2021 PRACE JĘZYKOZNAWCZE XXIII/3
ISSN 1509-5304 DOI 10.31648/pj.6838 87–104

Daniel Borysowski
Uniwersytet Opolski
ORCID: https://orcid.org/0000-0001-6594-9047
e-mail: dborysowski@uni.opole.pl

Web crawling dla celów lingwistycznych. Wybrane aspekty gromadzenia i analizy danych tekstowych na przykładzie rosyjskojęzycznych newsów internetowych

Web crawling for linguistic purposes. Selected aspects of collecting and analyzing text data on the example of Russian-language Internet news

Abstrakt
Autor niniejszego artykułu zgromadził ok. 2,7 mln rosyjskojęzycznych newsów internetowych. Zasadnicze cele tego tekstu stanowią: omówienie pojęcia web crawlingu w odniesieniu do pozyskiwania internetowych danych tekstowych, omówienie kwestii strukturyzacji takich danych w nieanotowanych korpusach tekstowych, a także przedstawienie wybranych aspektów analizy danych strukturyzowanych w ten sposób. Autor rozpatruje newsy internetowe jako połączenie tekstu zasadniczego oraz identyfikujących i charakteryzujących go metadanych (wyróżnionych podczas automatycznej ich ekscerpcji ze stron internetowych). Rozdział newsów na tekst zasadniczy i metadane stwarza możliwość przeprowadzenia ich analizy z dwóch perspektyw – tekstowej oraz metainformacyjnej (dodatkowo, np. w odniesieniu do badań chronologizacyjnych, z perspektywy uwzględniającej oba te poziomy). Zarys możliwych badań lingwistycznych zgromadzonego materiału uzupełnia autor ewaluacją wybranych wielowyrazowych całostek, wydobytych z tych tekstów z wykorzystaniem delimitacyjnej funkcji cudzysłowu.

Słowa kluczowe: web crawling, korpus plików tekstowych, news internetowy, ogranicznik tekstu, cudzysłów, reprodukt, związki wielowyrazowe

Abstract
The author of the article collected nearly 2.7 million excerpts of Russian-language Internet news. The main objectives of the article include: discussing the concept of web crawling in relation to the acquisition of online text data, addressing issues related to structuring such data in unannotated text corpora, as well as presenting selected aspects of analyzing data structured this way. The author considers Internet news to be a combination of the main text and metadata that identifies and characterizes it (acquired during automatic extraction from websites). The categorization of news into the main text and metadata creates an opportunity to analyze it from two perspectives – textual and meta-information

(and an additional perspective that combines these two, for example for the purpose of chronological studies). An outline of possible linguistic research into the collected material is supplemented with evaluating selected multi-word tokens extracted from these texts based on the delimitation function of quotation marks.

Key words: web crawling, corpus of text files, Internet news, text delimiter, quote, re-product, multi-word expressions

1. Wprowadzenie: terminologia, założenia, cele

Niniejszy tekst jest z jednej strony sprawozdaniem z realizacji szeregu działań informatyczno-programistycznych, podjętych przez nas w celu zdobycia znacznej liczby naturalnych tekstów rosyjskojęzycznych, z drugiej zaś – prezentacją wyników analizy lingwistycznej przeprowadzonej na tych tekstach. Zgromadzona przez nas kolekcja internetowych danych tekstowych składa się z blisko 2,7 mln newsów internetowych wyzyskanych z czterech wybranych źródeł. W związku z powyższym powstają trzy pytania: jak definiujemy termin „news internetowy"?, jakie źródła newsów wykorzystaliśmy (i dlaczego takie, a nie inne)?, czym w istocie jest zebrana przez nas kolekcja tekstów?

Punktem wyjścia dla charakterystyki pojęcia newsa internetowego jest dla nas definicja tekstu zaproponowana przez Jerzego i Stanisławę Bartmińskich (odwołujemy się do niej z pełną świadomością prowadzonej wśród badaczy dyskusji na temat aktualizacji definicji tekstu w dobie tzw. nowych mediów; zob. Witosz 2016). Można stwierdzić, iż news jako pewien wzorzec tekstowy (w literaturze genologicznej i w dziennikarstwie stawiany obok wzmianki, flesha, nowiny; zob. Kudra 2010) „ma swój podmiot", „intencję", „nacechowanie gatunkowe i stylowe", „poddaje się całościowej interpretacji", „wykazuje [w pewnym stopniu] **integralność strukturalną** oraz spójność semantyczną", podlega podziałowi – zarówno semantycznemu, jak i, w większości wypadków, logicznemu i kompozycyjnemu (Bartmiński, Niebrzegowska-Bartmińska 2009: 36; uzup. i wyróż. – D.B.). Nie w pełni oczywista wydaje się wyróżniona w powyższym cytacie integralność strukturalna newsów internetowych (czy w ogóle tekstów będących w istocie dokumentami hipertekstowymi i skorelowanych z interfejsem konkretnej strony internetowej). W wypadku takich tekstów pewne ich elementy bywają traktowane jako metadane (ang. *metadata*) niebędące w sensie technicznym integralną częścią tekstu zasadniczego (są to np.: adres URL identyfikujący dany tekst w Internecie, data i godzina jego publikacji, autor tekstu, kategoria tekstu, powiązane z nim tagi oraz odnośniki typu *więcej na ten temat*, *powiązane artykuły* itp.), cała strona zaś – wraz z jej interfejsem – stanowi swoisty

paratekst dla prezentowanych na niej treści (Loewe 2007: 217–219). Strony z newsami można rozpatrywać jako gazety elektroniczne (Grzenia 2006: 174), same newsy natomiast – jako prymarnie elektroniczne realizacje tekstów okołoprasowych z gatunku bliskiego czy tożsamego ze wzmianką, fleshem lub nowinami. Z czysto genologicznego punktu widzenia takie podejście jest uzasadnione, jednak w kontekście web crawlingu ważniejsze jest uwzględnienie masowości powstawania newsów internetowych oraz przytaczane wyżej swoiste rozczłonkowanie pewnych ich elementów. Przez news internetowy (jako obiekt web crawlingu) będziemy rozumieć dowolny tekst prasy elektronicznej o charakterze informacyjnym (czy też parainformacyjnym), który w przestrzeni Internetu (bezpośrednio w niej lub po jego wyzyskaniu z tej przestrzeni, np. dla celów badawczych) jako całościowy znak powinien być traktowany jedynie w połączeniu z uzupełniającymi go metadanymi. Tylko dzięki tym danym daje się on w pierwszej kolejności wyodrębnić z masywu tekstowej inforzeczywistości, w dalszej – klasyfikować oraz zestawiać z innymi takimi tekstami (również w warstwie treściowej).

Głównym celem lingwistycznym przeprowadzonego przez nas badania jest analiza frazematyczna newsów internetowych, co można zaliczyć do działań z obszaru szeroko rozumianej frazeologii czy ściślej – frazematyki. Są to działania polegające na identyfikacji i ewaluacji ciągów wyrazowych podejrzanych o frazematyczność, wyodrębnionych dzięki wykorzystaniu jednego z tzw. wskaźników lub pomocników formalnych, jakim jest cudzysłów (Chlebda 2003: 238–242, 2010: 21–23). O innych celach językoznawczych, którym może służyć pozyskiwanie znacznych ilości tekstów internetowych, będziemy jedynie wzmiankować podczas szczegółowego opisu stworzonego przez nas korpusu[1].

W tym miejscu doprecyzowania wciąż wymaga termin „web crawling". Na język polski może on zostać przetłumaczony jako 'indeksowanie sieci'. Tłumaczenie to, szczególnie lingwistom niekomputerowcom, może wydawać się niezbyt klarowne. Zasadniczo web crawling polega na pozyskiwaniu pewnych informacji na temat stron internetowych, choć sposób i zasięg tych działań zależy od celu, w jakim zostały przeprowadzone (lub są stale prowadzone)[2].

[1] Zasadne wydaje się traktowanie jako korpus zarówno zasobów internetowych w ogóle, jak też pewnych wyodrębnionych, spójnych pod jakimś względem, mniejszych kolekcji tekstów (Kilgarrif, Grefenstette 2003: 333–334). Od roku 2008 za sprawą organizacji Common Crawl (commoncrawl.org) tekstowe dane internetowe gromadzone są na wielką skalę. Zasoby te w ramach różnych przedsięwzięć lingwistycznych bywają przetwarzane i udostępniane np. jako kolekcje jednojęzyczne. Por. choćby korpus OSCAR (Ortiz Suárez, Romary, Sagot 2020).

[2] Za przykłady takich działań weźmy web crawling prowadzony przez twórców serwisów Frazeo i Monco (frazeo.pl; monco.frazeo.pl – zob. Pęzik 2020) czy też korpusu Taiga (Shavrina, Shapovalova 2017). Zasoby Frazeo i Monco nierzadko wykorzystywane

W naszym wypadku web crawling dotyczył przede wszystkim pozyskiwania tekstów, a nie indeksowania stron źródłowych, z których teksty te ekscerpowaliśmy.

2. Baza materiałowa: źródła, ekstrakcja danych, strukturyzacja korpusu

Przeprowadzony przez nas web crawling odbywał się dwuetapowo i dotyczył czterech popularnych rosyjskojęzycznych serwisów informacyjnych (Vesti, Interfax, Lenta, Fontanka; zob. strony vesti.ru, interfax.ru, lenta.ru, fontanka.ru). Podstawowym kryterium doboru dokładnie tych źródeł była właśnie ich popularność (ustalona na podstawie danych ze strony liveinternet.ru). Dodatkowym kryterium była nowoczesność witryn internetowych serwisów, co ułatwiało generowanie list odnośników do poszczególnych newsów oraz samo wyzyskiwanie ich treści (tekstu zasadniczego wraz z metadanymi). Na całościowy proces ekscerpcyjny kompletnych tekstów newsów składały się następujące etapy oraz ich subetapy:

1. Etap półautomatyczny:

a) ustalenie początku istnienia serwisu i daty pierwszej publikacji;
b) stworzenie listy wszystkich możliwych dat w przedziale od pierwszej publikacji do 31.12.2019;
c) wygenerowanie odnośników do podstron danego źródła z listami wszystkich dostępnych dla danej daty newsów.

2. Etap automatyczny[3]:

a) odwiedzanie przez program komputerowy (crawler[4]) poszczególnych podstron;
b) sczytywanie wszystkich dostępnych na danej podstronie odnośników do newsów;
c) odwiedzanie strony danego newsa;
d) ekstrakcja wybranych danych;
e) zapis danych do pliku korpusu.

Liczbowe podsumowanie powyższych działań prezentujemy w tabeli 1.

---

są w badaniach nad językiem polskim – zwykle w celu analizy jakiegoś trendu językowego (Kozioł-Chrzanowska 2015: 31–34) czy analizy o charakterze kolokacyjno-statystycznym (Falkowska 2019: 30), zasoby Taiga z kolei zostały włączone (obok materiałów Narodowego Korpusu Języka Rosyjskiego) do zbiorów korpusu CoCoCo i także stanowią podstawę licznych badań językowych (np. Kopotev, Escoter, Kormacheva, Pierce, Pivovarova 2015).

[3] Poprzedzony napisaniem odpowiednich programów (crawlerów) w języku Python.

[4] Crawler, czyli tzw. robot internetowy, to program wyzyskujący z Sieci pewne informacje (np. o strukturze i treści stron internetowych).

Dla poszczególnych zasobów pewna liczba newsów nie zawierała żadnego tekstu. Przypadki te zwykle związane były z umieszczeniem w treści danego newsa (przez jego autora) elementów nietekstowych, np. pojedynczego zdjęcia, galerii zdjęć, filmu video czy tzw. widgetu z mediów społecznościowych (np. wpisu lub komentarza z Twittera, Facebooka i in.). Pełny wykaz adresów tych newsów umieszczamy na stronie dborysowski.info/advocem (zob. *Indeks pustych tekstów*). Ich liczbę podajemy w tabeli 1. Wyjątkowo duża ich liczba dotycząca zasobu Fontanka związana jest z taką budową blisko 47 tys. dokumentów HTML tego serwisu, której nie przewidywał nasz crawler (powstał on na podstawie analizy kilkunastu losowo wybranych dokumentów). We wszystkich tych wypadkach korpus zawiera jednak kompletne metadane danego newsa, dlatego jest możliwe ich odtworzenie (na potrzeby powstania niniejszego tekstu zostały one pominięte).

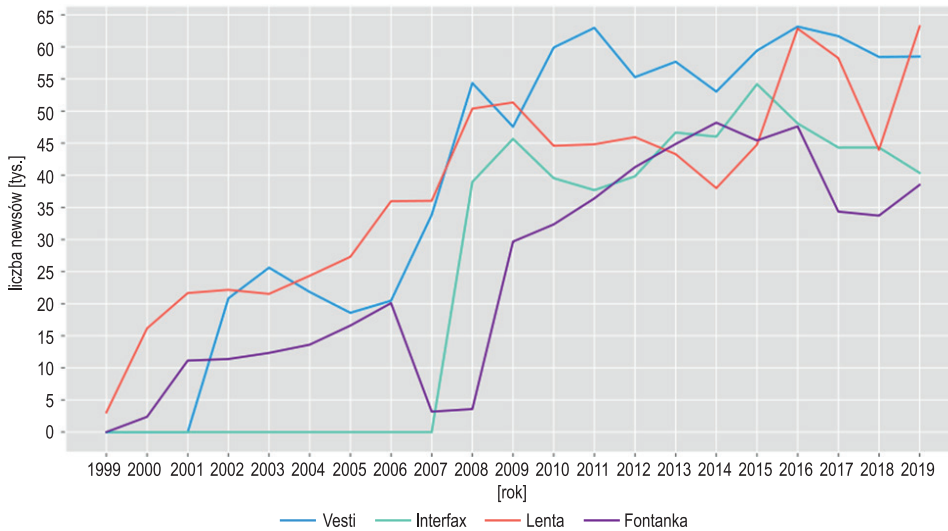
Tabela 1. Liczbowe podsumowanie przeprowadzonego crawlingu

Źródło	Pierwsza publikacja	Liczba wszystkich dat	Liczba wydobytych newsów		
			wszystkie	puste	niepuste
Vesti	18.02.2002	6 526	835 416	946	**834 470**
Interfax	11.02.2008	4 342	525 723	13	**525 710**
Lenta	31.08.1999	7 428	799 797	143	**799 654**
Fontanka	14.06.2000	7 140	573 523	46 839	**526 684**
SUMA			2 734 459	47 941	**2 686 518**

Dodatkowych informacji, np. o tendencjach wzrostu bądź spadku liczby newsów publikowanych w konkretnym serwisie, może dostarczyć podział danych przedstawionych w tabeli 1 na poszczególne lata czy miesiące. Wykorzystanie również godzin publikacji wszystkich newsów (oraz powiązanie dat ich publikacji z dniami tygodnia) umożliwia przedstawienie nawet bardziej szczegółowych statystyk[5], jednak tak dokładna analiza chronologizacyjna nie zostanie tutaj przeprowadzona. Statystyki dotyczące ekscerpcji newsów dla lat prezentujemy na wykresie 1[6]. Znaczący spadek

[5] Por. podobne dane w serwisie Frazeo (np. przyrost danych w perspektywie czasu pod adresem monco.frazeo.pl/stats#facets czy też *Cykle w życiu i języku* – frazeo.pl/blog/details/3).

[6] Warto dodać, iż ekscerpcja danych z różnych źródeł rzadko sprowadza się do prostego rozmieszczenia tych danych w odpowiednim polu korpusu. Ważnym elementem takich prac jest normalizacja ekscerptów. Rzecz w tym, aby wszystkie pozyskiwane dane (np. teksty wraz z ich metainformacją) były porównywalne strukturalnie, a w odniesieniu do pewnych ich elementów, jak choćby zapisu dat, przyjmowały identyczny format. W wypadku zasobów internetowych utrudniają to różnice w strukturze samych dokumentów HTML – różnice



Wykres 1. Liczba pozyskanych newsów w ujęciu rocznym
Źródło: Opracowanie własne.

liczby tekstów pozyskanych z serwisu Fontanka w latach 2007–2008 wynika ze wspomnianych wyżej błędów crawlingu.

Omawiany tu korpus 2 686 518 rosyjskojęzycznych tekstów (dalej zwany również NewsRu) zapisany jest w czterech osobnych plikach w formacie JSON Lines[7]. Format ten służy do przechowywania danych zgodnie ze specyfikacją

w sposobie strukturyzowania, klasyfikowania i wypełniania treścią wybranych bloków tych dokumentów (zaznaczmy, iż dotyczy to przede wszystkim obranej przez nas metodologii ekscerpcyjnej); w wypadku innych danych, np. danych elektronicznych ujętych w takich systemach, jak dLibra, o podobnych trudnościach decyduje nierzadko brak ustaleń „w zakresie standardu notacji chronologizacyjnej" (Wierzchoń 2013: 107). W tym kontekście musimy się przyznać do popełnienia błędu: podczas crawlingu danych wszystkie newsy z zasobów Vesti, Interfax, Lenta dotyczące marca zostały włączone do korpusu jako kwietniowe.

[7] W tym miejscu konieczne jest choćby zwięzłe przybliżenie stosowanego przez nas terminu „korpus" – z zastrzeżeniem, iż dotyczy on konkretnego, omawianego w niniejszym tekście, zbioru newsów internetowych. W tym ujęciu możemy mówić jedynie o **korpusie doraźnym** czy **oportunistycznym** (czy nawet o tzw. korpusie „domowym" – ang. *home-grown corpora*; zob. Pęzik 2013a). Wobec języka rosyjskiego korpus ten w żadnym stopniu nie stanowi reprezentatywnej próbki tekstów, lecz tylko pewien wycinek rosyjskojęzycznej internetowej rzeczywistości informacyjnej. Jest to wycinek niemały, liczący ok. 0,5 mld segmentów wyrazowych i prawdopodobnie będący językowym odzwierciedleniem wydarzeń w Rosji i na świecie w pewnym przedziale czasowym (od ok. roku 2000 do końca 2019). Jednym z istotnych elementów charakterystyki korpusów w ogóle są opisujące je dane liczbowe. Wybrane z nich na temat zbioru NewsRu prezentujemy pod adresem dborysowski.info/advocem (*Korpus NewsRu w liczbach*). Choć korpus ten jest mniejszy od innych dostępnych korpusów języka rosyjskiego (por. tatianashavrina.github.io/2018/08/30/datasets/), precyzja wyzyskiwania zarówno samych tekstów, jak i powiązanych z nimi metadanych

JSON (JavaScript Object Notation; zob. json.org), z uwzględnieniem jedynie takiego wymogu, aby każdy oddzielny rekord tych danych zajmował dokładnie jedną linię pliku tekstowego (zob. jsonlines.org). Taką jedną linię (którą dla przejrzystości prezentujemy poniżej w układzie blokowym) można utożsamiać z pojęciem rekordu w bazach danych. Do podstawowych typów elementów pojedynczego rekordu należą obiekty (zwane niekiedy słownikami) oraz tablice (zwane niekiedy listami). Zdecydowanie techniczny wydźwięk powyższej wypowiedzi z pewnością zrównoważą konkretne egzemplifikacje. Obiekty (słowniki) mogą przyjąć następującą postać:

```
{
"metadata":
        {
        "url": "https://www.vesti.ru/doc.html?id=3226219&cid=8",
        "date": "2019-12-31",
        "time": "23:39:00",
        "category": "Происшествия",
        "author": "No author",
        "title": "У погибшего в Ингушетии полицейского остались четверо маленьких детей"
        },
"text": "Установлены личности участников нападения на пост ДПС на въезде в Магас; один из них получил ранения (…)"
}
```

Granice obiektów wyznaczają nawiasy klamrowe stanowiące ramy dla odpowiednich pól (tzw. kluczy) oraz wartości przyporządkowanych tym polom[8]. Powyższy przykład obrazuje strukturę naszego korpusu (konkretniej – plików, które dotyczą zasobów Vesti i Interfax). Dane pozyskane z Lenta są dodatkowo wzbogacone o podkategorie (klucze „subcategory"), a wraz z danymi z Interfax uwzględniają również tagi (klucze „tags"). Właśnie tagi umieszczaliśmy w tablicach (lub listach), które mogą przyjąć następującą postać (ich granice wyznaczają nawiasy kwadratowe):

```
"tags":
        ["Общество", "Преступность", "Происшествия", "Конфликты", "Страноведение", "Выборы"]
```

może świadczyć o jego oryginalności – w odniesieniu do badań jednoosobowych, efektem podobnych projektów wieloosobowych bowiem, tzn. dotyczących web crawlingu dla celów korpusowo-lingwistycznych, jest np. korpus ГИКРЯ (por. webcorpora.ru; Pipierski 2013). Korpus NewsRu w całości dostępny jest w serwisie OSF na prawach licencji CC BY-NC 4.0 (zob. osf.io/697md/).

[8] Jak widać, obiekty mogą być umieszczone jeden wewnątrz drugiego, tzn. mogą stanowić wartość odpowiedniego klucza obiektu wyższego rzędu.

Tak strukturyzowane dane można przeszukiwać za pomocą narzędzia jq (stedolan.github.io/jq). Polecenia tego narzędzia w połączeniu ze standardowymi poleceniami wiersza poleceń systemów Unix/Linux oraz podstawową znajomością wyrażeń regularnych (ang. *regular expressions*) dają niemały wachlarz możliwości filtrowania korpusu – zarówno na poziomie metadanych, jak i samych tekstów newsów. Przykładowo filtr [1]:

```
[1] jq'. | select(.text == "") | .metadata.url' Vesti.jsonl
```

pokaże wszystkie odnośniki do newsów, dla których wartość klucza „text" jest tzw. pustym łańcuchem. Do powyższego filtra [1] można dodać przekierowanie do polecenia zliczającego wyniki (np. wc -l), aby dowiedzieć się, ile pustych tekstów zawiera nasz korpus (dla danego zasobu). Z kolei filtr [2]:

```
[2] jq -r .text Lenta.jsonl | grep -oE "^(\w+ ){5}"
```

wyświetli pięć pierwszych wyrazów każdego tekstu zawartego w zasobie Lenta (o ile wcześniej nie pojawi się jakiś znak interpunkcyjny), natomiast filtr [3]:

```
[3] jq -r .text Lenta.jsonl | grep -oE "^[^,|\.]*"
```

pokaże początek każdego z tych tekstów do pierwszego wystąpienia przecinka lub kropki.

Podobne filtry mogą służyć badaniu zgromadzonych newsów pod kątem zawartości zarówno samych tekstów zasadniczych ([2] i [3]), jak i związanych z tymi tekstami metadanych ([1]).

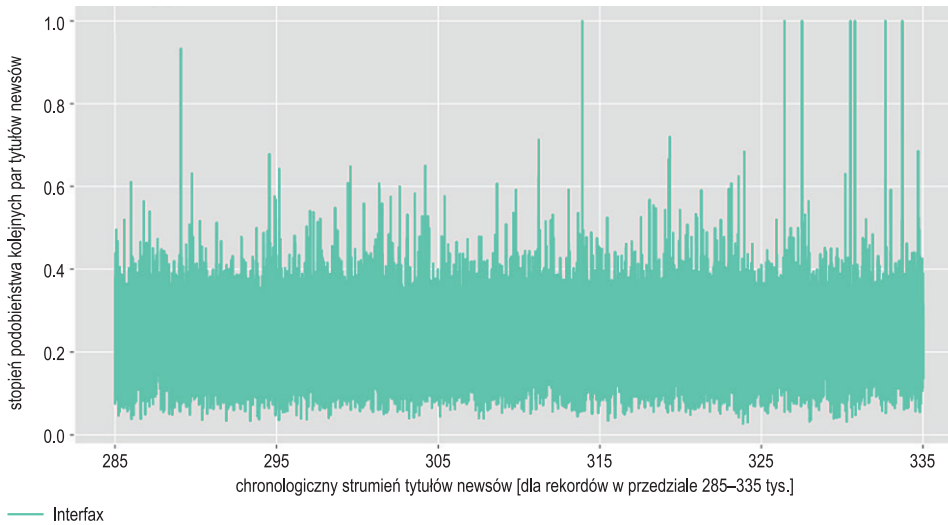
***

Korpusy tekstów internetowych często poddawane są tzw. deduplikacji. Wykres 2 pokazuje, w jaki sposób newsy pozyskane z wybranego źródła – tu: Interfax – łączą się w pewne ciągi (strumienie) informacyjne, zidentyfikowane na podstawie podobieństwa ich tytułów[9]. Do wykresu 2 wybraliśmy wycinek 50 tys. kolejnych newsów (z 575 710), poczynając od 285-tysięcznego (czyli mniej więcej połowy). Grupowanie się zbliżonych tytułów może np. świadczyć o relacjonowaniu jednego wydarzenia w kilku różnych newsach. Dla podobieństwa bliskiego 90% obrazują to dwa poniższe przykłady (podajemy je w zapisie zgodnym z oryginalnym wraz z własnym tłumaczeniem):

[9] W tym celu skorzystaliśmy z modułu difflib języka programowania Python (moduł ten wykorzystuje nieznacznie zmodyfikowany algorytm znany jako Gestalt Pattern Matching, por. docs.python.org/3/library/difflib.html; zob. także Ratcliff, Metzener 1988).

(1) *Число погибших при пожаре на рынке в Москве увеличилось до девяти человек.*
Liczba ofiar pożaru na targu w Moskwie wzrosła do dziewięciu.

(2) *Число погибших при пожаре на рынке в Москве увеличилось до 12 человек – МЧС.*
Liczba ofiar pożaru na targu w Moskwie wzrosła do 12 – Ministerstwo do spraw sytuacji wyjątkowych.



Wykres 2. Stopień podobieństwa kolejnych par tytułów newsów z zasobu Interfax
Źródło: Opracowanie własne.

Dla podobieństwa na poziomie ok. 80% podobną tendencję obrazują kolejne dwa tytuły – (3) i (4) – w nieco inny sposób informujące o wprowadzeniu kar za kąpiel w niedozwolonych miejscach w obwodzie moskiewskim:

(3) *В Подмосковье собрались ввести штрафы за купание в неположенных местах.*
Rozważane jest wprowadzenie kar za kąpiel w niedozwolonych miejscach w obwodzie moskiewskim.

(4) *В Подмосковье задумались о штрафах за купание в неположенном месте.*
tłumaczenie jw.

Dzięki porównaniu tytułów newsów w części korpusu NewsRu dotyczącej źródła Interfax odnaleźliśmy 215 duplikatów sąsiadujących ze sobą tytułów, co oznaczało identyczność newsów również na poziomie treściowym, lecz rozbieżność identyfikujących te dokumenty odnośników. Wykorzystanie do podobnych celów modułu difflib (zob. przyp. 8) jest zaledwie jedną z wielu możliwości (w zależności od potrzeb stosowane są inne narzędzia, np. moduł

Levenshtein, czy też inne miary, takie jak współczynnik Sørensena, indeks Jaccarda, odległość Hamminga; więcej o tym zob. np. Manku, Jain, Sarma 2007; Zielenkow, Siegałowicz 2007; Leskovec, Rajarman, Ullman 2010–2014).

3. Analiza frazematyczna: pojęcie reproduktu, bi- i trigramy cudzysłowowe

Prezentowana tutaj analiza dotyczy wyodrębniania z zebranego przez nas korpusu newsów pewnych całostek wielowyrazowych ujętych w cudzysłów. Wobec każdego takiego tworu w kontekście ekscerpcji automatycznej można powiedzieć, iż jest on „podejrzany o bycie wielowyrazowym reproduktem" (Chlebda 2010: 22). Termin „reprodukt" jego autor zdefiniował swego czasu jako:

> jednostkę języka (składnik systemu danego języka etnicznego) wyodrębnioną z tekstów, sformułowanych w tym języku, w rezultacie stwierdzenia jej regularnej powtarzalności w tych tekstach w funkcji werbalizatora tego, „co autor chciał powiedzieć", czyli pewnej wiązki sensów, intencji, emocji, określonego zespołu treściowego (Chlebda 2010: 15)[10].

Identyfikacja cudzysłowowych związków wielowyrazowych nawet podczas analizy wzrokowej jakiegoś tekstu jest dość łatwa (całostki te wyodrębniają się w pewnym sensie same, por. Chlebda 2003: 240). W wypadku analizy automatycznej lub zautomatyzowanej cudzysłów z jednej strony ułatwia proces identyfikacyjny zawartych między nim segmentów tekstu, z drugiej zaś – generuje pewne problemy techniczne, związane z dużą dowolnością jego stosowania w tekstach internetowych bądź po prostu cyfrowych (Wierzchoń 2010: 91–94). Pomimo względnej łatwości w identyfikacji wielu cudzysłowowych całostek wielowyrazowych (i wielu ich wystąpień – dodatkowo w odpowiednio dużym zbiorze tekstów) nie jest możliwe arbitralne przyznanie im statusu jednostek systemowych tylko na podstawie wysokiej ich powtarzalności. Status jednostek języka (Bogusławski 1976: 357–359) czy jednostek znaczenia (Sinclair 1996: 94) mogą one zyskać w drodze szczegółowej analizy, której podstawą jest skomplikowany i wysoce kompetencyjny proces o podłożu neurolingwistycznym (Wierzchoń 2010: 88). Metody półautomatyczne lub automatyczne mogą jedynie wspomóc wstępną ekscerpcję (z jakiegoś zbioru tekstów) tych całostek, które o systemowość będą jedynie podejrzane (Fiedoruszkow 2010: 60–61). Bardzo ważnym kryterium

[10] W swej istocie pojęcie reproduktu bliskie jest rozumieniu tzw. *multiword expressions*, głównie w odniesieniu do tych ich realizacji, które określane są jako pragmatycznie oraz statystycznie idiomatyczne (por. Baldwin, Kim 2010: 6–7).

dalszej ewaluacji każdego takiego segmentu jest ustalenie **regularności** jego powtarzalności. Manualne jej stwierdzenie jest z oczywistych powodów utrudnione, ponadto zawsze będzie nosić znamiona subiektywności[11]. W ramach paradygmatu frazeologii dystrybucyjnej, której przykładem jest przytoczona wyżej koncepcja reproduktu (Pęzik 2013b: 145–148), może zostać przeprowadzona bardziej zobiektywizowana ewaluacja – tu: dzięki wykorzystaniu danych korpusowych segmentowanych względem czasu (roku, miesiąca, tygodnia, dnia, a nawet godziny). Dla odnalezionych przez nas bi- oraz trigramów cudzysłowowych (czyli dwu- i trzywyrazowców ujętych w cudzysłów) postaramy się zastosować niektóre z założeń tego podejścia. Warto przy tym pamiętać, iż na wszelkie nasze ustalenia należy patrzeć przez pryzmat sporych ograniczeń. Dotyczą one odmiany językowej i gatunku (news), zasięgu czasowego (lata od ok. 2000 r. do 2019 r.; w jednym wypadku od 2008 r.) oraz zróżnicowania źródeł (tylko cztery wybrane serwisy informacyjne). Poniższe operacje dotyczą tylko tekstów zasadniczych newsów.

Z technicznego punktu widzenia proces wyszukiwania interesujących nas bi- i trigramów cudzysłowowych polegał na zastosowaniu nieznacznie różniących się od siebie filtrów (jednego dla bi-, drugiego zaś – dla trigramów) częściowo złożonych z poleceń narzędzia jq, częściowo z poleceń Terminala[12], częściowo zaś – z wyrażeń regularnych[13]. Filtry te uwzględniały pewne zabiegi normalizacyjne oraz ograniczenia formalne. Były to kolejno:

1. Normalizacja:

a) zmiana wielkości liter na małe;
b) ujednolicenie cudzysłowów (wszystkie podwójne do postaci ["], pojedyncze zaś – [']).

2. Ograniczenia formalne:

a) pomijanie wyników z cyframi/liczbami;
b) pomijanie wyników ze znakami alfabetu innymi niż cyrylica.

Ostatecznie zastosowane przez nas filtry przyjęły formę, którą można zawrzeć w swoistej instrukcji: wybierz wszystkie teksty z danego zasobu, dokonaj ujednolicenia cudzysłowów, znajdź N-gramy (bi- lub trigramy cudzysłowowe), zamień wielkie litery na małe, pomiń wyniki z cyframi/liczbami,

[11] Podobnie jak kwestia ustalania granic reproduktów wielowyrazowych, które tutaj wyznacza cudzysłów (również w ogóle ocena takich tworów werbalnych pod kątem stwierdzenia ich frazeologiczności czy systemowości; por. Pajdzińska 2006: 231; Bańko 2001: 154–159).

[12] W dużym uproszczeniu Terminal to program służący do obsługi systemu operacyjnego z poziomu poleceń tekstowych.

[13] Stosowanie wyrażeń regularnych do wyszukiwania pewnych regułowych połączeń wyrazowych było wykorzystywane w polskim językoznawstwie niejednokrotnie, por. choćby (Wierzchoń 2002; Małek 2006).

pomiń wyniki ze znakami spoza cyrylicy (nie licząc cudzysłowów), pokaż tylko unikatowe wystąpienia, zliczając ich frekwencję, posortuj malejąco według frekwencji, zapisz do pliku. Wynikiem filtrowania były N-gramy wyłącznie w całości zawarte w cudzysłowach. Tabela 2 podsumowuje przeprowadzone operacje filtrowania dla wszystkich zasobów z podziałem na bi- i trigramy (F oznacza tutaj frekwencję; wyniki nie zostały zlematyzowane).

Tabela 2. Podsumowanie wyników filtrowania korpusu NewsRu pod kątem bi- i trigramów cudzysłowowych

Zasób	Bigramy			Trigramy		
	SUMA	F >= 100	F <= 3	SUMA	F >= 100	F <= 3
Vesti	**126 501**	464	112 505	**61 571**	68	57817
Interfax	**63 873**	302	55 435	**33 228**	37	30 646
Lenta	**125 716**	522	110 111	**69 095**	71	64 192
Fontanka	**106 074**	372	93 384	**51 627**	46	48 111

W tabeli 2 oprócz ogólnej liczby tzw. okazów dwu- i trzywyrazowców wyszukanych w naszym korpusie (z podziałem na cztery zasoby) uwzględniamy również ich liczbę z frekwencją na poziomie 3 lub mniej powtórzeń oraz 100 lub więcej powtórzeń. Te dodatkowe dane liczbowe pokazują, iż odpowiednio 89% i 94% (Vesti), 87% i 92% (Interfax) oraz 88% i 93% (Lenta i Fontanka) wszystkich bi- i trigramów to związki wielowyrazowe rzadkie (o niskiej lub bardzo niskiej frekwencji). Dane te potwierdzają prawidłowość twierdzenia Zipfa, zgodnie z którym iloczyn rangi danego okazu i jego frekwencji jest wartością względnie stałą (rzadkich okazów będzie zawsze zdecydowanie więcej niż częstych).

W kolejnym kroku połączyliśmy cztery listy rangowe bi- i trigramów, dokonując ich lematyzacji (zob. pypi.org/project/pymystem3/), a także sumując frekwencję tych powtarzających się w co najmniej dwóch zasobach. W tabeli 3 uwzględniamy dziesięć najczęstszych oraz dziesięć rzadkich dwu- i trzywyrazowców po tym połączeniu (zachowujemy pisownię znormalizowaną, a także cudzysłów; N-gramy rzadkie – ze względu na sortowanie – dotyczą końca alfabetu)[14].

Wstępna analiza wzrokowa wyników cząstkowych podpowiada nam, iż podobne wielowyrazowe całostki można w jakimś stopniu klasyfikować. W pierwszej kolejności zauważamy, iż najczęstsze bigramy są m.in.

[14] Lista wszystkich bi- i trigramów (oraz dodatkowo uni-, tetra- i pentagramów) została umieszczona na stronie dborysowski.info/advocem (zob. *Listy cudzysłowowych N-gramów*). Dla czytelności w tabeli 3 podajemy bazowy wariant systemowy, np. „единая россия" zamiast „единый россия".

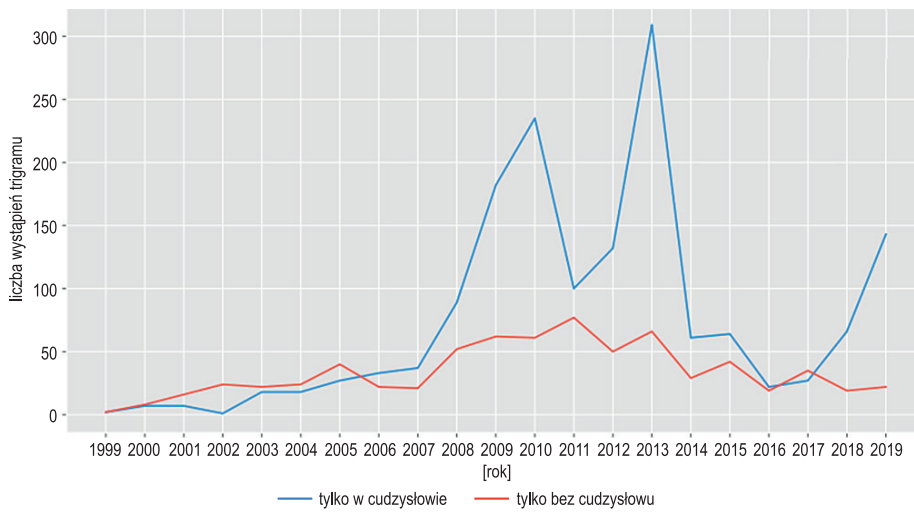
Tabela 3. Dziesięć najczęstszych oraz dziesięć rzadkich bi- i trigramów cudzysłowowych odnalezionych w korpusie NewsRu

F	Najczęstsze bigramy	F	Najczęstsze trigramy
50720	„единая россия”	3604	„русская служба новостей”
42057	„вести фм”	3457	„российские железные дороги”
28134	„исламское государство”	1735	„блок петра порошенко”
18655	„справедливая россия”	1580	„за права человека”
18642	„эхо москвы”	1394	„тур де франс”
9192	„скорая помощь”	1253	„союз правых сил”
8305	„манчестер юнайтед”	1234	„город без наркотиков”
8100	„правый сектор”	1230	„вор в законе”
7725	„новая газета”	1178	„незаконный оборот оружия”
7539	„нафтогаз украины”	1156	„гражданские самолеты сухого”
	Rzadkie bigramy		Rzadkie trigramy
F = 3 (koniec alfabetu)	„ящик доверия”	F = 3 (koniec alfabetu)	„я танцевать хочу”
	„ясловске богунице”		„я не верю”
	„я пришел”		„я все помню”
	„японский синдром”		„эффект сильвио берлускони”
	„яйцо фаберже”		„это только версия”
	„язык общения”		„это совершенно ясно”
	„языковая шизофрения”		„это семейный праздник”
	„языковое гнездо”		„это не наше”
	„я живой”		„это моя мама”
	„ядерная полуночь”		„это был теракт”

nazwami radiostacji (*Wiesti FM*, *Echo Moskwy*), partii politycznych (*Jedna Rosja*, *Sprawiedliwa Rosja*), zespołów piłkarskich (*Manchester United*) lub państw (*Państwo Islamskie*). Wśród trigramów odnajdujemy ponadto nazwy organizacji, ruchów obywatelskich (np. *W Obronie Praw Człowieka*, *Miasto Bez Narkotyków*), a także terminy (dotyczące kodeksu karnego, np. *nielegalny handel bronią*). Są to z pewnością reprodukty. Ich wysoka frekwencja nie dziwi w kontekście specyfiki materiału źródłowego, jakim są newsy internetowe. Dokładne ustalenie najczęstszych znaczeń wymaga oczywiście analizy kontekstów i w wypadku 1580 tekstów dla trigramu *„за права человека”* pochłonie sporo czasu. Możliwe jest przyspieszenie takiej analizy poprzez zbadanie, czy w jakiś sposób powtarzają się mikrokonteksty z lewej bądź prawej strony tej całostki (np. wyrazy poprzedzające dany reprodukt wielowyrazowy lub następujące po nim; por. Pęzik 2018: 4299). Pozwoli to na ustalenie następujących najczęstszych połączeń: *движение*

„за права человека”, *организация „за права человека”* (odpowiednio z wyrazami *ruch* i *organizacja*).

W odniesieniu do chronologizacji pewnych całostek warto badać ich rozmieszczenie w korpusie względem czasu. Dostarcza to ważnych informacji o zasięgu ich funkcjonowania w wybranym wycinku tekstowej rzeczywistości (takim, który został uwzględniony w danym korpusie). Wykres 3 przedstawia frekwencję trigramu *„за права человека”* w ujęciu chronologicznym. Ponadto w zestawieniu uwzględniamy wystąpienia tej całostki bez cudzysłowu (łączna ich liczba to 890 powtórzeń), ponieważ, jak się wydaje, może być on reprodukowany także lub głównie poza kontekstem onimicznym (w sumie odnotowano 2470 wystąpień bez cudzysłowu i z cudzysłowem).



Wykres 3. Rozkład chronologiczny trigramów *„за права человека”* oraz *за права человека* w korpusie NewsRu
Źródło: Opracowanie własne.

Wykres 3 tylko częściowo potwierdza nasze przypuszczenia. Frekwencja użycia danego trigramu bez cudzysłowu jest dość stabilna, co może oznaczać jego regularną odtwarzalność w języku newsów internetowych w ogóle[15]. Powtórzenia cudzysłowowe z kolei mogą wskazywać na powstanie lub zwiększoną aktywność ruchu/organizacji o nazwie *За права человека* (dlatego w wybranych okresach jest ich znacznie więcej niż niecudzysłowowych).

W ramach badań korpusowych nie mniej ważna jest analiza zjawisk rzadkich oraz bardzo rzadkich (również takich, których zasięg temporalny

[15] Przypomnijmy, że na mniejszą frekwencję w latach 1999–2002 może mieć wpływ brak danych z tych lat dla dwóch zasobów – zob. tabela 1.

jest niewielki – np. tzw. efemeryd leksykalnych; por. Graliński, Wierzchoń 2017: 103). Weźmy w tym celu przykłady uwzględnione w tabeli 3 i postawmy pytanie, czy wybrane z nich w korpusie NewsRu są po prostu rzadkie, czy jednak są rzadkie z powodu ujęcia ich w cudzysłów. Wyniki eksperymentu polegającego na porównaniu frekwencji związków dwuwyrazowych *„ящик доверия”*, *„японский синдром”*, *„яйцо фаберже”*, *„языковая шизофрения”*, *„ядерная полуночь”*, *„эффект сильвио берлускони”* z frekwencją uzyskaną dla tych samych całostek, lecz bez cudzysłowu, przedstawiają się następująco:

a) *ящик доверия* (*skrzynka zaufania*; inicjatywa polegająca na stworzeniu skrzynek, do których obywatele mogą wrzucać pytania lub prośby skierowane bezpośrednio do władz): **5 wystąpień** (2011; jedno z 5 użyć dotyczyło zbierania anonimowych pomysłów pracowników firmy Toyota, choć mogło być motywowane wcześniejszym tekstem z 4 użyciami);
b) *японский синдром* (*syndrom japoński*; inaczej również – *syndrom paryski*, dolegliwość polegająca na rozczarowaniu się japońskich turystów Paryżem): **3 wystąpienia** (2010 rok – jeden tekst);
c) *яйцо фаберже* (*jajo Fabergé*): **178 wystąpień** (wszystkie lata oprócz 1999 i 2002);
d) *языковая шизофрения* (*językowa schizofrenia*; sformułowanie ukraińskiej pisarki Larysy Nicoj dotyczące używania przez Ukraińców rosyjskiego wyrazu *кулич* zamiast ukraińskiego *паска* do nazwania wielkanocnej paschy): **11 wystąpień** (lata 2017, 2018);
e) *ядерная полуночь* (*jądrowa północ*; określenie amerykańskich naukowców, współautorów bomby atomowej, którzy stworzyli symboliczny zegar odliczający czas do atomowej apokalipsy; przesuwają wskazówkę zegara do przodu lub ją cofają w zależności od sytuacji jądrowej na świecie): **7 wystąpień** (lata 2012, 2015, 2018);
f) *эффект сильвио берлускони* (*Efekt Silvio Berlusconi*; tytuł książki, którą Władimir Putin podarował Silvio Berlusconi – książka traktuje o karierze włoskiego polityka): **3 wystąpienia** (2002 rok).

Jak pokazuje przeprowadzony eksperyment, tylko jeden cudzysłowowy bigram (*яйцо фаберже*) ma znacznie większą frekwencję bez cudzysłowu (użycia cudzysłowowe były zresztą związane z nazwą balonu oraz metaforami w odniesieniu do kształtu budynku i drogiego smartfona). Pozostałe całostki są rzadkie w ogóle, co, naszym zdaniem, nie umniejsza ich wagi w kontekście np. badań leksykalnych (być może niektóre z nich zasługują na miano jednostek języka, a – jak podkreślał Andrzej Bogusławski – „jednostki języka naturalnego NIE są dane. Czekają dopiero na odkrycie”; Bogusławski 1989: 167).

3. Podsumowanie

Przedstawiony zarys analizy cudzysłowowych reproduktów wielowyrazowych może być jedynie przyczynkiem do przeprowadzenia badań na większą skalę. Naszym celem było głównie pokazanie, w jaki sposób można gromadzić wybrane internetowe dane tekstowe, w jaki sposób można je strukturyzować (tak, aby możliwe było ich przeszukiwanie bez użycia zaawansowanych wyszukiwarek korpusowych) oraz jakie możliwości badawcze z takiego sposobu wynikają (np. analiza newsów internetowych pod kątem ich metadanych, wykrywanie duplikatów newsów, ustalanie trendów chronologicznych pewnych całostek, badanie zależności znaczeń związków wielowyrazowych od użycia cudzysłowu i in.). Przedstawione badania będą kontynuowane w przyszłości, choć w pierwszej kolejności zgromadzony przez nas korpus musi zostać uzupełniony i skorygowany (pod kątem wszelkich niedociągnięć, o których wzmiankowaliśmy).

Źródła internetowe

dborysowski.info/advocem
commoncrawl.org
docs.python.org/3/library/difflib.html
fontanka.ru
frazeo.pl
interfax.ru
json.org
jsonlines.org
lenta.ru
liveinternet.ru
monco.frazeo.pl
osf.io/697md/
pypi.org/project/pymystem3/
stedolan.github.io/jq
tatianashavrina.github.io/2018/08/30/datasets/
vesti.ru
webcorpora.ru

Literatura

Baldwin T., Kim S. N. (2010): *Multiword Expressions*. [W:] *Handbook of Natural Language Processing*. Red. N. Indurkhya, F. J. Damerau. Boca Raton, s. 267–292.

Bańko M. (2001): *Z pogranicza leksykografii i językoznawstwa. Studia o słowniku jednojęzycznym*. Warszawa.

Bartmiński J., Niebrzegowska-Bartmińska S. (2009): *Tekstologia*. Warszawa.

Bogusławski A. (1976): *O zasadach rejestracji jednostek języka*. „Poradnik Językowy" nr 8, s. 356–364.

Bogusławski A. (1989): *Preliminaria gramatyki operacyjnej*. „Polonica" R. XIII, s. 163–223.
Chlebda W. (2003): *Elementy frazematyki. Wprowadzenie do frazeologii nadawcy*. Łask.
Chlebda W. (2010): *Nieautomatyczne drogi dochodzenia do reproduktów wielowyrazowych*. [W:] *Na tropach reproduktów. W poszukiwaniu wielowyrazowych jednostek języka*. Red. W. Chlebda. Opole, s. 15–35.
Falkowska M. (2019): *Derywatywy słowotwórcze od rzeczownika empatia w tekstach współczesnej polszczyzny. Analiza semantyczna*. [W:] *Book of Abstracts. Polish Cognitive Linguistics Association Conference 2019. Cognitive Linguistics in the Year 2019*. Białystok, s. 30.
Fiedoruszkow J. (2010): *Metody automatyzacji ekscerpcji konstrukcji atrybutywnych języka rosyjskiego*. [W:] *Na tropach reproduktów. W poszukiwaniu wielowyrazowych jednostek języka*. Red. W. Chlebda. Opole, s. 59–85.
Graliński F., Wierzchoń P. (2017): *Jak powstaje słownik efemeryd leksykalnych polszczyzny XIX i XX wieku?* [W:] *Wokół „300 tysięcy polskich słów". Wstęp do hasłownikologii*. Red. J. Wawrzyńczyk, P. Wierzchoń. Warszawa, s. 101–118.
Grzenia J. (2006): *Komunikacja językowa w Internecie*. Warszawa.
Kilgarrif A., Grefenstette G. (2003): *Web as Corpus: Introduction*. „Computational Linguistics" vol. 29/3, s. 333–347.
Kopotev M., Escoter L., Kormacheva D., Pierce M., Pivovarova L., Yangarber R. (2015): *CoCoCo: Online Extraction of Russian Multiword Expressions*. [W:] *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing*. Sofia, s. 43–45.
Kozioł-Chrzanowska E. (2015): *„Przekrojowa" rubryka Heca hecą jako źródło potocznych reproduktów języka polskiego*. Kraków.
Kudra A. (2010): *News jako funkcja*. „Acta Universitatis Lodziensis. Folia Litteraria Polonica" nr 13, s. 399–404.
Leskovec J., Rajarman A., Ullman J. D. (2010–2014): *Mining of Massive Datasets*. Cambridge.
Loewe I. (2007): *Gatunki paratekstowe w komunikacji medialnej*. Katowice.
Małek E. (2006): *Filtry Wierzchonia jako narzędzie badawcze filologa*. Łódź.
Manku G. S., Jain A., Sarma A. D. (2007): *Detecting Near-Duplicates for Web Crawling*. [W:] *WWW'07: Proceeding of the 16th International Conference on World Wide Web*. New York, s. 141–149.
Ortiz Suárez P. J., Romary L., Sagot B. (2020): *A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages*. [W:] *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Red. D. Jurafsky, J. Chai, N. Schluter, J. Tetreault. ACL, s. 1703–1714.
Pajdzińska A. (2006): *Granice związku frazeologicznego jako problem leksykograficzny*. [W:] *Studia frazeologiczne*. Red. A. Pajdzińska. Łask, s. 222–231.
Pęzik P. (2013a): *Wybrane aspekty reprezentatywności małych i średnich korpusów*. [W:] *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*. Red. W. Chlebda. Opole, s. 45–58.
Pęzik P. (2013b): *Paradygmat dystrybucyjny w badaniach frazeologicznych. Powtarzalność, reprodukcja i idiomatyzacja*. [W:] *Metodologie językoznawstwa. Ewolucja języka. Ewolucja teorii językoznawczych*. Red. P. Stalmaszczyk. Łódź, s. 143–160.
Pęzik P. (2018): *Increasing the Accessibility of Time-Aligned Speech Corpora with Spokes Mix*. [W:] *LREC 2018: Eleventh International Conference on Language Resources and Evaluation*. Miyazaki, s. 4297–4300.
Pęzik P. (2020): *Budowa i funkcje korpusu monitorującego MoncoPL*. „Forum Lingwistyczne" nr 7, s. 133–150. Online: https://www.journals.us.edu.pl/index.php/FL/article/view/10335/7978. Pipierski A. Cz. (2013): *Gienieralnyj intierniet-korpus russkogo jazyka i poniatije riepriezientatiwnosti w korpusnoj lingwistikie*, „Gienieralnyj Intierniet-Korpus Russkogo Jazyka", <http://www.science-education.ru/pdf/2013/5/14.pdf>, dostęp: 19.10.2020.

Ratcliff J. W., Metzener D. (1988): *Pattern Matching: The Gestalt Approach*. „Dr. Dobb's Journal" vol. 7, s. 46.

Shavrina T., Shapovalova O. (2017): *To the Methodology of Corpus Construction for Machine Learning: „Taiga" Syntax Tree Corpus and Parser*. [W:] *Trudy mieżdunarodnoj konfieriencyi „Korpusnaja lingwistika – 2017"*. Red. V. P. Zakharov, M. V. Khokhlova. St. Petersburg, s. 78–84.

Sinclair J. (1996): *The search for units of meaning*. „Textus" vol. 9/1, s. 75–106.

Wierzchoń P. (2002): *Automatyzacja ekscerpcji definiowanych połączeń wyrazowych. Filtry wyrażeń regularnych*. [W:] *Przestrzenie informacji*. Red. P. Nowak, W. Krzemińska. Poznań, s. 119–184.

Wierzchoń P. (2010): *Pięć bardzo skutecznych (sprawdzonych) sposobów na masowe wyodrębnianie wielowyrazowych segmentów podejrzanych o frazematyczność (czyli reproduktów).* [W:] *Na tropach reproduktów. W poszukiwaniu wielowyrazowych jednostek języka*. Red. W. Chlebda. Opole, s. 87–125.

Wierzchoń P. (2013): *Druga dekada XXI wieku będzie dekadą „małej diachronii"*. [W:] *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*. Red. W. Chlebda. Opole, s. 97–111.

Witosz B. (2016): *Lingwistyczne koncepcje tekstu wobec wyzwań komunikacji wirtualnej*. [W:] *Język w internecie. Antologia*. Red. M. Kita, I. Loewe. Katowice, s. 101–112.

Zielenkow Ju. G., Siegałowicz I. W. (2007): *Srawnitielnyj analiz mietodow opriedielenija nieczetkich dublikatow dla Web-dokumientow*. „Gienieralnyj Intierniet-Korpus Russkogo Jazyka", <http://rcdl2007.pereslavl.ru/papers/paper_65_v1.pdf>, dostęp: 20.10.2020.