RAFAŁ L. GÓRSKI

Baza materiałowa, cytowanie, źródła i korpus dla *Wielkiego słownika języka polskiego*

1. Wprowadzenie

WSJP ma w założeniu — przynajmniej w dużej mierze — powstać w oparciu o korpus. Niniejszy tekst ma za zadanie przedstawić propozycje dotyczące szeroko pojętej bazy materiałowej tego opus magnum współczesnej leksykografii.

Ostatnim słownikiem współczesnej polszczyzny stworzonym wyłącznie w oparciu o korpus jest SJPDor. Słownik ten, mimo krytyki, jaka spada nań w ostatnich latach, jest wciąż punktem wyjścia w dyskusji nad leksykografią. Szczegółowe badania Jachimczak i Porosły (1989) pozwoliły ustalić, że — mówiąc w dużym uproszczeniu — Doroszewski cytował poezję z 2 poł. XVIII i 1 poł. XIX w., prozę z 2 poł. XIX w., publicystykę i literaturę fachową sobie współczesną, a okres międzywojenny zignorował. Ponadto, jak ustalił Lubaś (1989), SPJDor jest słownikiem opartym raczej na literaturze pięknej niż tekstach niebeletrystycznych, co jest zresztą zgodne z polską tradycją leksykograficzną.

2. Źródła pozakorpusowe

Słownik powinien zawierać również terminy przestarzałe w rodzaju *maszyna cyfrowa*. Jest ich zapewne niewiele (być może więcej dotyczących nauk społecznych), niemniej należy je troskliwie udokumentować. Na potrzeby projektowanego słownika jest tworzona lista słowników, quasi-słowników i opracowań, które powinny być źródłami pomocniczymi[1]. Trzeba jednak pamiętać, że zasób leksykalny tych źródeł musi być poddany ostrej selekcji, ponieważ tego rodzaju publikacje z zasady zawierają przede wszystkim słownictwo wybitnie specjalistyczne i/lub środowiskowe. Z drugiej strony nie znaczy to, że nie warto takiego słownictwa gromadzić, choćby w postaci zwykłych list, bez próby nawet pobieżnego opisu leksykograficzego,

[1] Lista ta liczy obecnie z górą 30 pozycji.

ponieważ listy tego rodzaju mogą stanowić cenne źródło uzupełniające dla różnego rodzaju badań leksykologicznych.

3. Korpus

3.1. Jeżeli mamy poważnie traktować datę graniczną słownika — rok 1945 — to należy stworzyć odpowiadający temu korpus, reprezentujący cały ten okres. Jest bowiem rzeczą istotną, by nie powtarzać wspomnianego wyżej błędu (bądź zamierzonej niekonsekwencji) Doroszewskiego. Stworzenie tego rodzaju korpusu jest możliwe, może być wszakże kosztowne. Decyzja zależy od roli, jaką przypiszemy korpusowi w procesie redagowania haseł. Jeśli ma być on przede wszystkim źródłem cytatów, to zapewne nie warto poświęcać na jego tworzenie zbyt wielkiej energii. Jeśli natomiast ma być rzeczywiście podstawą do tworzenia części haseł, to stworzenie tego rodzaju korpusu musi być uznane za jeden z istotnych etapów poprzedzających, a częściowo towarzyszących tworzeniu słownika. Pewne nadzieje budzą plany stworzenia Narodowego Korpusu Języka Polskiego, którego zasoby byłyby udostępnione zespołowi słownika.

3.2. Autorzy projektu słownika (Dunaj, Przybylska, Żmigrodzki 2006) przewidują, że korpus będzie spełniał co najmniej dwie role. Explicite mówią, że jeśli wystąpią wahania dotyczące fleksji, to na podstawie korpusu będzie się ustalało, która z form przeważa. Druga przewidziana przez projekt słownika rola dla korpusu to wychwytywanie kolokacji. Wprawdzie nie zostało to powiedziane w omawianym tekście wprost, ale kolokacje z zasady dają się obserwować jedynie w tekstach.

Zasadniczo słowniki powstają na podstawie korpusu tekstów (bazy źródłowej), a nie kompetencji językowej redaktorów. Rola ich kompetencji ma się sprowadzać do interpretacji źródeł[2].

Kilgarriff i Rundell (2002) opisują program komputerowy Word Sketches, który służy do wychwytywania typowych kolokatów w korpusie. Program ten posługuje się nie tylko prostą frekwencją danego połączenia wyrazowego, ale także wiedzą gramatyczną (np. szukanie połączeń przymiotnik–rzeczownik) i pewnymi równaniami matematycznymi, które — mówiąc bardzo ogólnikowo — uwzględniają to, czy dwa wyrazy sąsiadują ze sobą często tylko dlatego, że są częste, czy też sąsiadują ze sobą częściej, niż by to miało wynikać z ich frekwencji. Takie narzędzie jest niezbędne, żeby słownikarze mogli szybko dokonać oględzin typowych połączeń wyrazu hasłowego i wychwycić te istotne. W konkluzji artykułu autorzy stwierdzają, że w wydawnictwie Macmillana program ten został stworzony do wyłapywania kolokacji, jednak w praktyce okazało się, że redaktorzy rozpoczynali pracę nad hasłem właśnie od sprawdzenia kolokacji w Word Sketches, ponieważ często właśnie typowe kolokacje

[2] W słowniku syntaktyczno-semantycznym czasowników polskich opracowywanym w IJP PAN pod kierunkiem R. Laskowskiego czy w *Słowniku syntaktyczno-generatywnym czasowników polskich* kompetencja autorów była traktowana na równi z bazą materiałową.

pozwalały wstępnie wydzielić znaczenia opracowywanego leksemu. Nie podejmuję się w tym miejscu rozstrzygać, czy akurat ten program jest optymalnym rozwiązaniem dla zespołu WSJP, niemniej nie unikniemy konieczności znalezienia programu, który będzie spełniał podobne funkcje[3].

3.3. Pojęcie reprezentatywności korpusu jest przedmiotem wielu dyskusji i sporów[4]. Ponadto, w korpusach, które stanowią bazę materiałową dla słowników, przypisuje się reprezentatywności mniejszą wagę, ponieważ i tak słownikarz nie ma dość czasu, żeby analizować wszystkie konteksty, ani nie przedstawia dokładnych liczb. Istotniejszą cechą jest zrównoważenie, tzn. taki udział ilościowy poszczególnych stylów funkcjonalnych, że żaden z nich nie dominuje. W naszym wypadku należałoby też zadbać o zrównoważenie chronologiczne — teksty reprezentujące poszczególne style funkcjonalne powinny też pochodzić równomiernie z całego okresu od 1945 do współczesności. Gdyby się zdecydować na tworzenie tego rodzaju korpusu, należałoby się liczyć z koniecznością skanowania czasopism, co jest zadaniem trudnym, choć wykonalnym. Jestem bowiem przekonany, że projektowany słownik powinien uwzględniać język propagandy Polski Ludowej, czy szerzej — styl ówczesnej publicystyki.

3.4. Wobec tego, że trzeba z góry założyć, iż korpus będzie miał ograniczoną wielkość, słownik będzie tworzony dwutorowo — hasła o wyższej frekwencji będą opracowywane na podstawie korpusu, te o niskiej zaś — innych źródeł. Ponadto autorzy projektu słownika zakładają opracowanie najpierw 15 000 najważniejszych haseł. Wydaje się więc, że jednym z pierwszych zadań będzie stworzenie słownika frekwencyjnego. Będzie to naturalnie słownik do użytku wewnętrznego, bardzo uproszczony i niedokładny, a jego jedynym zadaniem będzie zorientowanie zespołu, które jednostki leksykalne powinny być opracowywane w pierwszej kolejności i na jakiej podstawie (korpus czy źródła pozakorpusowe). W tym miejscu należy zauważyć, że SFPW liczy z górą 10 000 słów. Uwzględnia on jedynie słowa o częstości większej niż 4 wystąpienia w korpusie liczącym 500 000 słowoform. W takim razie można oczekiwać, że słowa te powinny w korpusie liczącym np. 100 000 000 słowoform wystąpić co najmniej 800 razy, co daje bardzo dobrą ilustrację materiałową. Oczywiście, jest to bardzo naiwna ekstrapolacja, w szczególności w odniesieniu do słów z dołu listy rangowej, niemniej daje ona orientację co do oczekiwanego bogactwa materiału, przynajmniej z dokładnością do rzędu wielkości.

Inne przybliżenie daje wypowiedź Krishnamurthy'ego (2002). W nieformalnej odpowiedzi na pytanie o stosunek wielkości korpusu do objętości słownika, zamieszczonej na liście dyskusyjnej językoznawstwa korpusowego, podaje on, że w leksykografii angielskiej przyjmuje się, iż korpus wielkości 18 mln słów wystarczy do sporządzenia słownika liczącego 20 000 haseł, 120 mln — 45 000 haseł, a 450 mln pozwala

[3] Pewne nadzieje budzą prace nad rozbudową funkcjonalności programu do obsługi korpusów Poliqarp. Wstępna wersja tego programu jest opisana w książce Przepiórkowskiego (2004).

[4] Przytoczenie całej literatury na ten temat znacznie przekracza ramy niniejszego opracowania.

opisać leksykograficznie około 100 000 haseł. Warto zauważyć, że nie mamy tu do czynienia z zależnością liniową — do stworzenia słownika o pięciokrotnie większej liczbie haseł potrzeba dwudziestopięciokrotnego wzrostu korpusu. Oczywiście tekst angielski zawiera znacznie więcej słów (rozumianych jako ciąg od spacji do spacji) niż jego polski odpowiednik, niemniej znów te obliczenia dają jakiś bardzo przybliżony obraz oczekiwań względem korpusu.

3.5. Z powyższego wynika, że nie należy zbierać przykładów wprost z Internetu, ponieważ zaburza to wspomniane zrównoważenie. Oczywiście pewna liczba tekstów internetowych (w szczególności form dla Internetu charakterystycznych, typu blogi, czaty, posty itp.) zostanie włączona do korpusu, ale na zasadach takich, jak inne teksty, i w ten też sposób będzie lokalizowana. Należy wszakże uczynić wyjątek dla następującej sytuacji: wiele okazjonalizmów pojawia się wyłącznie w Internecie i oczywiście w takim wypadku powinny być one odnotowane. Jest też drugi, choć mniej istotny powód, dla którego Internet nie może stanowić podstawowego źródła dla słownika. Otóż teksty ściągane bezpośrednio z Internetu nie są zaopatrzone w anotację fleksyjną (morfosyntaktyczną), która ogromnie ułatwia analizowanie materiału.

4. Ilustracja materiałowa i chronologia cytatów

By ustalić odpowiedź na pytanie o zakres ilustrowania haseł, należy wcześniej zapytać, w jakim celu pewne słowniki podają cytaty. Generalnie rzecz biorąc, cytaty pełnią dwie funkcje: 1. są one rodzajem naukowego potwierdzenia, dowodem, że dane znaczenie istotnie w języku funkcjonuje, a nie zostało wymyślone przez leksykografa (a więc jest to rodzaj dowodu naukowego), 2. stanowią swego rodzaju uzupełnienie definicji słownikowej.

Obie te funkcje nie muszą się pokrywać, np. jeżeli się założy, że ilustracja materiałowa ma odzwierciedlać chronologię źródeł i należy zacytować najstarszy przykład, to często może się okazać, że jest on akurat mało ilustratywny. Wiele słowników, które nie mają ambicji naukowych, posługuje się przykładami tworzonymi przez leksykografów. W „papierowej" leksykografii nie bez znaczenia jest fakt, że ścisła lokalizacja cytatów zajmuje sporo miejsca.

Wspomnę wreszcie o trzeciej funkcji — słownik z cytatami stanowi rodzaj korpusu, szczególnie w odniesieniu do słowników historycznych, które cytują obficie bądź nawet przytaczają wszystkie przykłady.

Należy z tego wysnuć następujące wnioski: w dzisiejszych czasach, gdy korpusy są powszechnie dostępne, ostatni z celów stracił rację bytu. Co więcej, wnikliwy czytelnik słownika mógłby próbować zweryfikować znaczenia wyróżnione przez leksykografa, gdyby dysponował korpusem, na którego podstawie został stworzony słownik.

Wydaje się zatem, że rozwiązaniem najracjonalniejszym jest ograniczenie do niezbędnego minimum liczby cytatów „ilustracyjno-potwierdzeniowych". Zgodnie z projektem słownika (Dunaj, Przybylska, Żmigrodzki 2006) kolokacje będą przytaczane osobno, tak więc cytaty nie będą musiały zdawać sprawy z typowych kolokacji

wyrazu hasłowego. Cytaty te będą podawane w sposób zgodny z zasadami cytowania (tzn. opuszczone wyrazy będą zaznaczane trzykropkiem). Eo ipso nazwy własne będą musiały być cytowane in extenso, o czym wspominam, ponieważ niekiedy unika się takiego cytowania przykładów z prasy. Cytaty będą lokalizowane z dokładnością do źródła (tzn. autor i tytuł, ale już nie strona). Przykłady z artykułów prasowych będą lokalizowane tak jak w SJPDor, który podawał jedynie tytuł periodyku, numer i datę. Podanie nazwiska autora tekstu jest często niemożliwe w odniesieniu do dzienników, w których znakomita część artykułów prasowych jest anonimowa bądź podpisywana inicjałami autora. Ale jeśli zdecydować się na pomijanie autora, to „odziera się" z autorstwa znanych publicystów czy literatów — jeśli np. felieton W. Szymborskiej w „Gazecie Wyborczej" zostanie zlokalizowany jedynie jako GW. Wydaje się, że można to zostawić rozsądkowi autora hasła — czy uzna on za stosowne do lokalizacji dodać nazwisko autora tekstu, jeśli daje się ono zidentyfikować.

Sprawa lokalizacji cytatów stanowi też kolejny powód, dla którego nie należy pozyskiwać przykładów z Internetu — dokładne cytowanie URL nic nie daje, ponieważ strona może zniknąć z serwera bardzo szybko i adres stanie się nieaktualny.

Cytaty powinny idealnie odzwierciedlać rozwój i zmiany znaczenia słowa. W praktyce jest to dość trudne — odnalezienie najstarszego cytatu może być pracochłonne i w wypadku większości słów ten nakład pracy jest całkowicie nieopłacalny. Ponadto, najstarsze cytaty niekoniecznie są najbardziej ilustratywne. Pozostaje jednak problem słów, których znaczenie ewoluowało przez ponad pół wieku, a które ma objąć słownik. Przykładem niech będzie słowo *projekt* w SJPSzym, definiowane jeszcze jako: 1. 'zamierzony plan działania *etc.*' i 2. 'plan, szkic czegoś *etc.*', a który obecnie oznacza również 'przedsięwzięcie', 'wytwór działalności artystycznej' czy 'ideę'.

Nonono... jak tylko skończę projekt SuperSolider to potworzę jakieś modele i dosyłam Very Happy.

Zespół działa jak organizacja, to nie to samo, co didżejowanie. Czasami rzeczywiście jest ciężko przez te wszystkie różnice w osobowościach, wydumane ego itd. Zmieniliśmy kilku muzyków, co też było trudno przetrwać. W końcu przecież chcesz skończyć projekt i przedstawić go światu.

W latach 90-tych uwierzył w Michnika niczym w Marksa, a projekt III RP potraktował jak Utopię.

Wydaje się, iż polityczny projekt III RP zakładał proste przełożenie kantyzmu na grunt polski.

Jeszcze większą trudność sprawiają znaczenia-efemerydy, typ *„zabezpieczyć robotnikom ubrania robocze"* (pomijam tu fakt, czy to znaczenie, jako tępione przez specjalistów w zakresie poprawności językowej, projektowany słownik ma uwzględniać) — znaczenia, które pojawiły się i zniknęły.

Zdawać sprawę z tego rodzaju ewolucji powinno się w inny sposób niż przez chronologiczny układ cytatów. Ponieważ jest to kwestia czysto redakcyjna, czuję się zwolniony z obowiązku jej rozstrzygnięcia.

Bibliografia

Dunaj B., Przybylska R., Żmigrodzki P., 2006, Zarys koncepcji wielkiego słownika języka polskiego, Polonica XXVI–XXVII, s. 5–16.

Jachimczak V., Porosło K., 1989, Źródła w Słowniku języka polskiego pod redakcją Witolda Doroszewskiego, [w:] Wokół słownika współczesnego języka polskiego II, Studia Leksykograficzne 3, red. W. Lubaś, Wrocław, s. 27–37.

Kilgarriff A., Rundell M., 2002, Lexical Profiling Software and its Lexicographic Applications — a Case Study, [w:] Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, red. A. Braasch, C. Povlsen, Copenhagen, s. 807–818.

Krishnamurthy R., 2002, Corpus size for lexicography, Corpora list archive, http://torvald.aksis.uib.no/corpora/2002-3/0254.html

Lubaś W., 1989, Źródła do słownika współczesnego języka polskiego (I. Źródła literackie), [w:] Wokół słownika współczesnego języka polskiego II, Studia Leksykograficzne 3, red. W. Lubaś, Wrocław, s. 7–26.

Przepiórkowski A., 2004, Korpus IPI PAN. Wersja wstępna, Warszawa.