

Mgr Joanna Rabeiga-Wiśniewska
Zakład Językoznawstwa Komputerowego
Instytut Języka Polskiego
Uniwersytet Warszawski
jwisniewska@uw.edu.pl

Autoreferat rozprawy doktorskiej

Formalny opis derywacji w języku polskim. Rzeczowniki i przymiotniki.

Promotor: Prof. dr hab. Marek Świdziński

Przedmiotem rozprawy jest formalny opis zjawisk słowotwórczych w języku polskim, a wynikiem analizy procesów derywacyjnych – zbiór reguł, które obejmują sufiksację, pseudo-sufiksację, konwersję, prefiksację i pseudo-prefiksację rzeczowników i przymiotników. Opis zjawisk słowotwórczych wyprowadzam z przygotowanego starannie zbioru formalnych wykładników derywacyjnych. Dobór wykładników uzasadniony jest wysoką frekwencją słownikową, a następnie zweryfikowany przez dane słownikowe i korpusowe. W opisie derywacji uwzględniam charakterystykę fleksyjną bazy słowotwórczej i derywatu, uważając ją za jedną z najważniejszych składowych procesów słowotwórczych. Zajmuję się także szeroko problemem alternacji.

1. Cele i zakres pracy

Celem pracy jest przedstawienie procesów słowotwórstwa polskiego w ramach lingwistyki komputerowej. Wśród znanych mi propozycji formalnego ujęcia morfologii są tylko prace dotyczące fleksji (Gruszczyński 1989, Bień 1991). Powstały w ostatniej dekadzie narzędzia komputerowe przypisujące słowom charakterystykę fleksyjną, a wraz z nią odgadujące leksem (lematyzacja). Do tego celu jednak nie jest technicznie wymagana segmentacja słowa ani na temat fleksyjny i końcówkę, ani tym bardziej na morfemy. Choć więc pewne intuicje automatycznej analizy i syntezy form wyrazowych pojawiały się w literaturze, nie było do tej pory opracowania derywacji z perspektywy lingwistyki formalnej i komputerowej.

Obecnie w pracach językoznawczych wykorzystuje się duże korpusy tekstów. Aby w pełni skorzystać z materiału empirycznego, którego dostarczają korpusy, badacz potrzebuje wyrafinowanych narzędzi automatycznej analizy tekstu i wyszukiwania. Ich powstanie jest warunkowane dostępnością opisów języka naturalnego, opracowanych z myślą o zastosowaniach technicznych. Formalne, wręcz techniczne ujęcie derywacji w tej pracy jest równocześnie pewną propozycją automatyzacji analizy (i syntezy) morfologicznej i może stanowić część systemu przetwarzania tekstów języka polskiego.

Problem automatycznej analizy (i syntezy) słów można postrzegać z kilku perspektyw. Zrezygnowałam z opisu relacji semantycznych między formą bazy i derywatu, ponieważ przewidywanie znaczeń z wysokim prawdopodobieństwem jest możliwe dla niewielu tradycyjnych kategorii słowotwórczych, i to zwykle z ograniczeniami. Nieczęsto również dana kategoria ma jeden właściwy wykładnik, co praktycznie uniemożliwia związanie jej opisu z formalną regułą derywacyjną. Jeśli tylko było to możliwe, starałam się stawiać hipotezę co do znaczenia nowo powstałej konstrukcji. Wstępnie, na potrzeby programu komputerowego analizującego nowe słowo (takie, którego nie ma w słowniku) taki opis jest wystarczający. Zakładam, że wprowadzenie opisu semantycznego derywatów będzie możliwe na dalszym etapie analizy automatycznej.

Drugim ważnym zagadnieniem wiążącym się z tematyką rozprawy jest wybór zakresu procesów poddanych automatycznej analizie. Teoretycznie najlepsze byłoby objęcie regułami derywacyjnymi tylko najbardziej seryjnych procesów, w pełni przewidywalnych. Słowotwórstwo nie dzieli się jednak tylko na seryjne i na jednostkowe – między tymi biegunami znajduje się wiele procesów pośrednich. Z punktu widzenia praktycznych zastosowań opisu należy wziąć pod uwagę możliwie dużo formacji, kierując się ustalonymi kryteriami. Zakładam, że istnieje pewien słownik obejmujący bazy derywacyjne, derywaty jednostkowe i rzadkie. Program analizy słów korzystałby za każdym razem z tego słownika, aby sprawdzić, czy badane słowo zostało tam umieszczone. Analizowane byłyby tylko takie, które się w nim nie znalazły. Decyzja o włączeniu procesu derywacyjnego do formalnego opisu derywacji jest arbitralna, ale ma oparcie w danych empirycznych.

Innym problemem poruszonym w pracy jest formalny sposób analizy derywatów. W opisie, który z założenia musi zrezygnować z kategoryzacji semantycznej, opis fleksyjny okazuje się jednym z najważniejszych składników procesów słowotwórczych. Wybrane wykładniki derywacyjne są powiązane z formami baz słowotwórczych poprzez reguły derywacyjne, które w sposób jednoznaczny charakteryzują gramatycznie zarówno bazy, jak i derywaty.

Formalne opracowanie derywacji języka polskiego i jego implementacja powinny pokazać, które z wykładników derywacyjnych biorą największy udział w formowaniu derywatów w różnych tekstach, które z nich są obecnie produktywne oraz jakie kombinacje wykładników są możliwe. Takie ujęcie słowotwórstwa polskiego uzupełnia opisy czysto lingwistyczne.

2. Materiał empiryczny i metoda

Zbiór wykładników derywacji oparłam na polskiej literaturze przedmiotu, przykłady formacji przez nie tworzonych wybrałam z tekstów Korpusu Języka Polskiego PWN oraz słowników języka polskiego: „Słownika Języka Polskiego” pod red. W. Doroszewskiego (SJPDor) i „Innego Słownika Języka Polskiego” pod red. M. Bańki (ISJP) oraz słownika analizatora fleksyjnego AMOR (Rabiega-Wiśniewska 2004). Dane, które zebrałam, pozwoliły mi oszacować liczbowo objęty opisem fragment derywacji w języku polskim.

Literatura słowotwórcza stała się dla mnie źródłem informacji w dwóch aspektach. Po pierwsze, wyekscepowalam z niej formalne wykładniki derywacji, z których następnie wybrałam według opracowanych kryteriów podzbiór będący podstawą reguł derywacyjnych. Po drugie, doświadczenia wcześniejszych badaczy naprowadziły mnie na formalne rozwiązania podstawowych dla opisu derywacji problemów. Ze względu na cel pracy, którym jest zbudowanie modelu automatycznej derywacji, ścisłego oszacowania wymagała również produktywność procesów derywacyjnych, związany z nią stosunek derywacji do fleksji oraz morfonologia.

Formalny opis derywacji na potrzeby automatycznej analizy wykonywanej przez program komputerowy stawia przed badaczem zadanie ścisłego określenia podstaw opisu.

Problem definicji formantów w ujęciu informatycznym poruszył między innymi W. Hoepfner (1980). Autor przedstawił warunki, które musi spełniać formalna definicja sufiksu, oraz na nich oparł klasyfikację niemieckich. Formalna klasyfikacja formantów sufiksalnych musi wychodzić z poziomu tekstu – sufiks jest wprawdzie traktowany jako ciąg liter, które na poziomie słownikowym stają się grafemami. Grafemicznie zdefiniowane sufiksy mogą mieć warianty, które reprezentują sufiks w formach-derywatach. Należy ująć warianty w modelu, opisując warunki ich występowania.

Przygotowanie listy sufiksów dla języka niemieckiego Hoepfner rozpoczął od zebrania wszystkich grafemicznych zakończeń form-derywatów, opierając się przy tym na literaturze przedmiotu oraz kilku założeniach formalnych. Lista kandydatów na sufiksy zawierała 73

jednostki. Opracowany następnie przez badacza zbiór reguł na różnych poziomach opisu lingwistycznego narzucił na grafemiczną definicję sufiksu ograniczenia (np. przypisanie rodzaju gramatycznego, klasy semantycznej, przypisanie bazy klasy gramatycznej, przypisanie formantu do klasy kombinacji sufiksów).

Przyjęłam założenia zaproponowane przez Hoepfnera, modyfikując je i rozszerzając tak, aby korzystać w pełni z fleksji języka polskiego oraz maksymalnie zwiększyć wydajność reguł derywacji podczas analizy. Warunki formalne wyboru formantów przedstawiam w punktach w formie dyskusji z założeniami niemieckiego badacza.

- (1) Liczba grafemicznie zdefiniowanych formantów powinna być nieduża, aby ich automatyczne rozpoznanie było wydajne. Podstawą wyboru formantów jest liczba wystąpień derywatów formowanych przez nie zarówno w słownikach języka polskiego, jak i w korpusie tekstów. Włączając wszystkie formanty o wysokiej frekwencji w tekstach, zwiększa się wydajność automatycznej analizy dużych tekstów.
- (2) Podstawą badań jest współczesny język polski; uwzględniam sufiksy rzeczownikowe rodzime i obce, sufiksy przymiotnikowe, pseudo-sufiksy rzeczownikowe, pseudo-prefiksy rzeczownikowe i przymiotnikowe, oraz konwersję rzeczownikową i przymiotnikową (trzema ostatnimi typami derywacji Hoepfner się nie zajmuje).
- (3) W opisie zostają pominięte: (a) sufiksy występujące wyłącznie w słownictwie specjalistycznym (np. -AN [AZOTAN]); (b) nieproduktywne, podlegające analizie, ale nielicznie reprezentowane sufiksy (np. -WO [SPOIWO]).
- (4) Warianty sufiksów nie powinny mieć osobnych reprezentantów w opisie. Brane są pod uwagę tylko warianty o liczącej się frekwencji w formacjach.
- (5) Sufiksy formujące derywaty od nazw własnych zostają w ogóle pominięte w opisie. Frekwencja derywatów od nazw własnych nie jest w słownikach wysoka (patrz punkt (1)), do rzetelnego opisu formantów należałoby najpierw zebrać odpowiedni materiał empiryczny.
- (6) Pseudo-sufiksy i pseudo-prefiksy zostają włączone do opisu, ponieważ ich frekwencja w tekstach polskich jest wysoka i należy podjąć próbę automatycznego rozpoznawania tego typu formacji w tekstach (choćby stawiania hipotez co do ich charakterystyki fleksyjnej, gdy analiza derywacyjna się nie powiedzie).

Każdemu z wyodrębnionych na podstawie powyższych założeń formantów zostają w modelu przypisane określone własności (prefiksacja i konwersja nie muszą mieć wszystkich).

- (1) Formantowi przypisana zostaje klasa gramatyczna, pełna charakterystyka fleksyjna oraz grupa fleksyjna w rozumieniu Tokarskiego (1973).
- (2) Każdy formant dołącza się do baz określonej klasy gramatycznej.
- (3) Formant dołącza się do określonej formy-bazy.
- (4) Formant może być reprezentowany przez warianty. Każdy wariant formantu ma opisaną dystrybucję.
- (5) Formalny opis formantu zawiera reguły ograniczające dystrybucję.
- (6) Formalny opis formantu zawiera opis alternacji powodowanych przez niego podczas derywacji.
- (7) Formalny opis formantu zawiera opis możliwości dalszego jego udziału w procesie derywacji (tworzenia formantów złożonych).

Na podstawie wymienionych założeń wybrałam z przygotowanych list formantów (sufiksów, prefiksów, pseudo-sufiksów, pseudo-prefiksów i konwerterów) spełniające wymagania jednostki i ich warianty.

Automatyczna analiza tekstu wymaga również słownika o szczególnej konstrukcji. Na potrzeby automatycznej analizy fleksyjnej (oraz składniowej) tekstu polskiego powstało dotychczas kilka narzędzi wykorzystujących wbudowane w nie słowniki (Hajnicz i Kupść 2001). Ze względu na funkcję, którą pełnią, słowniki te zawierają jednostki zwykle opisane w odmienny

sposób od tradycyjnego, m. in. zarzucono w nich morfologiczny podział na rdzeń i końcówki fleksyjne (Rabiega-Wiśniewska i Rudolf 2003). Taka konstrukcja leksykonu ogranicza możliwości jego wykorzystania w zasadzie tylko do rozpoznawania charakterystyki fleksyjnej badanych form wyrazowych. Dobry słownik elektroniczny powinien być jednak niezależny od sposobu jego wykorzystania. Aby tak się stało, należy wrócić do opisu jednostki słownikowej bliskiego morfologii.

W dysertacji przedstawiam koncepcję słownika opartego na rdzeniach. Słownik o takiej konstrukcji może być wykorzystany zarówno w automatycznej analizie fleksyjnej, jak również słowotwórczej. Opis jednostki tego leksykonu pozwala na wydzielenie rdzenia wraz z alternacjami tematowymi oraz oddzielnie końcówek fleksyjnych. Dzięki temu w formalnym modelu derywacji można było zaproponować reguły derywacyjne o dużym stopniu szczegółowości (Rabiega-Wiśniewska 2005).

Projekt jednostki słownikowej w leksykonie rdzeni wykorzystuje elementy teorii dwupoziomowości morfologii K. Koskenniemi (1983). Głównym założeniem opisu hasła słownikowego jest oddzielenie od siebie tematów fleksyjnych danego leksemu oraz wszystkich końcówek fleksyjnych właściwych formom wyrazowym tego leksemu w taki sposób, aby możliwe było ich syntezowanie i analizowanie. Hasło słownikowe, podobnie jak we wspomnianej pracy, jest reprezentowane na dwóch poziomach, powierzchniowym i graficznym. Reprezentanta wszystkich tematów fleksyjnych danego leksemu nazywam rdzeniem. Rdzeń wraz ze zbiorem końcówek i ewentualnie uporządkowanym zbiorem alternacji wewnątrztematowych tworzy poziom graficzny słownika. Poziom powierzchniowy reprezentowany jest bezpośrednio przez etykietę leksemu oraz pośrednio przez każdą formę wyrazową, do której ma dostęp program komputerowy wykonujący proste instrukcje na poziomie graficznym. Oto przykład hasła słownikowego *wariat*, na które składa się forma hasłowa, rdzeń wraz z kodem rodzaju, zbiorem alternacji oraz zbiorem końcówek.

```
wariat  
wariaT r:l  
{T:t(N,G,D,A,B,g,d,a,b,l),ci(L,V),c(n,v)}  
k(a,owi,a,em, e,e,i,ów,om,ów,ami,ach,i)
```

Elektroniczny słownik rdzeni pokazuje nowy sposób opisu jednostek leksykalnych, uwzględniający zarówno ich fleksję, jak i alternacje występujące wewnątrz tematów fleksyjnych. Opis ten może nieznacznie odbiegać od tradycyjnego opisu tematów fleksyjnych, nie opiera się bowiem na fonetycznych właściwościach morfemów. Powodem pozostania w opisie na poziomie tekstu jest cel, któremu ma służyć słownik. Jego zadaniem jest gramatyczna identyfikacja słowa w tekście na podstawie jego kształtu i dlatego jedynym praktycznym uogólnieniem dla pewnych grup liter są grafemy. Ponieważ opis hasła jest niezależny od sposobu zapisu danych, słownik rdzeni może być podstawą różnych narzędzi analizy tekstowej.

W pracy formalny opis wykładników derywacji i konwersji ilustruję przykładami haseł ze słownika rdzeni. Opracowanie tego słownika umożliwiło mi opracowanie reguł alternacyjnych dla sufiksów i konwerterów. Słownik rdzeni został przeze mnie wykorzystany tylko jako pomoc empiryczna przy tworzeniu modelu słowotwórstwa. Opracowałam jednak ten słownik przede wszystkim z myślą o jego implementacji.

3. Budowa modelu

Materiał empiryczny został ułożony według malejącego stopnia komplikacji opisu formantów. Przyjmuję, że najbardziej złożoną budowę mają sufiksy i pseudo-sufiksy, następnie konwertery, a prefiksy i pseudo-prefiksy zamykają opis.

Formalny opis formantu otwiera tabela, w której podana jest krótka charakterystyka derywatu formowanego przez dany wykładnik:

- (1) klasa gramatyczna derywatu i wartości kategorii fleksyjnych formy hasłowej derywatu oraz informacja o jego przynależności do grupy deklinacyjnej (Tokarski 1973),
- (2) klasa gramatyczna bazy,
- (3) informacja o obecności reguł alternacyjnych w opisie sufiksu,
- (4) informacja o zdolności derywatu do podlegania dalszym procesom derywacyjnym,
- (5) lista wariantów.

W dalszej części opisu szczegółowo prezentuję formowanie derywatów przez dany formant. Określam formalne wymagania wykładnika względem form-baz, takie jak:

- (1) rodzaj gramatyczny,
- (2) grupa deklinacyjna bądź koniugacyjna,
- (3) zakończenie formy-bazy,
- (4) typy alternacji w temacie fleksyjnym formy-bazy.

Często poszerzam opis derywatu o informację semantyczną. Opis kończę podając przykład fragmentu hasła ze słownika rdzeni, który ilustruje wybór odpowiedniej formy-bazy (lub form-baz) do reguł derywacji.

Większość przedstawionych wyżej informacji o danym wykładniku ujmuję następnie w reguły derywacyjne. Reguła derywacyjna to schemat, w którym temat fleksyjny wybranej form-bazy łączy się z sufiksem w procesie derywacyjnym, wynikiem jest forma-derywat. Każdą regułę otwiera oznaczenie, na przykład, sufiksu, jednoznacznie przyporządkowując przekształcenie derywacyjne tylko temu wykładnikowi. Każdej regule sufiksacji towarzyszy przykład derywatu, który można zanalizować podaną regułą. Reguły obejmują sufiksację rodzimą, obcą oraz pseudo-sufiksy. Reguły każdej podgrupy zostały ponumerowane w sposób ciągły (przyjmując notację: Sr, So, Sp). Przykład jednej z reguł dla sufiksu $-K(a)_1$:

Sr 2. $-K(a)_1 : FN.gen.plu.fem + -K(a)_1 \rightarrow FN.nom.sing.fem$ (KÓZKA)

Reguły alternacyjne uzupełniają reguły derywacji, są przyporządkowane sufiksom przez oznaczenie na początku reguły (przykład poniżej). Reguły mogą przedstawiać zmianę graficzną tematu fleksyjnego formy-bazy bądź też zawiadywać dystrybucją wariantów sufiksu. Przy regułach alternacyjnych podaję przykłady przekształceń form wyrazowych spełniających podane w nich warunki. Reguły alternacyjne zostały ponumerowane w ramach podgrup sufiksacji rodzimej i obcej (przyjmując notację: A-Sr, A-So, A-Sp).

A-Sr 27. $-IK(\emptyset)_1F$: jeżeli $FN.loc.sing.masc$ jest zakończona na 'c' \rightarrow zamień na 'cz' (*chłopc-* na *chłopc-*)

Informacje zbiorcze o liczbie i rodzaju wykładników derywacji oraz liczbie reguł przedstawia tabela:

Wykładnik derywacji	Liczba jednostek	Liczba reguł derywacyjnych	Liczba reguł alternacyjnych
Sufiksy rzeczownikowe rodzime	35	89	116
Sufiksy rzeczownikowe obce	13	31	61
Sufiksy przymiotnikowe	17	32	41
Pseudo-sufiksy rzeczownikowe	10	10	-
Konwertery rzeczownikowe	10	10	9
Konwertery przymiotnikowe.	4	11	9
Prefiksy rzeczownikowe i przymiotnikowe	23	70	2
Pseudo-prefiksy rzeczownikowe i przymiotnikowe	35	139	-
Razem	147	392	238

Po scharakteryzowaniu formantu prezentuję dane ilościowe: udział derywatów w słownikach SJPDor i ISJP oraz frekwencję form-derywatów w korpusie tekstów.

Oglądem objęłam liczne derywaty i chociaż zaproponowane przeze mnie reguły nie obejmują zebranego przeze mnie zbioru w całości, wykazałam, że polskie słowotwórstwo można włączyć w system automatycznej analizy tekstu. Informacje ilościowe o zbiorze derywatów zebranych przy opracowywaniu formalnego modelu derywacji przedstawia tabela:

Wykładnik derywacji	Liczba derywatów w słownikach	Liczba form-derywatów w korpusie
Sufiksy rzeczownikowe rodzime	22950	17833
Sufiksy rzeczownikowe obce	6161	6494
Sufiksy przymiotnikowe	16005	39889
Pseudo-sufiksy rzeczownikowe	1520	894
Konwertery rzeczownikowe	13896	13113
Konwertery przymiotnikowe.	1354	3780
Prefiksy rzeczownikowe i przymiotnikowe	3591	5241
Pseudo-prefiksy rzeczownikowe i przymiotnikowe	2338	2227
Razem	67815	89471

Wszystkie wyszukane w słownikach derywaty poddaje następnie automatycznej analizie fleksyjnej, wykorzystując program AMOR. Nie rozpoznane przez analizator słowa są podstawą dalszych badań. Z pomocą reguł derywacyjnych (i alternacyjnych) analizuję derywaty i stawiam hipotezę co do kształtu form-baz. W następnym kroku formy-bazy są ponownie analizowane przez program AMOR, w wyniku czego pozostają, być może, pewne nie zinterpretowane słowa. Stosunek procentowy nie rozpoznanych słów w pierwszym kroku analizy oraz słów nie rozpoznanych w ogóle do wszystkich derywatów z listy słownika jest podstawą wniosków o ewentualnej przydatności reguł derywacji w procesie automatycznej analizy tekstu polskiego. Skuteczność analizy z wykorzystaniem przygotowanych reguł derywacji przedstawiają wykresy dołączone do pracy.

Opis formantu zamyka informacja o udziale formowanych przez niego derywatów w innych procesach derywacyjnych. Podaję przykłady wykładników, które mogą brać udział w dalszych krokach derywacyjnych. Zebranie i przetestowanie wszystkich możliwych zależności między wykładnikami derywacji możliwe będzie dopiero po implementacji systemu analizy derywacyjnej.

Przedstawiona analiza nie daje z założenia opisów semantyczno-słowotwórczych. Analiza ta służy przede wszystkim postawieniu hipotezy bazy (dokładniej – hipotez). Nie rozstrzygam, która hipoteza jest semantycznie poprawna, ani czy w ogóle baza została odpowiednio słowotwórczo dobrana, jeżeli w wyniku analizy derywat otrzymuje odpowiednią charakterystykę fleksyjną, np. DEGENERACYJNY ← DEGENERACJA | GENERACYJNY.

Poniżej prezentuję formalny opis sufiksu rzeczownikowego $-C(a)$.

33. $-C(a)$

ogólna charakterystyka sufiksu	wartość
cechy fleksyjne (grupa fleksyjna)	N, Nom, sing, masc (mz)
klasa gramatyczna bazy	V
obecność alternacji	tak
udział w procesach derywacyjnych	tak
wariant	$-ca$

OPIS

Sufiks $-C(a)$ formuje derywaty odczasownikowe.

UWAGI FORMALNE

Sufiks $-C(a)$ formuje derywaty od tematów fleksyjnych form-baz czasowników należących w większości do czterech grup koniugacyjnych: cI (*pochleb-*), cVI (*chwal-*), cVIII (*wychowyw-*) i cIX

(*nadaw-*) oraz okazjonalnie należących do grupy cIV (*naśladow-*). W procesie derywacji biorą też udział czasowniki nieregularne (*wynalaz-*).

Formant *-C(a)* wiąże najwięcej form-baz z grup cI i cVIa, wywołując niewiele zmian alternacyjnych na granicy sufiksu i tematu fleksyjnego formy-bazy. Wysoką frekwencję tekstową mają również nieliczne derywaty z tym sufiksem związane formalnie z grupami cIV, cVIII i cIX. Ich budowa nie jest tak regularna, jak wcześniejszych. Ponieważ są one również bazami dla wielu złożzeń, umieszczam je w słowniku rdzeni (np. WYCHOWAWCA).

REGUŁA DERYWACYJNA

Sr 1. *-C(a)* : FV.imp.praes.3.sing + *-C(a)* → FN.nom.sing.mz (POCHLEBCA)

Sr 2. *-C(a)* : FV.perf.praes.3.sing + *-C(a)* → FN.nom.sing.mz (ZWYCIĘZCA)

REGUŁA ALTERNACYJNA

A-Sr 1. *-C(a)F* : jeżeli FV.praes.3.sing jest zakończona na 'i' → usuń 'i' (*pochlebi-* na *pochleb-*)

A-Sr 2. *-C(a)F* : jeżeli FV.praes.3.sing jest zakończona na 'dz' → zamień na 'd' (*doradz-* na *dorad-*)

A-Sr 3. *-C(a)F* : jeżeli FV.praes.3.sing jest zakończona na 'odz' → zamień na 'ód' (*dowodz-* na *dowód-*)

A-Sr 4. *-C(a)F* : jeżeli FV.praes.3.sing jest zakończona na 'z' → zamień na 'z' (*zwycięz-* na *zwycięz-*)

A-Sr 5. *-C(a)F* : jeżeli FV.praes.3.sing jest zakończona na 's' → zamień na 'ś' (*roznoś-* na *roznoś-*)

A-Sr 6. *-C(a)F* : jeżeli FV.praes.3.sing jest zakończona na 'dz' → zamień na 'dź' (*wychodz-* na *wychodź-*)

TESTY

Derywaty z sufiksem *-C(a)* mają średni udział w obu słownikach. Słownik ISJP okazał się za mały do testów. Automatyczna analiza fleksyjna listy derywatów z SJPDor zawiodła w 51% analiz. Dzięki regułom derywacyjnym i alternacyjnym udało się zmniejszyć ten odsetek do 12%. Mimo że derywacja z sufiksem *-C(a)* podlega wielu wyjątkom i alternacjom, reguły podnoszą skuteczność analizy morfologicznej. Mały Korpus PWN dostarczył wyższej liczby form-derywatów niż słowniki, wzrasta tym samym prawdopodobieństwo znalezienia nowych derywatów. Reguły derywacyjne powinny w tym pomóc.

Częstość słownikowa derywatów <i>-C(a)</i>		
SJPDor		194
ISJP		89
Częstość tekstowa form-derywatów <i>-C(a)</i>		
Mały Korpus PWN		424
Analizator AMOR		
Liczba derywatów nie rozpoznanych		
	SJPDor (procent wszystkich)	99 (51%)
	ISJP (procent wszystkich)	3 (3%)
Liczba podstaw nie rozpoznanych (po zastosowaniu reguły)		
	SJPDor (procent wszystkich)	23 (12%)
	ISJP (procent wszystkich)	1 (1%)

ŁĄCZLIWOŚĆ

Sufiks *-C(a)* bierze udział w dalszych procesach derywacyjnych. Derywaty z tym sufiksem podlegają sufiksacji (np. z sufiksem *-IN(i)*) oraz wymianie (np. z sufiksem *-STW(o)₁*).

4. Zastosowania i perspektywy

Korzyści płynących z formalnego opisu zjawisk słowotwórczych jest co najmniej kilka – zarówno praktycznych, jak i teoretycznych.

Pierwszą i zasadniczą jest nowa koncepcja słownika gramatycznego, wywodzącego się z analizatora fleksyjnego AMOR. Nowy, mniejszy słownik rdzeni zawiera sieć zależności między leksemami, co pozwala użytkownikowi wyszukiwać jednostki tekstu ze względu na zawierane przez nie morfemy, np. BALON: *balonik, balonowy, baloniasty, ...*. Wykorzystanie

reguł derywacyjnych pozwala również na analizę jednostek nieznanymi słownikowi. Można stawiać hipotezy, z jakich komponentów składa się słowo i jaka jest prawdopodobna jego charakterystyka fleksyjna, np. *literackość*. Automatyczna analiza derywacyjna służy wtedy przynajmniej lematyzacji (oznaczaniu tekstu) jako dodatkowy moduł systemu przetwarzania tekstu. Gdy analiza nowej jednostki się powiedzie, system podaje pełny opis derywatu. W szczególności zastosowanie to można wykorzystać przy przeszukiwaniu dużego korpusu tekstu, ponieważ podając na wejściu leksem podstawowy, można spodziewać się w wyniku zbioru różnych leksemów powiązanych z zadaniem ustalonymi regułami derywacyjnymi, np. ZAB: *ząbek, ząbkować, ząbkowanie, ząbkujący, ...*. Na koniec, formalne opracowanie derywacji może zostać wykorzystane przy budowie słowników nie tylko elektronicznych, ale i specjalistycznych, np. słowników terminologii.

Zastosowaniem automatycznej derywacji szczególnie interesującym dla językoznawcy jest testowanie poprawności neologizmów słowotwórczych. Dzięki oszacowaniu liczby derywatów podlegających danej regule derywacyjnej można sprawdzić, czy dany okazjonalizm (lub neologizm) ma szansę na wejście do słownika ogólnego i czy wpisuje się w przyjęty model słowotwórstwa. Również badanie produktywności wykładników derywacji może pokazać dynamikę ich łączliwości z różnymi klasami gramatycznymi oraz z innymi wykładnikami derywacji. Konstrukcja słownika rdzeni pozwala natomiast na całościowy ogląd alternacji fleksyjnych i słowotwórczych w języku polskim, dając badaczom bezpośredni dostęp do danych.

Formalny opis derywacji obejmuje klasę rzeczowników i przymiotników. Wykorzystanie modelu derywacji w pełni będzie możliwe, gdy opis zostanie uzupełniony o czasowniki i złożenia. W najbliższym czasie możliwe jest skonstruowanie analizatora morfologicznego (obejmującego fleksję i derywację), który przetestowałby działanie reguł derywacyjnych dla opisanych klas. W przyszłości analizator należy rozbudować, dołączając wykładniki derywacji czasownikowej i złożzeń. Przedstawiony przeze mnie słownik rdzeni natomiast może wejść do analizatora fleksyjnego, który podawałby nie tylko charakterystykę fleksyjną badanej formy wyrazowej, ale również jej podział na temat fleksyjny i końcówkę. Podział ten można też wykorzystać do automatycznego wyszukiwania zadanych tematów fleksyjnych, reprezentantów grup fleksyjnych i typów alternacji fleksyjnych.

Prezentowany formalny opis derywacji przygotowałam z myślą o jego zastosowaniu w narzędziach automatycznej analizy tekstu. Mam nadzieję, że opis ten ukazuje system słowotwórstwa polskiego w nowej perspektywie.

Warszawa, dnia 12 listopada 2006