Mirosław Górny, Piotr Wierzchoń, *Polish Digital Libraries As a Philologists' Tool. Based on 666 Adjectives from the Digital Library of Wielkopolska*, Poznań 2010, ss. 261.

Książka dwóch poznańskich profesorów – bibliotekoznawcy (Mirosław Górny) i językoznawcy (Piotr Wierzchoń) – powiększa zbiór polskich publikacji dotyczących lingwistyki korpusologicznej i kwantytatywnej. W wymiarze szczegółowym poświęcona jest ona zilustrowaniu możliwości wykorzystania zasobów coraz prężniej rozwijających się w Polsce bibliotek cyfrowych w badaniach językoznawczych, zwłaszcza lingwochronologicznych. W pierwszym zdaniu *Przedmowy*[1] Autorzy deklarują wprost:

> „Cel tej książki określamy najprościej, jak to możliwe: chcemy *zakomunikować* społeczności filologów, jak bardzo wartościowe zbiory tekstów są zgromadzone w powstających obecnie *bibliotekach cyfrowych*"[2] (s. 7; wyróżnienia autorów).

Publikacja składa się z dwóch zasadniczych części, obudowanych *Przedmową*, *Posłowiem* i *Bibliografią*. Część pierwsza – opisowa – o tytule *Polskie biblioteki cyfrowe jako narzędzie filologa* została rozczłonkowana na cztery mniejsze rozdziały: *Wprowadzenie (do biblioteki)*, *Wyszukiwanie*, *Filtrowanie*, *Podsumowanie*. Część druga zatytułowana *...na podstawie 666 przymiotników z Wielkopolskiej Biblioteki Cyfrowej* zawiera z kolei materiał ilustracyjny.

W *Przedmowie* Autorzy wyrażają przekonanie, że badania nad językiem, zwłaszcza nad jego ewolucją, winny opierać się na tekstach, nie zaś na danych słownikowych (listach haseł). W związku z tym postanawiają zaprezentować biblioteki cyfrowe jako narzędzie zdobywania danych z tekstów, narzędzie wygodne, bo oparte na tzw. formacie Dublin Core[3], umożliwiającym m.in. przeszukiwanie dużych korpusów tekstów według daty publikacji.

Następująca dalej *Część pierwsza* zawiera uwagi dotyczące wykorzystywania zasobów polskich bibliotek cyfrowych.

Górny i Wierzchoń rozpoczynają od przedstawienia ich zalet, do których zaliczają: 1) nieograniczony w czasie i przestrzeni dostęp do materiałów w nich zgromadzonych (w przeciwieństwie do bibliotek tradycyjnych), 2) wysoką skuteczność wyszukiwania, pozwalającą badaczowi zaoszczędzić czas, 3) trwałość (tradycyjne źródła drukowane, pochodzące sprzed II wojny światowej i starsze, są często w złym stanie fizycznym, co utrudnia dostęp do nich).

Rozdział *Wyszukiwanie* Autorzy poświęcili omówieniu metodyki wyszukiwania potrzebnych lingwiście danych historycznych (czyli ustalenia

[1] W niniejszym tekście posługuję się polskimi nazwami poszczególnych partii książki. Na polski przekładam także ich tytuły.

[2] Wszystkie tłumaczenia moje – R.Z.

[3] Jest to powszechnie stosowany w opisie zasobów bibliotecznych standard metadanych, w wersji podstawowej uwzględniający 15 parametrów: tytuł, twórca, temat i słowa kluczowe, opis, wydawca, współtwórca, data, typ zasobu, format, identyfikator zasobu, źródło, język, odniesienie, zakres, zarządzanie prawami.

jak najwcześniejszych wystąpień określonego słowa). Procedurę wyszukiwawczą porównują do pracy archeologa – tak jak archeolog ma do dyspozycji „teren wykopalisk", tak lingwochronolog pracuje na „zbiorze czasopism" (te badacze uznają za niezwykle ważne dla badań językoznawczych), przy czym różnica polega na tym, że nie wszystko, co archeolog odnajdzie na swoim terenie, daje się precyzyjnie datować, natomiast „artefaktom" językowym, wyekscerpowanym z czasopism, można z góry przypisać moment zaistnienia. Co więcej, cyfrowy „teren wykopalisk" lingwisty jest łatwy do podzielenia na sektory, takie jak: data wystąpienia, obszar geograficzny, liczba poświadczeń, typ tekstu itd., co pozwala odpowiednio sprofilować materiał wyjściowy przed przystąpieniem do dalszych badań. W dalszej części mowa jest o praktycznych sposobach wyszukiwania interesujących badacza słów. Autorzy prezentują najpierw najprostsze wyszukiwanie manualne, po czym omawiają możliwości jego zautomatyzowania i trudności, jakie z tym się wiążą (np. zła jakość druku, co powoduje, że współczesne algorytmy cyfrowe nie zawsze są w stanie właściwie odczytać daną sekwencję znaków). Celem automatyzacji jest, rzecz jasna, przyspieszenie procesu wyszukiwania. Ostatnia część rozdziału poświęcona jest przedstawieniu możliwości zawężenia wyszukiwania w zależności od celu, który stawia sobie badacz. Górny i Wierzchoń omawiają sposoby ustawienia zapytania tak, aby przynosiło odpowiedź na konkretne pytanie badawcze, np. o najpóźniejsze (przykład: *komisya* – sic!) lub najwcześniejsze (przykład: *telewizja*) wystąpienie słowa w druku.

Rozdział kolejny *Filtrowanie* rozpoczynają autorzy wyrażeniem przekonania, że narzędzia cyfrowe winny służyć lingwistom nie tylko do suplementowania słowników, ale też opisywania dynamiki ewolucji języka, np. poprzez wykrywanie najczęstszych procesów słowotwórczych w danym czasie, ustalania, które języki dostarczały najwięcej zapożyczeń czy też które sufiksy były w pewnym momencie historycznym najbardziej produktywne. Badania tego typu mogą być prowadzone szybciej przy użyciu współczesnego oprogramowania, niż gdyby trzeba było je prowadzić tradycyjną metodą filologiczną. W drugiej części tej partii publikacji Czytelnik otrzymuje informacje (w tym opis działania) o istniejących narzędziach cyfrowych pozwalających przeszukiwać duże korpusy zdigitalizowanych tekstów. Są to portugalski program NeoTrack, służący do półautomatycznego wyszukiwania neologizmów, oraz aplikacja NeoloSearch, za pomocą której można wyszukiwać anglicyzmy w norweskich tekstach prasowych. Autorzy dalej twierdzą, że w podobny sposób można przeszukiwać zasoby polskich bibliotek cyfrowych, gdyż baza zdigitalizowanych obecnie (2010 rok) dokumentów liczy ok. 300 000 publikacji, na które składają się w 60% gazety i czasopisma (ok. 1,8 mln stron), w 10% – książki (ok. 6 mln stron), w pozostałej części – druki ulotne (ok. 8 mln stron)[4].

[4] Nie ma tu rozbieżności między szacunkami procentowymi a liczbowymi, gdyż dane procentowe dotyczą bezwzględnej liczby dostępnych publikacji, a dane dotyczące liczby stron – objętości publikacji (książki mają więcej stron niż czasopisma czy druki ulotne).

Rozdział zamyka krótki opis takiej metody konstruowania korpusu, która pozwoli automatycznie konwertować go na listę jednostek językowych, zapewniającą możliwość dalszej analizy, np. frekwencyjnej.

W *Podsumowaniu* poznańscy badacze postulują, aby w przygotowywaniu cyfrowych baz bibliotecznych, tak aby służyły jak najlepiej celom badawczym lingwistów, uczestniczyli pospołu filologowie, lingwiści[5] i bibliotekarze. Zdaniem Górnego i Wierzchonia filologowie powinni mieć wpływ na rodzaj gromadzonych materiałów, powinno się im także zapewnić dostęp do nich na innych zasadach (w domyśle – korzystniejszych) niż zwykłym użytkownikom. Rozdział wieńczą refleksje na temat funkcjonowania bibliotek cyfrowych w Polsce – Autorzy podkreślają, że jest ich tylko ok. 40, z czego najstarsza ma niewiele więcej ponad 5 lat. Są one ponadto rozproszone, a każda z nich ma raczej małe zasoby, na które z powodu ochrony praw autorskich składają się w większości publikacje sprzed 1939 roku. Dalej autorzy poruszają zagadnienie niewystarczającego subsydiowania przedsięwzięć związanych z digitalizacją ważnych dla kultury polskiej zbiorów (ich zdaniem odnotowano pewną poprawę w tym zakresie dzięki funduszom unijnym i dotacjom ministerialnym), upominają się także o konieczność prawnej regulacji działania polskich bibliotek cyfrowych – chodzi im o to, aby digitalizacja była statutowym obowiązkiem największych polskich bibliotek. Przyczyną tak zdecydowanego stanowiska jest to, że praktycznie wszystkie polskie teksty z XIX i XX wieku wydrukowano na tzw. kwaśnym papierze, co oznacza, że wkrótce ulegną one całkowitemu fizycznemu rozpadowi, dlatego trzeba podjąć działania, aby ocalić nieoceniony zbiór tekstów, pozwalający badaczom na znaczne poszerzenie wiedzy o XIX- i XX-wiecznej historii polskiego społeczeństwa.

Do tych myśli wracają Autorzy w *Posłowiu* (następującym już po *Części drugiej*) – przeprowadzają w nim błyskotliwe porównanie: ich zdaniem koszt zakupu przez państwo polskie jednego myśliwca F-16 jest równy kosztowi zdigitalizowania zawartości wszystkich polskich XIX-wiecznych czasopism, niewiele więcej zaś potrzeba (badacze piszą o „dodatkowych kilku Rosomakach[6]"), by utrwalić cyfrowo – czyli „ocalić od zapomnienia" – zawartość wszystkich dostępnych obecnie w polskich tradycyjnych bibliotekach czasopism z okresu międzywojennego. Dziwią się też Autorzy, że corocznie w Polsce przeznacza się ok. 120 milionów złotych na konserwację pomników, a trudno wygospodarować 10 milionów złotych w skali rocznej na ochronę (czyli digitalizację) tego, co bezcenne i szybko ulegające degradacji, czyli polskich tekstów drukowanych w wieku XIX i w pierwszej połowie XX. Książkę kończy dramatyczna niemalże przepowiednia:

[5] Nie jest dla mnie jasne wprowadzone przez Autorów rozróżnienie na filologów i lingwistów, jednakże w tym miejscu ograniczam się jedynie do ujęcia sprawozdawczego, toteż nie wdaję się w rozważania mające na celu usunięcie tej niespójności.

[6] *Rosomak* to nazwa kołowego transportera opancerzonego będącego na wyposażeniu polskiej armii.

„Jest wysoce prawdopodobne, że za dwadzieścia czy trzydzieści lat nie pozostanie nic, co by się nadawało do zeskanowania..." (s. 257).

Część druga – jak wspomniałem wyżej – zawiera materiał ilustracyjny, czyli przykłady najwcześniejszych użyć 666 przymiotników nie odnotowanych w *Słowniku języka polskiego* pod redakcją Witolda Doroszewskiego, poprzedzony krótkim opisem metodologii opracowania tego materiału. Autorzy ograniczyli się jedynie do ekscerptów z tekstów składających się na zasoby Wielkopolskiej Biblioteki Cyfrowej, zastrzegając się jednocześnie, że ich celem było jedynie zademonstrowanie bogactwa i różnorodności nieskodyfikowanego materiału leksykalnego, który może być pozyskany dzięki przeszukaniu zasobów bibliotek cyfrowych. Każdy leksem prezentowany jest w ten sam sposób: Autorzy zamieszczają wycinek fotokopii z kontekstem użycia danego przymiotnika, a także oryginalną winietę przedwojennego czasopisma, z którego ów wycinek pochodzi.

* * *

Trudno sformułować uwagi ściśle recenzenckie w odniesieniu do omówionej wyżej publikacji. Pierwszą tego przyczyną jest to, że nie ma ona charakteru analitycznego – Autorzy nie stawiają żadnej wyraźnej tezy, która organizowałaby przeprowadzony w książce wywód. Książkę należy raczej odbierać jako rodzaj praktycznego poradnika (może nawet przewodnika) dla filologów nie korzystających dotąd z zasobów bibliotek cyfrowych. Drugi powód utrudniający recenzentowi zadanie to ten, że prezentowana Czytelnikowi wiedza, nawet jeśli ma walor nowości, to jest nietrudna do przyswojenia, bo oparta jest na prostym doświadczeniu empirycznym, co wszak nie przekreśla pożyteczności ww. publikacji.

Tym, co może budzić zastrzeżenia, są proporcje między treścią merytoryczną a ilustracyjną. Część merytoryczna zajmuje ok. 20 stron (licząc z *Przedmową* i *Posłowiem*), czyli nieznacznie tylko przekracza objętość typowego artykułu naukowego, część ilustracyjna niemal 230 (średnio trzy przymiotniki wraz kontekstami na stronę druku). Domniemywam, że Autorzy dążyli w ten sposób nie tylko do osiągnięcia wysokich walorów estetycznych publikacji, ale zamierzali także unaocznić Czytelnikowi sposób utrwalenia danych pozyskanych z przeszukiwania zasobów bibliotek cyfrowych.

Wątpliwości wzbudza też fakt, dlaczego książka dotycząca polszczyzny i polskich bibliotek, pisana przez Polaków dla Polaków (ewentualnie dla zagranicznych slawistów), jest napisana po angielsku. Sądzę, że – niczego nie ujmując Autorom – gdyby zdecydowali się napisać ją po polsku, mogliby liczyć na znacznie szersze rozpropagowanie swojej idei wśród rodzimych polonistów.

Co do materiału ilustracyjnego – w opinii niżej podpisanego – 666 przymiotników zilustrowanych przez Górnego i Wierzchonia to derywaty systemowe, nie pomieszczone w *Słowniku* Doroszewskiego być może dlatego, że nie znajdowały poświadczenia w bazie tekstów, którą dysponował Doroszewski, a być może dlatego, że leksykograf uznał, że ich regularność (systemowość) i niska frekwencja w uzu-

sie są czynnikami wystarczającymi, by je pominąć w opisie słownikowym. Tym większą więc wartość ma dokumentacyjna publikacja Górnego i Wierzchonia, którzy zwrócili swoją uwagę na to, co rzadkie i niepozorne. Dla zobrazowania pomysłu badawczego Autorów przedstawiam wyimek z listy przymiotników, który tworzą pierwsze leksemy z grup zaczynających się na tę samą literę alfabetu (pełna lista na s. 36-43): *adhortatywny*, ***bałałajkowy***, *całopłócienny*, ***dendrometryczny***, ***elektrobiologiczny***, *faryzeuszowy*, *harmonizacyjny*, ***imigrancki***, *jasnobarwny*, ***kilkucentowy***, *lusterkowy*, ***małoazjatycki***, *nadbałkański*, *ochronkowy*, *paleoetnograficzny*, *radiomedyczny*, *samoofiarny*, *talizmaniczny*, *ultracarski*, ***wazelinowy***, *zachodniomorawski*, *żółtomięsny*. Tylko siedem przymiotników (wyróżniam je pogrubieniem) z tej listy nie jest podkreślanych przez standardowy Word 2007, co należy wszakże uznać za dowód, że warto prowadzić cyfrowe „wykopaliska" językowe.

Rafał Zimny

Od Redaktora Naukowego LB

Fakt, iż książka M. Górnego i P. Wierzchonia napisana została po angielsku wzbudza nie tylko wątpliwość (jak to ujął w swoim omówieniu R. Zimny), ale jest – moim zdaniem – główną wadą tej publikacji (choć może i jedyną). A jeśli już, z nieznych Czytelnikowi powodów (a warto by w tej sytuacji o nich wspomnieć choćby na wstępie *Przedmowy*, nawet tylko w przypisie), Autorzy wybrali język angielski, by przedstawić filologom (polskim) możliwości, jakie stwarzają dla badań językoznawczych (polskie) biblioteki cyfrowe, to warto było dołączyć streszczenie w języku polskim (tak, jak w wypadku prac pisanych po polsku dołącza się do nich streszczenia w językach obcych).

Hans Jürgen Heringer, *Fehlerlexikon. Deutsch als Fremdsprache. Aus Fehlern lernen: Beispiele und Diagnosen*, Cornelsen Verlag, Berlin, 2001, s. 310

Beim Erlernen jeder Sprache sind diverse sprachliche Fehler nicht zu vermeiden. Die Fehleranalyse und die Fehlerdiagnose gehören demnach zum festen Bestandteil der wissenschaftlichen Forschungen im Bereich des Fremdsprachenlernens. H. J. Heringer versucht mit Hilfe und am Beispiel real auftretender Fehler, bezogen auf Deutsch als Fremdsprache, in diese Problematik einzuführen.

Die folgende Arbeit verfolgt das Ziel, häufige Lernerfehler zu beschreiben, die mit den Schwierigkeiten der deutschen Sprache zu tun haben. Die Hauptidee des Lexikons ist: den Benutzer auf typische Fehler und Fehlerquellen beim Lernen des Deutschen hinweisen und ihre Ursachen erklären. Der Verfasser gibt einen Überblick über sprachliche Strukturen, mit denen häufig sowohl Anfänger als auch Fortgeschrittene Schwierigkeiten beim Erlernen des Deutschen haben.

Der Autor führt Beispiele der Fehler an, die aus verschiedenen Quellen stammen: Fehler von Lernenden des Goethe-Instituts auf der ganzen Welt, Fehler von Lernern in Deutschland und Fehler aus Sammlungen zu