RAFAŁ L. GÓRSKI

Charakterystyka chronologiczna i stylistyczna korpusu dla *Wielkiego słownika języka polskiego*

1. Pojęcia reprezentatywności i zrównoważenia korpusu

W kontekście tworzenia korpusów i badań korpusowych z reguły padają dwa określenia-klucze (albo: wytrychy), mianowicie reprezentatywność i zrównoważenie. Nestor badań korpusowych John Sinclair napisał na ten temat wprost (Sinclair 1996): „Questions of balance and representativeness recur in the discussion of reference corpora. They are extremely difficult to define, and yet fairly easy to work with”[1]. Co to znaczy, że korpus jest reprezentatywny? Albo lepiej: co korpus reprezentuje? Od razu możemy odrzucić pierwszą nasuwającą się odpowiedź: język. Korpus nie może reprezentować języka, ponieważ ten jest po części bytem abstrakcyjnym, dyspozycją psychiczną. Korpus nie reprezentuje bezpośrednio kompetencji językowej czy saussurowskiego langue[2]. Korpus jest zbiorem tekstów, a więc reprezentuje parole. Należy przeto oczekiwać, że będzie reprezentował populację tekstów.

Na początek zróbmy pewne zastrzeżenie: na co dzień percypujemy przede wszystkim teksty mówione. Nie znam żadnych badań dotyczących tego, ile słów wypowiadanych słyszy przeciętny użytkownik języka w stosunku do liczby słów przeczytanych, intuicyjnie jednak można stwierdzić, że przewaga tych pierwszych jest miażdżąca. Tym niemniej stopień trudności w pozyskaniu tekstów sprawia, że zazwyczaj proporcje są dokładnie odwrotne, tzn. teksty mówione nie stanowią więcej niż kilka procent korpusu. Kwestię więc tekstów mówionych poruszę w dalszej części artykułu.

Można sobie wyobrazić kilka sposobów realizacji postulatu „reprezentatywności” korpusu:

(a) włączamy do korpusu dużą liczbę przypadkowo dobranych tekstów;
(b) wyznaczamy pewne kategorie tekstów, a następnie zapełniamy każdą z nich tekstami o jednakowej sumarycznej objętości;

[1] Pytania o zrównoważenie i reprezentatywność powracają przy dyskusjach na temat korpusów referencyjnych. Bardzo trudno je [te pojęcia] zdefiniować i niełatwo z nimi pracować.

[2] Co oczywiście nie znaczy, że nie może on służyć do tego, by badać langue.

(c) staramy się, by korpus ściśle odzwierciedlał populację tekstów drukowanych;
(d) staramy się, by korpus odzwierciedlał produkcję tekstów w ramach danej społeczności językowej;
(e) staramy się, by korpus odzwierciedlał percepcję tekstów w ramach danej społeczności językowej.

Ad (a). Jeżeli korpus byłby dostatecznie duży i starano by się pozyskiwać teksty z różnych źródeł, to można by się spodziewać, że taki korpus da szeroką perspektywę języka. Byłaby to jednak nadzieja płonna; podobnie jak badania socjologiczne prowadzone na podstawie losowanych numerów telefonów są najczęściej bardzo mylące, choćby dlatego, że w ciągu dnia w domu przebywają osoby nieczynne zawodowo, a na wsi jest znacznie mniej telefonów niż w mieście. Podobnie w wypadku takiego trybu tworzenia korpusu jak opisany w punkcie (a) do korpusu najprawdopodobniej weszłyby łatwe do pozyskania teksty ściągnięte z Internetu, a nie znalazłyby się w nim dużo trudniejsze do zdobycia teksty rzadkie, np. literackie. Nie oznacza to wszakże, że tak stworzony korpus byłby bezwartościowy. Jeśli byłby naprawdę duży, to mógłby służyć za dodatkowe, kontrolne źródło ujawniające rzadkie słowa czy dające lepszy wgląd w pewne kolokacje. Metodą tą posługują się więc leksykografowie, tworząc tzw. „korpusy monitorowe”[3]. Tego rodzaju korpusy jednak powinny uzupełniać, a nie zastępować korpusy zrównoważone. Nawiasem mówiąc tego rodzaju korpusem jest Internet, zaś rolę programu wyszukiwawczego pełni najczęściej Google.

Metodą (b) posłużyli się twórcy pierwszego polskiego korpusu elektronicznego, mianowicie korpusu SFPW. Dodam, że korpus ten jest współcześnie dostępny w kilku wersjach, w tym w wersji dla Poliqarpa. Tą drogą również zamierzał pójść W. Lubaś, tworząc pierwszy korpus IJP PAN.

Droga ta, jakkolwiek w jakiś sposób metodologicznie podbudowana, też nie wydaje się najlepsza. Przedstawię to za pomocą analogii w badaniach opinii publicznej. Wyobraźmy sobie, że zadajemy jakieś pytanie równej liczbie osób odpowiednio z wykształceniem podstawowym, ponadpodstawowym i wyższym. Tego rodzaju ankieta da skrzywiony obraz społeczeństwa, ponieważ grupa osób ze średnim wykształceniem jest znacznie liczniejsza niż dwie pozostałe. Podobnie tu, pewne typy tekstów będą nadreprezentowane a inne niedoreprezentowane. Jeżeli teksty prasowe, naukowe i literackie mają mieć ten sam udział procentowy, to łatwo postawić zarzut, że tych pierwszych pisze się i czyta znacznie więcej niż pozostałych. Swego rodzaju wariantem tej metody jest arbitralne przypisanie objętości poszczególnym stylom funkcjonalnym. Tą drogą poszli twórcy korpusu PWN i British National Corpus.

Metoda (c) wydaje się mieć najlepsze uzasadnienie metodologiczne. Taki bowiem sposób postępowania korpus rzeczywiście odzwierciedla (= reprezentuje) populację tekstów, szczególnie jeśli się uwzględni ich nakład. O ile jednak zalecałbym tę metodę w odniesieniu do korpusów historycznych, o tyle w wypadku współczesnego języka jest ona obciążona pewną niedającą się pokonać wadą. Otóż współcześnie prasa stanowi przytłaczającą większość tekstów, tak w odniesieniu do liczby „wyprodukowanych”

[3] Ang. *monitor corpus*.

słów, jak i w odniesieniu do nakładów. Udział tekstów pozaprasowych byłby tak minimalny, że ginąłby wśród twórczości dziennikarskiej[4].

Wariantem opisanej wyżej metody jest metoda (d). Różnica jest dość subtelna — tutaj mianowicie populację stanowiliby członkowie jakiejś wspólnoty językowej, a korpus miałby odzwierciedlać proporcje tekstów tworzonych przez statystycznego członka tej wspólnoty językowej. Podejście to do niedawna mogłoby jedynie zostać zbyte wzruszeniem ramion, ponieważ przytłaczająca większość społeczeństwa nie publikuje żadnych tekstów. Ta sytuacja jednak zmienia się dzięki Internetowi — coraz większa liczba ludzi umieszcza swoje teksty w postaci stron domowych, blogów, tzw. postów, czy wreszcie artykułów w Wikipedii.

Ostatnią z prezentowanych metod zastosował konsekwentnie (wg mojej wiedzy) František Čermak. Zaproponował on mianowicie, by tym, co korpus reprezentuje, była percepcja tekstów przez daną społeczność językową. W praktyce wykonanie tego postulatu wygląda tak, że bada się strukturę czytelnictwa danej społeczności. I tak — nieco trywializując — jeśli np. w ciągu roku statystyczny obywatel czyta 4 powieści i 2 poradniki, to udział powieści w korpusie będzie dwakroć większy niż poradników.

Realizacja tego — zdawałoby się — prostego postulatu w praktyce napotyka wiele trudności. Przede wszystkim nie jest rzeczą prostą w wiarygodny sposób odtworzyć strukturę czytelnictwa. Zasadniczo daje się to zrobić za pomocą badań ankietowych. Zamawianie tego rodzaju badań jest dość kosztowne, przy tym ich wiarygodność zapewne nie zawsze jest wysoka[5]. Przetworzenie tych danych na obraz korpusu też nie jest sprawą trywialną. Mimo to takie właśnie podejście wydaje się najlepiej ugruntowane metodologicznie, dlatego też opowiadam się za nim. Taki też model reprezentatywności zamierzam postulować dla Narodowego Korpusu Języka Polskiego.

Niezależnie od przyjętej metodologii rodzi się pytanie o sposób zapełniania wyznaczonych stylów funkcjonalnych. Znów możemy sobie wyobrazić co najmniej 3 podejścia:

— całkowicie losowy dobór tekstów;
— do korpusu włączane są teksty o wysokim nakładzie lub np. często wypożyczane w bibliotekach;
— tworzona jest jakaś procedura wyboru tekstów „istotnych" czy „nośnych kulturowo".

Jeśli chodzi o pierwsze podejście, to znów można je rozumieć na dwa sposoby. Jeden to rzeczywiście losowy dobór, z zastosowaniem generatora liczb losowych itp.[6] Tak był tworzony korpus SWPF. Mimo że jest to procedura metodologicznie najbardziej poprawna, stosowana nie tylko w językoznawstwie, ale przede wszystkim w tych naukach i praktycznych dziedzinach życia, gdzie idealne dobranie próbki jest sprawą

[4] Dla porównania dodam, że codzienna średnia objętość dużych dzienników (Gazeta Wyborcza, Rzeczpospolita) wraz z dodatkami jest zbliżona do objętości książki.

[5] Dokładny informacje na ten temat mogą dać jedynie badania książeczkowe, tzn. takie, w których respondenci w specjalnych zeszytach szczegółowo notują czas czytania gazety (za tę cenną informację dziękuję w tym miejscu Prof. Waleremu Pisarkowi).

[6] Procedury te są szczegółowo opisane we wstępie do SFPW.

kluczową, w praktyce nie daje się ona zastosować do korpusu, który tworzymy. Trzeba bowiem przypomnieć, że korpus SFPW liczył zaledwie pół miliona słów; ponadto w dobie składu ręcznego i tak wszystkie teksty dygitalizowano przepisując je ręcznie, a kwestia praw autorskich nie była stawiana tak ostro jak obecnie. Narodowy Korpus Języka Polskiego będzie kilkaset razy większy, co narzuca bardziej „pragmatyczne" podejście — jeżeli mamy do wyboru tekst w wersji elektronicznej i inny, który trzeba skanować, zdecydujemy się raczej na ten pierwszy. Ponadto twórcy korpusu są związani prawami autorskimi. Nie wchodząc w szczegóły w odniesieniu do każdego tekstu zasadniczo powinno się uzyskać zgodę osób dysponujących prawami autorskimi. Należy się więc spodziewać, że w wypadku większości tekstów wylosowanych, przynajmniej jedno kryterium nie zostanie spełnione (tzn. nie dotrzemy do właścicieli praw autorskich albo nie pozyskamy tekstu w postaci elektronicznej). To wszystko sprawia, że pojęcie losowy oznacza tutaj przypadkowy („ten, który uda się pozyskać"), a nie losowany zgodnie z procedurami. Oczywiście w takim wypadku łatwo możemy się spotkać z zarzutem, że na dobór tekstów ma wpływ postawa twórcy (jedni autorzy chętniej udzielą zgody, inni będą przeciwni włączeniu ich tekstu do korpusu) albo wydawców — niektórzy będą niechętni włączaniu do korpusu bestsellerów, obawiając się piractwa. Nie są to zarzuty całkowicie nieistotne, jednak — powtórzmy — metoda ściśle losowa, taka jaką zastosowano tworząc korpus SFPW nie daje się zastosować w odniesieniu do dużych korpusów. Mamy tu więc do czynienia z konfliktem tego co idealne, z tym co możliwe.

Metoda druga zakładałaby dawanie preferencji książkom i gazetom o wysokim nakładzie ewentualnie częściej wypożyczanym w bibliotekach. Tym samym zostałoby uwzględnione zróżnicowanie percepcji poszczególnych tekstów przez społeczność językową.

Trzecia wreszcie metoda rzadko jest stosowana przy tworzeniu współczesnych korpusów, niemniej stoi ona u podstaw tworzenia bazy materiałowej dawniejszych słowników. Tą też metodą starał się posłużyć W. Lubaś, tworząc podwaliny korpusu IJP PAN. Nie wchodząc w szczegóły, wspomnę, że stworzono rodzaj kwestionariusza, w którym uwzględniano datę urodzenia, pochodzenie geograficzne, światopogląd autora, a także pewne cechy leksykalne jego dzieł (np. użycie regionalizmów, wyrazów potocznych itp.). W sumie komponent „artystyczny"[7] miał się składać z tekstów dających możliwie zróżnicowane odpowiedzi na pytania kwestionariusza. Poza wspomnianymi powyżej ograniczeniami w pozyskiwaniu tekstów metodzie tej można zarzucić pewną arbitralność, wypracowanie bowiem rygorystycznych kryteriów doboru nie jest sprawą prostą. Tym niemniej, jeśli korpus ma być bazą materiałową dla słownika, to próba wychwycenia tekstów o zróżnicowanej leksyce, zwłaszcza jeśli planowany korpus miał być stosunkowo niewielki, należy ocenić jako słuszne. Trzeba jednak pamiętać, że metoda ta podlega dokładnie tym samym ograniczeniom co dobór ściśle losowy.

[7] Używam tutaj ad hoc terminu artystyczny, ponieważ do tej kategorii należą również dzieła reprezentujące literaturę faktu, ale niepozbawione artystycznych ambicji, np. *Zniewolony umysł* Miłosza czy *Hipnoza* Hanny Krall.

1.2. Budowa wybranych korpusów

Dla zilustrowania różnic w budowie rozmaitych korpusów referencyjnych w poniższej tabeli przedstawiam udział procentowy poszczególnych stylów funkcjonalnych:

	książki niebeletrystyczne	**książki beletrystyczne**	**prasa**	**inne**
BNC	45	15	25	
węgierski	14	20	45	21
syn2000 (czeski)	25	15	60	
syn2005 (czeski)	27	40	33	
włoski (Cordis)	22	25	38	15
słoweński (FIDA)	6	18	76	

Jak z powyższego widać, mimo zróżnicowania metodologicznego nie da się dostrzec dramatycznie wielkich różnic w budowie poszczególnych korpusów.

1.3. Propozycje metodologiczne dla korpusu słownikowego

W niniejszym artykule chciałbym jeszcze raz podkreślić z naciskiem, że ścisła reprezentatywność nie jest tak istotna w wypadku korpusu leksykograficznego, jak to ma miejsce w wypadku korpusów służących do badań, w których opis ilościowy odgrywa istotną rolę. Leksykograf zasadniczo nie robi ścisłych zestawień natury statystycznej. Stosunkowo najbardziej wrażliwym na błędy statystyczne punktem pracy leksykografa jest analiza kolokacji, która — jak się planuje — może odgrywać istotną rolę również w wyznaczaniu poszczególnych znaczeń. Trzeba wszakże pamiętać, że metody ilościowe mogą odgrywać tylko wstępną rolę, ostateczną instancją jest kompetencja językowa słownikarza.

Również z naciskiem chciałbym podkreślić, że korpus dla zastosowań leksykograficznych jest korpusem dosyć specyficznym. Stawiamy mu inne wymagania niż tzw. „korpusowi referencyjnemu" (czyli ogólnemu korpusowi „narodowemu", który znajduje zastosowania w bardzo różnych dziedzinach — od gramatyki i rekonstrukcji językowego obrazu świata, poprzez lingwistykę stosowaną po komputerowe przetwarzanie języka naturalnego). Szczególnie istotne jest zróżnicowanie leksykalne, które można uzyskać, dbając m.in. o zróżnicowanie tematyczne; nie można też pominąć kwestii tzw. autorytetu literackiego — polski użytkownik słownika oczekuje, że ze szczególną pieczołowitością potraktuje leksykograf wielkie dzieła literackie. Mniejsze też znaczenie przypisujemy reprezentatywności. Wreszcie zupełnie inne podejście mamy do rozwarstwienia chronologicznego korpusu. Korpus referencyjny powinien zapewne ograniczać się do języka naprawdę współczesnego, gdzie naturalną cezurą byłby np. rok 1989 lub 1990 (to i tak jest okres niemal dwu dekad, wspomniany wyżej korpus Uniwersytetu

Browna uwzględniał wyłącznie teksty powstałe w roku 1964). Zwróćmy uwagę, że jeśli mówimy, iż korpus ma odzwierciedlać percepcję języka, to umieszczanie w nim tekstów prasowych lub powieści, do których nikt już nie wraca, jest błędem. Z drugiej strony umieszczanie w takim korpusie lektur obowiązkowych z przełomu XIX i XX w. daje się uzasadnić. Wreszcie korpus referencyjny nie może pomijać tak istotnego typu tekstów, jakim są teksty ściśle naukowe. Ich obecność niekoniecznie jest pożądana w korpusie leksykograficznym[8].

1.4. Proponowana koncepcja reprezentatywności

Jak wspomniałem, opowiadam się za koncepcją reprezentacji struktury czytelnictwa (por. Górski 2008 w druku). W tym celu posłużymy się publikowanymi co dwa lata raportami Zakładu Badań nad Czytelnictwem BN. Od razu należy zastrzec, że te badania trudno przekładają się na to, czego chcemy się dowiedzieć: one badają raczej wzorce kulturowe niż recepcję poszczególnych stylów funkcjonalnych (stąd takie kategorie, jak „książki szkolne", które mogą oznaczać zarówno podręczniki szkolne, jak i lektury, czyli beletrystykę); niemniej nie stać nas na zamówienie badań, poza tym te, które są, dają się wykorzystać. Z kolei dane dotyczące czytelnictwa prasy pochodzą z ankiet Ośrodka Badań Prasoznawczych. Dodajmy jeszcze jedno ograniczenie, mianowicie, staramy się odzwierciedlić strukturę czytelnictwa nie całego społeczeństwa, ale przeciętnego inteligenta. Zastrzeżenie jest podyktowane m.in. tym, że znaczny procent społeczeństwa w ogóle nie czyta książek.

1.5. Przyszła budowa korpusu

Bazą materiałową dla WSJP będzie Narodowy Korpus Języka Polskiego. Projekt ten został już opisany gdzie indziej (Przepiórkowski i in. 2008), więc nie będę powtarzał wszystkich informacji, przytoczę tylko to, co jest istotne z punktu widzenia słownika. Korpus ten jest wspólnym przedsięwzięciem Instytutu Podstaw Informatyki PAN, Instytutu Języka Polskiego PAN, Katedry Języka Angielskiego UŁ i Wydawnictwa Naukowego PWN. W pierwszej fazie projektu zakłada się zespolenie istniejących korpusów w jeden[9]. Wedle planów powinno się to zbiec w czasie z rozpoczęciem intensywnych prac nad WSJP. Tak więc słownik ten od początku będzie oparty na relatywnie dużym i zrównoważonym korpusie. Docelowo zrównoważona jego część ma liczyć 250 do 300 milionów słowoform, natomiast niezrównoważona 800–1000 milionów słów. Tę wielkość osiągnie on wszakże wtedy, gdy prace nad słownikiem (w ich pierwszej części) będą dobiegały końca.

---

[8] Osobiście uważam, że należy uwzględnić w korpusie leksykograficznym pewną liczbę tekstów naukowych — nie dla słownictwa ściśle fachowego, ale by reprezentowały wspólny dla różnych dziedzin wiedzy żargon naukowy.

[9] W chwili gdy piszę niniejszy tekst, proces scalania istniejących korpusów w jeden zbliża się do końca, stąd w dalszej części artykułu będę pisał o nim jako o już istniejącym.

Korpus nie może być projektowany pod kątem użyteczności dla słownika, ponieważ ma służyć wielu różnorodnym celom naukowym. Tym niemniej niewątpliwie pewne postulaty słownikarzy będą mogły być uwzględnione, o ile nie będą stały w sprzeczności z ogólnymi celami korpusu.

Ponadto IJP PAN jako jeden z partnerów projektu Narodowego Korpusu Języka Polskiego będzie mógł tworzyć dla własnych celów badawczych podkorpusy składające się z tekstów pozyskanych przez NKJP. W praktyce oznacza to, że możliwe będzie uzupełnienie korpusu o teksty szczególnie istotne z punktu widzenia pracy nad słownikiem. Niewątpliwie będzie pożądane włączenie do tego korpusu tekstów zgromadzonych w toku najwcześniejszego etapu prac nad korpusem IJP PAN. Objętość tych tekstów jest proporcjonalnie bardzo niewielka (nieco ponad 1 mln słowoform), niemniej były one od początku selekcjonowane, w toku żmudnej procedury ustalania ich bogactwa leksykalnego, jako baza materiałowa dla słownika. Trzeba jednak tutaj z naciskiem podkreślić, że partia tekstów o niewielkiej objętości „ginie" w korpusie. Przede wszystkim wtedy, gdy użytkownik korpusu korzysta z narzędzi statystycznych, choćby po to, by wychwycić kolokacje. Teksty stanowiące niewielki procent całego korpusu w oczywisty sposób w niewielkim stopniu wpływają na wyniki tego rodzaju obliczeń. Ponadto leksykograf (czy w ogóle użytkownik korpusu) zwykle nie ma czasu na analizowanie wszystkich konkordancji. Im liczniejszy materiał, tym mniejsza szansa, że zatrzyma się na tej właśnie, a nie innej linijce konkordancji.

1.6. Propozycje dotyczące zrównoważenia chronologicznego

Nim przedstawię propozycje zrównoważenia chronologicznego, zwrócę uwagę na pewien problem teoretyczny. Otóż wcześniej opowiedziałem się za modelem zrównoważenia, który polegałby na odzwierciedleniu struktury czytelnictwa. Tymczasem ta struktura ulegała przez ponad 50 lat dość istotnym przeobrażeniom. Co gorsza nie dysponujemy wiarygodnymi danymi dotyczącymi przeszłości. Stąd — zapewne słuszny — postulat, by poszczególnym warstwom chronologicznym przyznawać różną budowę (odmienne proporcje poszczególnych stylów funkcjonalnych), odpowiadającą ówczesnej strukturze czytelnictwa uznamy za zbyt daleko idący.

Sytuacja idealna to taka, w której podzielilibyśmy okres, który opisuje WSJP, na kilka podokresów (1945–1956, 1957–1970, 1970–1980, 1981–1989, 1990 — współczesność). Okresy te wyróżniłem oczywiście na podstawie zupełnie zewnętrznych, historycznych cezur.

Każdy z tych podokresów byłby reprezentowany mniej więcej jednakową „porcję" tekstów, może z wyjątkiem ostatniego, który jest wyraźnie dłuższy, w związku z tym byłby reprezentowany przez większą „porcję" tekstów. Z góry jednak uznajmy tę propozycję za nierealną — stworzenie takiego korpusu jest przedsięwzięciem zbyt pracochłonnym i kosztownym.

Realną propozycją jest korpus współczesny „z perspektywą wstecz". Oznacza to uzupełnienie NKJP o pewną liczbę tekstów starszych, realistycznie rzecz biorąc, będzie to mniejsza część objętości korpusu. W NKJP literatura piękna (o wyższych ambicjach

literackich) ostatniego półwiecza jest nieźle reprezentowana, optowałbym raczej za uzupełnianiem komponentu prasowego; styl publicystyczny może być zresztą reprezentowany przez książkę publicystyczną, która zapewne będzie łatwiejsza do pozyskania. Wydaje się, że trafnym pomysłem jest sugestia Renaty Przybylskiej (informacja ustna), by w szczególności pozyskać gazety z okresów przełomowych (Październik, Marzec, Grudzień, okres Solidarności i stanu wojennego, a także Okrągłego Stołu)[10]. Ponadto korpus powinien być uzupełniony o pewną liczbę książek popularnonaukowych i hobbystycznych z tego okresu, które nie są — przynajmniej na razie — reprezentowane w zasobach przekazanych do NKJP. Może to mieć pewne znaczenie dla wychwytywania zmian w terminologii technicznej wspólnoodmianowej. Istotne jest zadbanie o to, by włączane do korpusu źródła były również chronologicznie zróżnicowane. Może to stanowić pewien problem wobec faktu, że dostępność książek wyraźnie starszych jest niższa. Trzeba też sobie wyraźnie powiedzieć: prasa będzie miała mniejszy udział w tekstach „starszych" niż w tekstach współczesnych, ponieważ — przeciwnie niż to ma miejsce w wypadku testów współczesnych — jej pozyskanie jest trudniejsze.

Ewentualne powiększanie korpusu mogłoby iść w dwu przeciwstawnych kierunkach: z jednej strony włączylibyśmy do niego pewną liczbę tekstów reprezentujących literaturę wysoką[11]. Z drugiej strony należy przede wszystkim uzupełnić korpus o ówczesną literaturę popularną. Jeśli chodzi o literaturę popularną, trudno jest wskazać trafne kryteria doboru tekstów, ponieważ w Polsce Ludowej nakład niekoniecznie odzwierciedlał rzeczywistą popularność. Z pewnością należałoby się odwołać też do popularnej literatury młodzieżowej (np. Niziurski), która, być może, w jakimś niewielkim zakresie oddaje język ówczesnej młodzieży. Kolejna kategoria tekstów, o którą należy uzupełnić korpus, to scenariusze filmowe i teksty słuchowisk.

2. Zróżnicowanie tematyczne i stylistyczne korpusu

Z zagadnieniem zrównoważenia korpusu łączy się jego zróżnicowanie tematyczne i stylistyczne. Zróżnicowanie tego rodzaju, z oczywistych względów, jest szczególnie istotne w wypadku korpusów o zastosowaniach leksykograficznych. Po pierwsze tylko w ten sposób udaje się uchwycić bogactwo leksyki, a z drugiej strony stwierdzić, co przynależy do warstwy leksyki wspólnoodmianowej — dopiero występowanie danego leksemu w różnych rodzajach tekstów pozwala potwierdzić taką przynależność.

Pozornie kontrolę zróżnicowania tematycznego mogłoby zapewniać przypisywanie każdego tekstu do kategorii wypracowanych przez bibliotekarzy, z klasyfikacją Deweya na czele. Jednak wydaje się, że przydatność tej klasyfikacji może być problematycz-

[10] Obawiam się, że ze względów technicznych włączenie do korpusu publikacji drugoobiegowych, metodologicznie i — by tak rzec — moralnie jak najbardziej słuszne, może być utrudnione, ponieważ teksty te trzeba będzie przepisywać, a nie skanować.

[11] Tutaj stosunkowo łatwo zastosować pewne kryterium mechaniczne, mianowicie włączać teksty, które zostały uhonorowane Nagrodą Kościelskich, są lekturami obowiązkowymi studentów polonistyki czy też licealistów.

na, jest ona bowiem bardzo dokładna w odniesieniu do literatury naukowej, ale dość zgrabnie klasyfikuje pozostałe rodzaje tekstów. Jak sądzę, rozsądną propozycją będzie potraktowanie klasyfikacji Deweya jako punktu wyjścia i uzupełninie jej o dokładniejszą klasyfikację, przede wszystkim literatury poradnikowo-hobbistycznej, która stanowi istotną część produkcji księgarskiej. Pomocne mogą być w tym zakresie również klasyfikacje stron WWW, tworzone przez portale.

2.1. Typologie stylów funkcjonalnych

Jeżeli reprezentatywność ma oznaczać odzwierciedlenie percepcji poszczególnych stylów funkcjonalnych przez Polaków, to najpierw należałoby ustalić sensowny podział na te style; dalej — musimy wypracować kryteria trafnego zaliczania danego tekstu do konkretnego stylu. Rzecz nie jest prosta, ponieważ, jak przyznaje Saloni (1993), „wyróżnianie poszczególnych stylów funkcjonalnych jest konwencjonalne i nie ustalone", tym bardziej więc każdorazowe przypisanie tekstu do stylu może być arbitralne. Ponadto należy ustalić istotne cechy stylów; znów odwołajmy się do reprezentatywności próbki w badaniu socjologicznym — jeżeli chcemy zbadać preferencje wyborcze, to przy konstruowaniu próbki ustalamy wiek, płeć, zawód respondenta, ale już nie wzrost, czy posiadany samochód. Podobnie w wypadku tekstów konieczne byłoby ustalenie pewnych obiektywnych kryteriów przynależności do danego stylu funkcjonalnego. Trzeba jednak przyznać, że przynajmniej jeśli chodzi o kryteria wewnątrzjęzykowe, to wiadomo niewiele. Wedle mojej wiedzy istnieją pewne badania dotyczące cech statystycznych tekstów angielskich. Są to głównie prace Bibera (a także innych autorów w duchu „Biberowskim").

W wielkim skrócie mówiąc, Biber wyznaczył pewną liczbę cech, głównie gramatycznych, które dają się policzyć w każdym tekście. Poszczególne teksty charakteryzują się wysokim lub niskim natężeniem tych cech. Co więcej, wysokiemu natężeniu pewnych cech towarzyszy wysokie natężenie kilku innych, i — przeciwnie — niskie natężenie kolejnych innych cech. Zestawienie frekwencji wspomnianych cech pozwala grupować teksty. Jeśli się porówna wyniki grupowania tekstów na podstawie kryteriów wewnątrzjęzykowych i funkcjonalnych, to nie zawsze się okaże, że oba te grupowania się pokrywają. Przykładowo, wg ustaleń Bibera, „literatura naukowa" jest językowo bardzo niespójna. Tę samą technikę statystyczną zastosował do badania tekstów litewskich A. Utka, z tą wszakże różnicą, że brał pod uwagę nie cechy gramatyczne, ale leksykalne; chodzi przy tym nie o wyrazy specjalistyczne, ale wspólnoodmianowe, z szerokim uwzględnieniem wyrazów synsemantycznych. Jego badania pokazują, że style funkcjonalne dają się scharakteryzować niską bądź wysoką frekwencją takich słów. O ile mi wiadomo, nie ma tego rodzaju badań dla polskich tekstów. Zresztą sam Biber przyznaje, że angielska stylistyka i „korpusologia" są w uprzywilejowanej pozycji, ze względu na to, że przy tworzeniu kolejnych generacji korpusów można bazować na badaniach poczynionych na wcześniejszych generacjach. Zauważmy, że tę szczęśliwą pozycję osiągnęliśmy też w Polsce, skoro projekt NKJP bazuje na kilku wcześniejszych korpusach. Zespół NKJP zamierza też wypracować pewne wewnątrzjęzykowe kryteria typologii tekstów.

2.2. Typologie tekstów wypracowane na potrzeby konstruowania korpusów

Nie ma tu miejsca na przytaczanie bogatej literatury dotyczącej klasyfikacji stylistycznych[12]. Przytoczę jedynie klasyfikacje, jakich dokonano na potrzeby wybranych korpusów. Wobec braku jednolitej typologii tekstów za każdym razem przy tworzeniu kolejnego korpusu dokonuje się innych klasyfikacji. Przykładowo przytoczę tu trzy. Pierwsza to klasyfikacja tzw. Brown Corpus:

I Teksty informacyjne: A. Prasa: reportaż, B. Prasa: teksty redakcyjne, C. Prasa: recenzje, D. Religia, E. Umiejętności i hobby, F. Literatura popularnonaukowa, G. Literatura piękna: biografie, wspomnienia, H. Różne (w tym: dokumenty rządowe, sprawozdania fundacji, sprawozdania instytucji przemysłowych itp.), J. Literatura naukowa.

II Fikcja literacka: K. Fikcja literacka ogólnie, L. Tajemnice i historie detektywistyczne, M. Teksty fantastycznonaukowe, N. Przygodowe i westerny, P. Romanse i historie miłosne, R. Humor (uwaga: O brak).

Z kolei Korpus Słownika Frekwencyjnego wyznaczał 5 stylów: 1) Teksty popularnonaukowe, 2) Drobne wiadomości prasowe, 3) Publicystyka (w tym stenogramy z posiedzeń organów państwowych i partyjnych!), 4) Proza artystyczna, 5) Dramat artystyczny.

Wreszcie moja propozycja, która została przyjęta w korpusie IPI PAN, wygląda następująco:

1. Styl artystyczny 1.1. Proza 1.2. Poezja 1.3. Dramat
2. Styl publicystyczny
3. Literatura faktu (pamiętniki, dzienniki, wspomnienia, biografie)
4. Styl naukowo-dydaktyczny — książki naukowe, popularnonaukowe i podręczniki:
   - 4.1. Naukowe humanistyczne
   - 4.2. Naukowe przyrodnicze
   - 4.3. Naukowe techniczne
   - 4.4. Popularnonaukowe
   - 4.5. Podręczniki

   (uwaga: 4.1.–4.3. podział wg KBN)
5. Styl urzędowo-kancelaryjny (ustawy, protokoły tzw. Komisji Rywina etc.)
   - 5.1. Protokoły (zawierający zapis tekstu mówionego)
   - 5.2. Ustawy etc.
6. Styl informacyjno-poradnikowy (poradniki, instrukcje, przewodniki turystyczne)
7. Styl potoczny (w tym blogi).

Chciałbym jednak podkreślić, że klasyfikacja ta zapewne ulegnie dalszym modyfikacjom w trakcie prac nad NKJP.

[12] Należy tu przytoczyć przede wszystkim kompendium Gajdy (1995).

2.4. Język mówiony

Zupełnie odrębnym zagadnieniem jest kwestia reprezentatywności języka mówionego. Wydaje się, że tutaj należy zastosować kryterium demograficzne, tzn. najpierw wytypować reprezentatywną próbkę do badań spośród mieszkańców dużych miast, dla wyłączenia tekstów gwarowych i zabarwionych gwarą. Pozyskane w tej grupie teksty rozmów i wypowiedzi włączyć do korpusu.

3. Podsumowanie

Nasze propozycje stanowią kompromis pomiędzy tym co idealne a tym co możliwe. Fakt, że powstało konsorcjum NKJP, jest dla planowanego słownika niezwykle szczęśliwy. Będzie on mógł być oparty na bardzo dużym korpusie. Jak się wydaje, część zrównoważona pod względem wielkości w zupełności zaspokoi potrzeby WSJP. Nieco mniej optymistycznie należy ocenić zrównoważenie chronologiczne.

W niniejszym artykule nie poruszyłem też pewnej bardzo istotnej kwestii, mianowicie sensownego sposobu pracy leksykografa z korpusem. Zagadnienie to znacznie przekracza ramy omawianego przeze mnie tematu, niemniej wymaga gruntownego przemyślenia.

Bibliografia

Biber D., 1988, Variation across speech and writing, Cambridge–New York.

Gajda S. red., 1995, Przewodnik po stylistyce polskiej, Opole.

Górski R.L., w druku, Representativeness of the written part of a large general-reference of Polish. Primary notes.

Przepiórkowski A., Górski R.L., Lewandowska-Tomaszczyk B., Łaziński M., 2008, Towards the National Corpus of Polish, [w:] Proceedings of the Sixth International Conference on Language Resources and Evaluation, Marrakech.

Saloni Z., 1993, Styl, [w:] Encyklopedia językoznawstwa ogólnego, red. K. Polański, Wrocław.

SFWP — Słownik frekwencyjny polszczyzny współczesnej, oprac. I. Kurcz, A. Lewicki, J. Sambor, K. Szafran, J. Woronczak, t. 1–2, Kraków 1990.

Sinclair J., 1996, EAGLES Preliminary recommendations on Corpus Typology. Version of May, 1996, http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Utka A., 2006, Common words as indicators of text functions, [w:] Prace Bałtystyczne 3, red. N. Birgiel, M. Kozak, s. 213–224.