http://dx.doi.org/10.17651/POLON.38.6

POLONICA XXXVIII
PL ISSN 0137-9712

DANIEL ŚLEDZIŃSKI
Uniwersytet im. Adama Mickiewicza w Poznaniu

Wielowarstwowy model podziału wyrazów ortograficznych języka polskiego na sylaby

1. Wprowadzenie

W artykule omówiono model przeznaczony do dzielenia wyrazów ortograficznych na sylaby. Przeprowadzone liczne badania ukazały wiele istotnych faktów dotyczących sylabifikacji w języku polskim (Śledziński 2013). Dzięki nim można było wskazać na czynniki, które są istotne dla tego procesu. Omawiany w niniejszej publikacji model jest skonstruowany w ten sposób, żeby można było w dowolnym stopniu uwzględniać poszczególne istotne czynniki. Zatem nie jest to rozwiązanie oparte na stałych regułach — celem autora było stworzenie mechanizmu elastycznego, który będzie można dostosować do różnych zadań. Na wstępie należy wyjaśnić kilka podstawowych pojęć związanych z niniejszą publikacją, w szczególności sposób rozumienia tytułowego modelu podziału na sylaby oraz związku tego pojęcia z innymi istotnymi zagadnieniami.

Wielowarstwowy model podziału wyrazów na sylaby precyzuje sposób przetwarzania wyrazów ortograficznych. Celem tego przetwarzania jest wstawianie znaczników granic sylab w tekstach zapisanych w formie elektronicznej. Zatem model obejmuje teoretyczne założenia dotyczące: kolejnych operacji wykonywanych na przetwarzanych wyrazach, odniesienia do definicji lingwistycznych, dodatkowych mechanizmów towarzyszących przetwarzaniu wyrazów, precyzuje on również składnię i funkcje reguł. Trzeba zaznaczyć, że sam model nie określa ostatecznego umiejscowienia granic sylab w konkretnych wyrazach ortograficznych, ponieważ wynika ono z użytych reguł. Natomiast model nie zawiera konkretnego zbioru reguł. Zatem można utworzyć dowolną liczbę różnych zbiorów reguł, które będą zgodne z prezentowanym modelem.

Program komputerowy również nie jest częścią omawianego modelu, jednak jest to techniczne narzędzie, które umożliwia praktyczną realizację założeń tego modelu. Program wykonuje podział na sylaby w konkretnych tekstach. Przekształcenia wyra-

zów wykonywane przez program powinny być zgodne z opisem zamieszczonym w tej publikacji (zob. par. 3.1).

Pod pojęciem systemu dzielącego wyrazy na sylaby rozumiane są te wszystkie elementy, które umożliwiają praktyczną czynność automatycznego podziału — zatem pojęcie to obejmuje zarówno określony zbiór reguł o konstrukcji zdefiniowanej w modelu (zbiór ten jest zapisywany w oddzielnym pliku tekstowym), jak i program komputerowy, który przetwarza wyrazy zgodnie z założeniami modelu. Można mówić o systemie dzielącym wyrazy na sylaby, który jest oparty na prezentowanym modelu wielowarstwowym. W niniejszej publikacji skupiono się na omawianiu teoretycznego modelu podziału wyrazów na sylaby. Zatem nie jest omawiany konkretny system, jednak przedstawione informacje są wystarczające dla stworzenia takiego systemu.

Trzeba też zaznaczyć, że podział na sylaby wykonywany przez dowolny system oparty na modelu wielowarstwowym ma charakter umowny. Takie systemy mają zastosowanie praktyczne — mogą być używane w rozwiązaniach technicznych, takich jak systemy automatycznego rozpoznawania mowy czy systemy syntezy mowy. Inne zastosowania praktyczne mają związek z dydaktyką — dotyczy to między innymi materiałów edukacyjnych dla dzieci. Podział wyrazów na sylaby jest również wykorzystywany w narzędziach terapeutycznych dla dzieci dotkniętych dysleksją.

W paragrafie drugim niniejszej publikacji omówiono dwa podstawowe czynniki lingwistyczne, które zostały uwzględnione w prezentowanym modelu — jest to fonologia oraz morfologia. W paragrafie trzecim zostały przedstawione podstawy modelu wielowarstwowego — najwięcej uwagi poświęcono omówieniu poszczególnych etapów zaprezentowanego algorytmu. W paragrafie czwartym omówiono kolejne kroki związane z przetwarzaniem trzech przykładowych wyrazów ortograficznych. Natomiast w paragrafie piątym przedstawiono wyniki testów, które zostały przeprowadzone przy użyciu prototypowego systemu podziału na sylaby, opartego na modelu wielowarstwowym. Artykuł zamknięto podsumowaniem zawierającym między innymi propozycje nazw warstw w przyszłych publikacjach dotyczących konkretnych systemów opartych na przedstawionym modelu.

2. Czynniki uwzględnione w modelu dzielenia wyrazów na sylaby

2.1. Fonologia

We wcześniejszych publikacjach podkreślano, że lingwistyczne definicje sylaby nie umożliwiają wyznaczania granic sylab w sposób jednoznaczny (Polański 1999, Trask 1996, Wierzchowska 1971). Z fonetycznej definicji sylaby wynika, że ta jednostka językowa ma ścisły związek z procesem artykulacyjnym, a na granice sylab przypadają momenty maksymalnego zwarcia narządów artykulacyjnych. W rzeczywistości określenie momentu maksymalnego zwarcia jest trudne lub niemożliwe. W szczególności dotyczy to wieloelementowych grup spółgłoskowych, które w języku polskim mają wyjątkowo złożoną strukturę (Dobrogowska 1984, 1990; Dunaj 1985; Jassem, Łobacz

1974). Fonologiczna zasada sonorności w sposób abstrakcyjny odzwierciedla proces artykulacyjny związany z wymawianiem sylab. Odbywa się to poprzez odniesienie do fonologicznej skali sonorności, która przypisuje do poszczególnych dźwięków mowy abstrakcyjne wartości odzwierciedlające przede wszystkim stopień rozwarcia narządów artykulacyjnych (Szpyra-Kozłowska 1998, 2002)[1]. W omawianym w tej publikacji modelu istnieje możliwość odniesienia do skal sonorności różnych autorów. Ponadto model ten zawiera definicję domyślnej skali sonorności (zob. par. 3.6)

Pojęcie zasady sonorności (ang. the Sonority Sequencing Principle lub the Sonority Sequencing Generalization) zostało użyte po raz pierwszy w pracach Selkirka (1980), jednak idea opisywania struktury sylaby przy użyciu wartości sonorności była obecna już w pracach Jespersena (1904). W klasycznym rozumieniu tej zasady wartość sonorności dźwięków mowy maleje w miarę oddalania się od ośrodka sylaby. Zatem w ramach nagłosu sylaby dopuszczalny jest tylko wzrost wartości sonorności, natomiast w ramach wygłosu sylaby dopuszczalny jest tylko spadek wartości sonorności. Natomiast zgodnie z fonologiczną zasadą maksymalnego nagłosu (ang. the Maximal Onset Principle) do nagłosu sylaby powinna być przypisana jak największa liczba dźwięków, o ile nie jest to sprzeczne z zasadą sonorności. Badanie przeprowadzone przez autora (omówione w publikacji przytoczonej we wstępie) wykazało, że na podstawie tych dwóch zasad fonologicznych można wyznaczyć granice sylab w ponad 99% wyrazów analizowanego korpusu tekstów. Ponadto zasada sonorności w sposób abstrakcyjny odzwierciedla treść fonetycznych definicji sylaby (spadek sonorności w miarę oddalania się od ośrodka sylaby ma związek ze spadkiem energii artykulacyjnej oraz zwiększaniem stopnia zwarcia narządów artykulacyjnych). Dlatego metoda wyznaczania granic sylab oparta na omówionych zasadach fonologicznych stanowi podstawę omawianego modelu wielowarstwowego. Każdy system oparty na tym modelu domyślnie dzieli wyrazy właśnie w ten sposób. Jednak podstawowa koncepcja tego modelu umożliwia zastąpienie tego podziału dowolnym podziałem alternatywnym (zob. par. 3.1).

Trzeba zaznaczyć, że istnieją też inne koncepcje fonologiczne, które mogą posłużyć za podstawę wyznaczania granic sylab w wyrazach. W publikacji: „Edge of Constituent Effects in Polish" (Rubach, Geert 1990) autorzy postulują, że dla języka polskiego dopuszczalne jest przypisanie tego samego poziomu sonorności dla wszystkich spółgłosek właściwych. Poza tym autorzy dopuszczają dowolną kolejność spółgłosek właściwych w ramach nagłosu oraz wygłosu sylaby.

Podział domyślny (wynikający z zasad fonologicznych) może być zastąpiony przez podział wynikający z badań poczucia subiektywnego związanego z dzieleniem wyrazów na sylaby. Badania przeprowadzone przez autora (Śledziński 2017) wykazały, że podział wynikający z poczucia subiektywnego często różni się od podziału wskazywanego przez fonologię.

[1] W przytoczonej publikacji autorka omawia również warianty skali sonorności zaproponowane przez innych badaczy. Swoje wersje przedstawili: Jespersen (1904), Hooper (1976), Selkirk (1984), Goldsmith (1990), Clements (1990, 1992).

2.2. Morfologia

W języku polskim pojęcie morfemu nie pokrywa się z pojęciem sylaby. Pomimo to niektóre granice związane z procesami słowotwórczymi są odczuwalne jako naturalne granice sylab. Można do nich zaliczyć:

— granicę między przedrostkiem i rdzeniem w derywatach,
— granicę między członami zrostów,
— granicę w złożeniach (za interfiksem).

Przewaga czynnika morfologicznego nad czynnikiem fonetycznym została podkreślona przy omawianiu zasad związanych z przenoszeniem wyrazów między liniami tekstu (Polański 2006). Również badania pilotażowe przeprowadzone przez autora potwierdziły, że obecność tego typu granic ma decydujący wpływ na poczucie subiektywne związane z dzieleniem wyrazów na sylaby[2]. Można jednak dyskutować o tym, czy osoby biorące udział w eksperymencie rzeczywiście wyznaczają granice sylab czy raczej granice morfemów, które są interpretowane jako granice sylab. Nie ulega wątpliwości, że ten rodzaj granic może mieć istotne znaczenie przy projektowaniu systemu podziału na sylaby. Warto zauważyć, że w języku polskim wiele wieloelementowych grup spółgłoskowych zostało uformowanych w procesie słowotwórczym. Dlatego w prezentowanym modelu przewidziano możliwość tworzenia reguł odnoszących się do struktury morfologicznej wyrazów.

3. Wielowarstwowy model sylabifikacji

3.1. Podstawy modelu wielowarstwowego

W poprzednich paragrafach omówiono czynniki, do których można się odnieść przy wyznaczaniu granic sylab w wyrazach ortograficznych języka polskiego. Zasady fonologiczne umożliwiają wyznaczanie granic sylab w prosty i jednoznaczny sposób, który nawiązuje do lingwistycznych definicji sylaby. Jednak podział oparty na zasadach fonologicznych często jest sprzeczny z odczuciem subiektywnym — w szczegól-

[2] Wyniki tej części badań nie zostały jeszcze opublikowane, zatem niezbędny jest komentarz. Badanie polegało na zapisywaniu określonych wyrazów w kolumnie arkusza (na komputerze). Osoby biorące udział w badaniu musiały wyznaczać granice sylab w tych wyrazach zgodnie z własnym subiektywnym poczuciem (spacja oznaczała granicę). W badaniu wzięło udział 50 osób (głównie studenci pierwszego roku kierunków filologicznych). Został użyty test zgodności chi-kwadrat. Niemal wszystkie uzyskane wyniki potwierdziły kluczowy wpływ struktury morfologicznej wyrazów na decyzje dotyczące umiejscowienia granic sylab (w przytoczonych tutaj przykładach wyników podano tylko odsetek odpowiedzi pokrywających się z granicą wynikającą ze struktury morfologicznej wyrazu). Struktura morfologiczna miała decydujące znaczenie, jeżeli granice wynikające z tej struktury były inne niż granice wynikające z zasad fonologicznych, np.: *ćwierć|wałek* (94%), *pod|władny* (86%), *nad|chloran* (96%), *ob|słuchać* (88%). Jeżeli z fonologii nie wynikało żadne umiejscowienie granicy, to uzyskane wyniki jednoznacznie wskazują na kluczową rolę struktury morfologicznej, np.: *roz|łzawić* (74%), *bez|rtęciowy* (98%), *roz|mnażać* (90%). Jedyne odstępstwo od omawianej tendencji dotyczy niektórych wyrazów zawierających prefiks zakończony samogłoską (te badania będą kontynuowane).

ności dotyczy to niektórych konstrukcji morfologicznych. Celem autora było stworzenie modelu sylabifikacji, który umożliwiłby uwzględnienie wymienionych czynników w dowolnym zakresie. W tym paragrafie omówiono podstawowe elementy tego modelu. Poza tym moduł projekcji wewnątrzwyrazowych ortograficznych grup spółgłoskowych na transkrypcję fonologiczną (zob. par. 3.6) został szczegółowo omówiony w oddzielnej publikacji (Śledziński 2016).

W prezentowanym modelu zastosowano podejście algorytmiczne oparte na regułach. Są to reguły dwuczłonowe, przy czym pierwszy człon każdej reguły zawiera ciąg znaków, który ma być zastąpiony przez ciąg znaków zapisany w drugim członie tej samej reguły. Trzeba jednak zaznaczyć, że dla każdej reguły można definiować zbiory wyjątków. Każdy wyjątek definiuje się poprzez określenie nagłosowych lub wygłosowych ciągów liter. Obecność jednego z takich ciągów w przetwarzanym wyrazie sprawi, że dana reguła nie będzie dla niego zastosowana.

Działanie prezentowanego modelu obejmuje następujące etapy:

1. Wyznaczenie wszystkich możliwych początkowych oraz końcowych ciągów znaków należących do przetwarzanego wyrazu.
2. Stosowanie reguł odnoszących się do morfologii dla fragmentów wyznaczonych w etapie pierwszym.
3. Wyznaczenie ośrodków sylab w przetwarzanym wyrazie ortograficznym.
4. Wstawienie znaczników granicy między sąsiadującymi ośrodkami sylab.
5. Wyznaczenie wewnątrzwyrazowych ortograficznych grup spółgłoskowych oraz pojedynczych spółgłosek (wraz z literą oznaczającą samogłoskę umiejscowioną w kontekście prawostronnym).
6. Stosowanie reguł modyfikacyjnych w odniesieniu do fragmentów wyznaczonych w etapie piątym — jeżeli poszczególne grupy nie zawierają znacznika granicy sylaby wstawionego w kroku drugim.
7. Wyznaczenie granicy na podstawie zasad fonologicznych dla wewnątrzwyrazowych ortograficznych grup spółgłoskowych — tylko wtedy, gdy takie fragmenty nie zawierają znacznika granicy sylab wstawionego wcześniej.
8. Jeżeli dla danej wewnątrzwyrazowej grupy spółgłoskowej (lub pojedynczej spółgłoski umiejscowionej pomiędzy ośrodkami sylab) granica nie została wyznaczona w poprzednich krokach, to znacznik granicy jest wstawiany przed tą grupą lub spółgłoską.

W tym opisie pominięto fakt, że przetwarzana jest kopia oryginalnego wyrazu, która zostaje zapisana przy użyciu małych liter. Do tej kopii zostaje dodany symbol kratki (#) — przed i po przetwarzanym wyrazie. Po zakończeniu przetwarzania uzyskane granice są przenoszone do wyrazu oryginalnego.

Analizując przebieg tej procedury łatwo zauważyć, że obejmuje ona cztery poziomy (warstwy) wstawiania granic sylab (w etapach: drugim, czwartym, szóstym i siódmym), przy czym reguły stosowane wcześniej uniemożliwiają wstawianie granic

w kolejnych krokach (dla poszczególnych wewnątrzwyrazowych grup spółgłoskowych przetwarzanego wyrazu). Wstawianie znaczników granicy w kroku drugim jest oparte na ciągach znaków należących do nagłosu oraz do wygłosu wyrazu. Natomiast w przypadku pozostałych warstw dla wstawiania znaczników granicy niezbędne jest wyznaczenie liter oznaczających ośrodki sylab oraz ortograficznych grup spółgłoskowych lub pojedynczych spółgłosek umiejscowionych pomiędzy tymi ośrodkami. W kolejnych częściach omówiono sposób konstruowania poszczególnych rodzajów reguł, a także wyjaśniono rozwiązania odnoszące się do wymienionych etapów działania algorytmu.

3.2. Konstrukcja reguł odnoszących się do morfologii (etapy 1 i 2)

Wyznaczenie wszystkich początkowych oraz końcowych ciągów znaków należących do danego wyrazu (etap pierwszy procedury omówionej w części 3.1) umożliwia stosowanie reguł odnoszących się do struktury morfologicznej tego wyrazu (zob. par. 2.2). W wyrazie *podjazd* wszystkie początkowe ciągi znaków to: *#podjazd*, *#podjaz*, *#podja*, *#podj*, *#pod*, *#po*, *#p*. Przykładowa reguła mogłaby wyglądać następująco: *#podja>#pod|ja*. Zatem każdy wymieniony ciąg znaków (w kolejności od najdłuższego do najkrótszego) jest porównywany ze zbiorem reguł. Jeżeli dany ciąg jest taki sam jak pierwszy człon określonej reguły, to jest on podmieniany w wyrazie przez ciąg znaków umiejscowiony w drugim członie tej samej reguły i sprawdzanie początkowych ciągów znaków zostaje przerwane. Podobny mechanizm jest stosowany dla ciągów znaków należących do wygłosu wyrazu (wyznaczone w ten sposób granice nie muszą pokrywać się z granicami morfologicznymi).

Warto zauważyć, że podana reguła: *#podja>#pod|ja* jest właściwa również dla innych wyrazów (na przykład dla wyrazu *podjadać*). Wynika to z zależności, którą trzeba wziąć pod uwagę przy konstruowaniu reguł omawianego typu: im krótszy jest ciąg liter użyty w pierwszym członie danej reguły, tym większy jest zbiór wyrazów, których ta reguła dotyczy.

Wszystkie reguły są zapisywane zgodnie z określoną składnią. Zapis omawianego typu reguł rozpoczyna się od słowa !MOR. Następnie po znaku dwukropka zapisywane są dwa człony reguły (rozdzielone znakiem >). Ewentualna lista wyjątków jest zapisywana po słowie !EXC i po znaku dwukropka. Poszczególne łańcuchy znaków na liście wyjątków są rozdzielone przecinkiem (bez spacji), natomiast na końcu reguły stoi średnik. Oto przykład reguły zapisanej zgodnie z tymi wytycznymi:

!MOR:#ćwierć>#ćwierć|!EXC:#ćwierć#;

Podana reguła jest właściwa dla wyrazów typu: *ćwierćcalowy*, *ćwierćdolarowy*, *ćwierćfinalista*, *ćwierćnutowy*, *ćwierćwiecze*. Natomiast umieszczenie ciągu znaków *#ćwierć#* na jednoelementowej liście wyjątków sprawi, że ta reguła nie będzie stosowana dla wyrazu *ćwierć.* Często w celu utworzenia właściwej listy wyjątków trzeba wykonać analizę słownika. Kolejny przykład dotyczy reguły, której lista wyjątków zawiera ciągi znaków pozwalające na identyfikację różnych form fleksyjnych:

!MOR:#przed>#przed|!EXC:#przedłuż,#przedsięb,#przedział,#przedziel,#przedźw,#przedzw,#przedźwig,#przed#;

Warto podkreślić, że umieszczenie na liście wyjątków ciągów znaków, które identyfikują określone wyrazy, wcale nie wyklucza podziału tych wyrazów zgodnie z daną regułą — może to nastąpić w jednym z następnych etapów procedury przedstawionej w części 3.1.

Trzeba zaznaczyć, że użycie omawianego w tej części typu reguł niekoniecznie musi mieć związek ze strukturą morfologiczną wyrazów (jest to ich podstawowa funkcja). Pierwszy człon każdej reguły omawianego typu może zawierać dowolny ciąg znaków — nawet konkretną formę fleksyjną. Zatem daje to nieograniczone możliwości blokowania podziału wykonywanego w późniejszych etapach (można w ten sposób odnieść się do każdego wyrazu i każdej grupy spółgłoskowej, która zostałaby podzielona w etapie siódmym).

3.3. Wyznaczanie ośrodków sylab w wyrazach ortograficznych (etap 3)

Przetwarzanie wyrazu po drugim etapie omawianej procedury dotyczy wewnątrzwyrazowych grup spółgłoskowych — grup, które są otoczone przez litery oznaczające samogłoskę. Algorytm wyznaczający litery oznaczające samogłoski został opracowany na podstawie publikacji Marii Steffen-Batogowej „Automatyzacja transkrypcji fonematycznej tekstów polskich" (1975). Można przyjąć, że litery: *a*, *o*, *ó*, *e* zawsze oznaczają samogłoskę. Również takie założenie można przyjąć w odniesieniu do liter *ą* oraz *ę*, pomimo że w rzeczywistości oznaczają one sekwens fonemów (jest on nierozdzielny na płaszczyźnie ortograficznej). Litera *u* po literach *a* lub *e* może oznaczać samogłoskę lub fonem /w/. Litera *i* może oznaczać samogłoskę /i/, półsamogłoskę /j/ lub nie oznaczać żadnego fonemu (tylko znak miękkości). Litera *y* najczęściej oznacza samogłoskę /y/, jednak w nielicznych przypadkach może też oznaczać fonem /j/ lub /i/.

Wyznaczanie liter oznaczających ośrodki sylab opiera się na zamkniętym zbiorze reguł (istnieje możliwość modyfikowania tego zbioru). Każda reguła z tego zbioru składa się z dwóch członów, przy czym pierwszy człon oznacza określony ciąg liter (np. *bia*), natomiast drugi człon zawiera ciąg o takiej samej długości, który złożony jest ze znaków: C oraz V (np. CCV). Każdej literze w pierwszym członie danej reguły odpowiada dokładnie jeden znak (C lub V) w drugim członie tej reguły. Oznaczenie C lub V informuje o tym, że odpowiednia litera jest spółgłoską (łac. *consonans*) lub samogłoską (łac. *vocalis*). Algorytm wyznaczający samogłoski w wyrazach ortograficznych najpierw wyznacza wszystkie możliwe ciągi znaków, należące do danego wyrazu, a następnie ciągi te są porównywane (w kolejności od najdłuższego do najkrótszego) ze zbiorem reguł. Jeżeli jakiś ciąg w wyrazie jest taki sam jak pierwszy człon określonej reguły, to jest on zamieniany przez ciąg znaków należący do drugiego członu tej samej reguły. Ostatecznie otrzymywany jest ciąg złożony ze znaków C i V, natomiast jego długość jest taka sama jak długość pierwotnego wyrazu ortograficznego. Na podsta-

wie tych danych wyznaczane są wewnątrzwyrazowe grupy spółgłoskowe wraz z prawostronnym kontekstem samogłoskowym (etap piąty omawianej procedury).

Dla przykładu zostanie omówiona procedura dotycząca przekształcania wyrazu *biały*. Najpierw wyznaczane są wszystkie możliwe ciągi liter należące do tego wyrazu: *#biały#*, *#biały*, *biały#*, *#biał*, *biały*, *iały#*, *#bia*, *biał*, *iały*, *ały#*, *#bi*, *bia*, *iał*, *ały*, *ły#*, *#b*, *bi*, *ia*, *ał*, *ły*, *y#*, *#*, *b*, *i*, *a*, *ł*, *y*. Następnie uzyskane ciągi są porównywane z dostępnymi regułami, przy czym zasadnicze znaczenie ma fakt, że odbywa się to w kolejności od ciągów najdłuższych do ciągów najkrótszych. Zbiór reguł zawiera między innymi następujące reguły:

bia>CCV
b>C
i>V
a>V
ł>C
y>V

Ze względu na rozmiar najpierw stosowana jest pierwsza wymieniona reguła, po czym wyraz przyjmuje następujący kształt: #CCV*ły*#. Dlatego trzy kolejne reguły z powyższej listy nie mogą być już użyte. Po zastosowaniu dwóch ostatnich reguł widniejących na powyższej liście wyraz przyjmuje ostateczny kształt: #CCVCV#[3].

Trzeba zaznaczyć, że omówione przekształcenie odbywa się na kopii oryginalnego wyrazu. Po przekształceniu tej kopii otrzymywany jest ciąg znaków C i V o długości takiej samej jak wyraz oryginalny. To pozwala na wyznaczenie w wyrazie oryginalnym grup spółgłoskowych i pojedynczych spółgłosek umiejscowionych między ośrodkami sylab[4] (na podstawie obecności ciągów znaku C otoczonych przez znaki V).

3.4. Wstawianie granic między sąsiadującymi ośrodkami sylab (etap 4)

Etap czwarty procedury omówionej w części 3.1 obejmuje wstawianie znaczników granicy między sąsiadującymi ośrodkami sylab. Umiejscowienie tego typu granic jest

[3] Półsamogłoski mają cechy akustyczne i artykulacyjne zbliżone do samogłosek, jednak, podobnie jak spółgłoski, nie mogą stanowić ośrodka sylaby. Dlatego często są traktowane przez lingwistów jako oddzielna kategoria dźwięków mowy. W omawianym tu rozwiązaniu przeważyło jednak kryterium funkcjonalne, dlatego półsamogłoski są traktowane jako komponent grup spółgłoskowych. Takie podejście jest bardzo wygodne dla zastosowań praktycznych, szczególnie dla praktycznej sylabifikacji wyrazów, ponieważ wyznacza ono dwa rozdzielne zbiory głosek — takich, które mogą stanowić tylko ośrodek sylaby, oraz takich, które mogą należeć tylko do marginaliów sylaby.

[4] Dana grupa spółgłoskowa lub pojedyncza spółgłoska jest wewnątrzwyrazowa, jeżeli w ramach jednego wyrazu jest ona otoczona przez samogłoski (ośrodki sylab). W przeciwnym razie jest to grupa (lub spółgłoska) umiejscowiona w nagłosie lub w wygłosie wyrazu. Zasadniczo w języku polskim w ramach nagłosowych i wygłosowych grup spółgłoskowych nie występują granice sylab. Jednak z góry nie można wykluczyć takiej interpretacji w przypadku niektórych wyrazów. Omawiany model umożliwia wstawianie granic sylab w ramach nagłosowych oraz wygłosowych grup spółgłoskowych tylko w drugim etapie omawianej procedury.

całkowicie zdefiniowane w prezentowanym modelu wielowarstwowym. To oznacza, że jest to jedyny poziom wstawiania granic, który nie może być regulowany w systemie dzielącym wyrazy na sylaby. Ten etap jest realizowany po wyznaczeniu w danym wyrazie liter oznaczających ośrodki sylab (zob. par. 3.3). Zasada dotycząca wyznaczania tego typu granic jest stosunkowo prosta — znacznik granicy jest wstawiany między dwiema sąsiadującymi literami oznaczającymi samogłoski, ale tylko jeżeli te litery są różne, na przykład w wyrazach: *aeroplan*, *geoida*, *samoistny*. Natomiast dwie identyczne sąsiadujące litery oznaczające samogłoski nie wymagają rozdzielenia znacznikiem granicy. Dotyczy to na przykład wyrazów: *kopii*, *anarchii*, *unii*.

3.5. Konstrukcja reguł modyfikacyjnych (etap 6)

Reguły modyfikacyjne stosuje się w odniesieniu do wewnątrzwyrazowych ortograficznych grup spółgłoskowych (wyznaczonych w etapie piątym) przed zastosowaniem podziału wynikającego z przyjętych zasad fonologicznych (zob. par. 3.6) oraz po zastosowaniu reguł odnoszących się do nagłosowych oraz do wygłosowych ciągów znaków (zob. par. 3.2). Zatem reguły modyfikacyjne są blokowane przez reguły stosowane w etapie drugim, jednak one same blokują podział wykonywany w etapie siódmym.

Nawiązując do informacji przedstawionych w części 2.1, można wymienić następujące przyczyny modyfikacji podziału wynikającego z przyjętych zasad fonologicznych:

— jeżeli osoba projektująca konkretny zbiór reguł będzie opierać się na innej koncepcji fonologicznej niż koncepcja przyjęta w modelu wielowarstwowym,
— poza tym reguły modyfikacyjne powinny być stosowane w przypadku grup spółgłoskowych, dla których nie da się wskazać granicy na podstawie przyjętych zasad fonologicznych. Mogą być one również stosowane, jeżeli z jakichkolwiek względów podział wynikający z fonologii zostanie uznany za niewłaściwy (na przykład ze względu na wynik badania poczucia subiektywnego).

Reguły modyfikacyjne mają prostą strukturę. Pierwszy człon każdej reguły tego typu zawiera konkretną ortograficzną wewnątrzwyrazową grupę spółgłoskową, natomiast człon drugi zawiera tę samą grupę z wstawionym znacznikiem granicy sylab, na przykład: *sk*>|*sk*. Podobnie jak w przypadku reguł omówionych w punkcie 3.2 dla reguł modyfikacyjnych można definiować zbiory wyjątków. Zapis dowolnej reguły modyfikacyjnej rozpoczyna się od słowa !MOD. Następnie po znaku dwukropka podawana jest reguła. Ewentualna lista wyjątków jest zapisywana w jednym ciągu (bez spacji) po słowie !EXC i po znaku dwukropka, na przykład:

!MOD:sk>|sk!EXC:#łask,#kask;

Funkcja reguł modyfikacyjnych może być zastąpiona przez reguły oparte na nagłosowych oraz wygłosowych ciągach znaków (etap drugi procedury omówionej w części 3.1). W paragrafie 3.2 już wspomniano, że takie reguły mogą dotyczyć dowolnych ciągów znaków i ich zastosowanie niekoniecznie musi mieć związek ze strukturą mor-

fologiczną wyrazów. Reguły modyfikacyjne dotyczą tylko ortograficznych wewnątrzwyrazowych grup spółgłoskowych (bez kontekstu samogłoskowego). Jeżeli zaistnieje potrzeba określenia szerszego kontekstu dla danej grupy spółgłoskowej, to można to zrobić tylko poprzez przypisanie reguły modyfikacyjnej do etapu drugiego omawianej procedury. Uzyskuje się to poprzez umieszczenia słowa !MOR przed zapisem reguły (zamiast słowa !MOD).

3.6. Odniesienie do zasad fonologicznych (etap 7)

W języku polskim występują różnice ilościowe i jakościowe między zapisem ortograficznym i transkrypcją fonologiczną. Jeden znak ortograficzny może oznaczać różne fonemy (w zależności od kontekstu) lub dwa fonemy, natomiast pojedyncze fonemy mogą być zapisywane przy użyciu dwóch lub nawet trzech znaków (Ostaszewska, Tambor 2002).

Wartości sonorności są przypisywane do dźwięków mowy, zatem, mając na uwadze wspomniane wyżej problemy, nie można ich przypisywać do znaków ortograficznych. Aby rozwiązać ten problem, został utworzony moduł projekcji zapisu ortograficznego wewnątrzwyrazowych grup spółgłoskowych na transkrypcję fonologiczną. Moduł ten przypisuje kolejne litery grup spółgłoskowych do odpowiednich fonemów. Algorytm wyznacza miejsce granicy w fonologicznym zapisie grupy spółgłoskowej. Dzięki modułowi projekcji ta granica może być przeniesiona na odpowiedni zapis ortograficzny. Granica jest wyznaczana na podstawie fonologicznej zasady sonorności i fonologicznej zasady maksymalnego nagłosu przy założeniu, że wartość sonorności kolejnych spółgłosek w wygłosie sylaby musi maleć, a w nagłosie sylaby musi ona wzrastać (nie jest dopuszczalna równa wartość sonorności w wygłosie oraz w nagłosie sylaby). W ten sposób można wyznaczyć tylko jedną granicę lub nie da się wyznaczyć żadnej granicy — nie ma możliwości wyznaczenia kilku różnych granic.

Zatem sposób wyznaczania granic na podstawie zasad fonologicznych jest stały (w etapie siódmym procedury omówionej w części 3.1). Ten podział może być modyfikowany przez reguły stosowane we wcześniejszych etapach (zob. par. 3.2 i 3.5). Jednak w etapie siódmym istnieje możliwość odniesienia do dowolnej skali sonorności. Moduł projekcji zapisu ortograficznego wewnątrzwyrazowych grup spółgłoskowych na transkrypcję fonologiczną posługuje się określonym inwentarzem fonologicznym oraz określoną transkrypcją fonologiczną SAMPA[5, 6]. W tabeli pierwszej wymieniono wszystkie fonemy spółgłoskowe użyte w omawianym module. Plik zawierający definicję reguł dla konkretnego systemu dzielącego wyrazy na sylaby (opartego na modelu wielowarstwowym) może zawierać również zapisy dotyczące wartości sonorności właściwej dla poszczególnych fonemów. Wyglądają one następująco:

[5] Nazwa SAMPA została pierwszy raz użyta przez Johna Wellsa: http://www.phon.ucl.ac.uk/home/wells/. Informacje dotyczące transkrypcji SAMPA: http://www.phon.ucl.ac.uk/home/sampa/.

[6] Istnieje kilka wersji transkrypcji SAMPA dla języka polskiego w ramach różnych rozwiązań praktycznych i technicznych (Bachan 2007).

!SON:/ts/=1;
!SON:/s/=2;

Natomiast przypisanie wartości sonorności do fonemów samogłoskowych (ośrodków sylab nie będących komponentami grup spółgłoskowych) wykonuje się w sposób następujący:

!SON:/V/=6;

Tabela 1. Transkrypcja użyta w module projekcji wewnątrzwyrazowych grup spółgłoskowych

Lp.	Fonem SAMPA	Przykład SAMPA	Przykład zapis ort.	Lp.	Fonem SAMPA	Przykład SAMPA	Przykład zapis ort.
1.	/w/	/p.u.w.k.a/	półka	15.	/z´/	/z´.a.r.n.o/	ziarno
2.	/j/	/j.e.d.e.n/	jeden	16.	/x/	/k.u.x.n´.a/	kuchnia
3.	/l/	/v.j.e.l.e/	wiele	17.	/p/	/p.a.l.e.ts/	palec
4.	/r/	/r.y.b.a/	ryba	18.	/b/	/b.u.d.a/	buda
5.	/m/	/m.o.Z.e/	morze	19.	/t/	/t.a.m.a/	tama
6.	/n/	/m.o.n.e.t.a/	moneta	20.	/d/	/d.o.m/	dom
7.	/n´/	/k.o.n´/	koński	21.	/k/	/p.o.k.u.j/	pokój
8.	/f/	/f.u.t.r.o/	futro	22.	/g/	/g.o.s´.ts´/	gość
9.	/v/	/v.j.a.t.r/	wiatr	23.	/ts/	/ts.y.r.k/	cyrk
10.	/s/	/v.y.s.o.k.i/	wysoki	24.	/dz/	/dz.v.o.n.e.k/	dzwonek
11.	/z/	/k.o.z.a/	koza	25.	/tS/	/tS.a.s/	czas
12.	/S/	/m.a.S.t/	maszt	26.	/dZ/	/dZ.u.m.a/	dżuma
13.	/Z/	/k.o.Z.e.n´/	korzeń	27.	/ts´/	/k.o.ts´.o.w/	kocioł
14.	/s´/	/s´.m.j.e.x/	śmiech	28.	/dz´/	/dz´.a.w.k.a/	działka

Można zrezygnować z definiowania skali sonorności i posłużyć się skalą zaproponowaną przez Jolantę Szpyrę-Kozłowską (została ona przytoczona w tabeli 2). W omawianym modelu wartości tej skali są domyślnie przypisane do poszczególnych fonemów, zatem w celu jej użycia nie trzeba wykonywać żadnych dodatkowych czynności.

Tabela 2. Domyślna skala sonorności

Klasa głosek	Elementy	Sonorność
Samogłoski	/V/	6
Półsamogłoski	/w/, /j/	5
Spółgłoski płynne	/l/, /r/	4
Spółgłoski nosowe	/m/, /n/, /n´/	3
Spółgłoski szczelinowe	/f/, /v/, /s/, /z/, /S/, /Z/, /s´/, /z´/, /x/	2
Spółgłoski zwarte: zwarto-wybuchowe oraz zwarto-szczelinowe	/p/, /b/, /t/, /d/, /k/, /g/, /ts/, /dz/, /tS/, /dZ/, /ts´/, /dz´/	1

Informacje przedstawione w tej części zostaną poparte przykładem. Ortograficzny wyraz *uschnąć* zawiera grupę spółgłoskową *schn*. Zapis tej grupy w module projekcji wygląda następująco: s.ch.n>s.x.n. Zatem literze *s* odpowiada fonem /s/, dwuznakowi *ch* odpowiada fonem /x/, natomiast literze *n* odpowiada fonem /n/[7]. Domyślna skala sonorności przypisuje do tych fonemów następujące wartości sonorności: 2 (do fonemu /s/), 2 (do fonemu /x/) oraz 3 (do fonemu /n/). Przy uwzględnieniu samogłosek otaczających tę grupę otrzymywana jest następująca struktura wartości sonorności: 6-2-2-3-6. Jedyne miejsce, od którego następuje wzrost wartości sonorności w obu kierunkach, znajduje się między drugim oraz trzecim elementem tego ciągu. Zatem granica wynikająca z omawianej metody zostaje wstawiona między fonemem /s/ oraz /x/, a następnie zostaje ona przeniesiona na zapis ortograficzny przetwarzanej grupy: *s*|*chn*.

3.7. Systemy probabilistyczne

We wcześniejszych paragrafach przedstawiono możliwości związane z tworzeniem reguł podziału dla różnych etapów działania algorytmu dzielącego wyrazy ortograficzne na sylaby. Omówiono sposób tworzenia reguł uwzględniających strukturę morfologiczną oraz reguł modyfikujących podział wynikający z zasad fonologicznych. Reguły te mają statyczny charakter — co oznacza, że po spełnieniu określonych warunków są one zawsze stosowane. Prezentowany wielowarstwowy model może wykorzystywać również podejście statystyczne. Dotyczy to zarówno reguł powiązanych ze strukturą morfologiczną wyrazu (reguł opartych na nagłosowych oraz na wygłosowych sekwensach ortograficznych), jak i reguł modyfikujących podział wynikający z zasad fonologicznych.

Tworzenie reguł statystycznych wiąże się z modyfikacją drugiego członu reguły. Powinien on zawierać listę różnych możliwości wstawiania znacznika granicy wraz z informacją o prawdopodobieństwie każdej możliwości (jest ono zapisywane przy użyciu liczby całkowitej w nawiasie okrągłym). Suma wszystkich wartości prawdopodobieństwa[8] na liście nie może przekroczyć 100. Jeżeli suma wartości prawdopodobieństwa wszystkich możliwości podziału wymienionych na liście jest mniejsza niż 100, to różnica między wartością 100 a tą sumą jest prawdopodobieństwem, że nie zostanie zastosowany żaden podział. Na przykład jeżeli suma prawdopodobieństwa wszystkich możliwości podziału wyniesie 60 (0,6), to prawdopodobieństwo tego, że nie zostanie wstawiony znacznik podziału, wynosi 0,4. Poniżej podano przykładowy zapis reguły statystycznej:

[7] Taka transkrypcja litery *n* należącej do wewnątrzwyrazowej ortograficznej grupy *schn* jest możliwa tylko przy założeniu, że w kontekście prawostronnym tej grupy znajduje się litera oznaczająca samogłoskę, która jest różna od litery *i*. W punkcie 3.1 podkreślono, że w etapie piątym procedury wyznaczana jest litera oznaczająca samogłoskę umiejscowioną w kontekście prawostronnym grupy spółgłoskowej. Ta informacja jest niezbędna dla prawidłowego działania modułu projekcji.

[8] Podane wartości wskazują na setne części wartości 1. Na przykład wartość 50 oznacza prawdopodobieństwo 0,5.

!MOD:mn>|mn(33),m|n(67);

Podana reguła dotyczy między innymi wyrazu *amnezja* (wartości dotyczące prawdopodobieństwa wynikają z badań, o których wspomniano w paragrafie drugim). Do takich struktur reguł można dołączyć listę wyjątków zgodnie z zasadami omówionymi w części 3.2.

3.8. Systemy ustalone

Omówiona w poprzednich częściach procedura umożliwia uwzględnienie różnych czynników mających wpływ na sylabifikację. Umożliwia ona również modyfikowanie wpływu tych czynników — przede wszystkim poprzez korekty reguł podziału. Jednak sama procedura jest dość złożona i dość wymagająca pod kątem obliczeniowym. Dla niektórych zastosowań możliwość korekty reguł podziału nie odgrywa znaczącej roli, natomiast istotna jest szybkość działania. Z myślą o takich zastosowaniach przewidziano możliwość generowania systemów ustalonych (lub inaczej: systemów stałych).

Systemy ustalone są generowane na podstawie wszystkich elementów omówionych w tym paragrafie oraz na podstawie określonego słownika. Słownik ten powinien zawierać wszystkie formy fleksyjne wyrazów[9]. Stały system dzielący wyrazy na sylaby składa się z pozycji, które obejmują dwa elementy: zapis ortograficzny danej formy fleksyjnej oraz zapis tej samej formy z wstawionymi znacznikami sylab. Zatem system ustalony można uzyskać poprzez zastosowanie omówionej procedury sylabifikacji w odniesieniu do wykazu wszystkich form fleksyjnych języka polskiego. Uzyskany wykaz umożliwia natychmiastową sylabifikację wyrazów tekstu — bez poddawania ich omówionej procedurze. Jednak system ustalony nie daje możliwości wprowadzania korekt w regułach, można jedynie korygować podział poszczególnych form fleksyjnych.

Możliwość generowania systemu ustalonego komplikuje się nieco, jeżeli reguły podziału są oparte na podejściu statystycznym. Reguły statystyczne mogą być stosowane w drugim oraz w szóstym etapie procedury omówionej w części 3.1. Poza tym dany wyraz może zawierać więcej grup spółgłoskowych, zatem ostateczne prawdopodobieństwo poszczególnych wersji podziału danego wyrazu może wynikać z wielu czynników. Dlatego generowanie systemu ustalonego w odniesieniu do wyrazów, dla których mają zastosowanie reguły statystyczne, oparte jest na metodzie empirycznej. Metoda ta zakłada, że dana forma fleksyjna poddawana jest procedurze sylabifikacyjnej 100-krotnie. Dzięki temu łatwo można obliczyć prawdopodobieństwo różnych wersji podziału tej formy fleksyjnej. To prawdopodobieństwo jest brane pod uwagę przy właściwej sylabifikacji przeprowadzanej na podstawie systemu ustalonego. Takie wyrazy są zapisywane w systemie ustalonym wraz z informacją o wszystkich wariantach podziału oraz informacją o prawdopodobieństwie tych wariantów.

[9] Dla języka polskiego dostępny jest słownik sjp.pl — licencja tego słownika umożliwia jego bezpłatne użycie do dowolnych celów oraz przekształcenia związane z dostosowywaniem tego słownika do różnych zadań.

3.9. Systemy generatywne

Istnieje możliwość tworzenia systemów, przy użyciu których można generować zbiory reguł, które spełniają określone założenia. To podejście opiera się na tworzeniu alternatywnych względem siebie różnych grup reguł. W ramach jednego systemu generatywnego można stworzyć dowolną liczbę zbiorów grup reguł. Zatem w ramach systemu generatywnego trzeba zdefiniować przynajmniej jeden zbiór grup reguł — należy mu nadać nazwę oraz ewentualnie załączyć opis. Każdy element takiego zbioru jest grupą reguł i każda grupa reguł ma swoją nazwę (oraz opcjonalnie opis).

W czasie użytkowania systemu generatywnego z każdego zdefiniowanego zbioru musi być wybrana dokładnie jedna grupa reguł. W ten sposób można utworzyć standardowy zbiór reguł dla systemu opartego na omawianym modelu wielowarstwowym. Warto zauważyć, że suma wszystkich (różnych) zbiorów reguł, które można utworzyć przy użyciu systemu generatywnego, jest równa iloczynowi liczebności wszystkich zbiorów grup reguł. Zbiór alternatywnych grup reguł definiuje się w następujący sposób:

```
!GEN:nazwa_zbioru_grup_reguł!DES:opis zbioru grup;
!GRP:nazwa_grupy_reguł_1!DES:opis_grupy_reguł_1;
reguła_1;
reguła_2;
…
!GRP:nazwa_grupy_reguł_2!DES:opis_grupy_reguł_2;
reguła_1;
reguła_2;
…
!END;
```

Definicja zbioru grup reguł dla systemu generatywnego rozpoczyna się od słowa !GEN, po którym podawana jest jego nazwa. W tej samej linii po słowie !DES można umieścić opis zbioru. Definicja grupy reguł rozpoczyna się od słowa !GRP. Tutaj również po słowie !DES można umieścić opis grupy. W kolejnych wierszach wymieniane są wszystkie reguły należące do danej grupy. Dany zbiór może zawierać dowolną liczbę definicji grup reguł. Zapis: !END; kończy definicję zbioru dla systemu generatywnego.

W ten sam sposób można utworzyć różne grupy definiujące różne wersje fonologicznej skali sonorności. Dla osiągnięcia tego celu w poszczególnych grupach (należących do danego zbioru) należy ująć przypisania wartości sonorności do poszczególnych fonemów (zob. par. 3.6).

W pliku tekstowym przechowującym definicję systemu generatywnego można również zamieszczać reguły poza sekcjami oznaczonymi słowem !GEN. Takie reguły będą dołączane do każdego standardowego zbioru reguł wygenerowanego przez dany system generatywny.

4. Przykłady użycia systemów opartych na modelu wielowarstwowym

W paragrafie trzecim zamieszczono dokładny opis operacji zdefiniowanych w ramach wielowarstwowego modelu podziału wyrazów na sylaby. Natomiast w tej części przedstawiono opisy dotyczące przetwarzania przykładowych wyrazów ortograficznych. Przy każdym opisie podano założenie dotyczące obecności konkretnych reguł — ponieważ żadne reguły nie są częścią prezentowanego modelu (są one utworzone zgodnie z założeniami tego modelu). Zatem każdy z trzech opisów może należeć do różnych systemów opartych na modelu wielowarstwowym. Zakłada się, że została użyta domyślna skala sonorności (zob. par. 3.6).

4.1. Przetwarzanie wyrazu *konto*

Pierwszy prezentowany opis dotyczy wyrazu *konto*. Jest to wyraz dwusylabowy, w którym granica sylab jest umiejscowiona w obrębie ortograficznej grupy *nt*. Nie ma potrzeby używania reguł dla przetwarzania tego wyrazu, ponieważ grupa *nt* może być podzielona na podstawie przyjętych zasad fonologicznych. Nawiązując do procedury omówionej w części 3.1 można wyznaczyć kolejne etapy związane z przetwarzanie wyrazu *konto*:

Etap 1

Zostają wyznaczone wszystkie początkowe oraz końcowe fragmenty w wyrazie *#konto#*:

#konto, *#kont*, *#kon*, *#ko*, *#k*,
konto#, *onto#*, *nto#*, *to#*, *o#*.

Etap 2

Zbiór reguł nie zawiera żadnej reguły oznaczonej etykietą !MOR, której pierwszy człon pokrywałby się z jakimkolwiek ciągiem znaków uzyskanym w etapie pierwszym.

Etap 3

Oddzielny algorytm wyznaczający ośrodki sylab (zob. par. 3.3) wskazuje na indeksy znaków oznaczających ośrodki sylab: 2 oraz 5 (przy założeniu, że początkowym indeksem jest liczba 0 oraz że symbol # również jest uwzględniany przy indeksowaniu znaków).

Etap 4

Sprawdzenie, czy przetwarzany wyraz nie zawiera dwóch sąsiadujących ośrodków sylab (w celu ewentualnego wstawienia znacznika granicy między tymi ośrodkami).

Etap 5

Na podstawie wartości indeksów otrzymanych w kroku trzecim następuje wyznaczenie wewnątrzwyrazowej grupy *nt*.

Etap 6

Zbiór reguł nie zawiera żadnej reguły oznaczonych etykietą !MOD, której pierwszy człon pokrywałby się ze zbitką *nt*.

Etap 7

W odniesieniu do grupy *nt* stosowane są reguły fonologiczne. Odbywa się to dzięki informacji o projekcji tej grupy na zapis fonologiczny: /n.t/. Na podstawie domyślnych wartości sonorności przyjętych dla fonemu /n/ oraz dla fonemu /t/ (zob. par. 3.6), a także na podstawie przyjętych założeń dotyczących stosowania zasady sonorności oraz zasady maksymalnego nagłosu granica zostaje umiejscowiona pomiędzy tymi fonemami. Dzięki mechanizmowi projekcji możliwe jest przeniesienie tej granicy na zapis ortograficzny grupy.

4.2. Przetwarzanie wyrazu *okołozwrotnikowy*

Kolejny opis dotyczy ortograficznego wyrazu *okołozwrotnikowy*. Zakłada się, że zbiór reguł zawiera następującą regułę:

!MOR:#około>#około|!EXC:#około#;

Podana reguła ma zastosowanie dla wyrazów typu: *okołosłoneczny*, *okołoksiężycowy*, *okołostatutowy*. Jednak nie odnosi się do wyrazu *około* (jest on umieszczony na liście wyjątków). Przetwarzanie wyrazu *okołozwrotnikowy* odbywa się w sposób następujący:

Etap 1

Zostają wyznaczone wszystkie początkowe oraz końcowe fragmenty przetwarzanego wyrazu:

#okołozwrotnikowy, *#okołozwrotnikow*, *#okołozwrotniko*, *#okołozwrotnik*, *#okołozwrotni*, *#okołozwrotn*, *#okołozwrot*, *#okołozwro*, *#okołozwr*, *#okołozw*, *#okołoz*, *#około*, *#okoł*, *#oko*, *#ok*, *#o*,

okołozwrotnikowy#, *kołozwrotnikowy#*, *ołozwrotnikowy#*, *łozwrotnikowy#*, *ozwrotnikowy#*, *zwrotnikowy#*, *wrotnikowy#*, *rotnikowy#*, *otnikowy#*, *tnikowy#*, *nikowy#*, *ikowy#*, *kowy#*, *owy#*, *wy#*, *y#.*

Etap 2

Wyznaczony w etapie pierwszym zbiór zawiera ciąg znaków: *#około*, który jest identyczny jak pierwszy człon reguły podanej na wstępie do tej części. Zatem reguła ta zostaje zastosowana, w wyniku czego przetwarzany wyraz przyjmuje następujący kształt: *#około|zwrotnikowy#.*

Etap 3

Następuje wyznaczenie ośrodków sylab w przetwarzanym wyrazie — są nimi znaki o indeksach: 1, 3, 5, 10, 13, 15, 17 (do znacznika granicy sylab wstawionego w poprzednim kroku również zostaje przypisany oddzielny indeks).

Etap 4

Następuje sprawdzenie, czy przetwarzany wyraz nie zawiera sąsiadujących ośrodków sylab.

Etap 5

Na podstawie wartości indeksów otrzymanych w kroku trzecim następuje wyznaczenie wewnątrzwyrazowych ortograficznych grup spółgłoskowych oraz pojedynczych spółgłosek umiejscowionych między ośrodkami sylab: *k*, *ł*, |*zwr*, *tn*, *k*, *w*.

Etap 6

Zbiór reguł nie zawiera żadnej reguły oznaczonych etykietą !MOD, której pierwszy człon pokrywałby się z jakimkolwiek fragmentem wyznaczonym w etapie piątym. Jeżeli zbiór reguł zawierałby regułę dla grupy *zwr*, to ona również nie byłaby zastosowana, ponieważ ten ciąg znaków różni się od ciągu znaków: |*zwr.*

Etap 7

Na podstawie przyjętych zasad fonologicznych oraz przy użyciu mechanizmu projekcji grup ortograficznych na transkrypcję fonologiczną zostaje wyznaczona granica przed grupą *tn*. Grupa |*zwr* zawiera już znacznik granicy, dlatego w jej przypadku granica nie jest wyznaczana ponownie.

Etap 8

Zgodnie z założeniami dotyczącymi funkcjonowania opisywanego modelu, jeżeli dla danej wewnątrzwyrazowej zbitki lub pojedynczej spółgłoski granica nie została wyznaczona w poprzednich krokach, to zostaje ona umiejscowiona przed grupą spółgłoskową (lub pojedynczą spółgłoską). W przetwarzanym wyrazie dotyczy to spółgłosek: *k*, *ł*, *k*, *w.* W przypadku tych liter granica nie została wstawiona w etapie siódmym, ponieważ mechanizm projekcji zapisu ortograficznego na transkrypcję fonologiczną dotyczy tylko grup spółgłoskowych.

4.3. Przetwarzanie wyrazu *obmyślić*

Ostatni prezentowany przykład dotyczy ortograficznego wyrazu *obmyślić*. Zostaje przyjęte założenie, że zbiór reguł zawiera następującą regułę:

!MOR:#ob>#ob|!EXC:#obfit,#obj,#obraz,#obrec,#obrącz,#obrus,#obron,#obroń;
#oblach,#obligator,#oblin,#obłąk,#obław,#obły#,#obłego#,#obłym#,#obłe#,
#obli#,#obnaż,#obrac,#obrad,#obraz,#obraź,#obraż,#obrecj,#obredl,#obronn,#obroń,
#obrot,#obrót,#obróż,#obryzg,#obrz,#obsad,#obscuru,#obserw,#obses,#obski,#obsług,
#obstruk,#obsydian,#obszar,#obszern;

Podana reguła powoduje wstawienie znacznika granicy za słowem ortograficznym *ob*. Zawiera ona szereg wyjątków, między innymi ciąg znaków *#obraz* — jego obec-

ność sprawia, że podana reguła nie będzie stosowana do żadnej formy fleksyjnej wyrazu *obraz*. Drugie założenie dotyczy obecności następującej reguły:

!MOD:śl>ś|l;

Podana reguła zapewnia podział ortograficznej zbitki *śl* w sposób alternatywny względem podziału wynikającego z przyjętych zasad fonologicznych. W dalszym ciągu omówiono etapy związane z przetwarzaniem wyrazu *obmyślić*:

Etap 1

Zostają wyznaczone wszystkie początkowe oraz końcowe fragmenty wyrazu *#obmyślić#*:

#obmyślić, *#obmyśli*, *#obmyśl*, *#obmyś*, *#obmy*, *#obm*, *#ob*, *#o*,
obmyślić#, *bmyślić#*, *myślić#*, *yślić#*, *ślić#*, *lić#*, *ić#*, *ć#*.

Etap 2

Wyznaczony w etapie drugim zbiór zawiera ciąg znaków: *#ob*, który jest identyczny jak pierwszy człon podanej we wstępie reguły. Poza tym żaden ciąg znaków podanych na liście wyjątków tej reguły nie jest częścią przetwarzanego wyrazu. Dlatego ta reguła może by użyta. Po jej zastosowaniu przetwarzany wyraz przyjmuje następujący kształt: *#ob|myślić#*.

Etap 3

Następuje wyznaczenie ośrodków sylab, w przetwarzanym wyrazie są nimi znaki o indeksach: 1, 5, 8 (do znacznika granicy wstawionego w poprzednim kroku również jest przypisany oddzielny indeks).

Etap 4

Następuje sprawdzenie, czy przetwarzany wyraz nie zawiera dwóch sąsiadujących ośrodków sylab.

Etap 5

Na podstawie wartości indeksów otrzymanych w kroku trzecim następuje wyznaczenie wewnątrzwyrazowych ortograficznych grup spółgłoskowych: *b|m*, *śl*.

Etap 6

Założony zbiór reguł zawiera regułę oznaczoną symbolem !MOD, która jest odpowiednia dla grupy *śl*. Po zastosowaniu tej reguły podana grupa przyjmuje następujący kształt: *ś|l.* Natomiast w odniesieniu do grupy *|bm* nie mogą już być wykonywane żadne operacje, ponieważ zawiera ona znacznik granicy wstawiony wcześniej.

Etap 7

Wszystkie wewnątrzwyrazowe grupy spółgłoskowe zawierają znacznik granicy sylab wstawiony we wcześniejszych etapach, dlatego zasady fonologiczne nie są stosowane.

5. Testowanie modelu

W czasie pisania niniejszego artykułu istniała już pierwsza (prototypowa) wersja systemu dzielącego wyrazy na sylaby opartego na omówionym modelu wielowarstwowym. Zgodnie z informacjami zawartymi we wstępie, dany system podziału na sylaby obejmuje zarówno program komputerowy, jak i konkretny zbiór reguł. Zatem można było przeprowadzić pierwsze testy, które wypadły pomyślnie. W niniejszym paragrafie przedstawiono niektóre wyniki tych testów, przy czym zostały one dobrane w taki sposób, aby ukazać realizację podstawowych założeń i mechanizmów związanych z przetwarzaniem wyrazów ortograficznych w omówionym modelu — przede wszystkim: osobne odniesienie do fonologii i morfologii, blokowanie podziału przez reguły stosowane wcześniej i stosowanie wyjątków. Przedstawiono również jeden przykład dotyczący podejścia statystycznego (użyto wyników badań, do których odniesiono się w paragrafie drugim).

Pierwszy etap testu dotyczył możliwości blokowania podziału wynikającego z zasad fonologicznych przez reguły modyfikacyjne. Utworzono następującą listę wyrazów (ortograficzne grupy spółgłoskowe, które są przedmiotem rozważań, zostały podkreślone):

ze<u>mst</u>a, *per<u>sp</u>ektywa*, *po<u>rtf</u>el*, *ma<u>jstr</u>em*, *administ<u>r</u>acja*, *e<u>gz</u>amin*, *by<u>dl</u>ak*, *agresyw<u>ny</u>*, *po<u>wn</u>osić*, *wy<u>dm</u>a*, *wy<u>dm</u>uchać*.

Przy pustym zbiorze reguł podział na sylaby opiera się na przyjętych zasadach fonologicznych. Program wygenerował listę tych samych wyrazów ze wstawionymi znacznikami granic sylab:

zems|ta, *pers|pek|ty|wa*, *por|tfel*, *majs|trem*, *a|dmi|nis|tra|cja*, *e|gza|min*, *by|dlak*, *a|gre|sy|wny*, *po|wno|sić*, *wy|dma*, *wy|dmu|chać.*

Zatem wszystkie przetworzone wyrazy dało się podzielić przy odwołaniu do fonologii. Następnie do zbioru reguł dodano następujące reguły modyfikacyjne:

```
!MOD:mst>m|st;
!MOD:rsp>r|sp;
!MOD:rtf>rt|f;
!MOD:jstr>j|str;
!MOD:str>|str;
!MOD:gz>g|z;
!MOD:dl>d|l;
!MOD:wn>w|n!EXC:#powno;
!MOD:dm>d|m!EXC:#wydmu;
```

Obecność tych reguł sprawiła, że uzyskany podział wyrazów jest inny niż przy pierwszym podejściu:

zem|sta, *per|spek|ty|wa*, *port|fel*, *maj|strem*, *ad|mi|ni|stra|cja*, *eg|za|min*, *byd|lak*, *a|gre|syw|ny*, *po|wno|sić*, *wyd|ma*, *wy|dmu|chać.*

Jednak zmiany nie obejmują wyrazów: *pownosić* oraz *wydmuchać*, ponieważ odpowiednie ciągi znaków pozwalające na ich identyfikację zostały zaliczone do wyjątków. Zatem te dwa wyrazy zostały podzielone przy odwołaniu do przyjętych zasad fonologicznych (w etapie siódmym procedury podziału). Dla uzyskania identycznego efektu można zrezygnować z definiowania wyjątków w regułach modyfikacyjnych i jednocześnie utworzyć odpowiednie reguły odnoszące się do struktury morfologicznej:

```
!MOR:#powno>#po|wno;
!MOR:#wydmu>#wy|dmu;
```

W tym przypadku podział wykonany w etapie drugim procedury blokuje możliwość podziału na podstawie reguł modyfikacyjnych (w etapie szóstym).

Kolejny etap testu dotyczył wyrazów zawierających wewnątrzwyrazowe grupy spółgłoskowe, dla których przyjęte zasady fonologiczne nie wskazują żadnego rozwiązania:

herbstem, *gangsterski*, *powściągliwy*, *tekstem*, *ekspres.*

Przetworzenie takich wyrazów przy pustym zbiorze reguł dało następujący rezultat:

he|rbstem, *ga|ngsters|ki*, *po|wścią|gli|wy*, *te|kstem*, *e|kspres.*

Zgodnie z informacjami podanymi w części 3.1 granice zostały wstawione przed tymi grupami dopiero w ósmym etapie omówionej procedury. Aby to zmienić, do zbioru reguł należy włączyć następujące reguły modyfikacyjne:

```
!MOD:rbst>rb|st;
!MOD:ngst>ng|st;
!MOD:wści>w|ści;
!MOD:kst>ks|t;
!MOD:kspr>ks|pr;
```

Obecność tych reguł w zbiorze reguł sprawiła, że wyrazy zostały podzielone w następujący sposób:

herb|stem, *gang|sters|ki*, *pow|ścią|gli|wy*, *teks|tem*, *eks|pres.*

Kolejny fragment testu dotyczył reguł odnoszących się do struktury morfologicznej wyrazów. Uwzględniono listę zawierającą następujące wyrazy:

dostudzić, *nadlecieć*, *nadworny.*

Podział tych wyrazów przy pustym zbiorze reguł jest oparty na przyjętych zasadach fonologicznych:

dos|tu|dzić, *na|dle|cieć*, *na|dwor|ny*.

Do zbioru reguł dodano dwie reguły odnoszące się do początkowych ciągów znaków:

!MOR:#dostudz>#do|studz;
!MOR:#nad>#nad|!EXC:#nad#,#nadwor;

Obecność tych reguł sprawiła, że rozpatrywane wyrazy zostały podzielone w następujący sposób:
do|*stu*|*dzić*, *nad*|*le*|*cieć*, *na*|*dwor*|*ny.*
Zatem umieszczenie ciągu znaków *#nadwor* na liście wyjątków drugiej wymienionej reguły sprawiło, że ta reguła nie została zastosowana dla wyrazu: *nadworny* (zatem został on podzielony przy odwołaniu do fonologii).

Ostatni prezentowany fragment testu dotyczy podejścia statystycznego. Założono, że zbiór reguł zawiera następującą regułę modyfikacyjną:

!MOD:rpl>r|pl(72),rp|l(28);

Zatem ta reguła zapewnia dwie możliwości podziału ortograficznej grupy *rpl*, przy czym prawdopodobieństwo pierwszej możliwości (r|pl) jest równe 0,72, natomiast prawdopodobieństwo drugiej możliwości (rp|l) wynosi 0,28.

Testowanie tej reguły polegało na jej 100-krotnym zastosowaniu w odniesieniu do wyrazu *ciepliwy*. Podział zgodny z pierwszą możliwością (*cier*|*pli*|*wy*) uzyskano 68 razy, natomiast drugi wymieniony wariant podziału (*cierp*|*li*|*wy*) wystąpił w przetworzonym tekście 32 razy.

6. Podsumowanie

Proponowany wielowarstwowy model sylabifikacji obejmuje szereg założeń dotyczących przetwarzania wyrazów ortograficznych, jednak definicja tego modelu nie precyzuje ostatecznego umiejscowienia granic sylab w konkretnych wyrazach ortograficznych, wynika ono bowiem z użytego zbioru reguł. Trzeba podkreślić, że tworzenie zbioru reguł dla dowolnego rozwiązania (systemu) opartego na modelu wielowarstwowym ma charakter obligatoryjny. To znaczy, że w ogóle żadna reguła nie musi być utworzona, żeby to rozwiązanie funkcjonowało. Jeżeli w danym systemie dzielącym wyrazy na sylaby zbiór reguł będzie pusty, to wstawianie znaczników granicy będzie się odbywało tylko w siódmym oraz w ósmym etapie procedury omówionej w części 3.1. To znaczy, że podział ten będzie oparty wyłącznie na przyjętych zasadach fonologicznych (etap siódmy). Natomiast w przypadku wewnątrzwyrazowych grup, dla których zasady fonologiczne nie wskażą żadnego rozwiązania, znacznik granicy będzie wstawiany przed tymi grupami (etap ósmy).

Wyznaczanie granic w etapie siódmym oparto na ścisłych założeniach dotyczących stosowania fonologicznej zasady sonorności oraz fonologicznej zasady maksy-

malnego nagłosu. Ta metoda w abstrakcyjny sposób odzwierciedla treść fonetycznych definicji sylaby, poza tym umożliwia wyznaczenie granic w niemal wszystkich wyrazach w tekstach. Dlatego jest to podstawowy sposób wyznaczania granic sylab w prezentowanym modelu wielowarstwowym. Jednak taki podział może zostać uznany za niewłaściwy, jeżeli projektant systemu dzielącego wyrazy na sylaby oprze się na innych koncepcjach fonologicznych lub uzna za priorytet podział wynikający z badań poczucia subiektywnego związanego z podziałem wyrazów na sylaby. Dlatego podział wynikający z fonologii może być zastąpiony przez podział sprecyzowany w regułach modyfikacyjnych, które są stosowane w etapie szóstym procedury omówionej w części 3.1. Kolejny poziom przesłaniania podziału wynikającego z fonologii jest realizowany jeszcze wcześniej — w etapie drugim. Struktura reguł dla tego etapu opiera się na początkowych oraz końcowych ciągach znaków w wyrazach ortograficznych. Podstawowa funkcja tych reguł ma związek ze strukturą morfologiczną wyrazu oraz z tzw. wyraźnymi granicami morfologicznymi, które najczęściej są odczuwane jako naturalne granice sylab. Jeżeli ten typ granic zostanie uznany za istotny dla danego rozwiązania, to można wybrać jedną z dwóch możliwości tworzenia reguł. Pierwsze podejście polega na tworzeniu reguł tylko dla konstrukcji morfologicznych, których podział jest sprzeczny z podziałem wynikającym z fonologii. Drugie podejście polega na utworzeniu reguł dla wszystkich struktur morfologicznych, które zostaną uznane za istotne — niezależnie od podziału wynikającego z fonologii. Reguły definiowane dla etapu drugiego (reguły oznaczone symbolem !MOR) mogą również posłużyć do blokowania podziału wynikającego z przyjętych zasad fonologicznych.

Informacje zawarte w poprzednim akapicie stanowią podsumowanie dotyczące prezentowanego modelu wielowarstwowego i ukazują jego nieograniczone możliwości w zakresie projektowania dowolnych systemów dzielących wyrazy języka polskiego na sylaby. Model ten daje możliwość kontroli podziału każdej ortograficznej wewnątrzwyrazowej grupy spółgłoskowej, a nawet każdej formy fleksyjnej. Wydaje się jednak, że najważniejszą cechą prezentowanego modelu jest wielowarstwowość, dzięki której podział związany z morfologią oraz z fonologią funkcjonuje oddzielnie. Mechanizm blokowania podziału gwarantuje, że granice wyznaczone w danym etapie nie mogą być modyfikowane w kolejnych etapach.

W artykule omówiono także dodatkowe zagadnienia związane z modelem wielowarstwowym — przede wszystkim możliwość tworzenia reguł statystycznych. Poruszono również zagadnienie systemów ustalonych, które przechowują informacje o podziale na sylaby wszystkich form fleksyjnych języka polskiego i jednocześnie umożliwiają pominięcie omówionej procedury podziału w czasie przetwarzania tekstu. Uzupełnieniem dla prezentowanego modelu jest koncepcja systemów generatywnych. Ich cel polega na generowaniu różnych zbiorów reguł.

W tej publikacji omówiono dokładnie wszystkie etapy związane z przetwarzaniem wyrazów. Wydaje się jednak, że w publikacjach przyszłych, które mogą dotyczyć konkretnych systemów opartych na zaprezentowanym modelu wielowarstwowym, taki poziom szczegółowości nie jest potrzebny. W opisach tych systemów można posługi-

wać się tytułowymi warstwami. W tym miejscu można zaproponować nazewnictwo dla poszczególnych warstw, które jest związane z ich podstawowymi funkcjami: warstwa morfologiczna (związana z podziałem wykonywanym w etapie drugim przedstawionej procedury), warstwa modyfikacyjna (związana z etapem szóstym) oraz warstwa fonologiczna (związana z etapem siódmym).

Bibliografia

Bachan J., 2007, Automatic Close Copy Speech Synthesis, Speech and Language Technology 9/10, s. 107––121.

Dobrogowska K., 1984, Śródgłosowe grupy spółgłosek w polskich tekstach popularnonaukowych, Polonica X, s. 15–34.

— 1990, Word internal consonant clusters in Polish artistic prose, Studia Phonetica Posnaniensia II, s. 43–67.

Dunaj B., 1985, Grupy spółgłoskowe współczesnej polszczyzny mówionej (w języku mieszkańców Krakowa), Prace Językoznawcze DCCCIII, s. 46–79.

Jassem W., Łobacz P., 1974, Fonotaktyczna analiza mówionego tekstu polskiego, Biuletyn Polskiego Towarzystwa Językoznawczego XXXII, s. 179–197.

Jespersen O., 1904, Lehrbuch der Phonetik, Leipzig, s. 1–276.

Polański E. (red.), 2006, Wielki słownik ortograficzny PWN z zasadami pisowni i interpunkcji, Warszawa, s. 85–87.

Polański K. (red.), 1999, Encyklopedia językoznawstwa ogólnego, Wrocław, s. 644.

Ostaszewska D., Tambor J., 2002, Fonetyka i fonologia współczesnego języka polskiego, Warszawa, s. 47––83.

Rubach J., Geert E.B., 1990, Edge of Constituent Effects in Polish, Natural Language & Linguistic Theory 8, No. 3, s. 427–463.

Selkirk E.O., 1980, The role of prosodic categories in English word stress, Linguistic Inquiry XI, s. 563––606.

Szpyra-Kozłowska J., 1998, The sonority scale and phonetic syllabification in Polish, Biuletyn Polskiego Towarzystwa Językoznawczego LIV, s. 63–82.

— 2002, Wprowadzenie do współczesnej fonologii, Lublin, s. 150–152.

Steffen-Batogowa M., 1975, Automatyzacja transkrypcji fonematycznej tekstów polskich, Warszawa, s. 78–90.

Śledziński D., 2013, Podział korpusu tekstów na sylaby — analiza polskich grup spółgłoskowych, Kwartalnik Językoznawczy XV, s. 48–100.

— 2016, Projekcja ortograficznych form wewnątrzwyrazowych grup spółgłoskowych na transkrypcję fonologiczną na potrzeby systemu dzielenia na sylaby wyrazów języka polskiego, Investigationes Linguisticae XXXIV, s. 51–72.

— 2017a, Badanie odczucia subiektywnego związanego z dzieleniem na sylaby wyrazów języka polskiego — podział grup złożonych ze spółgłosek właściwych, Acta Universitatis Lodziensis — Folia Linguistica LI, s. 93–107.

— 2017b, Badanie odczucia subiektywnego związanego z dzieleniem na sylaby wyrazów języka polskiego — podział grup złożonych ze spółgłosek sonornych, Acta Universitatis Lodziensis — Folia Linguistica LI, s. 109–127.

Trask R.L., 1996, A dictionary of phonetics and phonology, New York, s. 345.

Wierzchowska B., 1971, Wymowa polska, wyd. II, Warszawa, s. 214.

SUMMARY

A multilayer model for the syllabification of Polish words written in orthographic form

Keywords: syllable, syllabification, phonology, the Sonority Sequencing Principle, the Maximal Onset Principle.

Słowa kluczowe: sylaba, podział na sylaby, sylabifikacja, fonologia, zasada sonorności, zasad maksymalnego nagłosu.

This paper presents a model for the syllabification of Polish words written in orthographic form. It raises some significant issues related to syllabification, including the linguistic definition of a syllable and certain phonological principles: the Sonority Sequencing Principle and the Maximal Onset Principle. The article describes the steps of a designed syllabification procedure. This procedure provides four layers of syllable boundary placement. The first layer is related to the morphological structure of the word (it concerns mainly boundaries between a prefix and a stem). The structure of the rules in the first layer is based on the initial and final strings in words. The goal of the second layer is to put boundaries between syllable nuclei. The next layer modifies the boundaries that result from the phonological principles. The last layer places boundaries resulting from the phonological principles. Rules applied earlier mask rules on the succeeding layers. The article also presents a structure of rules based on probability. A description of static syllabification systems is also given — these work much faster, but it is not possible to modify the rules associated with them. The last described feature of the presented solution is the ability to create generative systems.