MONIKA CZEREPOWICKA,
ZYGMUNT SALONI

Co skreślano i co dopisywano w korpusie *Słownika frekwencyjnego polszczyzny współczesnej*

Słownik frekwencyjny polszczyzny współczesnej[1] (dalej SFPW lub *Słownik*) jest dziełem pionierskim. Ponieważ założono z góry pełne opracowanie wybranego z tekstów materiału, przyjęto system oznaczeń, który z powodzeniem można by zastosować do wszelkich tekstów polskich. Założono, że każdy tekst polski jest sekwencją (ciągiem skończonym) znaków elementarnych, należących do wyznaczonego zbioru. Zbiór ten nie został bezpośrednio zdefiniowany. Składały się nań wszystkie 32 litery podstawowego alfabetu polskiego, 3 litery alfabetu łacińskiego stosowane w tekstach polskich tylko w sytuacjach specjalnych (q, v, x), znaki przestankowe (. , ; : oraz w wersji wzbogaconej ! ? ( ))[2], znaki pomocnicze (łącznik, apostrof oraz kropka skrótowa – zgodnie z tradycją nieodróżniana od kropki traktowanej jako znak przestankowy). Przyjęcie takich założeń było konieczne ze względu na zaplanowane opracowanie całości materiału metodami komputerowymi (maszyny miały wykonać wszystkie czynności identyfikacyjne i obliczeniowe) – przy bardzo ograniczonych w okresie gromadzenia i porządkowania materiału możliwościach technicznych. Wystarczy przypomnieć, że cały materiał – po wstępnym przepisaniu na mechanicznej maszynie do pisania i uzupełnieniu tekstu o zakodowane oznaczenia gramatyczne i funkcyjne – został wprowadzony na nośnik infor-

[1] Współautor niniejszego artykułu brał udział w pracach nad wydaniem *Słownika* (był jego redaktorem).

[2] Odtworzenie wykrzykników, pytajników i nawiasów było możliwe dzięki dokładnym korektom korpusu (por. Nazarczuk 1997).

macji (papierową taśmę perforowaną) za pomocą dalekopisu[3]. Na nośnik tekst wpisywano z fiszek papierowych, na które przepisywano tekst bezpośrednio ze źródeł. Mieliśmy dostęp do papierowych fiszek, będących pierwotnie w posiadaniu prof. Andrzeja Marii Lewickiego, a obecnie w dyspozycji mgr Doroty Lewandowskiej. Egzemplarz ten daje postać tekstu znacznie uboższą niż wprowadzona na nośnik informacji, np. z rzadka tylko występują w niej kodowe oznaczenia gramatyczne. Jest to prawdopodobnie kopia zapasowa, a nie ta, którą autorzy posługiwali się podczas prac nad korpusem.

Dostosowanie wymagań ścisłego opisu do ówczesnych możliwości było dziełem Jerzego Woronczaka, który zaplanował pracę i jej kolejne etapy. Według wypracowanych przez niego założeń wprowadzono na komputerowy nośnik informacji tekst polski składający się z 500 tys. słów polskich wraz z dopisaną zakodowaną informacją gramatyczną. Zbiór ten był bazą zamierzonego dzieła, a dodatkowo – po licznych technicznych operacjach porządkujących, dzięki którym korpus dostępny jest obecnie na współczesnych nośnikach informacji (np. w postaci płyty CD)[4] – stał się wzorcowym korpusem języka polskiego, wykorzystywanym w pracach studyjnych przy testowaniu nowo tworzonych narzędzi do przetwarzania obszernych zbiorów tekstów[5]. Korpus więc żyje.

Zanalizowanie przyjętych przed laty założeń ma zatem nie tylko historyczne znaczenie.

Dla twórców SFPW było oczywiste, że postulowana przez nich postać tekstu

[3] Z powodu ograniczeń technicznych tego urządzenia na etapie wprowadzania materiału została zgubiona informacja o rozróżnianiu wielkich i małych liter – starano się ją jednak odtworzyć. Niektóre wielkie litery dało się przywrócić na mocy ortograficznej definicji zdania (wielka litera po kropce lub znaku równorzędnym). Odtworzenie innych, jak nazwy indywidualne ludzi, zwierząt, przedmiotów, instytucji, tytuły książek, czasopism, filmów, nazwy geograficzne, możliwe było albo dzięki korektom korpusu, albo dzięki właściwemu kodowaniu. Pojawiły się i nieuzasadnione wielkie litery, np.:

```
Warszawski[11] Zakład[111] Elektrotechniczny[211] A-trzydzie-
ści[+] Jeden[/] w[66] Glinie[/][161]
```

Można przypuszczać, że w tej sytuacji winny jest system kodowania. Znacznik [/] informujący o wielkiej literze został na pewnym etapie przetwarzania korpusu niewłaściwie zinterpretowany. Oznaczenie to powinno się pojawiać bezpośrednio po wyrazie pisanym wielką literą. Znacznik [+] informuje, że ciąg *jeden* tworzy z poprzedzającym go *A-trzydzieści* jedną słowoformę. Zatem – zgodnie z ogólną zasadą ortograficzną – w tym miejscu nie powinna się pojawić wielka litera. Wynik taki nie jest regułą, np. `Boeing-siedemset [+] siedem[/].`

[4] Prace wykonał zespół pod kierunkiem prof. Janusza S. Bienia. Skompresowany obraz płyty jest dostępny w Internecie pod adresem: http://www.mimuw.edu.pl/polszczyzna/wksf/wksf.iso.bz2. Płyta zawiera m.in. tekst artykułu: J.S. Bień, M. Woliński, *Wzbogacony korpus „Słownika frekwencyjnego polszczyzny współczesnej"*, w którym podana jest obszerna bibliografia prac związanych z tematem.

[5] Por. M. Ogrodniczuk 2003b. Oboje autorzy niniejszego artykułu uczestniczyli w dyskusji nad instrukcją służącą do anotacji morfosyntaktycznej korpusu *Słownika frekwencyjnego*, prowadzonej w ramach grantu KBN realizowanego w Instytucie Podstaw Informatyki PAN (2001–2004), kierowanego przez dra Adama Przepiórkowskiego (współautorka należała do grupy anotatorów korpusu). Autorzy posługiwali się więc korpusem w wersji zaktualizowanej, dostępnej obecnie na komputerach osobistych.

polskiego odbiega w niektórych punktach od rzeczywistości. Przede wszystkim teksty polskie zawierały znaki nie należące do – uznanego za podstawowy – zbioru znaków elementarnych. Trzeba się było ich pozbyć. Wyjścia były dwa: albo usunąć je (skreślić) z tekstu, albo przetworzyć na ciągi znaków dopuszczalnych.

Ponieważ nie zadawano z góry żadnego słownika, tylko alfabet znaków, za słowa polskie można było uznać wszelkie, nawet najbardziej dziwaczne, nazwy przejęte z innych języków, jeśli tylko nie zawierały liter dla nich specyficznych (np. é, è, ê, ë, ě itd.). Niedopuszczalne były też wszelkie znaki alfabetów innych niż łaciński, znaki matematyczne lub notacji specjalnych. Przewidziano więc, że skreśli się je z tekstu. Odpowiednie fragmenty tekstów zostały opatrzone oznaczeniem [~] (por. Nazarczuk 1997), np.:

(1) Czynnik [~] zwiększał bowiem wartość regatową jachtów bulbkilowych.
(2) Szybkość przepływu przez przekroje [~] i [~], które są mniejsze od przekroju [~], musi być odpowiednio większa.
(3) …częstotliwości zgodnej z [~] obwodu głównego…

Symbol [~] został wprowadzony w miejsce fragmentu tekstu wykreślonego z próbki. We wstępie do SFPW zbiór elementów skreślanych został zdefiniowany jako klasa resztkowa – po omówieniu zasad przepisywania „elementów obcych" jako słów polskich (Wstęp, s. xviii; Saloni 1973). Elementami obcymi zajmiemy się w drugiej kolejności.

Odtworzenie, co faktycznie zostało opuszczone, wymaga dużej pracy. Tak zwane próby gronowe, każda złożona z ok. 50 słów, są wprawdzie ponumerowane i oznaczone w przepisanym materiale i na nośniku, ale zlokalizowanie ich w oryginalnych tekstach źródłowych, ze względu na atomizację odcinków, jest bardzo kłopotliwe. Przyjrzyjmy się przykładowo próbce, którą udało nam się znaleźć.

(4) Punkt pracy tego tranzystora ustala opornik [~]. Opornik ten w przypadku zastosowania innego tranzystora może mieć wartość od osiem i dwie dziesiąte do piętnastu [~]. Wzmocniona energia prądów małej częstotliwości zostaje przeniesiona przez transformator do drugiego stopnia pracującego na głośnik. Wzmacniacz małej częstotliwości wykonany w układzie [~] zapewnia dużą stabilność pracy i duże wzmocnienie mocy[6].

Na miejscu pierwszego [~] wystąpiło R_5, drugiego – *kΩ*, trzeciego – *WE*. Co więcej, fragment „wartość od osiem i dwie dziesiąte do piętnastu" został wprowadzony zamiast oryginalnego „od 8,2 do 15".

[6] S. Wolszczak, *Amatorskie odbiorniki tranzystorowe*, Warszawa 1966, s. 59, w. 7.

Niekiedy przyjmujemy, że opuszczane elementy obce są niewiadome, choć w niektórych wypadkach ich odtworzenie – odgadnięcie – jest stosunkowo łatwe, np.:

(5) Litera [~] podobnie jak[9] [~] jest skrótem wyrazu łacińskiego[221]: radix[$][7].

Na miejscu drugiego [~] było pewnie *ρ*, na miejscu pierwszego – *r*, które z powodzeniem mogło być uznane za wyraz polski, tak jak potraktowane zostały *A* i *D* w przykładach:

(6) Zapotrzebowanie[111] na[64] witaminę A[/]...
(7) Witamina D[/] przeciwdziała[5] krzywicy[131].

Do elementów obcych zaliczono zatem również litery alfabetu łacińskiego w funkcji metajęzykowej (jak *R* i *r* w przykładach (4) i (5)). Jednak analogiczne literowe symbole witamin zostały potraktowane inaczej – po pierwsze pozostawiono je, po drugie, ciągi typu *witamina A* nie zostały zinterpretowane jako całości znaczeniowe[8].

Często jednak trudno stwierdzić, jakie fragmenty pominięto, np.:

(8) Gdy dowiedziano[501] się, że pułkownik Suzuki[/][111] w[66] ogóle nie uczy[501] się angielskiego[221] [~] pozwolono mu wyjechać.
(9) …tłumaczą [~] pośpiechem…
(10) Nowe[211] kino[111] buduje[501] się już [~] osiem[34] lat.

W niektórych wypadkach skreślono zapewne symbole literowe, które można było uznać za słowa polskie jak *r* w przykładzie (5). A oto kilka analogicznych sytuacji:

(11) Mały[211] kwadrat[111] umieszczony[211] po[66] literce[161] [~] oznacza…
(12) Jednostką napięcia jest wolt w skrócie [~].
(13) Zdaniem [~] Davenporta[/][121]…
(14) [~] Wierzbołowski[/] uczynił to[44] na[66] płaszczyźnie[161] metodologicznej[261]…

W przykładzie (12) opuszczono zapewne *V*. Pominięcie tego symbolu zgodnie z instrukcją („Konwencjonalne skróty nazw jednostek miar rozpisywano, stosując pełny zapis formy wynikającej z kontekstu, np. kg → kilogram…". Wstęp, s. xvii)

---

[7] Znacznik [$] określa występujące w korpusie nieprzyswojone wyrazy obce (bezpośrednie cytaty z języków obcych).

[8] Ciągi typu *witamina A*, *tul-170* uznajemy za analogiczne. W dalszej części artykułu wykażemy, że mimo zbieżności zostały potraktowane różnie.

dałoby zdanie dziwaczne: *Jednostką napięcia jest wolt w skrócie wolt*. Nic dziwnego, że osoba przepisująca tekst nie zastosowała się do instrukcji. Jeśli idzie o zdania (13), (14) można przypuszczać, że Davenport i Wierzbołowski mają imiona zawierające litery nienależące do alfabetu podstawowego (René, Jiří?). A może w tekście zamiast imion wystąpiły inicjały? Dlaczego je skreślono?

Na pewno byłoby lepiej, gdyby w takich wypadkach (a pewnie i w wielu innych) pozostawić tekst oryginalny, zamiast zastępować go znakiem opuszczenia. W tekście pozostawiono bowiem parę nazw jawnie obcych, jeśli tylko złożone były ze znaków alfabetu podstawowego. Autorzy zdawali sobie zresztą sprawę, że nie byli w tym wypadku konsekwentni (Wstęp, s. xxvi). Redagując tomy *Słownika* zrezygnowano z obcych cytatów zdaniowych, pozostawiono natomiast cytaty leksykalne, jak zwroty, wyrażenia, nazwy własne (por. Wstęp, s. xxvi). Wiele zostało zinterpretowanych jako słowoformy kilkuwyrazowe (podkr. 15–21). Porównaj przykłady:

(15) w[66] czasie i zaraz po[66] Vaticanum[+] Secundum[/].
(16) in[+] minus
(17) nota[+] bene
(18) Ceskoslovensky[/][+] Spisovatel[/],
(19) „Yedioth[+] Achronoth[$]” (przedruk[111] „Vorwurts[&]”)
(20) [„]Moderato[+] Cantabile[$][”][9]
(21) Na[66] placu[161] la[+] Defense[/], Montparnasse[/] wyrastały w[64] górę

Kłopot sprawiały nazwy obce zawierające niedopuszczalne litery ze znakiem diakrytycznym. W zapisanym tekście zostały one opuszczone, być może – zgodnie z zapisem w tekście źródłowym (por. przykład (18)). W obecnej postaci korpusu (tekście w stronie kodowej Windows CP1250) znajdują się jednak litery niedopuszczalne z punktu widzenia danych SFPW, np.:

(22) W[66] kronikach Quauhtinchán[/] mówi się[41] o[66] dwudziestu[36] miastach założonych[262] przez[64] Tolteków[142]…

Przykład ten udało nam się sprawdzić na fiszce papierowej. Występuje tam *á* ze znakiem diakrytycznym oddanym odręcznie przez przepisującego. Na nośnik informacji musiał on być wprowadzony później – podczas korekt korpusu – nie w fazie wprowadzania danych za pomocą dalekopisu (por. Nazarczuk 1997).

[9] Przy okazji wypada zwrócić uwagę, że *Ceskoslovensky Spisovatel* w odróżnieniu od *Yedioth Achronoth*, *Vorwurts* i *Moderato Cantabile* nie został zinterpretowany jako nazwa obca. Świadczy o tym brak oznaczenia [$] przy *Ceskoslovensky*.

Wiele wątpliwości dotyczy również elementów dopisywanych. Chodzi tu przede wszystkim o rozpisywanie literami polskimi liczb zapisanych w tekstach cyframi arabskimi i rzymskimi (Wstęp, s. xvii, punkt 1). Był to punkt dyskusyjny już w momencie opracowywania materiału *Słownika*. Autorzy zdawali sobie sprawę, że decyzja ta pociągała za sobą także niepożądane konsekwencje, a w szczególności wpływała na wyniki liczbowe.

Po pierwsze, w tekstach źródłowych występowały liczbowe oznaczenia dat, zwłaszcza lat – były to lata stosunkowo bliskie momentowi zapisania tekstu. SFPW notuje więc następującą frekwencję leksemów odpowiadających wielokrotnościom liczby 100:

DWIEŚCIE – 161,
TRZYSTA – 101,
CZTERYSTA – 65,
PIĘĆSET – 102,
SZEŚĆSET – 67,
SIEDEMSET – 51,
OSIEMSET – 85,
DZIEWIĘĆSET – 572.

Rzecz jasna, że gdyby badać tymi metodami analogiczne teksty polskie w roku 2003, frekwencja przytoczonych leksemów byłaby inna. Co więcej, można wątpić, czy zapisane liczby odpowiadają faktycznej frekwencji odpowiednich leksemów w latach sześćdziesiątych XX wieku, czy nie odzwierciedlają tylko wartości zapisów lat liczbami.

Po drugie, jak piszą autorzy, „postać słowa w korpusie zależała w wielu wypadkach […] od osoby przepisującej tekst, np.:

Taka liczba słów zajmuje ogółem **ok. 8 000** stron. →
Taka liczba słów zajmuje ogółem ***około ośmiu tysięcy*** *stron.*
albo
Taka liczba słów zajmuje ogółem ***około osiem tysięcy*** *stron.*" (Wstęp, s. xviii).

Zanalizujmy tę kwestię nieco dokładniej.

W korpusie znajdujemy następujące wypowiedzenia:

(23) Załoga liczy czterysta pięćdziesięciu ludzi.
(24) …jedno z dział kalibru sto dwadzieścia siedem milimetrów…
(25) …od tysiąc dziewięćset do trzech tysięcy trzystu sztuk na godzinę…

Przytoczone powyżej zdania w oryginale zapisane były następująco:

(23') Załoga liczy 450 ludzi[10].

[10] W. Supiński, *Sto lat okrętów wojennych*, Warszawa 1965, s. 166, w. 11.

(24') …jedno z dział kalibru 127 mm[11]…

(25') …od 1900 do 3300 sztuk na godz.[12]

To przypadek, że w korpusie otrzymują reprezentacje (23)–(25). Możliwe byłyby również zapisy:

(23_1) Załoga liczy czterysta pięćdziesiąt ludzi.

(23_2) Załoga liczy czterystu pięćdziesięciu ludzi.

(24_1) …jedno z dział kalibru stu dwudziestu ośmiu milimetrów…

(25_1) …od tysiąca dziewięćset do trzech tysięcy trzystu sztuk na godzinę…

(25_2) …od tysiąca dziewięciuset do trzech tysięcy trzystu sztuk na godzinę…

(25_3) …od tysiąca dziewięciuset do trzy tysiące trzysta sztuk na godzinę…

(25_4) …od tysiąc dziewięćset do trzy tysiące trzysta sztuk na godzinę…

Niektóre z zaproponowanych tu rozwiązań mogą się wydawać sztuczne (np. 23_1, 24_1). Tak samo ocenić można jednak zdania występujące w korpusie – w przykładach 23 i 25 brakuje konsekwencji w formach liczebników. Działał tu pewnie mechanizm automatycznego odczytywania liczb – przepisujący zaczynał od formy neutralnej, nie patrząc na dalszy kontekst, potem zaś dostosowywał formy do wymagań dalszych członów składniowych, a nie poprawiał tego, co już było napisane (na mechanicznej maszynie do pisania!). Dowodzi tego sztuczność wprowadzonej do pamięci maszyny postaci zdań typu 23 i 25.

Na postać tekstu zakodowanego jeszcze większy wpływ miała decyzja, że „część cyfrową napisu złożonego z cyfr i liter przepisywano słownie, np.: 50-lecie → pięćdziesięciolecie, 3-dniowy → trzydniowy" (Wstęp, s. xvii).

W wyniku słownego zapisywania liczb zostały rozpisane specjalistyczne oznaczenia sprzętu, związków chemicznych itp. W ten sposób powstały słowoformy takie, jak: `XS-jeden, F-pięć[+] a[/]` oraz `tulu-sto[+] siedemdziesiąt[121], europu-sto[+] pięćdziesiąt[+] pięć[121]` i `Boeing-siedemset[+] siedem[/]`. Dziwi ich pozostawienie w materiale wobec skreślania znacznie prostszych napisów złożonych z samych liter.

Każdy z wymienionych ciągów stanowi w korpusie oddzielną słowoformę. Są to konsekwencje bądź decyzji segmentacyjnych (łącznik nie jest uważany za separator tekstu[13]), bądź dalszych prac nad korpusem – oznaczenie [+] (plus w nawiasie kwadratowym) informuje, że sąsiadujące słowa tworzą jedną słowoformę.

[11] *Ibidem.*

[12] *Rozwój techniki w PRL*. Praca zbiorowa, Warszawa 1965, s. 490, w. 5.

[13] Instrukcja traktowania wyrazów z apostrofem i łącznikiem nie była prawdopodobnie wystarczająco jasna. Widać to w zdaniach z przypadkami zależnymi nazwiska de Gaulle. Forma dopełniacza została wpisana i oznaczona na 3 różne sposoby:

- z końcówką oddzieloną od tematu; forma nie zinterpretowana pod względem przypadka: `dla[62] de[+] Gaulle'[/][+] a;`
- z końcówką zinterpretowaną łącznie z tematem; forma zinterpretowana pod względem przypadka: `powitanie[111] prezydenta[121] de[+] Gaulle'a [/][121];`

Inne podobnie złożone słowoformy nie zostały tak zinterpretowane (np. *witamina A*). Jednak ciągi *tul-sto siedemdziesiąt, europ-sto pięćdziesiąt pięć* zinterpretowano jako kilkuwyrazowe formy rzeczownikowe. Por.:

```
(26) Przyrząd[111] składa[501] się z[62] cynkowe-
     go[221] zasobnika, w[66] którym[261] umieszczo-
     na jest[57] aluminiowa probówka z[65] niewiel-
     ką[251] ilością tulu-sto[+] siedemdziesiąt[121]
     lub europu-sto[+] pięćdziesiąt[+] pięć[121].
```

W materiale źródłowym cytat ten ma postać:

(26') Przyrząd składa się z cynkowego zasobnika, w którym umieszczona jest aluminiowa probówka z niewielką ilością tulu-170 lub europu-155[14].

Oczywiście żadna z wyżej wymienionych form nie przekroczyła wymaganego progu 4 wystąpień i w *Słowniku* nie ma podobnych dziwolągów. Istotne jest jednak to, że w korpusie są, zwłaszcza w pliku tekstów popularnonaukowych. Widać od razu, że znacznie naturalniejsze byłyby tu dwie decyzje skrajne: albo skreślenie podobnych nazw literowo-cyfrowych, albo pozostawienie ich w postaci oryginalnej (co wymagałoby włączenia do alfabetu podstawowego cyfr).

Wprowadzanie symbolu [+] było równoznaczne ze skreślaniem istotnego elementu z tekstu źródłowego – spacji[15]. Operację taką zastosowano też, niestety, do

---

• z błędem technicznym, bez oddzielenia końcówki apostrofem; forma zinterpretowana pod względem przypadka: `apel[111] de[+] Gaullea[/][121].`

Przy okazji zwróćmy uwagę, że decyzje segmentacyjne zaważyły na interpretacji nazwisk obcych z cząstkami *de* i *von*. Postanowiono, że ciągi *de Gaulle*, *von Hassel* będą traktowane jako jedna słowoforma – niejednorodna przecież pod względem morfologicznym. Niesamodzielne cząstki *de*, *von* nie ulegają bowiem odmianie, w odróżnieniu od nazwisk, które się deklinują, np. Gaulle'**a**, Hassel**a**. Cząstki *de*, *von*, *van* są często pomijane – zdecydowanie częściej usłyszymy i przeczytamy, że orkiestra wykonała V symfonię Beethovena niż van Beethovena. Potwierdzają to dane z Archiwum Rzeczpospolitej za rok 1999: Beethoven – 21 wystąpień, van Beethoven – 8.

Wydaje się, że w tym wypadku zabrakło konsekwencji, dla porównania – skrótowce w formach przypadków zależnych były rozpisywane, np. `w[66] PGR[+] ach[=]`. Skoro łącznik nie był traktowany jako separator tekstu (Wstęp, s. xvii), to ciągi typu *PGR-ach* powinny być w ten sam sposób interpretowane gramatycznie.

[14] Z. Kazimierczuk, *Izotopy – nieznani czarodzieje*, Warszawa 1964, s. 76, w. 13.

[15] Decyzja łączenia słów („skreślania spacji") była tym bardziej zaskakująca, że nie dopuszczano operacji przeciwnej – dzielenia słowa tekstowego na dwie słowoformy. Autorzy wprawdzie we Wstępie wspominają o przyczasownikowych końcówkach ruchomych, jednak nie podają ani reguł ich występowania, ani listy leksemów, do których mogą one być doklejane, a w korpusie odpowiednio ich nie kodują. O ile interpretacja przyczasownikowych *-m*, *-ście* itp. nie stwarza trudności, o tyle końcówki osobowe przy formach innych leksemów odmiennych mogą co najmniej dziwić, np. ciągi *wszyscyśmy*, *wszyscyście* były traktowane jako jedna słowoforma – przymiotnika WSZYSTEK. Stąd w tomach *Słownika* są spójniki, przysłówki, przymiotniki i zaimek odmienne przez osobę (por. hasła: ABY, WSZYSTEK, KIEDYŚ, CÓŻ), a zatem nie spełniające kryteriów definicyjnych odpowiednich klas leksemów podanych we Wstępie. W korpusie formy tego typu nie są zaopatrzone w oznaczenia gramatyczne. Por. przykłady (podkr. własne):

sporej grupy wyrażeń złożonych z co najmniej dwóch słów tekstowych, będących najzwyklejszymi słowami polskimi. Należą do nich wyrażenia złożone z dwu lub więcej słów, uznawane powszechnie, wedle autorów *Słownika*, za jednolite całości. Jest ich wiele (Wstęp, s. xxiii–xxv), np.:

- połączenia imiesłowów przymiotnikowych i odsłowników z cząstką *się*, np. *odnoszący się*, *uchylaniem się*;
- połączenia leksemów CO i JAK z formami przysłówków w stopniu najwyższym, np. *co najwyżej*, *jak najbardziej*;
- skostniałe wyrażenia przyimkowe, np. *na zewnątrz*, *w prawo*;
- wyrażenia złożone z niektórych form przyimków i przymiotników, np. *za duży*, *po polsku*, *po drugie*, *za młodu*;
- inne wyrażenia o charakterze indywidualnym, np. *gdzie indziej*, *z dala*, *po omacku*, *podczas gdy*, *przede wszystkim*.

Traktowanie jako jednej słowoformy połączeń imiesłowów przymiotnikowych i regularnych rzeczowników odsłownych z cząstką *się*, jak udowadnia E. Awramiuk[16], znacznie wpływa na wyniki statystyczne *Słownika* – umieszczenie *się* w artykułach czasownikowych, rzeczownikowych czy przymiotnikowych, a pomijanie w artykule SIĘ (lub analogicznym) zaniża jego frekwencję. Dlatego przy odpowiednich hasłach autorka proponuje zaznaczać występowanie obligatoryjne lub fakultatywne morfemu. I tak zamiast dwu haseł: WYCOFANIE i WYCOFANIE SIĘ (analogicznie dla pary ZMIENIAJĄCY i ZMIENIAJĄCY SIĘ) proponuje jedno WYCOFANIE (SIĘ), podające osobno formy z *się* i bez *się*, a zamiast hasła ZBLIŻANIE SIĘ – ZBLIŻANIE z *się* zaznaczonym we wszystkich formach. Rozwiązanie to nie zmieni pozycji *się* na liście rangowej, ale da rzeczywisty obraz jego częstości występowania.

Dla przykładu przytoczmy jeszcze hasła z dwu fragmentów SFPW:

s. 377		s. 723	
	PO CIEMKU		ZA BARDZO
	PO DRUGIE		ZA DARMO
	PO LUDZKU		ZA DŁUGO
	PO OMACKU		ZA DUŻO
	PO PIERWSZE		ZA DUŻY

I przywiozłam ci[43] taki[241] drobiazg[141]. Abyś nie zapomniał naszych[222] chwil.

Uklęknij i módl[501] się, abyśmy nie splamiły rąk krwią sióstr.

…na[64] mnie[44], ani na[64] Franusia[/][141], kiedyśmy przechodzili[53]…

…ja sobie[43] wyobrażałam twoje[241] życie[141], kiedyś była[53] jeszcze całkiem malutka…

Dosłuchałem[501] się: cóżem[44] ci[43] uczynił[53]?

Znacznik przy formie *cóżem* – [44] – wskazuje na klasę zaimkową leksemu oraz formę biernika, nie informuje natomiast o aglutynacyjności formy i charakterystyce zakończenia *-em*.

Przy formach czasownikowych zabrakło informacji o połączeniu z końcówką ruchomą. Podobnie nie wyodrębniono partykułowego *-że/-ż*. W konsekwencji wątpliwości budzi artykuł hasłowy zaimka CÓŻ. Słownik notuje dwie formy zinterpretowane jako biernik liczby pojedynczej, o kształtach: *cóż*, *cóżem*.

[16] Awramiuk 2001.

s. 377	PO POLSKU	s. 723	ZA MAŁO
	PO PROSTU		ZA MAŁY
	PO TRZECIE		ZA MŁODY
			ZA PÓŹNO
			ZA SŁABY
			ZA WCZEŚNIE
			ZA WIELE

Jak się wydaje, na takie posegmentowanie tekstu zbyt silny wpływ miały pewne założenia przyjęte *a priori*.

Jednostkę ZA traktowano jako przyimek, nie dopuszczano, aby nie wpływała ona na przypadek formy następującej po niej. Nawet bez wstępnej analizy wymienionej serii, po uwzględnieniu możliwości substytucji drugiego członu, uznać trzeba, że należy tu wyróżnić odrębną jednostkę językową. W wymienionych połączeniach słowo ZA pełni inną – nieprzyimkową – funkcję, jest wymienne ze *zbyt*. Musimy się zgodzić z E. Awramiuk, że do *Słownika* powinna była być dodana jednostka ZA II, należąca do klasy przysłówków (zgodnie z podanym we Wstępie podziałem na części mowy).

Serie z PO nie są tak liczne, ale też wyraźnie produktywne, łatwo można je kontynuować: *po czwarte*, *po piąte*, …; *po angielsku*, *po rosyjsku* … I tu mamy do czynienia z parami jednostek: przyimek *po* i formy leksemu odmiennego.

Podobnie połączenia *co najwyżej*, *jak najbardziej* składają się z dwu słowoform: partykułowej *co* lub *jak* oraz formy przysłówkowej *najbardziej*, *najwyżej*. Ciągi *gdzie indziej*, *podczas gdy* proponowalibyśmy rozbić na dwie formy partykułowe. Jest to właściwość systemu gramatycznego polszczyzny (reguły tworzenia odpowiednich jednostek z jednostek niższego rzędu – połączenie 2 przysłówków, np. *bardzo ładnie*), dlatego nie wymagają jednostkowego potraktowania w słowniku[17].

Decyzje wstępnego opracowania korpusu, sprowadzające się do wykreślania lub rozpisania jego elementów, wpływały więc istotnie na postać tekstu, a w pewnej mierze również na zawartość *Słownika*: zasób uwzględnionych leksemów i uzyskiwane dane liczbowe. Doświadczenie zdobyte przy opracowaniu SFPW było jednak bardzo cenne. Wskazywało ono bowiem zagadnienia i problemy, które powinny być rozstrzygnięte przed opracowaniem tekstu polskiego wprowadzanego do maszyny. Problemy wskazane tutaj były przedmiotem dyskusji i analiz przed opracowaniem zasad kodowania następnych korpusów tekstów języka polskiego[18].

---

[17] Oddzielny problem stanowią pozostałe wyrażenia zaliczone do jednostek dwusegmentowych. Dzielenie tylko połowicznie rozwiązuje problem. Trudno bowiem bez wnikliwszych studiów zahasłować formy typu **zewnątrz*, **dala*, **omacku*, **ciemku*. Z pewnością wyrażenia *po omacku*, *na zewnątrz*, *z dala* na tyle utrwaliły się w języku i świadomości użytkowników, że „siekanie" ich może wydawać się zbędne. Jednak pierwsze człony tych wyrażeń to regularne przyimki, zatem bliższej obserwacji wymagają człony pozostałe.

[18] Por. Przepiórkowski 2002.

Bibliografia

Awramiuk E. (2001), *Wpływ odstępstw od segmentacji ortograficznej na wyniki statystyczne Słownika frekwencyjnego polszczyzny współczesnej*, „Roczniki Humanistyczne" Uniwersytetu w Białymstoku, z. 6, s. 31–43.

Bień J.S., Woliński M. (2001), *Wzbogacony korpus „Słownika frekwencyjnego polszczyzny współczesnej"*, tekst zamieszczony na płycie *Wzbogacony korpus „Słownika frekwencyjnego polszczyzny współczesnej"*, Warszawa. Skompresowany obraz płyty dostępny w Internecie pod adresem: http://www.mimuw.edu.pl/polszczyzna/wksf/wksf.iso.bz2.

Kurcz I., Lewicki A., Sambor J., Szafran K., Woronczak J. (1990), *Słownik frekwencyjny polszczyzny współczesnej*, Kraków.

Nazarczuk M. (1997), *Wstępne przygotowanie korpusu „Słownika frekwencyjnego polszczyzny współczesnej" do dystrybucji na CD-ROM*. Praca magisterska napisana pod kierunkiem dra hab. Janusza S. Bienia. Warszawa. Instytut Języka Polskiego Uniwersytetu Warszawskiego. Praca dostępna na płycie *Wzbogacony korpus „Słownika frekwencyjnego polszczyzny współczesnej"*, pod adresem: http://www.mimuw.edu.pl/polszczyzna/wksf/wksf.iso.bz2.

Ogrodniczuk M. (2003a), *Nowa edycja wzbogaconego korpusu Słownika frekwencyjnego*, [w:] *Językoznawstwo w Polsce. Stan i perspektywy*, pod red. nauk. S. Gajdy, Opole, s. 181–190.

Ogrodniczuk M. (2003b), *Rozszerzenie opisów morfologicznych w tekstach korpusu słownika frekwencyjnego polszczyzny współczesnej*. Złożone do *Tomu pamiątkowego Profesor Jadwigi Sambor*, pod red. J. Linde-Usiekniewicz, R. Huszczy.

Przepiórkowski A. (2002), *Składniowe uwarunkowania znakowania morfosyntaktycznego w korpusie IPI PAN*. Złożone do „Poloników".

Saloni Z. (1973), *Alien Elements in Texts of Natural Language*, „Biuletyn Polskiego Towarzystwa Językoznawczego" XXXI, s. 141–152.

Saloni Z. (1991), *Słownik frekwencyjny polszczyzny współczesnej*, „ComputerWorld", 4 listopada, s. 16–17.

Woliński M. (2002), *System znaczników morfosyntaktycznych w korpusie IPI PAN*. Złożone do „Poloników".