PRZEGLĄD RUSYCYSTYCZNY
2010, nr 3 (131)

Ewelina Alwasiak
Kraków

ZABURZENIA SYNTAKTYCZNE WEWNĄTRZ FRAZ NOMINALNYCH W KOMPUTEROWEJ KONWERSJI TEKSTU Z JĘZYKA ANGIELSKIEGO NA POLSKI I ROSYJSKI. SPOJRZENIE GENERATYWNE

1.1. Wstęp

Artykuł ten jest kontynuacją moich rozważań nad specyfiką modelu języka polskiego wykorzystywanego w wyszukiwarkach internetowych do automatycznego tłumaczenia ciągów z języka angielskiego na polski w porównaniu z rozwiązaniami proponowanymi w modelach języka rosyjskiego. W artykule poświęconym zaburzeniom morfologicznym[1], przyjmując, że wszystkie omawiane przeze mnie modele językowe wyszukiwarek generują ciągi dewiacyjne, starałam się podkreślić mniejszą liczbę zaburzeń w modelach rosyjskich, zwłaszcza w modelu wyszukiwarki Google i translatora wyszukiwarki Bing. Przytoczone przeze mnie argumenty świadczą, że geneza zaburzeń w procesie przekładu automatycznego nie wynika zatem z charakteru bądź nieprzekładalności faktów językowych z języka rodziny germańskiej na język grupy słowiańskiej — wówczas obydwa modele, tj. polski i rosyjski, generowałyby podobną liczbę aberracji. Tymczasem jakość tłumaczenia z języka angielskiego na rosyjski i propozycje pewnych rozwiązań inspirują do refleksji nad naturą „polskich" zaburzeń. Dlatego też, po uprzednim omówieniu zaburzeń morfologiczno-akomodacyjnych, niniejsze opracowanie stanowi wstęp do badań nad generowanymi przez wyszukiwarki zaburzeniami syntaktycznymi polskich ciągów. Rozważania te rozpoczynam analizą dewiacji wewnątrz fraz nominalnych, ponieważ błędna identyfikacja podmiotu jest często przyczyną nieprawidłowej generacji predykatu, a w efekcie, błędnych uzgodnień akomodacyjnych. Analizie

[1] E. Alwasiak: *Zaburzenia relacji akomodacyjnych w automatycznym tłumaczeniu z języka angielskiego na polski*. „Polonica" 2010, R. XXX, s. 73–85.

poddaję problemy z rozpoznawaniem podmiotu (często szeregowego) oraz hierarchii elementów budujących grupę rzeczownikową. Omówione zostały również zaburzenia koordynacji węzłów. Niniejszy artykuł jest zatem próbą rewizji reguł syntaktycznych dotychczasowego modelu i propozycją ich reorganizacji w ujęciu generatywnym, mającą na celu opracowanie modelu korygującego.

1.2. Cel analizy

Niniejszy artykuł ma przede wszystkim charakter teoretyczny. Moim zamiarem jest zasygnalizowanie problemów związanych z konwersją komputerową fraz nominalnych i naszkicowanie wstępnych rozwiązań w celu zreorganizowania obecnego modelu, a tym samym zdefiniowania reguł mechanizmu selekcyjnego automatycznie odrzucającego ciągi dewiacyjne.

Na obecnym etapie rozwoju technologii systemy tłumaczenia komputerowego stanowią rozwiązanie alternatywne w takich sytuacjach, gdy liczy się czas i koszty, a absolutna adekwatność przekładu nie jest priorytetem. Stąd też rosnące zapotrzebowanie na szybkie i tanie generowanie tłumaczenia zawartości stron internetowych. Jednak jakość tego typu tłumaczenia często bywa niesatysfakcjonująca pod względem poprawnościowym. Dlatego też zasadniczym celem wspomnianych badań porównawczych nad jakością przekładu komputerowego wyszukiwarek jest próba odpowiedzi na pytanie, dlaczego w modelu języka polskiego dochodzi do generowania ciągów dewiacyjnych, podczas gdy liczba zaburzeń generowanych przez model języka rosyjskiego jest znacznie mniejsza. Zaobserwowane zjawisko dotyczy zwłaszcza rozbudowanych fraz nominalnych i werbalnych, gdzie modelowi polskiemu, stosowanemu przez wyszukiwarki Google i Bing, najtrudniej rozpoznać człony główne i zależne. Translator Yahoo nie posiada opcji tłumaczenia ciągów w relacji z języka angielskiego na polski, a jedynie z języka angielskiego na rosyjski, w związku z czym prezentacja zaburzeń dotyczy wyłącznie modelu rosyjskiego. Są to głównie problemy z identyfikacją składniowych funkcji wyrazów tekstowych, hierarchią elementów budujących konkretny węzeł i koordynacją rozbudowanych węzłów.

Jestem przekonana, że po ustaleniu hierarchii zaburzeń możliwe będzie określenie ich statusu, co w efekcie powinno stworzyć podstawę do opracowania propozycji korekty obecnego modelu, tj. propozycji takiej jego przebudowy, by reguły nim rządzące były punktem wyjścia do generowania struktur poprawnych i automatycznego odrzucania struktur nieakceptowalnych.

1.3. Przedmiot analizy

Podobnie jak w poprzednim badaniu, analizą zostały objęte ciągi generowane automatycznie przez wyszukiwarkę Google, która dotychczas jako jedyna w Polsce proponuje automatyczne tłumaczenie zawartości całych stron internetowych z języka angielskiego na polski poprzez usługę *Translate this page*. Pozostałe samodzielne wyszukiwarki takiej funkcji nie posiadają — użytkownicy zmuszeni są korzystać z odrębnych translatorów, np.: Yahoo!: http://babelfish.yahoo.com/ czy Bing: http://www.microsoft-translator.com/Default.aspx

Dla potrzeb niniejszego opracowania materiał został zaprezentowany w wyjątkowo okrojonej formie — za pomocą przykładów najbardziej charakterystycznych z syntaktycznego punktu widzenia. Następnie, aby rozszerzyć możliwości porównawcze, omawiane ciągi, błędnie tłumaczone przez model wyszukiwarki Google, poddano procesowi tłumaczenia komputerowego w modelach proponowanych przez wyszukiwarki Yahoo i Bing.

Przedmiotem analizy są jednostki leksykalne przetłumaczone przez model poprawnie, czyli takie, w których określonemu fragmentowi tekstu w języku angielskim (tu: pełna fraza) zostały przypisane odpowiedniki w języku polskim.

1.4. Metoda analizy

Przy omówieniu zagadnienia zaburzeń składniowych wewnątrz fraz nominalnych, zwłaszcza z podmiotem szeregowym, mamy do czynienia głównie z naruszeniem reguł szyku, które są niezwykle istotne dla języków z rozbudowaną fleksją, takich jak np. język rosyjski i polski. W języku angielskim, odwrotnie — szyk jest stały. Stąd też tak duża liczba przekłamań w tłumaczeniu komputerowym, zwłaszcza w konstrukcjach z przydawkami rzeczownymi i przymiotnymi. W tej dziedzinie pomocne mogłyby być rozwiązania proponowane przez bezkontekstową gramatykę struktur frazowych. W anglojęzycznych pracach generatywistycznych koncepcja ta znana jest pod nazwą „x-bar syntax", natomiast polscy językoznawcy używają dwóch terminów: w terminologii Ireneusza Bobrowskiego — „składnia kategorii wzmocnionych"[2], w terminologii Kazimierza Polańskiego — „składnia frazowa wielostopniowa"[3]. Jest to alternatywna teoria języka, która — podobnie jak gramatyka generatyw-

[2] I. Bobrowski: *Gramatyka generatywno-transformacyjna (TG) a uogólniona gramatyka struktur frazowych (GPSG)*. Wrocław: Ossolineum 1988, s. 10.

[3] Tamże, s. 22.

no-transformacyjna — ma charakter generatywny, jednak nie zawiera ani transformacji przenoszącej, ani reguł stylistycznych, składa się zaś jedynie z bezkontekstowych frazowych reguł przepisywania[4].

U podstawy tej teorii leży przekonanie jej twórców, że język naturalny może być opisany wyłącznie z pomocą bezkontekstowych reguł struktur frazowych, zatem gramatyką języka naturalnego może być jedynie zbiór bezkontekstowych reguł struktur frazowych[5]. Rozwiązanie to wydaje się inspirujące szczególnie dla dziedziny tłumaczeń maszynowych, bowiem model konwertujący tekst nie może odwołać się ani do faktów pozajęzykowych, ani do kontekstu wypowiedzi. Informacje potrzebne do generowania tekstu musi zostać wydobyte ze struktury frazy, na podstawie hierarchii, w jakiej zostały ujęte jej składniki. Prace polskich generatywistów, jak np. Ireneusza Bobrowskiego[6], a przede wszystkim badania Geralda Gazdara, Donki Farkasa i Almerindo Ojedy, którzy „udowodnili, że pewne fakty językowe można wyjaśnić przez odwołanie się do relacji pomiędzy drzewkami struktur frazowych a strukturami wewnętrznymi kategorii”[7], są propozycjami ścisłego i formalnego ujęcia relacji pomiędzy morfologią a składnią, a tym samym stanowią inspirację do generatywnego spojrzenia na próby eliminacji zaburzeń w tłumaczeniach komputerowych.

2. Zaburzenia składniowe

2.1. Szyk elementów składowych węzłów

Pytanie o naturę zaburzeń wewnątrz grupy nominalnej jest zasadniczo pytaniem o porządek, czyli sformalizowane reguły łączenia składników jej członów zależnych. W niniejszym opracowaniu kwestia szyku elementów występujących w obrębie polskiej frazy nominalnej i związane z nim problemy w ujęciu metodologicznym są jedynie wspomniane i naszkicowane. Więcej informacji na ten temat można znaleźć np. w rozprawach Zuzanny Topolińskiej, Władysława Śliwińskiego i Stanisława Jodłowskiego[8].

W gramatyce generatywnej porządek elementów frazy jest ustalony już na poziomie reguł przepisywania. Analiza przedstawionego materiału

[4] Tamże, s. 30.

[5] Tamże, s. 30.

[6] Zob.: I. Bobrowski: *Składniowy model polszczyzny*. Kraków: Lexis 2005.

[7] I. Bobrowski: *Gramatyka generatywno-transformacyjna (TG)*..., s. 56.

[8] Zob.: *Gramatyka współczesnego języka polskiego. Składnia*. Red. Z. Topolińska. Warszawa: PWN 1984; W. Śliwiński: *Łączliwość składniowo-semantyczna przymiotników z rzeczownikami we współczesnym języku polskim*. Kraków: Uniwersytet Jagielloński 1990; S. Jodłowski. *Podstawy polskiej składni*. Warszawa: PWN 1976, s.167–169.

dowodzi, że przyczyną generowania dewiacyjnych fraz nominalnych są głównie zaburzenia szyku wewnątrz węzłów. Aplikacja reguł składni kategorii wzmocnionych (*x-bar syntax*), czyli ograniczeń dotyczących kierunku poprzedzania składników, mogłaby skutecznie odrzucać pewne typy struktur ciągów semantycznie nieakceptowalnych.

Przystępując więc do generowania związków składniowych wewnątrz grup nominalnych, model powinien bazować na jednostkach leksykonu, które posiadają zdefiniowane: 1) cechy kategorialne +N (rzeczownikowość), +A (przymiotnikowość), +P (przyimkowość); 2) otoczenie, w którym mogą występować, czyli informację o miejscu w strukturze wyjściowej, w której może się pojawić węzeł N (i dominujący nad nim węzeł NP), węzeł A (i dominujący nad nim węzeł AP) oraz węzeł P (i dominujący nad nim węzeł PP). Przełożenie tych informacji na reguły formalne, gdzie węzły rozumiane są jako zespoły cech[9], a następnie przełożenie angielskich struktur na ich polskie odpowiedniki z pewnością nie prowadziłoby do występujących obecnie zaburzeń.

Potrzebę weryfikacji dotychczasowych reguł generowania struktur frazowych w prezentowanych niżej przykładach (1, 2) wyraża akceptowalny charakter grupy rzeczownikowej w konwersji tekstu z języka angielskiego na rosyjski oraz ich dewiacyjność w przekładzie na język polski.

1) Council for American Private Education (CAPE)[10]
Rada ds. Amerykańskiej Edukacji Niepublicznej

Google (pl)	Rady dla Ameryki prywatne Edukacja (CAPE) (nieprawidłowo)
Google (ru)	Американский совет по частному образованию (CAPE) (prawidłowo)

2) American Architectural Foundation[11]
Amerykańska Fundacja Architektoniczna

Google (pl)	American Foundation architektury (nieprawidłowo)
Google (ru)	Американский архитектурный фонд (prawidłowo)

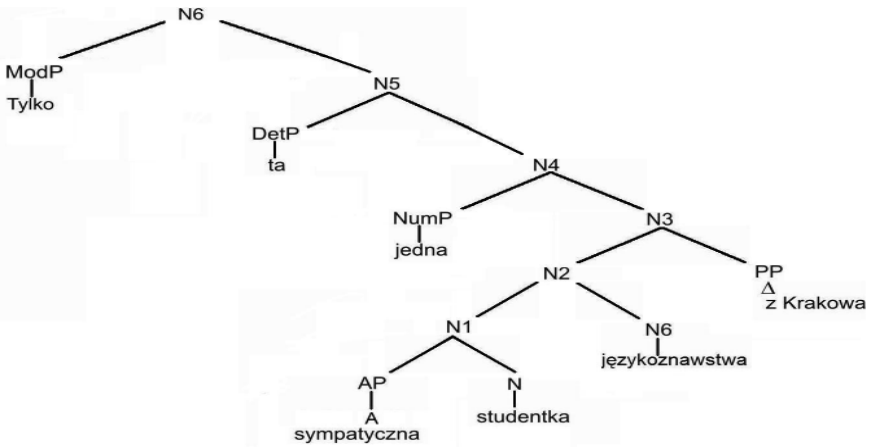
[9] I. Bobrowski: *Gramatyka generatywno-transformacyjna (TG)*…, s. 86, s. 155.
[10] <www.capenet.org> (data dostępu: 21.04. 2009).
[11] <www.archfoundation.org> (data dostępu: 26.05.2009).

2. 2. Weryfikacja obecnych reguł modelu

Przytoczone przykłady przekłamań generowanych przy automatycznej konwersji tekstu powstają w wyniku przeplatania się elementów należących do różnych węzłów frazy nominalnej. Ten typ zaburzeń nie jest generowany w uogólnionej gramatyce struktur frazowych, ponieważ zbiór reguł kategorii pośrednich, na których bazuje (wspomniana w pkt. 1.4) składnia kategorii wzmocnionych, automatycznie odrzuca je jako niedopuszczalne. Szersze omówienie tej teorii można znaleźć m.in. w pracach Bobrowskiego[12] i Katarzyny Węgrzynek[13]. W niniejszym opracowaniu prezentuje ją następująca fraza: *Tylko ta jedna sympatyczna studentka językoznawstwa z Krakowa*.

N6 → (ModP) N5 (tylko) studentka
N5 → (DetP) N4 (ta) studentka
N4 → (NumP) N3 (jedna) studentka
N3 → N2 (PP) studentka (z Krakowa)
N2 → N1 (N6) studentka (językoznawstwa)
N1 →{ AP N, N AP} (sympatyczna) studentka

Omawiana fraza miałaby zatem następującą strukturę wyjściową:



Powyższy przykład struktury wyjściowej, generowanej dzięki omawianym tu regułom definiującym pozycje członów zależnych od rzeczownika, ilustruje szyk przyłączania konkretnych składników z otoczenia przyrze-

[12] I. Bobrowski: *Składniowy model polszczyzny*..., s. 191–200.

[13] K. Węgrzynek: *O możliwości redukcji tzw. słów łącznikowych. Przyczynek do studiów nad cechami przymiotników polskich w ujęciu generatywno-transformacyjnym*. „Polonica" 1994, R. XVI, s. 127–145.

czownikowego. Na poziomie N6 przyłączane są modalizatory, np. *tylko*, *wyłącznie*. Na poziomie N5 — determinatory, np. *ten*, *żaden*, *wszyscy*, *nikt*. Na poziomie N4 — liczebniki, np. *jeden*, *cztery*, *trzysta dwie*. Na poziomie N3 — przyimki, np. *z*, *na*, *w*. Na poziomie N2 — rzeczowniki (najczęściej w funkcji przydawki rzeczownej), natomiast dopiero na poziomie N1 przyłączane są przymiotniki (najczęściej w funkcji przydawki przymiotnej). Rządek AP N zarezerwowany jest dla neutralnych połączeń typu *wspaniały nauczyciel*, *sympatyczna studentka*. Natomiast rządek N AP zarezerwowany jest dla połączeń — w terminologii Katarzyny Węgrzynek[14] — uwikłanych frazeologicznie, typu *stan podgorączkowy*, *układ kostny*, *logarytm naturalny*. Z analiz wynika, że z rozpoznawaniem tej grupy fraz nie radzi sobie model języka polskiego wyszukiwarki Google. Zjawisko ilustrują przykładowe frazy:

3) high school[15]
szkoła średnia

Google (pl)	wysoka szkoła (niepoprawnie)
Google (ru)	средняя школа (poprawnie)

4) school kids[16]
uczniowie

Google (pl)	szkoła dzieci (niepoprawnie)
Google (ru)	ученики (poprawnie)

Prezentowane przykłady źle świadczą o jakości polskiego modelu, a tym samym o słabym opracowaniu komponentu frazeologicznego leksykonu i reguł występowania rzeczowników w składniowej funkcji przydawki (tu: przymiotnej). Zasadniczym zatem wyzwaniem, stojącym przed twórcami korekty dotychczasowego modelu, będzie opracowanie mechanizmów tłumaczenia frazy poprzez przypisanie jej konkretnej struktury wyjściowej, uwzględniającej szyk elementów i koordynację jej węzłów. Model musi posiadać informację o tym, czy jest ona, czy też nie jest, uwikłana

[14] Tamże, s. 129.
[15] <www.cnn.com/EDUCATION> (data dostępu: 17.02.2009).
[16] <volcano.oregonstate.edu> (data dostępu: 22.10.2009).

frazeologicznie. W pracach Andrzeja Bogusławskiego[17] i Macieja Grochowskiego[18] połączenia frazeologiczne ze względu na swoje znaczenie globalne uznawane są za odrębne, niepodzielne jednostki leksykalne. Ich składniki mogą podlegać jedynie ograniczonej substytucji, a w wielu przypadkach substytucja ta w ogóle nie jest możliwa. Andrzej Maria Lewicki[19] proponuje nawet, aby rozróżniać składnię wewnętrzną i zewnętrzną tego typu połączeń.

Podsumowując tę część rozważań, chciałabym podkreślić, że model, konwertując frazę z języka angielskiego na polski, powinien koniecznie posługiwać się sformalizowanymi zasadami zachowania szyku jednostek frazowych oraz odpowiednio zdefiniowaną bazą leksykalną. Leksykon powinien zawierać dokładny opis cech kategorialnych jednostek, łącznie z informacjami dotyczącymi np. wymagań rekcyjnych obligatoryjnych zwłaszcza dla niektórych przymiotników. Informacje na temat możliwości i sposobu łączenia ze sobą jednostek tekstowych model powinien ekstrahować zarówno z reguł bazowych komponentu gramatycznego, jak i ze słownika. Tworzenie poprawnych związków składniowych polegałoby zatem na ścisłej współpracy tych dwóch komponentów — niedopracowanie przynajmniej jednego z nich powoduje liczne zaburzenia o charakterze składniowo-semantycznym dla całego modelu.

2.3. Zaburzenia relacji jednostek frazowych poziomu [N2 N1 N6] i [N1 AP, N]

2.3.1. Zaburzenia w konwersji przydawki rzeczownej

Aby dodatkowo nie zaciemniać obrazu zaburzeń, kwestia obligatoryjnej obecności determinatorów (the, a, an) we frazach angielskich została pominięta. W badaniach nad pełnym opisem konwersji struktur frazowych z języka angielskiego na polski należy koniecznie do tego zagadnienia powrócić. Analizę rozpoczynamy od zaburzeń relacji rzeczownik–rzeczownik, gdzie model ma największe trudności z odróżnieniem rzeczownika głównego od zależnego, który pełni składniową rolę przydawki rzeczownej. Zjawisko to ilustrują następujące frazy:

[17] A. Bogusławski: *O zasadach rejestracji jednostek języka*. „Poradnik Językowy" 1976, nr 8, s. 356–364.

[18] M. Grochowski: *Polskie partykuły. Składnia, semantyka, leksykografia*. Wrocław: Ossolineum 1986, s.27.

[19] A.M. Lewicki: *Składnia związków frazeologicznych*. „Biuletyn PTJ" 1986, nr XL, s. 75–83.

5) The president's initiatives[20]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	Inicjatywy prezydenta
Język rosyjski	Инициативы президента

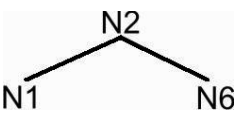
Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	Prezydent inicjatywy (niepoprawnie)
Bing	Inicjatyw Przewodniczący (niepoprawnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	Инициативы президента (poprawnie)
Yahoo!	Инициативы президента (poprawnie)
Bing	Инициативы президента (poprawnie)

Poprawny schemat struktury frazy, za generację której odpowiada reguła [N2 N1 N6], przedstawia się następująco:



Modele języka rosyjskiego poprawnie zakwalifikowały leksemy *inicjatywy* jako N1 i *prezydenta* jako N6. Modele języka polskiego nie rozpoznają funkcji składniowej obydwu rzeczowników. Analogiczne nieprawidłowości obserwuje się w poniższych zaburzeniach konwersji:

6) earth images[21]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	obrazy ziemi
Język rosyjski	изображения земли

Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	ziemia obrazów (niepoprawnie)
Bing	obrazów Ziemi (składniowo poprawnie, morfologicznie niepoprawnie)

[20] <www.ed.gov> (data dostępu: 17.02.2009).

[21] <www.earthscienceworld.org> (data dostępu: 22.05.2009).

Wersje modeli rosyjskich omawianego ciągu:

Google	изображения Земли (poprawnie)
Yahoo!	изображения земли (poprawnie)
Bing	изображения земли (poprawnie)

7) world history[22]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	historia świata
Język rosyjski	история мира

Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	Świat w historii (niepoprawnie)
Bing	Historia świata (poprawnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	История в мире (niepoprawnie)
Yahoo!	история мира (poprawnie)
Bing	Всемирная история (akceptowalnie)

Poważnym problemem obydwu modeli (rzadziej rosyjskiego) jest konwersja leksemów uwikłanych frazeologicznie (wspomniane w pkt. 2.2) i nazw własnych, których leksykon nie definiuje jako całostek tekstowych. Zaburzenia tego typu ilustruje poniższy przykład:

8) North America rivers[23]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	rzeki Ameryki Północnej
Język rosyjski	реки Северной Америки

Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	Ameryka Północna rzek (niepoprawnie)
Bing	Ameryka Północna rzek (niepoprawnie)

[22] <www.historyworld.net> (data dostępu: 22.05.2009).

[23] <www.americanrivers.org> (data dostępu: 26.05.2009).

Wersje modeli rosyjskich omawianego ciągu:

Google	Северная Америка рек (niepoprawnie)
Yahoo!	Реки Северная Америка (składniowo poprawnie, morfologicznie niepoprawnie)
Bing	Рек Северной Америки (składniowo poprawnie, morfologicznie niepoprawnie)

Podobnym wyzwaniem dla obu modeli jest koordynacja węzłów frazy zbudowanej z trzech rzeczowników. Ten typ aberracji prezentowany jest następującą frazą:

9) Development and Advocacy Organization[24]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	Organizacja Rozwoju i Rzecznictwa
Język rosyjski	Организация развития и защиты

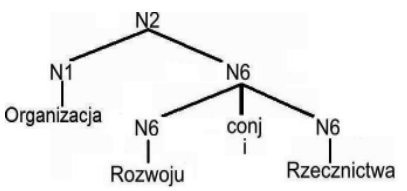
Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	Rozwój i wspieranie organizacji (niepoprawnie)
Bing	Organizacja rozwoju i Rzecznictwo (składniowo poprawnie, morfologicznie niepoprawnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	Организация развития и пропаганды (akceptowalnie)
Yahoo!	Организация развития и защиты (akceptowalnie)
Bing	Развитие и информационно-пропагандистской деятельности Организации (niepoprawnie)

W bezkontekstowej gramatyce reguł frazowych ten typ zaburzeń blokowany jest przez regułę [N2 N1 N6] oraz schemat koordynacji: $\alpha \rightarrow \alpha 1$ conj αn, gdzie α jest dowolną kategorią oprócz spójników oraz partykuł. Symbol *conj* oznacza spójnik (podrzędny lub współrzędny). Po aplikacji wspomnianych reguł model wygenerowałby następującą strukturę wyjściową frazy:



[24] <www.worldvision.org> (data dostępu: 22.10.2009).

2.3.2. Zaburzenia w konwersji przydawki przymiotnej

W omawianej grupie zaburzeń relacji dwóch rzeczowników i przymiotnika model nie rozpoznaje członu głównego i nie przypisuje mu składników zależnych. Zjawisko ilustruje poniższy przykład:

10) Humankind's current crisis[25]
Proponowana postać konwersji na język polski i rosyjski:

Język polski	Obecny kryzys człowieczeństwa
Język rosyjski	Кризис нынешнего человечества

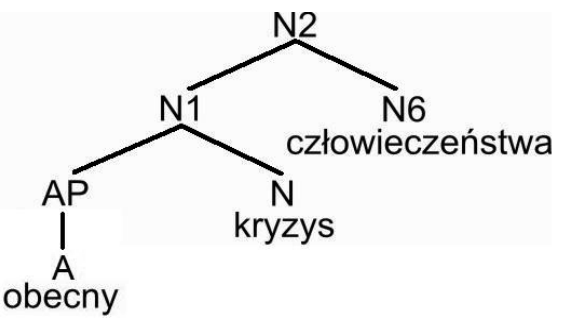
Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	Człowiek obecnego kryzys (niepoprawnie)
Bing	Obecny kryzys ludzkości (poprawnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	Человечество нынешнего кризиса (niepoprawnie)
Yahoo!	Кризис течения человечества (niepoprawnie)
Bing	Нынешний кризис человечества (akceptowalnie)

Poprawnie derywowana struktura wyjściowa frazy, którą definiują reguły: [N2 N1 N6] oraz [N1 AP N], jest następująca:



Analogiczne zaburzenia występują w poniższym materiale przykładowym:

[25] <www.crisis-forum.org.uk> (data dostępu:05.05.2009).

11) your leading information source[26]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	twoje wiodące źródło informacji
Język rosyjski	ваш ведущий источник информации

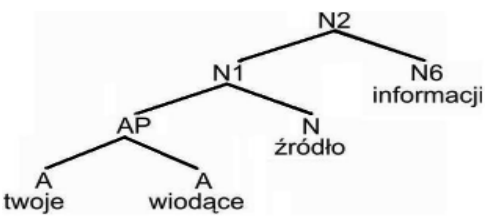
Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	twoje źródło informacji prowadzącej (niepoprawnie)
Bing	wiodące źródła informacji (składniowo poprawnie, morfologicznie niepoprawnie)

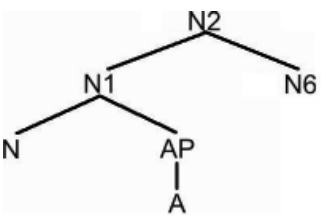
Wersje modeli rosyjskich omawianego ciągu:

Google	Ваш ведущий источник информации (poprawnie)
Yahoo!	ваш ведущий источник информации (poprawnie)
Bing	ваш ведущий источник информации (poprawnie)

Struktura wyjściowa frazy, którą definiują reguły: [N2 N1 N6], [N1 AP N] i [AP (AP)A], jest następująca:



W grupie fraz zawierających jednostki uwikłane frazeologicznie ostatnia reguła, oprócz wspomnianej [N2 N1 N6], posiada formę [N1 N AP] [AP (AP)A], dlatego ostatni stopień struktury wyjściowej omawianych ciągów stanowi lustrzane odbicie poprzedniej (przykład 10):



[26] <www.computerworld.com> (data dostępu: 22.10.2009).

Analogiczną strukturę na poziomie nieterminalnym będą zatem posiadały następujące frazy:

12) satellite altitude[27]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	wysokość satelitarna
Język rosyjski	спутниковая высота

Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	satelita wysokości (niepoprawnie)
Bing	wysokość satelitarne (składniowo poprawnie, morfologicznie niepoprawnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	спутниковая высота (poprawnie)
Yahoo!	спутниковая высота (poprawnie)
Bing	Высота спутник (niepoprawnie)

13) membership organization[28]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	organizacja członkowska
Język rosyjski	членская организация

Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	członkowstwo organizacji (niepoprawnie)
Bing	członkostwo w organizacji (niepoprawnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	организация (niepoprawnie)
Yahoo!	организация членства (niepoprawnie)
Bing	членская организация (poprawnie)

[27] <worldwind.arc.nasa.gov> (data dostępu: 22.10.2009).

[28] <www.asq.org> (data dostępu: 26.05.2009).

14) chemistry industry[29]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	przemysł chemiczny
Język rosyjski	химическая промышленность

Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	chemia branży (niepoprawnie)
Bing	przemysłu chemii (niepoprawnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	химической промышленности (składniowo poprawnie, morfologicznie niepoprawnie)
Yahoo!	индустрия химии (niepoprawnie)
Bing	химическая промышленность (poprawnie)

15) research level[30]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	poziom badawczy
Język rosyjski	исследовательский уровень

Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	badanie poziomu (niepoprawnie)
Bing	poziom badań (akceptowalnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	исследовательский уровень (poprawnie)
Yahoo!	уровень исследования (akceptowalnie)
Bing	исследование уровня (niepoprawnie)

Bez mechanizmów prymarnej selekcji węzłów N1 i N6 dla frazy dwuelementowej model konsekwentnie generuje błędną konwersję fraz trzy- i czteroelementowych. Zjawisko to ilustruje poniższy przykład:

[29] <www.americanchemistry.com> (data dostępu: 26.05.2009).

[30] <matchworld.wolfram.com> (data dostępu: 22.05.2009).

16) This crisis communication plan[31]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	Ten plan komunikacji kryzysowej
Język rosyjski	Этот план кризисной коммуникации

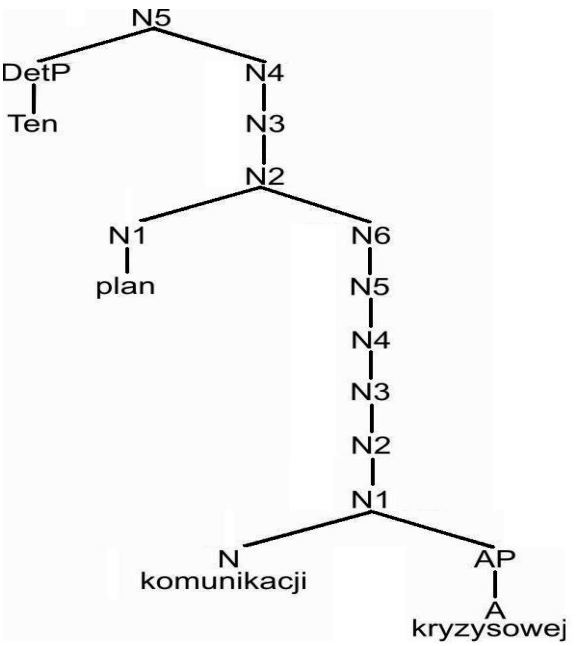
Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	Kryzys ten plan komunikacji (niepoprawnie)
Bing	Plan ten kryzys komunikacji (niepoprawnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	Этот кризис план (niepoprawnie)
Yahoo!	Этот план связи кризиса (niepoprawnie)
Bing	Этот кризис Коммуникационный план (niepoprawnie)

Reguły [N5 DetP N4], [N4 (NumP) N3], [N3 N2 (PP)], [N2 N1 N6], [N1 N AP] odpowiedzialne są za następującą strukturę wyjściową:



[31] <www3.niu.edu/newsplace/crisis.html> (data dostępu: 05.05.2009).

Analogicznie generowana jest struktura wyjściowa frazy trzyelementowej, którą definiują reguły [N2 N1 N6], [N1 N AP]. Brak ich aplikacji obfituje ponadgeneracjami typu:

17) water policy topics[32]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	tematy polityki wodnej
Język rosyjski	темы водной политики

Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	polityka wodna tematów (niepoprawnie)
Bing	woda tematy polityki (niepoprawnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	вода политика темы (niepoprawnie)
Yahoo!	темы политики воды (niepoprawnie)
Bing	вода политики темы (niepoprawnie)

Kolejną ważną grupą ponadgeneracji powstających jako wtórny efekt niezdolności modelu do rozpoznawania węzłów głównych i zależnych wewnątrz frazy nominalnej, które zawierają przydawki przymiotne, są zaburzenia w koordynacji węzłów. Ilustruje je poniższy materiał przykładowy:

18) all living and non-living things[33]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	wszystkie żywe i nieżywe rzeczy
Język rosyjski	все живые и не живые вещи

Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	życia i nie wszystkie żyjące rzeczy (niepoprawnie)
Bing	wszystkie rzeczy życia i non życia (niepoprawnie)

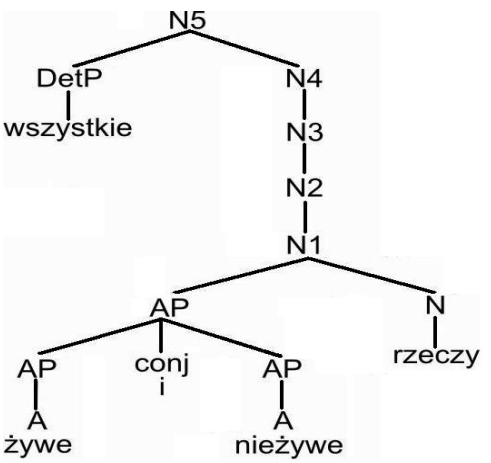
[32] <www.worldwatercouncil.org> (data dostępu: 22.05.2009).

[33] <en.wikipedia.org/wiki/Environment> (data dostępu:06.02.2009).

Wersje modeli rosyjskich omawianego ciągu:

Google	все живые и не живые вещи (poprawnie)
Yahoo!	все живущие и non-living вещи (niepoprawnie)
Bing	все вещи, живой и неживой (niepoprawnie)

Podobnie jak w przykładzie 17), za generację poprawnej wersji frazy odpowiedzialne są następujące reguły: [N5 DetP N4] blokująca zaburzenia w szyku determinatorów, [N4 (NumP) N3] blokująca zaburzenia w kolejności przyłączania liczebników [N3 N2 (PP)] regulująca pozycję grupy przyimkowej oraz [N2 N1 N6],i [N1 N AP] regulujące pozycje zależnych rzeczowników i przymiotników, które występują tu w składniowej roli przydawek rzeczownych i przymiotnych.



19) The world's second largest organization of physicists[34]
Proponowana postać konwersji na język polski i rosyjski:

Język polski	Druga największa światowa organizacja fizyków
Język rosyjski	Вторая крупнейшая мировая организация физиков

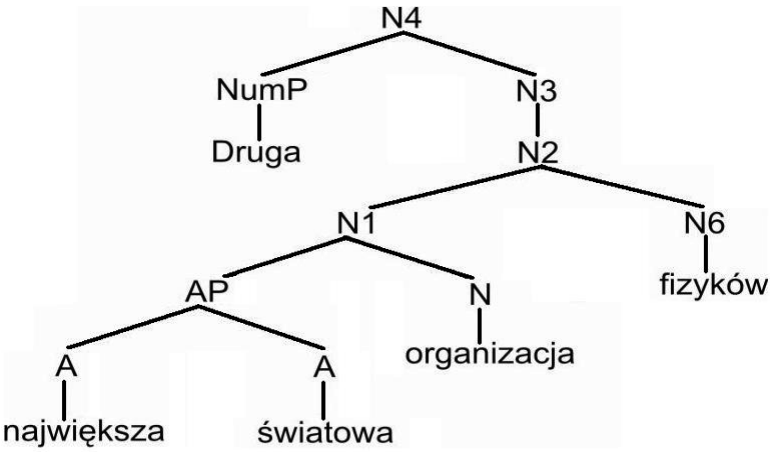
Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	Na świecie drugi co do wielkości (niepoprawnie)
Bing	Drugim największym światowym organizacji fizycy (niepoprawnie)

[34] <www.asp.org> (data dostępu: 26.05.2009).

Wersje modeli rosyjskich omawianego ciągu:

Google	Вторая крупнейшая мировая организация физиков (poprawnie)
Yahoo!	Организация мира второй по величине физиков (niepoprawnie)
Bing	Второй по величине организации мира физиков (niepoprawnie)



Za poprawną generację tej frazy odpowiadają reguły: [N4 NumP N3], która blokuje zaburzenia w szyku liczebników; następnie [N3 N2 (PP)], [N2 N1 N6] odpowiadające za poprawne przypisanie rzeczownikom składowej funkcji przydawki rzeczownej oraz [N1 AP N] i [AP A A], które odpowiadają za selekcję przydawek przymiotnych. Ponadto wydaje się, że aplikacja reguł koordynacji skutecznie ograniczałaby zaburzenia w szyku spójników. Kwestią zasadniczą na tym poziomie ustalania szyku elementów frazy jest zdolność modelu do rozpoznania właściwej pozycji spójnika, czyli do aplikacji reguł koordynacji węzłów.

Słuszności tej propozycji dowodzi również poniższy przykład zaburzenia konwersji:

20) Presidential politics and political news[35]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	Polityka prezydencka i wiadomości polityczne
Język rosyjski	Президентская политика и политические новости

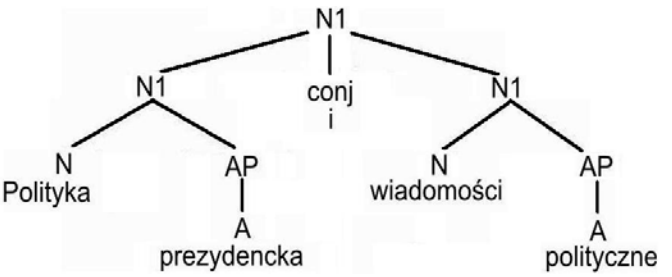
Wersje konwersji omawianej frazy w modelach języka polskiego:

[35] <www.foxnews.com/politics/index.html> (data dostępu: 17.02.2009).

Google	Prezydencki polityka informacyjna i politycznych (niepoprawnie)
Bing	Prezydencki Ustrój polityczny i politycznych grup dyskusyjnych (niepoprawnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	Президентская политика и политические новости (poprawnie)
Yahoo!	Президентская политика и политическая новость (składniowo poprawnie, morfologicznie niepoprawnie)
Bing	Президента политика и политические новости (niepoprawnie)



2.3.3. Zaburzenia relacji N2 z grupą przyimkową na poziomie N3

Problemy identyfikacji związków składniowych wewnątrz frazy, tj. między składnikiem konstytutywnym (tu: rzeczownik główny) a pozostałymi składnikami grupy nominalnej, komplikują się jeszcze bardziej, kiedy w otoczeniu przyrzeczownikowym pojawia się grupa przyimkowa. Model nie rozpoznaje statusu przynależności przyimka, a jego obecna selekcja i klasyfikacja do jednego z węzłów frazy nominalnej wydaje się przypadkowa. Zjawisko, któremu skutecznie przeciwdziałałyby reguły [N3 N2 PP], [N2 N1 N6], przypisujące przyimkom odpowiednie miejsce w otoczeniu przyrzeczownikowym, ilustruje przykład 21), natomiast reguły koordynacji skutecznie blokujące zaburzenia szyku spójników wewnątrz fraz nominalnych ilustruje przykład 22).

21) People`s problems with city government[36]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	Problemy ludzi z zarządem miasta
Język rosyjski	Проблемы людей с правительством города

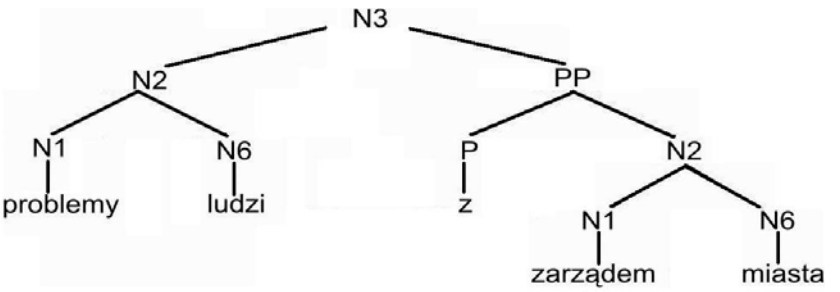
[36] <pubadvocate.nyc.gov> (data dostępu: 07.05.2009).

Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	Ludzie z problemami miasta rządu (niepoprawnie)
Bing	Ludzie problemy z rządem miasta (niepoprawnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	Проблемы народа с правительством города (poprawnie)
Yahoo!	Проблемы людей с городским правительством (akceptowalnie)
Bing	Проблемы людей с правительством города (poprawnie)



22) Healthy and Problematic Expectations in Relationship[37]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	Zdrowe i problematyczne oczekiwania w relacji/związku
Język rosyjski	Здоровые и проблематичные ожидания в отношениях

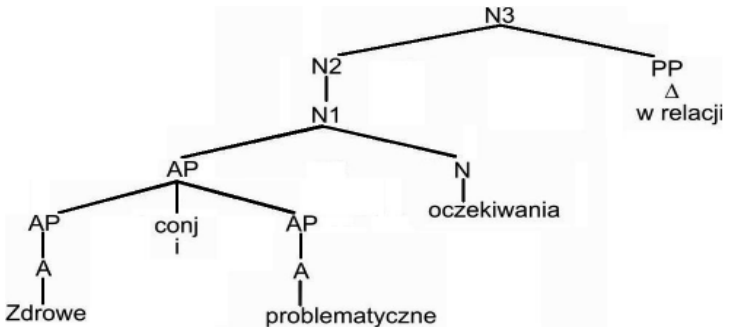
Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	Zdrowe i problemów Oczekiwania w relacji (niepoprawnie)
Bing	Zdrowe i problematyczna oczekiwania w relacji (składniowo poprawnie, morfologicznie niepoprawnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	Здоровые и проблематичные ожидании в отношениях (składniowo poprawnie, morfologicznie niepoprawnie)
Yahoo!	Здоровые и проблемные ожиданности в отношении (składniowo poprawnie, morfologicznie niepoprawnie)
Bing	Здоровый и проблематичные ожидания в связи (składniowo poprawnie, morfologicznie niepoprawnie)

[37] <cmhc.utexas.edu/healthyrelationships.html> (data dostępu: 28.05.2009).



Analogiczną strukturę na poziomie nieterminalnym, tym razem bez aplikacji reguł koordynacji, posiada fraza:

23) Outside Pressures on the Relationship[38]
Proponowana postać konwersji na język polski i rosyjski:

Język polski	Zewnętrzne presje/naciski na relację/związek
Język rosyjski	Внешние давления на отношения

Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	Poza presji na Powiązania (niepoprawnie)
Bing	Poza oddziaływań na relacji (niepoprawnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	Давление извне на взаимоотношения (akceptowalnie)
Yahoo!	Внешние давления на отношении (składniowo poprawnie, morfologicznie niepoprawnie)
Bing	За пределами давление на отношения (niepoprawnie)

Bez zdolności rozpoznawania przydawek przymiotnych i rzeczownych oraz bez umiejętności przypisywania hierarchii poszczególnym elementom rozbudowanej frazy nominalnej w dalszym ciągu model generować będzie następujące dewiacje składniowo-semantyczne:

24) The oldest and the largest library association in the world[39]
Proponowana postać konwersji na język polski i rosyjski:

[38] Tamże.
[39] <www.ala.org> (data dostępu: 26.05.2009).

Język polski	Najstarsze i największe stowarzyszenie bibliotek na świecie
Język rosyjski	Старейшая и крупнейшая ассоциация библиотек в мире

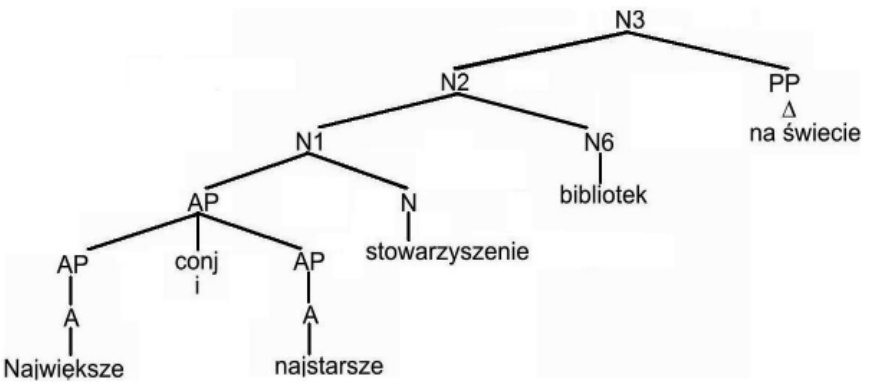
Wersje konwersji omawianej frazy w modelach języka polskiego:

Google	Najstarsza i największa biblioteka stowarzyszenia na świecie (niepoprawnie)
Bing	Najstarszą i stowarzyszenia biblioteki największy na świecie (niepoprawnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	Старейшая и крупнейшая библиотека ассоциации во всем мире (składniowo poprawnie, morfologicznie niepoprawnie)
Yahoo!	Самая старая и самая большая ассоциация архива в мире (składniowo poprawnie, morfologicznie niepoprawnie)
Bing	Старейшим и крупнейшим объединением библиотеки в мире (składniowo poprawnie, morfologicznie niepoprawnie)

Wspomniane wyżej typy zaburzeń kumulują się w rozbudowanych strukturach frazowych z wielostopniową koordynacją węzłów i podmiotem szeregowym, które ilustruje przykład 25) oraz rozbudowaną grupą przyimkową — przykład 26).



25) The ideas and knowledge of indigenous peoples and their social, economic and political status[40]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	Pomysły i wiedza rdzennej ludności i ich status społeczny, ekonomiczny i polityczny
Język rosyjski	Идеи и знания коренных народов и их социальный, экономический и политический статус

[40] <www.cwis.org> (data dostępu: 22.05.2009).

Wersje konwersji omawianej frazy w modelach języka polskiego:

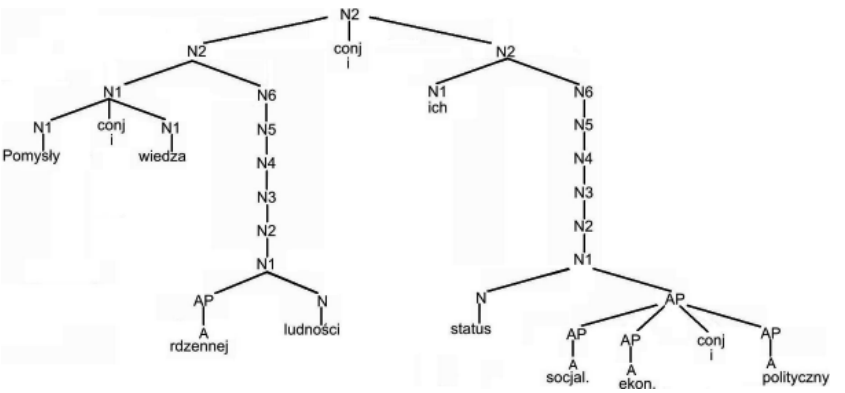
Google	Myśli i wiedzy rdzennej ludności i społecznej, statusu ekonomicznego i politycznego (niepoprawnie)
Bing	Pomysły i wiedzy na temat ludności tubylczej i ich statusu społecznego, gospodarczego i politycznego (niepoprawnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	Идей и знаний коренных народов и их социального, экономического и политического статуса (niepoprawnie)
Yahoo!	Идеи и состояние знания коренные народности и их социальное, хозяйственное/политическое (niepoprawnie)
Bing	Идей и знаний коренных народов и их социального, экономического и политического статуса (niepoprawnie)

Powyższy przykład zaburzenia konwersji dowodzi tego, w jakim stopniu obecne modele języków polskiego i rosyjskiego, wykorzystywane w wyszukiwarkach, nie radzą sobie z poprawnym przekładem z języka angielskiego fraz nominalnych z podmiotem szeregowym. W tym przypadku istotna jest nie tylko aplikacja odpowiednich reguł — równie ważna jest kolejność ich wdrażania. Po pierwsze, model powinien zastosować regułę koordynacji węzłów: $\alpha \rightarrow \alpha 1$ conj αn dla węzłów kategorii frazowych N2 i AP. Po drugie, powinien uruchomić regułę [N2 N1 N6], definiującą właściwe miejsce przydawek rzeczownych. Oba zabiegi, których kolejność nie powinna być aleatoryczna, skutecznie przeciwdziałałyby tego typu zaburzeniom.

Poprawnie wygenerowana struktura wyjściowa dla omawianej frazy nominalnej powinna w modelu języka polskiego przyjąć następującą postać:



Powyższe rozważania kończy przykład zaburzonej konwersji rozbudowanej frazy nominalnej złożonej z grupy rzeczownikowej i przyimkowej. Oba modele, tj. model języka polskiego i rosyjskiego, w zależności od typu wyszukiwarki, generują bądź ciągi dewiacyjne, bądź ciągi o słabym stopniu akceptowalności.

26) Irish national centre for development of best practice in public administration[41]

Proponowana postać konwersji na język polski i rosyjski:

Język polski	Irlandzkie krajowe centrum rozwoju najlepszych praktyk w administracji publicznej
Język rosyjski	Ирландский центр по разработке передовой практики в области государственного управления

Wersje konwersji omawianej frazy w modelach języka polskiego:

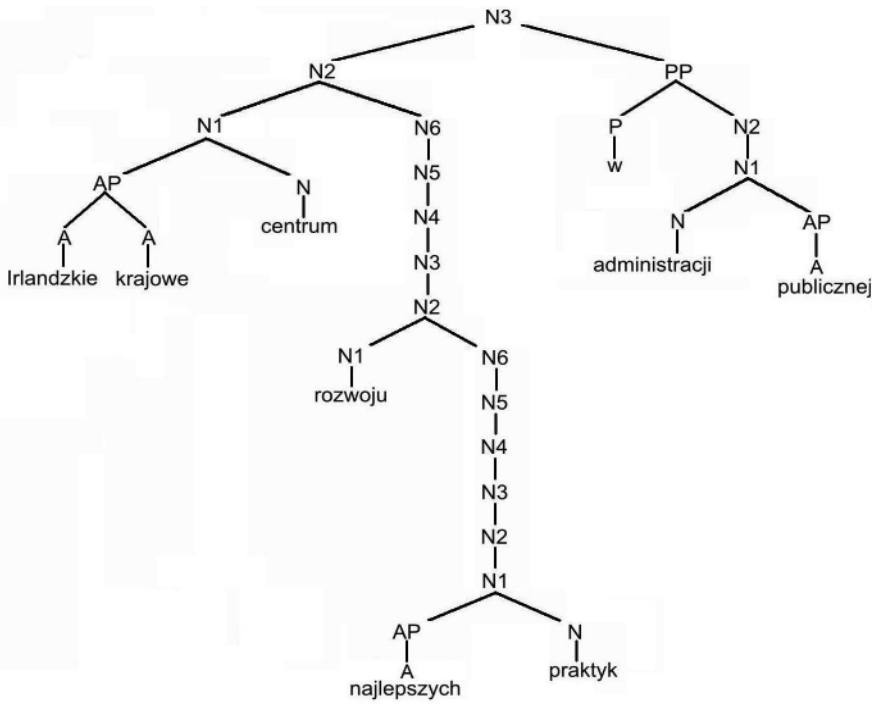
Google	Obywatelstwo irlandzkie centrum rozwoju najlepszych praktyk w administracji publicznej (niepoprawnie)
Bing	Irish krajowe centrum na rzecz rozwoju najlepszych praktyk w administracji publicznej (niepoprawnie)

Wersje modeli rosyjskich omawianego ciągu:

Google	Ирландский центр по разработке передовой практики в области государственного управления (poprawnie)
Yahoo!	Ирландский национальный центр для развития передовой практики в государственном управлении (akceptowalnie)
Bing	Ирландский Национальный центр разработки наилучшей практики в области государственного управления (akceptowalnie)

Dla powyższej frazy, podobnie jak w przykładzie 25), reguły [N2 N1 N6] oraz [N1 N AP] i [N1 AP N] skutecznie ograniczałyby problemy identyfikacji podmiotu (N2) i jego członów zależnych (N1 i N6). Jednocześnie zaś przykład 26) ilustruje tak charakterystyczną cechę modeli generatywnych, jak rekursja, czyli możliwość wielokrotnego rozwijania tego samego symbolu.

[41] <www.ipa.ie/-11k>.



3. Zakończenie

Niniejszym opracowaniem o charakterze wyłącznie teoretycznym włączam się do trwającej obecnie dyskusji nad mechanizmami precyzującymi automatyczną konwersję tekstów, proponowaną przez czołowe wyszukiwarki internetowe. Jestem świadoma stojących przed analitykami wyzwań oraz potrzeby dalszych badań nad rozwiązaniami proponowanymi w niniejszym artykule.

Pominęłam tutaj bardziej szczegółowe omówienie problematyki przekładu konstrukcji atrybutywności i posesywności z języka angielskiego na polski wymagającej osobnego opracowania. Mam jednak nadzieję, że przeprowadzenie dodatkowych badań o charakterze translatologicznym potwierdzi główną tezę tego artykułu.

Na podstawie prezentowanego materiału starałam się udowodnić potrzebę wprowadzenia do przyszłego zreorganizowanego modelu konwersji tekstu z języka angielskiego na polski reguł tzw. składni kategorii pośrednich (wzmocnionych) między węzłami fraz nominalnych, czyli NP i N. Zasadnicza bowiem różnica między proponowanymi strukturami fraz a dotychczasowymi grupami rzeczownikowymi konwertowanymi przez wyszukiwarki

polega na nieakceptowalności lub na słabym stopniu akceptowalności tych ostatnich. Składnia kategorii wzmocnionych pozwala na przełożenie formalnych informacji na temat struktury frazy nominalnej, tj. identyfikacji członów głównych i zależnych, hierarchizacji elementów składowych oraz koordynacji węzłów na informacje o charakterze dystrybucyjnym, jakimi są pozycje przyrzeczownikowe jednostek tekstowych występujących w składniowej roli przydawek rzeczownych i przymiotnych. Zatem są to informacje dotyczące zachowania poprawnego szyku elementów frazy nominalnej, co na obecnym etapie rozwoju technologii informatycznych dla omawianych wyszukiwarek stanowi poważny problem. Dodatkową zaletą proponowanego modelu jest jego rekursywność, ponieważ ograniczona liczba reguł bazowych umożliwia konwersję nawet bardzo rozbudowanych grup rzeczownikowych. Naturalnie, przyjęcie tej tezy wymaga dodatkowych badań, zwłaszcza na poziomie organizacji słownika danego modelu. Jak wspomniano wyżej, taki leksykon, będący drugim, równoważnym komponentem przyszłego modelu, powinien zawierać dokładny opis cech kategorialnych, czyli informacje na temat możliwości i sposobu łączenia ze sobą jednostek frazowych. Modele generatywne ekstrahują tego typu informacje z reguł bazowych komponentu gramatycznego i leksykalnego. Konwersja poprawnych związków składniowych jest bowiem uwarunkowana ścisłą wymianą danych między tymi dwoma komponentami, z których prymarny jest jednak zbiór reguł derywujących struktury frazowe. Prezentowany artykuł wpisuje się w początek badań nad jego rewizją.

Ewelina Alwasiak

СИНТАКСИЧЕСКИЕ НАРУШЕНИЯ ВНУТРИ НОМИНАЛЬНЫХ ФРАЗ В КОМПЬЮТЕРНОМ ПЕРЕВОДЕ С АНГЛИЙСКОГО ЯЗЫКА НА ПОЛЬСКИЙ И РУССКИЙ. ГЕНЕРАТИВНАЯ ТОЧКА ЗРЕНИЯ

Резюме

Настоящая работа является очередной в ряду статьей, предвиденных в рамках подготовки большего исследовательского проекта. Главная цель настоящего анализа — попытка создать механизмы по улучшению качества компьютерного перевода, производимого интернет-поисковиками. Автор работы старается также ответить на вопрос, почему синтаксические нарушения внутри номинальных фраз появляются в процессе автоматического перевода поисковика Google с английского языка на польский. С целью расширения области сравнения анализируются также результаты перевода на русский, производимого поисковиками Google, Yahoo и Bing. В работе, для этой области исследований, представлена генеративная точка зрения (теория *x-bar syntax*) как предложение одного из потенциальных способов избегнуть такого рода нарушений синтаксиса в компьютерном переводе.

Ewelina Alwasiak

SYNTACTIC DISORDERS IN NOMINAL PHRASES IN AUTOMATIC TRANSLATION FROM ENGLISH TO POLISH AND RUSSIAN LANGUAGES. THE GENERATIVE POINT OF VIEV

Summary

Presented article is the next one of several planned presentations of a larger research project. The overall goal of the presented analysis is to contribute toward improvement of the fully automatic machine translation integrated within search engines. The author tries to answer the question why language syntactic disorders occur in the process of automatic translations of phrases by the Google search mechanism. For comparison selected version of translation — Google, Yahoo and Bing has been analyzed. In the presented article a generative view (the theory of *x-bar syntax*) on this method theme is proposed as a possible way to avoid such kind of disorders in computer translations.