ROCZNIKI HUMANISTYCZNE
Tom XLIX-L, zeszyt 6 – 2001-2002

ELŻBIETA AWRAMIUK

WPŁYW ODSTĘPSTW OD SEGMENTACJI ORTOGRAFICZNEJ NA WYNIKI STATYSTYCZNE *SŁOWNIKA FREKWENCYJNEGO POLSZCZYZNY WSPÓŁCZESNEJ*

Słowniki frekwencyjne można podzielić na kilka typów [por. Sambor 1972: 22]. *Słownik frekwencyjny polszczyzny współczesnej* (SFPW) jest słownikiem leksykalno-morfologicznym. Oznacza to, że uwzględnia się w nim zarówno częstości jednostek leksykalnych, jak i ich form fleksyjnych, pomija natomiast informacje składniowo-semantyczne. W słowniku przyjęto zasadę maksymalnej ścisłości i precyzji. Z tego powodu starano się w możliwie najmniejszej liczbie wypadków odstępować od danej z zewnątrz segmentacji na słowa[1]. Od tej zasady poczyniono wszakże kilka wyjątków.

Celem niniejszej pracy jest przedyskutowanie wpływu, jaki wywarły na wyniki badań odstępstwa od ortograficznego podziału na słowa. Moim zamiarem nie jest podważenie rozwiązań przyjętych w słowniku, ale przedstawienie alternatywnych rozwiązań segmentacyjnych i prześledzenie ich wpływu na zmiany frekwencji i rangi poszczególnych haseł.

SFPW jest ostatnim etapem wieloletnich badań (początki prac nad słownikiem datuje się na rok 1967). W latach 1974-1977 wydano w jedenastu wolu-

Dr Elżbieta AWRAMIUK – adiunkt Instytutu Filologii Polskiej, Zakład Współczesnego Języka Polskiego Uniwersytetu w Białymstoku; adres do korespondencji: Plac Uniwersytecki 1, 15-420 Białystok; e-mail: awramiuk@hum.mwb.edu.pl.

[1] Pojęcie *słowo*, a także *hasło* i *słowoforma*, stosuję za autorami SFPW. *Słowo* oznacza jednostkę unilateralną, *słowoforma* – jednostkę bilateralną (zinterpretowane pod względem gramatycznym i znaczeniowym słowo), a *hasło* – zbiór słowoform (przybliżenie leksemu). Przedmiotem operacji i obliczeń były słowoformy, a hasła pełniły jedynie funkcję porządkującą [SFPW: xix].

minach *Słownictwo współczesnego języka polskiego. Listy frekwencyjne* (SWJP), obejmujące słownictwo korpusu o długości 500 000 słów. Była to baza materiałowa omawianego SFPW. Decyzją autorów hasła o frekwencji niższej niż 4 nie zostały włączone do SFPW[2].

Decyzje segmentacyjne podejmowano przy opracowywaniu list frekwencyjnych. Autorzy przyjęli zasadę, aby jak najrzadziej odstępować od danej z zewnątrz segmentacji ortograficznej [SFPW: xxi]. Jako ciągi słów potraktowano zestawienia (połączenia składające się z dwóch lub kilku wyrazów tworzących całość znaczeniową), których przynajmniej jeden człon wchodzi w opozycje fleksyjne z innymi jednostkami (np. *biały kruk, tysiąc dziewięćset siedemdziesiąt osiem*), formy analityczne stopnia przysłówków i przymiotników (np. *bardziej męski, najbardziej męczący*), formy złożone czasowników (*będzie pisać, byłbym jechał, jest obserwowany, niech pisze, bili się*).

Od zasad ortograficznej segmentacji poczyniono kilka wyjątków [SFPW: xxiii-xxiv]. Jako formy wyrazowe potraktowano następujące ciągi słów: połączenia *się* z formą imiesłowu przymiotnikowego (np. *znajdującą się*) i regularnego rzeczownika odsłownego (np. *pojawienie się*), nazwiska zawierające pisaną osobno cząstkę typu *de, von* (np. *de Gaulle*)[3], połączenia cząstek *co* i *jak* z przysłówkami i przymiotnikami w stopniu najwyższym (np. *co najmniej, jak największy*), zleksykalizowane wyrażenia przyimkowe (np. *na dziś*) i inne ustabilizowane wyrażenia (np. *między innymi*). Trafność tych decyzji i ich konsekwencje zostaną przedyskutowane poniżej.

Cząstkę *się* uznawano w słowniku na ogół za osobne hasło. Wyjątkiem były sytuacje, w których *się* występowało w połączeniu z formą imiesłowu przymiotnikowego (SFPW zawiera 19 takich haseł o łącznej frekwencji F = 186) albo regularnego rzeczownika odsłownego (15 haseł o łącznej frekwencji F = 82). Występujące w tych słowoformach *się* nie było uwzględnione przy ustalaniu frekwencji hasła **się** *Prn.*

Rozwiązanie zastosowane w słowniku nakazuje traktować *się* w omawianych połączeniach jako odmienne od pozostałych użyć, w których wystąpienia słowa *się* zostały potraktowane jako słowoformy należące do hasła **się** *Prn* (rzadko – do hasła **siebie**). Jednak zamiast o homonimii dwóch różnych elementów można mówić tu o polifunkcyjności jednego. Logicznym

[2] Włączenie ich w zakres niniejszej pracy stało się niezbędne przy badaniu wyrażeń nieciągłych. Chcąc obliczyć frekwencję haseł powstałych w wyniku rozbicia tychże wyrażeń, musiałam podsumować wszystkie ich wystąpienia, także te znajdujące się w SWJP.

[3] W tej pracy nie omawiam problemów segmentacyjnych związanych z nazwami własnymi, gdyż nie weszły one do SFPW.

następstwem takiego sądu jest potraktowanie wszystkich typów konstrukcji z *się* jako związków wyrazów, a więc uznanie cząstki *się* za samodzielne słowo [por. Saloni 1976: 118] i bezwyjątkowe stosowanie segmentacji ortograficznej.

Proponowane rozwiązanie doprowadziłoby do potraktowania połączeń z *się* w analogiczny sposób jak czasownikowych form zwrotnych (zaliczono do nich wszystkie połączenia form czasownika z *się* nie pełniącym funkcji bezosobowej i nie blokującym pozycji mianownikowej). W tym wypadku *się* było traktowane jako samodzielne słowo, jego wystąpienia podliczono w haśle **się** *Prn*, ale przy słowoformach czasownikowych sygnalizowano ich obligatoryjną łączliwość z cząstką *się*.

Zastosowanie ortograficznej segmentacji do wszystkich typów połączeń z *się* wniosłoby do obecnej wersji słownika dwa rodzaje zmian. Po pierwsze, w słowniku pojawiłyby się nowe hasła, których artykuł hasłowy od obecnie istniejącego różniłby się tylko innym zapisem: brakiem cząstki *się* przy haśle i zaznaczaniem jej obligatoryjności przy słowoformach. Przykładowo, hasło **zbliżanie się** *S* zostałoby zastąpione hasłem **zbliżanie** *S*. Słowoformy pozostałyby identyczne, zmieniłby się jedynie sposób zaznaczenia obligatoryjnej łączliwości z *się* poprzez umieszczenie tej cząstki w nawiasie, np. *zbliżanie [się]*. Zmiany dotyczyłyby następujących haseł (w nawiasie podaję ich frekwencję): **domagający** (4)**, pojawienie** (14), **rozchodzenie** (5), **rozchodzący** (5) oraz **ciągnący** (8), **odnoszący** (5), **rozpoczynający** (5), **rozwijający** (17), **rysujący** (4), **wznoszący** (6), **znajdujący** (43)[4].

Po drugie, obok haseł, których słowoformy obligatoryjnie przyłączają *się*, w słowniku pojawiłyby się też takie hasła, których słowoformy różniłyby się między sobą pod tym względem. Oznacza to, że w jednym artykule hasłowym znajdowałyby się słowoformy z zaimkiem zwrotnym i bez niego. Frekwencja takich haseł byłaby sumą wystąpień istniejących w słowniku haseł z *się* i bez *się*. Przykładowo, z istniejących haseł **kształtowanie** (22) i **kształtowanie się** (6), powstałoby nowe – **kształtowanie** (28), które zawierałoby słowoformy z zaznaczoną obligatoryjnością występowania elementu *się* (*kształtowanie* [*się*]) oraz formy pozbawione tej informacji (*kształtowanie*). Zmiany dotyczyłyby następujących haseł (w nawiasie najpierw podaję frekwencję hasła bez *się*, a następnie – hasła z *się*): **dający** (6+5), **utrzymanie**

[4] Grupa ostatnich siedmiu haseł potencjalnie mogłaby zawierać słowoformy bez zaznaczonej obligatoryjności *się* (por. zdanie: *Nie zwracał uwagi na ćwiczenia rozwijające ruchliwość ogólną dziecka*), jednakże SFPW i SWJP odnotowują jedynie hasła z cząstką *się*.

(41+4), **wycofanie** (7+4), **zaangażowanie** (8+5), **zajmujący** (5+14), **zmieniający** (4+9). Do omawianej grupy należą też hasła, których frekwencja byłaby sumą hasła ze SFPW i hasła znajdującego się w SWJP, np. (w nawiasie na drugiej pozycji podano częstość wystąpień haseł z SWJP): **nadający** (5+2), **odbywający** (4+3), **odcięcie** (4+2), **palący** (4+1), **podporządkowanie** (4+2), **poruszający** (8+1), **posługiwanie** (8+1), **przeciwstawienie** (5+2), **składający** (18+1), **toczący** (9+1), **uczenie** (4+1), **wyrzeczenie** (6+3), **zapoznanie** (5+3), **zastanowienie** (4+3), **zbliżający** (13+2).

Istnieje możliwość, że w SFPW znajduje się imiesłów przymiotnikowy lub rzeczownik odprzymiotnikowy bez *się*, natomiast w SWJP jest jego odpowiednik z zaimkiem zwrotnym. Podobnie mogło się zdarzyć, że ze względu na niską frekwencję oba hasła pozostały jedynie w SWJP. Przy zastosowaniu nowej segmentacji frekwencja takich haseł powinna być uwzględniona.

Przy segmentacji traktującej element *się* jako obligatoryjny, choć oddzielny, frekwencja hasła *się* ulegnie zwiększeniu o 268 (łączna frekwencja rzeczowników odsłownych i imiesłowów przymiotnikowych z *się*), a więc z 9302 wzrośnie do 9570[5]. Na liście rangowej według wskaźnika F **się** *Prn* zajmuje miejsce 4 i przy takim wzroście frekwencji nie ulegnie ono zmianie.

Proponowane tu rozwiązanie spełnia podstawowe założenia autorów SFPW o rozróżnianiu jednostek homonimicznych na podstawie cech gramatycznych, nie zaś znaczeniowych [SFPW: xlix] i o przyjęciu segmentacji danej z zewnątrz. Jego zaletą jest fakt, iż bez odstępstw od nadrzędnej reguły segmentacyjnej podaje się te same informacje gramatyczne, które można wyczytać z obecnej wersji słownika, nie zaniżając przy tym frekwencji żadnej z jednostek.

Liczną grupę wyrażeń nieciągłych stanowią w omawianym słowniku połączenia cząstek *co* i *jak* z przysłówkami i przymiotnikami w stopniu najwyższym (w interpretacji SFPW – słowami homonimicznymi z przysłówkami i przymiotnikami w stopniu najwyższym). Przy założeniu, że wspomniane połączenia są ciągiem słów, cząstki *co* i *jak* należałoby traktować jako partykuły (w SJP PWN cząstki te w połączeniach typu *co najmniej, jak najwięcej* są scharakteryzowane jako partykuły wzmacniające), a człony drugie – jako

[5] Operuję tu hipotetycznymi danymi frekwencyjnymi. Pisząc, że jakieś hasło przy zastosowaniu nowej segmentacji miałoby przykładowo frekwencję 130 i rangę 16, mam na myśli, że na istniejącej liście rangowej znalazłoby się między 15 a 16 miejscem. Dane te są czysto orientacyjne, gdyż nie uwzględniam ewentualnych przesunięć na wyższych pozycjach listy rangowej. Ukazanie zmian listy frekwencyjnej ma charakter przykładowy, gdyż niemożliwe jest skorelowanie całości listy rangowej.

specyficzne (bo obligatoryjnie łączące się ze słowem *co* lub *jak*) słowoformy przymiotnikowe lub przysłówkowe, np. w haśle **mały** *Adj* znalazłyby się słowoformy *najmniejszy* oraz *[jak] najmniejszy*. W SFPW istnieją już hasła **jak** *Prt* (225) i **co** *Prt* (51). W wyniku zastosowania segmentacji rozłącznej ich frekwencja wzrosłaby odpowiednio do 304 (na liście rangowej oznacza to przesunięcie z miejsca 224 na 166) i do 140 (znaczny skok na liście rangowej: z miejsca 1245 na 427). Włączenie słowoform typu [*jak*] *najmniejszy* do odpowiednich haseł przymiotnikowych lub przysłówkowych również spowodowałoby zmianę wyników statystycznych. Przykładowo, przy potraktowaniu wyrażeń *co najmniej, co prawda* jako ciągów słów przesunięcia kształtowałyby się w sposób następujący: frekwencja hasła **mniej** *Adv* wzrosłaby z 125 do 168 (na liście rangowej przysłówek przesunąłby się z pozycji 447 na 345); frekwencja hasła **prawda** *S* wzrosłaby z 189 do 219 (na liście rangowej przesunięcie z miejsca 300 na 254).

Kolejny wyjątek od segmentacji ortograficznej stanowią w SFPW skostniałe połączenia przyimków (autorzy słownika nazywają je *słowem homonimicznym z przyimkiem*) z formą o nieokreślonym przypadku (najczęściej *słowem homonimicznym z przysłówkiem lub partykułą*) oraz z przymiotnikiem (*elementem „o postaci przymiotnikowej”*). Decyzja o potraktowaniu tych ciągów słów jako jednolitych słowoform była podyktowana trudnościami w zaklasyfikowaniu poszczególnych słów do klasy gramatycznej.

Przykładem takiej trudności jest *za* w połączeniu *za bardzo*, *za słaby*. Słowo wyglądające jak tradycyjny przyimek w tym wypadku nie ma w stosunku do słowa następującego po nim wymagań składniowych takich, jak przy połączeniu typu *za domem*. Połączenia takie jak *za bardzo* uznano w słowniku za jednolite słowoformy przymiotnikowe lub przysłówkowe. W tymże słowniku wątpliwości dotyczące przynależności danej jednostki do części mowy rozwiewano poprzez substytucję. Przykładowo, słowo *koło* kwalifikowano jako słowoformę przyimkową, jeśli można było zastąpić je w tekście – bez zmiany sensu zdania – przyimkiem *obok* (*koło wozu* = *obok wozu*). W połączeniach typu *za bardzo*, *za słaby* słowo *za* może być zastąpione przysłówkiem *zbyt* (por.: *Za ciężki ten bagaż dla ciebie* oraz *Zbyt ciężki ten bagaż dla ciebie*).

Uzasadnienie przysłówkowej kwalifikacji *za* w połączeniach typu *za słaby* odnajdujemy również w pracach językoznawczych [por. Grochowski 1986: 52; Kamińska-Szmaj 1990] i w innych słownikach. W SJPDor *za* z podanych użyć zostaje opisane jako wyraz dodawany do przymiotników lub przysłówków i nadający im odcień zbyt wielkiego natężenia cechy. W SJP PWN klasyfikacja jest już bardziej jednoznaczna: *w funkcji przysłówka*.

W SFPW znajdują się **za** *Prp* o frekwencji F = 1336 i **za** *Prt* o F = 68. Przy bezwyjątkowym stosowaniu segmentacji ortograficznej należałoby wyodrębnić trzecie hasło: **za** *Adv*.

W omawianym słowniku znalazło się 12 haseł typu **za długo, za mały** o łącznej frekwencji F = 115. Przy ustalaniu częstości wystąpień *za* przysłówkowego należało zbadać także odpowiednie hasła w SWJP. Wystąpiły tam 33 hasła o łącznej frekwencji F = 41. Niektóre z nich wymagają kilku słów komentarza. Celownik przymiotnika dawnej deklinacji rzeczownikowej *młodu*, pochodzący ze zleksykalizowanego połączenia *za młodu*, należałoby wyodrębnić w samodzielne hasło z jedną słowoformą *[za] młodu* (1). *Za* pochodzące z tego połączenia jest wyraźnie przyimkowe. Z kolei wyrażenia *za bardzo, za darmo*, choć niezbyt poprawnie brzmiałyby tu substytucje *za* na *zbyt*, można uznać za analogiczne do takich jak *za późno, za mało.*

Rezultatem zastosowania segmentacji ortograficznej w omawianych sytuacjach byłoby powstanie nowego hasła **za** *Adv* o F = 155 (a więc wcale nie małej). Na liście rangowej znalazłoby się ono na 372 pozycji. Nastąpiłby także wzrost frekwencji odpowiednich haseł przymiotnikowych i przysłówkowych.

W przypadku haseł znajdujących się w górnych warstwach częstości (np. **bardzo**) nawet wzrost frekwencji rzędu kilkudziesięciu jednostek nie powoduje zmiany rangi. Im dalsze miejsce na liście rangowej, tym znaczniejsze są przesunięcia rangi spowodowane niewysokim wzrostem frekwencji (np. **mało, wiele**). Oczywiście, w wyniku zastosowania nowej segmentacji do słownika weszłyby także wyrazy, które w połączeniu z *za* miały frekwencję niższą niż 4. Zostałyby włączone do istniejących już haseł, ale nie spowodowałyby znacznych przesunięć na liście rangowej. Na przykład frekwencja hasła **dobrze** *Adv* wzrosłaby z 392 do 395 (na liście rangowej oznacza to przesunięcie o jedno miejsce), a frekwencja hasła **wielki** *Adj* wynosząca 465 wzrosłaby o 1, nie zmieniając jego pozycji na liście rangowej.

Do wyjątków segmentacyjnych zaliczono również połączenia *po* z archaiczną formą przymiotnikowego celownika (np. *po polsku, po ludzku*) i z biernikową formą liczebnika porządkowego (np. *po pierwsze, po drugie*). Stanowisko takie nie jest odosobnione [por. Miodunka 1989: 69; GWJP: 19; Zarębina 1985: 18], jednakże można wskazać argumenty przemawiające za potraktowaniem takich połączeń jako ciągów słów.

Formy typu *polsku, ludzku* wywodzą się z przymiotników, ale w połączeniach z *po* występują w funkcji nietypowej, bo nie przy rzeczowniku. Jeśli brać pod uwagę budowę omawianych połączeń, są one niewątpliwie wyrażeniami przysłówkowymi złożonymi z przyimka i przymiotnika. Można więc

traktować je jako ciągi słów i frekwencję przyimka odnotować przy haśle **po** *Prp*, a drugie człony wpisywać w odpowiednie hasła przymiotnikowe wraz ze specjalnym oznaczeniem, np. w haśle **polski** obok słowoform istniejących znalazłaby się słowoforma *[po] polsku*. Wzbogacenie paradygmatu przymiotnika o formy typu *polsku* postuluje Saloni [1992], motywując to ich dużą produktywnością. Badania frekwencyjne po trosze potwierdzają tę produktywność (odnalazłam 5 haseł w SFPW i 25 w SWJP), ale wydaje się, że tkwi ona raczej w potencji języka niż w faktycznych użyciach.

Umieszczenie drugiego członu wyrażenia w istniejącym haśle nie zawsze jest możliwe. W SFPW nie istnieje hasło **omacek**, kłopotliwe więc staje się rozdzielne traktowanie zleksykalizowanego połączenia *po omacku*. Warto jednak pamiętać, iż hasła są wyróżnionymi w sposób arbitralny ciągami liter i pełnią rolę pomocniczą, porządkującą [por. Bogusławski 1987]. Ich zadaniem jest ułatwienie odnalezienia właściwych jednostek opisu, którymi w SFPW są słowoformy. Ponadto proponowane rozwiązanie nie jest nowe. W SJPDor występuje hasło **omacek** z kwalifikatorem *daw.* i podhasłem **po omacku**.

Konsekwencją zastosowania w stosunku do omawianych połączeń segmentacji ortograficznej byłby wzrost frekwencji hasła **po** *Prp* z 1728 do 1968 i przesunięcie na liście rangowej z pozycji 24 na 20.

Podobne problemy jak przyimek *po* nasuwa przyimek *z*. Często wchodzi on w skład wyrażeń, których członem są z kolei dopełniacze dawnej deklinacji rzeczownikowej przymiotników (np. *z bliska, z daleka*). Motywacja do umieszczenia form typu [*z*] *bliska* w odpowiednim istniejącym haśle jest mniejsza niż dla form typu [*po*] *polsku*, gdyż nie są to konstrukcje produktywne we współczesnej polszczyźnie, rozwiązanie takie byłoby jednak konsekwentne w stosunku do przedstawionych wyżej zasad opisu wyrażeń typu [*po*] *polsku*[6].

W omawianym słowniku znajduje się 7 haseł przysłówkowych, w skład których wchodzi przyimek *z*. Ich frekwencja wynosi 64. W SWJP haseł takich jest 13, o łącznej frekwencji F = 24. Pewną niekonsekwencję w segmentacji stanowi hasło **z tak daleka**. We wstępie do słownika zaznaczono, że dopuszczano wyłącznie interpretację jako jednolitych słowoform ciągów słów bezpośrednio po sobie następujących, a nie dopuszczano sytuacji, aby części jednej

[6] Istnieje jeszcze rozwiązanie pośrednie, a mianowicie liczenie wystąpień cząstki *z* w tych połączeniach jako wystąpień przyimka oraz wydzielenie hasła **bliska** *Adv* (lub **[z] bliska** *Adv*) zawierającego jedną słowoformę [*z*] *bliska*.

słowoformy były przedzielone innymi wyrażeniami [SFPW: xxv]. Wyrażenie *z tak daleka* jest przedzielonym słowem *tak* wyrażeniem *z daleka*. Podobnie można przecież tworzyć inne wyrażenia: *z bardzo daleka, z niezmiernie daleka* itp.

Zleksykalizowane wyrażenia *z czasem, z powrotem* są połączeniami zaimka z rzeczownikiem. Ich podział nie powinien nastręczać trudności. Natomiast przypisanie drugiego członu wyrażenia *z kretesem* określonemu hasłu jest problematyczne. Konsekwentnie do pozostałych tego typu wypadków (np. *po omacku*) należałoby utworzyć hasło **kretes**, które miałoby jedną słowoformę z informacją o obligatoryjnej łączliwości z przyimkiem *z*. Rozwiązanie takie nie jest nowością. W SJP PWN istnieje hasło **kretes**, a w artykule hasłowym podano informację: tylko w wyrażeniu przyimkowym. Podobnie należałoby postąpić z wyrażeniem *z zewnątrz*. Słowoformy nowego hasła miałyby informację o łączliwości z określonym przyimkiem ([*z*] *zewnątrz* o F = 5, [*na*] *zewnątrz* o F = 16).

Przy zastosowaniu segmentacji rozłącznej frekwencja hasła **z** *Prp* wzrosłaby z 8310 do 8395, a na liście rangowej nastąpiłoby przesunięcie z miejsca 7 na 6.

Spośród kilkuwyrazowych haseł najwięcej w słowniku jest wyrażeń, w skład których wchodzi przyimek *na*. W SFPW występują 24 nieciągłe hasła zawierające *na* o łącznej frekwencji F = 398, a w SWJP jest ich 44, o F = 66. W innych badaniach frekwencyjnych traktowano te wyrażenia jako ciągi słów. Halina Zgółkowa [1983] połączenie *na długo* potraktowała jako dwie słowoformy i umieściła je odpowiednio w hasłach **na** *Prp* z symbolem *ndm* (informującym o specyficznej funkcji tego przyimka) i **długo** *Adv*. Dzięki takiemu zapisowi frekwencja przyimka odpowiada rzeczywistej liczbie jego wystąpień w korpusie, a symbol gramatyczny sygnalizuje użycia nietypowe. Podobne rozwiązanie przyjęła Maria Zarębina [1985], uzasadniając je tym, iż każdy wyraz graficzny może wystąpić i jako forma, i jako hasło [Zarębina 1985: 18].

Wśród wyróżnionych w SFPW wyrażeń nieciągłych z *na* część wymaga kilku słów komentarza. Wyrażenie *na chybił trafił* należałoby rozdzielić na dwa: *na* + *chybił trafił* (drugi człon jest dwuwyrazowy według przyjętych w słowniku kryteriów odnoszących się do zestawień [por. SFPW: xxi]), podobnie *na łapu capu*. Oczywiście, przy słowoformie należałoby zaznaczyć obligatoryjną łączliwość z przyimkiem *na*. Połączenia **na jak długo** i **na tak długo** (podobnie jak omówione wyżej **z tak daleka**) stanowią przykłady dyskusyjne z punktu widzenia przedstawionych we wstępie do słownika kryteriów. Pozostałe człony wchodzące w skład wyrażeń przyimkowych z *na*

powinny znaleźć się w odpowiednich hasłach przysłówkowych (**długo, dzisiaj, krótko**), przymiotnikowych (*dobre* w **dobry**, *marne* w **marny**), rzeczownikowych (**pych, pół**) i liczebnikowych (**ile, raz**). Zaszłaby też konieczność utworzenia nowych haseł, często o jedynej słowoformie z zaznaczoną obligatoryjnością elementu *na* (np. [*na*] *przemian*, [*na*] *odwrót*, [*na*] *pewno*).

Konsekwencją zastosowania nowej segmentacji byłby wzrost frekwencji hasła **na** *Prp* z 8600 do 8998. Na liście rangowej przyimek ten pozostałby na miejscu 5.

Pozostałe wyrażenia przyimkowe występujące w słowniku jako jedno hasło to połączenia, w skład których wchodzą przyimki *od, do* oraz *w*. W wyniku zastosowania segmentacji rozłącznej ich frekwencja uległaby następującemu wzrostowi:

od *Prp* z 1780 do 1918 (przesunięcie na liście rangowej z miejsca 23 na 21);
do *Prp* z 5845 do 5890 (na liście rangowej pozostałoby na miejscu 9);
w *Prp* z 16318 do 16554 (przyimek ten jest na pierwszym miejscu listy rangowej i, oczywiście, na takim by pozostał).

Drugie człony omawianych wyrażeń wraz z dodatkowym oznaczeniem powinny zostać włączone odpowiednio do istniejących haseł przysłówkowych (np. w haśle **dziś** *Adv* znalazłyby się słowoformy *dziś* (224), [*do*] *dziś* (16), [*na*] *dziś* (7), [*od*] *dziś* (2), [*po*] *dziś* (3)) i przymiotnikowych (np. [*jak*] *największy* w haśle **większy** *Adj*) lub do haseł nowych (np. [*w*] *zamian* w haśle **zamian** *Adv*).

Ostatnią grupę wyjątków stanowią wyrażenia nieciągłe powszechnie uznawane za jednolite jednostki. Wśród nich wyodrębniłam dwie podgrupy.

Jedną z nich są wyrażenia, w skład których wchodzi cząstka *nie*. W SFPW znalazły się cztery: **nie lada** *Adv* (4), **nie sposób** *V* (8), **nie tyle** *Cnj* (14) i **nie ma** *V* (304). Pierwsze z trzech wymienionych wyrażeń mogą zostać rozdzielone według przedstawionych już zasad. *Nie* z tych połączeń podniosłoby frekwencję hasła **nie** *Prt* z 8341 do 8369 (policzono także częstość hasła **nie opodal** (2) z SWJP[7]). Powstałyby też dwa nowe hasła **lada** *Adv* i **sposób** *V*, a *tyle* zostałoby włączone do **tyle** *Cnj*.

Połączenie *nie ma* jest jedynym czasownikiem zaprzeczonym potraktowanym jako oddzielne hasło. O podjęciu takiej decyzji zadecydowała specyfika tej formy: nie jest ona zaprzeczeniem od *mieć* (zaprzeczenie od *mieć* brzmi, oczywiście, tak samo, ale zostało potraktowane jako ciąg słów), tylko zaprze-

[7] Na marginesie dodam, iż za rozwiązaniem z SWJP, by wyrażenie *nie opodal* traktować jako jedną jednostkę obliczeniową, przemawia ostatnia kodyfikacja ortograficzna.

czeniem od *być*. Wydaje się, że rozwiązanie traktujące połączenie *nie ma* jako jednolitą słowoformę jest konieczne, jednakże *nie* z omawianego połączenia w dalszym ciągu pozostaje zaprzeczeniem, a więc słusznie można by było doliczyć jego frekwencję do **nie** *Prt*. Wzrosłaby ona wtedy, biorąc pod uwagę wcześniejsze wyliczenia, do 8673, co spowodowałoby przesunięcie na liście rangowej o jedno miejsce. Zaprzeczone formy czasownika *mieć* proponuję umieścić w odrębnym haśle (jest to zgodne z decyzją autorów SFPW o wyodrębnianiu form supletywnych w oddzielne hasła). W jego artykule hasłowym znalazłyby się dwie słowoformy: [*nie*] *ma* i [*nie*] *masz*. Połączenie *nie masz* pojawiło się w SWJP. Skoro zostało potraktowane jako całość, należy przypuszczać, iż jest to zaprzeczenie od **być**, choć w specyficznej, nie występującej we współczesnej polszczyźnie formie.

Drugą podgrupę wśród omawianych wyrażeń nieciągłych stanowią spójniki złożone (*jak gdyby, jako że, mimo iż, mimo że, nie tyle, o tyle, podczas gdy, tyle że*) i wyrażenia, które są traktowane jako utarte połączenia (*bądź co bądź, coś niecoś, między innymi, o wiele, przede wszystkim, raz dwa, raz po raz, tak samo*). Przy zastosowaniu segmentacji rozłącznej ich człony zasiliłyby istniejące już hasła lub stałyby się podstawą do wyodrębnienia nowych. Przykładowo, złożony spójnik *jak gdyby* według SJP PWN to „połączenie *jak* w funkcji zaimka ze spójnikiem *gdyby*”[8]. **Jak** *Prp* nie wystąpiło w słowniku, natomiast **gdyby** *Cnj* miało frekwencję 223 (przy zastosowaniu nowej segmentacji wzrosłaby o 39). Z kolei potraktowanie jako ciągu słów wyrażenia *przede wszystkim* spowodowałoby wzrost frekwencji hasła **przed** *Prp* z 682 do 874, skok na liście rangowej z miejsca 67 na 50, a także wzrost frekwencji hasła **wszystko** *Prn* z 525 do 717. Wśród nowych haseł pojawiłoby się hasło **indziej** (zaliczyć je należy do klasy resztkowej i dla uproszczenia oznaczać *Adv*) o słowoformach [*kiedy*] *indziej* (4) i [*gdzie*] *indziej* (15).

Przeprowadzone w niniejszym artykule rozważania prowadzą do jednoznacznego wniosku, iż segmentacja tekstu wyraźnie oddziałuje na wyniki statystyczne. Odstępstwa od reguł ortograficznej segmentacji tekstu w SFPW wywarły wpływ na frekwencję przede wszystkim haseł przyimkowych, a także przymiotnikowych, przysłówkowych i partykułowych.

Błędem byłoby jednak stwierdzenie, iż sposób segmentacji przyjęty w SFPW fałszuje obraz polskiej leksyki. Zafałszowaniem można nazwać

[8] Jest to kwalifikacja co najmniej dyskusyjna, ale przytaczam ją tutaj jako przykład jednego z możliwych sposobów interpretacji omawianego wyrażenia nieciągłego.

np. liczenie frekwencji *ma* z połączenia *nie ma* jako słowoformy tylko hasła **mieć** lub też segmentację traktującą element *-m* z połączeń typu *kiedym przyjechał* jako nieodłączną część formy czasownikowej (*kiedy przyjechałem*). W pierwszym wypadku zawyżona zostałaby frekwencja, w drugim – zniekształcony obraz polszczyzny pisanej.

Konsekwentne stosowanie w słowniku segmentacji ortograficznej nie wniosłoby zbyt wielu zmian do obecnych wyników statystycznych, gdyż wyrażenia nieciągłe (jest ich w słowniku 130) stanowią nikły procent całego materiału (ok. 1,25%). Ponadto jedynie sześć spośród nich można zaliczyć do słownictwa bardzo częstego (czyli takiego, którego frekwencja jest wyższa niż 100). Są to następujące hasła: **między innymi** *Prt* (191), **nie ma** *V* (304), **na pewno** *Adv* (147), **po prostu** *Adv* (128), **przede wszystkim** *Prt* (192) i **w ogóle** *Adv* (156). Wystąpienia wyrażeń nieciągłych znajdujących się w słowniku stanowią zaledwie 0,5% korpusu. Rozłączne potraktowanie pozostałych wyrażeń nieciągłych (także tych znajdujących się w SWJP) nie spowodowałoby znaczących zmian w wynikach statystycznych, przede wszystkim z tego powodu, iż większość z nich to hapax-, dis- lub trislogomeny (czyli wyrażenia o frekwencji 1, 2 lub 3).

Tworzenie nowej listy rangowej na bazie istniejącego słownika wydaje się czynnością zbędną, jednakże dla zobrazowania efektów stosowania bezwyjątkowej segmentacji ortograficznej podaję pierwszą dziesiątkę haseł listy rangowej uwzględniającej wszystkie przesunięcia (w ostatniej kolumnie podano frekwencję otrzymaną w wyniku zastosowania nowej segmentacji).

1.	**w** *Prt*	16316	16554
2.	**i** *Cnj*	12385	12385
3.	**być** *V*	9621	9621
4.	**się** *Prn*	9302	9570
5.	**na** *Prp*	8600	8998
6.	**nie** *Prt*	8341	8673
7.	**z** *Prp*	8310	8397
8.	**on** *Prn*	6650	6650
9.	**do** *Prp*	5854	5890
10.	**ten** *Adj*	5743	5743

W wyniku bezwyjątkowego stosowania segmentacji ortograficznej w pierwszej dziesiątce na liście rangowej nie nastąpiłyby żadne przesunięcia, jedynie wzrosłaby frekwencja niektórych haseł. W drugiej dziesiątce jedyną zmianą byłoby pojawienie się na miejscu dwudziestym hasła **po** *Prp*. Z powyższego widać, iż wyjątków od reguł ortograficznej segmentacji tekstu jest w słowniku tak mało, że uwzględnienie bardziej konsekwentnych zasad niewiele by

zmieniło. Dane statystyczne podane w SFPW są więc z punktu widzenia polskiej leksyki wiarygodne.

Powyższe stwierdzenie nie podważa zasadności stosowania przy badaniach frekwencyjnych segmentacji danej z zewnątrz. Proponowany sposób opracowania materiału nie zubaża informacji słownikowej zawartej w istniejącej wersji słownika, gdyż z artykułu hasłowego dzięki odpowiednim symbolom gramatycznym i oznaczeniom można byłoby odczytać frekwencję tych połączeń, które obecnie stanowią oddzielne hasła. Ponadto stosowanie segmentacji danej z zewnątrz jest najłatwiejsze. Nie znaczy to jednak, że zawsze najlepsze i możliwe do zaakceptowania przez każdego badacza. W wypadkach pogranicznych (a takie były przedmiotem analizy w niniejszym tekście) rozstrzygnięcie problemu, czy dany element jest niesamodzielną cząstką wyrażenia nieciągłego, czy też samodzielnym wyrazem, jest zawsze konwencjonalne i zależy od przyjętych kryteriów i celów. Niejednorodność kryteriów stosowanych przez różnych badaczy przy wydzielaniu jednostki obliczeniowej zmniejsza (lub przekreśla) porównywalność wyników ich badań. Z przeprowadzonych rozważań płynie więc wniosek, iż istnieje potrzeba opracowania takiego algorytmu postępowania przy wydzielaniu jednostek obliczeniowych, który mógłby zostać zaakceptowany przez wszystkich badaczy zajmujących się statystyką językoznawczą.

BIBLIOGRAFIA

B o g u s ł a w s k i A. (1987), *Obiekty leksykograficzne i jednostki języka*, w: *Studia z polskiej leksykografii współczesnej*, red. Z. Saloni, t. II, Białystok, s. 13-34.

Gramatyka współczesnego języka polskiego. Morfologia, red. R. Grzegorczykowa, R. Laskowski, H. Wróbel, Warszawa 1984 – GWJP.

G r o c h o w s k i M. (1986), *Polskie partykuły. Składnia, semantyka, leksykografia*, Wrocław.

K a m i ń s k a - S z m a j I. (1990), *Różnice leksykalne między stylami funkcjonalnymi polszczyzny pisanej. Analiza statystyczna na materiale słownika frekwencyjnego*, Wrocław.

M i o d u n k a W. (1989), *Podstawy leksykologii i leksykografii*, Warszawa.

S a l o n i Z. (1976), *Cechy składniowe polskiego czasownika*, Wrocław.

S a l o n i Z. (1992), *Rygorystyczny opis polskiej deklinacji przymiotnikowej*, „Uniwersytet Gdański. Prace Językoznawcze" 16, s. 215-228.

S a m b o r J. (1972), *Słowa i liczby. Zagadnienia językoznawstwa statystycznego*, Wrocław.

Słownik frekwencyjny polszczyzny współczesnej, t. I-II, Kraków 1990 – SFPW.
Słownik języka polskiego, red. W. Doroszewski, t. I-XI, Warszawa 1958-1969 – SJPDor.
Słownik języka polskiego, red. M. Szymczak, t. I-III, Warszawa 1978-1981 – SJP PWN.
Słownictwo współczesnego języka polskiego. Listy frekwencyjne, oprac. I. Kurcz, A. Lewicki, J. Sambor, J. Woronczak, t. I-V, Warszawa 1974-1977 – SWJP.
Z a r ę b i n a M. (1985), *Próba statystycznej analizy słownictwa polszczyzny mówionej (synteza danych liczbowych)*, Wrocław.
Z g ó ł k o w a H. (1983), *Słownictwo współczesnej polszczyzny mówionej. Lista frekwencyjna i rangowa*, Poznań.

THE INFLUENCE OF DEPARTURES FROM ORTHOGRAPHIC SEGMENTATION ON THE STATISTIC RESULTS OF THE FREQUENCY DICTIONARY OF CONTEMPORARY POLISH

S u m m a r y

The paper deals with the influence of qualitative interpretation (here: segmentation of test) on quantitative interpretation (here: statistic results). Those segmentation decisions have been discussed which we can find in the *Frequency Dictionary of the Contemporary Polish Language*, which are departures from orthographic segmentation. The simulation of the changes of results with a unexceptional application of segmentation given from without proves that the segmentation of text affects statistic results. The introduction of exceptions affected the frequency of some units (mainly prepositions, adjectives, and adverbs), but the material gathered in SFPW may be regarded as reliable, for the non-linear units constitute only 0.5 per cent in it. The paper ends with a postulate to work out an algorithm of procedure when calculatory units are separated, an algorithm that could be accepted by all researchers who deal with linguistic statistics. Such a study could make the results of statistic examination more comparable.

Translated by Jan Kłos

Słowa kluczowe: segmentacja, statystyka językoznawcza, jednostki nieciągłe, frekwencja.

Key words: segmentation, linguistic statistics, non-linear units, frequency.