Projekt KBN 8 T11C 002 13
Zestaw testów do weryfikacji i oceny analizatorów języka polskiego
Sprawozdanie merytoryczne

dr hab. Janusz S. Bień, prof. UW
kierownik projektu

luty 2000

1 Wprowadzenie

Wynikiem projektu jest załączona **płyta CD-ROM** pod tytułem ***Lingwistyczne zasoby polszczyzny*** zawierająca udokumentowane zestawy testów pozwalających na obiektywną weryfikację i ocenę analizatorów języka polskiego. **Płyta ta zawiera również niniejszy tekst** w formie plików `sprawozd.pdf` i `sprawozd.ps`.

Najbardziej nowatorski charakter ma zestaw wyników analizy syntaktycznej (ang. *treebank*), dalej nazywany bankiem rozbiorów gramatycznych, zawierający drzewa analizy syntaktycznej dla obszernego zbioru zróżnicowanych zdań; analizy te zostały dokonane zgodnie z formalną gramatyką języka polskiego autorstwa Marka Świdzińskiego (1992) — patrz s. 14 i 21. Choć prace tego typu są już od dłuższego czasu prowadzone dla języka angielskiego (Leech, Garside 1991) i od kilku lat np. dla języka czeskiego (Hajič 1998), dla polszczyzny dokonano dotąd jedynie ręcznej symulacji analizy syntaktycznej, przyporządkowującej zdaniom tylko najbardziej podstawowe własności składniowe (Świdziński 1993, 1996).

Przygotowanie tego zestawu okazało się o wiele trudniejsze niż oczekiwano. W celu zadowalającego rozwiązania napotkanych problemów stworzono niemal od podstaw nowy analizator syntaktyczny, co stanowi istotne rozszerzenie planowanego zakresu prac — patrz s. 15.

2 Geneza i cele projektu

2.1 Geneza

Opisywany projekt stanowi bezpośrednią i naturalną kontynuację projektu KBN nr 8 S503 032 27 pt. *Analizator morfologiczno-syntaktyczny dla obszernego pod-*

zbioru języka polskiego, realizowanego w latach 1994–1996 w Instytucie Informatyki UW pod kierunkiem dr. hab. Janusza S. Bienia z udziałem dr K. Szafrana — por. (Bień 1996, 1996a), (Szafran 1996). W projekcie wykorzystano również wyniki projektu KBN nr 1 1188 91 02 pt. *Słownik gramatyczny współczesnego języka polskiego*, realizowanego w latach 1992–1994 w Wyższej Szkole Pedagogicznej w Olsztynie pod kierunkiem prof. dr. hab. Zygmunta Saloniego, oraz projektu nr 1 P104 030 04 pt. *Ukierunkowana gramatycznie tekstowa baza danych: korpus wypowiedzeń współczesnej polszczyzny*, realizowanego w latach 1993-96 w Instytucie Języka Polskiego UW pod kierunkiem dr. hab. Marka Świdzińskiego (1996). Wszystkie wymienione wyżej osoby brały udział również w realizacji niniejszego projektu; należy przy tym podkreślić bardzo istotną rolę mgr Marcina Wolińskiego, który formalnie był tylko pracownikiem pomocniczym.

Z potrzeby sformułowania obiektywnych kryteriów oceny analizatorów składniowych zdawano sobie sprawę już dawno. Dla języka angielskiego pierwsze znane nam sformułowanie problemu pochodzi z 1983 r., a dla języka polskiego postawiono ten problem 2 lata później — por. (Bańko 1990) s. 56 i 59. W miarę rozwoju lingwistyki komputerowej i inżynierii językowej potrzeba dysponowania odpowiednimi zestawami testów staje się coraz bardziej paląca. Dowodem tego jest m.in. realizowany w latach 1993-1996 grant Komisji Europejskiej LRE-62-089 pt. *Test Suites for Natural Language Processing* mający na celu przygotowanie odpowiednich testów dla języka angielskiego, francuskiego i niemieckiego — por. (Lehmann et al. 1996) i `http://tsnlp.dfki.uni-sb.de/tsnlp/`, oraz zapoczątkowany w 1998 roku cykl międzynarodowych konferencji poświęconych zasobom językowym i ich ocenie (Rubio et al. 1998) oraz towarzyszących im warsztatów (Caroll et al. 1998).

Należy tutaj podkreślić, że dla wielu języków od dłuższego czasu na potrzeby badań lingwistycznych, a także prac leksykograficznych, tworzone są tzw. *korpusy*. W węższym znaczeniu korpus to kolekcja danych lingwistycznych — tekstów pisanych lub transkrypcji wypowiedzi mówionych — która może stanowić punkt wyjścia do opisu lingwistycznego lub do weryfikacji pewnych hipotez dotyczących języka[1]. Z nieco innego punktu widzenia korpus można określić jako kolekcję naturalnie występujących tekstów języka naturalnego, dobranych w celu scharakteryzowania aktualnego stanu języka naturalnego lub jego różnorodności[2]. Otóż korpusy takie mogą być wykorzystywane do testowania analizatorów, ale mają również pewne wady związane z tym, że częstość użycia różnych konstrukcji jest bardzo nierównomierna — aby zatem dobrze przetestować pewną rzadką własność, należy przetworzyć lub w inny sposób przejrzeć duże ilości mało interesujących danych. Z tego właśnie powodu do testowania analizatorów tworzy się dodatkowo *test suits* — w braku dobrego odpowiednika tego angielskiego terminu będziemy takie zestawy danych nazywać *korpusami sztucznymi* (korpusy w sensie podanych wyżej definicji będziemy nazywac *korpusami naturalnymi*).

[1]Definicja ta pochodzi z pracy: David Crystal, *A Dictionary of Linguistics and Phonetics*, Blackwell, 3rd Edition, 1991.

[2]Ta definicja pochodzi z pracy: John Sinclair, *Corpus, Concordance, Collocation*, Oxford University Press 1991.

Na marginesie warto wspomnieć, że w niektórych krajach uważa się korpusy za tak istotne dla badań językowych, że tworzone są korpusy narodowe (m.in. brytyjski i czeski), niekiedy zarządzane i uaktualniane przez specjalnie do tego celu powołane instytucje. W Polsce nadal jedynym korpusem ogólnego przeznaczenia dostępnym do prac badawczych jest — omówiony niżej — stosunkowo niewielki korpus słownika frekwencyjnego.

2.2 Cele projektu

Uwzględniając aktualny stan inżynierii językowej w Polsce i jego specyfikę, podjęliśmy się realizacji następujących zadań.

1. Przygotowanie testowego zestawu zdań polskich zgromadzonych doświadczalnie, na podstawie korpusu Vetulaniego — patrz (Vetulani 1990) — i korpusu polskiego słownika frekwencyjnego — patrz (Kurcz i in. 1990).

2. Przygotowanie symetrycznego (tj. zawierające zarówno zdania poprawne, jak i zbliżone do nich zdania niepoprawne) testu analizatorów syntaktycznych opartego na pracach Szpakowicza i Świdzińskiego — por. m.in. (Szpakowicz 1986, Szpakowicz, Świdziński 1990, Świdziński 1992).

3. Przygotowanie zestawu wyników analizy syntaktycznej wybranych zdań z korpusów wymienionych wyżej w dwojakim celu — zilustrowania trudniejszych aspektów formalnego opisu języka polskiego i dokonania oceny adekwatności lingwistycznej wykorzystywanego analizatora.

W trakcie pracy dokonaliśmy pewnej korekty sformułowanych zadań, w szczególności z powodów wyjaśnionych dalej analizie syntaktycznej poddaliśmy tylko jeden z utworzonych korpusów, a także wprowadziliśmy dodatkowo zestaw testów oparty na akademickim podręczniku składni polskiej (Saloni, Świdziński 1998).

3 Zastosowanie SGML do zapisu danych lingwistycznych

Obecnie coraz powszechniej stosowanym formalizmem do reprezentowania tekstów wraz z ich strukturą są różne aplikacje Standardowego Uogólnionego Języka Adiustacyjnego (ang. *Standard Generalized Markup Language*) znanego głównie pod jego angielskim skrótem SGML; język ten jest zdefiniowany przez normę międzynarodową ISO 8879. Podstawowym podręcznikiem SGML jest (Goldfarb 1990), w Internecie podstawowym źródłem informacji są strony `http://www.oasis-open.org` i `www.xml.pl/sgml.html`.

Zastosowania SGML w lingwistyce sięgają roku 1987, kiedy rozpoczął działalność wspólny projekt ACH (*Association for Computers and the Humanities*), ACL (*Association for Computational Linguistics*) i ALLC (*Association for Literary and Linguistic Computing*) pod nazwą *Text Encoding Initiative* (TEI),

zakończony opublikowaniem specyfikacji (Sperberg-McQueen, Burnard 1994) — por. także `www.uic.edu/org/tei`. Ta bardzo obszerna specyfikacja doczekała się z czasem różnych modyfikacji do konkretnych zastosowań, takich jak np. *Corpus Encoding Standard* (CES) — por. np. (Erjavec, Lawson 1998). Każda taka szczegółowa aplikacja jest zdefiniowana przez tzw. definicję typu dokumentu (ang. *Document Type Definition*), nazywaną w skrócie DTD. Definicja ta określa m.in. jakie znaczniki (ang. *tags*) mogą być użyte w zapisie tekstu i jakie są między nimi zależności.

SGML stosuje się coraz częściej do zasobów lingwistycznych również w Polsce, m. in. w ramach projektów TELRI (*TransEuropean Language Resources Activities Network*) i STEEL (*Specialized Tools for foreign language translation/understanding for East European Languages* — por. (Erjavec, Lawson 1998) i (Głowińska, Woliński w druku).

Stosowane zestawy znaków i inne szczegóły formalizmu tzw. składni konkretnej określane są przez definicję SGMLową (*SGML definition*), która nie jest obowiązkowa, domyślnie przyjmuje się bowiem wzorcową składnię konkretną (*concrete reference syntax*). Ponieważ w składni tej litery narodowe mogą być reprezentowane tylko w pośredni sposób za pomocą tzw. całostek (*entities*), będziemy stosować składnię konkretną definiującą zestaw znaków dokumentu jako kod ISO Latin-2. ***Definicja ta znajduje się na płycie*** w katalogu `korpusy` i — dla wygody — w jego podkatalogach.

Warto dodać, że pierwotna wersja standardu SGML — nazywana niekiedy SGML 1986 — nie pozwala na stosowanie w nazwach elementów (*generic identifiers*) liter narodowych. Mozliwość taka pojawiła się dopiero po wprowadzeniu w 1997 roku modyfikacji związanych ze stosowaniem uniwersalnych zestawów znaków (UNICODE, ISO/IEC 10646), ale postanowiliśmy z niej nie korzystać, aby nie wprowadzać zbędnych komplikacji. W związku z tym dalej zamiast `Przykład` piszemy `Przyklad` itp.

W niniejszym opracowaniu nie dyskutujemy szczegółów technicznych stosowanych definicji typów dokumentów, koncentrujemy się natomiast na informacjach, które za pomocą tych DTD są reprezentowane.

4 Korpusy naturalne

4.1 Korpus słownika frekwencyjnego

4.1.1 Podstawowe informacje

Korpus słownika frekwencyjnego to liczący 500 000 słów zestaw próbek wylosowanych z autentycznych tekstów na potrzeby słownika frekwencyjnego języka polskiego (Kurcz i in. 1990); choć próbki te pochodzą z lat 1963–1967 i stosowane w nich słownictwo jest częściowo przestarzałe, korpus ten nadal ma dużą wartość m.in. dla badań składniowych. Do celów badawczych jest on udostępniany przez autorów bezpłatnie, w najbliższych miesiącach przewidywane jest usprawnienie jego dystrybucji dzięki włączeniu go do zasobów europejskiego projektu ELAN (*European Language Activities Network*) — por.

http://solaris3.ids-mannheim.de/elan/.

Na potrzeby projektu ELAN korpus słownika frekwencyjnego został w Instytucie Informatyki UW przekształcony do formatu SGML. Zgodnie z założeniami tego projektu tekst korpusu został zapisany zgodnie z DTD PAROLE poziom 1.[3] Polega to w praktyce na oznaczaniu wyłącznie końców zdań i akapitów (bliższe informacje o tym formacie można znaleźć pod adresem http://svenska.gu.se/~ridings/textrep/textrep.html). Na potrzeby naszego projektu taki zapis byłby jednak zbyt ubogi.

Ze względu na to, że dostępne na nośniku oryginalne próbki nie uwzględniają późniejszych korekt i poprawek, ograniczyliśmy się do fragmentu zweryfikowanego w ramach wspomnianego na wstępie projektu prof. Świdzińskiego. We fragmencie tym pominięto wprawdzie kody informacji morfologicznej, ale dla naszych potrzeb nie ma to istotnego znaczenia. Fragment ten składa się z 1 000 próbek o łącznej objętości około 50 000 słów składających się na 8 000 zdań; uzyskano je wybierając co dziesiątą próbkę z korpusu słownika frekwencyjnego. Jednocześnie próbki te uzupełniono o wprowadzoną ręcznie informację o strukturze składniowej próbek.

Dla odróżnienia od oryginalnego korpusu słownika frekwencyjnego opisany wyżej jego podzbiór będziemy nazywać *korpusem wypowiedzeń współczesnej polszczyzny.*

4.1.2 Korpus w formacie SGML

Korpus słownika frekwencyjnego — a w konsekwencji korpus wypowiedzeń — składa się z pięciu części, odpowiadających odpowiednio tekstom popularnonaukowym, drobnym wiadomościom prasowym, publicystyce, prozie artystycznej i dramatowi artystycznemu. Części takie nazywa się niekiedy transzami.

Oto przykład zapisu konkretnego wypowiedzenia:

```
<Przyklad Zrodlo=KWWP-DR>
<NrProbki>1
<NrWypowiedzenia>1
<Segmenty><seg3>Co<\seg3> <seg1>myślał<\seg1> <seg2>ten artysta<\seg2>?
<Tresc>Co myślał ten artysta?</Tresc>
```

Atrybut `Zrodlo` elementu `Przyklad` wskazuje na to, że przykład pochodzi z korpusu wypowiedzeń współczesnej polszczyzny (skrót KWWP), a konkretnie z transzy dramatu artystycznego (skrót DR). Widzimy także, że jest to pierwsze wypowiedzenie pierwszej próbki tej transzy. Właściwe wypowiedzenie stanowi zawartość elementu `tresc`; element ten zaczyna się znacznikiem początkowym `<tresc>` i jest zakończony jawnie znacznikiem końcowym `</tresc>`. Jest to wygodne ze względu na to, że niektóre wypowiedzenia są zapisane w kilku wierszach lub kończą się wielokropkiem. Pozostałe elementy nie mają znaczników końcowych; dzięki wykorzystaniu tzw. własności minimalizacji koniec elementu

[3] PAROLE — wymawiane jak słowo włoskie, a nie francuskie — jest stowarzyszeniem powołanym w 1995 r. i stanowiącym sobie za cel gromadzenie, udostępnianie i popularyzację zasobów językowych — w szczególności korpusów — na skalę europejską.

rozpoznaje się po początku następnego elementu z tego samego poziomu struktury.

Element `Segmenty` zawiera dodatkowe informacje składniowe, dodane do korpusu w ramach wspomnianego wcześniej projektu Świdzińskiego. Oryginalna notacja Świdzińskiego wykorzystywała m.in. nawiasy kątowe, prostokątne i okrągłe, którym w naszym zapisie są reprezentowane odpowiednio przez elementy `seg1`, `seg2`, `seg3`. Znaczenie tych oznaczeń jest wyjaśnione w książce (Świdziński 1996). Książka ta zawiera również inne obszerne informacje na temat korpusu wypowiedzeń[4].

Pliki korpusu wypowiedzeń znajdują się na płycie w podkatalogu `KWWP` katalogu `KORPUSY`[5].

4.2 Korpus dialogów

4.3 Zawartość korpusu

Ponieważ omówiony wyżej korpus wypowiedzeń jest oparty wyłącznie na tekstach pisanych, uznaliśmy za wskazane uzupełnienie naszego zestawu danych o transkrypcje wypowiedzi mówionych. Praktycznie jedynym dostępnym korpusem tego typu jest korpus dialogów stworzony pod kierunkiem prof. Zygmunta Vetulaniego na Uniwersytecie im. Adama Mickiewicza w Poznaniu; prace nad nim były finansowane przez fundację Humboldta oraz przez projekty CPBP 08.15 i CPBP 08.05. Korpus ten razem z obszernymi komentarzami został opublikowany w książce (Vetulani 1990), a sam korpus udostępniony zainteresowanym na dyskietce. Niestety, wersja dyskietkowa korpusu została zapisana w ASCII, czyli bez polskich znaków diakrytycznych, co było dla nas nie do zaakceptowania. Na szczęście prof. Vetulani zachował tekst źródłowy książki (przygotowanej za pomocą edytora Chi-Writer) i uprzejmie nam go udostępnił. Pozwoliło to dokonać konwersji interesujących nas fragmentów książki i w rezultacie otrzymać korpus ze znakami diakrytycznymi i różnymi dodatkowymi informacjami. Prof. Vetulani uprzejmie upoważnił nas również do swobodnej dystrybucji tak uzyskanego korpusu.

Objętościowo korpus jest nieduży, liczy tylko 30 krótkich dialogów. Są to tzw. dialogi konsultatywne, odbywające się między dwoma partnerami, z których jeden posiada, a drugi pragnie uzyskać pewne informacje. Dialogi zostały uzyskane drogą eksperymentu, w którym jeden partner dysponował pełnym rysunkiem pewnej sytuacji, a drugi tylko pewnymi fragmentami tego rysunku; pytania dotyczyły elementów sytuacji nieznanych drugiemu partnerowi.

Oto przykład fragmentu dialogu w wersji książkowej (Vetulani 1980:23), przytoczony z dokładnością do wyróżnień typograficznych:

[4]Mówiąc ściślej, książka ta — jak widać po tytule — zajmuje się wypowiednikami, nie chcemy jednak wnikać tutaj w subtelne różnice między wypowiedzeniami (termin powszechnie przyjęty) a wypowiednikami (termin wprowadzony przez Świdzińskiego).

[5]Płytę można odczytać zarówno w systemach operacyjnych Microsoftu jak i w systemie UNIX i pochodnych; nie stosujemy ani zapisu `KORPUSY\KWWP` ani `KORPUSY/KWWP`, ponieważ sugerują one tylko jeden z wymienionych typów platform.

A.1.1. `Co trzyma w ręce św. Mikołaj?`
$X_{subst,a}$; $V_{f,p(3)}$;$_{<w>}$ N_1; N_n
(?)[Ar_1: Św. Mikołaj;P: trzyma;Ar_2: ?; Ar_3: ¡w¿ ręku]
[[Ar_1: $_4(N_n)$; P: $_2(V_{f,p(3)})$; Ar_2:$_1(X_{subst,a})$; Ar_3: $_3(_{<w>}N_1)$];
P=TRZYMAĆ–CZYMŚ(Ar_1,Ar_2,Ar_3)]
B.1.1. `Św. Mikołaj trzyma książkę.`

Jak widać, fragment ten ma złożoną strukturę. Symbol A.1.1. oznacza, że jest to pierwsza wypowiedź partnera oznaczonego literą A w dialogu nr 1; symbol B.1.1 analogicznie oznacza wypowiedź partnera B. W ogólnym wypadku wypowiedzi partnera B mogą być wielozdaniowe (por. np. wypowiedź B.29.15), sporadycznie wielozdaniowe są także wypowiedzi partnera A (por. wypowiedź A.10.18).

Formalne zapisy znajdujące się zawsze po wypowiedzi partnera A nazywane są przez autora opisem syntaktycznym na poziomie formalnym i poziomie predykat-argument. Należy w związku z tym podkreślić istotną różnicę terminologiczną. W niniejszym opracowaniu przez składnię rozumiemy zawsze tzw. składnię powierzchniową, bardzo bliską szkolnemu rozumieniu tego słowa, natomiast autor korpusu ma na myśli tzw. składnię głęboką, którą można często utożsamiać z formalną reprezentacją znaczenia.

4.3.1 Korpus w formacie SGML

Format SGML zastosowany do korpusu dialogu omówimy podając zapis fragmentu dialogu przedstawionego powyżej. Chociaż na obecnym etapie nie interesują nas formalne reprezentacje składni głębokiej, zachowujemy je w korpusie stosując mechaniczną transkrypcję użytej do tego notacji; ponieważ transkrybowane zapisy są dość obszerne, a jednocześnie w tej chwili mało istotne, pominęliśmy je w przedstawionym niżej przykładzie.

```
<Dialog NrStr=23>
  <Tytul>1
  <ZapytOdp>
       <IdZap NrStr=23>
          <Osoba>A
          <NrDialogu>1
          <NrZapytania>1
       </IdZap>
       <TrescZap>
          <Wiersz>Co trzyma w ręku św. Mikołaj?</Wiersz>
       </TrescZap>
       <Modele>
           ...
       </Modele>
       <IdOdp NrStr=23>
          <Osoba>B
```

```
            <NrDialogu>1
            <NrOdpowiedzi>1
         </IdOdp>
         <TrescOdp>
            <Wiersz>Św. Mikołaj trzyma książę.</Wiersz>
         </TrescOdp>
   </ZapytOdp>
```

Podstawową jednostką jest element `Dialog`, którego atrybut wskazuje na numer strony w książce, na której dany dialog się rozpoczyna. Element ten posiada tytuł, będący po prostu numerem dialogu, po którym następuje ciąg elementów `ZapytOdp` (zapytanie—odpowiedź). Element ten ma z kolei 5 składników: identyfikator zapytania (`IdZap`), treść zapytanie (`TrescZap`), transkrypcję reprezentacji składni głębokiej (`Modele`), identyfikator odpowiedzi (`IdOdp`) i treść odpowiedzi (`TrescOdp`). Identyfikatory mają atrybut podający, na której stronie książki zaczyna się dana wypowiedź, oraz podelementy: numer dialogu i numer wypowiedzi. Jest tutaj świadomie dopuszczona pewna redundancja — numer dialogu pokrywa się z tytułem, a numer zapytania z numerem odpowiedzi, chcieliśmy jednak nie odchodzić zbyt daleko od oryginału; poza tym zgrupowanie pełnej informacji przy konkretnej wypowiedzi może okazać się pożyteczne, gdy wypowiedzi te będą przetwarzane w izolacji. Inny rodzaj redundancji wprowadzają elementy `<wiersz>`, które informują po prostu o podziale na wiersze w oryginalnym tekście książki.

Pliki korpusu dialogów znajdują się na płycie w podkatalogu `CCD` katalogu `KORPUSY`.

5 Korpusy sztuczne

5.1 Korpus gramatyki Świdzińskiego

5.1.1 Gramatyka formalna języka polskiego

Książka (Świdziński 1992) zawiera prezentację najobszerniejszej i najbardziej szczegółowej gramatyki formalnej języka polskiego, jaka dotąd została opracowana; gramatykę tę dalej oznaczamy skrótowo GFJP.

Stosowany w książce formalizm wywodzi się z gramatyki metamorficznej, wprowadzonej przez Alaina Colmerauera[6] (Colmerauer 1978) — por. także (Kluźniak, Szpakowicz 1983, 1985) — nie jest jednak z nią identyczny. Oryginalna gramatyka metamorficzna dopuszcza występowanie po lewej stronie symboli terminalnych w charakterze prawego kontekstu, w gramatyce Świdzińskiego stosowany jest sporadycznie również lewy kontekst. Gramatyka metamorficzna umożliwia nakładanie pewnych warunków, które w gramatyce Świdzińskiego stosowane są w nieortodoksyjny sposób, co jest konsekwencją świadomego ignorowania przez jej autora obliczeniowych aspektów stosowanego formalizmu (patrz cytat na s. 24).

[6] Jest on znany przede wszystkim jako twórca języka Prolog.

Przykłady reguł gramatyki Świdzińskiego znajdują się w punkcie 6.4 na s. 15.

5.1.2 Podkorpusy GFJP-A i GFJP-B

Gramatyka Świdzińskiego jest obszernie ilustrowana przykładami zarówno zdań poprawnych, jak i niepoprawnych; ze względu na to, że przykłady te są w istotnie inny sposób podawane w aneksie, a inaczej we właściwej treści książki, rozróżniamy dwa zestawy przykładów: GFJP-A (przykłady z aneksu) i GFJP-B (pozostałe przykłady).

Charakterystyczną cechą przykładów z aneksu (GFJP-A) jest wyraźne przyporządkowanie im konkretnych reguł gramatyki. Niemal każda reguła jest ilustrowana bowiem przez zdania poprawne, które wymagają zastosowania tej reguły dla otrzymania poprawnej analizy, i przez bardzo podobne zdania poprawne, do których jednak dana reguła się nie stosuje. Jest oczywiste, że takie przykłady idealnie nadają się do testowania, czy analizator syntaktyczny wiernie realizuje zadaną gramatykę formalną, a także do wykrywania ewentualnych błędów w samej gramatyce. Z tego powodu temu zestawowi przykładów poświęciliśmy najwięcej uwagi. Zawiera on 660 przykładów, z czego 515 poprawnych i 145 niepoprawnych — jak widać, nie jest on wprawdzie symetryczny w ścisłym znaczeniu tego słowa, ale przykłady niepoprawne stanowią jednak znaczącą część podkorpusu. Długość przykładów wynosi od 2 do 22 słów, średnia długość przykładu wynosi 6 słów.

Przykłady z właściwej treści książki (GFJP-B) mają bardziej różnorodny charakter, niektóre np. ilustrują zjawiska lingwistyczne świadomie nie uwzględnione w gramatyce formalnej. W niektórych przypadkach jednak z kontekstu użycia przykładu można wywnioskować, jaką regułę gramatyki formalnej może on objaśniać. W sytuacjach takich staraliśmy się odpowiednię informację zapisać w sposób jawny. Podkorpus GFJP-B zawiera 1376 przykładów, w tym 1054 poprawnych, 296 niepoprawnych i 26 wątpliwych. Dodatkowo w tym podkorpusie wyróżnione są — niezależnie od oceny poprawności — zdania akceptowane przez gramatykę (979 przykładów) i nieakceptowane przez gramatykę (371 przykładów). Długość przykładów wynosi od 1 słowa do 59 słów, średnia długość przykładu wynosi 7 słów.

Ze względu na intensywne wykorzystywanie korpusu w trakcie tworzenia i testowania analizatora okazało się bardzo wskazane dysponowanie możliwością szybkiego odnalezienia konkretnego przykładu w książce. Służą do tego dwa rodzaje lokalizacji — przez numer strony oraz (w zestawie GFJP-A) przez wskazanie odpowiedniego podpunktu aneksu.

5.1.3 Postać korpusu w formacie SGML

Podobnie jak w poprzednich przypadkach, przykłady te zostały zapisane w formacie SGML będącym pewnym uszczegółowieniem specyfikacji TEI. Poniżej omówimy przykład pochodzący z podkorpusu GFJP-A, przykłady z z podkorpusu GFJP-B mają identyczną strukturę, różnią się tylko brakiem pewnych informacji.

```
<Przyklad Zrodlo=GFJP-A>
<NrPrzykladu>583
<Kompletnosc>T
<Poprawnosc>T
<Akceptowalnosc>T
<Lokalizacja>A5-14.5.1
<NrStrony>405
<IdReguly>PS19
<Zasieg>F
<Segmenty><seg1>Gdzie</seg1> macie szałas?
<Tresc>Gdzie macie szałas?</Tresc>
```

Podstawową jednostką korpusu jest element `Przykład`, posiadający atrybut wskazujący na źródło przykładu, w naszym wypadku jest to podkorpus GFJP-A. Element `NrPrzykładu` jednoznacznie wskazuje na zawartość elementu `Tresc` — jeśli przykłady o tej samej treści się powtarzają, to występują w korpusie tylko jeden raz pod jednym numerem.

Element `Kompletnosc` określa, czy chodzi o samodzielne zdanie, czy też o jego fragment (np. frazę nominalną, przymiotnikową itp.). Element `Poprawnosc` zawiera ocenę poprawności przykładu przez jego autora; może on przyjmować co najmniej 3 wartości: poprawne, niepoprawne i wątpliwe. W ideale wszystkie przykłady poprawne powinne być akceptowane przez gramatykę, a niepoprawne odrzucane, w praktyce występują jednak rozbieżności, stąd konieczność stosowania dodatkowego elementu `Akceptowalnosc`.

Następna para elementów, `Lokalizacja` i `NrStrony`, w razie potrzeby mogą się powtarzać dowolną liczbę razy. O ile funkcja numeru strony jest oczywista, element `Lokalizacja` wymaga pewnego komentarza — pozwala on mianowicie odnaleźć przykład również w pracy (Świdziński 1987), stanowiącej wcześniejszą wersję książki (Świdziński 1992).

Para elementów `IdReguły` i `Zasieg` (zasięg) również może być w razie potrzeby powtórzona. Element `IdReguły` to oczywiście identyfikator reguły gramatyki formalnej, która jest ilustrowana przez dany przykład. Element `Zasieg` wskazuje dodatkowo, czy reguła ta odnosi się do całej treści przykładu, czy tylko do jego frementu, stanowiącego zawartość podelementu `seg1` w elemencie `Segmenty`.

Element `Segmenty`, oprócz wymienionej wyżej funkcji, pozwala zapisać też niektóre inne informacje o strukturze przykładowego zdania podane w oryginale.

Pliki korpusu gramatyki Świdzińskiego znajdują się na płycie w podkatalogu `GFJP` katalogu `KORPUSY`.

5.2 Korpus gramatyki Szpakowicza

5.2.1 Gramatyka Szpakowicza

Przez gramatykę Szpakowicza rozumiemy opracowaną przez niego pierwszą formalną gramatykę języka polskiego, opisującą nietrywialny podzbiór języka, i przedstawioną w jego pracy doktorskiej (Szpakowicz 1978). Na podstawie tej pracy powstała książka (Szpakowicz 1986), której charakter był jednak nieco

inny — praca doktorska była broniona w Instytucie Informatyki i siłą rzeczy była adresowana do informatyków, natomiast książka była przeznaczona dla lingwistów. Po powstaniu gramatyki Świdzińskiego wydawało się początkowo, że gramatyka Szpakowicza stała się całkowicie przestarzała, natomiast w miarę upływu czasu coraz wyraźniej stały się widoczne jej zalety dydaktyczne — o ile gramatyka Świdzińskiego ze względu na swoją dużą złożoność i trudną przyswajalność nie wywarła większego wpływu na środowisko lingwistyczne, to książka Szpakowicza stała się klasyczną pracą z tej dziedziny, a fragmenty jego gramatyki doczekały się kilku niezależnych realizacji komputerowych.

Ponieważ książka Szpakowicza od dawna była wyczerpana, a praca doktorska z natury rzeczy była trudno dostępna, już kilka lat temu rozpoczęto przygotowania do stworzeniem elektronicznej wersji tych prac. Ze względu na niską jakość maszynopisu pracy doktorskiej, a także niską jakość druku książki wydanej techniką małej poligrafii, konieczne było ręczne wprowadzenie tekstu książki do komputera. To istotne utrudnienie postawiono obrócić w zaletę, uzupełniając w trakcie wprowadzania tekst książki o pewne dodatkowe informacje, pozwalające wykorzystać następnie wyrafinowane możliwości systemu składania tekstów TEX. Dużą uwagę przywiązano również do opracowania redakcyjnego książki, starając się ustalić konwencje przydatne także dla innych publikacji o podobnym charakterze. Funkcję redaktora merytorycznego pełnił J. S. Bień, szczegółowe informacje na ten temat można znaleźć w opracowanym przez niego posłowiu do elektronicznych wersji tych prac.

Dzięki odpowiedniemu wykorzystaniu systemu TEX możliwe jest uzyskanie tekstu wynikowego w kilku postaciach: DVI (*DeVice Independent file* specyficzny dla systemu TEX), PostScript, Portable Document Format i HTML (Pietrzak 1999).

Na płycie postanowiliśmy udostępnić — oczywiście za zgodą autora — wersję w formacie PDF (z zablokowaną możliwością drukowania) dla pracy doktorskiej i książki; ponieważ obecne prawo autorskie wymaga jawnego określenia w umowie wydawniczej tzw. dziedziny eksploatacji, a oczywiście udostępnianie elektroniczne nie było jawnie wymienione w typowej umowie, pełne prawa do wydań elektronicznych przysługują autorowi — z tego powodu zgoda Szpakowicza jest w pełni wystarczająca. Planowane jest całkowite rozwiązanie umowy z Wydawnictwami UW, co pozwoli udostępnić publicznie również inne elektroniczne formy tych tekstów.

Pliki z pracą doktorską i książką Szpakowicza znajdują się na płycie w katalogu `TEKSTY` w plikach `aaspzp.pdf` (doktorat) i `foszp.pdf` (książka).

5.2.2 Korpus przykładów

W trakcie składu opisanych wyżej elektronicznych publikacji wszystkie zawarte w nich przykłady są automatycznie wypisywane na odpowiedni plik wraz z pewnymi informacjami dodatkowymi. Po odpowiednim przekształceniu otrzymujemy zapis przykładów w formacie SGML. Ich postać jest bardzo zbliżona do przykładów z korpusu gramatyki Świdzińskiego, ale zawierają one mniej informacji; omówimy obecnie jeden z nich.

```
<Przyklad Zrodlo=FOSZP>
<Zasieg>F
<Lokalizacja>
<NrStrony>52
<IdReguly>frzpw6
<Segmenty><seg2>szukam ostatnich</seg2> ośmiorga dzieci
<Tresc>szukam ostatnich ośmiorga dzieci</tresc>
```

Atrybut `Zrodlo` elementu `Przyklad` zawiera skrót tytułu pracy i w tym wypadku wskazuje, że przykład pochodzi z książki (Szpakowicz 1986), a nie z pracy doktorskiej (Szpakowicz 1978). Wartość `F` (fragment) elementu `Zasieg` wskazuje, że podana reguła odnosi się tylko do tego fragmentu elementu `Segmenty`, ktory znajduje się na zewnątrz podelementu `seg2`, reprezentuje on bowiem w wersji drukowanej ujęcie w nawiasy okrągłe uzupełnienia właściwego przykładu do pełnego zdania.

Doktorat zawiera 787 przykładów, a książka 860; ich użyteczność jest bardzo różna. Te z nich, które zawierają — jak powyższy przykład — numery reguł, są bardzo pożyteczne do testowania analizatorów. W korpusie jednak, wskutek jego automatycznego tworzenia, znajdują się również przykłady w postaci np. pojedynczych słów, które do celów składniowych są mało przydatne.

Pliki korpusu gramatyki Szpakowicza znajdują się na płycie w podkatalogach `AASPZP` (przykłady z doktoratu) i `FOSZP` (przykłady z książki) katalogu `KORPUSY`.

5.3 Korpus „niebieskiej gramatyki”

5.3.1 Książka

„Niebieską gramatyką” (od koloru okładki) nazywany jest potocznie akademicki podręcznik składni polskiej autorstwa Zygmunta Saloniego i Marka Świdzińskiego (1998). Wyróżnia się on nowoczesnym podejściem do opisywanej problematyki — jak czytamy w najnowszym wydaniu, ambicją autorów jest *dostarczenie czytelnikowi kompendium gramatycznego na miarę ery rewolucji komputerowej. Potrzeby coraz szerszych rzesz informatyków zajmujących się automatycznym przetwarzaniem tekstów wymagają opisu lingwistycznego różnego od tych, jakie wdraża się w edukacji szkolnej.*

Autorzy książki wyrazili uprzejmie zgodę na udostępnienie jej w wersji elektronicznej (do czego — jak wspominaliśmy wcześniej — mają wystarczające prawo z punktu widzenia obowiązującego prawa autorskiego).

Plik z tekstem podręcznika znajduje się na płycie w katalogu `TEKSTY`. Jest to plik w formacie PDF (bez możliwości drukowania) o nazwie `swjp.pdf`.

5.3.2 Korpus

W związku z tym, że skład tej książki był wykonany za pomocą systemu TEX przez uczestnika niniejszego grantu Marcina Wolińskiego, postanowiliśmy skorzystać z okazji i zastosować podobną technikę tworzenia korpusu przykła-

dów, co w przypadku prac Szpakowicza. Informacja towarzyszące poszczególnym przykładom z niebieskiej gramatyki jest jednak jeszcze uboższa niż w przypadku gramatyki Szpakowicza. Jest to konsekwencją faktu, że w przypadku prac Szpakowicza decyzja o automatycznym generowanie plików z przykładami była podjęta jeszcze w trakcie opracowywania tekstu w komputerze, dzięki czemu można było ten tekst w miarę możliwości odpowiednio uzupełnić. W przypadku niebieskiej gramatyki dysponowaliśmy wprawdzie tekstem źródłowym dla systemu TEX, ale niestety zorientowanym wyłacznie na skład tradycyjny. Na szczęście w podręczniku tym — zgodnie z tradycją lingwistyczną — odróżnia się jawnie przykłady poprawne od niepoprawnych.

Oto jeden z przykładów w formacie SGML:

```
<Przyklad Zrodlo=SWJP>
<NrPrzykladu>132
<Poprawnosc>N
<Tresc>Jaś pilnuje albo.</Tresc>
```

Jak widać, ma on analogiczną postać, jak w przypadku już omawianych korpusów.

Korpus ten liczy 1904 przykłady, w tym 1703 poprawne, 184 błędne i 17 wątpliwych.

Pliki korpusu niebieskiej gramatyki znajdują się na płycie w podkatalogu `SWJP` katalogu `KORPUSY`.

5.4 Korpus fraz nominalnych Szpakowicza i Świdzińskiego

Ten zestaw przykładów pochodzi z artykułu pod tytułem *Formalna definicja równorzędnej grupy nominalnej we współczesnej polszczyźnie pisanej* (stąd stosowany niżej skrót FDRGN) autorstwa Stanisława Szpakowicza i Marka Świdzińskiego, który można traktować jako rozwinięcie gramatyki Szpakowicza; dostępny od 1981 w formie powielonego maszynopisu, później ukazał się drukiem (Szpakowicz, Świdziński 1990). Przykłady te zostały wprowadzone do komputera jeszcze w latach osiemdziesiątych na potrzeby pracy magisterskiej Mirosława Bańki (1985), a obecnie zostały przekształcone do formatu SGML.

Oto jeden z przykładów:

```
<Przyklad Zrodlo=FDRGN>
<NrPrzykladu>1
<Poprawnosc>T
<Lokalizacja>
<NrStrony>
<IdReguly>R2
<Tresc>Zarówno chłopiec, jak i dziewczyna przyszli.</Tresc>
```

Zestaw ten liczy 217 przykładów oryginalnych i 17 przykładów dodanych przez Mirosława Bańkę. Przykładów poprawnych jest łącznie 145, niepoprawnych 88, wśród przykładów dodanych znajduje się jeden wątpliwy. Choć korpus

ten jest najmniejszy, jest on jednocześnie najbardziej symetryczny, co przesądza o jego przydatności do testowania analizatorów.

Pliki tego korpusu znajdują się na płycie w podkatalogu `FDRGN` katalogu `KORPUSY`.

6 Bank rozbiorów gramatycznych

6.1 Analizator morfologiczny SAM-98

Do stworzenia banku rozbiorów gramatycznych zamierzaliśmy użyć analizatora AMOS, opracowanego w latach 1994–1996 w ramach projektu KBN (Bień 1996a). Analizator ten składa się z dwóch stosunkowo niezależnych modułów: analizatora morfologicznego SAM (Szafran 1996, 1997) i właściwego analizatora syntaktycznego. Na potrzeby niniejszego projektu SAM został rozbudowany, mianowicie jego słownik został rozszerzony o informacje stanowiące rozszerzenie i rozwinięcie niepublikowanych wyników innego projektu KBN, mianowicie projektu *Słownik gramatyczny współczesnego języka polskiego*, zrealizowanego w latach 1992–1994 pod kierunkiem prof. dr. hab. Zygmunta Saloniego. W odróżnieniu od wcześniejszej wersji (dostępnej od dłuższego czasu w Internecie), nową wersję nazywamy SAM-98.

6.2 Analizator syntaktyczny AMOS

Jak już wspominaliśmy, gramatyka Świdzińskiego ma bardzo nietypowy charakter z formalnego punktu widzenia (Bień 1996b). Podstawowym założeniem analizatora syntaktycznego AMOS było zachowanie gramatyki w zasadzie niezmienionej, wykorzystanie dostępnej w Prologu obsługi gramatyk metamorficznych (a ściślej, ich wariantu zwanego *definite clause grammars*) i odpowiednia obsługa wprowadzonych przez Świdzińskiego rozszerzeń, wykorzystująca między innymi „korutyny” (predykat *freeze*). Ponieważ podstawowym celem była weryfikacja gramatyki Świdzińskiego, a nie analizator przeznaczony do celów praktycznych, efektywność działania była traktowana jako drugorzędna.

Ponieważ początkowe wyniki były zachęcające, mieliśmy nadzieję, że dysponując dostatecznie dużą mocą obliczeniową (oszczędności uzyskane dzięki zakupowi tańszej drukarki przeznaczyliśmy na dodatkowe komputery) będziemy w stanie przetworzyć w wymaganym czasie dostatecznie dużą liczbę przykładów. Niestety, założenie to było błędne — okazało się, że analizator ten jest bardzo podatny na eksplozje kombinatoryczne prowadzące do kilkudniowych obliczeń dla pojedynczego przykładu. Próby dodatkowej optymalizacji analizatora (polegające m.in. na coraz bardziej wyrafinowanej faktoryzacji reguł) nie wpływały istotnie na polepszenie efektywności, wprowadzały natomiast nowe problemy: struktura analizatora stawała się coraz bardziej zawiła i trudniejsza do ogarnięcia, zaczęły pojawiać się trudne do zlokalizowania zakleszczenia itp.

Eksperymenty z analizatorem AMOS doprowadziły do lepszego poznania różnych aspektów gramatyki Świdzińskiego, w szczególności pozwoliły wychwy-

cić znajdujące się w niej różne drobne omyłki i przeoczenia, a także nabrać praktycznego doświadczenia w automatycznej analizie syntaktycznej języka naturalnego. Tym niemniej, z punktu widzenia celów niniejszego projektu, w pewnym momencie prace nad wykorzystaniem systemu AMOS znalazły się w ślepym zaułku.

6.3 Analizator syntaktyczny AS

Wobec zasadniczych trudności z wykorzystaniem analizatora AMOS do zamierzonych celów podjęliśmy radykalną decyzję opracowania nowego analizatora opartego na istotnie odmiennych założeniach. Szczegółową koncepcję tego analizatora opracował, a następnie zrealizował mgr Marcin Woliński; analizator ten został nazwany AS. Podstawowym założeniem była konwersja gramatyki na gramatykę równoważną, ale o wiele lepiej poddającą się standardowym technikom analizy syntaktycznej; otrzymane wyniki są następnie poddawane operacji nazywanej żartobliwie „świdzińskizacją", dzięki czemu użytkownik analizatora ma wrażenie, że korzysta on z gramatyki oryginalnej; wynik analizatora AS różni się od wyników analizy przeprowadzonej ręcznie tylko nieistotnymi szczegółami.

Proces konwersji gramatyki okazał się nietrywialny, ale dzięki niemu można było zastosować łatwo metodę analizy wstępującej (*bottom-up*) — por. np. (Gazdar, Mellish 1989), która okazała się bardziej efektywna. Została ona dodatkowo uzupełniona o zapamiętywanie częściowych wyników analizy; realizacja tego postulatu również nie była trywialna ze względu na skomplikowane struktury danych niezbędne do właściwej reprezentacji gramatyki Świdzińskiego.

Praca nad analizatorem nie jest jeszcze zamknięta, ponieważ widzimy potrzebę i możliwości wprowadzenia dodatkowych ulepszeń i udoskonaleń; z tego też powodu nie omawiamy tutaj dokładniej samego analizatora, lecz przedstawiamy tylko — zgodnie z założeniami projektu — jego wyniki.

6.4 Wizualizacja wyników analizy syntaktycznej

Wizualizacja drzewa rozbioru gramatycznego jest dokonywana zgodnie z konwencją opracowaną na potrzeby systemu AMOS — por. (Bień 1996). Konwencję tę przedstawimy poniżej na kilku konkretnych przykładach.

Niektóre aspekty stosowanej notacji są odziedziczone po języku programowania Prolog, w którym zrealizowano analizator syntaktyczny.

Najpierw omówimy drzewo analizy dla przykładu

Gdzie macie szałas?

Jak wskazuje nagłówek, jest to pierwsza z pięciu możliwych analiz przykładu nr 583 zawartego w aneksie do książki Świdzińskiego, w punkcie 5-14.5.1 (analiza ta znajduje się w katalogu `DRZEWA.PDF` w pliku `ap0560d.pdf` na stronie 3658). Przykład ten służy do zilustrowania funkcji reguły PS19, przy czym reguła ta odnosi się konkretnie do tego fragmentu przykładu, który jest ujęty w nawiasy kątowe, czyli do słowa *Gdzie*.

Aby uzyskać możliwie zwartą i oszczędną reprezentację drzew, drukujemy je w postaci dwóch kolumn. Lewa z nich zawiera odpowiednie symbole terminalne i nieterminalne, prawa zaś — oznaczenia zastosowanych reguł. Symbol początkowy gramatyki znajduje się w lewym górnym rogu, jest to symbol `wypowiedzenie`. W tym samym wierszu w prawej kolumnie znajduje się oznaczenie `[w1]`, mówiące o tym, że zastosowano regułę W1 (Świdziński 1992:334):

```
WYPOWIEDZENIE
    = # %                                                    (W1)
      ZR (wf, a, c, t, rl, o, neg, i, z)
      ZNAKKOŃCA (z).
```

Reguła ta stwierdza, że wypowiedzenie składa się ze zdania równorzędnego i znaku końca. W konsekwencji w odpowiednich miejscach lewej kolumny znajdujemy odpowiednio symbole

```
zr(os,nd,ter,ozn,_2625/mno,2,tak,ni,p,0)
```

i

```
znakkońca(p),
```

stanowiące pierwszy — nie licząc wierzchołka — poziom drzewa analizy. Dla przejrzystości są one wcięte w stosunku do symbolu nadrzędnego i połączone pionową linią.

Symbol nieterminalny `zr` ma 10 parametrów. Tzw. wyróżnik fleksyjny przyjmuje wartość osobową (`os`), aspekt ma wartość niedokonaną (`nd`), czas ma wartość teraźniejszą (`ter`), tryb — wartość oznajmującą (`ozn`). Liczba poprzedzona podkreśleniem to Prologowy zapis zmiennej wolnej, wskazującej, że pewien parametr — w tym wypadku rodzaj gramatyczny — nie został ustalony; liczba gramatyczna ma wartość mnogą (`mno`), wartością zaś kategorii osoby jest osoba druga (`2`). O ile wymienione dotąd kategorie są znane nawet laikom z gramatyki szkolnej, trzy następne parametry są specyficzne dla gramatyki Świdzińskiego: negatywność ma wartość pozytywną (`tak`), inkorporacyjność ma wartość „negatywną" (`ni`), zależność ma wartość pytajną (`p`). Ostatni parametr o wartości `0` został dodany ze względów technicznych i nie należy go brać pod uwagę przy interpretacji drzewa.

Ogólnie rzecz biorąc, dla znających książkę Świdzińskiego interpretacja symboli parametrów powinna być oczywista, ponieważ wprowadzone ze względów technicznych zmiany są bardzo niewielkie. Tak więc wartości ustalone parametrów (podobnie zresztą jak same symbole) są zapisywane małymi, a nie dużymi literami, niekiedy litery polskie są zastąpione odpowiednimi literami angielskimi, w parametrze *rodzaj–liczba* zamiast kropki mamy ukośnik `/`. Pozostałe różnice, zwłaszcza oznaczenia wartości kategorii rodzaju i typy wymagania składniowego, są omówione niżej.

Symbolowi `znakkonca(p)` odpowiada w prawej kolumnie oznaczenie `int1`, co oznacza, że zastosowano regułę (Świdziński 1992:432)

```
ZNAKKOŃCA (P) = # ? .  (INT1)
```

Reguła ta demonstruje przejście symbolu nieterminalnego na terminalny. W tym wypadku jest to po prostu znak zapytania, wypisany w drzewie pod symbolem `znakkonca(p)` z odpowiednim wcięciem. Warto zwrócić uwagę, że wiersze odpowiadające symbolom terminalnym nie mają nic w prawej kolumnie, co ułatwia ich dostrzeżenie w skomplikowanym drzewie.

W analogiczny sposób możemy odczytać, że zdanie równorzędne przeszło na zdanie szeregowe za pomocą reguły R1 (Świdziński 1992:335):

```
ZR (wf, a, c, t, rl, o, neg, i, z)
     = ZSZ (wf, a, c, t, rl, o, neg, i, z).                (R1)
```

Analogicznie zdanie szeregowe przechodzi na zdanie jednorodne za pomocą reguły S1, zdanie jednorodne na zdanie proste za pomocą reguły J1, zdanie proste na zdanie elementarne za pomocą reguły P1.

Na następnym poziomie drzewa napotykamy rozgałęzienie będące skutkiem zastosowania reguły E1 (Świdziński 1992:356):

```
ZE (wf, a, c, t, rl, o, wa, wb, wc, neg, i, z, ow)
     = FL (a, c, rl, o, neg, i, z1)                          (E1)
       < FF (wf, a, c, t, rl, o, wa, wb, wc, k, neg, NI, z, ow)
         FW (wa, k, a, c, rl, o, neg, NI, z2)
         FW (wb, k, a, c, rl, o, neg, NI, z3)
         FW (wc, k, a, c, rl, o, neg, NI, z4) >
       $ RÓWNE (z, P.P'.P")
       $ RÓWNE (z1, P.NP)
       $ RÓWNE (z2, P.NP)
       $ RÓWNE (z3, P.NP)
       $ RÓWNE (z4, P.NP)
```

Jest to jedna z bardziej skomplikowanych reguł, dopuszcza bowiem dowolną kolejność składników ujętych w nawiasy kątowe, a także nakłada pewne warunki na wartości parametrów.

Przyglądając się frazie luźnej (`fl`) stanowiącej pierwszy składnik zdania elementarnego zauważymy, że przechodzi ona kolejno na symbole nieterminalne noszące umowne nazwy frazy luźnej właściwej (`fl1`), frazy przysłówkowej (`fps`), konstrukcji przysłówkowej z frazą przyimkową (`kpspm`), konstrukcji przysłówkowej z frazą przysłówkową (`kpsps`), frazy przysłówkowej z inkorporacją (`kpsink`), konstrukcji przysłówkowej (`kprzysl`), zaimka pytajnego (`zaimpyt`) i w końcu na tzw. jednostkę elementarną w postaci zaimka przysłownego (`zaimprzys`). Jednostki elementarne z reguły przechodzą bezpośrednio na symbole terminalne, w tym wypadku na wyraz *Gdzie*. Ponieważ konstrukcja przysłówkowa (`kprzysl`) przechodzi przy tym na zaimek pytajny (`zaimpyt`) zgodnie z wymienioną w nagłówku regułą PS19, a następnie na wyróżniony w danych wejściowych napis *¡gdzie¿*, potwierdza to, że omawiana analiza jest zgodna z intencjami autora gramatyki.

Drugim składnikiem zdania elementarnego jest fraza finitywna reprezentowana przez symbol nieterminalny

`ff(os,nd,ter,ozn,_2624/mno,2,[np(bier)],_2018,tak,ni,p,br)`

Widzimy tu najbardziej istotne odstępstwo od oryginalnej gramatyki, a mianowicie inną reprezentację fraz wymaganych przez czasownik, i w konsekwencji inną reprezentację elipsy (pominięcia) frazy wymaganej. W cytowanej wyżej regule E1 widać, że na wymagania czasownika zarezerwowane są trzy parametry: `wa`, `wb`, `wc`. W oryginalnym drzewie analizy — które pod tym względem było wierniej reprezentowane w systemie AMOS, por. (Bień 1996:158) — z wierzchołka reprezentującego zdanie elementarne wychodzą zawsze trzy gałęzie reprezentujące frazy wymagane, przy czym dla naszego przykładu dwie z nich miałyby realizacje puste. W systemie AS zamiast trzech parametrów mamy jeden, którego wartością jest lista wymagań (listy zapisujemy — zgodnie z konwencją języka Prolog — w nawiasach prostokątnych). Dla konkretnej analizy lista ta zawiera tylko te wymagania, które są rzeczywiście zrealizowane w danym wypadku, i w konsekwencji drzewo analizy nie zawiera również gałęzi odpowiadających frazom wymaganym o realizacji pustej. Za takim rozwiązaniem przemawiały przede wszystkim względy techniczne, ale wydaje się, że zyskała na tym również czytelność wyników.

Zmianie uległa też postać zapisu wymagań, która jest inspirowana skrótami stosowanymi w podręczniku (Saloni, Świdziński 1998). Tak więc w omawianym wierszu widzimy wymagania czasownika w postaci listy zawierającej tylko jeden element `np(bier)`, który oznacza frazę nominalną (`np` jest skrótem o międzynarodowym charakterze angielskiego *noun phrase*) w bierniku. Widzimy też symbol `br` oznaczający brak tzw. ograniczenia wewnętrznego i nieokreśloną wartość kategorii korelatywności (obie te kategorie są specyficzne dla gramatyki Świdzińskiego); znaczenie pozostałych parametrów jest identyczne, jak w symbolu zdania równorzędnego. Jawne oznaczanie braku ograniczenia wewnętrznego za pomocą symbolu `br` jest odstępstwem od oryginalnej notacji i zostało wprowadzone ze względów technicznych.

Fraza finitywna `ff` przechodzi kolejno na symbole o nazwach: fraza finitywna właściwa (`ff1`), fraza werbalna (`fwe`), konstrukcja werbalna z negacją (`kweneg`), konstrukcja werbalna z inkorporacją (`kweink`), konstrukcja werbalna (`kwer`) i konstrukcja werbalna właściwa (`kwer1`). Następnym etapem jest jednostka elementarna w postaci formy czasownikowej (`formaczas`), która w gramatyce Świdzińskiego jest opisana nieformalnie; konkretne reguły dla niej — i dodatkowe symbole nieterminalne jak `formaczas1` — zostały zdefiniowane specjalnie na potrzeby analizatora; dla odróżnienia ich oznaczenia zaczynają się od napisu `n_`. Ostatecznie fraza finitywna przechodzi na wyraz *macie*. Ani z formy czasownikowej *macie*, ani z kontekstu nie można wywnioskować wartości rodzaju, dlatego cały czas jest ona reprezentowana przez zmienną wolną.

Trzeci składnik zdania elementarnego to fraza wymagana (`fw`). Przechodzi ona kolejno na frazę wymaganą właściwą (`fw1`), frazę nominalną (`fno`), konstrukcję nominalną z dopełniaczem (`knodop`), konstrukcję nominalną z frazą przyimkową (`knopm`), konstrukcję nominalną z atrybutem (`knoatr`), konstrukcję nominalną (`knom`) i jednostkę elementarną w postaci formy rzeczownikowej

(`formarzecz`), ta zaś z kolei na symbol terminalny w postaci wyrazu *szałas*. Warto zwrócić uwagę na to, że w ogólnym wypadku słowo to może stanowić również formę mianownika, tutaj zaś — dzięki analizie składniowej — jest ono poprawnie opisane jako biernik; rzeczownik ten jest rodzaju męskorzeczowego, który za Świdzińskim oznaczamy skrótem *mnż* (od *męski nieżywotny*).

Obecnie omówimy drzewo analizy przykładu nr 304

Choćbym przyszedł, nie zostawaj.

Ma ono o wiele bardziej skomplikowaną strukturę, dlatego analizator znalazł aż 20 możliwych interpretacji, a drzewa analizy są bardziej obszerne. Tutaj omówimy interpretację dwunastą (analiza ta znajduje się w katalogu `DRZEWA.PDF` w pliku `ap0200d.pdf` na stronie 1995).

W poniższym omówieniu skoncentrujemy się na najciekawszych elementach drzewa analizy syntaktycznej, a mianowicie na węzłach z rozgałęzieniami. Pierwszy taki węzeł odpowiada wypowiedzeniu, drugi — zdaniu elementarnemu. Jak widać, zdanie elementarne jest zbudowane zgodnie z regułą E4 (Świdziński 1992:357):

```
ZE (wf, a, c, t, rl, o, wa, wb, wc, neg, i, z, ow)
     = FL (a, c, rl, o, neg, i, NP)                          (E4)
       < FF (wf, a, c, t, rl, o, wa, wb, wc, k, neg, NI, z, ow)
         FW (wa, k, a, c, rl, o, neg, NI, NP)
         FW (wb, k, a, c, rl, o, neg, NI, NP)
         FW (wc, k, a, c, rl, o, neg, NI, NP) >
       $ RÓŻNE (z,   BY".CHOĆBY.CO.CZYŻBY.GDYBY.JAKBY.JAKI.JAKOBY.
                KTO.KTÓRY.P.P'.P".PZ.ŻEBY)
```

W drzewie symbol `ze` (i kilka innych) ma dodatkowy parametr w formie samego podkreślenia; w ten sposób jest widoczna zmienna „anonimowa" wprowadzona ze względów technicznych.

Pierwszy składnik zdania elementarnego, a mianowicie fraza luźna (`fl`), przechodzi na frazę luźną właściwą (`fl1`), a ta na frazę zdaniową (`fzd`). Fraza zdaniowa przechodzi na frazę zdaniową szeregową, frazę zdaniową jednorodną i frazę zdaniową z korelatem (`fzdkor`).

Pierwszy składnik frazy zdaniowej z korelatem to `przecsp`, ktory może być realizowany jako przecinek lub spójnik; w naszym przypadku to przecinek. Zgodnie z gramatyką jest to przecinek o realizacji pustej, ponieważ rozpoczęcie zdania od przecinka jest błędem ortograficznym. Ze względów technicznych (strategia *bottom-up* nie obsługuje reguł o pustych prawych stronach) jest on jednak zapisany w drzewie w taki sam sposób, jakby rzeczywiście wystąpił w danych wejściowych.

Drugi składnik frazy zdaniowej z korelatem do fraza zdaniowa elementarna. Stosuje się do niej reguła ZD43 (Świdziński 1992:419):

```
FZDE (tfz, a, c, t, neg, i)
     = SPÓJ (PO, tfz, NI)                                    (ZD43)
       ZR (wf, a, c, t, rl, o, neg, i, tfz)
```

```
$ RÓWNE (tfz, CHOĆBY.GDYBY.JAKBY.JAKOBY.ŻEBY).
```

Fraza ta składa się więc ze spójnika oraz zdania równorzędnego; na uwagę zasługuje fakt, że pojedyncze słowo *Choćbym* jest w ten sposób rozbite na spójnik *Choćby* i element aglutynacyjny należący do zdania równorzędnego. Rekurencyjne użycie symbolu zdania równorzędnego zostanie przekształcone kolejno na symbol zdania szeregowego, zdania jednorodnego, zdania prostego i wreszcie zdania elementarnego zbudowanego zgodnie z regułą E7 (Świdziński 1992:358):

```
ZE (wf, a, c, t, rl, o, wa, wb, wc, neg, i, z, ow)
     = FL (a, c, rl, o, neg, i, z)                          (E7)
       < FF (wf, a, c, t, rl, o, wa, wb, wc, k, neg, NI, z, ow)
         FW (wa, k, a, c, rl, o, neg, NI, NP)
         FW (wb, k, a, c, rl, o, neg, NI, NP)
         FW (wc, k, a, c, rl, o, neg, NI, NP) >
       $ RÓWNE (z, BY".CHOĆBY.CZYŻBY.GDYBY.JAKBY.JAKOBY. ŻEBY).
```

Fraza luźna stanowiąca pierwszy składnik zdania elementarnego przejdzie na frazę właściwą, a następnie na aglutynant zgodnie z regułą LU8 (Świdziński 1992:370):

```
FL1 (a, c, rl, o, neg, i, z)
     = AGL (rl, o, i)                                       (LU8)
       $ RÓWNE (z, BY".CHOĆBY.CZYŻBY.GDYBY.JAKBY.JAKOBY.ŻEBY).
```

Następnie aglutynant przechodzi na aglutynant właściwy (Świdziński 1992:431):

```
AGL (rl, o, NI) = AGL1 (rl, o).                             (AGL2)
```

Kolejny krok to zastąpienie aglutynantu właściwego jednostką elementarną w postaci morfemu aglutynacyjnego:

```
AGL1 (rl, 1)    = MORFAGL (M, rl, 1)                        (AGL3)
                $ RÓWNE (rl, (MOS.POJ).(MZW.POJ).(MNŻ.POJ).
                             (ŻEŃ.POJ).(NIJ.POJ))
```

Następny etap to oczywiście już symbol terminalny w postaci słowa *m*.

Ponieważ z nagłówka naszej analizy wynika, że omawiane zdanie ilustruje właśnie regułę LU8 odnoszącą się do wyróżnionego w danych wejściowych napisu *¡m¿*, wiemy dzięki temu, że co najmniej omówiony fragment analizy jest zgodny z intencjami autora gramatyki.

Warto zwrócić uwagę, że w omówionym wyżej fragmencie drzewa rodzaj jest oznaczony symbolem `m`. Jest to wprawdzie zgodne z (Bień 1991), ale symbol taki nie występuje w gramatyce Świdzińskiego. Należy go traktować jako oznaczenie nie konkretnego rodzaju, ale zbioru trzech wartości tej kategorii, mianowicie rodzaju męskosobowego, męskozwierzęcego i męskorzeczowego. Jest to jedyny istotny wyjątek od zasady stosowania oznaczeń zgodnych z książką (Świdziński 1992); można tę konwencję traktować jako zwięzłą reprezentację kilku drzew analizy różniących się tylko w określonych miejscach wartościami kategorii rodzaju.

6.5 Wyniki analizy syntaktycznej

Przetworzeniu przez analizator AS poddaliśmy wszystkie przykłady z korpusu gramatyki Świdzińskiego o łącznej liczbie 2037, które były traktowane przez analizator jako samodzielne wypowiedzenia. Dla każdego zaakceptowanego przykładu analizator znalazł od kilku do kilku tysięcy możliwych drzew analizy syntaktycznej.

W przypadku bardzo dużej liczby znalezionych drzew interpreter Prologu przerywał pracę analizatora (prawdopodobnie z powodu przekroczenia jego ograniczeń pamięciowych); miało to miejsce dla 5% przykładów podkorpusu GFJP-A i 8% przykładów podkorpusu GFJP-B (łącznie dla 146 zdań). Przerwanie pracy analizatora uniemożliwiało wypisanie informacji takich jak np. czas analizy, dlatego większość danych liczbowych odnosi się do pozostałych 1991 przykładów, dla których praca analizatora zakończyła się normalnie. Ich przetworzenie zajęło łącznie 711 037 sekund (około 197 godzin) czasu procesora na komputerze PC z procesorem Pentium II 266MHz.

Liczba drzew wyprodukowanych dla podkorpusu GFJP-A wynosi 37 204, dla podkorpusu GFJP-B jest równa 89 461, co daje łącznie 126 665. Średnia liczba drzew dla jednego przykładu jest praktycznie nieistotna, ponieważ jest znacznie zawyżona przez nieliczne przypadki skrajne; korzystając z pojęcia mediany możemy natomiast powiedzieć, że liczba drzew dla typowego przykładu jest rzędu 10 (interesujące jest też, że dla przykładów nieprzerwanych typowy czas analizy jest rzędu 10 sekund).

Dla 515 zdań określonych w podkorpusie GFJP-A jako poprawne zostały zaakceptowane 275 zdania (53%). Przyczyny niepowodzeń przy analizie pozostałych 47% tych zdań wymagają szczegółowej analizy. Można mieć nadzieję, że w większości wypadków powodem jest fakt, że analizator AS dysponował bardzo ubogim słownikiem własności gramatycznych. Oczywiście, niewykluczone są również błędy w implementacji gramatyki. Dalszy rozwój analizatora AS powinien doprowadzić do tego, że akceptowanych będzie praktycznie 100% zdań określonych jako poprawne w GFJP-A; jedynym powodem braku akceptacji takiego zdania mogą być tylko ewentualne pomyłki samego autora gramatyki (np. jakiś subtelny błąd w regułach lub nieświadome zilustrowanie pewnej reguły zbyt skomplikowanym przykładem).

Dla 1082 zdań określonych w podkorpusie GFJP-B jako poprawne zostało zaakceptowanych 356 zdań (33%). Tutaj również dalszy rozwój analizatora — a przede wszystkim jego słownika gramatycznego — powinien doprowadzić do znacznego zwiększenia liczby akceptowanych zdań. Ta część korpusu zawiera jednak zdania poprawne, o ktorych z góry wiadomo, że nie są z założenia opisywane przez gramatykę, osiągnięcie więc 100% akceptowalności nie jest możliwe.

Choć niski procent akceptowalności zdań poprawnych trochę nas rozczarował, nie jest to problem tak poważny, jak zalew interpretacji błędnych lub nadmiarowych, co czyni merytoryczną analizę otrzymanych wyników nadzwyczaj trudną.

Pewna liczba alternatywnych drzew jest nie do uniknięcia — istnieją zdania wieloznaczne również dla człowieka, a często wieloznaczności syntaktyczne

rozstrzygane są na podstawie znaczenia zdania, do którego nasz analizator nie ma oczywiście dostępu. Skala wykrytych przez analizator wieloznaczności jest jednak dla nas dużym zaskoczeniem[7]. Jak się wydaje, problem ten powinien być rozwiązywany dwutorowo.

Po pierwsze, niezbędne jest stworzenie narzędzi informatycznych do obsługi bazy danych zawierających drzewa analizy syntaktycznej; niezależnie od ich wartości merytorycznej uzyskany zbiór ponad stu tysięcy drzew stanowi dobry materiał do testowania tego typu narzędzi. Narzędzia takie powinny przede wszystkim pozwalać na wyszukiwanie drzew o odpowiednich własnościach. Przykładem prostej kwerendy może być np. zapytanie: jaki przykład ilustruje regułę *we31*, przy czym wartość parametru zależności równa jest *p"*? Przykładem kwerendy bardziej skomplikowanej może być zapytanie: w których drzewach co najmniej jedna reguła jest użyta rekurencyjnie? Bardzo pożądana jest też możliwość sortowania drzew według różnorodnych kryteriów, takich jak podobieństwo struktury, występowanie części wspólnych itp.

Po drugie, niezbędna jest lingwistyczna analiza przyczyn pojawiania się nadmiarowych drzew. Wymienimy tutaj dwie takie przyczyny, ale stopień, w jakim zwiększają one liczbę zbędnych drzew, jest dla nas jeszcze nieznany.

Pierwsza przyczyna to wspomniany już wcześniej brak dostatecznie obszernego słownika własności gramatycznych, który powinien w szczególności podawać tzw. wymagania czasownikowe (upraszczając sprawę, czasownik *kupić* wymaga *coś za coś*, *mówić* wymaga m.in. *o czymś* itp. — są to indywidualne własności poszczególnych czasowników). Rozpoczynając projekt mieliśmy nadzieję, że przynajmniej niektóre własności składniowe zostaną wywnioskowane przez analizator z kontekstu użycia słowa. Wyniki eksperymentu pokazały, że nie jest to możliwe (przynajmniej przy obecnie stosowanej gramatyce).

Druga przyczyna to pietyzm, z jakim odnosiliśmy się do oryginalnej gramatyki Świdzińskiego. Ze względów dydaktycznych chcieliśmy w równym stopniu pokazać, że prawidłowo opisuje ona złożone zdania i konstrukcje, jak i unaocznić praktyczne konsekwencje pewnych luk i niekonsekwencji w gramatyce, i to również tych, które przez czytelnika książki mogą być uznane za uprawnione skróty myślowe. Jaskrawym przykładem jest sprawa słowa *nie*, dla którego analizator morfologiczny podaje dwie interpretacje: partykuła zaprzeczenia i poprzyimkowa forma zaimka *one* (np. *Patrzę na nie.*). Niestety, dla gramatyki Świdzińskiego forma *nie* zaimka *one* jest nieodróżnialna od formy *je*[8] (np. *Widzę je*), w związku z czym każdy zaprzeczony czasownik może mieć alternatywną interpretację jako rzeczownikowa fraza luźna w bierniku z niezaprzeczonym czasownikiem. Dla przykładu, zdanie *Nie czytam* oprócz swojej naturalnej interpretacji otrzyma również drugą, w której *nie* jest traktowane analogicznie, jak słowo *godzinę* w zdaniu *Czytam godzinę*. Obecnie ten cel dydaktyczny został już osiągnięty i nic nie stoi na przeszkodzie, aby gramatykę Świdzińskiego odpowiednio uzupełnić.

[7]Nie pomogły nam tutaj doświadczenia z systemem AMOS, ponieważ wtedy zadowalaliśmy się znalezieniem tylko jednego drzewa analizy.

[8]Swoją drogą, wyraz *je* może być również formą czasownika *jeść* — tego typu wieloznaczności są znacznie częstsze niż to się może wydawać, i też mogą prowadzić do alternatywnych interpretacji.

Oprócz tych oczywistych przyczyn występowania błędnych i nadmiarowych interpretacji istnieją zapewne jeszcze inne, których zidentyfikowanie wymaga wnikliwej analizy otrzymanych wyników.

Wobec przedstawionych wyżej faktów uznaliśmy za niecelowe poddawanie analizie syntaktycznej pozostałych korpusów. Choć od strony technicznej nie powinno to sprawiać zasadniczych trudności, byłoby to zadanie czasochłonne i uciążliwe, a jednocześnie dostarczające w znikomym tylko stopniu takich informacji, które nie są już zawarte w wynikach otrzymanych dla korpusu gramatyki Świdzińskiego.

Pliki z wynikami analizy syntaktycznej znajdują się na płycie w katalogu `DRZEWA.PDF` i — jak wskazuje rozszerzenie nazwy katalogu — są one zapisane w formie *Portable Document Format* firmy Adobe. Do ich przeglądania należy w związku z tym stosować bezpłatny program Acrobat Reader w wersji 3 lub wyższej. Warto podkreślić, że program ten jest dostępny również dla MS Windows 3.x, zatem uzyskane wyniki mogą być wykorzystywane rownież przez lingwistów nie dysponujących nowoczesnym sprzętem.

Pliki z wynikami mają nazwy zbudowane zgodnie ze schematem `?p????d.pdf`; pierwsza litera wskazuje, czy chodzi o podkorpus A czy B, cztery cyfry podają natomiast numer pierwszego przykładu, który znajduje się w danym pliku.

Zawarte w plikach drzewa są wizualizowane zgodnie z opisaną wyżej konwencją, a nawigację ułatwiają zakładki pozwalające szybko przejść do analizy następnego przykładu. Niestety, wykorzystana wersja formatu PDF nie dopuszcza stosowania liter narodowych w zakładkach, które w związku z tym na pierwszy rzut oka mogą wydawać się niepoprawne (ograniczenie to zostało usunięte bardzo niedawno, stąd obawa, że użycie najnowszej wersji formatu zawęzi dostęp do wyników).

Oczywiście, zawarte w plikach PDF wyniki można drukować — również w kolorze. Należy jednak pamiętać, że czytelny wydruk analizy skomplikowanych przykładów może wymagać stosowania papieru w formacie A3.

Na płycie znajduje się również analiza statystyczna uzyskanych wyników. Jest to tekst autorstwa mgr. Łukasza Dębowskiego zapisany w katalogu `DODATEK` jako plik `dodatek.rtf`. Wykorzystane wyżej dane liczbowe pochodzą właśnie ze wspomnianego opracowania.

7 Wykorzystanie wyników projektu

Zgromadzone przez nas korpusy stanowią przede wszystkim doskonały materiał do testowania analizatorów syntaktycznych. Są one na tyle obszerne i różnorodne, że ich obiektywny charakter nie może być kwestionowany. Ich dodatkową wartością jest to, że te z nich, dla których było to możliwe, zostały udokumentowane przez udostępnienie elektronicznych wersji publikacji, z których zostały one zaczerpnięte.

Szczególnie cenne są te korpusy, które oprócz zdań poprawnych zawierają również odpowiednio oznaczone zdania niepoprawne, co umożliwia dokonywanie oceny adekwatności lingwistycznej analizatorów zgodnie z metodą zapro-

ponowaną przez Mirosława Bańkę (1985, 1990), por. także (Bień 1998). Jak pokazał Bańko, przy spełnieniu odpowiednich założeń metoda ta może służyć do obiektywnego porównywania analizatorów nawet wtedy, gdy są one oparte na odmiennych teoriach lingwistycznych.

Utworzony dla jednego z korpusów bank rozbiorów gramatycznych stanowi znaczący krok w kierunku zrealizowania naszych planów przedstawionych w referatach na międzynarodowych konferencjach w Grenadzie (Bień 1998) i w Lipsku (Bień, Szafran, Woliński w przygotowaniu). Chodzi mianowicie o wnikliwe porównanie gramatyki Szpakowicza i Świdzińskiego zarówno pod względem ich adekwatności lingwistycznej w sensie Bańki, jak też biorąc pod uwagę inne aspekty, np. wpływ stylu pisania gramatyki na efektywność opartego na niej analizatora. Jak się nam wydaje, taka głęboka analiza aktualnie istniejących gramatyk formalnych istotnie ułatwi stworzenie nowej gramatyki, odpowiadającej współczesnym potrzebom i uwzględniającej aktualny stan wiedzy lingwistycznej.

Niezależnie od tych planów bank rozbiorów gramatycznych ma już obecnie bardzo duże znaczenie dydaktyczne, pozwala bowiem studentom lingwistyki w wygodny sposób zapoznać się z praktycznym działaniem reguł gramatyki Świdzińskiego; mogą oni w ten sposób zarówno pogłębiać swoją znajomość tej konkretnej gramatyki, jak i uczyć się zasad tworzenia formalnych gramatyk języka naturalnego.

8 Podsumowanie

Postawione cele zostały osiągnięte pomimo napotkanych trudności. Mamy nadzieję, że stworzony przez nas zestaw testów będzie pożyteczny nie tylko dla prac prowadzonych przez zespoły związane z wykonawcami tego projektu, ale zostanie uznany za przydatny również dla innych zespołów, zarówno nowo powstających, jak i konkurencyjnych.

Stworzenie korpusu rozbiorów syntaktycznych udowodniło, że następujące stwierdzenie Marka Świdzińskiego — sformułowane ponad 10 lat temu (Świdziński 1987), ale powtórzone również później (1992:58):

> Opis przedstawiony w niniejszej pracy ukierunkowany jest w większym stopniu lingwistycznie (empirycznie) niż informatycznie. Przyjmuję tutaj tak wysoki stopień szczegółowości empirycznej, że bezpośrednia implementacja nawet fragmentów podanej w tej pracy gramatyki nie wydaje się możliwa.

nie było słuszne. Mamy nadzieję, że opisane wyniki pozwolą rozwiać mogące jeszcze pokutować jeszcze wśrod lingwistów wątpliwości co do przydatności komputerów do weryfikowania złożonych teorii lingwistycznych. Dzięki udostępnieniu przykładowych rozbiorów gramatycznych w formie plików PDF, które można oglądać bez trudu nawet na wolnych i przestarzałych komputerach PC, wyniki te mają szanse dotrzeć do szerokiego kręgu lingwistów, również tych, których umiejętności informatyczne są wiecej niż skromne.

Trzeba także stwierdzić, że stworzony w ramach projektu bank rozbiorów gramatycznych stanowi jeden z pierwszych takich banków dla języków słowiańskich.

Na zakończenie chcielibyśmy także podkreślić, że najważniejsze wyniki projektu zostaną udostępnione w Internecie (pod adresem `ftp://ftp.mimuw.edu.pl/pub/users/polszczyzna`), a ich dostępność będzie zaanonsowana na krajowej liście pocztowej poświęconej przetwarzaniu języka naturalnego `nlp-l@uci.agh.edu.pl` oraz na międzynarodowej liście pocztowej dotyczącej formalnego opisu języków słowiańskich `fdsl@main.amu.edu.pl`.

9 Literatura cytowana

Bańko, M. 1985. Analiza polskich fraz rzeczownikowym testem adekwatności i efektywności parsera Szpakowicza. Praca magisterska (opiekun J. S. Bień), Instytut Informatyki UW 1985.

Bańko, M. 1990. Niektóre problemy oceny adekwatności gramatyk (na przykładzie fragmentu gramatyki Szpakowicza). *Studia Gramatyczne* IX (1990), s. 55-72.

Bień, J.S. 1991. *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*. Wydawnictwa Uniwersytetu Warszawskiego, Warszawa 1991.

Bień, J. S. 1996. Komputerowa weryfikacja opisu składni polskiej. Raport Instytutu Informatyki Uniwersytetu Warszawskiego TR 96–06 (227), maj 1996. `ftp://ftp.mimuw.edu.pl/pub/users/polszczyzna/AMOS-95`

Bień, J.S. 1996a. Komputerowa weryfikacja formalnej gramatyki Świdzińskiego. *Biuletyn Polskiego Towarzystwa Językoznawczego*, zeszyt LII (1996), s. 147–164.

Bień, J. S. 1996b. Processing Polish with metamorphosis grammars. Tübingen Workshop on Computational Linguistics, Tübingen, 15–28 September 1996, `ftp://ftp.mimuw.edu.pl/pub/users/jsbien/tybinga96/pol-mg.ps`.

Bień, J.S. 1998. Evaluating Analysers of Polish. In A. Rubio et al. (eds.), Proceedings of First International Conference on Language Resources and Evaluation, European Language Resource Association: Grenada 1998, pp. 951–955.

Bień, J.S., Szafran, K., Woliński, M., w przygotowaniu. An experimental analyser of Polish. Proceedings of the Third European Conference on Formal Description of Slavonic Languages, Lepizug, 1–3 December 1999.

Caroll J. et al. 1978. Caroll J., Basili R., Calzolari N., Gaizauskas R., Grefenstette G. (organisers), Proceedings of the Workshop on the Evaluation of Parsing Systems at the first International Conference on Language Resources and Evaluation, Grenada, Spain, May 26, 1998.

Colmerauer, A. 1978. Metamorphosis grammar. In: L. Bolc (ed), *Natural Language Communication with Computers*, Lecture Notes in Computer Science 63, Springer-Verlag 1978, pp 133-189.

Erjavec, T., Lawson, A. (eds) 1998. *East meets West — A Compendium of Multilingual Resources.* CD-ROM. TELRI Association [1998].

Gazdar, G., Mellish, Ch. 1989. *Natural Language Processing in PROLOG.* Addison–Wesley: Wokingham 1989.

Głowińska, K., Woliński, M. w druku. Angielsko-polski słownik elektroniczny XeLDA. Materiały międzynarodowej konferencji *Problemy leksykografii dwujęzycznej w językach słowiańskich i bałkańskich*, Toruń, 25-26 maja 1999, w druku.

Goldfarb, Ch. F. 1990. *The SGML Handbook.* Clarendon Press 1990.

Hajič, J. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. *Issues of Valency and Meaning (Festschrift for Jarmila Panevová).* Carolina, Charles University, Prague 1998, pp. 106-132.

ISO 8879:1986 *Information Processing – Text and Office Systems – Standard Generalized Markup Language (SGML).*

Kluźniak, F, Szpakowicz, S. 1983. *Prolog.* WNT 1983.

Kluźniak, F, Szpakowicz S. 1985 *Prolog for Programmers.* Academic Press 1985, (dodruk z dyskietką 1987).

Kurcz i inni 1990. I. Kurcz, A. Lewicki, J. Sambor, K. Szafran, J. Woronczak, *Słownik frekwencyjny współczesnej polszczyzny pisanej.* Instytut Języka Polskiego PAN, Kraków 1990.

Leech, G., & Garside, R. (1991). Running a GRAMMAR factory: The production of syntactically analysed corpora or 'treebanks'. In S. Johansson & A. B. Stenströvm (Eds.), *English computer corpora: Selected papers and research guide* (pp. 15-32). New York: Mouton de Gruyter.

Lehmann S. et al, 1996. *TSNLP — Test Suites for Natural Language Processing.* International Conference on Computational Linguistics COLING 96.

Pietrzak, M. 1999. Wykorzystanie systemu LaTeX do tworzenia publikacji elektronicznych. Praca magisterska (opiekun J. S. Bień), Instytut Informatyki UW 1999.

Rubio, A., Gallardo, N., Castro, R., Tejada, A. (eds). Proceedings of First International Conference on Language Resources and Evaluation. European Language Resources Association: Grenada 1998.

Saloni, Z., Świdziński, M. 1998. *Składnia współczesnego języka polskiego.* Wydanie czwarte, zmienione. Wydawnictwo Naukowe PWN: Warszawa 1998.

Sperberg-McQueen, C. M., & Burnard, L. (Eds.). (1994). *Guidelines for electronic text encoding and interchange (TEI P3).* Text Encoding Initiative, Chicago-Oxford 1994.

Szafran, K. 1996. *Analizator morfologiczny SAM-95 — opis użytkowy.* Raport Instytutu Informatyki Uniwersytetu Warszawskiego TR 96–05 (226), maj 1996. `ftp://ftp.mimuw.edu.pl/pub/users/polszczyzna/SAM-95`

Szafran, K. 1997. Automatyczne hasłowanie tekstu polskiego. *Polonica*, tom XVIII. IJP PAN: Kraków 1997, s. 51–63.

Szpakowicz, S. 1978. Automatyczna analiza składniowa polskich zdań pisanych. Praca doktorska (promotor S. Waligórski), Instytut Informatyki UW 1978.

Szpakowicz, S. 1986. *Formalny opis składniowy zdań polskich*, 2. wyd. Wydawnictwa Uniwersytetu Warszawskiego 1986.

S. Szpakowicz, M. Świdziński, 1990. *Formalna definicja równorzędnej grupy nominalnej we współczesnej polszczyźnie pisanej.* Studia Gramatyczne IX (1990), s. 9-54.

Świdziński, M. 1987. Formalny opis polskich zdań o składniku zdaniowym. Praca habilitacyjna. Maszynopis powielony, Instytut Języka Polskiego UW, Warszawa 1987.

Świdziński, M. 1992. *Gramatyka formalna języka polskiego.* Wydawnictwa Uniwersytetu Warszawskiego 1992.

Świdziński, M. 1993. Od interpretacji do faktów językowych: weryfikacja empiryczna gramatyki formalnej. *Biuletyn Polskiego Towarzystwa Językoznawczego* zeszyt XLIX, s. 15–24.

Świdziński, M. 1996. *Własności składniowe wypowiedników polskich.* Warszawa 1996.

Vetulani, Z. *Corpus of consultative dialogues.* Experimentally collected source data for A[rtificial]I[ntelligence] application. UAM Press, Poznań 1990.