

Uniwersytet Warszawski
Wydział Polonistyki
Instytut Języka Polskiego

Marta Nazarczuk-Błńska

Wstępne przygotowanie korpusu
«Słownika frekwencyjnego polszczyzny współczesnej»
do dystrybucji na CD-ROM.

Praca magisterska napisana pod kierunkiem
dra hab. Janusza S. Bienia
(Instytut Informatyki Uniwersytetu Warszawskiego)

Warszawa 1997

Wstęp

Celem pracy jest udostępnienie uporządkowanej, jednolitej, komputerowej wersji pewnego zbioru tekstów języka polskiego, zwanego dalej korpusem „Słownika frekwencyjnego polszczyzny współczesnej” [korpus SFPW]. Korpus, którego oryginał powstał w końcu lat sześćdziesiątych, stanowi materiał cenny dla prac językoznawczych, między innymi ze względu na oznaczenia gramatyczne istniejące w jego tekście przy formach homonimicznych.

Niniejszy opis zdaje relację z wykonanych czynności, związanych z porządkowaniem korpusu. Ze względu na podobieństwo struktury poszczególnych części, praca dotyczy wybranego podkorpusu, zawierającego teksty stylu popularno-naukowego. Stanowi ona opracowanie wzorcowe, które może zostać wykorzystane przy opracowywaniu pozostałych części korpusu.

Rozdział 1. pracy, mający charakter wprowadzenia, poświęcony jest historii korpusu.

W rozdziale 2. można zapoznać się ogólnie z charakterem wykonanej pracy. Znajdują się tu również informacje dotyczące wykorzystanych programów komputerowych.

Rozdział 3. opisuje proces gromadzenia i porównywania istniejących, dających się odczytać na komputerach zgodnych z **IBM PC** plików podkorpusu popularno-naukowego.

W rozdziale 4. przedstawiony został opis ujednoliconych oznaczeń metatekstowych. Przy każdym oznaczeniu podaję przykład użycia.

Rozdział 5. stanowi komentarz do zamieszczonego w dodatku opisu zasad kodowania w korpusie form homonimicznych.

Rozdział 6. omawia czynności związane z dołączeniem do ujednoliconej wersji podkorpusu stylu popularno-naukowego informacji bibliograficznej, wskazującej na źródła wykorzystanych w nim tekstów.

Przy końcu rozdziałów podaję przybliżone oszacowanie czasu potrzebnego

do starannego wykonania danej czynności. Informacje te można wykorzystać przy opracowywaniu pozostałych części korpusu.

Zamieszczone w Dodatku fragmenty Wstępu z: *Ida Kurcz Andrzej Lewicki Jadwiga Sambor Jerzy Woronczak, Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom III. Styl popularnonaukowy. Warszawa 1974*, prezentują zasady kodowania w korpusie form homonimicznych. Tekst ten stanowi równocześnie najlepszą istniejącą dokumentację plików korpusu.

Dołączony do pracy aneks zawiera wydruk treści ujednoliconego pliku korpusowego — uporządkowanych, opatrzonych informacją bibliograficzną próbek podkorpusu stylu popularno-naukowego.

Poniżej podaję definicje podstawowych terminów. Za najbardziej istotne dla pracy uważam wyrazy: *fiszka*, *korpus*, *podkorpus*, *próbka* oraz *słowoforma*.

Słowo *fiszka* oznacza najczęściej małą karteczkę, na której spisano dla celów naukowych lub bibliotecznych notatki na określony temat. W mojej pracy *fiszka* to pojedyncza luźna karteczka z cytatem z książki lub czasopisma i informacją bibliograficzną o nim. Zbiór owych cytatów stanowi materiał wykorzystany do stworzenia korpusu SFPW.

Określenia *próbka* odnosi się do całości zawartego na jednej fiszce, analizowanego tekstu, będącego cytatem z książki lub gazety. Informacje dotyczące bibliografii, również znajdujące się na fiszce, nie należą do próbki.

Zbiór próbek stanowi *korpus*. Ze względu na zawartość jest on podzielony na części — *podkorpusy* poświęcone różnym stylom funkcjonalnym języka polskiego.

Główną jednostką korpusu jest *słowoforma*, czyli *forma fleksyjna wyrazu*. Terminu tego używam zgodnie z założeniami zawartymi w: [*Kurcz i in. 1974*]. Jedna *słowoforma* to — z pewnymi wyjątkami — jeden wyraz ortograficzny, czyli ciąg liter zawartych między dwiema kolejnymi spacjami. Wyjątki od tej reguły — *słowoformy* złożone z więcej niż jednego wyrazu ortograficznego — szczegółowo omówione zostały w Dodatku.

1 Korpus SFPW

Oryginał korpusu, zwanego w pracy umownie korpusem „Słownika frekwencyjnego polszczyzny współczesnej” [korpusem SFPW], powstał pod koniec lat sześćdziesiątych. Stanowił podstawę do stworzenia list frekwencyjnych, zdających sprawę z frekwencji słownictwa we współczesnej polszczyźnie.

Korpus składa się z 10 000 papierowych fiszek. Jest podzielony na 5 części (po 2000 fiszek w każdej z nich), prezentujących różne style funkcjonalne współczesnej polszczyzny: styl popularnonaukowy, drobnych wiadomości prasowych, publicystyki, prozy artystycznej i dramatu artystycznego. Każda fiszka korpusu zawiera materiał o długości około 50 wyrazów.¹

Na oryginalnej wersji fiszek dokonywano ołówkiem licznych wykreśleń pewnych fragmentów tekstu, które nie podlegały analizie. Dotyczyło to głównie znaków interpunkcyjnych: dwukropków, cudzysłówów, myślników, nawiasów, pytańników, średników, wielokropków i wykrzykników. Wykreśleniom podlegały także większe fragmenty tekstu, np. wzory (matematyczne, chemiczne etc). Skreślano też wyrazy występujące przy końcu fiszki, na próbkach dłuższych niż 50 wyrazowych. Dopisywano natomiast przy homonimach kod gramatyczny pozwalający odróżnić poszczególne formy. Tak opracowany materiał wpisano do maszyny cyfrowej **Eliott 803** i poddano analizie.²

¹Przy uważnym zapoznaniu się z tekstem korpusu, okazuje się, że 50 wyrazów stanowi przybliżoną długość znajdującej się na fiszce próbki tekstu. Zdaniem prof. Marka Świdzińskiego: *Autorzy starali się nie ucinać próbki w środku wypowiedzenia, tak więc najkrótsza próbka liczy 33 słowa, a najdłuższa 79. (M. Świdziński, Własności składniowe wypowiedników polskich. Warszawa 1997, s.13-14)* Do kwestii umieszczania na próbce zdań niedokończonych wróć jeszcze w dalszej części pracy.

²Przy opracowywaniu tekstu korpusu uzyskano kolejno: konkordancje czyli alfabetyczne listy słowoform z lokalizacjami w tekście, listę alfabetyczną i rangową, listę a tergo oraz alfabetyczną listę haseł z formami wyrazowymi hasła (pochodzącymi od niego słowoformami). Programy do analizy tekstu zostały przygotowane przez Jerzego Woronczaka. [*Kurcz*

Kiedy maszyny typu **Eliott 803** wyszły z użycia, materiał korpusu zapisany na papierowych taśmach był odczytywany na komputerze **Odra 1204**. Znalazł się on również na taśmach magnetycznych, z których korzystał egzemplarz komputera **Odra 1204**, wykorzystywany przez Bronisława Rośławskiego (Uniwersytet Gdański). W 1991 roku — Krzysztof Szafran (Instytut Informatyki Uniwersytetu Warszawskiego) przeniósł materiał korpusu na dyskietki odczytywane przez komputery zgodne z **IBM PC**. Początkowo zawartość jednego pliku odpowiadała prawdopodobnie zawartości jednej taśmy papierowej³. W posiadaniu tej najstarszej istniejącej wersji plików jest prof. Marek Świdziński (Zakład Językoznawstwa Komputerowego Wydziału Polonistyki Uniwersytetu Warszawskiego).

Oryginał korpusu znajduje się w Katedrze Językoznawstwa Ogólnego Wydziału Polonistyki Uniwersytetu Warszawskiego.⁴

Z wyjątkiem dwóch wersji — wspomnianej najstarszej⁵ oraz pliku `po.txt`, otrzymanych od prof. Marka Świdzińskiego, dla potrzeb niniejszej pracy wszystkie wymieniane przeze mnie pliki, — zostały mi udostępnione przez dr. hab. Janusza Stanisława Bienia. Znajdują się one na jednym CD-ROMie.

i in. , op. cit.] Materiał otrzymany w wyniku analizy posłużył później do stworzenia zbiorczego tomu słownika frekwencyjnego. [Ida Kurcz, Andrzej Lewicki, Jadwiga Sambor, Jerzy Woronczak, Krzysztof Szafran, *Słownik frekwencyjny polszczyzny współczesnej*. Kraków 1990]

³W latach osiemdziesiątych jeden z pięciu podkorpusów przeniesiony został na taśmę magnetyczną, ale dopiero od roku 1991 Krzysztof Szafran z Instytutu Informatyki Uniwersytetu Warszawskiego zapisał cały korpus na nośniku elektronicznym w formacie IBMa. Całość ma postać 200 plików tekstowych (z rozszerzeniem *.dat); każdy zawiera przeciętnie 50 ponumerowanych próbek, rozpoczyna się zaś kodem podkorpusu, po którym następuje numer pierwszej i ostatniej próbki zawartej w danym pliku. [M. Świdziński, *op. cit.*, s.15.] Zdaniem dr. Krzysztofa Szafrana, jest prawdopodobne, że na jednej taśmie mieściło się około 50 próbek.

⁴W katedrze tej pracuje obecnie jedna z współautorek list frekwencyjnych — prof. Jadwiga Sambor.

⁵Dla stylu popularnonaukowego — 44 małe pliki o nazwach D???.dat, gdzie ??? oznacza kolejną liczbę od 133 do 177.

2 Stosowane metody i narzędzia pracy

Opracowywanie danych tekstowych wymaga uwagi i precyzji. Znaczenie mają wszelkie modyfikacje tekstu. W wypadku danych przechowywanych na nośnikach magnetycznych (dyskietkach) lub na twardym dysku komputera wprowadzanie automatycznie dużej ilości zmian jest proste. Jeżeli nie sporadzi się dokładnych notatek informujących o charakterze modyfikacji, po pewnym czasie przestaje się dostrzegać powstałe w trakcie modyfikacji różnice. Bez starannie prowadzonej dokumentacji poprawek usunięcie niepożądanых zmian (jeżeli na przykład zmieniliśmy koncepcję pracy, lub po prostu pomyliliśmy się) jest o wiele trudniejsze niż ich wprowadzenie. Z tego też powodu należy pracować nie na oryginalnym pliku zawierającym interesujący nas materiał, lecz na jego kopii. Warto również zachowywać poprzednią wersję — bez naniesionych poprawek.

Niniejsza praca dotyczy ujednolicania kilku wersji określonego materiału tekstowego — podkorpusu popularno-naukowego korpusu SFPW. Obejmuje ona proces gromadzenia plików, ich porównywania i selekcji; korektę całego, ujednoliconego już materiału (tzw. wersji-matki, czyli jedynej oficjalnej wersji podkorpusu, powstałej po uważnym przejrzaniu wszystkich dostępnych plików); opracowanie opisu aktualnych, używanych w tekście próbek podkorpusu oznaczeń metatekstowych; dołączenie do pliku informacji bibliograficznej. Elementem pracy jest również nagranie na CD-ROM pliku z uporządkowaną wersją podkorpusu popularno-naukowego korpusu SFPW.

Ze względu na pokaźne rozmiary materiału i automatyczny charakter wykonywanych czynności, korzystałam odpowiednie programy komputerowe. Starałam się posługiwać programami darmowymi. W większej części prace prowadzone były w systemie *Linux*. Jako podstawowy edytor został użyty *GNU Emacs*. Umożliwiał on pracę w trzech oknach równocześnie, co stanowiło ogromną zaletę przy wpisywaniu bibliografii. Pozwalał także na

łatwe porównywanie plików, dzięki wykorzystaniu narzędzia o nazwie *diff*. (Program ten wyświetla na ekranie zawartość porównywanych plików, zaznaczając odpowiednimi kolorami różnice między wersjami.)

Przy jednej z bardziej żmudnych czynności — ujednolicaniu w tekście próbek zapisu kodu gramatycznego poprzez dopisanie wokół kodu nawiasów kwadratowych — bardzo pomocne okazały się skrypty napisane w języku programowania *Perl*, stworzonym specjalnie do przetwarzania danych tekstowych.

Brakujące informacje bibliograficzne zostały zapisane w plikach tekstowych. Tak przygotowywany materiał stał się później treścią tabeli bazy danych, zarządzanej przez system o nazwie *Postgres95 (Data Base Management System)*.

Posłużenie się komputerem znacząco przyspieszyło pracę.

3 Pliki korpusu

3.1 Gromadzenie plików

Istniejący w postaci plików zgodnych z komputerami **IBM PC** materiał korpusu był dostępny dla zainteresowanych osób. Każda z nich mogła, po otrzymaniu plików z treścią próbek tekstowych korpusu, modyfikować je, dostosowując do własnych potrzeb. W ten sposób powstało kilka wersji plików, w których w odmienny sposób zapisywano często te same informacje. Później, powstałe w różnych miejscach zmienione⁶ kopie były udostępnione przez osoby, które wprowadzały modyfikacje, dr. hab. Januszowi Stanisławowi Bieniowi z Instytutu Informatyki Uniwersytetu Warszawskiego. W ten sposób, w jednym miejscu, znalazło się kilka różniących się od siebie wersji plików, zawierających ten sam tekst.

Pierwszą ważną czynnością było uzgodnienie wersji. Należało porównać wszystkie dostępne pliki korpusowe, odczytać zawarte w nich informacje, wprowadzić ujednolicony zapis oznaczeń metatekstowych.

Prace były prowadzone w systemie Linux.

Poszczególne wersje plików znajdują się na wspomnianym już CD-ROMie w katalogach, których nazwy wskazują jednoznacznie, kto zajmował się daną wersją — nazwy katalogów pochodzą od nazwisk osób modyfikujących zawartość plików lub po prostu będących w posiadaniu danej wersji. Pliki skopiowane zostały do jednego katalogu o mnemotechnicznej nazwie **popularno-naukowy**. Kopiując z CD-ROMu do wspólnego katalogu w komputerze, starałam się

⁶Ingerencje osób, które modyfikowały zawartość plików częściowo dotyczyły technicznych problemów zapisu zawartych w plikach informacji, np. wystąpienia spacji przed zapisem kodu gramatycznego, użycia nawiasów kwadratowych dla odznaczenia kodu. Robert Wołosz próbował przywrócić do tekstu polskie litery. Istotną zmianę wprowadzili, niezależnie od siebie, Janusz S. Bień i Robert Wołosz, dokonując w plikach korekty literówek.

(Janusz S. Bień) JSBIEN \longrightarrow DYSKIETK \longrightarrow 1 \longrightarrow P0.3 \longrightarrow po3.zip \longrightarrow
 popul.3
 (Krzysztof Szafran) SZAFRAN \longrightarrow KORPUS \longrightarrow popul.zip \longrightarrow popul.dat
 (Robert Wołosz) WOŁOSZ \longrightarrow NOWY \longrightarrow (Robert Wołosz) WOŁOSZ \longrightarrow
 popular.txt
 (Robert Wołosz) WOŁOSZ \longrightarrow NOWY \longrightarrow ORYGINAL \longrightarrow popul.dat
 (Robert Wołosz) X \longrightarrow KORPUS \longrightarrow WOŁOSZ \longrightarrow ORYGINAL \longrightarrow popul.zip
 \longrightarrow popul.dat
 (Robert Wołosz) X \longrightarrow KORPUS \longrightarrow WOŁOSZ \longrightarrow ORYGINAL \longrightarrow WOŁOSZ \longrightarrow
 popular.txt
 (Robert Wołosz) WOŁOSZ97 \longrightarrow popular.frq
 (Robert Wołosz) WOŁOSZ97 \longrightarrow popul.txt
 (Marek Świdziński) MSWIDZ \longrightarrow po.txt
 (Marek Świdziński) MSWIDZ91 \longrightarrow D???.dat

Rycina 1: Oryginalne nazwy plików

zachować oryginalne nazwy dla poszczególnych podkatalogów. Jeżeli jakieś różniące się od siebie pliki pochodziły od jednej osoby, ale z różnego okresu — mogło się zdarzyć, iż znalazły się w dwóch oddzielnych, podkatalogach, z których nazwa jednego niosła informację, nie tylko o nazwisku osoby modyfikującej plik, ale też o czasie (wystarczyło podanie roku) ostatnich modyfikacji (por. podkatalog WOŁOSZ i WOŁOSZ97).

Rycina 1 zawiera informacje o otrzymanych plikach. Po prawej stronie w nawiasie zapisane jest imię i nazwisko osoby, od której pochodzi nazwa podkatalogu. Po nim, drukowanymi literami zapisałam oryginalne (lub analogiczne stworzone przeze mnie) nazwy podkatalogów. Dalej może nastąpić nazwa archiwum, w którym przechowywany był interesujący mnie plik lub — bezpośrednio — nazwa pliku.

W przypadku istnienia kilku takich samych nazw plików (np. popul.dat

```

MSWIDZ → po.txt → po.txt
MSWIDZ91 → d133.dat + d134.dat ... + d177.dat → dpopul.dat
JSBIEN → ...popul.3 → popul.3
SZAFRAN → ...popul.dat → popul.dat
WOLOSZ → ...popular.txt → npopular.txt
WOLOSZ → ...ORYGINAL → popul.dat → populorg.txt
WOLOSZ97 → popular.frq → popular.frq
WOLOSZ97 → popul.txt → popul.txt
X → KORPUS → WOLOSZ → ORYGINAL ... → popul.dat → xpopul.dat
X → KORPUS → WOLOSZ → ORYGINAL → WOLOSZ → popular.txt →
xpopular.txt

```

Rycina 2: Nowe nazwy plików

z katalogu WOLOSZ i popul.dat z katalogu SZAFRAN), część nazw została zmieniona na nowe — w miarę możliwości mnemotechniczne, np: populorg.dat (kopia pliku popul.dat z podkatalogu ORYGINAL), npopular.txt (kopia pliku popular.txt z podkatalogu NOWY). W ten sposób otrzymałam, 10 plików o różniących się od siebie nazwach, zawierających zmodyfikowane w odmienny sposób wersje tego samego tekstu.

Rycina 2 przedstawia stare (obok nazw podkatalogów) i nowe (po prawej stronie) nazwy plików. Kopie plików D???.dat zostały połączone w 1 plik o nowej nazwie dpopul.dat.

Dla zachowania bezpieczeństwa, w katalogu popularno-naukowy został utworzony podkatalog roboczy, w którym umieściłam kopie 10 plików opatrzonych nowymi nazwami.

3.2 Porównywanie, selekcja

Uważne porównywanie wersji plików miało na celu odczytanie wszystkich zawartych w nich metatekstowych informacji i sporządzenie wzorcowej wersji komputerowej podkorpusu.

Roboczo przyjąłam stronę kodową ISO-Latin-2.⁷

Dobrym narzędziem, służącym do porównywania dwu lub trzech plików, okazał się diff. Najbardziej wygodne było użycie go w środowisku X-windows pod edytorem GNU Emacs. Uruchomiony w edytorze diff wyświetla na ekranie zawartość porównywanych plików, zaznaczając odpowiednimi kolorami różnice między wersjami. W oddzielnym oknie podaje ilość występujących między wersjami różnic, oraz liczbę określającą, którą z nich właśnie oglądamy.

Na początku przejrzałam 10 zgromadzonych w katalogu **roboczym** plików i zanalizowałam występujące w nich różnice. Ponieważ niektóre z nich charakteryzowały się pewną regularnością, zostały automatycznie usunięte. Dotyczy to głównie problemu: spacja – brak spacji przed kodem gramatycznym lub: nawiasy kwadratowe – ich brak wokół kodu. Przyjęłam za poprawne występowanie nawiasów kwadratowych wokół kodu gramatycznego oraz pojawienie się spacji jedynie między słowoformami lub przed kolejnym wyrazem słowofromy wielowyrazowej. Zmian dokonywałam przy pomocy jednolinijkowych skryptów w języku perl. Rycina 3 przedstawia treść komend wpisywanych w linii poleceń. Składają się one z ułożonych w odpowiedniej kolejności części: uruchamiającej interpreter języka perl (napis **perl**), opcji (w tym wypadku tekst **-pe**), skryptu w języku perl (tekst ograniczony z dwu stron apostrofem), nazwy pliku, który opracowujemy oraz nazwy pliku, który uzyskujemy w wyniku zastosowania skryptu.

⁷Sposób kodowania znaków w komputerze, zgodny z normą ISO-8859-2.

Usunięcie zbędnych spacji

```
perl -pe 's/(\s+)(\d+)/$2/g;' nazwa_źródłowa > nazwa_docelowa
```

Dodawanie nawiasów

```
perl -pe 's/([^\s])(\d+)/$1\[$2\]/g;' nazwa_źródłowa > nazwa_docelowa
```

Wyrzucenie nadmiarowych nawiasów (wokół numeru fiszki)

```
perl -pe 's/*[(\d+)]*$/g;' nazwa_źródłowa > nazwa_docelowa
```

Zamiana miejscami oznaczeń kolejnej fiszki

```
perl -pe 's/(\*)(\d+)/$2$1/g;' nazwa_źródłowa > nazwa_docelowa
```

```
perl -pe 's/*(d+)/$1*/g;' nazwa_źródłowa > nazwa_docelowa
```

Zamiana końca linii na spację

```
perl -pe 's/\n\s/g;' nazwa_źródłowa > nazwa_docelowa
```

Dodanie końca linii przed oznaczeniem kolejnej fiszki

```
perl -pe 's/(\d+*)/\n$1/g;' nazwa_źródłowa > nazwa_docelowa
```

Wprowadzenie dużych liter na początku zdania

```
perl -pe 's/(\.\s)(.)/$1\u$2/g;' nazwa_źródłowa > nazwa_docelowa
```

Wprowadzenie dużej litery w pierwszym zdaniu fiszki

```
perl -pe 's/(\*)(.)/$1\u$2/g;' nazwa_źródłowa > nazwa_docelowa
```

Rycina 3: Wprowadzenie regularnych zmian

Udało się również w prosty sposób wyeliminować duplikaty. W przypadku porównywania ze sobą duplikatów diff informował, iż istnieje między nimi 0 różnic. Po przejrzeniu plików, ich ilość zredukowana została do 5. Okazało się bowiem, że:

```
popul.dat = populorg.dat = xpopul.txt  
npopular.txt = xpopular.txt.
```

Zawartość pliku `dpopul.dat` różni się od zawartości pliku `po.txt` jedynie napisami określającymi numer próbki (lub kolejnych 50 próbek), a w pliku `popul.txt` występuje tylko jednolita treść próbek w postaci ciągłego tekstu, bez jakichkolwiek oznaczeń, nawet informacji o numerze fiszki. Dlatego został on wykluczony z przeznaczonych do dalszego porównywania plików i zachowany jako cenny plik-źródło w katalogu `czysty_tekst`.

Oto nazwy 5 pozostałych plików:

```
po.txt,  
popul.dat,  
popul.3,  
npopular.txt,  
popular.frq.
```

W tabeli podaję podstawowe cechy różnicujące treść wymienionych 5 plików.

nazwa pliku	zawartość
po.txt	<ul style="list-style-type: none"> • brak polskich liter • brak dużych liter na początku zdań • duże litery tylko w miejscu niektórych polskich liter • użycie = dla oznaczenia skrótowców, np nrd= • użycie / dla oznaczenia nazw własnych • użycie litery z apostrofem w miejscu niektórych polskich liter, np. wskaz'nika, sLon'ca • użycie dużych liter w charakterze niektórych polskich liter np. kotLOW, z'rOdLem • brak spacji przed oznaczeniami

popul.dat	<ul style="list-style-type: none"> • liczne literówki • nazwy własne oznaczone / • użycie = dla oznaczenia skrótowców, np. NRD= • brak dużych liter na początku zdań • brak spacji przed oznaczeniami • polskie litery
npopular.txt	<ul style="list-style-type: none"> • polskie litery • duże litery • nazwy własne pisane dużą literą • wyrazy nieznane zapisane w nawiasach klamrowych (wyrazy niezrozumiałe dla analizatora Pomor) • kod gramatyczny poprzedzony spacją

popular.frq	<ul style="list-style-type: none"> • kod gramatyczny poprzedzony spacją • polskie litery • duże litery • poprawione literówki
popul.3	<ul style="list-style-type: none"> • kod gramatyczny w nawiasach kwadratowych • kod gramatyczny poprzedzony spacją • polskie litery • poprawione literówki • oznaczenie nazw własnych / wewnątrz nawiasów kwadratowych przy formach homonimicznych np. Aleksandra[/141]

Między 5 wymienionymi plikami można było dostrzec pewne podobieństwa. Spostrzeżenie to okazało się istotne przy podejmowaniu decyzji, w jakiej kolejności porównywać wersje. Najbardziej ekonomiczne okazało się:

1. prównanie i ujednolicenie wersji:

popular.frq + npopular.txt → plik1.txt

2. prównanie i ujednolicenie wersji:

popul.dat + popul.3 → plik2.txt

3. prównanie i ujednolicenie wersji:

`plik1.txt + plik2.txt → plik3.txt`

4. prównanie i ujednolicenie wersji:

`plik3.txt + po.txt → wersja_matka.txt`

Na tym etapie pracy ważne było sporządzenie jednolitego zapisu oznaczeń, które mogą pojawić się w tekście. Zapis taki znaleźć można w rozdziale 4 niniejszej pracy.

3.3 Korekta

Część poprawek można było wprowadzić dopiero po sporządzeniu próbnego wydruku treści pliku uznanego za wersję-matkę i zrobieniu korekty na papierze. Należało porównać uzyskany wydruk z oryginalnym tekstem fiszek. Wydruk próbny liczy 663 strony, oryginał próbek podkorpusu znajduje się na 2000 fiszek. Uważne przeglądanie strona po stronie kartek wydruku i fiszek okazało się najbardziej żmudnym fragmentem pracy. Korekta obejmowała poprawienie nielicznych literówek, przywrócenie usuniętych znaków interpunkcyjnych (myślników, średników, dwukropków, wielokropków, apostrofów, nawiasów, cudzysłowów, znaków zapytania i wykrzykników), wprowadzenie ujednoliconych oznaczeń, między innymi w miejscu wykreślonych formuł (zdań urwanych, wykreśleń wewnątrz i na końcu tekstu próbki). Pracę utrudniał brak jakiegokolwiek opisu wykreślonych formuł. Dla ustalenia rodzaju usuniętych znaków interpunkcyjnych pomocnym okazało się sporządzenie tabeli [ryc.4]. Została ona opracowana na podstawie zamieszczonych w słowniku ortograficznym zasad pisania znaków interpunkcyjnych [12].

Zgromadzenie, porównanie i selekcja plików oraz stworzenie jednolitej, przygotowanej do korekty wersji zajęło około 240 godzin. Korekcie (mam na myśli korektę na wydruku, a następnie w pliku — zapisanie na wydruku uwag o błędach, naniesienie na ich podstawie poprawek do kopii pliku wersji-matki) poświęconych było kolejnych 300 godzin.

Znak interpunkcyjny — opis	Kształt graficzny znaku interpunkcyjnego	Wystąpienie wewnątrz zdania	Wystąpienie na końcu zdania	Wykreślenie z fiszek
kropka	.		Tak	
wielokropek	...	Tak	Tak	Tak
pytajnik	?	Tak	Tak	Tak
wykrzyknik	!		Tak	Tak
przecinek	,	Tak		
średnik	;	Tak		Tak
dwukropek	:	Tak		Tak
myślnik	—	Tak		Tak
nawias	()	Tak		Tak
nawias	/ /	Tak		
nawias	[]	Tak		
nawias	{ }	Tak		
cudzysłów	„ ”	Tak		Tak
cudzysłów	« »	Tak		
cudzysłów	» «	Tak		
cudzysłów	” ”	Tak		
maszynowy cudzysłów	’ ’	Tak		

Rycina 4: Występowanie w tekście znaków interpunkcyjnych

4 Oznaczenia metatekstowe

Celem porównań wszystkich plików korpusowych było uporządkowanie komputerowej wersji korpusu. Jednym z istotnych zadań było sporządzenie jednolitych oznaczeń, w których udałooby się zachować wszystkie istotne informacje. Poniżej podaję opis takich oznaczeń:

[/] — nazwy własne, pojawia się bezpośrednio po wyrazie; jeśli forma składa się z części połączonych dywizem (np. nazwiska), oznaczenie pojawia się na końcu formy, np.: *Joliot-Curie*[/]; w przypadku form analitycznych — po każdym wyrazie występuje osobne oznaczenie; często pojawia się tylko przy wyrazach, które były niezrozumiałe dla analizatora, np. *Jack*[/] *Woolams*[/], *Morze Beringa*[/]

[&] — urwany tekst ostatniego zdania próbki; zdanie nie kończy się kropką, znakiem zapytania lub wykrzyknikiem, na końcu tekstu próbki występuje wielokropek, lub nie ma żadnego znaku interpunkcyjnego; przed oznaczeniem występuje spacja, np. *W*[66] *następnym*[261] *okresie zmniejszenie*[111] *wskaźnika zostało*[57] *uzyskane*[211] *przede*[+] *wszystkim dzięki*[63] *wprowadzeniu*[131] *jednostek bardziej nowoczesnych*[222], [&]

[#] — oznaczenie końca próbki, z oryginału której wykreślono więcej niż jeden znak; przed oznaczeniem występuje spacja; pojawia się na końcu próbki, np. *Otrzymuje przeto ona światło*[141] *zmodulowane*[241] [#]

[~] — w treści próbki brak fragmentu tekstu; występuje w miej-

scu wykreślonych formuł, np. wzorów, symboli etc.; przed oznaczeniem występuje spacja, np. *Dwie[31] cechy[112] fizyczne[212] gwiazd dają[501] się zaobserwować w[64] sposób[141] najbardziej bezpośredni[241] i dość pewny[241]: ich[42] temperatura [~] i jasność[111] (dzielnosc[111] promieniowania[121]) [~]*.

[||] — pojawia się przy końcu próbki, w miejscu, w którym na papierowej fiszce między słowoformami zaznaczono pionową kreskę; prawdopodobnie zaznaczenie granicy po 50 wyrazie, np: *Jeszcze sto[34] lat temu[8] zużycie[111] drewna[121] dla[62] celów energetycznych[222] było dwukrotnie większe[211] [||] niż zużycie[111] węgla*.

[+] — występuje pomiędzy wyrazami jednej słowoformy analitycznej, bezpośrednio po pierwszym wyrazie słowoformy; po oznaczeniu występuje spacja, np.: *przede[+] wszystkim*

[=] — oznaczenie skrótowców, np.: *NRD[=]*

[\$] — określa użyte w tekście fiszki nie przyswojone wyrazy obce (bezpośrednie cytaty z języków obcych), np.: *baba[\$]* (słowo arabskie)

[>] — występuje na początku próbki, jeżeli pierwsze zdanie na papierowej fiszce nie rozpoczynało się dużą literą (możliwe, że nie przepisano początku zdania z pozycji źródłowej, gdyż występował na poprzedniej stronie); oznaczenie pojawia się na początku fiszki, np. *[>] wszystkie[212] wymiary[112] w[66] metrach lub stopach angielskich[262]*.

[„] [”] — cudzysłów występujący wokół tytułów, bez względu na to, czy pojawia się na oryginalnej fiszce, np.: *[„]Koszty wytwarzania[”]*

„ ” — cudzysłów oryginalny, występujący na papierowej fiszce

korpusu, np.: „*Manfred[+] Weiss[/]*”, „*bariery[121] cieplnej[221]*”, „*olśnienia[121]*”

[221] — przykładowy kod fleksyjny

[/][121] — przykładowy kod fleksyjny nazwy własnej

(dwie[31] odmiany[112]) — przykładowe użycie nawiasu

Trudno określić czas opracowywania systemu oznaczeń metatekstowych. Propozycje oznaczeń powstawały w trakcie przeglądania treści plików, podczas ich porównywania. Opis oznaczeń mógł powstać dopiero po uważnym przeczytaniu oryginalnych fiszek korpusu.

5 Kilka uwag dotyczących opisu kodu korpusu

Cytowany w dodatku fragment wstępu [Kurcz i in. 1974] można uznać za dokumentację plików korpusu. Wymaga on jednak pewnego komentarza, zakradły się bowiem do niego drobne nieścisłości. Nużące też wydaje się czytanie kilkunastu stron ciągłego tekstu (nie można ograniczać się do przejrzenia zamieszczonego na końcu „Wykazu symboli używanych w opisie morfologicznym słowoform”). Aby uniknąć nadmiernego powtarzania informacji, uwagi moje ograniczę do fragmentów, które moim zdaniem wydają się nieprecyzyjne lub nieczytelne.

Między oznaczeniami w plikach i symbolami opisywanymi we wstępie „List frekwencyjnych” występują pewne rozbieżności:

Imię własne znaczymy znakiem + — informacja ta odnosi się jedynie do SFPW i list frekwencyjnych. W wersji komputerowej znak + (w nawiasie kwadratowym) pojawia się pomiędzy wyrazami jednej słowoformy. Ostatecznie nazwy własne oznaczałam [/]

nazwy geograficzne, np. (...) Morze + Czarne + — instrukcja zawarta we wstępie sugeruje oznaczanie znakiem + każdego wyrazu należącego do nazwy własnej, w komputerowej wersji korpusu oznaczenie nazwy własnej nie pojawia się przy wszystkich wyrazach, np.: *Ocean Lodowaty*[/], *Nowy Jork*[/], *Los* [/] *Angeles*[/]

tytuły książek, czasopism, filmów, referatów — zgodnie z instrukcją, na fiszkach oznaczano tytuły podobnie jak nazwy własne +. (W plikach: / lub [/]). Znak ten stawiano po każdym wyrazie wchodzącym w skład tytułu. Ponadto — wszystkie wyrazy tytułu rozpoczynano dużą literą. Po ujedno-

liceniu — wokół tytułu występuje cudzysłów ujęty w nawiasy kwadratowe, zaś poprzednie oznaczenie zostało usunięte, np.: [„]Warszawski Tygodnik Ilustrowany[”]

Opisowy fragment poświęcony zawartemu w tekście próbek kodowi gramatycznemu (*wstęp do [Kurcz i in.]*) wydaje się nieprecyzyjny. Po jego odczytaniu próbowałam ujednolicić informacje dotyczące występowania konkretnych cyfr:

KOD POZYCYJNY — [xyz]

X – określenie części mowy

Y – zależy od wartości X

Z – zależy od wartości X

X = 1 – rzeczownik

X = 2 – przymiotnik

X = 3 – liczebnik

X = 4 – zaimek

X = 5 – czasownik

X = 6 – przyimek

X = 7 – wykrzyknik

X = 8 – partykuła

X = 9 – spójnik

X – może być liczbą od 1 do 9.

Y – może przyjmować wartość Y_0 , Y_1 , Y_2

Z – może przyjmować wartość Z_0 , Z_1 , Z_2

Jeżeli X przyjmuje wartość 1, 2, 3, 4, 6 to $Y = Y_0$.

Jeżeli X przyjmuje wartość 7, 8, lub 9 to $Y = Y_2$.

Jeżeli X przyjmuje wartość 1 lub 2 to $Z = Z_0$.

Jeżeli X przyjmuje wartość 3, 4, 6, 7, 8, 9 to $Z = Z_2$.

Jeżeli $X = 5$, to $Y = Y_1$, a $Z = Z_1$.

Y_0 – oznaczenie przypadku

Z_0 – oznaczenie liczby

Y_0 – może przyjmować wartość od 1 do 7

Z_0 – może przyjmować wartość 1 lub 2

Y_1 – oznaczenie formy czasownika

Z_1 – określenie zwrotności/ niezwrotności czasownika

$Y_1 = 1$ – bezokolicznik w funkcji form czasu przyszłego

$Y_1 = 2$ – składnik form czasu przyszłego na *-t*

$Y_1 = 3$ – czas przeszły z ruchomą końcówką

$Y_1 = 4$ – tryb warunkowy z ruchomą partykułą *-by*

$Y_1 = 5$ – użycie form czasu teraźniejszego w funkcji trybu rozkazującego

$Y_1 = 6$ – *być, bywać, zostać* jako człon składowy w czasach złożonych

$Y_1 = 7$ – *być, bywać, zostać* jako człon składowy strony biernej

$Z_1 = 1$ – czasownik zwrotny

$Z_1 = 0$ – czasownik niezwrotny

Warto zwrócić uwagę, iż na drugiej pozycji w kodzie gramatycznym, przy wyrazach deklinujących się, może wystąpić cyfra 7. W zdaniu *W obrębie*

rzeczowników [1] i przymiotników [2] rozróżniamy homonimię przypadków (symbole 1-6 na II pozycji) (...) — występują moim zdaniem dwa błędy: na II pozycji mogą pojawić się symbole 1-7, a homonimia przypadków odróżniana jest w korpusie także w obrębie liczebników, zaimków i przyimków.

6 Dane bibliograficzne

6.1 Brakujące dane bibliograficzne

W plikach zawierających próbki korpusu nie umieszczono informacji o źródłach wykorzystanych tekstów. Dane bibliograficzne można było odnaleźć jedynie na papierowych fiszkach korpusu. Pierwszą istotną czynnością było ich przepisanie z papierowych fiszek do komputera.

Treść próbek podkorpusu stylu popularno-naukowego, pochodzi z różnych, wybranych drogą losowania książek. Jedna książka mogła stanowić źródło dla więcej niż jednej próbki. Często te same dane bibliograficzne należało dopisać do kilku próbek. Jeżeli jedna książka została wykorzystana np. do sporządzenia dziesięciu próbek, trzeba było dziesięć razy wpisać dane autora, tytuł książki, rok jej wydania i nazwę wydawnictwa, w którym się ukazała. Taki sposób uzupełniania tekstu wydawał się mało ekonomiczny. Zaznaczanie fragmentów tekstów i kopiowanie ich w wybranym miejscu także nie wydawało się zadawalające. Do stworzenia podkorpusu stylu popularno-naukowego wykorzystano 602 książki. Często brano pod uwagę kilka książek tego samego autora. Należało zminimalizować ilość przepisywanego tekstu. Wystąpienie dwu książek, takiego samego tytułu i tego samego autorstwa, wydanych w jednym roku przez to samo wydawnictwo wydawało się niemożliwe. Po uważnym przejrzaniu fiszek okazało się, że w podkorpusie nie ma książek, które podważyłyby przyjęte przeze mnie założenie.

Można było podzielić dane na kategorie tematyczne. Informacje o autorze (imię i nazwisko), książce (tytuł, wydawnictwo, rok wydania, numer strony, numer wersu) oraz o samej fiszce (numer fiszki, treść próbki) wpisywane były osobno w oddzielnych plikach. Informacja o stylu (części) korpusu — **C** występujące w lewym górnym rogu fiszki — została zanotowana w nazwach plików tekstowych: `autor_c.txt`, `fiszki_c.txt`, `ksiazk_c.txt`.

C.

111

Lew J.: Sto wcieleń żyroskopu. MON 1963 str.36 w.7

Wirnik 111 żyroskopu wraz z 65 układem zawieszenia 121, a więc z 65 wszystkimi obrotowymi ramkami, nazywa 501 się węzłem lub podzespołem żyroskopowym 261. Podzespół 111 żyroskopowy 211 jest podstawowym 251 elementem przyrządu żyroskopowego 221 - jego 42 sercem. Dlatego też specjalna uwaga przy 66 budowie 161 i produkcji 161 przyrządów żyroskopowych 222 zwrócona jest 57 na 64 ten 241 właśnie podzespół 141.

Rycina 5: Wygląd przykładowej fiszki

W plikach tych przechowywane są podzielone tematycznie dane bibliograficzne:

o książce, z której pochodzi próbka (ksiazk_c.txt):

numer książki~tytuł~nazwa wydawnictwa~rok wydania~numer autora

o autorze (autor_c.txt):

nazwisko inicjał imienia.

ewentualnie:

nazwisko inicjał imienia., nazwisko inicjał imienia.

o fiszce (fiszki_c.txt):

numer fiszki~numer strony~numer wersu~numer książki

Tyldy oddzielają od siebie poszczególne dane.

Książki numerowane są w kolejności występowania na papierowych fiszkach. Umieszczenie pliku `ksiazk.c.txt` numeru autora pozwoliło na stworzenie w odpowiednim momencie relacji między danymi autora i książki. Informację o zależnościach między fiszkami i książkami uzyskano dzięki wprowadzeniu do pliku `fiszki.c.txt` numeru książki.

Wpisywanie danych bibliograficznych odbywało się w edytorze GNU Emacs, w trzech oknach (buforach), w których umieszczono trzy wspomniane pliki. Były one uzupełniane równocześnie.

Tekst wpisanej informacji bibliograficznej jest zgodny z tym, który istnieje na oryginalnych fiszkach.

Średnie tempo wpisywania — 50 fiszek na godzinę. Czas wpisania pełnej informacji bibliograficznej jednego podkorpusu — około 40 godzin.

6.2 Połączenie informacji bibliograficznej i próbek podkorpusu

Po wpisaniu do komputera danych bibliograficznych, należało połączyć je z treścią próbek. Najwygodniejsze okazało się użycie bazy danych typu Postgres [8]. Zostały sporządzone 4 tabele, w których następnie umieszczono odpowiednie dane: o autorze, książce, fiszce oraz treść próbki:

```
CREATE TABLE autor (numer_autora int, autor text);
```

```
CREATE TABLE tresc (numer_fiszki int, tresc text);
```

```
CREATE TABLE ksiazka (numer_ksiazki int, autor int, tytul text,  
wydawnictwo text, rok_wydania text);
```

```
CREATE TABLE fiszka (numer_fiszki int, strona text, wers text,  
numer_ksiazki int);
```

Aby postgres mógł wpisać dane do tabeli, należało odpowiednio przygotować zapisany uprzednio w plikach `autor.c.txt`, `fiszki.c.txt`, `ksiazk.c.txt` tekst. Zawartość wymienionych plików została zmieniona za pomocą perlowego skryptu.

```
perl -pe 's/(\d+)\~(.*)/INSERT INTO autor  
VALUES ($1,\x27$2\x27)/g;' <autor.nrm >autor.sql
```

```
perl -pe 's/(\d+)\~(.*)/INSERT INTO tresc  
VALUES ($1,\x27$2\x27)/g;' <tresc.nrm >tresc.sql
```

```
perl -pe 's/(\d+)\~(\d+)\~(.*)\~(.*)\~(.*)/ INSERT INTO ksiazka
VALUES ($1,$2,\x27$3\x27,\x27$4\x27,\x27$5\x27)/g;'
<ksiazka.nrm >ksiazka.sql
```

```
perl -pe 's/(\d+)\~(.*)\~(.*)\~(\d+)/ INSERT INTO fiszka
VALUES ($1,\x27$2\x27,\x27$3\x27,$4)/g;' <fiszka.nrm >fiszka.sql
```

Wpisane do tabel dane udało się połączyć dzięki napisanemu w języku SQL zapytaniu dotyczącemu wszystkich tabel. Zapytanie to zostało zadane Postgresowi przy pomocy specjalnego programu, o nazwie psql (program służący do komunikowania się z bazą danych).

```
select fiszka.numer_fiszki, autor.autor, ksiazka.tytul,
ksiazka.wydawnictwo, ksiazka.rok_wydania, fiszka.strona,
fiszka.wers, tresc.tresc
from fiszka, ksiazka, autor, tresc
where fiszka.numer_ksiazki=ksiazka.numer_ksiazki
and ksiazka.autor=autor.numer_autora
and fiszka.numer_fiszki=tresc.numer_fiszki;
```

Odpowiedź na zapytanie — tabela zawierająca wszystkie interesujące mnie dane — została zapisana w pliku tekstowym, następnie jego treść zmodyfikowano. Istotą zmian było wprowadzenie komend programu T_EX , stanowiącego system do składu tekstu. Tak przygotowany materiał można było wykorzystać do sporządzenia wydruku, na którym znalazła się treść próbek podkorpusu, z pełną informacją bibliograficzną i ujednoliconymi oznaczeniami metatekstowymi.

7 Podsumowanie

Niniejsza praca ma charakter interdyscyplinarny. Opisuje czynności związane z opracowywaniem danych tekstowych przy użyciu komputera. Efektem działań jest ujednolicenie wersji podkorpusu stylu popularno-naukowego korpusu SFPW i sporządzenie wydruku zawartości jego fiszek. Łączny czas wykonanych czynności wynosi 630 godzin. Do tekstu pracy dołączony został CD-ROM, zawierający między innymi ujednoliconą wersję podkorpusu. Wydruk stanowi aneks niniejszej pracy.

Dodatek

Opis kodu korpusu

W każdym tomie zdającym sprawę z prac nad listami frekwencyjnymi [Kurcz i in.], we wstępie zamieszczono instrukcje wpisywania tekstów na fiszki oraz „Opis gramatyczny”. Jest to cenne źródło informacji o kodzie, użytym w korpusie do odróżnienia form homonimicznych. Tekst drukowany w latach siedemdziesiątych wydaje się jednak mało czytelny⁸. Poniżej przytaczam fragmenty wstępu trzeciego tomu, z wprowadzonymi nawiasami w miejsce pojedynczych ukośników i pojedynczych ukośników w miejsce ukośników podwójnych. Podkreślenia wyrazów linią przerywaną, odróżniające w tekście przykłady analizowanych form, zastąpiłam przez pochylony krój czcionki. Używam polskiego cudzysłowu. Cyfry stanowiące oznaczenia gramatyczne, jeżeli zawierają tzw. cyfrę na pierwszej pozycji (oznaczenie części mowy), podaję w nawiasach kwadratowych (tak, jak w komputerowej wersji próbek).

Wstęp

1 Jednostka badania

Ze względu na leksykalno-morfologiczny typ podjętych badań frekwencyjnych podstawową jednostką analizy w tekście jest słowoforma, czyli forma fleksyjna wyrazu. Wyraz definiujemy — z pewnymi wyjątkami — według definicji graficznej, określając go jako ciąg liter zawartych między dwoma

⁸Autorzy książki nie mieli możliwości pisania pewnych znaków, np. nie mogli używać nawiasów — w ich miejsce wpisywali ukośnik; posługiwali się także podwójnym ukośnikiem.

kolejnymi pauzami. Takie ujęcie było konieczne przy maszynowym opracowaniu danych. Z tego też względu formy czasów złożonych typu *będę jechał*, *pojechali byli* traktuje się jako składające się z dwóch odrębnych słowoform; funkcje tych słowoform jako składowych form czasów złożonych wskazuje jedynie specjalny kod liczbowy (por. Kod fleksyjny str. 40). Podobnie — zgodnie z zasadą analityczności — wyróżnia się dwie odrębne słowoformy w zestawieniach nominalnych typu *Stany Zjednoczone*; jako ciągi osobnych słowoform traktujemy także zestawienia liczebników porządkowych i głównych typu *sześćdziesiąty piąty*, *tysiąc dwieście czternaście* itp.

Od reguły tej stosowano jednakże wyjątki, ujmując niekiedy jako jedną słowoformę segment tekstu składający się z dwóch formalnie różnych członów. Dotyczyło to następujących wypadków:

1. Dwa formalnie odrębne człony stanowią jednostkę zleksykalizowaną, przy czym przynajmniej jeden z członów utracił samodzielność, stanowiąc archaizm leksykalny bądź fleksyjny — są to wyrażenia typu *po prostu*, *za młodu*, *po omacku*, *na bakier* itp.
2. Dwa formalnie odrębne człony nie dopuszczają wstawienia między nie innego członu, np. *w ogóle*, *przede wszystkim*, *na razie* itp. Tę zasadę stosowano tylko do niektórych leksemów wchodzących w skład utartych połączeń.
3. Częstki *de*, *von* stojące przed obcymi nazwiskami stanowią wraz z nazwiskiem jedną słowoformę (*de Gaulle*, *von Hassel*).

Jako jedną słowoformę traktuje się także:

1. imiesłowy przymiotnikowe i rzeczowniki odczasownikowe z zaimkiem *się* (*skradający się*, *skradanie się*);
2. częstki *co* i *jak* stojące przed przymiotnikami i przysłówkami w stopniu najwyższym wraz z danymi częściami mowy (*co naj-*, *jak naj-*);

3. cząstkę *do* + przysłówkę (typ *do dziś*);
4. cząstkę *na* + przysłówkę (typ *na gorąco, na bieżąco, na nowo, na zawsze*);
5. cząstkę *od* + przysłówkę (typ *od wczoraj, od dziś*);
6. cząstkę *z* + przysłówkę (typ *z zewnątrz*);
7. cząstkę *za* + przysłówkę/przymiotnik (typ *za dużo, za mało, za duży, za mały*).

Jeszcze inaczej potraktowano czasowniki zwrotne, zaliczając do nich z punktu widzenia formalnego wszystkie czasowniki z zaimkiem *się*. Ze względu na definicję graficzną wyrazu oraz biorąc pod uwagę fakt, że informacja o częstotliwości zaimka *się* jako odrębnej słowoformy jest ważna dla językoznawców, interpretowano czasownik zwrotny jako składający się z dwóch słowoform (typ *biją się*). Aby jednakże nie utracić informacji językowej o zwrotności danego czasownika, umieszczano przy słowoformie czasownika symbol 501 jako znak kodu fleksyjnego wskazujący na zwrotność (*bili 501 się*). Stąd też wszystkie słowoformy czasowników zwrotnych (z wyjątkiem imiesłówów przymiotnikowych traktowanych jako osobne hasła) opatrzone są symbolem 501. Program hasłowania pozwalał w toku opracowania na przekształcenie symbolu 501 na zaimek *się* przy hasle, stąd „pełna” postać haseł czasowników zwrotnych (por. hasło *bić się*).

W wielu wypadkach wątpliwych dotyczących postaci hasła przyjmowano rozstrzygnięcia zawarte w „Słowniku Języka Polskiego” pod red. W. Doroszewskiego.

Instrukcja I

1 Rozpisywanie tekstów

1. Skróty rozwiązywano, traktując je jako normalne słowoformy: *prof.* = *profesor*, *dla dra Kowalskiego* = *dla doktora Kowalskiego*, podobnie *mgr*, *doc.*, *tow.*, *ob.* i wszelkie inne — takie jak: *np.* = *na przykład*, *itd.* = *i tak dalej*, *br.* = *bieżącego roku*, *itp.* = *i tym podobne*.
2. Jednostki miar pisane skrótami rozwiązywano: *kg* = *kilogram*, *cm* = *centymetr*, *godz.* = *godzina*.
3. Liczby pisane cyframi arabskimi i rzymskimi przepisywano słownie:

po XX zjeździe = *po dwudziestym zjeździe*, *o godz. 9* = *o godzinie dziewiątej*, *w 1966 r.* = *w tysiąc dziewięćset sześćdziesiątym szóstym roku*.

4. Złożenia z członem cyfrowym przepisywano słownie: *50-lecie* = *pięćdziesięciolecie*, *3-dniowy* = *trzydniowy*.
5. Skrótownice (literowce, głoskowce, sylabowce) nie były rozwiązywane i traktowano je jako osobne słowoformy: *bombardowanie DRW* = *bombardowanie DRW*, *rocznica PKWN* = *rocznica PKWN*, *prezydium PAN* = *prezydium PAN*, *w PAN-ie* = *w PAN-ie*, *w PGR-ach* = *w PGR-ach*, *Pafawagiem* = *Pafawagiem*.
6. Pominięto w tekście wszelkie wzory oraz symbole typu H_2O , odcinek *AB* itp.

Instrukcja II

1 Imiona własne

Imię własne znaczymy znakiem +.

1. Za imię własne uważa się:

- (a) imiona i nazwiska ludzi i zwierząt (*Jan + Kowalski +, Azor +*); także, jeśli są homonimiczne z apelatywami (np. *Puzon +, Gołąb +*);
- (b) nazwy umowne instytucji, pojazdów lądowych i wodnych (*kawiarnia Sarenka +, statek Orzeł +, ekspres Błękitna + Fala +*);
- (c) nazwy geograficzne, np. *Wielka + Brytania +, Związek + Radziecki +, Stany + Zjednoczone +, Morze + Czarne +, Bagno + (dzielnica Warszawy), Bliski + Wschód +*;
- (d) tytuły książek, czasopism, filmów, referatów (*Ogniem + i + mieczem +, Życie + Gospodarcze +, Tam + gdzie + rosną + poziomki +, O + aktualnej + sytuacji + politycznej + w + świecie +*).

2. Nie są imionami własnymi:

- (a) pełne, nieumowne nazwy instytucji typu *Spółdzielnia pracy szewców i cholewkarzy, Centrala Handlowa Przemysłu Mleczarskiego*;
- (b) przymiotniki utworzone od imion własnych, np. *bliskowschodni, amerykański, warszawski*;
- (c) nazwy mieszkańców danego regionu typu *poznanianin, Amerykanin, Chińczyk*;
- (d) nazwy firmowe samochodów, papierosów, słodczy, np. *moskwicz, wołga, warszawa, sporty, giewonty, katarzynki* itp.

3. Znakiem + znaczone także umowne symbole klas sportowych (*drużyna zakwalifikowała się do klasy a +*) oraz witamin (*witamina b +*).

Opis gramatyczny

1 Kod fleksyjny

Leksykalno-morfologiczne dane frekwencyjne z naszego opracowania dotyczą częstości zarówno haseł, jak i form fleksyjnych. Przy ustalaniu haseł na podstawie słowoform występujących w izolacji napotykamy na różne typy form homonimicznych. Jak wiadomo, istnieją 3 typy homonimii:

1. homonimia leksykalna typu *pokój, zamek*;
2. homonimia morfologiczna dotycząca form fleksyjnych; jest nią w obrębie nomen homonimia przypadków i liczb (*sprawności* jako gen., dat., loc., sg. i gen. pl., *dobrym* jako instr., loc. sg. oraz dat. pl.);
3. homonimia syntaktyczna dotycząca poszczególnych części mowy, polegająca na tym, że dany wyraz może pełnić funkcje syntaktyczne rzeczownika bądź czasownika (*brak, potrzeba*), rzeczownika bądź przysłówka (*czasem*), rzeczownika bądź przymiotnika (*chory, pośpieszny*), czasownika bądź przysłówka (*wolno*), przysłówka bądź przyimka (*blisko, obok, wokół*) itd.

W opracowaniu naszym pominięto homonimię leksykalną, przede wszystkim ze względu na trudności teoretyczne w odróżnianiu homonimii leksykalnej od polisemii. Homonimię morfologiczną i syntaktyczną usuwano, różniując formy homonimiczne przez dopisywanie umownych symboli cyfrowych. Ze względu na ograniczoną pojemność pamięci maszyny liczba symboli gramatycznych została ograniczona do 63; ułożony kod jest kodem pozycyjnym.

Zwracamy uwagę, że w opracowywanym materiale leksykalnym symbolizacja odpowiednich cech gramatycznych dotyczy tylko słowoform i haseł homonimicznych, nie obejmuje więc ona całości badanego słownictwa. Słowoformy i hasła nie kodowane oznaczają więc formy nie będące homonimami, których funkcje morfologiczno-syntaktyczne łatwo odczytać z samej postaci wyrazu (por. *domami*).

Tylko w dwóch wypadkach — przy rzeczownikach i przy liczebnikach brak kodu ma jeszcze inne, odrębne znaczenie. Przy rzeczownikach oznacza on także nieodmienność członu, pełniącego funkcję przydawki rzeczownikowej (przykłady typu *w hucie Pokój +*, *do spółdzielni Jedność +*), a także występującego przy odmianie tytułów i nazwisk żeńskich (*z doktor Kowską +*, *nagrodę przyznano Janinie + Bator +*). Przykłady dotyczące liczebników wymagają szerszego omówienia i podane są niżej (por. str. 50).

Homonimie syntaktyczną usuwamy, umieszczając przy homonimicznym hasle symbol części mowy. Są to cyfry od 1 do 9 występujące *w pierwszej pozycji* po słowoformie. Interpretuje się je następująco:

1. „rzeczownik” — symbol przypisywany rzeczownikom i substantywizowanym przymiotnikom;
2. „przymiotnik” — symbol przypisywany przymiotnikom, liczebnikom i zaimkom przymiotnikowym oraz imiesłowom przymiotnikowym;
3. „liczebnik” — symbol przypisywany liczebnikom głównym i zbiorowym oraz zaimkom liczebnym (*ile, tyle, wiele*);
4. „zaimek” — symbol oznaczający zaimki rzeczownikowe;
5. „czasownik” — symbol oznaczający formy czasownika i prädicativa (*potrzeba, brak*);
6. „przyimek” — symbol oznaczający przyimki pierwotne i wtórne (*blisko, obok*);

7. „wykrzyknik” — symbol odróżniający wykrzykniki od spójników (*a, o*);
8. „przysłówek” lub „partykuła” — symbol podawany przy słowoformach nieodmiennych nie będących przyimkami ani spójnikami;
9. „spójnik”.

Pozycja II występuje jedynie wówczas, gdy w pozycji I użyto symboli 1–6. W obrębie rzeczowników [1] i przymiotników [2] odróżniamy homonimie przypadków (symbole 1–6 na II pozycji) i liczb (symbole 1–2 na III pozycji). Nie uwzględniono natomiast homonimii w obrębie kategorii rodzaju, gdyż pomnożyłoby to znacznie liczbę symboli.

Przy czasowniku [5] II pozycja, na której występują cyfry od 1 do 4, pozwala odróżnić użycie bezokolicznika w funkcji składnika form czasu przyszłego od innych jego użyć, oraz trzy różne funkcje formy na -ł (składnik form czasu przyszłego, czas przeszły z ruchomą końcówką i tryb warunkowy z ruchomą partykułą). Symbol 5 na II pozycji wyróżnia użycie form czasu teraźniejszego w funkcji trybu rozkazującego. Z kolei symbole 6–7 na II pozycji zarezerwowano dla wyróżnienia niektórych funkcji gramatycznych czasowników *być*, *bywać*, *zostać* (człony składowe w czasach złożonych i w formach strony biernej).

Na III pozycji występuje cyfra 1 przy czasownikach zwrotnych lub 0 (czyli brak symbolu cyfrowego) przy czasownikach niezwrotnych.

2 Homonimia syntaktyczna

Ze względu na brak wypracowanych ścisłych kryteriów pozwalających na odróżnienie poszczególnych części mowy, przyjęto przy kodowaniu pewne sprawdziany o charakterze roboczym. Starano się ustalać kryteria wyłącznie syntaktyczne, abstrahowano całkowicie od kryteriów semantycznych. Szcze-

gółowy opis przyjętych przez nas zasad rozróżniania części mowy został opublikowany osobno (Biuletyn PTJ nr XXIX, 1971), tu ograniczymy się do skrótowego omówienia tychże zasad.

Przyjęte kryteria będziemy podawać, omawiając kolejno parami części mowy, których funkcje mogą pełnić opisywane homonimy syntaktyczne. Dla jasności oznaczamy określony rodzaj homonimii syntaktycznej symbolami (SV = subst./verbum = 1/5, SAdv, VA itp.).

2.1 Rzeczownik czy czasownik (typ SV = 1/5)

Homonimia ta dotyczyła wyrazów *potrzeba*, *brak*, *szkoda*, *strach*, *żał*, *wstyd*.

Kryterium: kwalifikator czasownika [5], jeżeli w danym zdaniu występuje względnie można wstawić człon (*było*) *nam* = zaimek osobowy w celowniku przy nieobecności innego członu predykatywnego.

S: *brak*[1] *ten odczuwano bardzo boleśnie*;

V: *w dyskusji brak*[5] *nam było ładu*.

2.2 Przymiotnik czy czasownik (typ AV = 2/5)

Homonimiami syntaktycznymi tego typu są hasła *winien* albo *winny*. Hasło *winien/winny* kwalifikowano:

1. jako czasownik [5], jeżeli w zdaniu towarzyszy mu bezokolicznik: *weryfikacja winna* [5] *objąć wszystkich członków*;
2. jako przymiotnik [2], jeżeli pełni funkcję przydawki lub orzecznika: *człowiek winny* [2] *zbrodni*; *winni* [2] *są rodzice, nie dzieci*; także *jest mi winien* [2] *sto złotych*.

2.3 Przysłówek czy czasownik (typ AdvV=8/5)

Homonimia ta dotyczy wyrazów *wolno* oraz *przeszło*.

Wolno kwalifikujemy:

1. jako przysłówek [8], jeżeli jest możliwa substytucja wyrazów *pomału*, *powoli*: *jechał wolno* [8];
2. jako czasownik [5], jeżeli w danym zdaniu występuje względnie można wstawić człon (*było*) *nam* = zaimek osobowy w celowniku: *nie wolno (nam palić)*.

Przeszło kwalifikujemy jako przysłówek [8], jeżeli jest możliwa substytucja wyrazów *więcej niż*, w pozostałych wypadkach kwalifikujemy jako czasownik [5]: *wydobyto przeszło [8] tysiąc ton węgla*.

2.4 Rzeczownik czy przymiotnik (typ SA = 1/2)

Homonimia ta dotyczy substantywizowanych przymiotników. Ze względu na różne stopnie tej substantywizacji przyjęto sformułowane przez S. Jodłowskiego formalne kryterium, które wskazuje przy przymiotnikach substantywizowanych na ograniczenie form rodzaju do dwóch lub nawet do jednego rodzaju. Np. *chory* ma tylko 2 rodzaje w funkcji rzeczownika [1]. Używając tego kryterium kwalifikowano jako rzeczowniki [1] takie wyrazy jak *przewodniczący*, *biały*, *młody* itp. A także słownictwo typu *wytyczna*, *średnia*, *dane*, również ułamki (*dziesiąta*, *setna*). Przykłady kontekstów: *otrzymano nowe wytyczne*[1], *średnia*[1] *krajowa w zakresie spożycia*, *dane*[1] *liczbowe* itp.

2.5 Rzeczownik czy przyimek (typ SPrp = 1/6)

Homonimia ta dotyczy niektórych rzeczowników, które w pewnych wypadkach mogą pełnić funkcje przyimków. W badanych tekstach były to rzeczowniki *celem*, *drogą*, *koło*, *skutkiem*, *względem*.

Jako kryteria pozwalające uznać dany człon za przyimek [6] przyjęto:

1. możliwość zastosowania określonych transformacji lub substytucji,

2. relację danego członu z innymi członami ciągu syntaktycznego (kontekst).

Celem + gen. nom. actionis jest przyimkiem [6], jeżeli jest możliwa transformacja tego wyrażenia w podrzędne zdanie okolicznikowe celu, tzn. gdy zachodzi relacja:

celem + gen. nom. act. \rightarrow *aby* + infinitivus/verbum finitum.

Przykład: *Rada Bezpieczeństwa zebrała się celem [6] przedyskutowania sytuacji politycznej* \rightarrow ... *aby przedyskutować sytuację polityczną*. Jeżeli podobna transformacja nie jest możliwa, *celem* uzyskuje kwalifikator rzeczownika [1].

Drogą + gen. nom. actionis jest przyimkiem [6], jeżeli jest możliwa substytucja wyrażenia przyimkowego *przez* + acc. nom. act.: *drogą [6] pertraktacji uzyskano ustępstwa* \rightarrow *przez pertraktację*. . . . Natomiast *drogą* jest rzeczownikiem [1], jeżeli:

1. jest członem określającym,
2. nie jest możliwa powyższa substytucja.

Skutkiem + gen. oraz *względem* + gen. kwalifikujemy jako przyimki [6], jeżeli są możliwe odpowiednie substytucje: dla *skutkiem* substytucja *wskutek*, dla *względem* substytucja *wobec*:

skutkiem [6] nieostrożności spowodowano pożar \rightarrow *wskutek nieostrożności*. . . , *mieć względem [6] kogoś zobowiązania* \rightarrow *mieć wobec kogoś*. . . . Jeżeli zaś substytucje te nie są możliwe, kwalifikujemy oba te człony jako rzeczowniki [1].

Koło + gen. kwalifikujemy jako przyimek [6], jeżeli jest możliwa substytucja wyrazu *obok*: *koło [6] lasu* \rightarrow *obok lasu*. W wypadku dwuznaczności (*koło wozu*) o możliwości lub niemożliwości tej substytucji rozstrzyga kontekst, a więc relacja z innymi członami ciągu syntaktycznego: *koło [6] wozu pasły się konie, ale pękło nam tylne koło [1] wozu*.

2.6 Zaimki

Zaimki klasyfikujemy wyłącznie na podstawie kryterium syntaktycznego, wyodrębniając tylko dwie klasy zaimków: zaimki o odmianie przymiotnikowej, mające homonimiczne formy przypadka i liczby (*których, jakie*) oraz zaimki o odmianie rzeczownikowej, mające homonimiczne formy przypadka (*kogo, czym*). Zaimki przymiotnikowe otrzymują symbol [2], zaimki rzeczownikowe — symbol [4]. Tak więc nie uwzględniamy — z bardzo małymi wyjątkami (por. „Hasła kodowane umownie”, str. 54) klasyfikacji semantycznej zaimków (zaimki względne, pytajne, nieokreślone itp.).

Rozróżnienie między zaimkami przymiotnikowymi i rzeczownikami nie jest kłopotliwe, ulega ono jedynie zmianie w odniesieniu do zaimków *to, wszystko, samo, jedno*. Mianowicie:

1. Zaimki *ten, ta, to; wszystek, wszystka, wszystko; sam, sama, samo; jeden, jedna, jedno* są traktowane jako zaimki przymiotne [2], jeżeli występują :
2. w liczbie mnogiej,
3. w liczbie pojedynczej w funkcji przydawki. Przykłady: *ci ludzie; zaprosź tych, których znasz; zastanawia mnie to zagadnienie; wszyscy ludzie; wszystkim się zdawało, że Wojski wciąż gra jeszcze; kot wypił wszystko mleko; w zebraniu uczestniczyli sami cudzoziemcy; dokonały tego zupełnie same, bez niczyjej pomocy; błąd tkwił już w samym założeniu; jedne pary szły na prawo, drugie na lewo; jedni chcą tak, inni inaczej; jedna jaskółka wiosny nie czyni; jedna pani powiedziała, że widziała twoją córkę z chłopcem.*
4. Zaimki *to, wszystko, samo, jedno* są traktowane jako zaimki rzeczownikowe [4], jeżeli występują w liczbie pojedynczej w funkcji podmiotu, dopełnienia lub przydawki dopełniaczowej. Przykłady:

to [4]: *dążył do tego, aby się zemścić; chodziło nam o to, aby jak najszybciej ukończyć studia; bardzo się temu sprzeciwiano; rzecz w tym, że już nie dysponujemy na to funduszami; także jest to niebagatelny problem;*

wszystko [4]: *troska o wszystko; niedobrze we wszystkim dopatrywać się zła; wszystkiego się można po nim spodziewać; wszystkiemu można by zaradzić;*

jedno [4]: *jedno jest dla mnie oczywiste — że brak im wytrwałości; przykład ten świadczy o jednym — o ich wzajemnym przywiązaniu;*

samo [4]: *to samo mam na myśli; tym samym projekt zatwierdzono; chodziło nam wszystkim o to samo.*

Uwaga: w połączeniach *to samo*, *to wszystko*, *to jedno* przyjęto umownie *to* jako człon określający (przydawkę = [2]), *samo*, *wszystko*, *jedno* jako człon określany (podmiot, dopełnienie = [4]): *to samo chciałem powiedzieć, mieliśmy szczęście w tym wszystkim; to jedno warto zapamiętać.*

Zaimek *to* jest także homonimiczny ze spójnikiem [9]: *jeżeli...to, chociaż...to, skoro...to, to...to*, jednakże odróżnienie *to* nie sprawiało przy opisie żadnego kłopotu.

Człon *to* kwalifikowano także jako partykułę [8], jeżeli w danym ciągu składniowym człon ten można było zerować: *coraz to* [8] *chłodniej ~ coraz chłodniej; gdyby to* [8] *chodziło o was ~ gdyby chodziło o was; kiedy to* [8] *wreszcie skończą się kłopoty ~ kiedy wreszcie skończą się kłopoty*; także: *brat długi zwrócił i to* [8] *jeszcze przed terminem ~ brat długi zwrócił i jeszcze przed terminem.*

Zaimek *tym* (= forma fleksyjna zaimka *ten*) jest homonimiczny ze spójnikiem [9] w połączeniach *i...tym*: *im* [9] *dalej w las, tym* [9] *więcej drzew* oraz ze spójnikiem w połączeniach stopnia wyższego przysłówków i przymiotników: *tym* [9] *bardziej, tym* [9] *większa radość* itp., odróżnienie *to* nie wymagało jednak formułowania osobnych kryteriów.

Bardziej szczegółowy opis uzyskał także zaimek rzeczownikowy *co*. Oprócz

normalnych wyróżnień homonimicznych form przypadków (*co* nom. i acc., *czym* inst. i loc. itp.) kwalifikatorem mianownika [41] opatrywano *co* występujące przy stopniu wyższym przysłówka (*co gorzej, co więcej*) oraz wyróżniano umownie symbolem [6] *co* występujące przy miarach czasu i odległości (typu *co* [64] *godzinę, co* [62] *roku*). Także umownie wyróżniano *co* w funkcji zaimka względnego, opatrując je symbolem [9]: *ten, co* [9] *nie miał szczęścia*, oraz *co* w funkcji partykuły [8]: *dopiero co*.

Na osobną wzmiankę zasługują wreszcie zaimki liczebnikowe *tyle, wiele*. Oba te człony traktowano:

1. jako liczebniki [3] mające homonimiczne formy przypadków, jeżeli były one członami określającymi przy nomen (*w tylu* [36] *sprawach, z tyloma* [35] *ludźmi, w wielu sprawach, z wieloma ludźmi*;
2. jako przysłówki [8], jeżeli były one nieodmiennymi członami określającymi przy verbum (*wiele* [8] *można zrobić; tyle* [8] *widziałem, że nie wiem, od czego zacząć*), a także — umownie — jeżeli następuje po nich wyrażenie przyimkowe: *to rozwiązanie pozostawia wiele* [8] *do życzenia; miał jeszcze wiele* [8] *do zrobienia*.

W połączeniach *o tyle ... o ile, o tyle ... że, nie tyle ... ile, tyle ... że* człon *tyle* wchodzi w skład większego wyrażenia, stanowiącego spójnik; podobnie jak w wypadku członu *to*, także i ten przykład homonimii syntaktycznej nie stwarzał trudności przy opisie i nie wymagał stosowania żadnych specjalnych kryteriów.

2.7 Liczebniki

Skomplikowany i nieregularny charakter tej kategorii w języku polskim nastręczał mnóstwo kłopotów w opisie. Jak już wspomniano, wyodrębniono w zasadzie 2 klasy liczebników: liczebniki o odmianie przymiotnikowej [2] mające homonimiczne formy w zakresie przypadku i liczby (por. *pierwszym, jednej*

oraz pozostałe mające homonimię tylko w zakresie przypadku (por. *pięciu, dwóch*).

Przy liczebnikach głównych odrębne znaczenie ma także brak kwalifikatora. Dotyczy to:

1. liczebników głównych występujących jako człony nieodmienne w wielo-
członowych zestawieniach mających funkcję liczebników porządkowych:
w sto dwudziestym pułku; od tysiąc dziewięćset piętego roku;
2. liczebników głównych w funkcji nieodmiennych przydawek ilościowych
(przykłady częste w prasie codziennej i w publicystyce): *zwyciężył Ko-
walski w czasie pięć godzin dziewięć minut piętnaście sekund; ze stopą
wzrostu osiem procent, w latach tysiąc dziewięćset sześćdziesiąt – sześć-
dziesiąt trzy; zwyciężyli Belgowie w stosunku pięć – cztery.*

Dodatkowy opis dotyczy dwóch funkcji liczebnika *jeden*. Jak już wspo-
mniano przy omawianiu zaimków, przy członach *jeden, jedna, jedno* nie roz-
różniamy funkcji liczebnika i zaimka nieokreślonego. Natomiast wyraz *jeden*
otrzymuje inne kwalifikatory w następujących dwóch wypadkach:

1. umownie kwalifikator [8], jeżeli jest on składnikiem odmiennego liczeb-
nika głównego: *w dwudziestu jeden [8] krajach, delegaci trzydziestu jeden
[8] państw;*
2. bez kwalifikatora, jeżeli występuje jako liczebnik główny w funkcji ilo-
ściowej przydawki nieodmiennej: *miara o długości jeden metr trzydzie-
ści centymetrów; w czasie dwie godziny dwadzieścia jeden minut.*

2.8 Przysłówek czy przyimek (typ Adv/Prp = 8/6)

Jest to rozróżnienie bardzo trudne, a dotyczyło takich członów jak *blisko, obok, dokoła, ponad, wokół, poniżej, powyżej*. W gramatykach polskich brak szczegółowego opisu odrębnych (dwojakich) funkcji tych członów, także w

opisie tych haseł w „Słowniku języka polskiego” pod red. W. Doroszewskiego nie rozgranicza się tych funkcji (podano je wyłącznie przy hasłach *około* i *obok*).

Kryterium przyjęte przez nas brzmi następująco:

1. badany człon zaliczamy do klasy przyimków [6], jeżeli przypadek następującego po nim nomen jest zdeterminowany przez człon badany;
2. badany człon zaliczamy do klasy przysłówków [8], jeżeli przypadek następującego po nim nomen jest zdeterminowany przez inny człon, a człon badany można zastąpić innym przysłówkiem. Kryterium to wyjaśnimy na przykładzie kilku spośród przytaczanych członów, cytując pod:
 - (a) klasyfikację danego członu jako przyimka [6],
 - (b) klasyfikację jako przysłówka [8].

Okolo:

1. *utwory te miały około [6] dwustu nagrań; ankieta dała około [6] tysiąca odpowiedzi; w strajku brało udział około [6] miliona pracowników; około [6] południa zaczęło zbierać się na deszcz;*
2. *człon około można zastąpić wyrazami mniej więcej, prawie: drużyna wróciła do kraju z około [8] dwustu punktami; cała liczba tomów równa się sumarycznie około [8] ośmiu tysiącom stron; artykuły przemysłowe stanowią około [8] czterdzieści procent obrotów.*

Blisko:

1. *pociąg zatrzymał się blisko [6] stacji; byli już blisko [6] celu swej wędrówki;*

2. człon *blisko* można zastąpić wyrazami *mniej więcej, serdecznie, unikliwie*: *wydobyto blisko [8] trzy tysiące ton węgla; blisko [8] połowę załogi stanowili kolorowi; byli z sobą blisko [8] już od wielu lat; trzeba go znać bardzo blisko [8], aby móc go ocenić.*

Ponad:

1. *istnieją obecnie próby wznoszenia komunikacji ponad [6] ziemię; ponad [6] miastem unosiły się ciężkie chmury;*
2. człon *ponad* można zastąpić wyrazem *przeszło*: *drużyna wróciła do kraju z ponad [8] dwustu punktami; na eksport przeznaczają się ponad [8] dziesięć procent produkcji; przeciętnie gospodarstwa te posiadają ponad [8] sto hektarów ziemi.*

Wokół:

1. *dyskusja rozwinęła się wokół [6] poruszonego w referacie zagadnienia; wokół [6] stadionu zbierały się tłumy;*
2. człon *wokół* jest członem określającym przy czasowniku: *nie orientował się, co się wokół [8] dzieje.*

Analogiczne kryteria dotyczą członów *obok, dokoła, poniżej, powyżej*.

Na zakończenie uwag dotyczących opisu liczebników oraz omówionych powyżej typów przysłówków/przyimków trzeba przypomnieć, iż decyzja rozpisywania w formie słownej występujących w tekstach liczebników w znacznej mierze uzależnia formy rozpisywane od indywidualnej wymowy maszynistek, co — ze względu na dużą chwiejność w zakresie uzusu — sprawia, iż frekwencje pewnych form liczebników uznać trzeba za mało dokładne.

2.9 Przyimek czy spójnik (typ Prp/Coni. = 6/9)

Homonimia ta dotyczyła członów *podczas* oraz *mimo* i *pomimo*. Kwalifikowano je jako przyimki [6], gdy dany człon determinował przypadek następu-

jącego po nim nomen: *podczas* + gen., *mimo* + gen., acc., *pomimo* + gen., acc. (*podczas zimy*, *mimo gniewu*, *mimo to*, *pomimo licznych próśb*, *pomimo to*). Jako spójniki [9] uznano natomiast w całości hasła: *podczas gdy*, *podczas kiedy*, *mimo że*, *mimo iż*, *pomimo że*, *pomimo iż*.

2.10 Spójnik czy partykuła (typ Coni./Part. =9/8)

Rozróżnienie to zastosowano tylko do morfemów *by*, *aby*, zaniechano go przy morfemie *a* ze względu na trudność znalezienia dostatecznie wyrazistego kryterium. *By* oraz *aby* są kwalifikowane jako spójniki [9], gdy zaczynają zdania podrzędne okolicznikowe celu lub zdania dopełnieniowe: *pojechał na stację, by* [9] *spotkać się na dworcu ze znajomym*; *prosił by* [9] *go możliwie szybko odwiedzić*; *proszono uczestników wycieczki, aby* [9] *nie oddalali się od miejsca zbiórki*; *wyjechał, aby* [9] *uregulować swoje sprawy majątkowe*.

Morfem *by* kwalifikowano jako partykułę [8], gdy pełnił funkcję ruchomego morfemu w formach trybu warunkowego: *on by* [8] *to zrobił*, *trzeba by* [8] *się na to zdecydować*. Równocześnie leksem danego czasownika otrzymuje kwalifikator trybu warunkowego z ruchomą partykułą (*zrobił* [54]). Natomiast morfem *aby* uznawano za partykułę [8] w wypadku, gdy w danym ciągu syntaktycznym morfem ten można było wyzerować: *czy aby* [8] *nie za późno na wizytę* ~ *czy nie za późno na wizytę*.

3 Hasła kodowane umownie

Zbyt mała liczba (63) symboli kodu fleksyjnego nie pozwalała na bardziej szczegółowy opis gramatyczny. Jednakże w kilkunastu wypadkach, chcąc objąć kodowaniem pewne formy nie uwzględnione w spisie rejestrowanych cech gramatycznych, wykorzystano przyjęte symbole w sposób całkowicie umowny, niezgodny z przyjętymi znaczeniami kodu. O niektórych tego typu wypadkach już wspomniano powyżej.

Chcąc ułatwić Czytelnikom korzystanie z opublikowanych materiałów, podajemy kodowanie umowne hasła w formie listy alfabetycznej, co pozwoli odszukać szybko odmienne znaczenie danego symbolu przy określonym hasle.

aby	8	w kontekstach typu <i>czy aby</i> [8] <i>nie za późno</i> ;
co	61-64	przy miarach czasu i odległości, gdy jest możliwa substytucja członu <i>każdy</i> : <i>co</i> [61] <i>godzina</i> , <i>co</i> [64] <i>godzinę</i> , <i>co</i> [62] <i>roku</i> ;
co	8	w kontekstach typu: <i>wielu co</i> [8] <i>ambitniejszych</i> , <i>nie ma się co</i> [8] <i>dziwić</i> ;
co	9	w funkcji zaimka względnego: <i>ten, co</i> [9] <i>nie miał szczęścia</i> ; <i>ci, co</i> [9] <i>dobiegli pierwsi</i> ;
ile	8	w kontekstach typu: <i>było tyle</i> [8] <i>wody</i> , <i>ile</i> [8] <i>się jej w wiadrze zmieści</i> ;
jak	8	w funkcji zaimka pytajnego: <i>jak</i> [8] <i>to zrobić?</i> z <i>jak</i> [8] <i>dużą dokładnością?</i> ; także jako partykuła w kontekstach typu: <i>udało mi się i to jak</i> [8];
jak	9	porównawcze: <i>gryka jak</i> [9] <i>śnieg biała</i> ; zaczynające zdanie niepytające: <i>jak</i> [9] <i>zdobędę skierowanie</i> , <i>przyjadę</i> ; <i>jak</i> [9] <i>wzeszło słońce</i> , <i>wojska ruszyły</i> ; + przysłówki w stopniu równym: <i>jak</i> [9] <i>dalece</i> , <i>jak</i> [9] <i>dotychczas</i> ;
jako	61-66	w kontekstach typu: <i>on jako</i> [61] <i>znawca</i> , <i>jego jako</i> [64] <i>znawcę</i> ; z <i>nim jako</i> [65] <i>ze znawcą</i> ;

jeden		bez kodu: w nieodmiennych przydawkach liczebnikowych odpowiadających na pytania który? jaki? — przykłady: <i>zwyciężył w czasie dziewięć godzin dwadzieścia jeden minut; w latach sześćdziesiąt jeden sześćdziesiąt dwa;</i>
jeden	8	jako składnik odmiennego liczebnika głównego: <i>w dwudziestu jeden [8] krajach; delegacje trzydziestu jeden [8] państw;</i>
liczebniki główne		bez kodu: jako nieodmienne człony w zestawieniach liczebników porządkowych (<i>w sto dwudziestym pułku; w tysiąc dziewięćset piątym roku</i>); w nieodmiennych przydawkach liczebnikowych odpowiadających na pytanie który? jaki? (<i>zwyciężył w czasie dziesięć godzin dwadzieścia minut piętnaście sekund; Stal pokonała Wartę trzy dwa</i>);
niczym	9	w kontekstach typu: <i>niczym [9] słońce wszędzie rosa oczy wyje;</i>
raz	8	w znaczeniu <i>jednokrotnie</i> : <i>raz [8] na cztery lata; jeszcze raz [8] spróbujemy;</i>
razem	8	jeżeli jest możliwa substytucja wyrazu <i>wspólnie</i> : <i>odkrycia tego dokonali Rosjanie razem [8] z Amerykanami;</i>
się	41	formy nieosobowe przy czasownikach niezwrotnych: <i>mówi się [41] po plaży chodzi się [41] boso;</i>
temu	8	przy miarach czasu: <i>godzinę temu [8]; kilka lat temu [8];</i>

tym	9	w połączeniach spójnikowych <i>im...tym</i> , + stopień wyższy przysłówka lub przymiotni- ka: <i>tym</i> [9] <i>bardziej</i> , <i>tym</i> [9] <i>większy</i> ;
ułamki		liczebniki porządkowe jako mianowniki ułamków mają kwalifikatory rzeczownika: <i>jedna setna</i> [1];
z	8	w kontekstach typu: <i>było tego z</i> [8] <i>dziesięć</i> <i>sztuk</i> , <i>było tego ze</i> [8] <i>sto par</i> ;
za	8	w kontekstach typu: <i>co za</i> [8] <i>pech</i> ;

Wykaz symboli używanych w opisie morfologicznym słowoform

111	rzeczownik w mianowniku l.poj.
112	rzeczownik w mianowniku l.mn.
121	rzeczownik w dopełniaczu l.poj.
122	rzeczownik w dopełniaczu l.mn.
131	rzeczownik w celowniku l.poj.
132	rzeczownik w celowniku l.mn.
141	rzeczownik w bierniku l.poj.
142	rzeczownik w bierniku l.mn.
151	rzeczownik w narzędniku l.poj.
152	rzeczownik w narzędniku l.mn.
161	rzeczownik w miejscowniku l.poj.
162	rzeczownik w miejscowniku l.mn.
171	rzeczownik w wołaczu l.poj.
172	rzeczownik w wołaczu l.mn.
211	przymiotnik w mianowniku l.poj.
212	przymiotnik w mianowniku l.mn.
221	przymiotnik w dopełniaczu l.poj.
222	przymiotnik w dopełniaczu l.mn.

- 231 przymiotnik w celowniku l.poj.
- 232 przymiotnik w celowniku l.mn.
- 241 przymiotnik w bierniku l.poj.
- 242 przymiotnik w bierniku l.mn.
- 251 przymiotnik w narzędniku l.poj.
- 252 przymiotnik w narzędniku l.mn.
- 261 przymiotnik w miejscowniku l.poj.
- 262 przymiotnik w miejscowniku l.mn.

- | | |
|-----------------------------|--------------------------|
| 31 liczebnik w mianowniku | 41 zaimek w mianowniku |
| 32 liczebnik w dopełniaczu | 42 zaimek w dopełniaczu |
| 33 liczebnik w celowniku | 43 zaimek w celowniku |
| 34 liczebnik w bierniku | 44 zaimek w bierniku |
| 35 liczebnik w narzędniku | 45 zaimek w narzędniku |
| 36 liczebnik w miejscowniku | 46 zaimek w miejscowniku |

- 5 czasownik w formie nie wyróżnionej niżej, o ile dana forma jest homonimiczna (np. *napaść*[5]/*napaść*[1])
- 501 czasownik zwrotny w formach nie wyróżnionych niżej
- 51 bezokolicznik jako forma składowa form czasu przyszłego, czasownik niezwrotny (*będzie pisać*[51])
- 511 jak 51, ale czasownik zwrotny (*będzie się bawić*[511])
- 52 forma na -ł jako składowa form czasu przyszłego, czasownik niezwrotny (*będzie pisał*[52])

- 521 jak 52, ale czasownik zwrotny (*będzie się ba-
wił*[521])
- 53 forma na -ł jako czas przeszły z ruchomą koń-
cówką, czasownik niezwrotny (*kiedyście przy-
jechali*[53])
- 531 jak 53, ale czasownik zwrotny (*dla czegoś się
spóźnił*[531])
- 54 forma trybu warunkowego z ruchomą party-
kułą, czasownik niezwrotny (*wczoraj bym na-
pisał*[54])
- 541 jak 54, ale czasownik zwrotny (*ale bym się
przestraszył*[541])
- 55 opisowa forma trybu rozkazującego 3 osoby,
czasownik niezwrotny (*niech pisze*[55])
- 551 jak 51, ale czasownik zwrotny (*niech się ba-
wi*[551])
- 56 czasownik *być* jako składowa form cza-
sów złożonych (*będzie*[56] *pisał*[52], *napisał
był*[56])
- 57 czasownik *być*, *bywać*, *zostać* jako składowe
form strony biernej (*jest*[57] *napisany*, *by-
wał*[57] *nabywany*, *zostanie*[57] *napisany*)
- 61 przyimek z mianownikiem
- 62 przyimek z dopełniaczem (*z*[62] *domu*)
- 63 przyimek z celownikiem (podkrku[63] *domo-
wi*)
- 64 przyimek z biernikiem (*na*[64] *stół*)
- 65 przyimek z narzędnikiem (*z*[65] *tobą*)
- 66 przyimek z miejscownikiem (*na*[66] *stole*)

- 7 wykrzyknik
- 8 partykuła lub przysłówek
- 9 spójnik

Spis treści

Wstęp	1
1 Korpus SFPW	3
2 Stosowane metody i narzędzia pracy	5
3 Pliki korpusu	7
3.1 Gromadzenie plików	7
3.2 Porównywanie, selekcja	10
3.3 Korekta	17
4 Oznaczenia metatekstowe	19
5 Kilka uwag dotyczących opisu kodu korpusu	22
6 Dane bibliograficzne	26
6.1 Brakujące dane bibliograficzne	26
6.2 Połączenie informacji bibliograficznej i próbek podkorpusu . .	29
7 Podsumowanie	31
 Dodatek	 32
Opis kodu korpusu	32
1 Jednostka badania	32
Instrukcja I	35
1 Rozpisywanie tekstów	35
Instrukcja II	36
1 Imiona własne	36

Opis gramatyczny	37
1 Kod fleksyjny	37
2 Homonimia syntaktyczna	39
2.1 Rzeczownik czy czasownik (typ SV = 1/5)	40
2.2 Przymiotnik czy czasownik (typ AV = 2/5)	40
2.3 Przysłówek czy czasownik (typ AdvV=8/5)	40
2.4 Rzeczownik czy przymiotnik (typ SA = 1/2)	41
2.5 Rzeczownik czy przyimek (typ SPrp = 1/6)	41
2.6 Zaimki	43
2.7 Liczebniki	45
2.8 Przysłówek czy przyimek (typ Adv/Prp = 8/6)	46
2.9 Przyimek czy spójnik (typ Prp/Coni. = 6/9)	48
2.10 Spójnik czy partykuła (typ Coni./Part. =9/8)	49
3 Hasła kodowane umownie	49

Literatura

- [1] Cameron Debra, Rosenbatt Bill, Raymond Eric, Learning GNU Emacs. 1996
- [2] Kurcz Ida, Lewicki Andrzej, Sambor Jadwiga, Woronczak Jerzy, Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom III. Teksty popularno-naukowe. Warszawa 1974
- [3] Kurcz Ida, Lewicki Andrzej, Sambor Jadwiga, Szafran Krzysztof, Woronczak Jerzy, Słownik frekwencyjny polszczyzny współczesnej. Kraków 1990
- [4] Knuth E. Donald, The T_EX Book, 1996
- [5] Leszczyński Krzysztof, Perl cz. 1 w: Linux Plus 1997 nr 2, s.20-23
- [6] Leszczyński Krzysztof, Perl cz. 2 w: Linux Plus 1997 nr 3, s.10-13
- [7] Leszczyński Krzysztof, Perl cz. 3 w: Linux Plus 1997 nr 5, s.10-12
- [8] Mosiewicz Michał, Postgres — wcielenie 6.1 w: Linux Plus 1997 nr 7, s.6-11
- [9] Sambor Jadwiga, Słowa i liczby, Wrocław – Warszawa 1972
- [10] Schwartz Randal L., Learning Perl. (The Llama Book) 1993
- [11] Stallman Richard, GNU Emacs Manual. Eleventh Edition, Updated for Emacs Version 19.29. Boston 1995
- [12] red. Szymczak Mieczysław, Słownik ortograficzny języka polskiego. Warszawa 1975
- [13] Świdziński Marek, Własności składniowe wypowiedników polskich. Warszawa 1997

- [14] Wall Larry, Christiansen Tom, Schwartz Randal L., Programming Perl.
(The Camel Book) O'Reilly 1996
- [15] Wall Larry, Perl Kit, Version 4.0