AUTOREFERAT

Maciej Ogrodniczuk

Weryfikacja korpusu wypowiedników polskich
(z wykorzystaniem gramatyki formalnej Świdzińskiego)

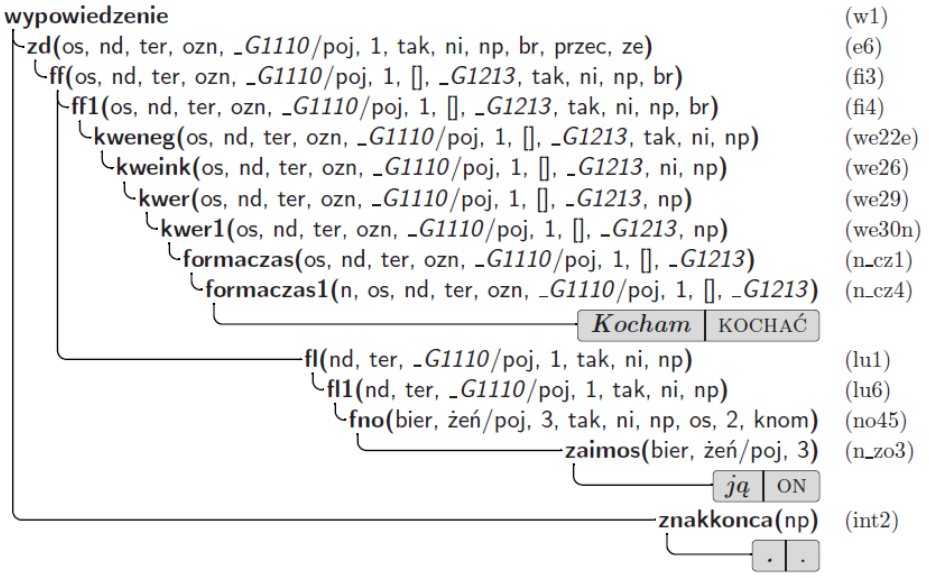
Rozprawa doktorska przygotowana w Katedrze Lingwistyki Formalnej
na Wydziale Neofilologii Uniwersytetu Warszawskiego
pod kierunkiem dra hab. Janusza S. Bienia, Warszawa 2006.

1. Założenia

Głównym celem pracy była weryfikacja danych tzw. korpusu wypowiedników polskich (czyli zdań i oznajmień) stworzonego pod kierunkiem prof. Marka Świdzińskiego i ręcznie uzupełnionego o oznaczenia struktury gramatycznej. Jako narzędzie wykorzystałem do tego zadania programy opracowane przez dra Marcina Wolińskiego: analizator morfologiczny Morfeusz i analizator syntaktyczny Świgra oparty na gramatyce formalnej Świdzińskiego, nazywanej dalej GFJP od tytułu książki *Gramatyka formalna języka polskiego*, w której została ona sformułowana.

Metoda weryfikacji — oprócz analizy danych zastanych — polegała przede wszystkim na konfrontacji dwóch opisów danego zbioru wypowiedzeń polskich: oryginalnych strukturalizacji dodanych ręcznie oraz rozbiorów, których dokonywałem automatycznie. W zadaniu tym gramatyka Świdzińskiego została po raz pierwszy wykorzystana do automatycznej analizy autentycznych tekstów polskich.

Gramatyka Świdzińskiego opiera się na koncepcji składników bezpośrednich, wynik analizy daje się więc przedstawić w postaci sparametryzowanych drzew składników (po prawej stronie wypisane są symbole reguł wykorzystanych do stworzenia drzewa):



Sama gramatyka liczy przeszło 460 reguł postaci:

```
ff(Wf, A, C, T, Rl, O, Wym, K, Neg, I, z(SwZ, NZ), Ow)
--> s(fi3),
    ff1(Wf, A, C, T, Rl, O, Wym, K, Neg, I, z(SwZ, Z), Ow),
    { zrozne( Z, ['p','px','pxx','pz'], NZ) },
    fl(A, C, Rl, O, Neg, ni, z(SwZ1, Z1)),
    { zrowne( Z1, [np], SwZ1 ) }.
```

Reguła ta mówi, że fraza finitywna może być realizowana jako fraza finitywna właściwa (`ff1`) i fraza luźna (`fl`), o ile spełnione są warunki zapisane w nawiasach klamrowych. Parametrami fraz są tradycyjne i nietradycyjne — wprowadzone przez Świdzińskiego — kategorie gramatyczne, takie jak wyróżnik fleksyjny (`Wf`), aspekt (`A`), czas (`C`), tryb (`T`), łączna kategoria rodzaju–liczby (`Rl`), korelatywność (`K`), negacja (`Neg`), inkorporacja (`I`) czy specjalna kategoria ograniczenia wewnętrznego (`Ow`).

2. Korpus wypowiedników

Powstały w latach 1993–96 korpus wypowiedników ma postać bazy danych liczącej 6721 rekordów zaopatrzonych w dodany ręcznie opis składniowy, w założeniu oparty na gramatyce formalnej Świdzińskiego. Teksty do analizy zostały wybrane z korpusu słownika frekwencyjnego, dziś dostępnego m. in. w Oxford Text Archive jako *korpus polszczyzny lat sześćdziesiątych XX wieku*.

Negatywną konsekwencją przyjętego sposobu reprezentacji tekstów w bazie jest brak bezpośredniego powiązania między wypowiednikami pochodzącymi z jednego wypowiedzenia; co więcej, odzyskanie tej informacji nie jest trywialne, gdyż jedyną wskazówką wspólnego pochodzenia grupy wypowiedników jest ta sama informacja lokalizacyjna źródła. Sama struktura wypowiedzenia jest ponadto wyróżniana w korpusie na dwa sposoby: poprzez podział na wypowiedniki oraz zapis struktury frazowej wewnątrz wypowiednika.

Niedogodność tę usunąłem poprzez automatyczne połączenie tekstów wypowiedników w wypowiedzenia metodą dopasowania fragmentów tekstów, a następnie zapisanie tak utworzonego zbioru w formacie XML. Powstały korpus, nazywany przeze mnie *korpusem wypowiedzeń*, liczący 3489 wypowiedzeń i 35556 słów, stał się jednym z wyników pracy i przedmiotem dalszej analizy.

Oprócz informacji o rozbiorach zdań korpus zawiera bogaty zasób pomocniczych danych składniowych, które wyekstrahowałem i opisałem w sposób rozszerzający oryginalny opis Świdzińskiego. Za najważniejsze informacje tego rodzaju uważam listę schematów zdaniowych (czyli „wzorców zdania lub oznajmienia elementarnego" w postaci typologii fraz wymaganych przez frazę finitywną) oraz informację o szyku składników zdania elementarnego, potwierdzającą dla polszczyzny typowość układu SVO.

Korpus jest też źródłem licznych materiałów pochodnych, z których za najważniejszy uważam słownik czasowników z informacją składniową zapisany w formacie zgodnym z wcześniejszym projektem Świdzińskiego opartym o analizę tradycyjnych słowników. Słownik zapisany przeze mnie powstał w odmienny sposób — metodą ekstrakcji danych z korpusu.

3. Wstępna weryfikacja korpusu

Weryfikacja warstwy typograficznej i morfologicznej Pierwszym etapem prac weryfikacyjnych było zagwarantowanie kompletności tekstu, spełnione w wyniku konfrontacji korpusu wypowiedników z innymi wariantami korpusu źródłowego, mianowicie tzw. wersją kanoniczną oraz wersją przygotowaną na potrzeby korpusu IPI PAN.

W procesie weryfikacji poprawności typograficznej próbek dokonałem sprawdzenia tak warstwy znakowej tekstu (postać interpunkcji, spacjowanie, parzystość różnych typów nawiasów używanych do wyróżniania składni, numeracja fraz itp.), jak również opisu parametrów (zgodność użytych wartości z dozwolonym zestawem).

Weryfikacja morfologiczna polegająca na skierowaniu tekstów próbek do analizy morfologicznej przez analizator Morfeusz doprowadziła w pierwszym rzędzie do wykrycia dodatkowych usterek typograficznych, a ponadto umożliwiła uzupełnienie słownika analizatora o formy pominięte. Dla słów nie rozpoznanych docelowe kody morfologiczne, niezbędne dla dalszego procesu analizy składniowej dopisałem ręcznie lub pozyskałem z wersji korpusu używanej w IPI PAN.

Weryfikacja składników frazowych Inicjalny etap poprzedzający analizę tekstów pełnych wypowiedników stanowiła analiza składniowa ich części wyodrębnionych strukturalnie w oryginalnej wersji korpusu. Weryfikację tego rodzaju przeprowadziłem dla wszystkich oznaczonych w korpusie typów składników: fraz finitywnych, podmiotowych, wymaganych, luźnych i tzw. członów innych. Z racji braku odpowiedniości między tą klasyfikacją a elementami gramatyki Świdzińskiego jednym z etapów procesu weryfikacji było jej ustalenie; w przypadku członów innych sprowadziło się ono do sprawdzenia wszystkich ich potencjalnych interpretacji, zakończonego prawie zawsze akceptacją względem co najmniej jednej jednostki GFJP.

Zakres opisu składniowego i wstępne wyniki analizy składniowej Ze względu na własności GFJP przeprowadzenie systematycznej analizy pełnych tekstów korpusu okazało się celowe wyłącznie dla pewnego podzbioru próbek, w całości poddających się opisowi. Odwołując się do własności gramatyki, z zestawu tego należało wykluczyć wszystkie wypowiedniki niezdaniowe oraz wypowiedniki zdaniowe zawierające składniki innego rodzaju niż zakładane przez projekt opisu (m. in. niektóre typy fraz nieciągłych, elips, konstrukcji porównawczych, ponadto mowę niezależną czy znaki interpunkcyjne spoza akceptowanego zestawu, takie jak średniki, pauzy, dwukropki czy wielokropki niekońcowe). Po przyjęciu tego założenia analizie mogły zostać poddane 4242 wypowiedniki o łącznej liczbie 43802 słów.

„Podstawowa” analiza składniowa, rozumiana jako zastosowanie oryginalnej gramatyki i aparatu analitycznego wzbogaconego jedynie o brakujące formy morfologiczne do korpusu tożsamego z oryginalnym, różniącego się jedynie korektą błędów typograficznych, dała w wyniku akceptację tylko ok. 31% wypowiedników.

4. Rozszerzenie gramatyki i usprawnienie analizy

W oryginalnej wersji GFJP nie są reprezentowane konstrukcje liczebnikowe oraz współrzędne grupy składniowe, w związku z czym niezbędne było dokonanie odpowiednich uzupełnień. Rozszerzenie GFJP o reprezentację liczebników polegało na zapisie reguł definiujących składnię wewnętrzną frazy liczebnikowej oraz uzgodnienia wewnętrzne i zewnętrzne frazy liczebnikowo-nominalnej. Grupy składniowe obejmują z kolei konstrukcje równorzędne różnych typów (nominalne, przymiotnikowe, przysłówkowe i przyimkowo-nominalne). Przy formułowaniu reguł wykorzystano w szczególności publikacje Saloniego, Szpakowicza, Świdzińskiego i innych.

Z obliczeniowego punktu widzenia ważnym etapem prac była eliminacja w gramatyce pięciu rodzajów cykli — zestawów zapętlonych reguł zawierających pojedynczy nieterminal w nagłówku i treści. Zadanie to zostało wykonane poprzez wprowadzenie wspólnego nieterminala dla każdego cyklu, z zachowaniem znaczenia pierwotnych reguł dzięki użyciu dodatkowego parametru.

Zostały zrealizowane również inne niezbędne rozszerzenia, takie jak reprezentacja grup apozycyjnych, form gerundialnych z *się*, zanegowanych fraz, rozszerzenie zakresu frazy luźnej o złożenia współrzędne, obsługa konstrukcji przysłówkowych z *coraz* czy *za*.

Wymienione uzupełnienia gramatyki, przede wszystkim w zakresie frazy liczebnikowej, pozwoliły na poszerzenie zestawu wypowiedników objętych opisem do 4716.

Własności gramatyki wynikowej W związku z niebezpieczeństwem nadmiernego rozluźnienia gramatyki w wyniku jej rozszerzania, równolegle z procesem optymalizacji GFJP prowadziłem testy weryfikujące identyczność jej siły wyrazu w porównaniu z gramatyką oryginalną; do tego celu wykorzystałem zbiór przykładów testowych używany wcześniej w pracach Wolińskiego. Zbiór ten, opracowany przez Świdzińskiego razem z gramatyką, jest bardzo przydatny dla badania adekwatności rozwijanej gramatyki ze względu na obecność w zestawie przykładów testowych zdań niepoprawnych, z definicji nie reprezentowanych w korpusie wypowiedników. Przeprowadzone testy wykazały daleką zgodność obu wersji gramatyki.

Rozbudowa słownika wymagań składniowych W ramach modyfikacji mechanizmu analizy dokonałem uzupełnienia słownika wymagań składniowych o 111 nowych zestawów wymagań (z czego 73 dotyczy czasowników nie występujących wcześniej w tym słowniku). Ograniczyłem się wyłącznie do wymagań niezbędnych do zapewnienia akceptacji odpowiednich zdań.

Końcowe wyniki analizy Ostatecznie wyniki analizy przedstawiają się w sposób następujący:

Wypowiedniki	**Liczba**	**Udział %**
akceptowane	3969	84,16 %
nieakceptowane	732	15,52 %
o czasie analizy przekraczającym 8 h	15	0,32 %

Jak widać, wyniki te są znacznie lepsze od uzyskanych z wykorzystaniem pierwotnej wersji gramatyki.

Łączny czas analizy zdań na komputerze PC z procesorem Intel Pentium 4M, taktowanym z częstotliwością 2.2 GHz, i z 1 GB pamięci RAM wyniósł ponad 50 godzin, jednak mediana czasów analizy nie przekracza 0,27 sekundy — innymi słowy, tyle samo zdań jest analizowanych w czasie krótszym, ile w czasie dłuższym od tej wartości. Zdania, dla których czas analizy przekroczył przyjętą arbitralnie wartość graniczną zawierają konstrukcje wielokrotnie złożone współrzędnie (w rodzaju *„Wszystko, co dzieje się wokół nas, wszystko, co słyszymy i widzimy, wszystko, co dzieje się z nami, z naszymi bliskimi i znajomymi, z tym, co do nas należy, z czym jesteśmy związani, jest odniesione do naszych nastawień.”*).

5. Udostępnione narzędzia informatyczne

Jednym z wyników pracy jest udostępnione na dołączonej płycie CD kompletne środowisko analizy składniowej dla systemu Microsoft Windows — *Świgra Live*. Oprócz analizatora składniowego Świgra w skład środowiska wchodzą: rozszerzona wersja GFJP będąca wynikiem pracy, korpusy wypowiedników i wypowiedzeń wraz z mechanizmem do ich przeglądania, dodatkowe narzędzia oraz wyniki testów i innych operacji, jak również zestaw stron WWW ze skrótową informacją o korpusie wypowiedników, gramatyce Świdzińskiego oraz analizatorze, a także instrukcją korzystania z mechanizmu analizy.

Dołączone programy mogą zostać wykorzystane do przetwarzania i masowej analizy składniowej innych korpusów, udostępniając m. in. takie funkcje jak możliwość wyboru formatu wyniku, sortowanie drzew wynikowych, ograniczanie czasu analizy czy wykrywanie reguły początkowej dla wyrażeń nie będących pełnymi wypowiedzeniami.

6. Podsumowanie

Głównym celem stworzenia korpusu wypowiedników była ręczna weryfikacja GFJP, zatem wyniki uzyskane przez użycie GFJP do automatycznej weryfikacji korpusu stanowią w istocie weryfikację tej weryfikacji.

Okazało się, że zakres i stopień szczegółowości opisu gramatycznego w GFJP i korpusie wypowiedników różnią się zasadniczo; oprócz ograniczonej odpowiedniości typów składników warto wspomnieć nie uwzględnione w gramatyce, a liczne w korpusie nieciągłości fraz czy sposób ich parametryzacji — w korpusie ograniczony do wskazania schematu zdaniowego z jego zgrubną charakterystyką, zaś w gramatyce formalnej opisany w sposób ścisły, niezbędny do przeprowadzenia pełnej analizy składniowej. Niezmiernie ważną różnicę stanowi także dokonywane w trakcie analizy ręcznej niejawne ujednoznacznianie analiz, zaniedbujące problem wieloznaczności.

Ręczna weryfikacja okazała się, moim zdaniem, zawodna i nieskuteczna, mimo to właściwy obiekt weryfikacji — gramatyka — po uwzględnieniu rozszerzeń okazuje się bardzo dobrym opisem rzeczywistych tekstów polskich.