Wstęp

Celem niniejszej pracy jest przedstawienie formalnego opisu składni obszernego podzbioru języka polskiego, który byłby przydatny w zadaniach przetwarzania języka naturalnego. Opis ten został zaimplementowany w postaci automatycznego analizatora składniowego Świgra 2, a jego weryfikację stanowi korpus składniowy Składnica.

Gramatyka formalna opisuje pewien zbiór zdań – język formalny. Zastosowana do opisu języka naturalnego może wskazywać, które zdania są poprawne, a które nie. Jednak celem stworzenia gramatyki dla języka naturalnego nie jest uzyskiwanie odpowiedzi binarnych. Dużo istotniejsze jest to, że gramatyka jawnie lub niejawnie przypisuje wypowiedzeniu pewną strukturę, która ma je reprezentować. Językoznawca w istocie myśli o konstrukcjach składniowych za pomocą tych struktur. W kontekście metod komputerowych reprezentacja struktury składniowej stanowi także dane wejściowe dla dalszych etapów przetwarzania, przede wszystkim do stworzenia reprezentacji semantycznej.

Ponieważ polszczyzna jest językiem fleksyjnym, analizę składniową wypowiedzeń trzeba poprzedzić analizą fleksyjną (do której konieczny jest opis fleksji, czyli odmiany wyrazów). Rozdzielenie opisu na etapy ułatwia pracę, a w wypadku implementacji komputerowej ma też dodatkowe uzasadnienie techniczne. Dla zapewnienia efektywności przetwarzania warto stosować możliwie najprostsze środki, jako że z rosnącą siłą formalizmów rośnie też złożoność obliczeniowa. Dlatego do opisu fleksji warto zastosować efektywne techniki związane z automatami skończonymi, podczas gdy do opisu składni potrzebne jest zastosowanie formalizmu o większej sile wyrazu.

Przedstawiony tu opis abstrahuje od semantyki, a więc uwzględnia tyle ze struktury języka, ile da się opisać poprzez interakcje cech formalnych, a nie znaczeń. Jego przedmiotem jest „gra kształtów", a nie „gra znaczeń".

Jest to opis języka ogólnego w wariancie pisanym, z naciskiem na jego staranną, redagowaną odmianę. Celem nie jest jednak formułowanie zaleceń poprawnościowych, w szczególności wyłapywanie wypowiedzeń niepoprawnych, lecz opisanie jak największej liczby konstrukcji faktycznie pojawiających się w tekstach. Z tego punktu widzenia opłaca się opisywać niektóre konstrukcje niepoprawne (dotyczy to w szczególności sposobu używania przecinków przez typowych użytkowników języka).

Niniejsza książka mieści się w nurcie prac nad opisem fleksyjnym i składniowym języka polskiego, w które autor jest zaangażowany od kilkunastu lat. Wcześniejszym etapem tych prac była implementacja gramatyki formalnej GFJP Marka Świdzińskiego (1992). Jej wynikiem był automatyczny analizator składniowy Świgra 1. Dzięki niemu udało się pokazać, że opis Świdzińskiego ma spory poziom spójności. Jednak, mimo że gramatyka ta jest bardzo rozbudowana, oparty na niej analizator akceptuje niewielki odsetek zdań polskich (około 30%).

Dlatego powstała koncepcja rozwinięcia tego opisu gramatycznego, aby osiągnąć większy odsetek zdań poprawnie analizowanych, i opracowania za jego pomocą korpusu oznakowanego informacją składniową, czyli tzw. banku drzew (ang. *treebank*). Przedmiotem pracy Woliński (2004) była możliwie wierna implementacja GFJP, natomiast tematem niniejszych rozważań jest przedstawienie nowej gramatyki wolnej od pewnych niedoskonałości tamtej. W sensie technicznym gramatyka ta jest w całości napisana na nowo. Za kształt poszczególnych reguł i zastosowane rozwiązania techniczne odpowiada autor niniejszej pracy. Co więcej, niektóre zasady opisu zostały wyraźnie zmienione w stosunku do GFJP. W ten sposób narodziła się Świgra 2[1].

Opracowany automatyczny analizator składniowy Świgra 2 jest wystarczająco sprawny, aby było możliwe przetwarzanie dziesiątek tysięcy wypowiedzeń. Pozwoliło to zbudować korpus składniowy Składnica. Świgra 2 była używana także w innych pracach z dziedziny inżynierii lingwistycznej. Łukasz Dębowski stosował ją w pracach dotyczących automatycznej ekstrakcji schematów walencyjnych czasowników (Dębowski i Woliński 2007). Elżbieta Hajnicz wykorzystywała analizę składniową do hipotetyzowania ram semantycznych (Hajnicz 2011). Gramatyka Świgra 2 stała się też podstawą gramatyki POLFIE (Patejuk 2015).

Określenie „analiza składnikowa języka polskiego" użyte w tytule tej pracy jest bardzo ogólne. Jednak każde przedsięwzięcie mające na celu opis języka naturalnego jest w jakiś sposób ograniczone. Przedstawiony tu opis można postrzegać jako pewien etap na drodze do celu: z pewnością z czasem pojawi się kolejny opis, reprezentujący wyższy poziom dokładności. Warto też podkreślić, że różne partie przedstawianego materiału są przemyślane w różnym stopniu. Opis fleksyjny jest wynikiem długiej ewolucji i można go traktować jako w znacznym stopniu zweryfikowany zarówno z punktu widzenia leksykografa (por. Saloni *et al.* 2015), jak i gramatyka. Niektóre elementy opisu składni

[1] W powstałych na wcześniejszym etapie rozwoju nowej gramatyki wspólnych pracach z Markiem Świdzińskim posługiwaliśmy się nazwą GFJP2. Nowa gramatyka formalna, stanowiąca część analizatora Świgra 2, jest jednak na tyle różna od GFJP, że nazwa ta jest myląca. Nowa gramatyka zasługuje na nową nazwę. Niestety nazwa Świgra 2 zdążyła zacząć funkcjonować w szerszym obiegu i na jej zmianę jest za późno. W związku z tym na przedstawiony tu opis będę używał określenia *gramatyka Świgra 2*. Z technicznego punktu widzenia gramatykę tę można utożsamić ze zbiorem reguł, których używa analizator Świgra 2.

przedstawione w rozdziale 2 są zupełnie nowe i stanowią dopiero pierwsze przybliżenie opisu formalnego.

Zakładanym odbiorcą książki jest informatyk zainteresowany technikami przetwarzania języka naturalnego. Jej autor nie jest z wykształcenia językoznawcą, dlatego dość szczegółowo wprowadza zastosowany system pojęć z zakresu fleksji i składni oraz ilustruje tekst wieloma przykładami językowymi, które powinny pomóc czytelnikowi w intuicyjnym uchwyceniu znaczenia poszczególnych pojęć. Osoby, które odbiorą to jako nadmiar oczywistych przykładów, proszone są o wyrozumiałość. Być może książka zainteresuje także językoznawców tym, że pokazuje, jak informatyk widzi pojęcia językoznawcze oraz jaki poziom ścisłości jest konieczny, aby opis „działał" jako program komputerowy.

Sposób sformułowania reguł gramatycznych zastosowany w tej pracy można nazwać inżynierskim, nie jest on głównym jej przedmiotem. Ważniejszą sprawą jest pokazanie proponowanego opisu gramatycznego. Książka stanowi przez to dokumentację struktur zawartych w korpusie Składnica, które można traktować jako niezależne od programu komputerowego, który je wygenerował.

Struktura książki

Rozdział 1 przedstawia przyjęte zasady powierzchniowego dystrybucyjnego opisu fleksji języka polskiego. Opis ten wywodzi się z koncepcji Zygmunta Saloniego, zwłaszcza jego klasyfikacji leksemów polskich. Istotnym aspektem niniejszej pracy jest pokazanie, że opis ten może być dobrą podstawą opisu składni. Do tego rozdziału mogą sięgnąć użytkownicy analizatora fleksyjnego Morfeusz 2 SGJP, aby poznać szczegóły zastosowanego systemu znakowania i stojące za nimi motywacje.

Tematem rozdziału 2 są struktury składniowe (drzewa) przypisywane wypowiedzeniom polskim przez omawianą tu gramatykę. Rozdział zawiera systematyzację opisanych konstrukcji składniowych i pokazuje, co zostało objęte opisem. Mogą doń sięgnąć osoby zainteresowane dalszym przetwarzaniem struktur składniowych generowanych przez analizator Świgra 2, w szczególności przetwarzaniem danych z korpusu składniowego Składnica.

W rozdziale 3 przedstawiono wykorzystywany w analizie automatycznej słownik walencyjny Walenty. O ile poprzedni rozdział dotyczy głównie systematycznych własności składniowych (przysługujących leksemom należącym do dużych klas wyrazów, np. klas gramatycznych), to słownik walencyjny notuje własności składniowe uwarunkowane leksykalnie, a więc charakterystyczne dla poszczególnych leksemów.

Rozdział 4 poświęcony jest implementacji gramatyki. Przedstawiono w nim istotne rozszerzenie formalizmu Definite Clause Grammar (DCG), a następnie omówiono sposób użycia go do realizacji gramatyki. Zaprezentowano też

kilka mechanizmów analizatora Świgra 2, przede wszystkim mechanizm realizowania wymagań składniowych.

Celem rozdziału 5 jest umieszczenie przedstawionego tu opisu na tle innych opisów składniowych języka polskiego. Porównanie dotyczy zarówno warstwy językoznawczej, jak i technicznych aspektów sposobu wyrażenia poszczególnych opisów.

W rozdziale 6 zaprezentowano korpus składniowy Składnica, stanowiący przykład wdrożenia przedstawionego tu opisu do zanalizowania pewnego zbioru tekstów. Korpus składniowy można traktować jako dokumentację adekwatności opisu, ponieważ zawarte w nim struktury składniowe zostały wygenerowane automatycznie za pomocą analizatora Świgra 2, a następnie ujednoznacznione i zweryfikowane przez ekspertów.

Tematem rozdziału 7 są techniki statystycznego ujednoznaczniania analiz składniowych. Program komputerowy, trenowany na danych korpusu składniowego, ma za zadanie wykonać ujednoznacznienie podobnie, jak to robili eksperci budujący korpus. Uzupełnienie analizatora regułowego o taki moduł pozwala zbliżyć się do ideału, czyli sytuacji, w której program komputerowy wskazuje dla danego wypowiedzenia dokładnie jedną strukturę składniową.

Konwencje notacyjne

Przytaczane przykłady wypowiedzeń i ich fragmenty (w szczególności wykładniki tekstowe form wyrazowych) składane są kursywą, np.:

(1) *Książka ukazała się w odpowiednim momencie, w okresie dyskusji nad nowymi programami.* [Skł.]

(2) **Książka ukazała się odpowiednim momentem.*

Gwiazdka, jak w przykładzie (2), sygnalizuje niepoprawność gramatyczną przytoczonego wypowiedzenia. Przy numerowanych przykładach w nawiasach kwadratowych podawane jest źródło:

[Skł.] – wypowiedzenie z korpusu Składnica (por. p. 6.1),

[NKJP1M] – wypowiedzenie z ręcznie znakowanego podkorpusu Narodowego Korpusu Języka Polskiego (NKJP, nkjp.pl) o wielkości jednego miliona segmentów,

[NJKP300] – wypowiedzenie ze zrównoważonego wariantu NKJP (300 milionów segmentów),

[NKJP1800] – wypowiedzenie z pełnego NKJP (1 800 milionów segmentów),

[Walenty] – przykład ilustrujący schemat walencyjny cytowany za słownikiem Walenty (zob. rozdz. 3).

Brak oznaczenia sygnalizuje przykład własny (skonstruowany).

Identyfikatory leksemów podawane są kapitalikami (CZYTAĆ, WARSZAWA). Definicje stosowanych pojęć z dziedziny fleksji są przytoczone w początkowych punktach rozdziału 1, a składni – rozdziału 2.

Podziękowania

Za zainteresowanie mnie problemami przetwarzania języka naturalnego jestem wdzięczny prof. Januszowi S. Bieniowi, który był animatorem wielu prac wspomnianych w tej książce. Profesorowi Zygmuntowi Saloniemu jestem niezmiernie zobowiązany za to, że pokazał mi, że możliwe jest rygorystyczne podejście do opisu fleksji, jak również nauczył mnie, że wszystkie pojęcia, które próbujemy przypasować do rzeczywistości językowej, to tylko modele, a nie część tej rzeczywistości. Profesorowi Markowi Świdzińskiemu zawdzięczam podróż w krainę składni. Niniejsza książka nie powstałaby bez bardzo wielu godzin dyskusji z profesorem Świdzińskim, a przede wszystkim bez Jego gramatyki jako punktu odniesienia. Dziękuję również wszystkim członkom Zespołu Inżynierii Lingwistycznej w Instytucie Podstaw Informatyki za wsparcie mnie w wysiłku nad tą pracą.

Szczególne podziękowania należą się czytelnikom jej fragmentów: Markowi Świdzińskiemu, Łukaszowi Dębowskiemu, Elżbiecie Hajnicz, Witoldowi Kierasiowi, Alinie Wróblewskiej i wreszcie Tomaszowi Obrębskiemu, który jako pierwszy przeczytał całość, a jego wnikliwe uwagi pozwoliły usunąć wiele niedociągnięć tekstu.