INVESTIGATIONES LINGUISTICAE, VOL. XIV; POZNAŃ, DECEMBER 2006

Model skończenie stanowy niemieckich wyrazów pojedynczo i wielokrotnie złożonych

A finite-state model of German compounds

Marcin Junczys-Dowmunt

Institute of Linguistics, Adam Mickiewicz University
ul. Międzychodzka 5, 60-371 Poznań, POLAND

junczys@amu.edu.pl

Abstract

This paper summarizes the results of my Master's thesis and the main points of a talk I presented at the seminar of the Department of Applied Logic at Adam Mickiewicz University in Poznań. It gives a short overview of the structure of German compounds and newer research concerning the role of the so-called interfixes. After an introduction to the concept of finite-state transducers the construction of a transducer used for naive compound segmentation is described. Tag-based finite-state methods for the further analysis of the found segments are given and discussed. Distributional transducer rules, for the construction of which I assume the existence of local and global morphological contexts, are proposed as means of disambiguation of the analyzed naive segmentation results.

1 Wstęp

Zjawisko kompozycji występuje w wielu językach, które różnią się jednak znacznie stopniem produktywności tego procesu słowotwórczego. W języku polskim złożenia mające więcej niż dwa tematy słowotwórcze są rzadko spotykane poza tekstami technicznymi lub poetyckimi (por. Grzegorczykowa et al. 1999: 457). Głównym środkiem tworzenia nowych polskich rzeczowników jest derywacja. Natomiast w języku niemieckim kompozycja, charakteryzująca się teoretycznie zarówno nieograniczoną produktywnością, jak i nieograniczoną złożonością, zajmuje pierwsze miejsce wśród procesów słowotwórczych niemieckich rzeczowników (por. Eichinger 2000: 71).

Nie można leksykograficznie zestawić wszystkich lub nawet większości niemieckich złożeń, co w życiu codziennym może spowodować duże prawdopodobieństwo spotkania się z wyrazami niefigurującymi w żadnym słowniku. W większości przypadków interpretacja takich okazjonalizmów nie stwarza problemów dla rodzimego użytkownika języka niemieckiego, który ma pewną wiedzę dotyczącą znaczenia poszczególnych członów złożenia. Pamięta także znaczenia innych podobnych złożeń i potrafi interpretować nowe złożenia według znanych mu wzorców kompozycyjnych. Jeżeli to nie wystarczy, korzysta z kontekstu zdaniowego i wiedzy ogólnej, by pozbyć się ostatnich wątpliwości.

Dla komputerowego przetwarzania języka niemieckiego złożenia te stanowią poważny problem. Żaden korpus nie byłby w stanie zawrzeć wszystkich potencjalnie istniejących złożeń, ponieważ każde złożenie zawarte w korpusie może zaistnieć w kolejnym złożeniu jako jeden z jego członów.

Takie rekurencyjne pojęcie złożenia pozwala na tworzenie dowolnie długich struktur kompozycyjnych. Do tego dochodzą nieograniczone możliwości kombinacji członów. W odpowiednim kontekście każde dwa rzeczowniki mogą brać udział w procesie kompozycji, przy czym kompozycja nie jest ograniczona do samych rzeczowników; każda część mowy może zaistnieć jako człon złożenia – bezpośrednio lub po nominalizacji.

Kolejną cechą utrudniającą analizę niemieckich złożeń jest ich łączna pisownia, stosowana mimo potencjalnie wielokrotnej złożoności. Wyodrębnienie poszczególnych członów złożenia wymaga przeprowadzenia segmentacji i ustalenia granic międzymorfemowych. Z drugiej strony na poziomie tekstu łączna pisownia pomaga odróżnić złożenia od fraz, co np. w języku angielskim, gdzie występują złożenia pisane rozdzielnie, jest zadaniem niełatwym.

Ograniczamy się w tej pracy do subkodu graficznego. Dla niemieckich złożeń oznacza to, że nie dysponujemy informacjami o właściwościach prozodycznych wyrazu, w szczególności o akcencie. Akcent ma duży wpływ na poprawne zrozumienie usłyszanych złożeń, pomaga bowiem w lokalizacji granicy między głównymi członami złożenia.

Przeprowadzamy analizę złożeń danych w postaci ciągów znakowych. Automaty i transduktory skończone są szczególnie wydajnymi formalizmami przeznaczonymi do przetwarzania takich ciągów.

W niniejszej pracy opisujemy konstrukcję transduktora wykorzystywanego do naiwnej segmentacji złożeń. Naiwną segmentacją nazywamy segmentację, która nie sprawdza poprawności gramatycznej znalezionych rozkładów. W dalszej części artykułu wprowadzamy metody analizy poszczególnych segmentów oparte na tzw. tagach. Zakładamy istnienie lokalnych i globalnych kontekstów morfologicznych, zachodzących wewnątrz i pomiędzy członami złożeń. Konteksty te wykorzystujemy do konstrukcji reguł dystrybucyjnych, które służą do eliminacji wieloznaczności wyników naiwnej segmentacji.

2 Struktura złożeń

W skład złożenia wchodzą przynajmniej dwa wyrazy. W większości przypadków łączą się one bezpośrednio w jeden wyraz bez pomocy dodatkowych elementów. Jednak w mniej więcej jednej trzeciej (por. Dudenredaktion 1998: 497) złożonych rzeczowników występują tzw. interfiksy, które również trzeba uwzględnić w trakcie segmentacji wyrazów złożonych. W literaturze polskiej interfiksy traktuje się jako ciągi fonologiczne niezaliczane do żadnego z głównych składników (por. np. Grzegorczykowa et al. (1999: 366), Polański (1999: 259)), pełniące różne funkcje strukturalne, zapobiegające pojawieniu się na granicy morfemów nieakceptowalnych ciągów fonologicznych. W nowszych źródłach niemieckich, wraz z wprowadzeniem pojęcia formy kompozycyjnej, poglądy te uległy pewnej zmianie, co omówimy niżej.

Z diachronicznego punktu widzenia owe interfiksy są pozostałościami morfemów fleksyjnych. W czasach historycznych powstawały zrosty, które początkowo zachowywały swoje końcówki fleksyjne i rodzaj fleksji wewnętrznej. Takie zrosty z czasem utraciły fleksję wewnętrzną i zaczęły służyć jako wzorce kompozycyjne dla złożeń właściwych, których nie można było już nazywać zrostami. Mimo to, iż te złożenia nie powstały ze zrostów, również miały interfiksy. Synchronicznie takie historyczne struktury oraz struktury tworzone analogicznie są traktowane jak złożenia, ponieważ nie da się ich już odróżnić od nowoczesnych struktur złożonych. Diachroniczne pokrewieństwo z morfemami fleksyjnymi oraz morfotaktyczne właściwości interfiksów, które w skrócie opiszemy w piątej części, pozwalają na opracowanie reguł dystrybucyjnych pomocnych przy eliminacji wieloznaczności segmentacji złożeń.

Każde złożenie ma strukturę binarną (por. Fleischer i Barz 1995: 93). Człon prawy (drugi) jest elementem głównym, którego wszystkie kategorie morfologiczne dziedziczy cała struktura. Jeżeli element główny jest rzeczownikiem o określonym rodzaju gramatycznym i określonej paradygmatyce fleksyjnej, to całe złożenie także będzie rzeczownikiem o tym samym rodzaju gramatycznym,

należącym do tej samej klasy fleksyjnej. W przypadku gdy element główny również jest złożeniem, to jego element główny określa rekurencyjnie kategorie morfologiczne nadrzędnego złożenia itd.

Wynika z tego, że w przypadku wielokrotnych złożeń wystarczy wyodrębnić najbardziej prawy człon niezłożony i ustalić jego kategorie morfologiczne, by ustalić kategorie morfologiczne całego złożenia. Hierarchiczna struktura binarna złożenia nie jest tutaj istotna.

W przypadku złożeń nadrzędno-podrzędnych ze względu na wzajemny stosunek członów element główny określa też klasę semantyczną całego złożenia. Złożenia współrzędne lub nadrzędno-podrzędne egzocentryczne, w których żaden z członów nie określa klasy semantycznej złożenia, są rzadkie i najczęściej zleksykalizowane.

Lewy (pierwszy) człon nie ma wpływu na kategorie morfologiczne złożenia. Podobnie jak element główny pierwszy człon może być prosty lub złożony pod względem słowotwórczym.

Do niedawna zakładano, że w kompozycji biorą udział formy podstawowe obu członów, i interfiksów nie wliczano *explicite* do pierwszego członu. W nowszej literaturze, np. u Fuhrhop (1998) i Langer (1998), zakłada się istnienie form kompozycyjnych[1] w paradygmatyce rzeczowników. Mianem formy kompozycyjnej określa się formę pierwszego członu biorącego udział w kompozycji. Zakłada się, że każdy rzeczownik ma co najmniej jedną taką formę. W jej skład oprócz formy podstawowej rzeczownika wchodzi również interfiks – o ile jest wymagany. W przypadku kompozycji bez interfiksu forma kompozycyjna rzeczownika jest równa formie podstawowej. Przyjęcie takiej formy ułatwia również interpretację złożeń, w których pierwszy człon jest czasownikiem. Jeśli chodzi o czasowniki, tylko ich rdzeń bierze udział w kompozycji, a nie bezokolicznik. Nazwanie takiej formy czasownika formą kompozycyjną rozwiązuje niektóre nieścisłości terminologiczne związane z tym, iż rdzeń czasownika nie jest jego formą podstawową. W przypadku innych części mowy forma podstawowa wchodzi w skład formy kompozycyjnej.

Argumentacja Fuhrhop (1998: 187), że interfiks należy zaliczać do formy kompozycyjnej, opiera się na trzech właściwościach interfiksów:

1. wybór interfiksu jest wyznaczany przez pierwszy człon;
2. postać interfiksu jest związana z systemem fleksyjnym pierwszego członu;
3. w przypadku koordynacji interfiks pozostaje przy pierwszym członie.

Każdy rzeczownik może mieć kilka form kompozycyjnych. Wynika to z możliwości łączenia się niektórych rzeczowników z różnymi interfiksami. Wybór istniejących form kompozycyjnych pierwszego członu zależy wtedy głównie od semantycznych właściwości drugiego członu.

Obok zleksykalizowanych złożeń istnieją zleksykalizowane formy kompozycyjne, które nie odpowiadają ogólnym regułom tworzenia form kompozycyjnych. Takie formy mogą być wykorzystane w tworzeniu nowych, niezleksykalizowanych złożeń.

3 Segmentacja złożeń

Wspominaliśmy, że niemieckie złożenia są pisane łącznie – niezależnie od liczby składników. Przeprowadzenie segmentacji wymaga ustalenia rodzaju segmentu, czyli jednostki segmentacji. Przy analizach słowotwórczych wyodrębnia się morfemy podstawowe i afiksalne. Elementy, których nie da się przyporządkować do żadnej z grup, nazywa się interfiksami. W przypadku złożeń, a w szczególności złożeń wielokrotnych, uważamy taką analizę za zbyt wnikliwą. Często już sama interpretacja złożenia na podstawie jego członów nie jest jasna, a interpretacja na podstawie jego morfemów byłaby jeszcze trudniejsza.

Człony złożeń mogą być simpleksami, derywatami lub kolejnymi złożeniami, lub mówiąc krócej, wyrazami[2]. Przyjmując granicę wyrazu jako granicę segmentu, człon złożenia, który sam nie jest

[1] Nasza propozycja na tłumaczenie pojęcia „Kompositionsstammform".

[2] Przyjmujemy w tym miejscu jeszcze intuicyjne, nieostre pojęcie wyrazu.

Tag	Znaczenie	Tag	Znaczenie
+N	rzeczownik	+V	czasownik
+A	przymiotnik	+I	interfiks
+MS	rodzaj męski	+FM	rodzaj żeński
+NT	rodzaj nijaki		
+S1	liczba pojedyncza na *-[e]s*	+S2	liczba pojedyncza na *-[e]n*
+S3	liczba pojedyncza na *-∅*	–S	*plurale tantum*
+P1	liczba mnoga na *-[e]*	+P2	liczba mnoga na *-∅*
+P3	liczba mnoga na *-[e]n*	+P4	liczba mnoga na *-er*
+P5	liczba mnoga na *-s*	–P	*singulare tantum*
+M	wielosylabowy	–M	jednosylabowy
+SS	sufiks wymagające *-s*	+AT	wygłos zawierający *-t*
+SN	sufiks bez interfiksu	+AE	wygłos zawierający *-e*
–SA	brak sufiksu/wygłosu		
#	koniec tagsetu		

Rysunek 1: Spis wykorzystanych tagów i ich znaczenie

złożeniem, jest dalej niepodzielny. Jeżeli człon jest natomiast kolejnym złożeniem, zostaną zidentyfikowane kolejne dwa wyrazy lub segmenty składowe.

Istnieją złożenia, których idiomatyzacja jest na tyle zaawansowana, że pojawia się potrzeba włączenia ich jako całości do słownika języka niemieckiego. Leksykalizacji podlegają również złożenia często używane, które tym samym przestają być okazjonalizmami. Wynika z tego, że pojęcie segmentu należy rozszerzyć na wyrazy zleksykalizowane, czyli leksemy. Z punktu widzenia analizy komputerowej złożeń wyraz jest zleksykalizowany, jeśli znajduje się w słowniku aplikacji. I odwrotnie: wyraz niezawarty w słowniku jest niezleksykalizowany i podlega dalszej segmentacji. W przypadku gdy wyraz jest niezleksykalizowany i proces segmentacji nie jest w stanie znaleźć odpowiednich segmentów w słowniku, segmentacja się nie powiedzie. Należy wtedy uzupełnić informacje zawarte w słowniku.

Na podstawie segmentacji można ustalić względnie prostą definicję okazjonalizmów. Jeśli wyraz nie jest zleksykalizowany, ale można przeprowadzić segmentację na leksemy składowe na podstawie słownika, to wyraz segmentowany jest okazjonalizmem.

Mówiliśmy o interfiksach jako elementach nienależących do żadnej szczególnej grupy morfemów. Trzeba je jednak uwzględnić w procesie segmentacji, ponieważ podobnie jak pozostałe segmenty, istnieją w postaci ciągów fonologicznych lub grafematycznych. Odróżnienie ich od segmentów o charakterze leksemowym odbywa się dopiero poprzez kategoryzację morfologiczną znalezionych segmentów.

Na potrzeby implementacji umieszczamy interfiksy w słowniku, ignorując tym samym lingwistyczną poprawność takiego rozwiązania. Segmentacja oraz odróżnienie interfiksów od leksemów odbywa się za pomocą słownika.

Zadaniem segmentacji nie jest tylko ustalenie granic między segmentami. Należy też przyporządkować znalezionym segmentom informacje zawarte w słowniku. W literaturze angielskiej i niemieckiej w przypadku, gdy informacje w słowniku ograniczają się do cech gramatycznych, taki proces nazywa się „tagging” (por. Glück 2000: 720). Określenie pochodzi od nazwy pojedynczej jednostki (cechy) gramatycznej „tag”. Ciąg cech gramatycznych nosi nazwę „tagset”. W dalszej części artykułu będziemy korzystali z tej terminologii. W przypadku leksemów do segmentów w postaci ciągów znakowych zostaną dołączone tagsety – w postaci kolejnych ciągów znakowych – odpowiadające informacjom o części mowy, do której należy dany leksem, i o różnych kategoriach morfologicznych

określonych dla danego rodzaju części mowy. Interfiksy zostaną wyróżnione specjalnym znacznikiem. Dopiero proces taggingu nadaje segmentom znaczenie gramatyczne. Na rysunku 1 zestawiliśmy wykorzystane przez nas tagi. Są one podzielone na grupy, z których w jednym tagsecie może występować maksymalnie jeden tag. Występowanie tagów określających przynależność do części mowy jest obligatoryjne. Każdy tagset kończy się znacznikiem „#".

4 Wieloznaczność strukturalna i leksykalna

Wyróżniamy dwa rodzaje wieloznaczności, które mogą się pojawić podczas segmentacji złożeń – wieloznaczność strukturalną i wieloznaczność leksykalną.

Wieloznaczność leksykalna pojawia się przy taggingu, gdy pojedynczemu segmentowi odpowiada więcej niż jeden zestaw danych w słowniku. Wpisy różnią się np. pod względem rodzaju gramatycznego lub klasy fleksyjnej. Wieloznaczność strukturalna występuje przy segmentacji niezależnie od taggingu. Jeżeli istnieje kilka możliwości rozkładu złożenia na podstawie różnych granic między segmentami, wszystkie należy uznać za poprawne, o ile nie istnieją inne kryteria eliminujące wieloznaczność strukturalną. Zasady dystrybucji interfiksów mogą pomóc w eliminacji wieloznaczności w rozpatrywanych przypadkach.

Przykładem niejednoznacznej segmentacji jest niemiecki wyraz złożony *Druckerwartung*. Czysta segmentacja wyznaczy trzy rozkłady:

(1) *drucker* + *wartung*
(2) *druck* + *erwartung*
(3) *druck* + *er* + *wartung*

Po taggingu znalezionych segmentów pojawią się dodatkowe rozkłady *Druckerwartung*. Segment *druck* może być rzeczownikiem lub rdzeniem czasownika, co powoduje podwojenie się liczby możliwych interpretacji rozkładów zawierających ten segment:

(1) *drucker* (N) + *wartung* (N)
(2) *druck* (N) + *erwartung* (N)
(3) *druck* (V) + *erwartung* (N)
(4) **druck* (N) + *er* (I) + *wartung* (N)
(5) **druck* (V) + *er* (I) + *wartung* (N)

Zbiór możliwych rozkładów w tym przykładzie zawiera kilka rozkładów gramatycznie niepoprawnych, które oznaczono gwiazdką (*). Pojawia się tutaj interfiks *-er* tworzący formy kompozycyjne jedynie z rzeczownikami z końcówką fleksyjną *-er* w liczbie mnogiej, do których *druck* nie należy. Analiza segmentu *druck* wykazuje, że jest to jednosylabowy rzeczownik rodzaju męskiego, bez charakterystycznego wygłosu lub sufiksu; takie rzeczowniki nie przyjmują żadnych interfiksów, co również eliminuje możliwość występowania *-er*. Z czasownikami występuje jedynie interfiks *-e*.

Z tych powodów trzeba odrzucić ostatnie dwa rozkłady. Pozostałe trzeba akceptować jako potencjalnie możliwe. Na tym prostym przykładzie widać, w jaki sposób można wykorzystać interfiksy – albo dokładniej: zasady tworzenia form kompozycyjnych za pomocą interfiksów – do eliminacji niegramatycznych wyników segmentacji bądź taggingu.

Kolejnym przykładem jest złożenie *Lieblingstier*. Wynikiem naiwnej segmentacji są następujące dwa rozkłady:

(1) *liebling* (N) + *s* (I) + *tier* (N)
(2) **liebling* (V) + *stier* (N)

Po zastosowaniu reguł dystrybucyjnych należy odrzucić drugi rozkład, ponieważ rzeczowniki zakończone sufiksem *-ling* tworzą formy kompozycyjne wyłącznie za pomocą paradygmatycznego interfiksu *-s*. Formy bez tego interfiksu są niepoprawne.

5 Dystrybucja interfiksów

Wyróżnia się następujące interfiksy występujące w niemieckich złożeniach: *-e*, *-en*/*-n*, *-er*, *-es* i *-s*. Istnieje jeszcze interfiks *-ens*, który występuje tylko przy niektórych wyrazach i nie zalicza się już do interfiksów produktywnych (por. Fuhrhop 1998: 194).

Występowanie interfiksów zależy według Dudenredaktion (1998: 495) od następujących właściwości pierwszego członu złożenia:

1. od części mowy, do której należy dany leksem;
2. od charakterystyki morfologicznej (np. klasa fleksyjna);
3. od struktury fonologicznej (liczba sylab, typ wygłosu);
4. od złożoności słowotwórczej (simpleks, derywat, złożenie);
5. od semantycznie umotywowanej liczby pierwszego członu;
6. od terytorialnych odmian języka.

Fleischer i Barz (1995: 137) stwierdzają, że ustalenie regularności dystrybucyjnych interfiksów pociąga za sobą pewne trudności. Wynikają one z konieczności pogodzenia dwóch koncepcji – orientacji według reguł gramatycznych z jednej strony oraz orientacji według wzorców leksykalnych z drugiej. Pojedyncze zleksykalizowane złożenia lub formy kompozycyjne nie podporządkowują się wyznaczonym regułom gramatycznym.

Zasady tworzenia form kompozycyjnych za pomocą poszczególnych interfiksów zostały wyczerpująco opisane przez Fuhrhop (1998: 187-220). Autorka zestawia najpierw produktywne interfiksy, a następnie przydziela im grupy form kompozycyjnych. Na podstawie tego podziału opracowaliśmy własne kryteria grupujące formy kompozycyjne według podobnych cech gramatycznych (wspólnych tagów). Następnie wyszczególniliśmy zawarte w nich interfiksy. Jest to podejście odwrotne do klasyfikacji Fuhrhop, orientujące się bardziej według kryteriów podanych powyżej przez Dudenredaktion (1998). Przy czym nie bierzemy pod uwagę terytorialnych odmian języka niemieckiego. Nie uwzględniamy również złożoności słowotwórczej wyrazów, ponieważ wynikające z nich zasady tworzenia form kompozycyjnych nie nadają się naszym zdaniem do opisu formalnego. Traktujemy te przypadki jak formy zleksykalizowane.

Fuhrhop skupia się na formach kompozycyjnych, w których występują jawne interfiksy. W związku z tym nie podaje zasad tworzenia form bez interfiksów lub z tzw. interfiksem zerowym. Kilka reguł dla złożeń bez interfiksów podali Fleischer i Barz (1995: 139) oraz Dudenredaktion (1998: 503).

Zestawienie według form kompozycyjnych. Podsumowując powyższe opisy, zestawiamy formy kompozycyjne według czterech głównych kryteriów wpływających na dystrybucję interfiksów, są to:

1. cechy morfologiczne (przynależność do części mowy, klasy fleksyjnej, informacje, czy wyraz jest *singulare*/*plurale tantum*) – ze względu na to, iż większość interfiksów występuje paradygmatycznie;
2. postać wygłosu lub pewnych sufiksów – wiele sufiksów tworzących derywaty rzeczownikowe wymaga występowania szczególnego interfiksu;

NK		Rodzaj	l.p.	l.m.	Struktura	Sufiks/Wygłos	Interfiksy
1	+N			+P1			-, -e
2	+N		+S2	+P3		–SA	-en
	+N		+S2	+P3		+AE	-n
3	+N		+S3	+P3		+AE	-n
4	+N			+P4			-, -er, -s
5	+N			+P5			-
6	+N		+S1		–M		-, -es
7	+N	+FM			–M		-
8	+N	+FM			+M	+AT	-s, -en
	+N	+FM		–P	+M	+AT	-s
9	+N	+NT/+MS			+M	+SN	-
10	+N	+NT/+MS			+M	+SS	-s
11	+N	+FM			+M	+SS	-s, -en
	+N	+FM		–P	+M	+SS	-s
12	+N	+FM			+M	+SN	-, -en
	+N	+FM		–P	+M	+SN	-
13	+N			–P		–SA	-
VK	+V						-, -e
AK	+A						-

Rysunek 2: Formy kompozycyjne według znaczących tagów

3. jedno- lub wielosylabowość pierwszego członu – wielosylabowe formy kompozycyjne wykazują większe tendencje przyjmowania interfiksów;

4. stopień leksykalizacji, np. złożenia zleksykalizowane w całości lub złożone ze zleksykalizowanymi formami kompozycyjnymi.

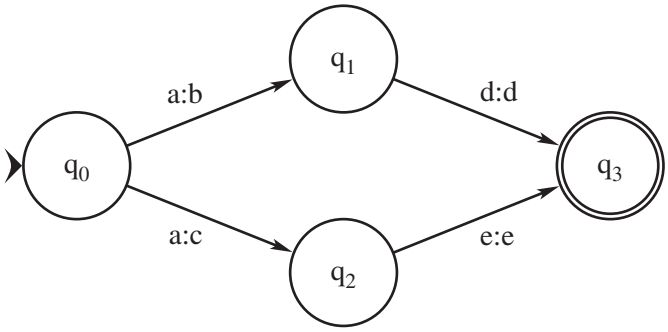
Pierwsze trzy kryteria przedstawiliśmy w postaci tabeli na rysunku 2. Uwzględniamy jedynie znaczące cechy form kompozycyjnych, tzn. umieszczamy w tabeli tylko takie tagi, które trzeba wziąć pod uwagę przy tworzeniu reguł dystrybucyjnych za pomocą transduktorów. Puste miejsca oznaczają brak wypływu danej cechy na wybór interfiksu. W ten sposób można wyznaczyć rozłączne zbiory form kompozycyjnych, gdzie wszystkie wyrazy, w których tagsetach występują wyznaczone tagi, tworzą jeden zbiór.

W słowniku umieszczamy jedynie informacje korespondujące z punktem a. Cechy segmentów z punktów b i c identyfikujemy za pomocą analiz wykorzystujących wiele różnych transduktorów (patrz część 8).

Zbióru form zleksykalizowanych nie można określić na podstawie ich tagów. Zasady wyboru interfiksów w tych złożeniach często są powiązane z historycznymi aspektami i nie sposób wyznaczyć kryteriów synchronicznych. Takie wyrazy trzeba traktować jako wyjątki od reguł określonych na rysunku 2. Reguły opisujące formy zleksykalizowane nie uwzględniają wobec tego postaci tagsetów danych wyrazów, tylko wyznaczają zasady dystrybucji na podstawie postaci ortograficznej wyrazu. Dokładniej omawiamy reguły zleksykalizowane w Junczys-Dowmunt (2005: str. 77).

6 Transduktory – podstawy formalne

Narzędzie, które wykorzystujemy do komputerowej analizy złożeń niemieckich, to transduktory skończone. Kilka podstawowych technicznych właściwości czyni z transduktorów wydajne narzędzia do modelowania regularności morfotaktycznych i do przetwarzania danych lingwistycznych. Transduktory określają pewną klasę grafów i pewną klasę odwzorowań języków formalnych na inne języki formalne. W postaci graficznej odpowiadają skierowanym grafom z przejściami zaetykietowanymi parami symboli. Następujące definicje i wyliczone własności transduktorów pochodzą w większości z Roche i Schabes (1997).



Rysunek 3: Przykładowy transduktor T

Formalnie transduktor skończony jest piątką (Σ, Q, q_0, F, E), gdzie

1. Σ jest pewnym skończonym alfabetem;
2. Q jest skończonym zbiorem stanów;
3. q_0 jest stanem początkowym;
4. $F \subset Q$ jest zbiorem stanów końcowych;
5. $E \subseteq Q \times \Sigma \cup \{\varepsilon\} \times \Sigma^* \times Q$ jest zbiorem krawędzi lub przejść.

Na przykład rysunek 3 jest graficzną postacią przykładowego transduktora T opisanego za pomocą poniższej piątki:

$$\begin{aligned} T \quad = \quad & (\{a,b,c,d,e\}, \{q_0,q_1,q_2,q_3\}, q_0, \{q_3\}, \\ & \{(q_0,a,b,q_1),(q_0,a,c,q_2),(q_1,d,d,q_3),(q_2,e,e,q_3)\}) \end{aligned}$$

Transduktory określają też odwzorowania słów[3] za pomocą rozszerzonego zbioru krawędzi $\hat{E}$ definiowanego poprzez następujący rekursywny związek:

- jeśli $e \in E$, to $e \in \hat{E}$
- jeśli $(q,a,b,q'),(q',a',b',q'') \in \hat{E}$, to $(q,aa',bb',q'') \in \hat{E}$.

Wtedy odwzorowanie f z Σ^* w Σ^* definiowane przez $f(w) = w'$, gdy istnieje $q \in F$ takie, że $(q_0,w,w',q) \in \hat{E}$ jest odwzorowaniem określonym przez transduktor T – pisze się też $f = |T|$. Gdy takie odwzorowanie zwraca dla jednego słowa wejściowego tylko jedno słowo wyjściowe, mówi się o funkcji, natomiast w przypadku wielu możliwych słów wyjściowych – o transdukcji.

[3] „słowo” w znaczeniu ciągu znaków należącego do języka formalnego.

W celu ułatwienia niektórych czynności wprowadza się pojęcia funkcji przejścia i funkcji emisyjnej. Funkcja przejścia $\delta : Q \times \Sigma \rightarrow 2^Q$ odwzorowuje parę złożoną ze stanu i symbolu wejściowego w zbiór stanów w następujący sposób:

$$\delta(q,a) = \{q' \in Q | \exists w' \in \Sigma^* \text{ oraz } (q,a,w',q') \in E\}$$

Funkcja emisyjna $\sigma : Q \times \Sigma \times Q \rightarrow 2^{\Sigma^*}$ odwzorowuje trójkę złożoną ze stanu wyjściowego, symbolu i stanu wynikowego na słowo złożone z symboli alfabetu Σ w następujący sposób:

$$\sigma(q,a,q') = \{w' \in \Sigma^* | (q,a,w',q') \in E\}$$

Jeśli zarówno funkcja przejścia, jak i funkcja emisyjna pewnego transduktora dla wszystkich danych wejściowych zwracają zbiory składające się z nie więcej niż jednego elementu, to taki transduktor nazywa się transduktorem deterministycznym. W przeciwieństwie do automatów skończonych nie można obliczyć dla każdego transduktora niedeterministycznego odpowiednika deterministycznego.

Zalety transduktorów deterministycznych to stosunkowo prosta implementacja i szybkość działania. Żeby znaleźć wartość transdukcji dla danego słowa, wystarczy podążać deterministycznie po przejściach pojedynczej ścieżki w transduktorze. W takim przypadku nie trzeba podawać w argumencie funkcji emisyjnej stanu wynikowego, ponieważ istnieje tylko jeden taki stan.

Każdy rodzaj transduktora można rozszerzyć o dodatkową końcową funkcję emisyjną $\rho : F \rightarrow \Sigma^*$ lub $\rho : F \rightarrow 2^{\Sigma^*}$, która odwzorowuje zbiór stanów końcowych na pojedyncze słowo lub skończony zbiór słów. Oznacza to, że po zakończeniu transdukcji zostają dołączone do słów wyjściowych wartości końcowej funkcji emisyjnej, zależne od stanów końcowych, w których skończyły się ścieżki dla danego słowa wejściowego.

Jeżeli transduktor deterministyczny jest rozszerzony o końcową funkcję emisyjną ρ, to mówimy o transduktorze subsekwencyjnym[4]. Transduktor nazywamy transduktorem *p*-subsekwencyjnym, jeżeli funkcja ρ zwraca zbiór słów, gdzie liczba *p* odpowiada mocy zbioru z największą ilością elementów.

7 Połączenie słownika z segmentacją

Symbolem Σ_D oznaczamy zbiór złożony ze znaków występujących w niemieckim alfabecie. Podobnie oznaczamy zbiór symboli odpowiadających różnym cechom gramatycznym (tagom) przez symbol Σ_T. Z elementów zbioru Σ_T składają się wszystkie tagsety przydzielone poszczególnym segmentom.

Implementacja słownika opiera się na koncepcji transduktorów z końcową funkcją emisyjną. Ten sam transduktor pełni funkcję słownika oraz jest odpowiedzialny zarówno za proces naiwnej segmentacji, jak i za tagging znalezionych segmentów.

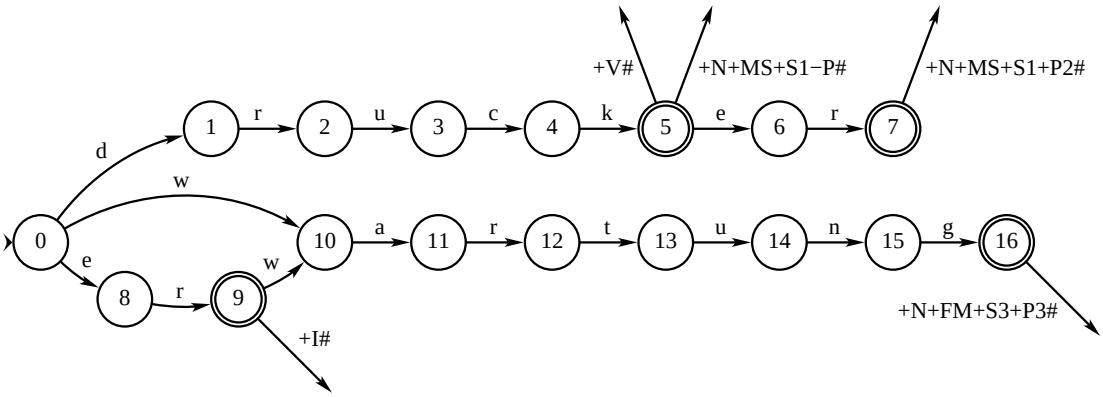
Transduktory są szczególnie wydajnymi formalizmami realizowania słowników. Sprawdzenie czy wyraz składający się z *n* znaków należy do słownika, w przypadku transduktora deterministycznego wymaga dokładnie *n* kroków. Rysunek 4 przedstawia transduktor 2-subsekwencyjny T_{dic}, który zawiera wszystkie możliwe segmenty wyrazu złożonego *Druckerwartung* i definiuje transdukcję określoną następująco:

$$|T_{\text{dic}}| : \Sigma_D^* \rightarrow 2^{(\Sigma_D^* \cdot \Sigma_T^*)}$$

Podążając przejściami łączącymi stan początkowy z jednym ze stanów końcowych, otrzymamy hasło słownikowe reprezentowane przez połączenie symboli tych przejść.

Słowo wyjściowe T_{dic} składa się ze słowa wejściowego i dołączonych wartości końcowej funkcji emisyjnej ρ dla tego słowa. Ponieważ słowa wejściowe nie podlegają modyfikacji, można założyć

[4] ang. *subsequential transducer*, Roche i Schabes (1997: 45).



Rysunek 4: Uproszczony słownik T_{dic}

znacznie uproszczoną główną funkcję emisyjną σ, gdzie $\sigma(q,a,q') = a$. Wynik transdukcji określonej przez transduktor T_{dic} dla segmentu *druck* wygląda następująco:

$$|T_{\text{dic}}|(\text{druck}) = \{\text{druck}_{\text{+N+MS+S1–P\#}}, \text{druck}_{\text{+V\#}}\}$$

Jeżeli zbiór zwróconych słów dla pojedynczego segmentu zawiera więcej niż jeden element, pojawia się wieloznaczność leksykalna. Tagging poszczególnych segmentów odbywa się więc poprzez deterministyczne podążanie po ścieżkach transduktora słownikowego i następnie dołączanie informacji lingwistycznych. Na tym poziomie następuje odróżnienie interfiksów od innych wyrazów słownikowych przez dołączenie odpowiednich informacji lingwistycznych.

Proces konstrukcji słownika jest oparty na przedstawionym w Daciuk et al. (1998) algorytmie do konstrukcji minimalnych acyklicznych automatów na podstawie uporządkowanych list wyrazów słownikowych. W celu zastosowania algorytmu do transduktorów z końcową funkcją emisyjną wprowadziliśmy dodatkowe kryterium równoważności stanów końcowych: stany końcowe $q, q' \in F$ nie są równoważne, jeżeli $\rho(q) \neq \rho(q')$. Taki dodatkowy warunek gwarantuje, że w trakcie stopniowej minimalizacji słownika informacje odpowiadające poszczególnym segmentom pozostaną właśnie przy tych wyrazach.

Pisaliśmy wcześniej, że złożenia mogą się składać z dowolnej liczby elementów. Acykliczny automat lub transduktor nie będzie w stanie opisać wszystkich możliwych złożeń, ponieważ nie może opisać struktur dowolnie długich. Problem ujęcia wszystkich możliwych wyrazów złożonych występuje również tutaj. Automat stanowiący zminimalizowaną reprezentację listy wyrazów słownikowych nie jest wystarczający do opisu wszystkich złożeń, ponieważ nie ma listy zawierającej wszystkie złożenia.

Teoretyczny słownik, który zawiera wszystkie możliwe sekwencje segmentów z transduktora T_{dic}, zaakceptuje oprócz wielu sekwencji niegramatycznych wszystkie gramatycznie poprawne złożenia. Jeżeli dodatkowo wprowadzimy możliwość zaznaczania granic między segmentami oraz możliwość taggingu wszystkich segmentów, to otrzymamy rodzaj naiwnego parsera. Następnie przedstawimy kolejne modyfikacje transduktora T_{dic}, na podstawie których otrzymamy transduktor o opisanych własnościach.

Transduktory traktowane jako automaty skończone ze złożonymi symbolami są zamknięte ze względu na operację Kleene'ego. Wykonując domknięcie Kleene'ego na T_{dic}, otrzymujemy transduktor $(T_{\text{dic}})^+$, który będzie akceptował wszystkie możliwe sekwencje słów wejściowych transduktora T_{dic}, nie będzie jednak zaznaczał granic między segmentami, a informacje lingwistyczne dołączy

tylko do ostatniego segmentu. Domknięcie Kleene'ego można zrealizować, zastępując funkcję przejścia δ przez $\hat{\delta}$:

$$\hat{\delta}(q,a) = \begin{cases} \delta(q,a) \cup \{q_0\} & \text{gdy } \exists p\,(p \in \delta(q,a) \wedge p \in F) \\ \delta(q,a) & \text{w pozostałych przypadkach} \end{cases}$$

Ponieważ informacje lingwistyczne mają się pojawić po każdym segmencie, nasuwa się pomysł, by wykorzystać te informacje również w celu zaznaczania granic między segmentami, rozwiązując w ten sposób dwa problemy równocześnie. Końcowa funkcja emisyjna ρ jest odpowiedzialna za dołączanie poszczególnych tagsetów po ostatnim znaku danych wyjściowych. Żeby zaznaczyć segmenty, trzeba zmusić ją do dołączania tagsetów zawsze, gdy osiągnie się koniec każdego z segmentów, nawet wtedy, gdy nie jest to koniec słowa wejściowego. Ponieważ po takiej modyfikacji funkcja ρ zmienia swój charakter, będziemy ją od tej pory nazywali funkcją taggingu. Trzeba więc sprawdzać każdy osiągnięty stan transduktora, czy nie jest przypadkiem stanem końcowym, czyli czy nie zawiera informacji lingwistycznych. Realizujemy to, zastępując główną funkcję emisyjną σ funkcją $\hat{\sigma}$, określoną w następujący sposób:

$$\hat{\sigma}(q,a,q') = \sigma(q,a,q') \cdot \hat{\rho}(q') = a \cdot \hat{\rho}(q')$$

Określona poniżej funkcja $\hat{\rho}$ jest rozszerzeniem funkcji ρ na zbiór wszystkich stanów Q, gdzie symbol ε oznacza słowo puste:

$$\hat{\rho}(q) = \begin{cases} \rho(q) & \text{gdy } q \in F \\ \varepsilon & \text{gdy } q \in Q \setminus F \end{cases}$$

W ten sposób określiliśmy funkcje $\hat{\delta}$, $\hat{\sigma}$ i $\hat{\rho}$ na podstawie klasycznych funkcji opisujących działanie transduktora T_{dic}. Powstały w ten sposób transduktor T_{seg} odwzorowuje każde złożenie składające się z haseł słownikowych transduktora T_{dic} na zbiór elementów złożonych z ciągów segmentów i odpowiednich tagsetów. Transduktor T_{seg} określa następującą transdukcję:

$$|T_{\text{seg}}| : \Sigma_D^* \rightarrow 2^{(\Sigma_D^* \cdot \Sigma_T^*)^*}$$

Przykładowa transdukcja złożenia *Druckerwartung* zwróci:

$|T_{\text{seg}}|$(druckerwartung) = {
drucker$_{\text{+N+MS+S1+P1\#}}$wartung$_{\text{+N+FM+S3+P3\#}}$,
druck$_{\text{+N+MS+S1–P\#}}$erwartung$_{\text{+N+FM+S3+P3\#}}$,
druck$_{\text{+V\#}}$erwartung$_{\text{+N+FM+S3+P3\#}}$,
druck$_{\text{+N+MS+S1–P\#}}$er$_{\text{+I\#}}$wartung$_{\text{+N+FM+S3+P3\#}}$,
druck$_{\text{+V\#}}$er$_{\text{+I\#}}$wartung$_{\text{+N+FM+S3+P3\#}}$
}

Otrzymany transduktor T_{seg} nie jest już deterministyczny lub subsekwencyjny, nie jest również *p*-subsekwencyjny. Stracił te własności poprzez dokonane modyfikacje. Każdy automat skończony można sprowadzić do deterministycznego odpowiednika, który będzie definiował ten sam język regularny. W przypadku transduktorów niedeterministycznych nie zawsze można uzyskać wersje deterministyczne, subsekwencyjne bądź *p*-subsekwencyjne. Metody sprawdzające, czy istnieje *p*-subsekwencyjny odpowiednik, są oparte na algorytmach analizujących całą strukturę danego transduktora (por. Mohri i Allauzen 2002).

W tym miejscu postaramy się dowieść, że transduktor T_{seg} nie ma odpowiedników deterministycznych lub *p*-subsekwencyjnych, wykorzystując jedynie definicje tych transduktorów oraz powiązania zachodzące pomiędzy strukturą słowa wejściowego i wynikami transdukcji definiowanej przez T_{seg} dla tego słowa.

Segmentacja wyrazu *Druckerwartung* zwraca pięć różnych rozkładów. Skupiając się jedynie na czystej segmentacji bez taggingu, otrzymujemy trzy różne podziały na segmenty.[5]

Transduktory deterministyczne bez końcowej funkcji emisyjnej oraz transduktory subsekwencyjne określają transdukcje będące funkcjami. Oznacza to, że jednemu słowu na wejściu transduktora odpowiada dokładnie jedno słowo wyjściowe. Ponieważ dla wyrazu *Druckerwartung* otrzymujemy trzy różne słowa wyjściowe, wynika z tego, że żaden transduktor deterministyczny lub subsekwencyjny nie będzie w stanie przeprowadzić segmentacji wyrazu *Druckerwartung*. Nie istnieje więc deterministyczny lub subsekwencyjny odpowiednik transduktora T_{seg}.

Za to transduktory *p*-subsekwencyjne są w stanie opisać odwzorowanie pojedynczego słowa wejściowego na zbiór słów wyjściowych o najwyżej *p* elementach. Załóżmy, że istnieje transduktor 3-subsekwencyjny T_3, który odwzorowuje *Druckerwartung* na trzy segmentacje tego słowa wejściowego. Wtedy za wygenerowanie pewnego wspólnego przedrostka jest odpowiedzialna główna funkcja emisyjna. Trzy różniące się fragmenty są dołączane przez końcową funkcję emisyjną.

Załóżmy więc, że istnieje wyraz złożony sam ze sobą *Druckerwartungdruckerwartung*. Transduktor T_{seg} przeprowadzi poprawną segmentację tego wyrazu. Zbiór słów wyjściowych będzie się składał z przynajmniej dziewięciu rozkładów na segmenty. Transduktor T_3 nie będzie w stanie opisać odpowiedniego odwzorowania, ponieważ jego zbiór słów wyjściowych może się składać z co najwyżej trzech elementów. Załóżmy w takim razie istnienie 9-subsekwencyjnego transduktora T_9, który poradzi sobie z tym zadaniem. Jednak nie będzie on mógł opisać wszystkich rozkładów potrójnego złożenia *Druckerwartungdruckerwartungdruckerwartung*, z którym T_{seg} nie ma problemów.

Złożenia mogą mieć dowolną długość, a transduktor T_{seg} może przeprowadzić segmentację dowolnie długiego wyrazu powyższego typu. Liczba zwróconych segmentacji jest proporcjonalna do długości wyrazów, w naszym przykładzie rośnie nawet eksponencjalnie. Żeby skonstruować transduktor *p*-subsekwencyjny, trzeba znać maksymalną liczbę zwróconych rozkładów, tym samym trzeba znać maksymalną długość słowa wejściowego. Dla transduktora T_{seg} nie ma takiej maksymalnej długości słowa wejściowego. Wynika z tego, że nie można skonstruować odpowiednika *p*-subsekwencyjnego transduktora T_{seg}. Można zatem stwierdzić, że ze strukturalnej wieloznaczności – lub strukturalnego niedeterminizmu – niemieckich wyrazów złożonych wynika brak możliwości determinizacji transduktora T_{seg}.

8 Dalsza analiza segmentów

W części dotyczącej dystrybucji interfiksów przedstawiliśmy kryteria mające wpływ na istnienie poszczególnych form kompozycyjnych pierwszych członów złożeń. W słowniku, jak widać, są zawarte informacje o części mowy, a jeśli hasło słownikowe jest rzeczownikiem, również o rodzaju gramatycznym oraz o klasach fleksyjnych w liczbie pojedynczej i w liczbie mnogiej. Liczba sylab oraz występowanie pewnego sufiksu lub wygłosu również mają znaczący wpływ na wybór interfiksu.

Liczba sylab. Dla niektórych form kompozycyjnych posiadanie wielu sylab decyduje o pojawieniu się interfiksu. Jeśli natomiast wyraz jest jednosylabowy, interfiks nie występuje mimo podobnych innych właściwości. Trzeba więc sprawdzić, czy pierwszy człon złożenia jest jedno- czy wielosylabowy, przy czym dokładna liczba sylab w przypadku wyrazu wielosylabowego nie jest istotna. Wykorzystujemy w tym celu również transduktory skończone.

Sylaba składa się z obligatoryjnego ośrodka, zwykle samogłoski, i z fakultatywnych marginaliów, czyli nagłosowej grupy spółgłoskowej zwanej następem oraz wygłosowej grupy spółgłoskowej zwanej zestępem (por. Polański 1999: 575). Podstawowym elementem sylaby jest ośrodek. Wystarczy sprawdzić, czy wyraz zawiera więcej niż jeden ośrodek, by stwierdzić, iż składa się z więcej niż

[5] W części 4. opisujemy wieloznaczność strukturalną i leksykalną wyrazu *Druckerwartung*. Tutaj bierzemy pod uwagę tylko wieloznaczność strukturalną.

jednej sylaby. Marginalia mogą, ale nie muszą, występować. Identyfikacja dokładnej pozycji granic pomiędzy sylabami również nie jest istotna.

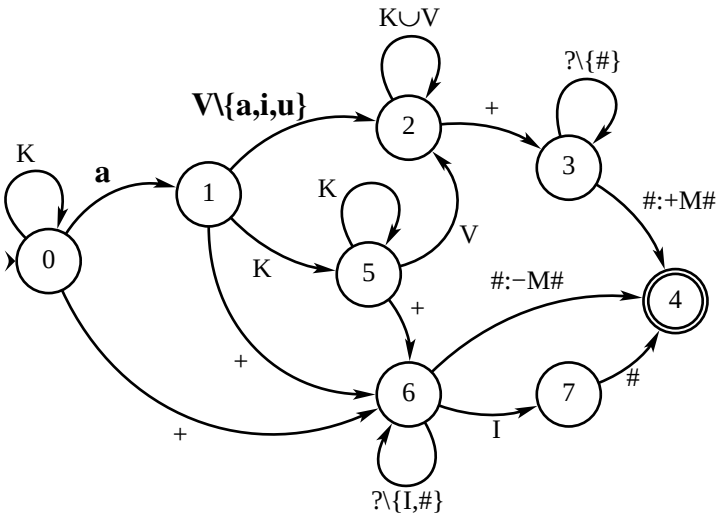
Na rysunku 5 zestawiliśmy reprezentacje grafemowe spółgłosek niemieckich, które mogą pełnić funkcję ośrodka sylaby:

pojedyncze:	*a, e, i, o, u, ä, ö, ü*
podwójne:	*aa, ee, ie, oo*
dyftongi:	*ai, au, äu, ei, eu*

Rysunek 5: Reprezentacja możliwych ośrodków sylabowych

W wyrazach zawierających podwójne samogłoski lub dyftongi mogą wystąpić problemy przy automatycznym rozstrzyganiu, czy w otoczeniu tych samogłosek mamy do czynienia z jedną czy z dwiema sylabami. Przypadki, gdzie pojawiają się dwa dowolne ośrodki na granicy segmentów, można pominąć, ponieważ poprzednio opisany proces segmentacji eliminuje ten problem. Możemy więc założyć, iż każdy ciąg złożony z dwóch samogłosek i odpowiadający jednemu z dyftongów lub z samogłosek podwójnych z powyższej tabeli jest pojedynczym ośrodkiem sylabowym. W przeciwnym razie mamy do czynienia z dwoma stykającymi się ośrodkami złożonymi z pojedynczych samogłosek i tym samym z dwiema sylabami.

Dokładna postać ośrodków jest istotna tylko dla pierwszej sylaby. Jeśli po identyfikacji pierwszego ośrodka wystąpi w segmencie jakakolwiek kolejna samogłoska, to mamy do czynienia z segmentem wielosylabowym. Nie trzeba dalej sprawdzać, czy jest ona częścią dyftongu lub samogłoską pojedynczą.

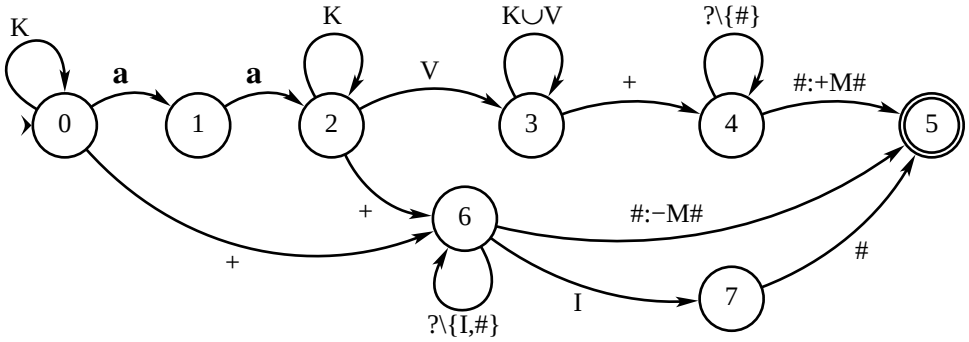


Rysunek 6: Transduktor A_1 dla sylab z pierwszym ośrodkiem *a*

Rysunek 6 przedstawia transduktor A_1, który dla wyrazów z pierwszym ośrodkiem *a* sprawdza, czy występują kolejne ośrodki. Jeżeli segment jest jednosylabowy, to do tagsetu segmentu transduktor dołącza tag $-M$, jeśli zaś wielosylabowy, to $+M$. Symbol V oznacza tutaj zbiór wszystkich samogłosek pojedynczych, symbol K – zbiór wszystkich pojedynczych spółgłosek, przy czym $K, V \subset \Sigma_D$.

Wszystkie transduktory A_1 do A_8 dla samogłosek pojedynczych mają podobną strukturę. Transduktory A_9 do A_{16} sprawdzają występowanie dalszych sylab po sylabach z podwójnymi samogłoskami i dyftongami. Ich struktura odpowiada transduktorowi przedstawionemu na rysunku 7.

Każdy transduktor A_n definiuje transdukcję $|A_n| : \Sigma_D^* \cdot \Sigma_T^* \rightarrow \Sigma_D^* \cdot \Sigma_T^*$ określoną dla pojedynczych segmentów złożonych z grafematycznej postaci segmentu oraz jego tagsetu. Łączymy wszystkie



Rysunek 7: Transduktor A_1 dla sylab z pierwszym ośrodkiem *aa*

transduktory za pomocą sumy określonej dla relacji zbiorów regularnych oraz stosujemy domknięcie Kleene'ego.

$$T_{\text{syl}} = \left(\bigcup_{i=1}^{16} A_i\right)^+$$

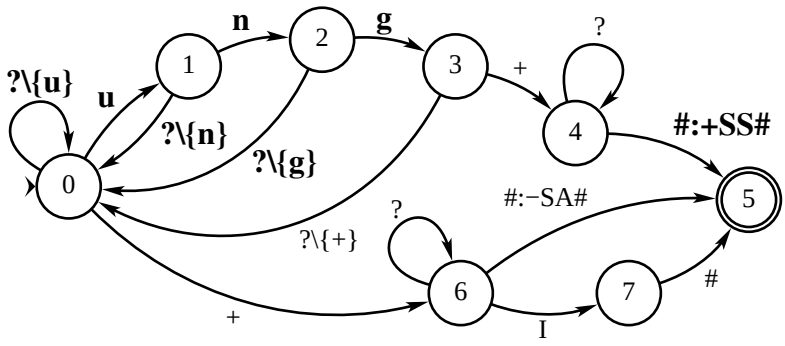
Otrzymany w ten sposób transduktor T_{syl} sprawdza dla wejściowego ciągu segmentów, czy każdy segment tego ciągu jest jedno- lub wielosylabowy, i zaznacza tę cechę. Ponieważ wynikiem segmentacji może być zbiór takich ciągów, rozszerzamy dziedzinę transdukcji określonej przez T_{syl} na zbiór W takich ciągów w następujący sposób:

$$|T_{\text{syl}}|(W) = \bigcup_{w \in W} |T_{\text{syl}}|(w)$$

Dopiero powyższy krok umożliwia złożenie transdukcji segmentującej $|T_{\text{seg}}|$ z transdukcją $|T_{\text{syl}}|$: $2^{(\Sigma_D^* \cdot \Sigma_T^*)^*} \rightarrow 2^{(\Sigma_D^* \cdot \Sigma_T^*)^*}$.

Identyfikacja sufiksów. Poza liczbą sylab sufiks i wygłos członu również mają wpływ na formy kompozycyjne i występowanie interfiksów. Sprawdzenie, czy dany segment zawiera jakiś poszczególny sufiks, sprowadza się do prześledzenia, czy dany ciąg znakowy zawiera jakiś podciąg, co jest typowym zadaniem dla automatów skończonych. Dodatkowo musi być spełniony warunek, że podciąg znajduje się na końcu sprawdzanego ciągu.

Na rysunku 8 widać transduktor B_{18} sprawdzający, czy dany segment kończy się sufiksem *-ung*. Przejście z symbolem $? \setminus \{u\}$ oznacza, że takie przejście może być wykorzystane przez wszystkie symbole należące do alfabetu Σ_D oprócz u. Transduktor B_{18} dołącza do końca tagsetu sprawdzanego



Rysunek 8: Transduktor B_{18} sprawdzający występowanie sufiksu *-ung*

segmentu tag $+SS$, jeżeli znalazł się szukany sufiks, lub $-SA$ w przeciwnym wypadku. Dla innych sufiksów stosujemy inne tagi.

Rozróżnienie sufiksów i różnych rodzajów wygłosu odbywa się podobnie. W konsekwencji wszystkie transduktory od B_1 do B_{19} dla sufiksów i B_{20}, B_{21} dla dwóch rodzajów wygłosu mają strukturę podobną do transduktora B_{18} na powyższym rysunku.

Wszystkie transdukcje $|B_n| : \Sigma_D^* \cdot \Sigma_T^* \rightarrow \Sigma_D^* \cdot \Sigma_T^*$ są określone dla pojedynczych segmentów. Dokonując modyfikacje analogiczne do tych z paragrafu dotyczącego analizy sylab, otrzymujemy transduktor T_{suf} określający transdukcję $|T_{\text{suf}}| : 2^{(\Sigma_D^* \cdot \Sigma_T^*)^*} \rightarrow 2^{(\Sigma_D^* \cdot \Sigma_T^*)^*}$.

9 Reguły dystrybucyjne

Do tej pory analiza złożeń skupiała się jedynie na pojedynczych segmentach, bez uwzględnienia związków dystrybucyjnych pomiędzy nimi zachodzących. W niniejszym artykule zakładamy istnienie dwóch poziomów kontekstu morfologicznego zachodzącego wewnątrz każdego złożenia – kontekstu globalnego oraz kontekstu lokalnego. Kontekst globalny jest realizowany przez zasady łączenia się poszczególnych członów. Można też tutaj mówić o pewnej syntagmatyce złożeń. Kontekst lokalny odpowiada paradygmatyce form kompozycyjnych i zachodzi wewnątrz członu złożenia. Są to zasady łączenia się form podstawowych danego członu z odpowiednim interfiksem.

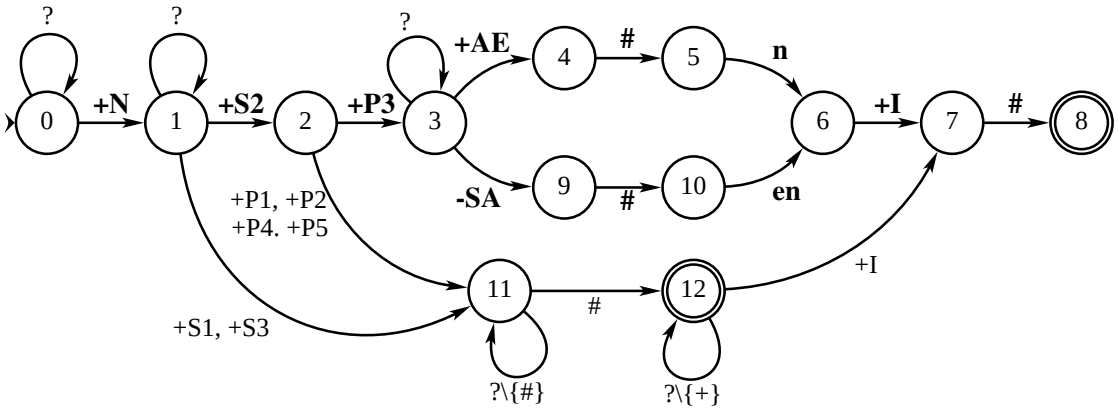
Wykorzystujemy oba konteksty do eliminacji wieloznaczności otrzymanych w wyniku naiwnej segmentacji złożenia. O ile do modelowania globalnych reguł dystrybucyjnych stosujemy jedynie własności zamkniętości automatów i transduktorów, o tyle w przypadku lokalnych zależności konstruujemy transduktor dla każdego rodzaju formy kompozycyjnej.

Lokalne reguły dystrybucyjne. Podstawą każdej lokalnej reguły jest zestawienie znaczących tagów z rysunku 2. Transduktor modelujący formy kompozycyjne danego rodzaju rzeczowników musi uwzględnić wszystkie tagi wyszczególnione dla tych rzeczowników. Jak widać po numeracji, uwzględniliśmy 13 różnych reguł do tworzenia form kompozycyjnych dla rzeczowników i po jednej dla czasowników i przymiotników.

Przez NK_2 oznaczamy regułę opisującą formy kompozycyjne członów nominalnych z tzw. słabą deklinacją (niem. *schwache Deklination*, por. Dudenredaktion 1998: 223). Do tej grupy zalicza się wszystkie wyrazy, które w liczbie pojedynczej odmieniają się według schematu S2, a w liczbie mnogiej według schematu P3. Pojawienie się odpowiednich tagów w tagsecie danego członu kwalifikuje go jednoznacznie do słabych rzeczowników. Można jednak w tej grupie wyodrębnić dwie rozłączne podgrupy, które przyjmują różne (choć allomorficzne) interfiksy – *-n* dla rzeczowników kończących się na *e* oraz *-en* we wszystkich pozostałych wypadkach.

Wprowadzimy w tym miejscu pojęcie tolerancji dla reguł lokalnych. Mówimy, że reguła jest tolerancyjna wobec wszystkich form kompozycyjnych, których nie opisuje, to znaczy, że uznaje je jako poprawne, ponieważ nie zawiera informacji, na podstawie których mogłaby orzec niepoprawność danej formy kompozycyjnej. Reguła jest nietolerancyjna wobec grupy form kompozycyjnych, którą dokładnie opisuje, to znaczy, że jeśli wyraz należy do tej grupy, musi tworzyć poprawną formę kompozycyjną, inaczej zostanie odrzucony. Dla reguły NK_2 oznacza to, że będzie akceptowała wszystkie formy kompozycyjne rzeczowników nieodmieniających się według słabej deklinacji, niezależnie od poprawności gramatycznej danej formy. Jeżeli jednak dany wyraz jest rzeczownikiem słabym, to odrzuci ona wszystkie niegramatyczne formy kompozycyjne.

Transduktor NK_2 z rysunku 9 implementuje odpowiednią regułę lokalną. Aby opisywany przypadek był bardziej czytelny stosujemy tutaj symbole złożone do oznaczania przejść pomiędzy stanami. Nie zmienia to funkcjonalności transduktora, ponieważ wystarczy w miejscu złożonych symboli założyć dodatkowe stany i przejścia oznaczone kolejno pojedynczymi znakami składowymi. Ścieżki z pogrubionymi symbolami opisują główną, nietolerancyjną część reguły, a niepogrubione – część tolerancyjną. Jak widać, transduktor identyfikuje daną grupę rzeczowników na podstawie znaczących



Rysunek 9: Formy kompozycyjne rzeczowników ze słabą deklinacją – reguła NK_2

tagów, po czym musi pojawić się jeden z odpowiednich interfiksów *-en* lub *-n*. Dla ścieżek tolerancyjnych nie określa się formy interfiksu, akceptuje się również jego brak.

Po konstrukcji pozostałych reguł nominalnych według podobnych zasad co powyżej wystarczy wykonać operację przecięcia określonej dla zbiorów regularnych:

$$NK = \bigcap_{i=1}^{13} NK_i$$

Zamkniętość transduktorów pod względem przecięcia nie jest gwarantowana dla wszystkich rodzajów transduktorów. Możemy ją jednak stosować w tym przypadku, ponieważ opisane reguły realizują odwzorowania identycznościowe, są więc równoważne z automatami skończonymi. Wszystkie automaty skończone są zamknięte na przecięcia (por. Roche i Schabes 1997: 6). Otrzymany w ten sposób transduktor *NK* jest równocześnie nietolerancyjny wobec wszystkich opisanych rzeczownikowych form kompozycyjnych oraz tolerancyjny wobec wszystkich pozostałych form.

Dla czasownikowych i przymiotnikowych form kompozycyjnych wystarczy założyć po jednej regule dystrybucyjnej, ponieważ dystrybucja interfiksów jest skomplikowana tylko dla rzeczowników. Transduktory *VK* i *AK* implementują odpowiednie reguły lokalne.

Globalne reguły dystrybucyjne. Konteksty globalne opisują zasady łączenia się formy kompozycyjnej pierwszego członu z drugim członem. Jeśli mamy do czynienia ze złożeniem wielokrotnym, to występuje ciąg form kompozycyjnych, który łączy się z pojedynczym członem głównym. Nie uwzględniamy w tym miejscu związków hierarchicznych, zachodzących między poszczególnymi członami.

Dla wyrazów pojedynczo złożonych zakładamy, że pierwszy człon może być rzeczownikową, czasownikową lub przymiotnikową formą kompozycyjną. Drugi człon będzie formą podstawową wyrazu należącego również do jednej z powyższych części mowy. Transduktory *NG*, *VG* i *AG* akceptują takie formy podstawowe, sprawdzając tylko tagi określające przynależność do danej części mowy. Ponieważ w słowniku zawarto jedynie formy podstawowe wyrazów, nie trzeba sprawdzać innych cech.

Konkatenując sumę transduktorów określających formy kompozycyjne pierwszego członu z sumą transduktorów akceptujących podstawowe formy drugiego członu, otrzymujemy globalną regułę dystrybucyjną dla wyrazów pojedynczo złożonych.

$$T_{\text{poj}} = (NK \cup VK \cup AK) \cdot (NG \cup VG \cup AG)$$

Taka reguła nie będzie akceptowała wyrazów wielokrotnie złożonych. Jednak zakładając, że w złożeniach wielokrotnych pierwszy człon jest potencjalnie nieskończonym ciągiem form kompozycyjnych składającym się z co najmniej jednej takiej formy, wystarczy zmodyfikować regułę T_{poj} do następującej postaci:

$$T_{\text{dis}} = (NK \cup VK \cup AK)^{+} \cdot (NG \cup VG \cup AG)$$

Powstały transduktor T_{dis} łączy wszystkie lokalne reguły dystrybucyjne w całość i ustala relacje pomiędzy nimi zachodzące. Określamy w ten sposób kolejność możliwych członów. Reguły lokalne gwarantują, że nie zostaną zaakceptowane niegramatyczne formy kompozycyjne, reguły globalne sprawdzają, czy poprawnie skonstruowane formy kompozycyjne pojawiają się w złożeniu w odpowiednim miejscu, odrzucając np. wszystkie wyniki segmentacji twierdzące, że znaleziono interfiks na końcu wyrazu złożonego.

Wszystkie transduktory modelujące reguły dystrybucyjne odwzorowują słowo wejściowe na identyczne słowo wyjściowe w przypadku, gdy zostało zaakceptowane, w przeciwnym przypadku zwracają puste słowo. Transduktor T_{dis}, który jest połączeniem wszystkich reguł, działa analogicznie oraz określa transdukcję $|T_{\text{dis}}| : (\Sigma_D^* \cdot \Sigma_T^*)^* \rightarrow (\Sigma_D^* \cdot \Sigma_T^*)^*$. Rozszerzając transdukcję podobnie jak poprzednio na zbiory słów wejściowych, otrzymujemy transdukcję $|T_{\text{dis}}| : 2^{(\Sigma_D^* \cdot \Sigma_T^*)^*} \rightarrow 2^{(\Sigma_D^* \cdot \Sigma_T^*)^*}$. Wtedy jeśli $W_{\text{we}} \in 2^{(\Sigma_D^* \cdot \Sigma_T^*)^*}$ będzie zbiorem słów wejściowych otrzymanym po segmentacji i analizie segmentów oraz $W_{\text{wy}} = |T_{\text{dis}}|(W_{\text{we}})$ będzie zbiorem słów wyjściowych po transdukcji, to $W_{\text{wy}} \subseteq W_{\text{we}}$ oraz W_{wy} powinien zawierać tylko gramatycznie poprawne złożenia.

Gramatyczna poprawność jest oczywiście pojęciem względnym i zależy od jakości wyodrębnionych reguł. Wspominaliśmy wcześniej o regułach zleksykalizowanych dla form kompozycyjnych niepodporządkowujących się ogólnym zasadom dystrybucji interfiksów. Każdą taką regułę można dołączyć do modelu za pomocą intersekcji reguły zleksykalizowanej z transduktorem opisującym formy kompozycyjne odpowiedniej części mowy. W ten sposób można sukcesywnie zmniejszyć tolerancję modelu dla przypadków nieujętych.

10 Połączenie wszystkich elementów

Opisaliśmy już wszystkie elementy modelu. Przedstawimy teraz łączne działanie wszystkich transduktorów dla naszego przykładowego złożenia *Druckerwartung*. Wyniki jednej transdukcji będą wykorzystywane jako zbiór słów wejściowych do następnej transdukcji. Konstruujemy w ten sposób tzw. kaskadę transduktorów.

Poniższy przykład segmentacji pojawił się już wcześniej w odpowiedniej części artykułu. Jak widać, otrzymamy pięć różnych rozkładów wyrazu wejściowego w postaci ciągów segmentów i dołączonych tagsetów. Występują tu zarówno wieloznaczności strukturalne ze względu na postać segmentów, jak i wieloznaczności leksykalne ze względu na różniące się tagsety homonimicznych segmentów. Zbiór W_1 zawiera wyniki tymczasowe naiwnej segmentacji, które będą analizowane w dalszej części procesu.

$$W_1 = |T_{\text{seg}}|(\text{druckerwartung})$$

$W_1 = \{$
(1) $\text{drucker}_{\text{+N+MS+S1+P1\#}}\text{wartung}_{\text{+N+FM+S3+P3\#}}$,
(2) $\text{druck}_{\text{+N+MS+S1–P\#}}\text{erwartung}_{\text{+N+FM+S3+P3\#}}$,
(3) $\text{druck}_{\text{+V\#}}\text{erwartung}_{\text{+N+FM+S3+P3\#}}$,
(4) $\text{druck}_{\text{+N+MS+S1–P\#}}\text{er}_{\text{+I\#}}\text{wartung}_{\text{+N+FM+S3+P3\#}}$,
(5) $\text{druck}_{\text{+V\#}}\text{er}_{\text{+I\#}}\text{wartung}_{\text{+N+FM+S3+P3\#}}$
$\}$

Analizę członów skupiliśmy w jednej transdukcji będącej złożeniem transdukcji $|T_{\text{syl}}|$ i $|T_{\text{suf}}|$. Do tagsetów poszczególnych członów zostają dołączone tagi informujące o liczbie sylab i o postaci sufiksu lub wygłosu. W wyniku analizy ilość elementów zbiorów W_1 i W_2 nie podlega zmianom.

$$W_2 = |T_{\text{syl}} \circ T_{\text{suf}}|(W_1)$$

$$\begin{aligned}
&W_2 = \{\\
&(1)\quad \text{drucker}_{\text{+N+MS+S1+P1+M–SA\#}}\text{wartung}_{\text{+N+FM+S3+P3+M+SS\#}},\\
&(2)\quad \text{druck}_{\text{+N+MS+S1–P–M–SA\#}}\text{erwartung}_{\text{+N+FM+S3+P3+M+SS\#}},\\
&(3)\quad \text{druck}_{\text{+V–P–SA\#}}\text{erwartung}_{\text{+N+FM+S3+P3+M+SS\#}},\\
&(4)\quad \text{druck}_{\text{+N+MS+S1–P–M–SA\#}}\text{er}_{\text{+I\#}}\text{wartung}_{\text{+N+FM+S3+P3+M+SS\#}},\\
&(5)\quad \text{druck}_{\text{+V–M–SA\#}}\text{er}_{\text{+I\#}}\text{wartung}_{\text{+N+FM+S3+P3+M+SS\#}}\\
&\}
\end{aligned}$$

Po umieszczeniu wszystkich wymaganych tagów może nastąpić analiza gramatycznej poprawności wyników naiwnej segmentacji. Służą do tego opisane powyżej lokalne i globalne reguły dystrybucyjne. Zadaniem naiwnej segmentacji jest znalezienie zbioru wszystkich możliwych rozkładów na segmenty. Analiza gramatyczności ma usunąć z tego zbioru wszystkie rozkłady niegramatyczne.

$$W_{\text{wy}} = |T_{\text{dis}}|(W_2)$$

$$\begin{aligned}
&W_{\text{wy}} = \{\\
&(1)\quad \text{drucker}_{\text{+N+MS+S1+P1+M–SA\#}}\text{wartung}_{\text{+N+FM+S3+P3+M+SS\#}},\\
&(2)\quad \text{druck}_{\text{+N+MS+S1–P–M–SA\#}}\text{erwartung}_{\text{+N+FM+S3+P3+M+SS\#}},\\
&(3)\quad \text{druck}_{\text{+V–P–SA\#}}\text{erwartung}_{\text{+N+FM+S3+P3+M+SS\#}}\\
&\}
\end{aligned}$$

Widzimy, że zostały usunięte rozkłady (4) i (5). Są one niepoprawne pod względem gramatycznym. Rozkład (4) został odrzucony przez regułę lokalną NK_{13}, która orzeka, że rzeczownik niemający specyficznego sufiksu oraz wygłosu, będący ponadto *singulare tantum*, tworzy formę kompozycyjną bez interfiksu. Reguła VK jest odpowiedzialna za wyeliminowanie rozkładu (5), ponieważ ustala ona, że formy kompozycyjne czasowników albo nie zawierają interfiksów, albo łączą się z interfiksem *-e*. Pozostałe rozkłady są zgodne ze sformułowanymi regułami lub mieszczą się w zakresie tolerancji wszystkich reguł równocześnie. Jeśli nie została określona żadna reguła wykluczająca dany rozkład, trzeba go akceptować. Dzieje się tak np. w przypadku rozkładu (1). Na tym kończy się opis naszego modelu.

Literatura

Daciuk, J., Watson, B. W. i Watson, R. E., 1998. Incremental Construction of Minimal Acyclic Finite State Automata and Transducers. W L. Karttunen (red.) *FSMNLP'98: International Workshop on Finite State Methods in Natural Language Processing*, str. 48–55. Somerset, New Jersey: ACL.

Dudenredaktion (red.) 1998. *Duden. Grammatik der deutschen Gegenwartssprache*, tom 4. Mannheim: Dudenverlag.

Eichinger, L. M., 2000. *Deutsche Wortbildung: Eine Einführung*. Tübingen: Gunter Narr Verlag.

Fleischer, W. i Barz, I., 1995. *Wortbildung der deutschen Gegenwartssprache*. Tübingen: Max Niemayer Verlag.

Fuhrhop, N., 1998. *Grenzfälle morphologischer Einheiten*. Rozprawa doktorska, Freie Universität Berlin.

Glück, H. (red.) 2000. *Metzler-Lexikon Sprache*. Stuttgart: Verlag J. B. Metzler.

Grzegorczykowa, R., Laskowski, R. i Wróbel, H. (red.) 1999. *Morfologia – Gramatyka współczesnego języka polskiego*. Warszawa: Wydawnictwo Naukowe PWN.

Junczys-Dowmunt, M., 2005. *Ein Finite-State-Modell für einfach und mehrfach zusammengesetzte Komposita*. Praca magisterska, Uniwersytet Kazimierza Wielkiego.

Langer, S., 1998. Zur Morphologie und Semantik von Nominalkomposita. W *Tagungsband KONVENS 98*, str. 83–97. Bonn.

Mohri, M. i Allauzen, C., 2002. p-Subsequentiable Transducers. W *Seventh International Conference CIAA 2002*, str. 24–34.

Polański, K. (red.) 1999. *Encyklopedia językoznawstwa ogólnego*. Wrocław: Zakład Narodowy im. Ossolińskich – Wydawnictwo.

Roche, E. i Schabes, Y., 1997. Introduction to Finite-State Devices in Natural Language Processing. W E. Roche i Y. Schabes (red.) *Finite State Language Processing*, str. 1–66. Cambridge: MIT Press.