# Przedmowa

## Cel, zakres, adresat

Celem niniejszej pracy jest przedstawienie formalnego opisu wybranych zjawisk języka polskiego oraz implementacji prototypu parsera (tj. analizatora składniowego) języka polskiego opartego na tym opisie. Praca ta adresowana jest przede wszystkim do następujących czterech grup czytelników:

- do osób zainteresowanych językiem polskim, jego składnią, zjawiskami z pogranicza składni i semantyki oraz z pogranicza składni i morfologii;
- do informatyków zainteresowanych automatycznym przetwarzaniem języków naturalnych, przede wszystkim języka polskiego;
- do matematyków i logików zainteresowanych zaksjomatyzowanym opisem języków naturalnych; oraz
- do lingwistów zainteresowanych współczesnymi teoriami lingwistycznymi.

Czytelnicy zainteresowani problemami składni języka polskiego znajdą tu nowe propozycje szczegółowych analiz zjawisk tak systematycznych i typowych dla języka polskiego, jak koordynacja, różnorakie związki zgody, czy nadawanie wartości przypadka, ale także opisy zjawisk, które nie zostały wcześniej tak szeroko opisane, takich jak zależności nielokalne w zdaniach względnych, dalekie uzgodnienie negacji, czy też koreferencja zaimków anaforycznych. Zjawiska te są opisane w sposób modularny i — mamy nadzieję — spójny: opisy poszczególnych zjawisk są w pewnej mierze niezależne, ale zasady gramatyczne postulowane w tych opisach odpowiednio współdziałają ze sobą.

Praca niniejsza zawiera także opis implementacji prototypu parsera języka polskiego opartego na zaproponowanych tutaj rozwiązaniach teoretycznych. Implementacja ta, udostępniona publicznie w Internecie, choć obejmuje zjawiska opisane w części teoretycznej, często odbiega od rozwiązań teoretycznych, zarówno z powodu pewnych różnic pomiędzy przyjętym tutaj formalizmem matematycznym i platformą implementacyjną, jak i z powodu czynników efektywnościowych.

Formalizm matematyczny wspomniany powyżej jest pewnym rodzajem logiki deskryptywnej (ang. *description logic*) z dobrze zdefiniowaną składnią i semantyką (por. pozycje cytowane w dodatku B). Choć w interesie dostępności niniejszej pracy dla czytelników mniej obeznanych z formalizmami logicznymi nie kładziemy w pracy dużego nacisku na formalizację, wszystkie zasady lingwistyczne zaproponowane w treści i częściowo sformalizowane w dodatkach do odpowiednich rozdziałów mogą być łatwo przetłumaczone na zdania języka tej logiki. Czytelnika polskiego zainteresuje zapewne fakt, że istnieją formalizmy będące połączeniem formalizmu wykorzystanego w niniejszej pracy z gramatykami kategorialnymi (Dörre i Manandhar, 1997).

Uważamy w końcu, że niniejsza praca powinna także zainteresować lingwistów zajmujących się współczesnymi teoriami lingwistycznymi, przede wszystkim zaś teoriami generatywnymi. Opis zjawisk języka polskiego zaproponowany tutaj opiera się na jednej z najbardziej dynamicznie rozwijających się teorii lingwistycznych, Head-driven Phrase Structure Grammar (HPSG). Prace nad opisem języków naturalnych w formalizmie HPSG prowadzone są (lub były) w licznych ośrodkach w Stanach Zjednoczonych (m.in., Stanford, Ohio, North Carolina, Buffalo, Texas), Niemczech (m.in., Saarbrücken, Tybinga, Bielefeld, Bochum, Jena), Wielkiej Brytanii (m.in., Londyn, Essex, Edynburg), Francji (Paryż, Lille), Holandii, Korei, Japonii, Turcji, Słowenii, Czechach i Bułgarii, zaś języki, nad którymi były prowadzone prace z zastosowaniem tego formalizmu to m.in.: angielski, niemiecki, holenderski, francuski, włoski, hiszpański, portugalski, serbsko-chorwacki, bułgarski, czeski, słoweński, walijski, koreański, japoński, turecki, grecki, hebrajski, warlpiri i amerykański język migowy. Choć praca niniejsza nie jest podręcznikiem HPSG, może ona być wykorzystana jako pomoc w uczeniu tej teorii, szczególnie na gruncie polskim.

Ponieważ niniejsza praca adresowana jest do tak różnych grup czytelników, staraliśmy się nie czynić zbyt daleko idących założeń dotyczących wiedzy czytelnika na temat informatyki, logiki czy lingwistyki. Pewna wiedza informatyczna dotycząca technik parsowania może być przydatna w rozdziale opisującym implementację, zaś znajomość technik obiektowych i podstawowych metod sztucznej inteligencji może nieco ułatwić zrozumienie formalizmu przyjętego w niniejszej pracy, nie jest jednak do tego niezbędna. W głównej części pracy nie zakładamy też znajomości systemów logicznych, poza zrozumieniem podstawowych operatorów logicznych (negacja, koniunkcja, dyzjunkcja, wynikanie, równoważność) i znajomością pojęć takich jak *zmienna* i *prawda*. Znajomość logiki predykatów powinna być w zupełności wystarczająca do zrozumienia formalizmu naszkicowanego w treści pracy. Choć bardziej precyzyjny opis formalizmu znajduje się w dodatku, czytelnik zainteresowany podstawami logicznymi przyjmowanego tu formalizmu powinien sięgnąć bezpośrednio do źródeł cytowanych w tymże dodatku.

Nieco większa powinna być natomiast wiedza lingwistyczna czytelnika. Zakładamy tutaj znajomość terminów lingwistycznych przynajmniej na poziomie szkoły podstawowej i średniej, choć niewątpliwie orientacja w zakresie precy-

zyjnych opisów polszczyzny (np. Saloni i Świdziński 1998 lub Bobrowski 1995, 1998) czy też w zakresie współczesnych teorii lingwistycznych (np. Bresnan 1982, 2000, Haegeman 1994, Webelhuth 1995, czy Sag i Wasow 1999) może ułatwić czytanie niniejszej pracy.

Należy tu podkreślić, że wielu ważnych terminów lingwistycznych, którymi posługujemy się w niniejszej pracy, nie definiujemy w sposób precyzyjny. Istnieją ku temu dwa powody. Po pierwsze, większość terminów używanych w lingwistyce jest tak wieloznaczna i niejasna, że konsekwentne wprowadzenie porządku terminologicznego wymagałoby napisania oddzielnej pracy.[1] Obawiamy się, że próba szczegółowego zdefiniowania wszystkich używanych tutaj terminów lingwistycznych tylko utrudniłaby czytanie niniejszej pracy. Po drugie, precyzyjne definiowanie używanych terminów jest niezbędne tylko w wypadku, tj. gdy należą one do języka, którym opisujemy pewne zjawiska, gdy należą one do języka teorii; gramatyki opisowe, posługujące się językiem naturalnym do opisania zjawisk składni czy semantyki, muszą w interesie precyzji definiować podstawowe pojęcia, którymi operują. Inaczej jednak wygląda sytuacja w wypadku gramatyki opisanej za pomocą pewnego formalizmu, tak jak to ma miejsce w niniejszej pracy. Tutaj pojęcia lingwistyczne mają znaczenie drugorzędne, służą tylko lepszemu zrozumieniu treści opisów formalnych i intuicyjnemu uzasadnieniu przyjętych rozwiązań; cała gramatyka jest natomiast opisana zdaniami języka formalnego, których interpretacja nie zależy od definicji poszczególnych terminów lingwistycznych.

Oczywiście opis zjawisk składniowych i z pogranicza składni zawarty w niniejszej pracy jest siłą rzeczy fragmentaryczny, podobnie jak w innych gramatykach języka polskiego (np. Świdziński 1992, Bobrowski 1995, 1998, Saloni i Świdziński 1998) i innych języków. Niemniej jednak różnorodność zjawisk, którym poświęciliśmy niniejszą pracę pozwala twierdzić, że metoda tu przedstawiona dobrze nadaje się do formalnego opisu polszczyzny.

## Układ pracy

Niniejsza praca zawiera dziesięć rozdziałów i dwa dodatki. Rozdział 1 stanowi krótki i raczej nieformalny opis formalizmu HPSG i teorii lingwistycznej HPSG. Następne dwa rozdziały, 2–3, zawierają liczne modyfikacje założeń teorii lingwistycznej HPSG dotyczących m.in. struktury słownika, struktur frazowych i zjawiska modyfikacji składniowej. Sens wielu z tych modyfikacji stanie się w pełni jasny dopiero w następnych rozdziałach, opisujących poszczególne zjawiska języka polskiego.

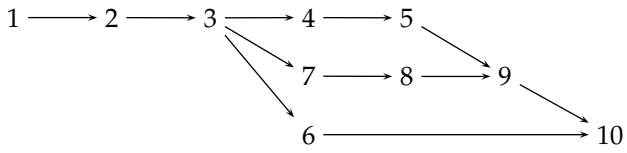
Rozdział 4 poświęcony jest różnorakim uzgodnieniom występującym w języku polskim; pokazujemy w nim m.in., że uzgodnienie podmiotu z tzw. orzeczeniem jest zjawiskiem z pogranicza składni i semantyki. Rozdział 5 opisuje

[1] Istnieją prace dyskutujące wielość interpretacji pojęć takich jak *zdanie* (np. Klemensiewicz 1963), *przypadek* (np. Mel'čuk 1986), *wyraz* (np. Bień i Saloni 1982) czy *język* (np. Chomsky 1986a).

składniowe reguły nadawania przypadka w języku polskim, z uwzględnieniem tak efemerycznych zjawisk jak nadawanie przypadka we frazach liczebnikowych i tzw. daleki dopełniacz negacji. Następny rozdział, 6, zawiera systematyczny opis zjawiska znanego w literaturze anglosaskiej jako *binding* (pol. *wiązanie*), polegającego na składniowym uwarunkowaniu interpretacji pewnej klasy zaimków (tzw. zaimków anaforycznych). Rozdział 7 opisuje mechanizm pozwalający na oddanie zjawisk nielokalnych, tj. takich, których zasięg wykracza poza proste zdanie, oraz ilustruje ten mechanizm szczegółową analizą konstrukcji względnych. Kolejny rozdział, 8, także wykorzystuje ten mechanizm, a mianowicie do analizy tzw. dalekiego uzgodnienia negacji. W rozdziale tym zaproponowane jest także potraktowanie negacji zdaniowej jako kategorii morfologicznej. Ostatni rozdział empiryczny, 9, poświęcony jest zjawisku koordynacji; zawiera on opis różnego rodzaju konstrukcji współrzędnie złożonych, a także uzgodnień, jakie zachodzą wewnątrz i „na zewnątrz" takich konstrukcji. Rozdział ten, stanowiący próbę spójnego opisu niezwykle skomplikowanego zjawiska, ma charakter bardziej wstępny od pozostałych rozdziałów; opis w nim zawarty będzie w przyszłości dalej rozwijany i modyfikowany. Dodatki do rozdziałów 1–9 zawierają podsumowanie opisów gramatycznych zawartych w tych rozdziałach.

Ostatni rozdział niniejszej pracy, 10, opisuje implementację eksperymentalnego parsera języka polskiego opartego na gramatyce formalnej przedstawionej w poprzednich rozdziałach, zaś dwa dodatki zawierają podsumowanie tej gramatyki formalnej (dodatek A) oraz krótki opis formalizmu logicznego leżącego u podstaw HPSG (dodatek B).

Poniższy rysunek oddaje graf zależności pomiędzy materiałem zawartym w poszczególnych rozdziałach. Wynika z niego w szczególności, że rozdziały 4–9 zakładają znajomość materiału rozdziałów 1–3, a rozdział 9 będzie w pełni zrozumiały dopiero po lekturze wszystkich innych rozdziałów, być może z wyjątkiem rozdziału 6. Ostatni rozdział niniejszej pracy, 10, opisujący różnice między teorią a implementacją, zakłada znajomość całej teorii opisanej we wcześniejszych rozdziałach.




Choć praca niniejsza powstała w wyniku zbiorowego wysiłku wszystkich autorów, głównymi autorami poszczególnych rozdziałów są: Anna Kupść — 8, 9; Małgorzata Marciniak — 6, 10; Agnieszka Mykowiecka — 7; Adam Przepiórkowski — pozostałe rozdziały i dodatki. Początek rozdziału 4 został częściowo oparty na wcześniejszych pracach Krzysztofa Czuby (Czuba, 1995, 1997; Czuba i Przepiórkowski, 1995).

## Podziękowania

Książka, którą oddajemy do rąk czytelników, jest wynikiem naszych wieloletnich prac nad zjawiskami języka polskiego, prac, które zaowocowały licznymi artykułami i wystąpieniami konferencyjnymi. Choć niniejszy tom stanowi oryginalną próbę *spójnego* opisu tych zjawisk, i rozwiązania w nim przedstawione często różnią się od rozwiązań zaproponowanych we wcześniejszych publikacjach, niewątpliwie uwagi do tych publikacji miały znaczący wpływ na jakość prezentowanej tu analizy. Chcielibyśmy za nie podziękować przede wszystkim prof. Markowi Świdzińskiemu (Uniwersytet Warszawski; UW), który przez kilka lat wiernie kibicował naszym pracom, a także następującym osobom: Anne Abeillé (Université Paris 7), Bob Borsley (University of Essex), Carl Pollard (Ohio State University), Frank Richter (Universität Tübingen), Ivan Sag (Stanford University), Manfred Sailer (Universität Tübingen).

Niniejsza praca wiele zyskała dzięki krytycznym uwagom i licznym sugestiom poprawek dr Magdaleny Derwojedowej (UW), dr Elżbiety Hajnicz (Instytut Podstaw Informatyki Polskiej Akademii Nauk; IPI PAN), Marcina Wolińskiego (IPI PAN) oraz Łukasza Dębowskiego (IPI PAN). Za komentarze do fragmentów pracy dziękujemy także Krzysztofowi Czubie (Carnegie Mellon University), dr. Krzysztofowi Szafranowi (UW), Monice Korczakowskiej (UW) i dr Petyi N. Osenovej (Bułgarska Akademia Nauk).

Nasze prace nad formalnym opisem polszczyzny i jego implementacją finansowane były przede wszystkim w ramach projektu Komitetu Badań Naukowych numer 8 T11C 011 10 „Logiczne podstawy inżynierii lingwistycznej" kierowanego przez prof. Leonarda Bolca (IPI PAN), a także, w mniejszym zakresie, w ramach grantu Unii Europejskiej CRIT-2. W czasie tych prac, Anna Kupść korzystała ze stypendium doktoranckiego rządu francuskiego na Université Paris 7 oraz ze stypendium Volkswagen Foundation (program CLaRK) na Universität Tübingen, zaś Adam Przepiórkowski korzystał ze stypendium doktoranckiego Deutsche Forschungsgemeinschaft (w ramach Graduiertenkolleg ILS) na Universität Tübingen.

*Anna Kupść*
*Małgorzata Marciniak*
*Agnieszka Mykowiecka*
*Adam Przepiórkowski*