Półrocznik Językoznawczy Tertium. Tertium Linguistic Journal 4 (1) (2019)
www.journal.tertium.edu.pl
10.7592/Tertium2019.4.1.Zabawa

O roli samodzielnie przygotowanych korpusów w badaniach językoznawczych (na przykładzie korpusu wykorzystującego zasoby internetowe)

Marcin Zabawa
Uniwersytet Śląski w Katowicach
marcin.zabawa@us.edu.pl

Streszczenie

Celem niniejszego artykułu, o charakterze teoretyczno-przeglądowym, jest omówienie problematyki związanej z budową własnego korpusu językowego. Badacz, chcący skupić się np. na analizie neologizmów, musi oprzeć swoje badania na określonych źródłach: o ile dawniej często wykorzystywano do tego celu prasę, o tyle obecnie znacznie częściej są to korpusy językowe (np. NKJP) oraz Internet. Autor artykułu stawia tezę, że zarówno NKJP, jak i Internet jako całość, nie są jednak najlepszym wyborem w wypadku chęci badania np. najnowszego słownictwa polszczyzny, a już na pewno nie są wystarczające. Jeszcze więcej problemów stwarza wybór języka mówionego jako podstawy analiz. Najlepszym wyjściem, choć jednocześnie najtrudniejszym i najbardziej czasochłonnym, jest budowa własnego korpusu językowego. W artykule wykazano, dlaczego użycie prasy czy Internetu jako całości niekoniecznie jest najlepszym rozwiązaniem, a także omówiono różnego rodzaju aspekty teoretyczne związane z budową własnego korpusu (np. wybór rodzaju tekstów, wielkość korpusu, wykorzystanie narzędzi informatycznych ułatwiających tworzenie korpusu).

Słowa kluczowe: korpus, językoznawstwo korpusowe, język Internetu, język prasy

Abstract

Self-compiled Corpora in Linguistic Research (On the Example of an Internet Corpus)

The aim of the present paper, which is of a theoretical character, is to discuss the problems related to the process of the compilation of one's own linguistic corpus. A linguist who wants to study e.g. neologisms must base his or her analysis on a certain source. Formerly, the language of the press was frequently used as such source; now, however, linguistic corpora and the Internet are utilized more frequently. The author of the paper points out that both the National Corpus of Polish (NKJP) and the Internet as a whole are not the best choices (and are definitely not sufficient) when a linguist intends to study e.g. the newest vocabulary items in Polish. The use of the spoken language as the main source is even more problematic. The best solution, albeit the most difficult and time-consuming at the same time, is the compilation of one's own linguistic corpus. The paper discusses the inadequacy of regarding the press or the Internet as a whole as the best sources and then proceeds to discuss various theoretical aspects connected

with the compilation of one's own corpus (such as the choice of the type of texts, corpus size, the use of computer tools intended to aid in corpus compilation, etc.)

Keywords: corpus, corpus linguistics, Internet language, press language.

1. Wprowadzenie

Truizmem będzie stwierdzenie, że językoznawstwo korpusowe przeżywa obecnie bujny rozwój i coraz więcej autorów wykorzystuje elektroniczne zbiory tekstów (czyli korpusy) w swojej pracy naukowej. Korpus można tutaj zdefiniować, za Crystalem, jako

> A collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language.
>
> (Crystal [1980] 2008: 117)
>
> Zbiór danych językowych, w formie albo tekstów pisanych, albo transkrypcji nagranej mowy, który może być użyty jako podstawa opisu językowego lub jako sposób weryfikacji hipotez dotyczących języka[1].

Warto w tym miejscu podkreślić, że obecnie nie do końca zasadne jest używanie konstrukcji rozłącznej (*albo…albo*). Wiele korpusów, w tym przede wszystkim tzw. korpusy narodowe (*national corpora*), posiada zarówno komponent pisany, jak i mówiony (ten pierwszy jest zwykle znacznie większy, oscylujący – w zależności od korpusu – wokół 70-90% jego zawartości). W różnego rodzaju badaniach językoznawczych najczęściej używane są korpusy „gotowe" (w wypadku angielszczyzny jest to głównie korpus COCA, a polszczyzny przede wszystkim NKJP). Sytuacja taka jest zrozumiała, gdy weźmiemy pod uwagę, że ww. korpusy są stosunkowo duże, adnotowane morfologicznie i dostępne bez opłat. Nie dziwi zatem, że stosunkowo rzadko badacze decydują się na budowę własnego korpusu. Praca taka jest czasochłonna, jednak w wielu przypadkach warto podjąć taki wysiłek. Dobrym przykładem mogą tutaj być badania nad nowymi jednostkami leksykalnymi – neologizmami, neosemantyzmami, czy też zapożyczeniami, zwłaszcza ograniczonymi do określonego pola semantycznego, np. związanego z komputerami i Internetem (korpus taki może być szczególnie przydatny np. do badań nad stopniem asymilacji zapożyczeń, poprzez porównywanie częstości

[1] Wszystkie tłumaczenia pochodzą od autora artykułu. Definicje autorów polskojęzycznych można znaleźć np. w pracach M. Podhajeckiej (2006) czy M. Hebal-Jezierskiej (2013).

różnych wariantów zapisu tego samego słowa: *chat / czat*, ale także do badań nad częstością zapożyczeń w ogóle czy znaczeń, w jakich się pojawiają). Inne przykłady badań, gdzie własnoręcznie zbudowany korpus byłby bardzo przydatny, to różnego rodzaju badania pragmatyczne, np. analiza zwrotów grzecznościowych w Internecie (np. na powitanie nowego użytkownika na forum), badania morfologiczne (np. nad użyciem cząstki *e-*, z ang. *electronic*), czy składniowe (np. nad szykiem przymiotnika: *wirtualna rzeczywistość / rzeczywistość wirtualna*). Celem niniejszego artykułu jest zatem omówienie teoretyczne problematyki związanej z budową własnego korpusu badawczego na przykładzie korpusu opartego o teksty zaczerpnięte z Internetu. Autor stoi bowiem na stanowisku, że chcąc np. badać najnowsze słownictwo, warto przede wszystkim sięgać właśnie do języka Internetu. Szczególnie dobrym wyborem, jak się wydaje, będzie budowa korpusu złożonego z tekstów internetowych pozyskanych z forów internetowych. Najpierw jednak, celem porównania, zostaną krótko przedstawione inne metody pozyskiwania danych językowych (np. wykorzystanie prasy, nagrania języka mówionego, wykorzystanie Internetu jako całości, tj. bez zastosowania metodologii badań korpusowych).

2. Język prasy

Dawniej za najlepszy materiał badawczy (służący m.in. do zbierania nowych jednostek językowych) była uważana prasa (chodziło tutaj naturalnie o prasę drukowaną; dziś, co warto zaznaczyć, termin ten obejmuje również prasę internetową). Opinia taka była formułowana przez licznych badaczy, por. np. słowa Smółkowej:

> Dla językoznawcy najlepszym źródłem pozyskiwania nowych wyrazów jest prasa. Cechy charakterystyczne dla tekstów pisanych, różne od przypisywanych tekstom mówionym i mieszanym, stwarzają dogodne warunki tworzenia nowych wyrazów.
>
> (Smółkowa 2000: 67)

Również w swojej późniejszej publikacji Smółkowa argumentuje, że „dramatycznie spada czytelnictwo książek" (2010: 5), a „teksty prasowe – być może – stają się lub nawet już są jedynymi dłuższymi tekstami pisanymi powszechnie czytanymi przez młodych dobrowolnie, bez przymusu" (Smółkowa 2010: 5). W efekcie, jak pisze autorka

> […] przede wszystkim teksty prasowe powinny być brane pod uwagę przy badaniu funkcjonowania normy gramatycznej i leksykalnej, określaniu stopnia innowacyjności w zakresie reguł gramatycznych oraz rodzaju i zasięgu zmian zachodzących w zasobie leksykalnym.
>
> (Smółkowa 2010: 5)

> Materiały prasowe są niezastąpionym źródłem poznania sposobów pomnażania leksyki. Są nimi: neologizmy słowotwórcze mające różną strukturę, neofrazeologizmy, neosemantyzmy, zestawienia oraz pożyczki.
>
> (Smółkowa 2010: 9)

Również inni językoznawcy podnosili takie argumenty, jak fakt, że prasa dobrze odzwierciedla nowości w języku, cechuje ją różnorodność tematyczna i stylistyczna, a także tani i łatwy dostęp. W efekcie pozwala ona na szybkie zgromadzenie wielu nowych jednostek leksykalnych (Deignan 2005, Krok 2011, Majkowska i Satkiewicz 1999, Mańczak-Wohlfeld 1994).

Takie podejście, niewątpliwie słuszne jeszcze 20 lat temu, dzisiaj nie wydaje się już jednak właściwe. Wybór prasy jako podstawy badań może być kuszący ze względów praktycznych: warto w tym kontekście przytoczyć opinię Algeo (1993), który pisze, że badacze często decydują się na badanie języka prasy bądź literatury nie dlatego, że źródła te są jakoś szczególnie odpowiednie, ale raczej dlatego, że tego typu materiał jest stosunkowo łatwy do zebrania i analizy, w odróżnieniu od np. języka potocznego. Tym niemniej, język prasy może być ciągle dobrym wyborem w przypadku np. badań nad stylistyką języka, ale najprawdopodobniej przestaje takim być w wypadku chęci tworzenia list i analizy najnowszych jednostek leksykalnych. Bardzo zawężony jest tutaj krąg autorów, a przecież istotne jest nie tylko to, jak piszą czy mówią ludzie zawodowo zajmujący się słowem (a zatem m.in. dziennikarze), ale także – a może wręcz przede wszystkim – tzw. zwykli ludzie, niezajmujący się językiem czy obróbką tekstów zawodowo. Z tego samego powodu prasa zupełnie nie spełnia swojej roli w wypadku chęci przeprowadzenia badań np. nad językiem młodego pokolenia.

Co więcej, obecnie bardzo wyraźnie spada czytelnictwo prasy drukowanej[2], a jej rolę przejmuje Internet, stając się podstawowym medium dla młodego pokolenia. Jak słusznie twierdzi Bugajski:

[2] Zob. np. https://www.tvp.info/12096519/najciekawsze-materialy/spada-czytelnictwo-prasy/, https://www.polskieradio.pl/130/2787/Artykul/1496227, Spada-czytelnictwo-prasy-papierowej, http://www.wirtualnemedia.pl/artykul/prasa-na-swiecie-traci-spadek-czytelnictwa-o-25-proc-w-4-lata (data dostępu do wszystkich stron internetowych wymienionych w artykule: 2 kwietnia 2019).

> […] trzeba zauważyć, że w ciągu kilkudziesięciu ostatnich lat kilkakrotnie zmienił się główny kanał publicznego komunikowania i co za tym idzie, zasadniczy nośnik języka. Najpierw – w dużym uproszczeniu – była to przede wszystkim literatura piękna […], następnie media, zwłaszcza wysokonakładowa prasa, a w dalszej kolejności radio i telewizja […] W ostatnich czasach język i kultura kształtują się w Internecie, będącym swoistym metamedium, łączącym w sobie wszystkie sposoby i środki komunikowania i stanowiącym najważniejsze źródło informacji.
>
> (Bugajski 2015: 68)

Wydaje się zatem, że prasa nie jest już najlepszym wyborem w przypadku chęci badania najnowszych zjawisk leksykalnych.

3. Język mówiony

Innym możliwym rozwiązaniem jest badanie języka mówionego. Niestety badania takie – co autor niniejszego artykułu może potwierdzić z własnego doświadczenia (Zabawa 2012) – są bardzo trudne do przeprowadzenia, przede wszystkim ze względu na konieczność zebrania odpowiedniej liczby próbek spontanicznego, potocznego języka mówionego (nie nadają się tutaj np. przemówienia sejmowe, które często mają charakter wystąpień przygotowanych wcześniej, a także – z tych samych powodów – dialogi filmowe czy serialowe, wywiady itp.), a następnie dokonania jego transkrypcji (co nie jest proste w sytuacji, gdy np. jednocześnie mówi kilka osób, co często zdarza się w sytuacjach nieformalnych). Problem stwarza tutaj, co trzeba wyraźnie podkreślić, także samo zebranie materiału językowego, gdyż proces ten rodzi liczne wątpliwości o charakterze prawno-etycznym: badacz staje choćby przed koniecznością podjęcia decyzji, czy zbieranie materiału, a zatem nagrywanie rozmów, powinno odbywać się otwarcie, czy raczej potajemnie. W pierwszym przypadku może dojść do wykrzywienia wyników badania (słynny „efekt obserwatora", polegający na tym, że osoby badane zachowują się nienaturalnie, gdy wiedzą, że są przedmiotem badania, a ich wypowiedzi są nagrywane), w drugim rodzą się liczne wątpliwości o charakterze etycznym, a być może także prawnym (zob. np. artykuły Chomczyńskiego 2006 czy Bartłomiejczyk 2012).

Co jednak istotniejsze, w niektórych sytuacjach badania języka mówionego nie przyniosą oczekiwanych rezultatów: o ile tego typu badanie może być dobrym wyborem np. w wypadku chęci analizy strategii komunikacyjnych czy sposobów przekonywania kogoś do swego zdania, to już w wypadku chęci zebrania nowych jednostek leksykalnych badanie takie niezbyt się sprawdzi. Język mówiony, zwłaszcza potoczny, cechuje się bowiem częstymi powtórzeniami, niskim zagęszczeniem nowych informacji itp. (zob. opis najistotniejszych cech języka

mówionego, Zabawa 2012: 18-23), a – co za tym idzie – stosunkowo niewiele znajdziemy tam zapożyczeń czy neologizmów, co potwierdzają wyniki badań na własnoręcznie tworzonych korpusach języka mówionego (Otwinowska-Kasztelanic 2000, Zabawa 2012). Innymi słowy, aby móc efektywnie badać nowe jednostki leksykalne w języku mówionym, potrzeba by zapewne bardzo dużego korpusu języka mówionego, zaś zebranie takiego – jak przedstawiono w poprzednim akapicie – jest czaso-, praco- i kosztochłonne, a w efekcie może się okazać, że wkład włożonej pracy jest niewspółmiernie wysoki w stosunku do uzyskanych wyników.

4. Narodowy Korpus Języka Polskiego (NKJP)

Biorąc pod uwagę trudności w badaniu języka mówionego i małą reprezentatywność języka prasy (w wypadku badania nowych jednostek leksykalnych), wydaje się, że najlepszym wyjściem będą badania nad językiem Internetu. Trzeba tutaj naturalnie podkreślić, że badania takie nie powinny odbywać się na zasadzie przypadkowego (losowego) czytania różnych tekstów i np. wyszukiwania nowych jednostek leksykalnych, bo byłoby to działanie metodologicznie niepoprawne (notabene ta sama uwaga dotyczy naturalnie badań opartych na języku prasy); w takim wypadku można bowiem oczywiście stwierdzić istnienie danego neologizmu czy zapożyczenia, ale niewiele można powiedzieć np. o częstości jego występowania. Najlepszym wyjściem będzie zatem wykorzystanie korpusu językowego.

Można naturalnie wykorzystać gotowy korpus polszczyzny (NKJP), który zawiera m. in. komponent internetowy (tj. teksty zaczerpnięte z Internetu, np. z forów i list dyskusyjnych). Komponent ten jest jednak stosunkowo niewielki, ponadto korpus ten nie nadaje się do badania słownictwa specjalistycznego, czyli np. neologizmów w języku informatyki czy medycyny.

Co jednak istotniejsze, NKJP nie udostępnia pełnych tekstów; nie jest to naturalnie krytyką NKJP, lecz raczej immanentną cechą korpusów, por. artykuł autorstwa Kuratczyk (2006: 70-71), która pisze o braku wglądu do pełnych wersji tekstów jako wręcz jednej z cech korpusu językowego (w odróżnieniu od korpusu tekstowego, będącego bardziej wirtualną czytelnią, oraz korpusu sieciowego, tworzonego przez wszystkie teksty opublikowane w Internecie, a zatem Internetu jako całości). Trzeba zresztą dodać, że – biorąc pod uwagę wielkość NKJP – udostępnianie pełnych wersji tekstów (pomijam tutaj kwestie związane z prawami autorskimi) nie byłoby celowe, gdyż ręczny dostęp do tak olbrzymiego zbioru tekstów byłby mało praktyczny, por. też artykuł Świdzińskiego i Rudolfa (2006: 31).

W efekcie NKJP nadaje się przede wszystkim do takich badań, gdzie forma wyrazu czy wyrażenia jest znana (a zatem nie do wyszukiwania nowych jednostek leksykalnych). Można

naturalnie wykorzystać któryś z dostępnych analizatorów morfologicznych (np. Morfeusz; http://sgjp.pl/morfeusz/) i przeszukiwać formy nierozpoznane przez analizator. Użycie tego typu analizatora nie rozwiązuje jednak najważniejszego problemu: od roku 2012 korpus nie jest uaktualniany, a zatem siłą rzeczy nie będzie zawierał najnowszych zapożyczeń czy neologizmów. W moim przekonaniu dowodzi to ostatecznie, że korpus ten nie nadaje się do badań nad najnowszymi jednostkami leksykalnymi (w tym miejscu ponownie chcę podkreślić, że nie jest to krytyką NKJP, lecz raczej zwróceniem uwagi na fakt, że nie każdy korpus językowy nadaje się do każdego typu badań). Nie oznacza to jednak, że korpus ten jest całkowicie nieprzydatny: wydaje się, że może być dobrym korpusem służącym do celów porównawczych. Innymi słowy, za pomocą NKJP możemy np. sprawdzić, czy znaleziona w innym źródle jednostka jest istotnie nowością leksykalną w języku (jeśli dana forma nie występuje w pełnym NKJP i nie ma charakteru wybitnie specjalistycznego, to ze sporym prawdopodobieństwem można założyć, że tak właśnie jest).

5. Język Internetu

Pisząc o badaniu języka Internetu, trzeba na początku zaznaczyć, że jeszcze do niedawna dominowało podejście krytyczne. Badania takie koncentrowały się często na (krytycznym) podkreślaniu jego odrębnego charakteru (w odniesieniu do języka bardziej tradycyjnego), a także rozluźnieniu norm poprawnościowych, np. użycia niestandardowej ortografii czy interpunkcji (zob. np. pracę Godzica (2000: 179), który pisze o „wzgardzie dla norm i reguł języka" czy Dunaja i Mycawki (2009: 71-73), którzy piszą o „rozchwianiu normy językowej" w Internecie czy też „świadomej kontestacji reguł poprawnościowych"). Internet był postrzegany – jak słusznie zauważa Data (2009: 132-133) – jako zagrożenie dla kultury i standardowej odmiany języka (używam tutaj czasu przeszłego, gdyż – jak się wydaje – takie podejście jest dziś coraz rzadsze). Trzeba jednak pamiętać, jak stwierdzają ci sami autorzy, że są także gatunki internetowe, gdzie kryteria poprawnościowe polszczyzny ogólnej są przestrzegane, jak np. internetowe wydania prasy.

O ile rozluźnienie norm poprawnościowych faktycznie mogłoby stanowić problem przy niektórych rodzajach badań, nie niweluje wielu korzyści, jakie niesie ze sobą wykorzystanie języka internetowego do badań językoznawczych, szczególnie nad nowymi jednostkami leksykalnymi, i to nawet pomimo słusznej skądinąd uwagi poczynionej przez Dunaja i Mycawkę, iż teksty internetowe są pełne „okazjonalnych, często dziwacznych neologizmów" (2009: 73).

Internet staje się obecnie dobrem powszechnie dostępnym: według danych GUS-u prawie 72% gospodarstw domowych (i aż 95% przedsiębiorstw) w Polsce ma dostęp do Internetu (Bugajski 2015: 69), zaś według badań CBOS-u (przeprowadzonych w maju 2015) Internet jest używany przez 97% Polaków w wieku 18-24 lat. Gdy weźmiemy pod uwagę poziom wykształcenia, odsetek ten może jeszcze wzrosnąć: wśród uczniów szkół średnich i studentów sięga on aż 99% (Feliksiak 2015; zob. też monografię Zabawy 2017: 98-102). Można zatem z całą pewnością stwierdzić, że wśród młodego pokolenia wykształconych Polaków (a tacy będą mieli największy wpływ na kształt polszczyzny w najbliższych dekadach) korzystanie z Internetu ma charakter absolutnie powszechny.

Badacze podkreślają również, że polszczyzna internetowa ma i będzie miała coraz większy wpływ na język ogólny (można w tym miejscu dodać, co jednak nie zmienia wymowy całości, że istnieją oczywiście pewne zjawiska, które wydają się ograniczone tylko i wyłącznie do potocznego języka internetowego, np. zapożyczenia ortograficzne, zob. Zabawa 2009: 74-75), gdyż „nie ma ona charakteru marginesowego socjolektu, lecz posługuje się nią ogromna i coraz liczniejsza, dynamiczna grupa młodej inteligencji, której zwykle przypisuje się rolę kulturotwórczą" (Libura 2006: 54). To właśnie w takich tekstach zapożyczenia czy neologizmy często pojawiają się najszybciej i przynajmniej część z nich przedostanie się również do języka ogólnego. Z dużą dozą prawdopodobieństwa można też założyć, że liczba nowych jednostek leksykalnych w polszczyźnie internetowej będzie wyższa niż w polszczyźnie ogólnej. Truizmem byłoby w tym miejscu stwierdzenie, że język Internetu staje się dziś coraz częściej przedmiotem badań językoznawczych[3]; dodać można, że z całą pewnością warto go badać.

6. Internet jako korpus

W wypadku badań nad językiem Internetu, najbardziej oczywistym rozwiązaniem wydaje się potraktowanie całości Internetu jako jednego wielkiego korpusu, a wyszukiwarki Google jako wyszukiwarki korpusowej. Takie podejście ma naturalnie pewnie zalety (odpowiednio długie poszukiwania z pewnością mogą zaowocować rzadkimi zapożyczeniami czy neologizmami, których nie znaleźlibyśmy nigdzie indziej), por. wypowiedzi językoznawców:

> […] Internet to największy zbiór tekstów, któremu pod względem wielkości żaden obecny ani najprawdopodobniej przyszły korpus nie jest i raczej nigdy nie będzie w stanie dorównać. Poza tym to właśnie w Internecie najszybciej można wychwycić najnowsze słownictwo, zaobserwować, jak

[3] Ze względu na szczupłość miejsca, rezygnuję tutaj z podawania wyczerpującej literatury przedmiotu, poprzestając jedynie na stwierdzeniu, że jest ona już bardzo bogata.

> wyrazy zapożyczone asymilują się w naszym języku, można niemal na bieżąco obserwować kształtowanie się uzusu.
>
> (Andrzejczuk i Czupryniak 2008: 18-19)

> Mimo oczywistych mankamentów zasoby internetowe umożliwiają jednak przeprowadzenie szeregu ciekawych analiz, szczególnie, że nowe tendencje językowe można łatwiej zaobserwować na podstawie tekstów internetowych niż tekstów publikowanych drukiem.
>
> (Podhajecka 2006: 340)

> […] teksty internetowe warto poddawać analizie, gdyż na ich podstawie można określić nowe tendencje w języku szybciej (wyszukanie materiałów nie jest czasochłonne) i bardziej wiarygodnie (teksty przetwarzane są komputerowo) niż na podstawie tekstów drukowanych czy tradycyjnych korpusów językowych, tworzonych częściowo w oparciu o teksty drukowane.
>
> (Podhajecka 2006: 347)

Mimo oczywistych zalet, badania na całości Internetu (w języku angielskim tego typu podejście określa się jako Web as Corpus) niosą ze sobą wiele niedogodności. Internet nie jest korpusem w ścisłym znaczeniu tego słowa (i nie takie jest jego przeznaczenie). Jest to raczej dosyć przypadkowy, a wręcz bezładny, zbiór tekstów (Andrzejczuk i Czupryniak (2008: 192-193) piszą o zaśmieceniu Internetu „bezwartościowymi quasi-tekstami" oraz o „mnogości nieoznakowanych tekstów o nieznanej proweniencji", zaś Podhajecka (2006: 340) wspomina o internetowym „grochu z kapustą"). Z całą pewnością zbiór ten nie jest też zrównoważony ani reprezentatywny, por. „nie jest to oczywiście korpus tradycyjny, ponieważ nie został stworzony przez żaden zespół, nie opiera się o żadne kryteria projektowe i nie można kontrolować jego zawartości" (Podhajecka 2006: 338). Nic również nie wiadomo o jego strukturze (nie wiadomo np. jaki odsetek stanowią teksty potoczne albo teksty nt. nowych technologii). Podkreśla się też, że język internetowy zaburza tradycyjny podział na język mówiony i język pisany (Baker 2010: 13), a zatem badania języka internetowego najprawdopodobniej nie są w pełni reprezentatywne ani dla języka mówionego, ani dla pisanego.

Wykorzystywanie Internetu jako korpusu językowego niesie ze sobą także inne problemy. Paradoksalnie, bardzo duża liczba wystąpień określonego słowa może powodować, że jakakolwiek sensowna analiza jest niemożliwa; trzeba też zaznaczyć, że wyszukiwarka Google ma tendencję do zawyżania liczby znalezionych wyników, a zatem nie można jej traktować z pełnym zaufaniem (zob. np. Podhajecka 2006: 343, Andrzejczuk i Czupryniak 2008: 194; o

wadach i zaletach wyszukiwarek typu Google szerzej pisze też Kuratczyk 2006: 74-79). Ta sama autorka (a także np. Andrzejczuk i Czupryniak 2008: 192) obszernie pisze o innych trudnościach: zwraca między innymi uwagę na fakt, iż nierzadko te same teksty pojawiają się na wielu stronach, co oczywiście wypacza uzyskane wyniki. Co więcej, określone słowa są sztuczne wstawiane na stronach internetowych, aby podnieść „widzialność" strony przez wyszukiwarki, a tym samym zapewnić jej lepsze pozycjonowanie (czyli wyświetlanie danej strony na górze wyników)[4]. Podkreśla się też, że Internet zawiera nieproporcjonalnie dużą liczbę tekstów nt. szeroko rozumianych nowych technologii, co powoduje, że słowa należące do tego pola semantycznego są nadreprezentowane.

Teksty internetowe, co oczywiste, nie są adnotowane morfologicznie ani składniowo, co znacząco utrudnia (a czasem nawet uniemożliwia) wyszukiwanie np. form odmienionych. Istnieją naturalnie różnego rodzaju analizatory morfologiczne (np. wspomniany wyżej Morfeusz), ale niekoniecznie będą one dobrze sobie radzić z nowymi jednostkami leksykalnymi, np. zapożyczeniami leksykalnymi, o częstokroć jeszcze nie do końca ustabilizowanej pisowni i/lub morfologii (a wtedy każdą potencjalną odmienioną formę trzeba by sprawdzać osobno, co oczywiście byłoby możliwe, aczkolwiek bardzo żmudne i długotrwałe). Wyszukiwarki internetowe nie zostały ponadto zaprojektowane do uwzględniania znaków interpunkcyjnych (co jest znaczną przeszkodą przy wyszukiwaniu np. związków frazeologicznych), a także zupełnie sobie nie radzą z formami wieloznacznymi (co jest dużą trudnością w wypadku badania języków bogatych morfologicznie, np. polszczyzny, natomiast stanowi mniejszy problem w wypadku np. angielszczyzny). Oczywiście nie da się również obliczyć częstotliwości znormalizowanej (ang. *normalized frequency*), gdyż nie znamy liczby słów całego korpusu, tj. – w naszym przypadku – polskiego Internetu (nawet gdybyśmy znali, to i tak taka wiedza byłaby *de facto* bezwartościowa, bo odzwierciedlałaby jedynie chwilę obecną: Internet jest medium dynamicznym, nieustannie pojawiają się nowe teksty, a stare mogą znikać). Jest to jednak zrozumiałe, gdy sobie uświadomimy, że wyszukiwarki nie powstały z myślą o językoznawcach (a zatem ich możliwości są bardzo ubogie w porównaniu do wyszukiwarek korpusowych), a sam Internet nie został stworzony z myślą o badaniach językoznawczych (Andrzejczuk i Czupryniak 2008: 192).

[4] Pomimo tych problemów, wyszukiwarka Google jest jednak dość często wykorzystywana do przeprowadzania badań porównawczych, por. np. artykuł A. Niepytalskiej-Osieckiej (2014), która sprawdza częstotliwość wybranych par zapożyczeń (przyswojonych i nieprzyswojonych, np. *lajk/like, fejk/fake, hejt/hate*) przeprowadzając, używając słów samej autorki, „prosty eksperyment" z użyciem wyszukiwarki Google. Takie podejście nie musi być oczywiście złe, zawsze jednak trzeba mieć w pamięci niereprezentatywność wyników.

7. Własnoręcznie zbudowany korpus

Idealnym zatem wyjściem, choć z całą pewnością nie najprostszym, byłoby stworzenie własnego korpusu badawczego opartego o teksty internetowe (w języku angielskim takie podejście określa się jako Web for Corpus). Korpus taki, skompilowany nierzadko na potrzeby jednego badania, bywa określany mianem korpusu doraźnego lub „domowego" (Pęzik 2013: 46). Jak pisze autor, zaletą takiego korpusu jest fakt, że można go budować w ramach tzw. dozwolonego użytku (a zatem bez różnego rodzaju ograniczeń wynikających z praw autorskich). Nie może być on jednak dystrybuowany (np. w Internecie), właśnie ze względu na prawa autorskie (Pęzik 2013: 46). Nie powinien to być naturalnie zbiór przypadkowo dobranych tekstów. Innymi słowy, ich zbieranie musi opierać się o jakieś uprzednio ustalone kryteria selekcji: pod uwagę trzeba wziąć czas powstania tekstów, ich typy, gatunki, stopień oficjalności/potoczności itp. (zob. artykuł Chapmana 2014: 85).

Adamczyk (2009: 172) wyróżnia dwa podstawowe typy komunikacji internetowej: głosową oraz tekstową. Ta druga jest następnie dzielona na odbywającą się w czasie rzeczywistym oraz asynchroniczną. Nieco inny podział wprowadza Grzenia (2007: 43), który pisze o trzech podstawowych typach komunikacji internetowej (zob. też monografię Zabawy 2017: 18): konwersacyjnym (czatowym, z ang. *chat*), korespondencyjnym (e-mailowym) oraz hipertekstowym. Typ konwersacyjny, zwany także synchronicznym, obejmuje różnego rodzaju rozmowy internetowe odbywające się w czasie rzeczywistym, np. czaty czy dyskusje z użyciem komunikatorów internetowych (np. GaduGadu, Skype). Typ korespondencyjny, zwany również asynchronicznym (lub tekstem konwersacyjnym typu nieaktualnego, Loewe 2006: 98), obejmuje rozmowy internetowe, które, w odróżnieniu od typu konwersacyjnego, nie są prowadzone w czasie rzeczywistym, a zatem nie wymagają jednoczesnej obecności dyskutantów. Wyróżnić tutaj można interakcję za pomocą poczty elektronicznej (e-maili), a także grup i forów dyskusyjnych. Typ ten jest zatem czymś pośrednim pomiędzy rozmową a korespondencją (podczas gdy typ pierwszy jest zadecydowanie bardziej zbliżony do tradycyjnej rozmowy), co czasem bywa odzwierciedlane nawet w tytułach publikacji, zob. np. artykuł autorstwa Sikory (2009) o tytule „E-mail – między listem a rozmową". Typ trzeci, hipertekstowy, obejmuje zasadniczo pozostałe formy komunikacji, nienależące do grup powyższych. Są to zatem przede wszystkim różnego rodzaju strony internetowe, zarówno oficjalne (np. firm, instytucji), jak i te tworzone przez osoby prywatne. Typ ten nie reprezentuje zatem, co należy podkreślić, typowej komunikacji internetowej o charakterze dwustronnym (rozumianej jako sytuację, gdzie nadawca i odbiorca wysyłają naprzemiennie komunikaty, a

zatem nadawca jest również odbiorcą, i odwrotnie). Jest to raczej komunikacja jednostronna (od nadawcy do odbiorcy).

W tym miejscu trzeba dodać, że przedstawiona wyżej klasyfikacja jest już dziś nieco przestarzała: nie uwzględnia ona bowiem np. prasy internetowej (internetowe wersje dzienników, tygodników itp.), która jest czymś pośrednim pomiędzy typem hipertekstowym (artykuły dziennikarzy) a korespondencyjnym (komentarze czytelników pod artykułem, odnoszące się zarówno do artykułu, jak i do siebie nawzajem). Z tego względu można dodać tutaj czwarty typ komunikacji: hipertekstowo-korespondencyjny (mieszany).

Internetową komunikację głosową pomijam tutaj całkowicie, jako bardzo trudną do wykorzystania w badaniach nad językiem, przede wszystkim ze względów praktycznych (trudność w dotarciu do tego typu materiałów). Komunikacja synchroniczna tekstowa również jest trudna do pozyskania, gdyż ma ona głównie charakter prywatny, pomiędzy dwiema osobami. W tym momencie badacz albo byłby zmuszony do pozyskiwania bazy materiałowej na postawie swoich rozmów (prowadzonych np. na czacie) z różnymi użytkownikami Internetu, co jednak niewątpliwie wypaczyłoby wyniki badań, a ponadto rodziłoby rozmaite wątpliwości o charakterze etycznym (czy badacz powinien się ujawniać przed osobą, z którą rozmawia, czy powinien raczej grać rolę przypadkowego internauty zainteresowanego np. poznaniem innej osoby?) (zob. artykuł autorstwa Chomczyńskiego (2006), gdzie można znaleźć bardzo interesujący opis tego typu dylematów etycznych badacza). Z kolei badania nad trzecim typem tekstów (hipertekstowym) byłoby dosyć podobne do badań języka prasy (bardzo ograniczona liczba autorów, nierzadka redakcja tekstu itp.). Nie ulega zatem wątpliwości, że najlepszym wyborem, ze względów praktycznych, ale także etycznych, będzie wybór typu drugiego lub czwartego, a zatem oparcie badania na komunikacji asynchronicznej. Do wyboru mamy zatem fora internetowe, e-maile lub quasi-fora (w formie komentarzy pod artykułami prasowymi).

W tym miejscu trzeba wspomnieć o jednym z podstawowych podziałów korpusów istotnych z punktu widzenia dalszych rozważań: należy bowiem wyraźnie odróżniać korpus ogólny od korpusu specjalistycznego. Korpus ogólny jest zwykle bardzo duży (złożony z wielu milionów, a obecnie nawet miliardów, słów) i zebrany z wielu różnych źródeł (pisemne, ustne, elektroniczne, publiczne, prywatne itp.). Można się spodziewać, że w przyszłości, wraz z rozwojem różnego rodzaju urządzeń do przechowywania danych, korpusy ogólne będą jeszcze większe (dobrze widać ten trend na przykładzie korpusów z rodziny BYU, obecnie znanych jako English Corpora; https://www.english-corpora.org/). Naturalnie wielkość poszczególnych typów tekstów w korpusach ogólnych nie jest taka sama: ze względów technicznych, prawnych i praktycznych zwykle niewiele jest np. spontanicznych tekstów mówionych. Baker (2010: 14)

pisze wręcz, że korpus prawdziwie ogólny (czyli reprezentatywny dla języka jako całości) w zasadzie nie istnieje, ponieważ każdy korpus jest w jakiś sposób specjalistyczny (np. zawiera 90% tekstów pisanych). W podobnym tonie wypowiada się Kuratczyk (2006: 75), która, pisząc o korpusie języka ogólnego, stwierdza, że w tego typu korpusach nierzadko „pewne typy tekstów (np. mówionych, potocznych) prezentowane są w niewielkim zakresie lub pomijane, inne z kolei uwzględnia się i przedstawia szerzej (zwykle nadreprezentowane są teksty literatury pięknej)". Widać zatem wyraźnie, że podstawą doboru tekstów w korpusach ogólnych często nie jest reprezentatywność, lecz raczej łatwość zebrania materiału. W efekcie, jak konstatuje Chapman (2014), być może korpusy w ogóle niezbyt nadają się do formułowania sądów o języku ogólnym, głównie ze względu na brak np. odpowiednio dużego komponentu mówionego, czy w ogóle prawdziwie potocznego i spontanicznego języka, niekoniecznie mówionego.

Korpusy specjalistyczne, jak się często podkreśla (zob. np. Baker 2010: 13-14), nie muszą być bardzo duże, aby być reprezentatywnymi dla określonej odmiany języka (najczęściej określonego socjolektu lub profesjolektu). Dla całości obrazu trzeba jednak podkreślić, że niektórzy badacze (np. Hebal-Jezierska 2013: 20) uważają, że ich zazwyczaj niewielki rozmiar powinien być postrzegany jako wada (choć określenie, co jest, a co nie jest dużym korpusem, jest naturalnie w znacznym stopniu subiektywne). Wydaje się jednak, że – co do zasady – korpus specjalistyczny (np. ograniczony do określonego pola semantycznego) nie musi być tak wielki jak korpus ogólny, a relatywnie duża reprezentatywność wydaje się znacznie łatwiejsza do osiągnięcia (pełna reprezentatywność, jak słusznie pisze Chapman (2014: 89), nie jest oczywiście możliwa); łatwiej także o stworzenie kryteriów doboru próbek, które zostaną włączone do korpusu.

Dobrym pomysłem będzie zatem ograniczenie badań do określonego pola semantycznego, np. badania nad językiem ludzi interesującym się kolarstwem, wspinaczką górską, komputerami czy koleją. Pożądaną sytuacją jest, gdy analizowane pole semantyczne należy do zainteresowań badacza (zob. np. Bartłomiejczyk 2012: 194). Jak zaznaczono wyżej, zdecydowanie łatwiej – z różnych względów – zbudować korpus o charakterze specjalistycznym (ograniczonym do wybranego pola semantycznego), niż korpus ogólny. Samodzielne zbudowanie tego drugiego byłoby bardzo trudne ze względu na konieczność zebrania bardzo dużej liczby tekstów; wymagałoby to zatem zaangażowania wielu osób w tego typu projekt.

Wracając do przedstawionej wyżej klasyfikacji typów komunikacji internetowej, wydaje się, że najlepszym wyborem będzie oparcie badania na języku forów internetowych[5]. Badania języka e-maili obarczone są bowiem podobnymi trudnościami, co badania języka czatów – mają one charakter prywatny, co rodzi problemy w dotarciu do takich tekstów, zaś w wypadku prasy internetowej komentarze są często nie na temat (bardzo często obserwuje się np. zjawisko przekierowywania dyskusji na tematy polityczne, pomimo, że artykuł wyjściowy nie był z nimi tematycznie związany), co utrudnia zbudowanie spójnego tematycznie korpusu i ograniczenie go do określonego pola semantycznego. Co więcej, o ile łatwo znaleźć artykuły prasowe z komentarzami internautów na tematy społeczno-polityczne (co oczywiście determinuje charakter komentarzy), o tyle sytuacja wygląda znacznie trudniej w wypadku bardziej specjalistycznych pól semantycznych.

Przed rozpoczęciem budowy korpusu konieczne jest podjęcie decyzji co do sposobu zbierania tekstów z forów internetowych: można to wykonywać albo ręcznie (za pomocą opcji kopiuj-wklej), albo automatycznie (funkcję umożliwiającą automatyczne „ściąganie" zawartości strony posiada np. bezpłatny program TextSTAT). Każda z tych metod ma swoje wady i zalety. Ręczne kopiowanie jest żmudne i czasochłonne; kopiowanie automatyczne może natomiast powodować, że kopiują się także różnego rodzaju dane, których nie potrzebujemy (np. cytaty poprzedniego użytkownika, informacje techniczne itp.), które następnie należy usunąć, co również może być pracą czasochłonną (i to pomimo istnienia programów, które mogą być pomocne przy tego rodzaju „czyszczeniu" materiału). Jak się wydaje, wybór metody powinien zależeć tutaj przede wszystkim od planowanej wielkości korpusu (a także wielkości próbek: im próbki mniejsze, tym metoda ręczna jest bardziej czasochłonna).

Aby analizować nowe jednostki leksykalne, badacz musi mieć dostęp do pełnych zbiorów tekstów, co oczywiście umożliwia własnoręcznie stworzony korpus. Konieczne jest również posiadanie odpowiedniego oprogramowania, które umożliwi np. wyszukiwanie kolokacji, podawanie częstotliwości wystąpień danej konstrukcji itp. Teoretycznie można sobie wyobrazić pracę jedynie z edytorem tekstu Microsoft Word, który jest wyposażony w funkcję zliczania wyrazów czy wyszukiwania odpowiednich fraz w tekście. Mogę jednak stwierdzić z własnego doświadczenia, że edytor ten niezbyt dobrze sobie radzi z bardzo dużymi zbiorami tekstów i niespecjalnie nadaje się do tego typu prac. Znacznie lepiej zdecydować się zatem na użycie specjalistycznego oprogramowania. Jak wskazują moje doświadczenia, nie zawsze istnieje

[5] Ze względu na brak miejsca rezygnuję tutaj ze szczegółowej charakterystyki forów internetowych. Więcej informacji na ww. temat można znaleźć np. w pracach A. Naruszewicz-Duchlińskiej (2009, 2011), artykule M. Karwatowskiej i B. Jarosz (2013), a także w monografii M. Zabawy (2017: 25-28).

bezwzględna potrzeba korzystania z komercyjnych (płatnych) programów do obsługi korpusów, takich jak np. WordSmith (http://www.lexically.net/wordsmith/). Do prostych badań leksykalnych wystarczające są programy bezpłatne, takie jak np. AntConc (http://www.laurenceanthony.net/software.html) czy – o stosunkowo niewielkich możliwościach, lecz bardzo prosty w obsłudze – TextSTAT (dostępny również w polskiej wersji językowej; http://neon.niederlandistik.fu-berlin.de/en/textstat/).

W tym miejscu koniecznie trzeba wspomnieć o innym, być może najbardziej przydatnym narzędziu wspomagającym i ułatwiającym tworzenie własnego korpusu językowego. Chodzi tutaj o aplikację Korpusomat (http://korpusomat.pl/; zob. Kieraś, Kobyliński i Ogrodniczuk 2018). Jest ona bardzo prosta w użyciu: użytkownik wskazuje pliki (program jest bardzo uniwersalny w tym zakresie, gdyż może współpracować z wieloma formatami plików tekstowych, m. in. txt, doc, docx, pdf, epub czy mobi), które mają wejść w skład korpusu, a program dokonuje ich automatycznej konwersji (kompilacji do postaci binarnej) i – co bardzo ważne – adnotacji fleksyjnej. Korpusomat oferuje ponadto różne inne przydatne funkcje, takie jak tworzenie listy frekwencyjnej, tworzenie listy charakterystycznego słownictwa i kolokatów, a także słów kluczowych. Co istotne, program zawiera także dość szczegółową instrukcję obsługi zawierającą m.in. opis składni zapytań.

Jednak podstawową wadą Korpusomatu jest fakt, że program niezbyt dobrze sobie radzi z rozpoznawaniem form fleksyjnych nowych jednostek leksykalnych (neologizmów, zapożyczeń itp.). Autor artykułu stworzył próbny korpus korzystając z tego narzędzia (korpus zawierał posty o tematyce komputerowej zebrane z wybranych forów internetowych). O ile np. polecenie wyszukania form odmienionych słowa *komputer* zadziałało bardzo dobrze, o tyle np. anglicyzm leksykalny *boxowy* (‘dostarczany w pudełku’, np. *wersja boxowa*) będący z całą pewnością przymiotnikiem, został zakwalifikowany przez analizator jako rzeczownik w mianowniku liczby mnogiej. Trzeba jednak dodać, że wiele anglicyzmów, także tych stosunkowo nowych (np. *mail*), jest adnotowana poprawnie. Ogółem można zatem powiedzieć, że Korpusomat jest narzędziem bardzo przydatnym przy budowie własnego korpusu, choć może niezbyt dobrze sprawdzać się przy adnotacji najnowszych czy mniej typowych zapożyczeń czy neologizmów. Nie jest to jednak krytyką tego narzędzia, a jedynie wskazaniem, że w pewnych obszarach jego przydatność może być ograniczona.

Kolejną problematyczną sprawą może być określenie docelowej wielkości korpusu. Na wstępie trzeba podkreślić, jak słusznie stwierdza Chapman (2014: 87), że samo pojęcie wielkości korpusu nie zawsze jest w pełni miarodajne i istotniejsze dla jego reprezentatywności mogą okazać się inne wskaźniki, takie jak liczba próbek włączonych do korpusu, średnia

wielkość jednej próbki, wielkość największej/najmniejszej próbki itp. Szczególnie istotne, jak podkreśla Chapman (2014: 89-90), jest uwzględnienie możliwie dużej liczby próbek. Lepiej zatem, jak się wydaje, jest uwzględnić większą liczbę krótszych próbek tekstu (za czym idzie zwykle uwzględnienie większej liczby autorów) niż mniejszą liczbę dłuższych. Zatem fora internetowe ponownie wydają się dobrym wyborem (dominują tam raczej wpisy krótkie; ubocznym skutkiem jest jednak większy nakład pracy, jaki jest potrzebny, aby skompilować korpus złożony z większej liczby tekstów krótkich). Trzeba też podkreślić, na co słusznie zwraca uwagę Mair (2015: 2), że odpowiednia wielkość korpusu zależy też od rodzaju badań, jakie będą na nim prowadzone.

Nie jest zatem łatwym zadaniem ustalenie, kiedy własnoręcznie budowany korpus jest już wystarczająco duży. Trzeba też wziąć pod uwagę fakt, że ciągły rozwój technik komputerowych powoduje, że korpus dawniej uważany za duży będzie dziś co najwyżej przeciętnej wielkości. Dla przykładu można tutaj przytoczyć artykuł Fernandez-Diaz (2008), która, opisując stworzony przez siebie korpus języka polityków UE, podkreśla, że ma on 190.916 słów, co jak stwierdza autorka (choć nie wprost), jest już wystarczającą wielkością według standardów językoznawstwa korpusowego (powołując się na prace językoznawcze opublikowane w 2002 roku). Dzisiaj taka wielkość, nawet w wypadku korpusów specjalistycznych, zostałaby z dużą dozą prawdopodobieństwa uznana za niewystarczającą. Można by w tym miejscu powiedzieć, że im korpus większy, tym lepszy, ale to niekoniecznie jest prawdą: bowiem w miarę powiększania korpusu spada tempo przyrostu nowych jednostek leksykalnych (zjawisko to znane jest jako nasycenie leksykalne; zob. Pęzik 2013: 47-50); innymi słowy, w pewnym momencie „nie opłaca" się już powiększać korpusu, bo – gdy np. celem badania jest znalezienie i analiza najnowszych jednostek leksykalnych – nie przyniesie to wielu nowych neologizmów czy zapożyczeń, a jedynie kolejne wystąpienia jednostek już znanych. Oczywiście nie sposób podać tutaj konkretnej wielkości, bo zależy to od bardzo wielu czynników. Z własnego doświadczenia mogę powiedzieć, że przy własnoręcznie zbudowanym korpusie potocznego języka informatyki nasycenie leksykalne osiągnąłem, gdy wielkość korpusu przekroczyła milion słów (rozumianych ortograficznie, tj. jako szereg liter oddzielonych spacją) (opis korpusu mojego autorstwa można znaleźć w monografii: Zabawa 2017).

Innym ważnym problemem związanym z badaniami korpusowymi jest szybkość starzenia się korpusu. Niestety, jak słusznie pisze Chapman (2014: 88), korpus może być nieaktualny zaraz po zakończeniu jego budowy, jeszcze przed przeprowadzeniem jakichkolwiek badań z jego użyciem. Jest to szczególnie istotne w wypadku badań nad najnowszymi leksemami w języku. Jedynym rozwiązaniem, jeśli badacz jest zainteresowany opisem języka nie tylko w

danym momencie, jest stałe uzupełnianie korpusu o nowe, niedawno stworzone teksty. Warto też w tym kontekście wyrazić ubolewanie, że NKJP nie jest uaktualniany.

Własnoręcznie zbudowany korpus może mieć zatem istotne zalety w stosunku do korpusu „gotowego". Badacz doskonale zna strukturę swojego korpusu i może w pełni kontrolować jego zawartość, znacznie łatwiej też np. oddzielać poszczególne sensy znalezionych wyrazów, gdyż badacz zna pełne konteksty użycia, a także ma w każdej chwili dostęp do pełnych tekstów. Jeszcze do niedawna sporym problemem byłby brak adnotacji składniowo-morfologicznej (np. oznaczanie części mowy), obecnie jednak, dzięki istnienia Korpusomatu, ta niedogodność została już w znacznej mierze naprawiona (choć trzeba naturalnie pamiętać, że na takiej automatycznej adnotacji nie można niestety w pełni polegać[6]). Korzystanie z własnego korpusu ma też naturalnie pewne wady: najistotniejszą wydaje się to, iż w małych korpusach problemem może być niska frekwencja określonych form, co może uniemożliwiać formułowanie wiążących wniosków (Andrzejczuk i Czupryniak 2008: 197). Andrzejczuk i Czupryniak (2008: 203) słusznie konkludują, że warto wykorzystywać Internet (jako całość) w charakterze korpusu uzupełniającego, a zatem własny korpus może stanowić podstawę badań, a całość Internetu – wraz z wyszukiwarką Google – może pełnić rolę uzupełniającą, przydatną np. do sprawdzania użycia form, które pojawiają się ze zbyt małą częstotliwością w korpusie podstawowym.

Pewne problemy rodzić będzie także samo wyszukiwanie form: w wypadku zbierania tekstów z forów internetowych, niewątpliwie kłopotliwa będzie stosunkowo wysoka frekwencja form zapisanych błędnie (chodzi tutaj nawet nie o typowe błędy ortograficzne, których – wbrew pozorom – zwykle nie ma aż tak wiele, ale głównie o różnego rodzaju literówki i brak znaków diakrytycznych)[7]. W wypadku korpusu języka informatyki, liczne będą np. przypadki zapisu *przegladarka* zamiast *przeglądarka*. W tym momencie badacz ma w zasadzie dwa wyjścia: albo poprawiać tego typu zapisy w tekście (co jednak może nie być dobrym wyjściem, bo taki korpus będzie bezużyteczny do badań np. nad odstępstwami od normy ortograficznej w tekstach internetowych), albo, konstruując zapytania, pamiętać o tego typu formach: chcąc zatem sprawdzić częstotliwość formy *przeglądarka* warto wykorzystać znak gwiazdki (*), który w większości programów do obsługi korpusów zastępuje dowolny ciąg znaków, i sformułować

[6] W tym miejscu można dodać, że rezygnacja z budowy własnego korpusu i korzystanie z profesjonalnie adnotowanych korpusów również nie oznacza, że wyniki można przyjmować w pełni bezkrytycznie. O takiej „zasadzie ograniczonego zaufania", z ciekawymi przykładami, piszą Piotrowski i Grabowski (2013).

[7] Na temat błędów w języku internetowym napisano bardzo wiele prac i niepodobna podać tutaj wyczerpującej listy: informacje na ww. temat, wraz z licznymi przykładami, można znaleźć – wymieniając kilka przykładów nowszych publikacji – w artykułach autorstwa Pachowicz (2012), Szymańskiego (2012), Urzędowskiej (2015) czy Zabawy (2014).

dwa zapytania: *przeglądar** i *przegladar** (dzięki takiemu zapytaniu otrzymamy również formy odmienione i derywaty). Trzeba wyraźnie podkreślić, że to niestety nie rozwiązuje problemu całkowicie (nadal mogą istnieć bowiem w korpusie formy typu *przeglądakra* czy *przeglądaka*). Tak więc uzyskane wyniki należy traktować jedynie orientacyjnie, nigdy zaś jako ostatecznego i niepodważalnego dowodu (uwaga ta dotyczy zresztą wszelkich badań korpusowych).

Jak już wspomniano wyżej, pełny opis korpusu mojego autorstwa, wraz ze szczegółowym opisem poszczególnych etapów procesu jego tworzenia można znaleźć w monografii mojego autorstwa (Zabawa 2017). Stworzony przeze mnie korpus (jego podstawą są teksty zebrane z 32 wybranych forów internetowych poświęconych komputerom i Internetowi) został wykorzystany do badań nad zapożyczeniami semantycznymi i kalkami w języku informatyki. Korpus tego typu może być przydatny do różnych aspektów badań nad zapożyczeniami i neosemantyzmami (czy w ogóle nowymi jednostkami leksykalnymi): przede wszystkim, jak się wydaje, jest to bardzo wygodne i rzetelne narzędzie do określania częstości zapożyczeń w języku, ich stopnia asymilacji (zwłaszcza na poziomie ortograficznym i morfologicznym), znaczeń, w jakich się pojawiają w języku (wraz z uszeregowaniem ich, w wypadku zapożyczeń polisemicznych, według częstości występowania), a także badania ich łączliwości frazeologicznej.

8. Podsumowanie

Językoznawca chcący badać np. najnowsze zjawiska leksykalne w języku (neologizmy i zapożyczenia) ma kilka możliwości: po pierwsze, mogą to być badania prowadzone niejako przy okazji lektury (gazet, forów internetowych, szyldów ulicznych itp.). Niestety badania takie mają charakter bardzo przypadkowy i nie sposób wtedy sformułować jakichkolwiek wniosków dotyczących np. frekwencji omawianych form czy kontekstów, w jakich się pojawiają. Po drugie, można wykorzystać gotowy korpus (NKJP). Korpus taki jednak, jak wykazano w tekście, niezbyt nadaje się do badań nad najnowszymi zjawiskami leksykalnymi. Po trzecie, można wykorzystać całość Internetu jako jeden wielki korpus, a wyszukiwarkę internetową jako wyszukiwarkę korpusową. Także jednak i to podejście ma rozliczne wady. Pozostaje wreszcie droga czwarta: budowa własnego korpusu. Jest to rozwiązanie najbardziej praco- i czasochłonne, ale i częstokroć dające najlepsze wyniki. W wielu wypadkach warto taki wysiłek podjąć.

Skróty nazw korpusów

COCA – Corpus of Contemporary American English, https://www.english-corpora.org/coca/.

NKJP – Narodowy Korpus Języka Polskiego, http://nkjp.pl/.

Bibliografia

Adamczyk, Małgorzata Joanna (2009) „Język sieciowych dyskusji w opiniach samych dyskutantów". [W:] Danuta Ulicka (red.), *Tekst (w) sieci. Tom 1: Tekst, Język, Gatunki*. Warszawa: Wydawnictwa Akademickie i Profesjonalne; 171–184.

Algeo, John (1993) *Fifty Years Among the New Words: A Dictionary of Neologisms, 1941−1991*. Cambridge: Cambridge University Press.

Andrzejczuk, Anna, Maciej Czupryniak (2008) „O wykorzystaniu zasobów internetowych w pracy językoznawcy". *Polonica XXIX*; 189−204.

Baker, Paul (2010) *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Bartłomiejczyk, Magdalena (2012) „O języku kierowców autokarów". *Socjolingwistyka XXVI*; 191–204.

Bugajski, Marian (2015) „Kultura języka w Internecie". *Poradnik Językowy 9*; 68–80.

Chapman, Richard (2014) „«Small is Beautiful?» Shakespeare's Sonnets as a Linguistic Corpus". *Litteraria Pragensia: Studies in Literature and Culture, 24/47*; 84–96.

Chomczyński, Piotr (2006) „Wybrane problemy etyczne w badaniach. Obserwacja uczestnicząca ukryta". *Przegląd Socjologii Jakościowej 2/1*; 68–87 [pobrane z: http://www.qualitativesociologyreview.org/PL/Volume2/PSJ_2_1_Chomczynski.pdf. Data ostatniego dostępu: 02.04.2019].

Crystal, David ([1980] 2008) *A Dictionary of Linguistics and Phonetics, 6th edition*. Oxford: Blackwell Publishing.

Data, Krystyna (2009) „Wpływ komunikacji sieciowej na współczesną polszczyznę". [W:] Danuta Ulicka (red.), *Tekst (w) sieci. Tom 1: Tekst, Język, Gatunki*. Warszawa: Wydawnictwa Akademickie i Profesjonalne; 131−138.

Deignan, Alice (2005) *Metaphor and Corpus Linguistics*. Amsterdam − Philadelphia: John Benjamins.

Dunaj, Bogusław, Mirosława Mycawka (2009) „Norma i kodyfikacja". [W:] Anna Piotrowicz, Krzysztof Skibski, Michał Szczyszek (red.), *Kształtowanie się wzorów i wzorców językowych*. Poznań: Wydawnictwo Poznańskie; 67−75.

Feliksiak, Michał (2015) *Komunikat z badań CBOS nr 90/2015. Internauci 2015* [pobrane z: http://www.cbos.pl/SPISKOM.POL/2015/K_090_15.PDF. Data ostatniego dostępu: 02.04.2019].

Fernandez-Diaz, Gabriela (2008) „A Political Language Corpus and its Applications in Language Teaching". *Respectus Philologicus, 1*; 46–57.

Godzic, Wiesław (2000) „Język w Internecie: Czy piszemy to, co myślimy?". [W:] Jerzy Bralczyk, Katarzyna Mosiołek-Kłosińska (red.), *Język w mediach masowych*. Warszawa: Oświata UN-O; 176−185.

Grzenia, Jan (2007) *Komunikacja językowa w Internecie*. Warszawa: Wydawnictwo Naukowe PWN.

Hebal-Jezierska, Milena (2013) „Podstawowe zasady korzystania z korpusów przy badaniu języka". [W:] Wojciech Chlebda (red.), *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*. Opole: Wydawnictwo Uniwersytetu Opolskiego; 17–30.

Karwatowska, Małgorzata, Beata Jarosz (2013) „Forum internetowe, czyli (cyber)komunikacja o ograniczonym zasięgu społecznym". *Polonica XXXIII*; 109–121."Korpusomat – a Tool for Creating Searchable Morphosyntactically Tagged Corpora". *Computational Methods in Science and Technology, 24(1)*; 21–27.

Krok, Iwona (2011) "Selected English Borrowings in Popular Contemporary Russian Press on the Example of *Ogonyok*". [W:] Ewa Willim (red.), *English in Action. Language Contact and Language Variation*. Kraków: Krakowskie Towarzystwo Edukacyjne – Oficyna Wydawnicza AFM; 43−52.

Kuratczyk, Magdalena (2006) „Narzędzia korpusowe w leksykografii dwujęzycznej". *Biuletyn Polskiego Towarzystwa Językoznawczego LXII*; 69−81.

Libura, Agnieszka (2006) „Internet. Między wielokulturowością a globalizacją kultury" [W:] Anna Dąbrowska, Anna Burzyńska-Kamieniecka (red.), *Wielokulturowość w języku [Język a kultura 18]*. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego; 45−58.

Loewe, Iwona (2006) „Internet i jego zasoby w polskich badaniach lingwistycznych. Rekonesans". *Biuletyn Polskiego Towarzystwa Językoznawczego LXII*; 93−103.

Mair, Christian (2015) "Parallel Corpora. A Real-time Approach to the Study of Language Change in Progress". *Diacronia 1*; 1–9. DOI: 10.17684/i1A6en.

Majkowska, Grażyna, Halina Satkiewicz (1999) „Język w mediach". [W:] Walery Pisarek (red.), *Polszczyzna 2000. Orędzie o stanie języka na przełomie tysiącleci*. Kraków: Ośrodek Badań Prasoznawczych; 181−196.

Mańczak-Wohlfeld, Elżbieta (1994) *Angielskie elementy leksykalne w języku polskim*. Kraków: Universitas.

Naruszewicz-Duchlińska, Alina (2009) „Internetowe grupy dyskusyjne. Wstępna charakterystyka gatunku". *Język Polski LXXXIX 3*; 191−198.

Naruszewicz-Duchlińska, Alina (2011) *Internetowe grupy dyskusyjne. Analiza językowa i charakterystyka gatunku*. Olsztyn: Wydawnictwo Uniwersytetu Warmińsko-Mazurskiego.

Niepytalska-Osiecka, Anna (2014) „O *fejku, lajku* i *hejcie* w polszczyźnie internetowej". *Język Polski XCIV 4*; 343–352.

Otwinowska-Kasztelanic, Agnieszka (2000) *A Study of the Lexico-semantic and Grammatical Influence of English on the Polish of the Younger Generation of Poles (19-35 Years of Age).* Warszawa: Wydawnictwo Akademickie Dialog.

Pachowicz, Małgorzata (2012) „W (nie)zgodzie z normą językową w portalach internetowych". *Język Polski XCII 1*; 29−36.

Pęzik, Piotr (2013) „Wybrane aspekty reprezentatywności małych i średnich korpusów". [W:] Wojciech Chlebda (red.), *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*. Opole: Wydawnictwo Uniwersytetu Opolskiego; 45−58.

Piotrowski Tadeusz, Łukasz Grabowski (2013) „Interpretacja danych frekwencyjnych z korpusów językowych: opis pewnych problemów (na kilku przykładach z życia wziętych)". [W:] Wojciech Chlebda (red.), *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*. Opole: Wydawnictwo Uniwersytetu Opolskiego; 59−71.

Podhajecka, Mirosława (2006) „Kilka uwag o wykorzystaniu zasobów internetowych do analiz korpusowych języka". *Język Polski LXXXVI 5*; 338−347.

Sikora, Agata (2009) „E-mail – między listem a rozmową". [W:] Danuta Ulicka (red.), *Tekst (w) sieci. Tom 1: Tekst, Język, Gatunki*. Warszawa: Wydawnictwa Akademickie i Profesjonalne; 245–252.

Smółkowa, Teresa (2000) „Nowe słownictwo w prasie". [W:] Jerzy Bralczyk, Katarzyna Mosiołek-Kłosińska (red.), *Język w mediach masowych*. Warszawa: Oświata UN-O; 67−78.

Smółkowa, Teresa (2010) „Prasa jako źródło wiedzy o języku". *Poradnik Językowy 5*; 5–14.

Szymański, Leszek (2012) „Konwencje zapisu wyrazów na czacie internetowym". *Język Polski XCII, 1*; 20−28.

Świdziński, Marek, Michał Rudolf (2006) „Narzędzia informatyczne obsługi wielkich korpusów tekstów: wyszukiwarka Holmes". *Biuletyn Polskiego Towarzystwa Językoznawczego LXII*; 31−43.

Urzędowska, Aleksandra (2015) „Poprawność języka w Internecie (na przykładzie facebookowych fanpage'y)". *Poradnik Językowy 9*; 94–104.

Zabawa, Marcin (2009) „«My blogasek bierze udział w konQursie» – czy polskie blogi internetowe są pisane po polsku?". [W:] Mirosław Filiciak, Grzegorz Ptaszek (red.), *Komunikowanie (się) w mediach elektronicznych – język, edukacja, semiotyka*. Warszawa: Wydawnictwa Akademickie i Profesjonalne; 60–78.

Zabawa, Marcin (2012) *English Lexical and Semantic Loans in Informal Spoken Polish*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Zabawa, Marcin (2014) „Subkultura Internetu: język internetowy najmłodszego pokolenia". [W:] Joanna Bierówka, Katarzyna Pokorna-Ignatowicz (red.), *Media – kultura popularna – polityka. Wzajemne oddziaływania i nowe zjawiska*. Kraków: Krakowskie Towarzystwo Edukacyjne – Oficyna Wydawnicza AFM; 223–244.

Zabawa, Marcin (2017) *English Semantic Loans, Loan Translations, and Loan Renditions in Informal Polish of Computer Users*. Katowice: Wydawnictwo Uniwerystetu Śląskiego.