Gramatyki wykorzystywane w analizie języka naturalnego

Ciekawe zastosowania

Anna Sikora <sikora.anna@gmail.com>
Janusz Głowiak <janusz.glowiak@gmail.com>

Gramatyka

W zakres gramatyki wchodzą:

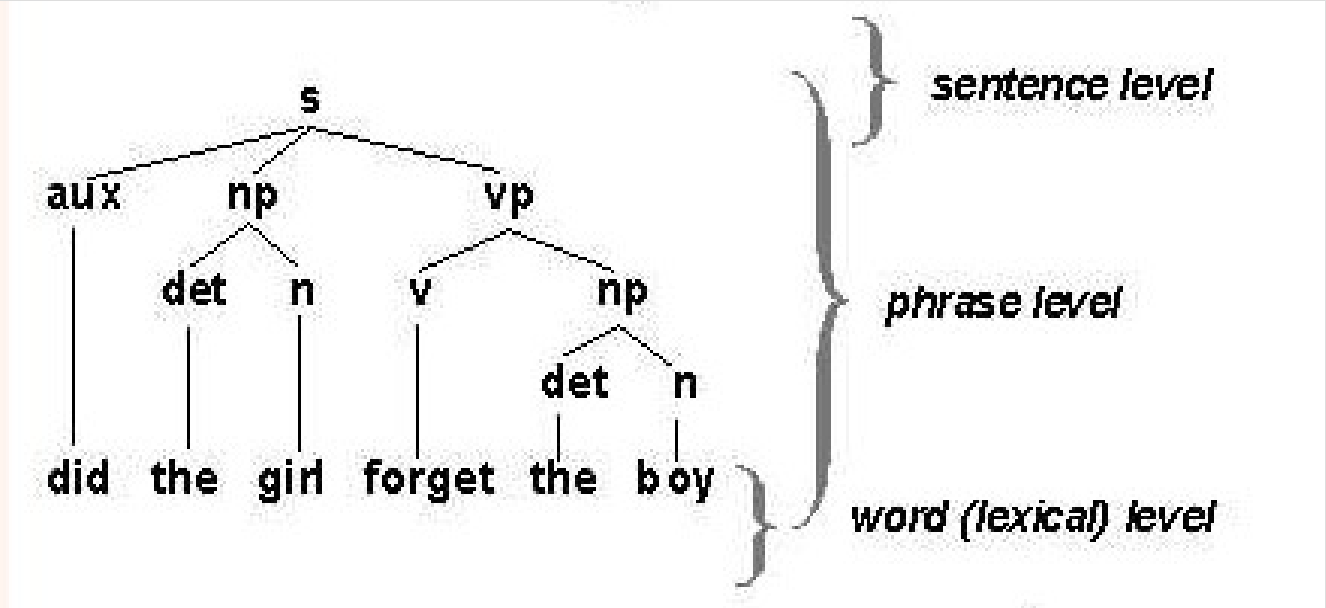
Leksykologia - nauka o słownictwie, o zasobie wyrazów i związków wyrazowych.
Semantyka - zajmuje się badaniem znaczenia słów, czyli interpretacją znaków oraz interpretacją zdań i wyrażeń języka.
Składnia - bada zasady, na jakich wyrazy łączone są w dłuższe wypowiedzi, na przykład zdania.
Morfologia – opisuje sposób, w jaki morfemy (części słów) mogą się ze sobą łączyć.
Fonetyka - jeden z działów lingwistyki zajmujący się badaniem dźwięków mowy ludzkiej.

Struktura językowa

Poziomy struktury językowej wg Chomsky'ego:

poziom wypowiedzi (discourse level) - zdania wymieniane pomiędzy dwiema osobami
poziom akapitu (paragraph level) - zdania połączone są w jedno zdanie za pomocą znaków przestankowych (kropka, znak zapytania, wykrzyknik) pomiędzy nimi
Przysłówki mogą występować aby wskazać na logiczne powiązanie pomiędzy zdaniami
poziom zdania (sentence level) - proste zdanie składa się z podmiotu i orzeczenia. Zdanie złożone składa się z dwu lub więcej prostych zdań połączonych spójnikiem współrzędnym (np. i, ale) lub spójnikiem podrzędnym (np. że)
poziom frazy (phrase level) - fraza składa się z jednsostki leksykalnej i powiązanych z nią modyfikatorów, np the, a... poprzedzajacych rzeczownik; very, too przed przymiotnikiem; will, can przed czasownoikiem.
poziom słowa (word level) - wszystko, co występuje w zdaniu i posiada białe znaki po obu stronach.
poziom morfemu (morpheme level) - morfem jest najmniejszą jednostką funkcjonalną języka.
poziom fonetyczny (ortography/phonetic level) - elementy są znakami pisarskimi albo dźwiękami języka

Struktura językowa



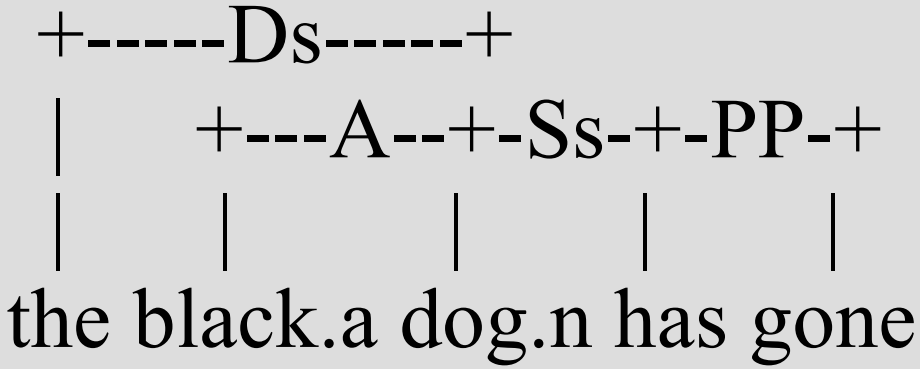
undoable:
morpheme level -> un-do-able
phonetic level -> u-n-d-o-a-b-l-e

Link grammar

- Gramatyka zajmująca się nie konkretnymi częściami zdania, ale relacjami między poszczególnymi słowami.
- Każdy typ słów ma swój zbiór połączeń do innych słów który może lub musi być wykorzystany
- Połączenia mają swój zwrot (przed lub za) oznaczonyy za pomocą znaków +/- oraz priorytet -wcześniej położone muszą się znajdować bliżej danego słowa. Priorytety liczą się tylko w ramach połączeń znajdujących się po jednej stronie danego wyrazu.
- Między połączeniami mogą występować operatory logiczne (i, lub, nie)

Link Grammar- przykład

Zdanie: “The black dog has gone.”



- Ds – połączenie rzeczownika z rodzajnikiem
- A – połączenie rzeczownika z przymiotnikiem stojącym przd nim
- Ss – łączy pomiot z czasownikiem
- PP – has/have z past participe

Lexical Functional Grammar

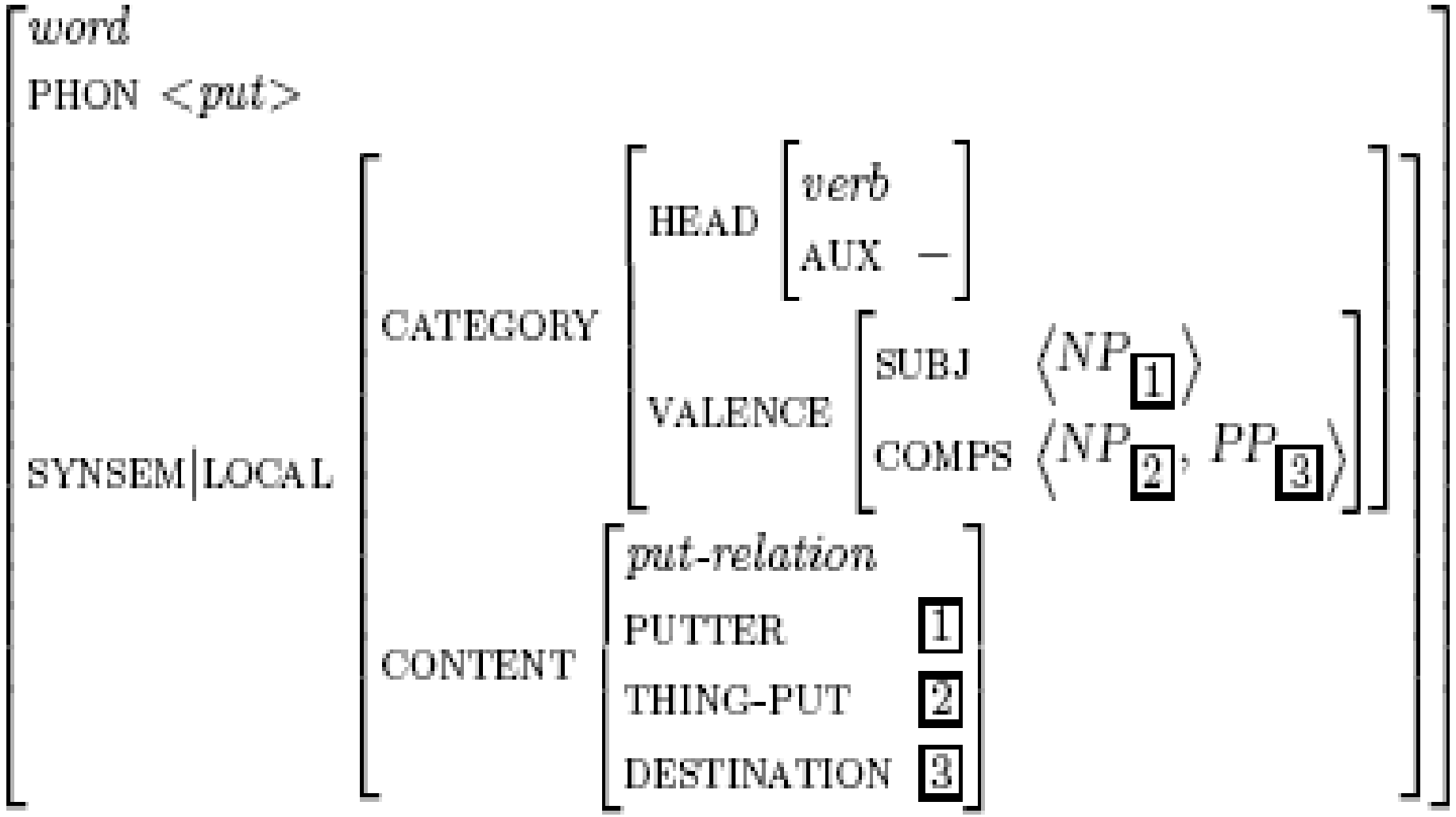
Operuje na dwóch rodzajach rozbiorów zdań:

- C-structure- ma postać gramatyki bezkontekstowej, obejmuje logiczną budowę zdania
- F-structue- (functional structure) jest zbiorem par atrybutów i ich wartości. Atrybutami mogą być np. Czas (tense), rodzaj (gender)
- Ważna jest równoległość dwóch powyższych struktur.
- Wzięcie pod uwagę F-structure daje tej gramatyce dużą przewagę nad gramatykami Chomskego, w niektórych przypadkach (np. passive)

Head-Driven Phrase Structure Grammar

- Łączy cechy wszystkich poprzednich gramatyk
- Bardzo rozbudowana
- Wymaga bardzo dokładnego i rozbudowanego opisu gramatyki:
  - zawiera leksykon podstawowych form wielu słów
  - własności odmian tych słów
  - reguły budowy zdań
  - zależności między poszczególnymi częściami zdania i frazami
  - fonologie słowa

Reprezentacja słowa “put”



Statistical Language Models

- Model języka który przyporządkowywuje prawdopodobieństwa do sekwencji słów
- Najprostsza wersja nie bierze pod uwagę kolejności słów- każda permutacja jest równie prawdopodobna
- Gramatyki te zazwyczaj nie są ręcznie pisane ale uczone na dużych zbiorach danych
- Bardzo przydatne przy obsłudze tekstu mówionego.
- Może być inna gramatyka do wczytania tekstu, inna do analizy, metoda do analizy może zawierać wolne miejsca na słowa -działać na niepełnych danych

SGS

W 1975 powstał formalizm Systemów Grup Składniowych (SGS). Reprezentacja składniowa zdania polega przedstawieniu zdania w postaci etykietowanego skierowanego grafu. W grafie tym węzły są grupami składniowymi, a krawędzie reprezentują relacje zależności.

W 1984 rozpoczął się rozwój formalizmu w celu modelowania języka naturalnego.
Zaletą SGS jest możliwość zastosowania do języków naturalnych, przede wszystkim o szyku swobodnym.

SGGP

Zadaniem SGGP (Syntactical Groups Grammar for Polish) jest określenie warunków grupowania GS i warunków powstania relacji składniowych.

Istnieją różne typy GS: ogólna, atrybutywna, imię, grupa liczebnika, spójnika, wyounktowania, rzeczownika, przyimka, przysłówka, zdania, trywialna złożona, czasownika, interpunkcyjna.

SGGP

SGGP operuje na skończonych zbiorach:
G – zbiór grup składniowych
I – zbiór indeksów
    indeks słowa - numer porządkowy słowa w tekście
    indeks GS - nazwa typu + numer porządkowy GS danego typu
K – zbiór typów GS
A – zbiór atrybutów GS (cechy morfo-syntaktyczne, semantyczne, itd.)
R – zbiór relacji składniowych

GS-gramatyka wykonuje operacje na iloczynie kartezjańskim GxIxKxAxR

SGGP

Aplikacje oparte o SGGP:

- **PolSumm** – program streszczania tekstów
- **Thetos** – automatyczny translator polskiego pisanego na polski język migowy
- **LAS** – Linguistic Analysis Server (serwer lingwistyczny)

LAS

Analiza podzielona jest na trzy etapy:

Analiza morfologiczna dzieli tekst źródłowy na słowa i ustala ich rodzaj i cechy morfologiczne.

Analiza syntaktyczna buduje drzewo na podstawie gramatyki SGGP (Syntactic Gropu Grammar for Polish) i znajduje syntaktyczne powiązania pomiędzy grupami. SGGP dzieli złożone zdania na zdania proste.

Analiza semantyczna określa semantyczną rolę każdej syntaktycznej grupy. Wykorzystywana jest informacja o semantycznym środowisku predykatu (czasownika).

Analizator używa słownika oraz reguł w poszukiwaniu składniowych grup zajmujących określone miejsca w odniesieniu do czasownika.

LAS (przykład)

Po analizie syntaktycznej zdania "Nieroztropność młodzi ludzie przypłacają zdrowiem" otrzymujemy cztery grupy syntaktyczne:

NG1 – nieroztropność
NG2 – młodzi ludzie
NG3 – zdrowiem
VG1 – przypłacają

Wszystkie te grupy składają się na zdanie S1 = {NG1,NG2,NG3,VG1} z podstawową grupą czasownikową VG1.

LAS (przykład)

Dla VG1 w słowniku został znaleziony schemat: Ngn1+VG1+Ngac-c2+NGi3. Dopasowujemy grupy do schematu.

Wynik jest następujący:
- podmiot: NG1 albo NG2
- orzeczenie: VG1
- rzeczownik w bierniku: NG1
- rzeczownik w narzędniku: NG3

Są dwie grupy mogące być podmiotem, ale tylko NG1 może być rzeczownikiem występującym w bierniku, tak więc podmiotem musi być NG2.

W tym przypadku pełna analiza semantyczna grup nie jest konieczna.

Analiza składniowa

Analiza składniowa koncentruje się na własnościach fleksyjnych wyrazów. Własności te mają zasadnicze znaczenie w procesie rozpoznawania związków składniowych pomiędzy wyrazami.
Na przykład zależność między rzeczownikiem i określającym go przymiotnikiem ma wykładnik formalny w postaci końcówek fleksyjnych charakterystycznych dla wspólnego obu wyrazom przypadka i dla wspólnej liczby.
W języku polskim pozycja wyrazu w zdaniu i grupie ma niewielki wpływ na składniową strukturę zdania.
Zupełnie inaczej jest w języku angielskim, gdzie głównym nośnikiem powiązań składniowych jest szyk wyrazów.

Analiza składniowa

Klasycznym przykładem jest zdanie: “John hit Paul”. W angielskim znaczenie zdania określa kolejność wyrazów, natomiast po polsku zdania: “John uderzył Paula” i “Johna uderzył Paul” mają różne znaczenia, które są zależne od odmiany wyrazów, chociaż ich kolejność pozostaje taka sama.

POLENG

W oparciu o analizę składniową powstały pierwsze wersje systeu automatycznego tłumaczenia polsko-angielskiego POLENG.
Gramatyka oparta jest głównie na opisie zawartym w książce S. Szpakowicza "Formalny opis składniowy zdań polskich".
Gramatyka ma swoje ograniczenia: na przykład dopuszczalne tylko zdania twierdzące (pojedyncze i złożone), niedopuszczalne są zdania z domyślnym orzeczeniem.
Nie są akceptowane konstrukcje eliptyczne - z domyślnym członem zdania innym niż podmiot.
Nie są dopuszczalne zdania niepoprawde pod względem interpunkcyjnym.

GFJP

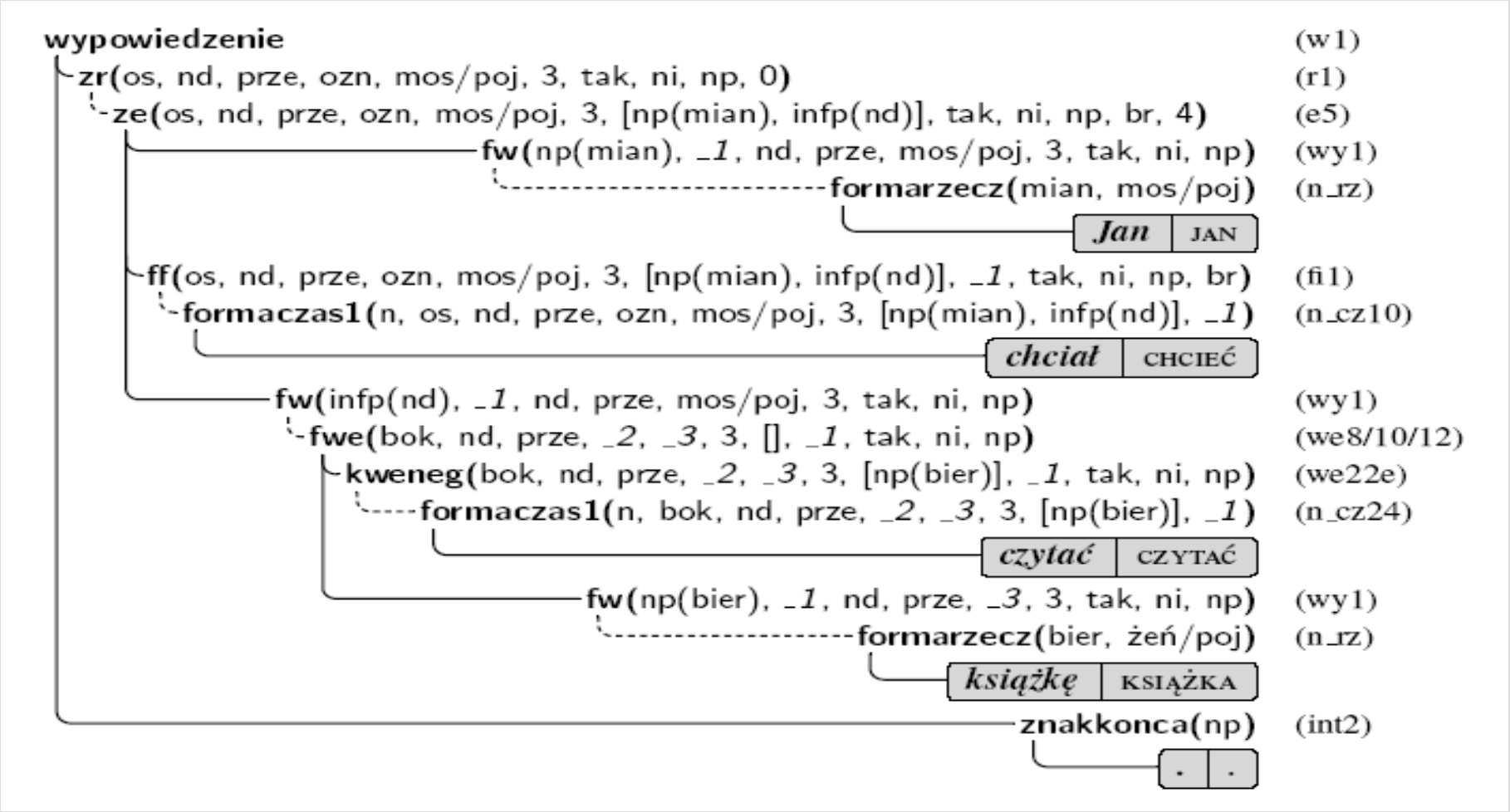
Największa i najbardziej szczegółowa gramatyka formalna polszczyzny: **gramatyka formalna języka polskiego** (GFJP) stworzona przez Marka Świdzińskiego (rozwinięta koncepcja Szpakowicza).
Gramatykę uważać można za gramatykę kontekstową skracającą.
Słownik terminalny tej gramatyki pomyślany został docelowo jako zbiór słów i znaków interpunkcyjnych.
Wszystkie jednostki składniowe nieterminalne reprezentowane są przez pary postaci <nazwa jednostki, komplet wartości parametrów>; nieliczne mają drugi element pary pusty. Parametry formalizują różnorodne cechy składniowe jednostek, takie jak charakterystyka fleksyjna, negacja, zależność, typ frazy zdaniowej, oznaczenie spójnika i in.

CFJP

Symbolem początkowym jest **WYPOWIEDZENIE.**
Hierarchia wygląda w uproszczeniu następująco. Wypowiedzenie realizowane jest jako **zdanie**. Zdanie zbudowane jest ze zdań niższego rzędu, zdanie zaś najniższego poziomu hierarchii jednostek zdaniowych składa się z **fraz**. **Frazy** redukowalne są również krok po kroku i nie wszystkie – do **form wyrazowych**. Ideę hierarchii można przedstawić, jak niżej:

wypowiedzenie
zdanie
...
fraza
...
forma wyrazowa

GFJP


wypowiedzenie (w1)
zr(os, nd, prze, ozn, mos/poj, 3, tak, ni, np, 0) (r1)
ze(os, nd, prze, ozn, mos/poj, 3, [np(mian), infp(nd)], tak, ni, np, br, 4) (e5)
fw(np(mian), _1, nd, prze, mos/poj, 3, tak, ni, np) (wy1)
formarzecz(mian, mos/poj) (n_rz)
Jan JAN
ff(os, nd, prze, ozn, mos/poj, 3, [np(mian), infp(nd)], _1, tak, ni, np, br) (fi1)
formaczas1(n, os, nd, prze, ozn, mos/poj, 3, [np(mian), infp(nd)], _1) (n_cz10)
chciał CHCIEĆ
fw(infp(nd), _1, nd, prze, mos/poj, 3, tak, ni, np) (wy1)
fwe(bok, nd, prze, _2, _3, 3, [], _1, tak, ni, np) (we8/10/12)
kweneg(bok, nd, prze, _2, _3, 3, [np(bier)], _1, tak, ni, np) (we22e)
formaczas1(n, bok, nd, prze, _2, _3, 3, [np(bier)], _1) (n_cz24)
czytać CZYTAĆ
fw(np(bier), _1, nd, prze, _3, 3, tak, ni, np) (wy1)
formarzecz(bier, żeń/poj) (n_rz)
książkę KSIĄŻKA
znakkonca(np) (int2)
. .

Świgra

Celem automatycznej analizy składniowej jest sprawdzenie, czy wypowiedzenie (dane jako napis) jest akceptowane przez daną gramatykę formalną i określenie jego struktury.

Program Świgra jest automatycznym analizatorem składniowym, co oznacza, że wypowiedzeniu danemu w formie napisu przyporządkowuje on rozbiory gramatyczne (w formie drzew analizy).

Działa on według reguł gramatyki Świdzińskiego.

Analizator składniowy pracuje na wypowiedzeniu podzielonym na segmenty, które zostały zanalizowane morfologicznie (fleksyjnie).

Gramatyka Świdzińskiego okazała się poddawać realizacji komputerowej, chociaż nie było zamiarem autora.

Jak działają boty?

Oto co odpowiedział jeden z nich na to pytanie:

> “Przetwarzając pytanie użytkownika biorę pod uwagę całą frazę oraz rozkładam ją na czynniki analizując poszczególne wyrazy, ich kolejność, istotność, częstość, formę gramatyczną, znaki przestankowe, kontekst poprzednich pytań, wcześniejszych rozmów oraz informacje zebrane na temat użytkownika (np. imię, ilość i długość rozmów, itp.), a następnie przeszukuje moje bazy wiedzy w poszukiwaniu najlepszej odpowiedzi na Twoje pytanie. Dzięki temu możesz zadawać mi pytania w języku naturalnym.”

Problemy przy analizie tekstu

- brak wiedzy (znaczeń wyrazów)
- części zdań domyślne
- przenośnie
- możliwość powiedzenia czegoś na parę różnych sposobów
- problemy z odmianą
- zasób słownictwa
- wiele znaczeń poszczególnych słów
- zrozumienie zdania wymaga znajomości kontekstu
- problemy z dialektami, błędami pisowni

Zastosowania

- Translatory z różnych języków
- Multimedialne książki, które “same się czytają” http://www.afb.org/Section.asp?SectionID=38&DocumentID=2429
- Boty - rozrywkowe i przemysłowe
- Automatyczne systemy obsługi - “naciśnij 1 aby odsłuchać wiadomość...”
- Sterowanie urządzeniami za pomocą rozkazów ustnych
- Uczenie się maszyn
- Wydobywanie informacji z tekstu

Eliza - bot-terapeutka

- Nieudana rozmowa
  - You: How are you today?
  - Eliza: Does that question interest you?
  - You: yes, it is really interesting
  - Eliza: You are sure?
  - You: yes
  - Eliza: You are sure?
  - You: no!
  - Eliza: You are being a bit negative.
  - You: I am not
  - Eliza: How long have you been not?
- Udana rozmowa
  - User: Men are all alike.
  - Eliza: In what way?
  - U: There're always bugging us about something or other.
  - E: Can you think of a specific example?
  - U: Well, my boyfriend made me come here.
  - E: Your boyfriend made you come here?
  - U: He says I'm depressed much of the time.
  - E: I am sorry to hear you are depressed

Snickers 4.0

- U:Jak sie dzisiaj masz?
- S: A dzięki, całkiem dobrze.
- U:Co lubisz robić w wolnym czasie?
- S: Lubię rozmawiać z ludźmi.
- U:A nie nudzą Cię rozmowy z ludźmi?
- S: Moja praca polega na rozmowach.
- U:Dlaczego nazywasz sie Snicker?
- S:Snickers to baton, ja nazywam się Snikers
- U:Jesteś batonem?
- S:Nie jestem żadnym batonikiem!
- U:Na pewno?
- S:Na pewno.

Fido

- F: Skoro już się znamy, to o czym sobie porozmawiamy?
- U: Jak się dzisiaj masz?
- F: Jak zwykle świetnie. Mam nadzieję, że u Ciebie też.
- U: Tak, wszystko w porządku
- F: To znakomicie
- U: Co lubisz robić w wolnym czasie?
- F: W pracy takiej jak moja nie ma zbyt wiele czasu dla siebie. Każdą chwilę spędzam rozmawiając. Ale wcale nie narzekam.
- U: Nie nudzą Cię rozmowy z ludźmi?
- F: Ludziom wiele zawdzięczam - przede wszystkim są moimi twórcami. Ludzie, a konkretnie programiści, stworzyli mnie na swoje podobieństwo... i pewnie stąd moje zainteresowanie piwem.

Prolog

- Język deklaratywny często używany do NLP.
- Jest to język predykatów pierwszego rzędu
- Nie mówimy **jak** coś ma być zrobione tylko **co** ma być zrobione i fakty jakimi może się posługiwać
- Program wyciąga nowe fakty z istniejących i dochodzi (lub nie) do poprawnego rozwiązania, poprawnych rozwiązań
- Łatwo w nim zapisywać gramatyki czy skojarzenia.
- Główne zastosowania poza NLP:
  - logika
  - sztuczna inteligencja

Linki

- Boty

Fido http://www.fido.pl/
Snikers http://www.snikers.pl/
Denise http://www.denise.cog-lab.com/
Eliza http://www-ai.ijs.si/eliza/eliza.html

- Translatory

POLENG http://www.poleng.pl/
Free Polish-English Translator http://www.poltran.com/

Bibliografia

- Natural Language Computing http://www.nyu.edu/pages/linguistics/anlcbind.htm
- AI in Information Retrieval and Language Problem http://www.phys.uni.torun.pl/~duch/IR.html#Agent
- Department of Language and Linguistics at the University of Essex, UK http://www.essex.ac.uk/linguistics/
- Wikipedia http://pl.wikipedia.org/
- Linguistics: An Introduction to Linguistics http://www.geocities.com/CollegePark/3920/
- Gramatyka formalna języka polskiego www.mswidz.republika.pl/pliki/materialy/GFJP_R3.doc
- Sztuczna inteligencja - analiza języka naturalnego www.phys.uni.torun.pl/~duch/Wyklady/AI/AI6-1.ppt
- Komputerowa weryfikacja gramatyki Świdzińskiego www.ipipan.waw.pl/~wolinski/publ/mw-autoref.pdf http://www.ipipan.waw.pl/~wolinski/publ/mw-phd.pdf
- Formalny opis składniowy zdań polskich http://sprocket.ict.pwr.wroc.pl/~banasiak/psdjn/foszp2.pdf
- Link Grammar http://www.link.cs.cmu.edu/link/
- Head-Driven Phrase Structure Grammar (HPSG) http://hpsg.stanford.edu/
- Lexical Functional Grammar. The Stanford Web Site. http://www-lfg.stanford.edu/lfg/