient*, *transwell*, *tumorigenicity*, *Tunel*, *turning*, *typical*, *ubiquitination*, *Uptake*, *various* ***assay***,

alteration, *bacterial*, *bacterioplankton*, *death*, *diatom*, *diverse*, *faunal*, *foraminiferal*, *fossil*, *insect*, *intricate*, *lithic*, *local*, *microbial*, *microfossil*,

mineral, *mixed*, *natural*, *palynological*, *pathogen*, *phytoplankton*, *pollen*, *tool*, *vertebrate* ***assemblage***,

previous, *recent* ***assertion***,

accurate, *analytical*, *environmental*, *global*, *harsh*, *hazard*, *independent*, *initial*, *multidisciplinary*, *national*, *performance*, *Phylogenetic*, *previous*, *quantitative*, *recent*, *regional*, *reliable*, *research*, *risk*, *scientific*, *Stock*, *visual* ***assessment***,

important, *valuable* ***asset***,

backbone, *chromosomal*, *correct*, *functional*, *gene*, *GO*, *incorrect*, *peptide*, *protein*, *resonance*, *satisfying*, *sequence*, *sequential*, *shift*, *spectral*, *Stereospecific*, *structure*, *tentative*, *truth*, *unambiguous* ***assignment***,

former, *research*, *robotic*, *teaching*, *technical* ***assistant***,

Biotechnology, *DAP*, *paired*, *research* ***associate***,

Agricultural, *allelic*, *archaeological*, *biotechnology*, *British*, *causal*, *centromere*, *chance*, *Chiropractic*, *complex*, *cytoskeletal*, *disease*, *dynamic*, *Figure*, *functional*, *gene*, *genetic*, *graphite*, *incorrect*, *independent*, *indirect*, *interchromosomal*, *intermolecular*, *interstitial*, *intimate*, *involving*, *lateral*, *learned*, *load*, *Medical*, *membrane*, *Molecular*, *mutualistic*, *National*, *negative*, *OB*, *optical*, *paired*, *phenotype*, *phylogenetic*, *physical*, *plant*, *positive*, *postdoctoral*, *preferential*, *Professional*, *protein*, *Psychiatric*, *Psychological*, *random*, *recurrent*, *Research*, *reward*, *scientific*, *significant*, *spatial*, *specific*, *stellar*, *strongest*, *symbiotic*, *syntrophic*, *telomeric*, *temporal*, *trade*, *word* ***association***,

arbitrary ***assortment***,

arbitrary, *basic*, *common*, *conservative*, *critical*, *fourth*, *fundamental*, *implicit*, *independence*, *key*, *minimal*, *model*, *plausible*, *previous*, *principal*, *priori*, *reasonable*, *scientific*, *simple*, *simplifying*, *theoretical*, *underlying*, *untested*, *widespread* ***assumption***,

sterility ***assurance***,

isolated, *microtubule*, *sperm* ***aster***,

black, *blue*, *green*, *red*, *single*, *triple*, *white* ***asterisk***,

belt, *binary*, *carbonaceous*, *Earth*, *known*, *parent*, *porous*, *stony*, *test*, *Trojan* ***asteroid***,

cortical, *hippocampal*, *human*, *multiple*, *primary*, *rat*, *Retinal*, *single*, *underlying* ***astrocyte***,

Apollo ***astronaut***,

amateur, *Australian*, *French*, *Indian*, *modern*, *observational*, *Planetary*, *professional*, *radio*, *Spanish*, *US*, *visiting* ***astronomer***,

Spanish ***astronomy***,

theoretical ***astrophysicist***,

spinocerebellar ***ataxia***,

northern ***Atlantic***,

brain ***atlas***,

anoxic, *CO*, *cometary*, *Cottrell*, *dwarf*, *extended*, *gas*, *helium*, *hydrogen*, *model*, *nitrogen*, *original*, *outer*, *planet*, *planetary*, *present*, *primitive*, *solar*, *stellar*, *terrestrial*, *upper*, *urban* ***atmosphere***,

central ***atoll***,

adsorbed, *Al*, *alkali*, *all*, *antihydrogen*, *antimony*, *Ar*, *argon*, *artificial*, *backbone*, *Bi*, *boron*, *bosonic*, *bound*, *bromine*, *caesium*, *carbon*, *chlorine*, *Co*, *condensate*, *convert*, *cooled*, *copper*, *deposited*, *deuterium*, *diffusing*, *distinguished*, *dopant*, *embedded*, *evaporated*, *excited*, *fermionic*, *focal*, *framework*, *gas*, *gold*, *guest*, *guided*, *halogen*, *hassium*, *heavy*, *helium*, *host*, *hydrogen*, *impurity*, *individual*, *iodine*, *iron*, *isolated*, *krypton*, *lanthanide*, *lanthanum*, *lattice*, *lithium*, *magnetic*, *mercury*, *metal*, *meter*, *Mo*, *moving*, *NE*, *neighbouring*, *neutral*, *nickel*, *nitrogen*, *observed*, *oxygen*, *phosphorus*, *platinum*, *probe*, *protein*, *pumped*, *recoiling*, *rubidium*, *Rydberg*, *scattering*, *selected*, *selenium*, *Si*, *silicon*, *silver*, *single*, *sodium*, *solute*, *state*, *sulphur*,

surface, *trapped*, *united*, *uranium*, *used*, *vanadium*, *vortex*, *wall*, *X*, *xenon*, *zinc* ***atom***,

ABC, *membrane*, *proteasomal*, *Rad*, *traffic*, *transport*, *vacuolar* ***ATPase***,

cell, *crossbridge*, *cytoskeletal*, *eyestalk*, *flexible*, *microtubule*, *muscle*, *social*, *specific*, *spindle*, *syntelic* ***attachment***,

anthrax, *asthma*, *autoimmune*, *biological*, *bioterror*, *bioterrorist*, *bioweapon*, *bomb*, *chemical*, *clinical*, *combined*, *enzyme*, *Florida*, *heart*, *herbivore*, *immune*, *immunological*, *infection*, *ischaemic*, *parasitoid*, *pathogen*, *personal*, *recent*, *September*, *subsequent*, *suicide*, *terrorist*, ***attack***,

educational ***attainment***,

ambitious, *breeding*, *earlier*, *failed*, *feeding*, *future*, *ingenious*, *initial*, *international*, *invasion*, *latest*, *legislative*, *mating*, *noble*, *penetration*, *preliminary*, *previous*, *recent*, *serious*, *successful*, *systematic*, *theoretical*, *unsuccessful* ***attempt***,

unwanted ***attention***,

ambiguous, *conservative*, *intellectual*, *positive*, *prevailing* ***attitude***,

patent, *Promega*, *university*, *US* ***attorney***,

axonal, *chemical*, *soluble* ***attractant***,

electrostatic, *mutual*, *tourist*, *weak* ***attraction***,

chaotic, *strange* ***attractor***,

common, *desirable*, *functional*, *key*, *selection*, *stimulus*, *unique*, *visual* ***attribute***,

general, *intended*, *lay*, *target* ***audience***,

management ***audit***,

Saint ***Augustine***,

crested ***auklet***,

metamorphic ***aureole***,

cerebellar ***auricle***,

southern ***aurora***,

Jane ***Austen***,

northern ***Australia***,

Aboriginal ***Australian***,

Robust ***Australopithecine***,

distinguished, *expert*, *lead*, *original*, *present*, *previous*, *published*, *senior*, *sole*, ***author***,

honorary ***authorship***,

essential, *islet* ***autoantigen***,

sealed ***autoclave***,

lag, *spatial*, *temporal* ***autocorrelation***,

electric ***automobile***,

intermolecular ***autophosphorylation***,

Emulsion ***autoradiogram***,

presynaptic ***autoreceptor***,

ancestral, *mouse* ***autosome***,

terrestrial ***autotroph***,

active ***auxin***,

lysine ***auxotroph***,

flux, *largest*, *spontaneous*, *triangular* ***avalanche***,

promising, *unexplored* ***avenue***,

annual, *arithmetic*, *basin*, *Calculated*, *class*, *ensemble*, *equilibrium*, *experimental*, *genome*, *global*, *group*, *Holocene*, *local*, *Movement*, *moving*, *population*, *radial*, *regional*, *sample*, *show*, *simple*, *spatial*, *time*, *wavevector*, *weighted*, *world* ***average***,

odour, *taste* ***aversion***,

basal ***avialan***,

basal ***avian***,

specific ***avnosmia***,

shade ***avoider***,

annual, *career*, *Development*, *funding*, *Innovator*, *Institute*, *investigator*, *Lasker*, *mentoring*, *Merit*, *Nobel*, *Postdoctoral*, *Program*, *research*, *Scientist*, *service*, *Society*, *training* ***award***,

grant ***awardee***,

hand ***axe***,

leaf ***axil***,

active, *cell*, *ChAT*, *commissural*, *cortical*, *damaged*, *developing, extending*, *giant*, *growing*, *guide*, *Guiding*, *individual*, *injured*, *longitudinal*, *mammalian*, *many*, *motor*, *myelinating*, *nerve*, *OB*, *olfactory*, *ORN*, *PC*, *peripheral*, *photoreceptor*, *pioneer*, *presynaptic*, *Regenerated*, *retinal*, *root*, *sensory*, *several*, *single*, *squid*, *temporal*, *transfected* ***axon***,

flagellar ***axoneme***,

solar ***azimuth***.

4.4.4.2 Lista kolokacji trójwyrazowych

Cancer Chromosome, *transverse chromatic* ***Aberration***,

last probable ***absence***,

semiconductor saturable ***absorber***,

induced optical, *narrow optical* ***absorption***,

average iron, *higher relative*, *high relative*, *host cell*, *larger weed*, *low diatom*, *low local*, *low metal*, *mantle source*, *observed elemental*, *peak population*,

relatively high, *relatively low*, *relative species*, *solar system*, *Total cell*, *total host*, *total Li*, *total Prochlorococcus*, *typical bacterial*, *very low* ***abundance***,

swollen thorn ***acacia***,

remarkable rate ***acceleration***,

laser wakefield, *powerful particle* ***accelerator***, *alternative electron*, *hydrogen bond*, *metabolic electron*, *potential proton*, *single electron*, *terminal electron* ***acceptor***,

annual snow ***accumulation***,

greatest scientific, *most significant* ***achievement***,

abundant amino, *acidic amino*, *additional amino*, *aliphatic amino*, *arginine amino*, *aromatic amino*, *basic amino*, *canonical amino*, *catalytic amino*, *cellular fatty*, *charged amino*, *common amino*, *Conserved amino*, *critical amino*, *cysteine amino*, *destabilizing amino*, *different amino*, *engineered nucleic*, *essential amino*, *essential fatty*, *excitatory amino*, *extractable mycolic*, *first amino*, *free amino*, *free fatty*, *hydrophobic amino*, *identical amino*, *Important amino*, *individual amino*, *intact amino*, *key amino*, *last amino*, *leucine amino*, *locked nucleic*, *lysine amino*, *major mycolic*, *meteoritic amino*, *natural amino*, *neutral amino*, *new amino*, *peptide nucleic*, *phospholipid fatty*, *phosphorylated amino*, *polar amino*, *polyunsaturated fatty*, *Purified nucleic*, *saturated fatty*, *serine amino*, *single amino*, *small amino*, *specific amino*, *target amino*, *tyrosine amino*, *underivatized amino*, *unnatural amino*, *unsaturated fatty*, *unusual amino*, *viral nucleic* ***acid***,

Clean Air ***act***,

transcription effector, *visually guided* ***action***,

yeast transcriptional ***activator***,

key feeding ***adaptation***,

death domain ***adapter***,

putative transcriptional ***adaptor***,

Saharan dust ***addition***,

multiple colorectal ***adenoma***,

research integrity ***adjudication***,

growth cone, *Holocene ice*, *most important* ***advance***,

presidential science, *previous science* ***adviser***,

externally mixed, *secondary organic*, *strongly absorbing* ***aerosol***,

early optical ***afterglow***,

instrumentally recorded ***aftershock***,

apparent ventilation, *cosmic dark*, *detrital mineral*, *feldspar cooling*, *high median*, *identical geological*, *Late Neolithic*, *major ice*, *martian ice*, *more precise*, *nominal exposure*, *precise absolute*, *recent ice*, *relatively old*, *relatively young*, *rescaled conventional*, *source formation*, *standardized median*, *stratospheric mean*, *surface exposure*, *surface reservoir*, *total gas*, *unstandardized median*, *weighted mean* ***age***,

chiral complexing, *Infectious Disease*, *potential biowarfare*, *registered patent* ***Agent***,

subsynaptic translation ***aggregate***,

Lower Cambrian ***agnathan***,

cannabinoid receptor, *melanocortin receptor*, *potent inverse* ***agonist***,

good quantitative, *material transfer*, *Property Rights*, *regional forest* ***agreement***,

achiral amino ***alcohol***,

complex scoring, *gene prediction*, *heuristic search* ***algorithm***,

acid sequence, *amino acid*, *multiple sequence*, *Protein sequence*, *west side* ***alignment***,

class I, *dominant negative*, *functional null*, *functional paternal*, *human disease*, *molecular null*, *most common*, *nucleotide polymorphic*, *null cad*, *paternally inherited*, *Paxton benthic*, *severe hypomorphic*, *unstable mutant* ***allele***,

grass pollen ***allergen***,

human platelet ***alloantigen***,

iron nickel ***alloy***,

amino acid, *neurotransmitter receptor* ***alteration***,

cloud forest ***altitude***,

Asian Pacific, *modern native* ***American***,

fatty acid ***amide***,

optically active ***amine***,

abnormally high, *arbitrarily small*, *daily sleep*, *extremely small*, *much smaller*, *relatively large*, *relatively small*, *REM sleep*, *reported sleep*, *very large*, *very small* ***amount***,

Josephson parametric, *optical parametric*, *patch clamp* ***amplifier***,

acoustic pressure, *action potential*, *CaT current*, *electric field*, *fault reflection*, *field potential*, *index modulation*, *instantaneous current*, *maximal current*, *maximum stack*, *peak current*, *slowly varying*, *spin wave*, *tail current*, *threshold strain* ***amplitude***,

fused silica ***ampoule***,

Iron deficiency ***anaemia***,

anticancer nucleoside, *closest modern*, *interstellar ice*, *more potent*, *organic zeolite*, *reactive sugar*, *Solar System*, *transition state* ***analogue***,

differential mobility, *infrared gas*, *PRISM genetic* ***analyser***,

last common, *putative wild*, *recent common*, *shared common* ***ancestor***,

putative basal ***angiosperm***,

average turning, *electron emission*, *equilibrium glide*, *faceplate rotation*, *field zenith*, *high emission*, *high phase*, *high zenith*, *mean turning*, *orbital inclination*, *polarization position*, *solar phase*, *solar zenith*, *wide viewing* ***angle***,

BAC transgenic, *bilaterally symmetrical*, *genetically modified* ***Animal***,

Supplementary Information ***Animation***,

doubly charged, *fatty acid*, *moderately hard*, *multiply charged*, *planar anthracene* ***anion***,

Gene Ontology ***Annotation***,

gold membrane ***anode***,

red harvester, *Saharan desert* ***ant***,

glutamate receptor, *glycine site*, *kainate receptor*, *neurokinin receptor*, *vanilloid receptor* ***antagonist***,

gravitational wave ***antenna***,

clinically important, *cyclic peptide*, *macrocyclic lactone* ***antibiotic***,

geometrically frustrated ***antiferromagnet***,

blood group, *cell surface*, *class I*, *Ebola virus*, *erythrocyte surface*, *exogenous lipid*, *histocompatibility leukocyte*, *human leukocyte*, *human nuclear*, *integral membrane*, *variant surface* ***antigen***,

sintered diamond ***anvil***,

African great, *captive great* ***ape***,

Russian wheat ***aphid***,

abnormal dorsal, *filamentous integumental* ***appendage***,

Golden Delicious ***apple***,

international patent, *most important*, *potential therapeutic*, *repeated brief*, *research grant*, *US Patent* ***Application***,

association mapping, *candidate gene*, *comparative genomic*, *completely different*, *different experimental*, *functional genomics*, *genetic complementation*, *Machine Learning*, *molecular genetic*, *more abstract*, *more focused*, *more rational*, *more reductionist*, *more sophisticated*, *most promising*,

most viable, *new therapeutic*, *positional cloning*, *quite different*, *radical new*, *reverse genetic*, *very different* ***approach***,

generalized gradient, *local density*, *reasonable first* ***approximation***,

plasma membrane ***aquaporin***,

elaborate terminal ***arbor***,

axon terminal ***arborization***,

fourth pharyngeal, *second pharyngeal* ***arch***,

complex molecular, *different protein* ***architecture***,

Los Alamos, *PubMed Central* ***archive***,

accessible surface, *active kitchen*, *auditory cortical*, *auditory processing*, *biggest growth*, *blue shaded*, *Breeding unit*, *broad research*, *buried surface*, *climatically suitable*, *coastal upwelling*, *coherently illuminated*, *Connolly surface*, *cortical motor*, *cortical surface*, *crater density*, *dark grey*, *different brain*, *dorsal premotor*, *forebrain auditory*, *frontal cortical*, *geographic sampling*, *gouge surface*, *greater surface*, *grey shaded*, *higher surface*, *highly competitive*, *highly endemic*, *high surface*, *ice sheet*, *inferior temporal*, *integrated cortical*, *known spawning*, *land surface*, *larger surface*, *large geographic*, *Large mode*, *large protected*, *large surface*, *lateral belt*, *lateral hypothalamic*, *lateral intraparietal*, *LIS core*, *major research*, *marine protected*, *much larger*, *new research*, *occlusal surface*, *olfactory cortical*, *particle surface*, *rapidly growing*, *relatively large*, *relatively small*, *sound input*, *specific surface*, *superior temporal*, *supplementary motor*, *temporal cortical*, *temporal visual*, *very small*, *visual cortical*, *visual processing* ***area***,

artificial lever, *inner dynein*, *muscle moment*, *outer dynein* ***arm***,

tandem gene ***arrangement***,

carbon nanotube, *cortical microtubule*, *crystalline colloidal*, *Cytokine Antibody*, *highly ordered*, *Human genome*, *Josephson junction*, *large millimetre*, *linear mixing*, *long tandem*, *nonlinear waveguide*, *optical sensor*, *ordered dopant*, *stable isotope*, *telomeric repeat*, *Very Large* ***array***,

cell cycle, *replication fork* ***arrest***,

inherited cardiac, *polymorphic ventricular* ***arrhythmia***,

Bold horizontal, *dashed red*, *large open*, *light blue*, *solid black*, *thin black* ***arrow***,

Old World ***artiodactyl***,

fine vitric ***ash***,

most important, *most interesting*, *most intriguing*, *most significant*, *most striking* ***aspect***,

actin assembly, *actin polymerization*, *alkaline comet*, *animal cap*, *Animal cap*, *bacterial ingestion*, *binary choice*, *blastomere injection*, *cellular toxicity*, *cell adhesion*, *cell attachment*, *Cell proliferation*, *chromatin immunoprecipitation*, *Chromatin immunoprecipitation*, *Colony formation*, *colony formation*, *competitive binding*, *convergent extension*, *Dicer activity*, *Disk diffusion*, *displacement helicase*, *dual luciferase*, *Dual luciferase*, *Electromobility shift*, *endosome fusion*, *enzyme activity*, *ERK activation*, *Filter binding*, *filter binding*, *filter overlay*, *fluorescence polarization*, *fluorometric competition*, *Gal transactivation*, *gel retardation*, *gel shift*, *Glucose production*, *Haematopoietic colony*, *hair growth*, *histone deacetylase*, *histone methyltransferase*, *immunocomplex kinase*, *kinase activity*, *leaf cell*, *Ligand overlay*, *Lipid kinase*, *Local adaptation*, *luciferase activity*, *Luciferase Reporter*, *membrane repair*, *microtubule formation*, *mobility shift*, *neurite outgrowth*, *nuclear microinjection*, *nuclease protection*, *nucleotide exchange*, *overlay binding*, *peptide library*, *Primer extension*, *protein binding*, *Protein hybridization*, *protein interaction*, *Protein kinase*, *protein kinase*, *receptor binding*, *Receptor binding*, *reporter gene*, *ribonuclease protection*, *ring sprouting*, *RNase protection*, *SCaM overlay*, *scintillation proximity*, *Slice overlay*, *Solution binding*, *sphere formation*, *telomerase activity*, *thymidine incorporation*, *transient expression*, *transient retrotransposition*, *transient transfection*, *transwell chemotaxis*, *transwell migration*, *tropism switch*, *turnover GTPase*, *ubiquitin ligase*, *vivo phosphorylation*, *western blot*, *wound migration* ***assay***,

local species, *mantle mineral*, *mixed diatom*, *planktic foraminiferal*, *stone tool* ***assemblage***,

scientific risk, *standard analytical*, *thorough risk* ***assessment***,

chemical shift, *putative functional*, *secondary structure* ***assignment***,

Postdoctoral Research ***Associate***,

German Medical, *National Hydrogen*, *statistically significant*, *stimulus stimulus*, *viral load* ***association***,

widely held ***assumption***,

highly porous, *Koronis family*, *main belt*, *Near Earth* ***asteroid***,

ATLAS Stellar, *terrestrial planet* ***Atmosphere***,

adsorbed sulphur, *amide nitrogen*, *any framework*, *carbonyl oxygen*, *Co wall*, *daughter chlorine*, *diphthamide NE*, *energetic neutral*, *extra oxygen*, *ground state*, *neutral hydrogen*, *single dopant*, *single magnetic*, *single slow*, *surface carbon* ***atom***,

possible bioterror, *potential bioterrorist*, *transient ischaemic* ***attack***,

European Patent ***Attorney***,

learned odour ***aversion***.

4.4.5 Dyskusja wyników

Główną zaletą metody przedstawionej w tej części rozprawy jest bardzo znaczna liczba wyodrębnianych kolokacji a – w związku z tym – bogactwo materiału językoznawczego, który można wykorzystać w dalszej analizie (na przykład badania nad kolokacjami i ich zmiennością (wymiennością członów kolokacji), wyodrębnianie n-gramów stanowiących najsilniejsze kolokacje (a więc najmniej zmiennych), ocena wykorzystania podobnych form przez użytkowników języka).

Ponadto, dzięki zastosowaniu takiej właśnie metody ekscerpcji, przeszukiwanie obejmuje bardzo szeroki zakres materiału zawartego w zbiorze tekstów (nie korzysta się na przykład z wcześniej przygotowanych list elementów kolokacji, na przykład wszystkich kolokacji, w których występuje dane słowo). Metoda pozwala na ekscerpcję wszystkich n-gramów (w tym przypadku bigramów i trigramów) występujących w poddawanym analizie zbiorze tekstów w liczbie pojedynczej i mnogiej. Ponadto zastosowane podejście umożliwia dokładne i jednoznaczne ustalenie ostatniego elementu kolokacji.

Metoda nie wymaga wykorzystania bardzo dużych zbiorów tekstów, ponieważ nie korzysta ze złożonych obliczeń statystycznych. Jedynym wymogiem jest występowanie danego n-gramu zarówno w liczbie pojedynczej, jak i mnogiej. Celem etapu generowania n-gramów jest jedynie uzyskanie z danych wejściowych listy, nie zaś obliczanie częstości ich występowania.

Metodę łatwo zautomatyzować. Przedstawiona w niniejszym rozdziale procedura nie jest, rzecz jasna, automatyczna, jednak wszystkie etapy – z istotnym wyjątkiem ostatniego – można (korzystając z umiejętności programistycznych) przekształcić w kompletny pakiet wykonujący etapy od pierwszego do czwartego (a więc z wyjątkiem etapu manualnego) w jednej sekwencji.

Największą wadą metody jest jednoznaczna konieczność uwzględnienia analizy manualnej do usunięcia wszystkich wierszy niebędących kolokacjami (patrz rozdział 4, punkt 4.3). Mimo że odsetek wierszy usuniętych jest stosunkowo niewielki (szacowany na 7,5%), osoba biegle posługująca się danym językiem musi z dużą dokładnością przejrzeć i zweryfikować listę uzyskaną w etapie 4. Istotnym czynnikiem wpływającym na jakość listy poddawanej analizie manualnej i liczbę n-gramów usuniętych jest zastosowany analizator morfologiczny. Im program ten jest bardziej precyzyjny (bogatszy słownik, precyzyjniejsze algorytmy przypisywania indeksów), tym mniejsza jest konieczność interwencji manualnej.

Inną niedogodnością jest to, że metoda nie pozwala stwierdzić, czy badana kolokacja nie składa się z liczby słów większej niż n (n jest liczbą słów w badanym n-gramie). Inaczej mówiąc, metoda nie pozwala jednoznacznie określić elementu inicjalnego kolokacji. Prawdopodobnie można by zaproponować metodę porównywania listy n-gramów z listą n+1-gramów (n+2…) i w ten sposób sprawdzać, czy być może szukana kolokacja jest w istocie na przykład n+1 gramem, nie zaś n-gramem. Ta niedoskonałość także zwiększa znaczenie dokładnego przeprowadzenia etapu analizy manualnej (wszystkie n-gramy uznane przez oceniającego za niepełne należy usunąć).

4.4.6 Możliwości rozwoju metody i dalszych badań

Zasadę, na której opiera się opisywana metoda, można zastosować także do innych języków. Ponieważ jednak jej podstawowym elementem jest wyrażenie regularne pozwalające wyodrębnić formy w liczbie pojedynczej i mnogiej (a także formy dopełniacza saksońskiego), jest ona bardziej niż pozostałe metody opisywane w pracy zależna od gramatyki danego języka, co oznacza że bez znacznych zmian, czyli nie tylko modyfikacji poszczególnych etapów, ale także wprowadzenia etapów nowych lub usunięcia dotychczasowych, nie można jej przenieść na inne języki. Łatwość bowiem zastosowania przedstawionej metody zależy od tego, czy końcówki liczby mnogiej występujące w języku można wyodrębnić w sposób konsekwentny (na tym etapie nie jest konieczna analiza o charakterze językoznawczym). W języku angielskim końcówki liczby mnogiej są w znakomitej większości przypadków jednolite, czyli tworzeniem liczby mnogiej rządzi prosta reguła (oprócz stosunkowo niewielkiej liczby wyjątków[75]), co oznacza, że mogą zostać wyodrębnione przy użyciu stosunkowo prostego wyrażenia regularnego. Zastosowana procedura pozwala z listy n-gramów uporządkowanych alfabetycznie wyodrębnić wszystkie n-gramy zakończone literą *s* (lub

[75] Np.: *wife – wives*, *life – lives*, *man – men*, *child – children*, *ox – oxen*, *alge – algae*, *formula – formulae*, *corpus – corpora*, *bacterium – bacteria*, *nucleus – nuclei*, *matrix – matrices*, *analysis – analyses*, *thesis – theses*, *libretto – libretti*, *cherub – cherubim* (por. też Loberger, Shoup 2001).

's w przypadku dopełniacza saksońskiego), przed którymi występuje taki sam n-gram bez litery *s* (lub znaków *'s*). Należy zauważyć, że lista może także obejmować czasowniki w czasie teraźniejszym prostym (końcówka *s* występuje w trzeciej osobie liczby pojedynczej czasu teraźniejszego prostego), jednak ponieważ stosuje się w kolejnym etapie analizę morfologiczną, wszystkie n-gramy tego typu zawierające czasownik jako element terminalny mogą zostać wyszukane i usunięte.

Podejście takie wydaje się niemożliwe w przypadku języków fleksyjnych, na przykład słowiańskich. W języku polskim, na przykład, nie występuje jednolita końcówka liczby mnogiej rzeczownika, tj. o identycznej postaci dla wszystkich rzeczowników. Dlatego też jedynym rozwiązaniem tego problemu byłaby najpierw analiza morfologiczna listy n-gramów, a następnie – na podstawie form zlematyzowanych – identyfikacja n-gramów występujących w liczbie pojedynczej i mnogiej. W ten sposób można by także wykorzystać występowanie przypadków gramatycznych w języku polskim i w ten sposób wyodrębniać interesujące nas jednostki (na przykład *dysk twardy*, *dysku twardego* itd.). Główną wadą takiego z kolei podejścia – choć wymuszonego przez właściwości języka – jest to, ze całą listę n-gramów, a więc plik w przybliżeniu n-krotnie większy od wejściowego (gdzie n jest liczbą wyrazów w n-gramie; w opisywanej metodzie n jest równe 2 lub 3), należałoby poddać analizie morfologicznej.

Jeśli chodzi o udoskonalenie metody przedstawionej w obecnym kształcie, mimo że dostarcza bardzo znacznej liczby interesujących danych, wydaje się, że byłoby to możliwe. Wartość progową rejestrowania n-gramów (w szczególności w odniesieniu do liczby pojedynczej, która powinna występować częściej) można by przyjąć nie jako jeden, ale dwa lub nawet więcej. (W przedstawionej metodzie uwzględniano także te n-gramy, które występują jednokrotnie, czyli wszystkie wykorzystywano jako dane wejściowe w kolejnym etapie.) W ten sposób wyodrębniano by tylko te n-gramy, które występują w tekście więcej niż raz w każdej liczbie gramatycznej. Z drugiej jednak strony takie podejście znacznie skróciłoby listę końcową n-gramów.

4.4.7 Porównanie wyników uzyskanych za pomocą metody ekscerpcji kolokacji w oparciu o akronimy (metoda A) oraz metody ekscerpcji kolokacji rzeczownikowych na podstawie obserwacji parametru końcówki liczby mnogiej (metoda B)

Wydaje się, że interesujące byłoby – także w perspektywie ewentualnych dalszych badań – porównanie wyników, które uzyskano za pomocą dwóch metod ekscerpcji kolokacji. Pytania, na które należałoby w tym miejscu odpowiedzieć, to:

1) Czy jedna z metod daje lepsze wyniki (więcej kolokacji, mniej wyników, które należy usuwać manualnie),

2) W jakim stopniu, jeśli w ogóle, wyniki uzyskane obiema metodami są zbieżne.

W analizie wykorzystano kolokacje wynikowe złożone z dwóch słów, w których słowem drugim jest *acid* (pierwszy przykład), *factor* (drugi przykład) oraz *receptor* (trzeci przykład). Wynik takiego porównania jest następujący (czcionką wytłuszczoną przedstawiono pokrywające się wyniki dla obu metod):

1. Acid:

Metoda w oparciu o akronimy:

amino acid, arachidonic acid, benzoic acid, bongkrekic acid, cinnamic acid, domoic acid, gibberellic acid, Hyaluronic acid, jasmonic acid, okadaic acid, phosphatidic acid, retinoic acid, salicylic acid

Metoda w oparciu o końcówki liczby mnogiej:

acetic, acylamino, **amino,** aqueous, **arachidonic,** aspartic, **benzoic,** bile, carbonic, carboxylic, deoxyribonucleic, dicarboxylic, dihydroxyeicosatrienoic, epoxyeicosatrienoic, fatty, general, glutamic, haloacetic, humic, lactic, Lewis, linoleic, linolenic, liquid, methacrylic, mugineic, mycocerosic, mycolic, nitric, nucleic, oleanonic, organic, oxalic, **phosphatidic,** phosphinic, phosphoamino, phosphonic, pinonic, ribonucleic, sialic, sulphenic, sulphinic, sulphonic, sulphuric, threonic, uronic, weak acid

2. Factor

Metoda w oparciu o akronimy:

Branching factor, complementing factor, elongation factor, fill factor, lachrymatory factor, lethal factor, Nod factor, nuclear factor, oedema factor, release factor, Retardation factor, retention factor, rheumatoid factor, tissue factor, transcription factor , trigger factor

Metoda w oparciu o końcówki liczby mnogiej:

accessory, acting, activated, additional, adhesion, aetiological, alignment, amplification, angiogenesis, angiogenic, antiviral, assembly, associated, auxiliary, basal, bHLHPAS, binding, biological, biotic, bound, **branching,** canonical, causal, causative, cell, cellular, chemotactic, CID, circulating, cleavage, cleavagepolyadenylation, climatic, coagulation, colonization, common, complement, complex, complicating, condition, confounding, conjugating, constant, containing, contains, contrast, contributing, contributory, control, controlling, conversion, cooling, correction, coupling, critical, crucial, cytoplasmic, cytosolic, death, decisive, depletion, determination, determining, differentiation, diffusible, diffusion, dilution, displacement, dominant, driving, duty, ecological, effectiveness, efficiency, **elongation,** emission, endogenous, energetic, enhancement, enrichment, environmental, epigenetic, essential, evolutionary, exacerbation, exchange, explanatory, export, external, extrinsic, filling, forcing, Forkhead, form, fractionation, GATA, general, genetic, geometric, global, Growth,

guidance, gust, helix, homeodomain, hormone, host, human, humoral, hyperpolarizing, impact, important, incompatibility, indicated, inducible, inducing, inflation, inhibitory, initiation, interacting, intrinsic, involved, key, kinetic, leading, **lethal,** licensing, limiting, local, loss, lytic, major, maternal, mendelian, missing, mitochondrial, molecular, mortality, mosquito, multiplicative, myogenic, negative, neuroprotective, neurotrophic, **Nod,** nodulation, normalization, novel, nucleation, numerical, **oedema,** organizing, paracrine, pathogenicity, permissive, phase, polyadenylation, possible, predisposing, predominant, preexponential, primary, prime, principal, processivity, prognostic, protection, protective, protein, psychological, Purcell, quality, random, recognition, recombination, reduction, redundant, regulatory, related, relaxing, **release,** releasing, relevant, reliability, remodeling, remodelling, repair, replication, repression, reprogramming, resistance, response, restriction, **rheumatoid,** risk, safety, scale, scaling, scattering, secreted, selective, selectivity, sigma, signalling, significant, single, social, soluble, specific, specificity, splice, splicing, statistical, steric, stimulating, stimulatory, stress, stretching, structural, structure, survival, systematic, temperature, tethering, thermal, time, toxic, trans, **transcription,** transcriptional, translation, transport, **trigger,** trophic, ultimate, unidentified, unknown, variable, virulence, Waller, weighting, factor

3. Receptor:

Metoda w oparciu o akronimy:

adrenergic receptor, androgen receptor, cytoplasmic receptor, death receptor, estrogen receptor, glucocorticoid receptor, insulin receptor, nuclear receptor, odorant receptor, olfactory receptor, progesterone receptor, scavenger receptor, SRP receptor

Metoda w oparciu o obserwację parametru końcówki liczby mnogiej:

ABA, accessory, acetylcholine, ACh, acid, acidkainate, activatable, activated, activating, activation, active, activin, Additional, adhesion, adiponectin, **adrenergic**, affinity, AGE, alternative, AMPAkainate, **androgen**, animal, antigen, Article, aspartate, associated, auditory, auxin, BAFF, basolateral, Benzodiazepine, binding, bound, brain, brassinosteroid, cAMP, candidate, cannabinoid, capsaicin, cargo, CB, CD, cell, cellular, certain, chemoattractant, chemokine, chemosensory, chemotaxis, chimaeric, cholinergic, chordotonal, cloned, cognate, complement, corresponding, costimulatory, coupled, cytokine, cytokinin, **death**, decoy, dependence, designed, distinct, domain, dopamine, ecdysone, elusive, endogenous, endothelial, endothelin, enterotoxin, entry, Eph, erythropoietin, **estrogen**, ethylene, export, expressed, extrasynaptic, family, fibronectin, folate, forbidden, four, frizzled, functional, fusion, ghrelin, **glucocorticoid**, glutamate, glutamatergic, glycine, gonadotropin, guidance, heptahelical, heterodimeric, heteromeric, histamine, homing, homomeric, hormone, host, human, hybrid, hypocretinorexin, identified, immune, immunoglobulin, import, individual, inhibitory, **insulin**, intact, integrin, intracellular, ionotropic, IP, kainate, key, kinase, Kit, known, labelled, lacking, leptin, leukotriene, liganded, lipoprotein, lymphocyte, lysis, macrocylic, macrophage, major, mannose, melanocortin, membrane, molecular, multiple, muscarinic, mutant, nACh, native, netrin, neurokinin, neuronal, neurotensin, neurotransmitter, nicotine, nicotinic, Nodal, noradrenergic, Notch, novel, **nuclear**, ob, **odorant**, odour, oestrogen, **olfactory**, opiate, opioid, orexin, orphan, oxytocin, PAMP, particular, peripheral, pheromone, phosphatidylserine, plexin, postsynaptic, preprotein, Presynaptic, primary, **progesterone**, prostanoid, protein, prototypic, PubMed, purinergic, putative, recognition, redundant, Reelin, related, respective, resulting, retinoid, ryanodine, **scavenger**, secretagogue, SEMA, semaphorin, sensory, serotonin, seven, several, signalling, single, soluble, somatostatin, sorting, specific, steroid, surface, sweet, synaptic, systemin, tachykinin, Tar, target, taste, tetraloop, three, thrombin, Tie, Toll, Tom, Torpedo, toxin, TRAIL, transferrin, transgenic, transmembrane,

transport, triggering, truncated, type, unidentified, unliganded, unoccupied, vasopressin, virus, vomeronasal, X, xenobiotic receptor

Dane liczbowe dotyczące porównania wyników uzyskanych za pomocą obu metod podaje tabela 9.

Tabela 9. Porównanie liczby uzyskanych kolokacji dla obu metod oraz kolokacji powtarzających się.

Kolokacje	Liczba kolokacji (metoda A)	Liczba kolokacji (metoda B)	Liczba powtarzających się kolokacji
acid	13	47	4
factor	16	240	9
receptor	13	238	11

Na podstawie uzyskanych wyników można jednoznacznie stwierdzić, że metoda B daje znacznie więcej wyników niż metoda A. Mimo jednak znacznej, kilku- a nawet kilkunastokrotnej różnicy, znaczna część kolokacji uzyskanych metodą A nie występuje w wynikach zastosowania metody B (ogółem 18 na 42 kolokacje). Oznacza to, że w danym zbiorze tekstów warto zastosować obie metody, aby uzyskać pełniejsze dane dotyczące kolokacji. Ponadto opisywanym wcześniej niedostatkiem metody B jest niepewność wyznaczenia elementu inicjalnego kolokacji, która nie występuje w przypadku metody A korzystającej z akronimów – wynika z tego konieczność większej ingerencji manualnej w uzyskane wyniki. Co więcej wyniki zastosowania metody B obejmują stosunkowo mniej niż w przypadku metody A fraz o charakterze jednostek wielowyrazowych.

4.5 Podsumowanie

W przypadku wszystkich trzech proponowanych i zastosowanych metod należało sformułować hipotezę (H) w ramach metodologii językoznawstwa stosowanego oraz postulat (P) badawczy, określić dyrektywy prowadzenia badań (D), zaprojektować procedury zastosowania dyrektyw oraz na podstawie wyników sformułować pytania (Q) będące punktem wyjścia do dalszej pracy.

Procedury zostały szczegółowo przedstawione w omówieniu poszczególnych metod, zaś hipotezy, postulaty, dyrektywy i pytania podsumowano w tabeli 10.

Tabela 10. Hipotezy (H), postulaty (P), dyrektywy (D) i pytania (Q) w przebiegu pracy badawczej.

Porównanie wybranych metod ekscerpcji jednostek nowych	**Metoda ekscerpcji kolokacji w oparciu o akronimy**	**Metoda ekscerpcji kolokacji rzeczownikowych na podstawie obserwacji parametru końcówki liczby mnogiej**
H: Zastosowanie wyróżników leksykalnych i interpunkcyjnych oraz analizatora morfologicznego i wyszukiwarki internetowej umożliwia wyodrębnienie jednostek nowych.	H: Zastosowanie wyróżników w postaci akronimów pozwala na wyodrębnienie kolokacji.	H: Zastosowanie wyróżników w postaci końcówki liczby mnogiej pozwala na wyodrębnienie kolokacji.
P1: W zbiorze tekstów naukowych występuje duża liczba jednostek nowych.	P1: W zbiorze tekstów naukowych występuje duża liczba kolokacji.	P1: W zbiorze tekstów naukowych występuje duża liczba kolokacji.
D1: Zbiór tekstów powinien opierać się na powszechnie uznawanym czasopiśmie naukowym. D2: Teksty są dostępne w formacie .html. D3: Należy przetworzyć teksty w format .txt. D4: Należy wyodrębnić frazy poprzedzone	D1: Zbiór tekstów powinien opierać się na powszechnie uznawanym czasopiśmie naukowym. D2: Należy wybrać teksty najnowsze. D3: Teksty są dostępne w formacie .html. D4: Należy wybrać teksty najnowsze. D5: Należy przetworzyć teksty w format	D1: Zbiór tekstów powinien opierać się na powszechnie uznawanym czasopiśmie naukowym. D2: Teksty są dostępne w formacie .html. D3: Należy wybrać teksty najnowsze. D4: Należy przetworzyć teksty w format .txt.

Porównanie wybranych metod ekscerpcji jednostek nowych	**Metoda ekscerpcji kolokacji w oparciu o akronimy**	**Metoda ekscerpcji kolokacji rzeczownikowych na podstawie obserwacji parametru końcówki liczby mnogiej**
lub ograniczone wyróżnikami. D5: Należy zastosować analizator morfologiczny w celu usunięcia wyrazów znanych. D6: Należy usunąć wyrazy zawierające wielką literę. D7: Należy zastosować przeglądarkę internetową w celu wyodrębnienia proponowanych jednostek nowych. D8: Należy dokonać manualnej weryfikacji uzyskanych jednostek nowych. D9: Należy przedstawić wyniki.	.txt. D6: Należy wyodrębnić frazy, po których następuje ich akronim. D7: Należy przedstawić wyniki w jednej kolumnie dla kolokacji dwuwyrazowych i w dwóch kolumnach dla kolokacji trójwyrazowych.	D5: Należy wyodrębnić frazy użyte w liczbie pojedynczej oraz mnogiej. D6: Należy dokonać manualnej weryfikacji uzyskanych kolokacji. D7: Należy przedstawić wyniki w postaci wyliczenia elementów inicjalnych kolokacji, zakończonego podaniem elementu terminalnego.
Q1: Zastosowanie metody do tekstów w innych językach. Q2: Zwiększenie wydajności i skuteczności metody.	Q1: Zastosowanie metody do tekstów w innych językach. Q2: Zwiększenie wydajności i skuteczności metody.	Q1: Zastosowanie metody do tekstów w innych językach. Q2: Zwiększenie wydajności i skuteczności metody.

Porównanie wybranych metod ekscerpcji jednostek nowych	**Metoda ekscerpcji kolokacji w oparciu o akronimy**	**Metoda ekscerpcji kolokacji rzeczownikowych na podstawie obserwacji parametru końcówki liczby mnogiej**
Q3: Opracowanie i zastosowanie innych wyróżników. Q4: Połączenie metody w jeden pakiet – pełna automatyzacja oprócz końcowej oceny.	Q3: Połączenie metody w jeden pakiet – pełna automatyzacja. Q4: Sprawdzenie, czy możliwe i skuteczne jest wyszukiwanie, jeśli akronim i jego rozwinięcie nie sąsiadują bezpośrednio ze sobą.	Q3: Połączenie metody w jeden pakiet – pełna automatyzacja oprócz końcowej oceny. Q4: Wprowadzenie metody statystycznej (częstość występowania kolokacji). Q5: Dokładniejsza metoda określania elementu inicjalnego kolokacji.

5 Zakończenie

Jednym z głównych celów niniejszej rozprawy było dążenie do pokazania na przykładzie konkretnych metod, że możliwa jest znaczna automatyzacja poszczególnych etapów ekscerpcji danych o znaczeniu leksykograficznym za pomocą metod niewymagających umiejętności programistycznych, a jednocześnie wykorzystujących cechy leksykalno-gramatyczne, nie zaś głównie metody statystyki matematycznej. Dzięki zastosowaniu metod językoznawczych zamiast matematycznych możliwe jest wyodrębnienie fraz występujących w tekście lub zbiorze tekstów jedynie raz.

Bez zastosowania wiedzy programistycznej nie udało się podanych metod zautomatyzować całkowicie, co byłoby teoretycznie możliwe, w przeważającej części pracy uniknięto jednak mechanicznego powtarzania czynności niezautomatyzowanych, w tym analizy tekstu źródłowego (z ważnym być może wyjątkiem ostatecznego etapu analizy manualnej, który – jak się wydaje – będzie niezbędny, dopóki wiedza badacza mieć będzie znaczenie i dopóki człowiek choć w niektórych aspektach badań językoznawczych będzie dysponował możliwościami większymi niż komputer).

Dzięki takiej automatyzacji wszystkie trzy zaproponowane metody dostarczyły oryginalnego (kryterium nowości leksykograficznej), bogatego (kryterium wydajności), ciekawego (kryterium kompozycji, tj. analizy morfologicznej oraz translacyjnej) i rzetelnego (kryterium autentyczności) materiału do ewentualnych dalszych studiów. Wyniki badań świadczą o tym, że możliwe jest zdefiniowanie wyróżników, które z powodzeniem można wykorzystać w ekscerpcji jednostek nowych (ściślej: agnonimów

słownikowych), kolokacji na podstawie akronimów i kolokacji rzeczownikowych za pomocą końcówki liczby mnogiej.

Jak zaznaczono w podsumowaniu metod, wydaje się, że jest możliwe ich rozszerzenie na inne języki (w szczególności metody pozyskiwania kolokacji na podstawie akronimów i jednostek nowych), co świadczyłoby o ich wszechstronności i elastyczności, mimo oczywistych ograniczeń, których nie zdołano wyeliminować, a które w przypadku każdej metody zostały omówione.

Zgodnie z przewidywaniami pełny, liczący około 45 milionów wyrazów zbiór tekstów z tygodnika *Nature* z lat 1997-2005 okazał się niezwykle bogatym źródłem danych o charakterze leksykograficznym i pozwolił na staranne przebadanie skuteczności wszystkich metod.

I tak metoda ekscerpcji jednostek nowych dostarczyła ponad 1000 wyrazów nowych, niezarejestrowanych, metoda ekscerpcji kolokacji w oparciu o akronimy daje ponad 6000 jednostek, zaś metoda ekscerpcji kolokacji wykorzystująca końcówkę liczby mnogiej dała ponad 110 tysięcy wyodrębnionych jednostek.

Można mieć nadzieję, że zaproponowane metody przyczynią się do wzbogacenia wiedzy językoznawczej w zakresie struktury tekstu naukowego w perspektywie neologiczno-kolokacyjnej, zaś ich wyniki zostaną wykorzystane w dalszych badaniach.

Summary

Retrieval of lexical data from electronic texts, based on data from a scientific journal

The principal objective of the thesis is to exemplify, based on three specific and verified methods, that automation of successive steps of lexicographic data retrieval using customised procedures is possible. The common feature of all the three methods is the use of linguistic discriminants which instead of statistics or other mathematical means make it possible to discern the words or phrases of interest. The lexicographic methods developed and tested in the thesis include the retrieval of putative new words or collocations from English-language texts.

Corpus

The texts used for method verification constitute a full collection of articles from the *Nature* scientific weekly between 1997 and 2005 with a total of almost 450 issues and nearly 45 million words. The selection of the input was based on several rational observations:

- *Nature* is focused on the latest developments in natural sciences and medicine (potential new words and new collocations);
- it has a very high Impact Factor which should translate into subject matter and linguistic accuracy;
- it is published every week which contributes to an extensive amount of data;

– it is available in the .html format rather than the PDF format, which enables easy further processing within method verification.

According to the typology discussed in Chapter 2.3, the corpus has the following attributes:

1. Specialised: it concerns a definitely specialised register of language, namely the scientific register.

2. Full-text: full articles make up the corpus with no sampling procedures used.

3. Written language: obviously, the corpus contains no spoken language data.

4. Monolingual: the corpus contains texts from an English-language journal (British English).

5. Non-annotated: no additional information within the corpus is provided; even though the initial data contain certain additional information (headings, footings, titles, etc.), it is removed during data processing.

6. Synchronic: the corpus includes texts published within a relatively short interval; however, when divided into month- or year-long sections, it could provide certain diachronic information.

The following original methods have been developed and tested:

New word retrieval method

Owing to the use of discriminants, detailed analysis of a text or a collection of texts does not require a full-force approach, that is the processing of the whole text.

Two types of discriminants are discussed and used within method application:

- lexical discriminants, or phrases which precede new words (*termed*, *called*, *defined as* and *known as*),
- punctuation discriminants, or those which mark neologisms (single and double quotation marks).

Instead of whole text processing and in a more refined way, extracted from a text are phrases in which the likelihood of the occurrence of a new word is higher than in the whole text (confirmed statistically with a set of lines retrieved from the collection and processed in an identical manner as the for the discriminants; random file). Through a sequence of automated steps, the method yields putative new words using first of all a morphological analyser and a Web browser (Google search engine). The morphological analyser lists all unknown words and the Web browser is used to list words whose frequency on the Internet is the lowest (with the limit arbitrarily set at 1000 hits). The method could possibly be used also for other languages using similar or different discriminants to optimise results.

Method for the retrieval of collocations based on acronyms

The discriminant for collocations used is the acronym of a collocation which follows the collocation, in a form of w_1 w_2 … w_n ($a_1a_2…a_n$), wherein w_n is the n-th word in a collocation and a_n is the first letter of the n-th word used in the acronym. Collocations are retrieved using an optimised regular expression with back references. The regular expression used to retrieve the collocations was the following:

```
\([A-Z]\)[a-z]+[-    ]\([A-Z]\)[a-z]+[-    ]\([A-Z]\)[a-z]+ \(\1\2\3\)
```

The regular expression retrieves strings, such as:

- active site (AS),
- Department of Energy (DOE),

- electron microscope tomography (EMT),
- European Medicines Evaluation Agency (EMEA).

Although the method obviously does not yield all collocations which occur in a text, it is highly reliable in that the beginning and end of the collocation is very precisely and unambiguously defined. Furthermore, it should provide a list of strong collocations, because acronyms are usually used when a collocation, or frequently a term in itself, is needed more than once in a text. Therefore, it should not be an ephemeral entity, but one having an established position in the text and probably also in the language. The method could be very easily transferred into certain other languages without any major modifications.

Method for the retrieval of noun collocations based on plural endings

The discriminant used in the retrieval procedure is the plural ending (and also the Saxon genitive ending). The text is processed to yield bigrams or trigrams (prospective two-word and three-word collocations) which are listed alphabetically. Subsequently, entries are retrieved which differ in the ending only (i.e. they are used more than once in the text or collection of texts and they must be used at least once in the plural and at least once in the singular form). This is effected using the following regular expression involving back references:

```
\([a-z]+\) \([a-z]+\), \1 \2(s|es|'s|s')
```

In the next step, the result is morphologically analysed to remove any entries whose last word is not a noun. While a disadvantage of the method is that it does not enable the beginning of collocations to be precisely determined, due to which the manual assessment step is absolutely necessary (however, as little as about 7.5 per cent of entries are removed), it is very productive, in that it yields an enormous number of collocations. Even though the principle should be applicable to other languages, the method itself cannot

be easily transferred without significant modifications, especially for inflection languages.

Automation

A key feature of the developed methods is the fact that they do not require any programming skills. Probably the most important qualitative point is that lexical and grammar attributes rather than purely mathematical methods are employed (the text is analysed linguistically rather than statistically). Therefore, linguistic methods used instead of mathematical formulae enable the retrieval of phrases found in a text or a collection of texts once only.

It appeared to be impossible to completely automate the developed methods without advanced programming skills. Theoretically, and probably also practically, this would be possible, but nevertheless the work within the thesis did not involve any mechanical repetition of non-automated operations. This includes the analysis of source texts. The final human analysis needed in the methods (especially in that based on plural endings) is a notable exception which, however, seems to be necessary as long as humans surpass computers in certain aspects of linguistic studies.

Principally owing to the automation the three proposed methods provided an extensively rich, interesting and reliable material for any possible further studies. The results prove that it is possible to define discriminants which can be successfully used for the retrieval of neologisms, collocations based on acronyms and noun collocations based on plural endings.

As discussed at length in the summary for each method, it is possible to use the methods for other languages as well. This should be the most convenient for the method for collocation retrieval based on acronyms and for the new word retrieval method. The possibility points to the versatility and flexibility of the methods developed, while taking into consideration certain obvious limitations, discussed after each method, which could not be avoided.

Conclusion

As expected, the complete collection of texts from the *Nature* weekly of approximately 45 million words proved to be a very rich source of lexicographic data and made it possible to test the efficiency of the methods thoroughly.

The new word retrieval method yielded more than 1000 new words, not recorded in dictionaries, the method for collocation retrieval based on acronyms yielded more than 6000 results and the method for collocation retrieval based on plural endings yielded more than 110,000 retrieved results.

It is hoped that the methods developed and tested will have their contribution, minor at least, to the enrichment of lexicographic knowledge and that they will be used in future studies.

6 Bibliografia

Aarts, J., Haan, P. de, Oostdijk, N. (red.) (1993). *English Language Corpora: Design*, *Analysis and Exploitation.* Amsterdam: Rodopi.

Aarts, J., Meijs, W. (red.) (1984). *Corpus linguistics: Recent developments in the use of computer corpora in English language research*. Amsterdam: Rodopi.

Aarts, J., van den Heuvel, T. (1982). Grammars and intuitions in corpus linguistics, w: Johansson, S. (red.) (1982), 66 – 84.

Abercrombie D. (1965). *Studies in phonetics and linguistics*. London: Oxford University Press.

Aijmer, K., Altenberg, B. (red.) (1991). *English Corpus Linguistics*. Studies in Honour of Jan Svartvik. London/New York: Longman.

Aijmer, K., Altenberg, B. (red.) (1993). *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London: Longman.

Algeo, J. (2006). *British or American English?* Cambridge: Cambridge University Press.

Allerton, D. J. (1984). Three or four levels of co-occurrence relations, *Lingua*, 63, 17 – 40.

Aston, G., Burnard, L. (1998). *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Bańko, M. (2001). *Z pogranicza leksykografii i językoznawstwa*. Warszawa: Wydawnictwo Wydziału Polonistyki UW.

Barlow, M. (1992). Using Concordance Software in Language Teaching and Research, w: Shinjo, W. et al. (red.) (1992), 365 – 373.

Barlow, M. (1995). ParaConc: a Concordancer for parallel texts, *Computer and Text*, 10, 14 – 16. Oxford: OPU.

Bellugi, U., Brown, R. (red.) (1964). *The Acquisition of Language*. Monographs of the Society for Research in Child Development 29.

Bień, J. S., Szafran, K. (2001). Analiza morfologiczna języka polskiego w praktyce, *Biuletyn Polskiego Towarzystwa Językoznawczego*, LVII, 171 – 184.

Blackwell, S. (1993). From dirty data to clean language, w: Aarts, J. et al. (red.) (1993): 97 – 106.

Bogusławski, A. (2010). *Dwa studia z teorii fleksji (i inne przyczynki)*. Warszawa: BEL Studio Sp. z o.o.

Bogusławski, A., Danielewiczowa, M. (2005). *Verba polona abscondita. Sonda słownikowa III*. Warszawa: Elma Books.

Brants, T. (2000). Tnt – a statistical part-of-speech tagger, w: *Proceedings of the 6th Applied Natural Language Processing Conference*, Seattle, WA.

Brin, S., Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. http://infolab.stanford.edu/~backrub/google.html (data pobrania 20 maja 2010).

Brill, E. (1992). A simple rule-based part-of-speech tagger, w: *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing*, 152 – 155, Trento, IT.

Buczyński, A. (2004). *Pozyskiwanie z Internetu tekstów do badań lingwistycznych*. Niepublikowana praca magisterska pod kierunkiem Janusza S. Bienia. Warszawa: Instytut Informatyki UW.

Burnage, G., Dunlop, D. (1993). Encoding the British National Corpus w: Aarts, J. et al. (red.) (1992), 79 – 95.

Burnard, L. (1988). The Oxford Text Archive: principles and prospects, w: Genet, J.P. (red.) (1988), 191 – 203.

Burnard, L. (1996). Introducing SARA: exploring the BNC with SARA (Paper presented at TALC 96).

Busemann, S. (red.) (2002). *KONVENS 2002*, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Saarbrücken.

Buttler, D. (1962). Neologizm i terminy pokrewne. *Poradnik Językowy*, 5-6: 235 – 244.

Buttler, D. (1993). Neologizmy z formantem -acja w powojennej polszczyźnie. *Przegląd Filologiczny*, 38, 7 – 15.

Bygate, M., Tonkyn, A., Williams, E. (red.) (1994). *Grammar and the Language Teacher*. Hemel Hempstead: Prentice Hall.

Carterette, E.C., Jones, M.H. (1974). *Informal Speech*. Berkeley: University of California Press.

Chafe W. (1988). Punctuation and the prosody of written language. *Written Communication*, 5, 396 – 426.

Chalker, S. (1992). Pedagogical Grammar: Principles and Problems. w: Bygate, M. et al. (red.) (1992), 31 – 44.

Charniak, E., Hendrickson, C., Jacobson, N., Perkowitz, M. (1993). Equations for part-of-speech tagging, w: *National Conference on Artificial Intelligence*, 784 – 789.

Chlebda, W. (1991). *Elementy frazematyki. Wprowadzenie do frazeologii nadawcy*. Opole: WSP.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. (1964). *Current Issues in Linguistic Theory*. The Hague: Mouton.

Chomsky, N. (1964). Formal discussion, w: Bellugi, U., Brown, R. (red.) (1964), 37 – 39.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (2004). (Interviewed by Andor, Jozsef). The master and his performance: An interview with Noam Chomsky. *Intercultural Pragmatics.* 1(1), 93 – 111.

Choueka, Y. (1988). Looking for needles in a haystack, w: *Proceedings*, *RIAO Conference on User-Oriented Context Based Text and Image Handling*, 609 – 623. Cambridge, MA.

Clark, A., Fox, Ch., Lappin, Sh. (red). (2010). *The Handbook of Computational Linguistics and Natural Language Processing*. Oxford: Blackwell Publishing.

Clopper, C.G., Pisoni, D.B. (2006). The Nationwide Speech Project: A new corpus of American English dialects, *Speech Communication* 48, 6: 633 – 644.

Coniam, D. (2004). Concordancing oneself: Constructing individual textual profiles, *International Journal of Corpus Linguistics* 9, 2: 271 – 298.

Cowie, A.P. (red.) (1998). *Phraseology. Theory*, *Analysis*, *and Applications*. Oxford: Oxford University Press.

Crowdy, S. (1995). The BNC spoken corpus, w: Leech, G. et al. (red.) (1995), 224 – 235.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press.

Cutting, D., Kupiec, J., Pedersen, J., Sibun, P. (1992). A practical part-of-speech tagger, w: *Third Conference on Applied Natural Language Processing* (ANLP-92), 133 – 140.

Čermák F. (2001). Substance of idioms: Perennial problems, lack of data or theory? *International Journal of Lexicography*, 14.1: 1 – 20.

Damerau, F.J. (1993). Generating and Evaluating Domain-Oriented Multi-Word Terms from Texts, *Information Processing & Management* vol. 29, No. 4, 433 – 447.

Daudaravičius, V., Marcinkevičienė, R. (2004). *Gravity Counts for the Boundaries of Collocations*, International Journal of Corpus Linguistics, 2004, No. 9-2, 321 – 348.

Dias, G. et al. (2000). Normalization of Association Measures for Multiword Lexical Unit Extraction, *International Conference on Artificial and Computational Intelligence for Decision Control and Automation in Engineering and Industrial Applications (ACIDCA'2000).* Monastir, Tunisia, 207 – 216.

Doroszewski, W. (red.) (1958 – 1969). *Słownik języka polskiego*. Warszawa: Wiedza Powszechna.

Dubisz, S. (red.) (2003). *Uniwersalny słownik języka polskiego.* Warszawa: Wydawnictwo Naukowe PWN.

Dzierżanowska, H. (1990). *Przekład tekstów nieliterackich na przykładzie języka angielskiego*. Warszawa: Państwowe Wydawnictwo Naukowe.

Edwards, J.A. (1993). Survey of electronic corpora and related resources for language researchers, w: Edwards, J.A., Lampert, M.D. (red.), 263 – 310.

Edwards, J.A., Lampert, M.D. (red.) (1993). *Talking Data: Transcription and Coding in Discourse Research*. Hillsdale: Lawrence Erlbaum.

Farrow, J. F. (1996). Alexander Cruden and his concordance. *Indexer*, 20, Apr. 1996, 55 – 56.

Feilke, H. (1996). *Sprache als soziale Gestalt. Ausdruck*, *Prägung und die Ordnung der sprachlichen Typik*. Frankfurt/Main: Suhrkamp.

Feilke, H. (2004). Kontext - Zeichen - Kompetenz. Wortverbindungen unter sprachtheoretischem Aspekt, w: Steyer, K. (red.), 41 – 64.

Fillmore, C. (1992). 'Corpus linguistics' or 'Computer-aided armchair linguistics', w: Svartvik, J. (red.) (1992), 35 – 60.

Fillmore, C., Ide, N., Jurafsky, D., Macleod, C. (1998). An American National Corpus: A Proposal. *Proceedings of the First International Language Resources and Evaluation Conference*, Granada, Spain, 965 – 970.

Francis, W.N., Kučera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Friedl, J. (1998). *Mastering Regular Expressions*. Sebastopol: O'Reilly& Associates.

Garside, R. (1987). The CLAWS Word-tagging System, w: Garside, R. et al. (red.) (1987), 30 – 41.

Garside, R., Leech, G., McEnery, A. (red.). (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London

Garside, R., Leech, G., Sampson, G. (red.) (1987). *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Garside, R., Smith, N. (1997) A hybrid grammatical tagger: CLAWS4, w: Garside, R. et al. (red.) (1997), 102 – 121.

Genet, J.P. (red.). (1988). *Standardisation et Exchange des Bases de Données Historiques*. Paris: Editions du CNRS.

Gitsaki, C., Taylor, R. (1999). English collocations and their place in the EFL classroom. *NUCB Journal of Language Culture and Communication*, 1(4), 83 – 94.

Godby, J. (1994). Two techniques for the identification of phrases in full text. Annual Review of OCLC Research.

Golding, A.R., Schabes, Y. (1996). Combining Trigram-based and Feature-based Methods for Context-Sensitive Spelling Correction. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*. Santa Cruz, CA.

Good, N. (2004). *Regular Expression Recipes: A Problem-Solution Approach*. Berkeley: Apress.

Greenbaum, S. (1991). The development of the International Corpus of English, w: Aijmer, K., Altenberg, B. (red.) (1991), 83 – 91.

Greenbaum, S. (red.) (1996). *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press

Greenbaum, S., Yibin, N. (1996). About the ICE Tagset, w: S. Greenbaum (red.) (1996), 92 – 109.

Greene, B.B., Rubin, G.M. (1971). *Automated Grammatical Tagging of English*. Providence, RI: Department of Computer Science, Brown University.

Gries, S. Th., Stefanowitsch, A. (2004). Extending collostructional analysis. A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*. 9:1, 97 – 129.

Gries, S. Th. (2006). Some proposals towards more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik*, 54(2), 191 – 202.

Grune, D., Jacobs, C.J.H. (1990). *Parsing Techniques – A Practical Guide*. Chichester: Ellis Horwood.

Habibi, M. (2003). *Java Regular Expressions: Taming the java.util.regex Engine*. Berkeley: Apress.

Hajnicz, E., Kupść, A. (2001). Przegląd analizatorów morfologicznych dla języka polskiego. *Prace Instytutu Podstaw Informatyki Polskiej Akademii Nauk.* Warszawa: Instytut Podstaw Informatyki PAN.

Halliday, M.A.K., Teubert, W., Yallop, C., Čermáková, A. (red.) (2004). *Lexicology and Corpus Linguisticc: An Introduction*. New York: Continuum Intl Pub Group.

Hann, M.L. (1975). Towards an Algorithmic Methodology of Lemmatization, *ALLC Bulletin* 3, 140 – 150.

Harris, Z. (1951). *Methods in Structural Linguistics*. Chicago: University of Chicago Press.

Hitchings, H. (2005). *Dr Johnson's Dictionary: The Extraordinary Story of the Book that Defined the World*. Edinburgh: John Murray.

Hocket, C. (1948). A note on structure, *International Journal of American Linguistics* 14, 269 – 271.

Hockey, Susan (1980). *A Guide to Computer Applications in the Humanities*, Baltimore: The Johns Hopkins University Press.

Hoey, M. (2000). A world beyond collocation: new perspectives on vocabulary teaching, w: Lewis, M. (red.) (2000), 224 – 243.

Hoover, D. (2002). Frequent Word Sequences and Statistical Stylistics. *Literary and Linguistic Computing* vol. 17, no. 2 (2002), 157 – 180.

Horrocks, G. (1987). *Generative Grammar*. London: Longman.

http://eur-lex.europa.eu/pl/index.htm (data pobrania: 28 marca 2007)

http://www.collins.co.uk/books.aspx?group=153 (data pobrania: 2 września 2007)

Hunston, S. (2002). *Corpora and Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S., Francis, G. (2000). *Pattern Grammar – A corpus-driven approach to the lexical grammar of English*. Amsterdam/ Philadelphia: John Benjamins.

Hunston, S., Gill, F. (2000). Pattern Grammar. A Corpus-Driven Approach to the Lexical Grammar of English, *Studies in Corpus Linguistics*. Vol. 4. Amsterdam/Philadelphia: John Benjamins.

Hunston, S., Gill, F. (2000). Pattern Grammar. A Corpus-Driven Approach to the Lexical Grammar of English, *Studies in Corpus Linguistics*. Vol. 4. Amsterdam/Philadelphia: John Benjamins.

Hussmann, S., Deng, P.W.P. (2005). A high-speed optical mark reader hardware implementation at low cost using programmable logic, *Real Time Imaging* (11), 1, 19 – 30.

Ide, N., Macleod, C. (2001). The American National Corpus: A Standardized Resource of American English, *Proceedings of Corpus Linguistics 2001*, Lancaster UK.

Ide, N., Suderman, K. (2004). The American National Corpus First Release, *Proceedings of the Fourth Language Resources and Evaluation Conference* (LREC), Lisbon, Portugal, 1681 – 1684.

Jacquemin, C., Bourigault, D. (2003). Term extraction and automatic indexing, w: Mitkov, R. (red.) (2003), 599 – 615.

Johansson, C. (1996). Good Bigrams, w: *Proceedings from the 16th International Conference on Computational Linguistics (COLING-96)*. Copenhagen, 592 – 597.

Johansson, S. (1991). Times change and so do corpora, w: Aijmer, K., Altenberg, B. (red.) (1991), 305 – 314.

Johansson, S. (red.) (1982). *Computer corpora in English language research*. Bergen: Norwegian Computing Centre for Humanities.

Johansson, S., Atwell, E., Garside, R., Leech, G. (1986). *The Tagged LOB Corpus: Users' manual*. Bergen: ICAME, The Norwegian Computing Centre for the Humanities, Bergen University, Norway.

Johansson, S., Stenström, A.-B. (red.) (1991). *English computer corpora. Selected papers and research guide.* Berlin: Mouton de Gruyter.

Johnson, M. (2006). *Essential Python for Corpus Linguistics*. Oxford: Blackwell Publishing.

Johnson, S. (1747). *The Plan of an English Dictionary*. Ed. by Lynch J. New Brunswick: Rutgers University.

Jung, Y., Choi, Y., Myaeng, S.-H. (2007). Determining Mood for a Blog by Combining Multiple Sources of Evidence, *IEEE/WIC/ACM International Conference on Web Intelligence* (WI'07), 271 – 274.

Kachru, B. B. (1992). *The Other Tongue: English across cultures*. Urbana-Champaign: University of Illinois Press.

Kaeding, F.W. (Hrsg.) (1897/98). *Häufigkeitswörterbuch der deutschen Sprache 1*, *2*. Selbstverlag, Berlin-Steglitz.

Karlsson, F. (1995). The formalism and environment of Constraint Grammar Parsing, w: Karlsson, F., Voutilainen, A., Heikkilä, J., Anttila, A. (red.) (1995), 41 – 88.

Karlsson, F., Voutilainen, A., Heikkilä, J., Anttila, A. (red.) (1995). *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Berlin: Mouton de Gruyter.

Karttunen, L., Chanod, J.-P., Grefenstette, G., Schiller, A. (1996). Regular Expressions for Language Engineering, *CUP Journals: Natural Language Engineering* 2 (4), 305 – 328.

Kaye, G. (1988). The design of the database for the Survey of English Usage, w: Kytö, M. et al. (red.) (1988), 145 – 168.

Keay, J. (2005). *Alexander the Corrector: the tormented genius whose Cruden's concordance unwrote the bible*. Woodstock: Overlook Press.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London: Longman.

Kermes, H., Heid, U. (2003). Using chunked corpora for the acquisition of collocations and idiomatic expressions, *Proceedings of the 7th International Conference on Computational Lexicography*, *COMPLEX'03*, *Budapest*, *April 2003*.

Kilgarriff, A., Rundell, M. (2002). Lexical Profiling Software and its Lexicographic Applications – a Case Study, w: Braasch et al. (red.) (2002), 807 – 818.

Kilgarriff, A., Tugwell, D. (2001). Word sketch: Extraction and display of significant collocations for lexicography, *Proceedings of the 39th ACL and 10th EACL workshop "Collocation: computational extraction*, *analysis and exploitation"*, Tolouse, 32 – 38.

Kis, A., Kis, B. (2003). A prescriptive corpus-based technical dictionary. development of a multi-purpose technical dictionary, *Papers in Computational Lexicography: Proceedings of COMPLEX 2003*, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, 47 – 56.

Kis, B., Villada Moiron, M.B., Bouma, G, Ugray, G., Biro, T., Pohl, G., Nerbonne, J. (2004). A New Approach to the Corpus-based Statistical Investigation of Hungarian Multi-Word Lexemes, *Proceedings of the 4th International Conference on Language Resources and Evaluation*, *European Language Resource Association*, Lisbon, 2004, 1677 – 1681.

Kjellmer, G. (1994). *A Dictionary of English Collocations*. Oxford: Clarendon Press.

Klein, S., Simmons, R.F. (1963). A computational approach to grammatical coding of English words, *Journal of the Association for Computing Machinery*. 10, 334 – 347.

Kłopotek, M., Wierzchoń, S., Trojanowski, K. (red.) (2006). *Intelligent Information Processing and Web Mining, IIS:IIPWM'06 Proceedings*. Berlin/Heidelberg: Springer-Verlag.

Knowles, G., Taylor, L., Williams, B. (1992). *A Corpus of Formal British English Speech*. London: Longman.

Kocourek, R. (1991). *La langue française de la technique et de la science. Vers une linguistique de la langue savante*. Wiesbaden: Oscar Brandstetter Verlag.

Kordek, N., K. Stroński (red.) 1998. *Artis linguisticae paululum. Libellum professori Georgio Bańczerowski sexagenario oblatum*. Poznań: Instytut Językoznawstwa.

Kosek, I. (2008). *Fleksja i składnia nieciągłych imiennych jednostek leksykalnych*. Olsztyn: Uniwersytet Warmińsko-Mazurski.

Krenn, B., Evert, S. (2001). Can we do better than frequency? A case study on extracting pp-verb collocations, *Proceedings of the ACL Workshop on Collocations*, Toulouse, France, 39 – 46.

Krzemińska, W., Nowak, P. (red.). (2002). *Przestrzenie informacji*. Poznań: Sorus.

Kublik, A. (2006). A4 będzie bezpłatna do 2010 roku, *Gazeta Wyborcza*, 17 listopada 2006, 27.

Kučera, H., Francis, W.N. (1967). *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Kurcz, I., Lewicki, A., Sambor, J., Szafran, K., Woronczak, J. (red. Zygmunt Saloni) (1990). *Słownik frekwencyjny polszczyzny współczesnej*. Kraków: Polska Akademia Nauk, Instytut Języka Polskiego.

Kurcz, I., Lewicki, A., Sambor, J., Woronczak, J. (1974 – 1977). *Słownictwo współczesnego języka polskiego. Listy frekwencyjne*. Warszawa: PAN.

Kytö, M. (1994). *Manual to the Diachronic Part of the Helsinki Corpus of English Texts*: *Coding Conventions and Lists of Source Texts*. Helsinki: Helsinki University Press for Department of English, University of Helsinki.

Kytö, M., Ihalainen O., Rissanen, M. (1988). *Corpus Linguistics*, *hard and soft.* Amsterdam: Rodopi.

Kytö, M., Ihalainen, O., Rissanen, M. (red.) (1988). *Corpus Linguistics: Hard and Soft*. *Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora*. Amsterdam: Rodopi.

Kytö, M., Rissanen, M. (1988). The Helsinki Corpus of English Texts. Classifying and coding the diachronic part, w: Kytö, M., Ihalainen, O., Rissanen, M. (red.) (1988), 169 – 180.

Lamel, L.F., Gauvain, J.-L., Eskenazi, M. (1991). BREF, a Large Vocabulary Spoken Corpus for French, w: Eurospeech 91. 2nd European Conference on Speech Communication and Technology. Genova, Italy, 24-26 September 1991. vol. 2, 505 – 508.

Landau, S. (2005). Johnson's Influence on Webster and Worcester in Early American Lexicography, *International Journal of Lexicography* 18 (2), 217 – 229.

Lebert, M. (2005). Project Gutenberg, from 1971 to 2005, publikacja prezentowana na Third International Colloquium on ICT-enhanced French Studies: Dialogues across languages and cultures.

Leech, G. (1991). The state of the art in corpus linguistics, w: Aijmer, K., Altenberg, B. (red.) (1991), 8 – 29.

Leech, G. (1992). Corpora and theories of linguistic performance, w: Svartvik, J. (red.) (1992), 105 – 122.

Leech, G. (1993). Corpus annotation schemes, *Literary and Linguistic Computing* 8(4): 275 – 281.

Leech, G., Garside, R., Bryant, M. (1994). CLAWS4: The tagging of the British National Corpus. Proceedings of the 15th International Conference on Computational Linguistics (COLING 94) Kyoto, Japan, 622 – 628.

Leech, G., Garside, R., Bryant, M. (1994). The large-scale grammatical tagging of text: experience with the British National Corpus, w: Oostdijk, N., Haan, P. (red.) (1994), 47 – 63.

Leech, G., Myers, G., Thomas, J. (red.) (1995). *Spoken English on computer: transcription*, *mark-up and application*. Harlow: Longman.

Leech, G., Rayson, P., Wilson, P. (2001). *Word Frequencies in Written and Spoken English: based on the British National Corpus*. London: Longman.

Levická, J., Garabík, R. (red.) (2009). *NLP, Corpus Linguistics, Corpus Based Grammar Research: Proceedings of the Fifth International Conference*, Smolenice, Slovakia, 25 – 27 November 2009, Slovko 2009.

Lewandowska-Tomaszczyk, B. (red.) (2005). *Podstawy językoznawstwa korpusowego*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Lewandowska-Tomaszczyk, B., Oakes, M., Schulte, F. (2001). *Foreign Language Teaching and Information and Communication Technology*. Frankfurt am Main: Peter Lang.

Lewandowska-Tomaszczyk, B., Osborne, J., Schulte, F. (2001). *Foreign Language Teaching and Information and Communication Technology*. Hamburg: Peter Lang,

Lewis, M. (red.) (2000). *Teaching Collocation. Further Developments in the Lexical Approach*. Hove: LTP.

Lin, D. (1998). Extracting Collocations from Text Corpora, *First Workshop on Computational Terminology*, Montreal, 57 – 63.

Loberger, G., Welsh, K. S. (2001). *Webster's New World English Grammar Handbook*. New York: Wiley.

Mallinson, G., Blake, B. J. (1981). *Language Typology: Cross-linguistic Studies in Syntax*. Amsterdam: North-Holland.

Manning, C. (1993). Automatic acquisition of a large subcategorization dictionary from corpora, w: *Proceedings of the 31th Annual Meeting of the Association of Computational Linguistics*, 235 – 242.

Manning, C.D., Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Matsumoto, Y., Asahara, M., Kawabe, K., Takahashi, Y., Tono, Y., Ohtani, A., Morita, T., (2005). ChaKi: An Annotated Corpora Management and Search System, w: *Proceedings from the Corpus Linguistics Conference Series*, Vol.1, no. 1, July 2005 (http://www.corpus.bham.ac.uk/PCLC).

Matthews, P. (1981). *Syntax*. Cambridge: Cambridge University Press.

Mays, E., Damerau, F.J., Mercer, R.L. (1991). Context based spelling correction, *Information Processing and Management*. 27 (5), 517 – 522.

McEnery, A.M., Wilson, A., Sanchez-Leon, F., Nieto-Serano, A. (1997). Multilingual Resources for European Languages: Contributions of the Crater Project,*Literary and Linguistic Computing*, 12 (4), 219 – 226.

McEnery, T., Wilson, A. (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T. (2003). Corpus linguistics, w: Mitkov, R. (red.) (2003), 448 – 462.

Mel'čuk, I. (1998). Collocations and Lexical Functions, w: Cowie, A.P. (red.) (1998), 23 – 53.

Merkel, M., Andersson, M. (2000). Knowledge-lite extraction of multi-word units with language filters and entropy thresholds, w: *Proceedings of RIAO'2000*. Collége de France, Paris, France, April 12-14, 2000, Volume 1, 737 – 746.

Merkel, M., Nilsson, B., Ahrenberg, L. (1994). A Phrase-Retrieval System Based on Recurrence, w: *Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2)*. Kyoto, 99 – 108.

Meyer, C.F. (2002). *English Corpus Linguistics. An Introduction*. Cambridge: Cambridge University Press.

Mikheev, A. (2003). Text segmentation, w: Mitkov, R. (red.) (2003), 201 – 218.

Mitkov, R. (red.) (2003). *Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press.

Moon, R. (1998). *Fixed expressions and idioms in English: a corpus-based approach.* Oxford: Clarendon Press.

Moszczyński, R. (2006). *Formal approaches to multiword lexemes*. Warszawa: Instytut Anglistyki UW.

Mugglestone, L. (2005). *Lost for Words: The Hidden History of the Oxford English Dictionary*. Yale: Yale University Press.

Nakano, Y. (2000). *Needs and Seeds in Character Recognition*, ICPR'00 (Vol. IV), 140 – 146.

Nunberg, G., Sag, I.A., Wasow, T. (1994). Idioms. *Language*, 70(3), 491 – 538.

O'Grady, W., Dobrovolsky, M., Katamba, F. (red.) (1997). *Contemporary linguistics: An Introduction*. London and New York: Longman.

Oliva, K., Květoň, P. (2002). (German) Corpus Representativity, Bigrams, and PoS-Tagging Quality, w: Busemann, S. (red.) (2002), 131 – 138.

Oostdijk, N. (1988). A corpus linguistic approach to linguistic variation. *Literary and Linguistic Computing* 3, 12 – 25.

Oostdijk, N., Haan, P. de (red.) (1994). *Corpus-Based Research into Language: In Honour of Jan Aaarts*. Amsterdam: Rodopi.

Otwinowska – Kasztelanic, A. (2000). *Korpus języka mówionego młodego pokolenia Polaków (19-35 lat)*. Warszawa: Wydawnictwo Akademickie DIALOG.

Padró, L. (1998). *A Hybrid Environment for Syntax-Semantic Tagging*. Rozprawa doktorska. Barcelona: Universitat Politècnica de Catalunya.

Partington, A. (1998). *Patterns and Meanings. Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins Publishing Company.

Pearce D. (2001). Synonymy in Collocation Extraction. *WordNet and Other Lexical Resources: Applications*, *Extensions and Customizations (NAACL 2001 Workshop)*, 41 – 46.

Pedersen, J. (1995). The Identification and Selection of Collocations in Technical Dictionaries. *Lexicographia* 11: 60 – 73.

Pedersen, T. (1996). Fishing for exactness, *Proceedings of the South-Central SAS Users Group Conference*, Austin, Texas, 188 – 200.

Peters, P. (2004). *The Cambridge Guide to English Usage*. Cambridge: Cambridge University Press.

Pluciński, J. (2000). *Makropolecenia w Word 2000 PL*. Warszawa: Wydawnictwo Mikom.

Poos, D., Simpson, R. (2002). Cross-disciplinary comparisons of hedging: Some findings from the Michigan Corpus of Academic Spoken English, w: Reppen et al. (red.) (2002), 3 – 23.

Porter, M.F. (1980). An Algorithm for Suffix Stripping, *Program* 14, 130 – 137.

Przepiórkowski, A. (2004). *Korpus IPI PAN – wersja wstępna*. Warszawa: Instytut Podstaw Informatyki PAN.

Przepiórkowski, A., Górski, R., Łaziński, M., Pęzik, P. (2009). Recent Developments in the National Corpus of Polish, w: Levická et al. (red.) (2009), 302 – 309.

Puppel, S. (red.). (2005). *Scripta Neophilologica Posnaniensa*. Tom VII. Poznań: Wydział Neofilologii UAM.

Quirk, R. (1960). Towards a description of English usage, *Transactions of the Philological Society*, 40 – 61.

Quirk, R. (1968). The Survey of English Usage, Chapter 7 of *Essays on the English Language: Medieval and Modern*. London: Longman.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Rabiega-Wiśniewska, J., Rudolf, M. (2002). AMOR – program automatycznej analizy fleksyjnej tekstu polskiego, *Biuletyn Polskiego Towarzystwa Językoznawczego* LVIII, 175 – 185.

Reddick, A. (1990). *The Making of Johnson's Dictionary 1746 – 1773*. Cambridge, Cambridge University Press.

Renouf, A. (1987). Corpus development, w: Sinclair, J. (red.) (1987), 1 – 40.

Reppen, R., Fitzmaurice, S., D. Biber (red.) (2002). *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins.

Rissanen, M., Kytö, M., Palander-Collin, M. (red.) (1993). *Early English in the Computer Age*. Berlin: Mouton de Gruyter.

Rożyński, P. (2006). Czym TP SA zaskoczy swoich Klientów, Gazeta Wyborcza, 17 listopada 2006, 29.

Saloni, Z., Wołosz, R. (2001). Nowa errata do *Indeksu a tergo* do *Słownika języka polskiego* pod redakcją Witolda Doroszewskiego, *Prace Filologiczne* XLVI, 503 – 534.

Sampson, G. (1992). Probabilistic parsing, w: Svartvik, J. (red.) (1992), 425 – 447.

Sampson, G. (2005). Quantifying the shift towards empirical methods. *International Journal of Corpus Linguistics*. 2005, Vol. 10 Issue 1, 15 – 36.

Say, B., Akman, V. (1997). Current approaches to punctuation in Computational Linguistics. *Computers and the Humanities*, 30, 457 – 469.

Schafroth, E. (2003). Kollokationen im GWDS, w: Wiegand, H. E. (red.) (2003), 397 – 412.

Schmied, J. (1993). Qualitative and quantitative research approaches to English relative constructions, w: Souter, C., Atwell, E. (red.) (1993), 85 – 93.

Scott, M. (1998). WordSmith Tools Manual, ver. 3.0, Oxford: Oxford University Press.

Seretan, V., Nerima, L. Wehrli, E. (2004). A tool for multi-word collocation extraction and visualization in multilingual corpora, *Proceedings of the Eleventh EURALEX International Congress (EURALEX 2004)*, Lorient, France, 755 – 766.

Seymore, K. (1999). Learning Hidden Markov Model Structure for Information Extraction, *AAAI 99 Workshop on Machine Learning for Information Extraction*.

Sharoff, S. (2003). Methods and tools for development of the Russian Reference Corpus, w: Wilson, A. et al. (red.) (2003), 167 – 180.

Shastri, S. V. (1985). A computer corpus of present-day Indian English: A preliminary report, *International Computer Archive of Modern and Medieval English Journal*, 9, 9 – 10.

Shastri, S. V. (1988). The Kolhapur Corpus of Indian English and work done on its basis so far, *ICAME Journal*, 12, 15 – 26.

Shimohata, S., Sugio, T., Nagata, J. (1997). Retrieving Collocations by Co-occurrences and Word Order Constraints, *Proceeding of the 35th Conference of the Association for Computational Linguistics (ACL'97)*, Madrid, 476 – 481.

Shinjo, I., Landhl, K., Macdonald, M., Noda, K., Ozeki, S., Shiozawa, T. Sugiura, M. (red.) (1992). *Proceedings of the Second International Conference on Foreign Language Education and Technology*. Kasugai, Japan: Language Laboratory Association of Japan and International Association of Learning Laboratories.

Siemens, R.G. (1996). Lemmatization and Parsing with *TACT* Preprocessing Programs, *Computing in the Humanities Working Papers*, A.1.

Siepmann, D. (2005). Collocation, Colligation and Encoding Dictionaries. Part I: Lexicological Aspects, *International Journal of Lexicography* 18 (4), 409 – 443.

Siepmann, D. (2006). Collocation, Colligation and Encoding Dictionaries. Part II: Lexicographical Aspects, *International Journal of Lexicography* 19 (1), 1 – 39.

Silva, P. (2005). Johnson and the OED, *International Journal of Lexicography*, 18 (2), 231 – 242.

Sinclair, J. (1991). *Corpus*, *concordance*, *collocation*. Oxford: Oxford University Press.

Sinclair, J. (1992). The automatic analysis of corpora, w: Svartvik, J. (red.) (1992), 379 – 397.

Sinclair, J. (1995). *Collins Cobuild English Language Dictionary*. London: Harper Collins.

Sinclair, J. (1996). *Eagles*, *Preliminary recommendations on Corpus Typology*, Expert Advisory Group on Language Engineering Standards Guidelines EAG-TCWG-CTYP/P.

Sinclair, J. (red.) (1987). *Looking Up. An Account of the Cobuild Project in Lexical Computing*. London: Collins ELT.

Sinclair, J. M. (2004). *Trust the Text. Language*, *Corpus and Discourse*. New York/London: Routledge.

Smadja, F. A. (1993). Retrieving collocations from text: Xtract, *Computational Linguistics*, 19(1), 143 – 177.

Smółkowa, T. (2001). *Neologizmy we współczesnej leksyce polskiej*. Kraków: IJP PAN.

Snarska, A. (2007). *Makropolecenia w Excelu. Nowe wydanie. Opis języka VBA na przykładach*. Warszawa: Wydawnictwo Naukowe PWN.

Sokirko, A. (2004). Морфологические модули на сайте www.aot.ru – *Международная конференция ДИАЛОГ'2004 Компьютерная лингвистика и интеллектуальные технологии*.

Souter, C., Atwell, E. (red.) (1993). *Corpus Based Computational Linguistics*. Amsterdam: Rodopi.

Sperberg-McQueen, C.M., Burnard, L. (red.) (1999). *Guidelines for Electronic Text Encoding and Interchange*. TEI P3 Text Encoding Initiative. Oxford: Oxford University Press.

Steyer, K. (red.) (2003). *Wortverbindungen – mehr oder weniger fest. (Jahrbuch de Instituts für deustche Sprache)*. Berlin: De Gruyter.

Stoberski, Z. (1976). O centralną rejestrację neologizmów naukowych. *Poradnik Językowy*, 4, 186 – 189.

Summers, D. (1991). *Longman/Lancaster Language Corpus: Criteria and Design.* Harlow: Longman.

Svartvik, J. (red.) (1990). *The London Corpus of Spoken English: Description and Research.* Lund Studies in English 82. Lund: Lund University Press.

Svartvik, J. (red.) (1992). Directions in Corpus Linguistics, *Proceedings of Nobel Symposium 82*, Stockholm, 4 – 8 August 1991. Berlin: Mouton de Gruyter.

Svartvik, J., Quirk, R. (1980). *A Corpus of English Conversation*. Lund: C.W.K. Gleerup.

Taghva, K., Stofsky, E. (2001). OCRSpell: An Interactive Spelling Correction System for OCR Errors in Text, *International Journal of Document Analysis and Recognition*, (3), 3, 125 – 137.

Tapanainen, P., Järvinen, T. (1998). Dependency concordances. *International Journal of Lexicography*, 11 (3): 187 – 203.

Taylor, Ch. (2008). What is *corpus linguistics*? What the data says. *ICAME Journal*, 32, 143 – 164.

Teubert, W. (2004). Language and corpus linguistics, w: Halliday, M.A.K., Teubert, W., Yallop, C., Čermáková, A. (red.) (2004), 73 – 112.

Teubert, W., Čermáková, A. (2004). Directions in Corpus Linguistics, w: Halliday, M.A.K., Teubert, W., Yallop, C., Čermáková, A. (red.) (2004), 113 – 165.

Teubert, W., Krishnamurthy, R. (2007). General introduction, w: Teubert, W., Krishnamurthy, R. (red.) (2007), 1–37.

Teubert, W., Krishnamurthy, R. (2007). *Corpus linguistics. Critical concepts in linguistics*. London and New York: Routledge.

Tokarski, J. (1993). *Schematyczny indeks* a tergo *polskich form wyrazowych (opracowanie i redakcja Zygmunt Saloni)*. Warszawa: Państwowe Wydawnictwo Naukowe.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.

TOSCA Research Group (1997). TOSCA tagger vs. 1.1.

Trudgill, P., Hannah, J. (2002). *International English: A Guide to the Varieties of Standard English*, 4th ed. London: Arnold.

Tzoukerman, E., Klavans, J., Strzalkowski, T. (2003). Information retrieval, w: Mitkov, R. (red.) (2003), 529 – 544.

Van Valin, R. D., Jr. (2001). *An introduction to syntax*. Cambridge: Cambridge University Press.

Voutilainen, A. (2003). Part-of-speech tagging, w: Mitkov, R. (red.) (2003), 219 – 232.

Waszakowa, K. (2005). *Przejawy internacjonalizacji w słowotwórstwie współczesnej polszczyzny*. Warszawa: Wydawnictwa UW.

Wawrzyńczyk, J. (1994). *Tak zwane nowe słownictwo polskie w świetle dokumentacji „Polskiego Informatorium Wyrazowego"*. Katowice: Śląsk.

Wawrzyńczyk, J. (1999). *Nowe słownictwo polskie. Fikcje i fakty*. Warszawa: UW.

Wawrzyńczyk, J. (2000). *Słownik bibliograficzny języka polskiego: wersja przedelektroniczna. T. 1*, *A – Ć*. Warszawa: Uniwersytet Warszawski. Instytut Informacji Naukowej i Studiów Bibliologicznych.

Weeber, M., Vos, R., Baayen, H. (2000). Extracting the lowest-frequency words: Pitfalls and possibilities, *Computational Linguistics*, 26(3), 301 – 317.

Whaley, L. J. (1997). *Introduction to Typology: The Unity and Diversity of Language*. Newbury Park: Sage.

Widdows, D., Dorow, B. (2002). A graph model for unsupervised lexical acquisition, *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan, 1093 – 1099.

Widdows, D., Dorow, B. (2005). Automatic Extraction of Idioms using Graph Analysis and Asymmetric Lexicosyntactic Patterns, *Proceedings of the*

ACL-SIGLEX Workshop on Deep Lexical Acquisition Ann Arbor, Michigan, 48 – 56.

Wiegand, H. E. (red.) (2003). *Untersuchungen zur kommerziellen Lexikographie der deutschen Gegenwartssprache I. „Duden. Das große Wörterbuch der deutschen Sprache in zehn Bänden"*. Tübingen: Max Niemeyer Verlag.

Wierzchoń, P. (1998). Wyraz (najmniejsza jednostka sporna), w: Kordek, N., Stroński, K. (red.) (1998), 127 – 170.

Wierzchoń, P. (2002). Automatyzacja ekscerpcji definiowanych połączeń wyrazowych. Filtry wyrażeń regularnych, w: Krzemińska, W., Nowak, P. (red.) (2002), 119 – 184.

Wierzchoń, P. (2003). *Z cudzysłowów do poczekalni leksykograficznej*. Warszawa: Katedra Literatury i Kultury Rosyjskiej Uniwersytetu Łódzkiego.

Wierzchoń, P. (2004). *Gramatyka diakrytologiczna. Studium ortograficzno-kwantytatywne*. Poznań: Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Wierzchoń, P. (2005a). *Z cudzysłowów do poczekalni leksykograficznej II*. Warszawa: TAKT.

Wierzchoń, P. (2005b). Automatyczne metody ekscerpcji neologizmów, czyli językoznawstwo faktograficzne, w: Puppel, S. (red.) (2005), 221 – 240.

Wilson, A., Archer, D., Rayson, P. (2003). *Language and Computers*, *Corpus Linguistics Around the World*. Amsterdam: Rodopi.

Wilson, A., Archer, D., Rayson, P., McEnery, T. (red.) (2003). *Proceedings of the Corpus Linguistics 2003 conference*, UCREL, Lancaster University.

Woliński, M. (2006). Morfeusz – a Practical Tool for the Morphological Analysis of Polish, w: Kłopotek, M. et al. (red.) (2006), 503–512.

Wołosz, R. (2005). *Efektywna metoda analizy i syntezy morfologicznej w języku polskim*. Warszawa: Akademicka Oficyna Wydawnicza Exit.

Wright, J. (red.) (1898-1905). *The English Dialect Dictionary*. Oxford: Henry Frowde.

Xunfeng, X., Kawecki, R. (2001). The Use of an On-line Trilingual Corpus for the Teaching of Reading Comprehension in French, 6th TELRI Seminar „Multilingual Corpus Research", Bansko, Bułgaria, 9 – 11.11.2001.

Yamamoto, M., Church, K.W. (1998). Using Suffix Arrays to Compute Term Frequency and Document Frequency for all Substrings in a Corpus, *Proceedings of the Sixth Workshop on Very Large Corpora*, Montreal, 28 – 37.

Zernik, U. (red.) (1991). *Exploiting On-Line Resources to Build a Lexicon*. Hillsdale: Lawrence Erlbaum Associates.

Zgółkowa, H., Bułczyńska, K. (1987). *Słownictwo dzieci w wieku przedszkolnym. Listy frekwencyjne*. Poznań: Wydawnictwo Naukowe UAM.

Zhou, J., P. Dapkus (1995). Automatic Suggestion of Significant Terms for a Predefined Topic, w: *Proceedings of the Third Workshop on Very Large Corpora*. Cambridge, 131 – 147.

Zhu, Q. (1989). A quantitative look at the Guangzhou Petroleum English Corpus, w: *ICAME Journal* 13, 28 – 38.

Zinsmeister, H., Heid, U. (2003). Identifying predicatively used adverbs by means of a Statistical Grammar Model, w: Wilson, A., Archer, D., Rayson, P., McEnery, T. (red.) (2003), 932 – 939.