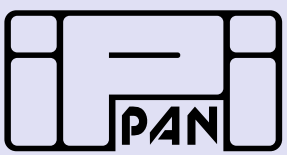
Parsowanie składniowe i jego zastosowania

Alina Wróblewska



INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES
ul. Jana Kazimierza 5, 01-248 Warszawa

Warszawa, 15 kwietnia 2015

Prowadząca: Alina Wróblewska

Stopnie naukowe

- 2014: Stopień doktora nauk technicznych w zakresie informatyki nadany przez Instytut Podstaw Informatyki PAN
- 2009: Stopień Magistra Artium w zakresie lingwistyki informatycznej i języka niemieckiego jako filologii języka obcego na Uniwersytecie w Heidelbergu

Doświadczenie

- 2014 – dzisiaj: adiunkt w IPI PAN
- 2009 – 2014: asystent w IPI PAN
- 2003 – 2008: asystent w dziale digtalizacji rękopisów i starodruków Biblioteki Uniwersyteckiej w Heidelbergu

Outline

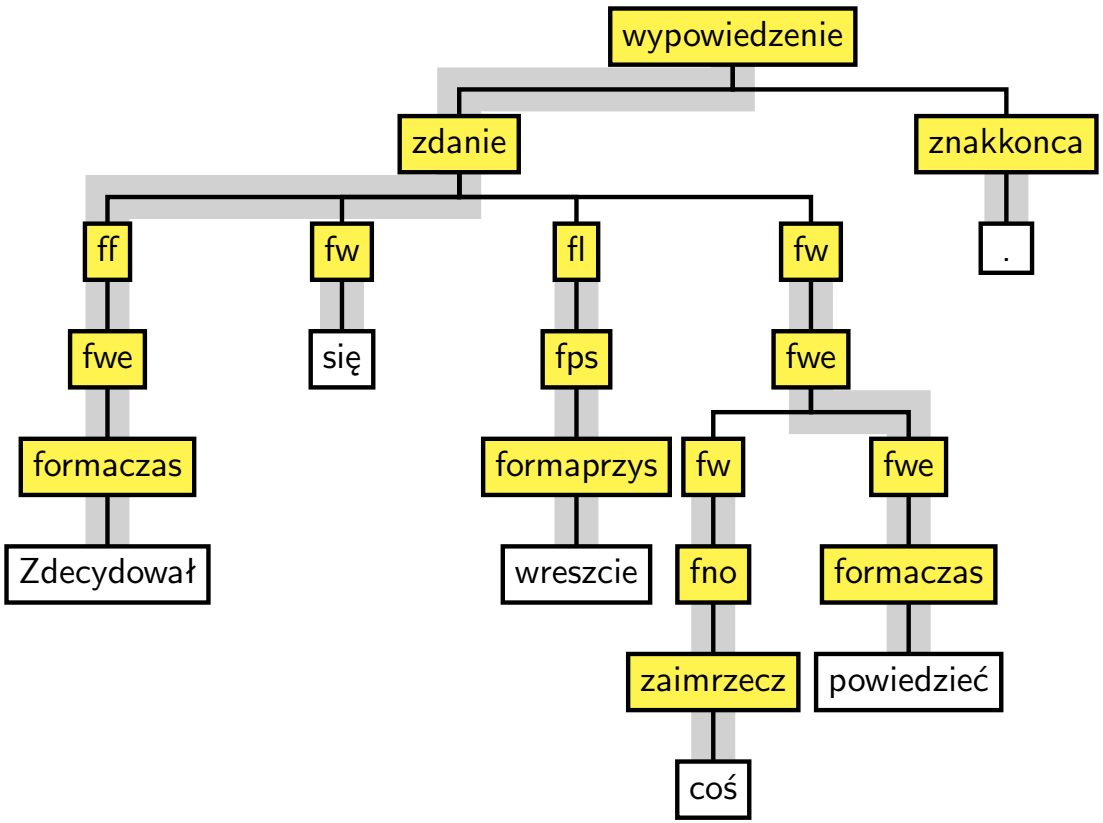
Parsowanie składniowe

Parsowanie składniowe to automatyczna analiza zdań i przypisanie im odpowiednich struktur składniowych.

Rodzaje analizy składniowej

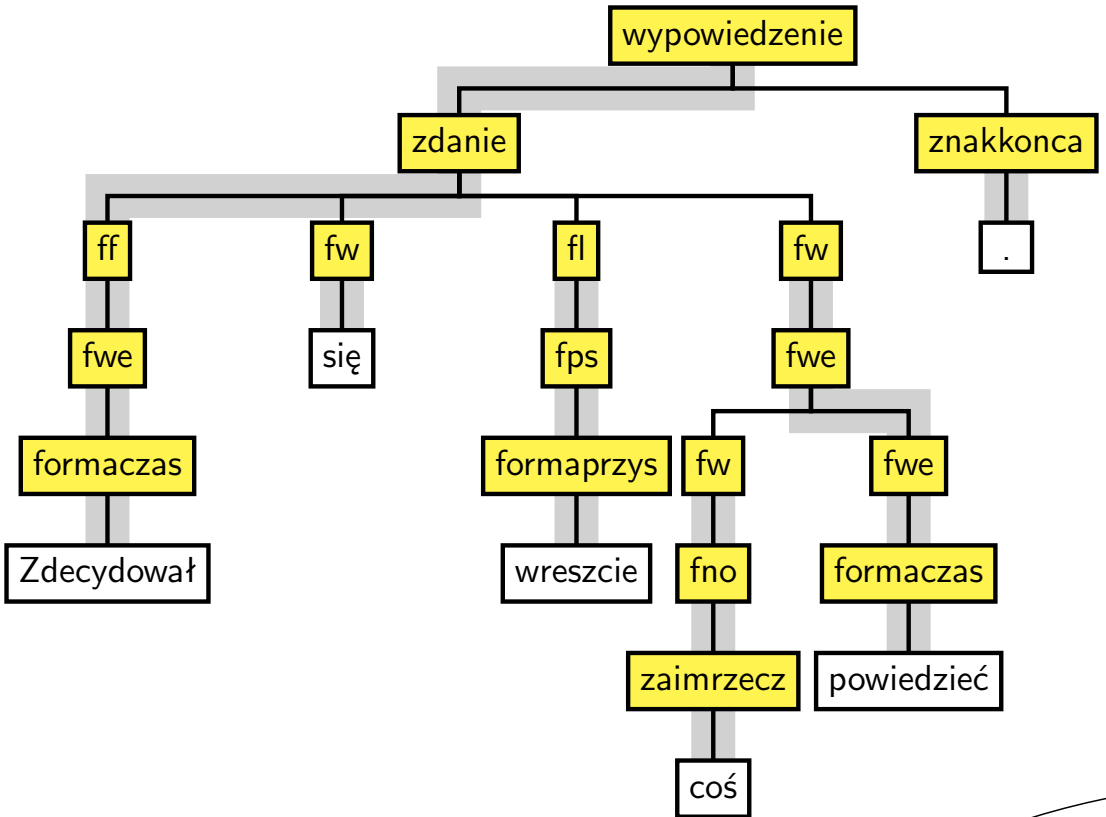
- **Parsowanie składnikowe** (ang. constituent parsing) polega na identyfikacji fraz (składników) w zdaniu oraz ich rekurencyjnej struktury.
- **Parsowanie zależnościowe** (ang. dependency parsing) polega na wyznaczeniu relacji pomiędzy wyrazami w zdaniu.

Struktura składnikowa

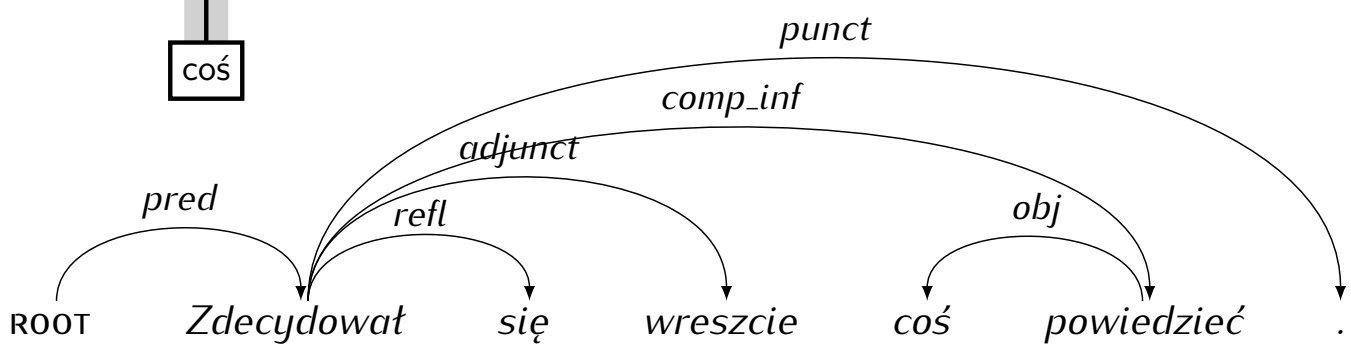


Struktura składnikowa i zależnościowa

Struktura składnikowa



Struktura zależnościowa



Zastosowanie parserów

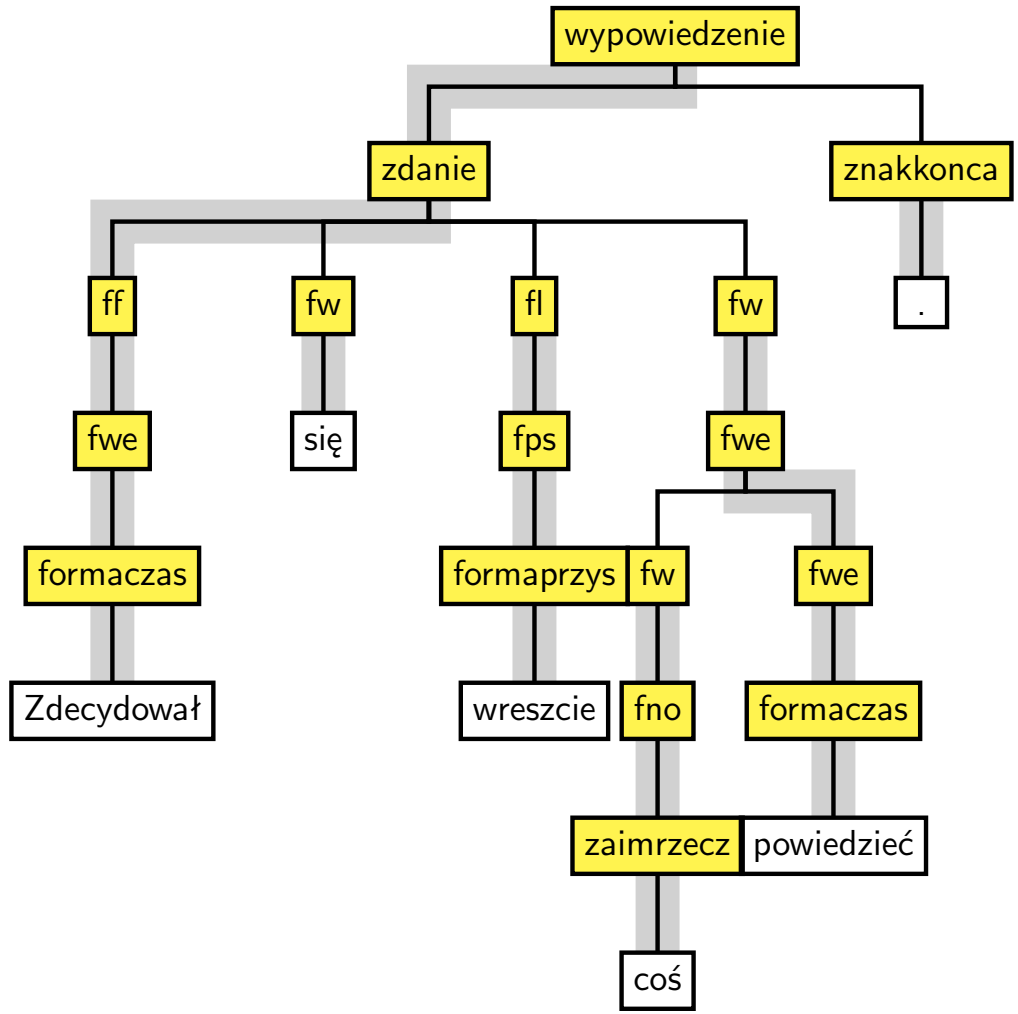
Zastosowanie parserów składniowych

- Parsowanie składniowe jest jednym z podstawowych elementów w zaawansowanych systemach przetwarzania języka naturalnego.
- Analiza składniowa jest wykorzystywana m.in. w:
  - systemach dialogowych,
  - systemach ekstrakcji wiedzy,
  - maszynowym tłumaczeniu,
  - analizie wydźwięku,
  - ...

Outline

Struktura składnikowa (1)

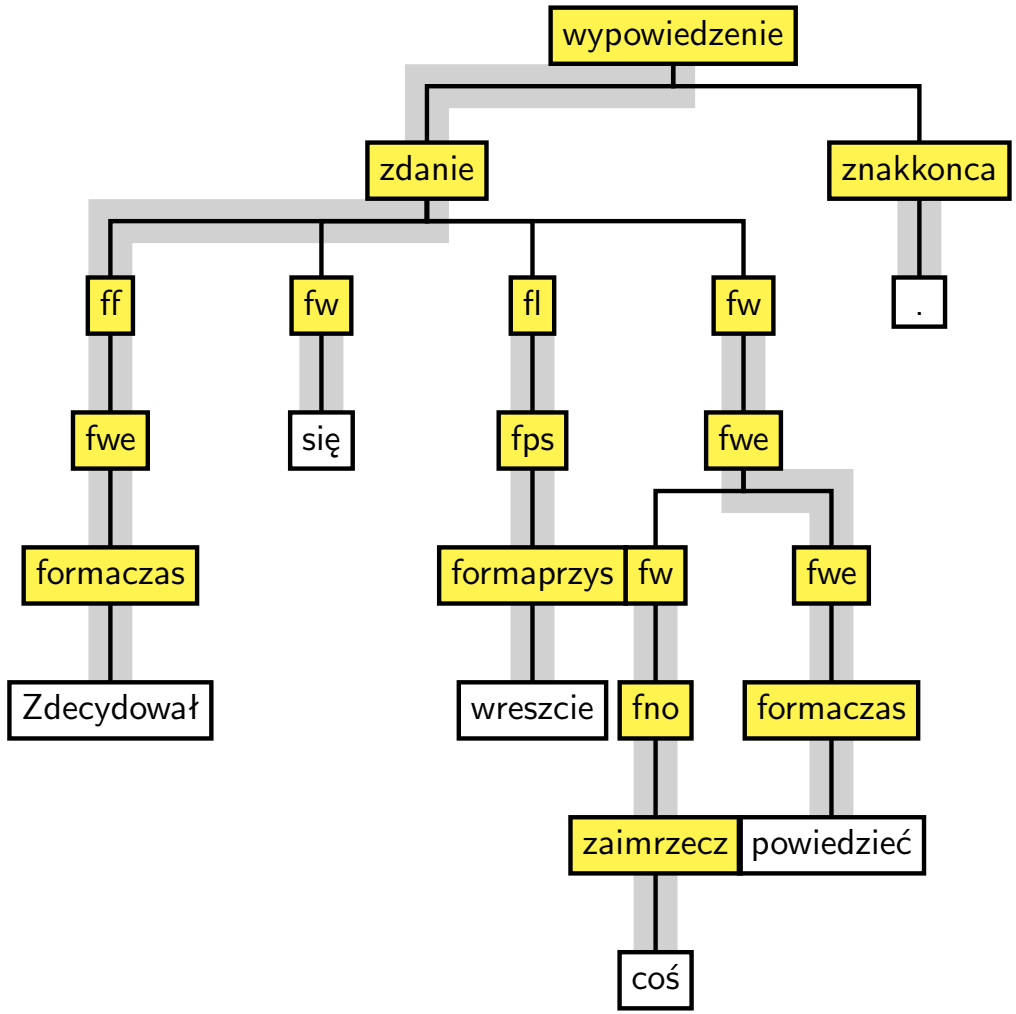
- Struktura składnikowa to drzewo, tutaj zbudowane zgodnie z gramatyką formalną języka polskiego (Świdziński, 1992).
- Liście w drzewie:
  - odpowiadają wyrazom (tokenom) w zdaniu,
  - kodują informację o formie podstawowej, części mowy i cechach morfologicznych danego wyrazu.



Struktura składnikowa (2)

Wewnętrzne węzły drzewa:

- zawierają informację o głowie frazy, funkcji podmiotu i kategoriach gramatycznych,
- mają etykietę odpowiadającą:
  - formie wyrazowej np. forma rzeczownikowa (*formarzecz*),
  - typowi składnika np. fraza werbalna (*fwe*), fraza nominalna (*fno*),
  - typowi subkategoryzacji – fraza wymagana (*fw*), fraza luźna (*fl*).



Parser Świgra (1)

Parser Świgra (Woliński, 2004)

Podstawę parsera stanowi implementacja zbioru reguł *Gramatyki formalnej języka polskiego* (GFJP, Świdziński, 1992) oraz słownik walencyjny (Świdziński, 1998).

Zmiany w gramatyce parsera w stosunku do GFJP

- mniej złożone i bardziej intuicyjne drzewa rozbioru,
- analiza zjawisk lingwistycznych nieuwzględnionych w GFJP (np. współrzędnie złożone frazy nominalne i przymiotnikowe, wymagania składniowe form rzeczownikowych i przymiotnikowych).

Słownik walencyjny parsera

- istotny element parsera,
- uzupełniony o dodatkowe czasowniki.

Parser Świgra (2)

Dostępność parsera Świgra

- Istnieje możliwość pobrania instalacji parsera Świgra ze strony `http://zil.ipipan.waw.pl/Świgra`.
- Informacje licencyjne znajdują się w pliku CZYTAJTO.

Zalety i niedoskonałości parsowania składnikowego

Zalety

- możliwość wyboru przez użytkownika właściwego rozbioru z proponowanych przez parser,
- wiarygodność zwracanych struktur, które są oparte na ręcznie zaimplementowanych regułach.

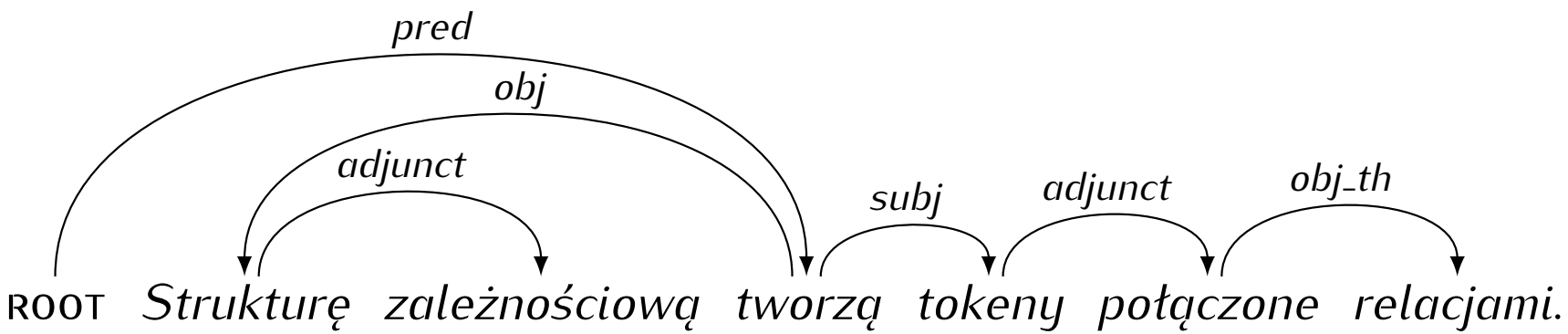
Niedoskonałości

- brak ujednoznaczniania – parsowanie zwraca wszystkie możliwe rozbiory dla danego zdania,
- problemy z reprezentowaniem niektórych zjawisk w języku o swobodnym szyku np. nieciągłości,
- w przypadku niezgodności zdania z regułami lub braku odpowiednich reguł parser nie zwraca żadnej analizy,
- definiowanie reguł gramatyki jest procesem bardzo czasochłonnym i kosztownym.

Outline

Struktura zależnościowa

- Struktura zależnościowa to drzewo (rozpinające):
  - każdy wierzchołek ma jedną krawędź wejściową,
  - korzeń ROOT nie ma krawędzi wejściowych i ma jedną krawędź wyjściową.
- Wierzchołki w drzewie odpowiadają tokenom w zdaniu.
- Krawędzie skierowane reprezentują relacje:
  - token, z którego wychodzi krawędź, jest nadrzędnikiem tokena, do którego wchodzi dana krawędź,
  - etykiety krawędzi odpowiadają funkcji gramatycznej podrzędnika.



Koncepcja struktury zależnościowej

Podstawy teoretyczne

Koncepcja struktury zależnościowej wywodzi się z założeń gramatyki zależnościowej:

- *Éléments de syntaxe structurale* (Tesnière, 1959) → zasada centralności czasownika, *connexion* (relacja zależnościowa) vs. *valence* (relacja walencyjna),
- teoria Mel'čuka (ang. Meaning-Text Theory; Mel'čuk, 1988) → model MTT (struktura lingwistyczna obejmuje wiele reprezentacji zależnościowych),
- teoria Sgalla (ang. Functional Generative Description; Sgall et al., 1986),
- teoria Hellwiga (ang. Dependency Unification Grammar; Hellwig, 1986, 2003),
- teoria Hudsona (ang. Word Grammar; Hudson, 1990),
- ...

Parsowanie zależnościowe

- Parsowanie zależnościowe to analiza składniowo-semantyczna, która wydobywa strukturę predykatywno-argumentową zdania.
- Parsowanie zależnościowe to proces automatycznego przypisania zdaniu wejściowemu odpowiedniej struktury zależnościowej.

Parsery zależnościowe

- Mogą opierać się na ręcznie stworzonej gramatyce.
- Mogą wykorzystywać metody statystyczne do wytrenowania modeli parsowania.

Gramatyki parserów zależnościowych

Bezkontekstowe gramatyki zależnościowe

- Link Grammar (Sleator i Temperley, 1991),
- Gramatyki bileksykalne (Eisner, 1996; Eisner, 2000).

Gramatyki zależnościowe z ograniczeniami

- Weighted Constraint Dependency Grammar (Menzel and Schröder, 1998; Foth et al., 2004),
- Probabilistic Constraint Dependency Grammar (Harper and Helzerman, 1995; Wang and Harper, 2004),
- Topological/Extensible Dependency Grammar (Duchier and Debusmann, 2001; Debusmann et al., 2004).

Statystyczne parsery zależnościowe

Trenowanie modeli parsowania

Na podstawie danych treningowych parsery uczą się analizować zdania i generować dla nich odpowiednie struktury zależnościowe.

Parsery trenowane za pomocą metod z nadzorem

- trenowane na banku ręcznie zaanotowanych drzew,
- osiągają najlepsze wyniki jak dotychczas,
- wymagają dużej liczby poprawnie zaanotowanych struktur.

Parsery trenowane za pomocą metod bez nadzoru

- mało efektywne,
- bardzo duża złożoność obliczeniowa,
- generują głównie drzewa bez etykiet na krawędziach przez co są niewystarczające dla wielu zadań NLP.

Parsery trenowane z nadzorem

Parsery oparte na przejściach (ang. transition-based parsers)

Parsery konstruują optymalną sekwencję przejść opierając się na wytrenowanym modelu.

Parsery grafowe

Dla danego zdania parser grafowy definiuje zbiór możliwych drzew zależnościowych (kandydaci), ocenia te drzewa na podstawie wytrenowanego modelu i wybiera najwyżej punktowane drzewo jako ostateczną analizę.

Literatura

S. Kübler, R.T. McDonald i J. Nivre (2009). *Dependency Parsing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Ogólnodostępne systemy parsujące

Parsery oparte na przejściach

- **MaltParser** (Nivre i in., 2007), `http://www.maltparser.org/`.
- Inne systemy: Attardi (2006), Attardi i Ciaramita (2007), Duan i in. (2007), Johansson i Nugues (2007), Titov i Henderson (2007).

Parsery grafowe

- Parser Eisnera (1996).
- **Parser MST** (McDonald i in., 2006), `http://sourceforge.net/projects/mstparser/`.
- **Parser MATE** (Bohnet, 2010), `https://code.google.com/p/mate-tools/`.
- Inne systemy: Carreras (2007), Koo i in. (2007), Nakagawa (2007), Smith i Smith (2007).

Jakość parsowania

Zawody (ang. shared task)

- CoNLL 2006 shared task (Buchholz i Marsi, 2006): 19 parserów dla 13 języków.
- CoNLL 2007 shared task (Nivre i in., 2007): 23 parsery dla 10 języków.

MaltParser i parser MST na CoNLL 2006

	MST	Malt
arabski	66.91	66.71
bułgarski	87.57	87.41
chiński	85.90	86.92
czeski	80.18	78.42
niemiecki	87.34	85.82
słoweński	73.44	70.30
szwedzki	82.55	84.58
turecki	63.19	65.68

Składnica zależnościowa

Składnica zależnościowa (Wróblewska, 2012, 2014)

- Struktury zależnościowe zostały przekonwertowane z drzew składnikowych (Woliński i in., 2011) w sposób automatyczny.
- Poszczególnym krawędziom w drzewach zostały przypisane etykiety na podstawie zbioru zdefiniowanych reguł.
- Poszczególne typy relacji zależnościowych zostały opisane na `http://zil.ipipan.waw.pl/FunkcjeZaleznosciowe`.
- Bank obejmuje ponad 8 tysięcy drzew.
- Składnica zależnościowa jest dostępna na stronie `http://zil.ipipan.waw.pl/Składnica`.

Parsowanie zależnościowe języka polskiego (1)

Model zależnościowy dla języka polskiego

- Kompatybilny z systemem parsującym MaltParser.
- Wytrenownany na Składnicy zależnościowej.
- Dostępny na stronie `http://zil.ipipan.waw.pl/PolishDependencyParser`.

Parsowanie zależnościowe języka polskiego (2)

Jakość parsowania języka polskiego

- System: parser MATE (Bohnet, 2010).
- Dane uczące: Składnica zależnościowa.
- Dane testowe:
  1. krótkie i proste zdania z ręcznie zaanotowanymi tokenami,
  2. rozbudowane zdania z półautomatycznie zaanotowanymi tokenami.
- Wyniki testów:
  - Test 1: 87,2 LAS* i 92,7 UAS**
  - Test 2: 70,3 LAS i 76 UAS

*LAS (labelled attachment score) – procent tokenów, którym został przypisany poprawny nadrzędnik i poprawna funkcja gramatyczna (etykieta relacji).
**UAS (unlabelled attachment score) – procent tokenów, którym został przypisany poprawny nadrzędnik.

Zalety i niedoskonałości parsowania zależnościowego

Zalety

- szybkość – zdania mogą być parsowane w czasie liniowym,
- ujednoznacznianie – parser zwraca tylko jedną strukturę dla zdania,
- łatwość w reprezentowaniu niektórych zjawisk lingwistycznych np. nieciągłości,
- istnieje wiele systemów parsujących, które można wykorzystać dla danego języka.

Niedoskonałości

- brak możliwości zastąpienia zwracanej struktury, która zawiera błędne relacje, przez inną,
- ograniczenia i uproszczenia w reprezentowaniu niektórych zjawisk np. podmiot logiczny, podnoszenie podmiotu,
- duże uzależnienie od jakości tagowania.

Outline

Gramatyka leksykalno-funkcyjna

Lexical Functional Grammar (LFG)

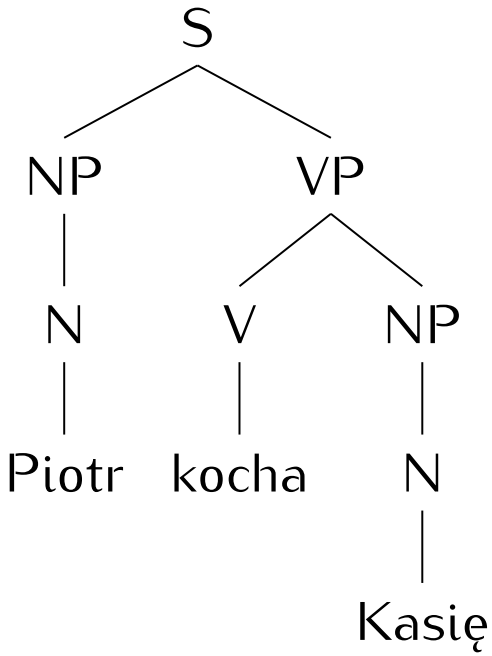
- Teoria, której rozwój zapoczątkowali Joan Bresnan i Ronald M. Kaplan w latach 70-tych XX w. (Kaplan i Bresnan, 1995).
- Formalizm LFG kładzie nacisk na składnię i jej związki z morfologią, semantyką i pragmatyką.
- LFG traktuje język jako wielowymiarową strukturę z wymiarami: składniowym, semantycznym i pragmatycznym.

Wymiar składniowy

Wymiar składniowy obejmuje trzy korespondujące ze sobą reprezentacje:

- struktura C (struktura składnikowa),
- struktura F (struktura funkcyjna),
- struktura A (struktura argumentowa).

Struktury C i F (1)



$$\begin{bmatrix} \text{PRED} & \text{`kochać}\langle \text{SUBJ,OBJ} \rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`Piotr'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{`Kasia'} \end{bmatrix} \end{bmatrix}$$

Struktury C i F (2)

Struktura C

Struktura C koduje strukturę składnikową i szyk zdania.

Struktura F

- Struktura F koduje:
  - funkcje gramatyczne
    - argumenty: SUBJ, OBJ, OBJ_θ, COMP, XCOMP i OBL_θ
    - nie-argumenty: modifikatory – ADJ, XADJ i funkcje dyskursu – TOPIC, FOCUS
  - cechy morfoskładniowe np. NUM, GEND, PERS, CASE, TENSE.
- Warunki poprawności np. warunek pełności (ang. completeness condition), warunek spójności (ang. cohrence condition) wykluczają niepoprawne struktury (Bresnan, 2001; Dalrymple, 2001).

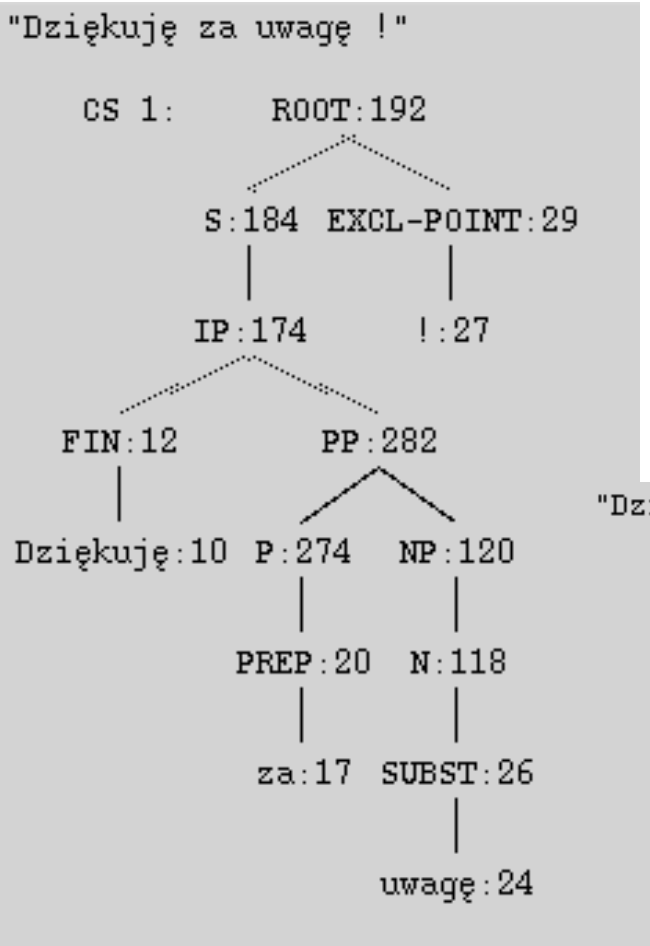
POLFIE – polska gramatyka LFG

POLFIE (Patejuk i Przepiórkowski, 2012)

- Reguły gramatyki POLFIE pokrywają główne lingwistyczne zjawiska języka polskiego zdefiniowane w gramatyce Świdzińskiego (1992).
- Aktualna wersja gramatyki POLFIE zawiera 65 reguł.
- Gramatyka POLFIE jest dostępna na stronie `http://zil.ipipan.waw.pl/LFG`.
- Gramatyka POLFIE jest kompatybilna z systemem parsującym XLE (Maxwell i Kaplan, 1993).
- Parser XLE z gramatyką POLFIE parsuje ok. 42% zdań z poprawnymi tagami morfoskładniowymi (test na 20 tysiącach zdań).

Dziękuję za uwagę!

Dziękuję za uwagę!



"Dziękuję za uwagę !"

```
      [PRED    'dziękować<[12-SUBJ:pro], [17:uwaga]>'                                  ]
      [SUBJ    [PRED 'pro'                                                             ]
      [        [CASE nom, NUM sg, PERS 1]                                              ]
      [      24[PRED  'uwaga'                                                          ]
      [      26[CHECK [CAT subst]                                                      ]
      [     118[                                                                       ]
      [OBL  120[NTYPE [NSEM [COMMON count]]                                           ]
      [      17[      [NSYN common]                                                   ]
      [      20[                                                                       ]
   27 [     274[CASE acc, GEND f, NUM sg, PERS 3, PFORM za, PTYPE nosem                ]
   29 [     282[                                                                       ]
   12 [                                                                               ]
  174 [CHECK    [CAT fin]                                                             ]
  184 [                                                                               ]
  192 [TNS-ASP  [ASPECT imperf, MOOD indicative, TENSE pres]                          ]
```

Literatura (1)

- Bohnet, B. (2010). Very High Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING 2010, pages 89–97.
- Bresnan, J. (2001). Lexical-Functional Syntax. Oxford: Blackwell.
- Dalrymple, M. (2001). Lexical Functional Grammar. Volume 34, Syntax and Semantics. Academic Press.
- J.M. Eisner (1996) Three new probabilistic models for dependency parsing: An exploration. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING), pp. 340–345.
- J.M. Eisner (2000) Bilexical grammars and their cubic-time parsing algorithms. In: H. Bunt and A. Nijholt (ed.), Advances in Probabilistic and Other Parsing Technologies, pp. 29–62. Kluwer.
- Hudson, R.A. (1990) English Word Grammar, Blackwell.
- S. Kübler, R.T. McDonald and J. Nivre (2009). Dependency Parsing. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
- R. McDonald, K. Lerman and F. Pereira (2006) Multilingual dependency analysis with a two-stage discriminative parser. In: Proceedings of the Tenth Conference on Computational Natural Language Learning. pp. 216–220.
- R. McDonald and G. Satta (2007) On the complexity of non-projective data-driven dependency parsing. In Proc. IWPT.
- Mel'čuk, I. (1988) Dependency Syntax: Theory and Practice, State University of New York Press.

Literatura (2)

- Nivre, J., J. Nilsson, J. Hall, A. Chanev, G. Eryigit, S. Kübler, S. Marinov & E. Marsi (2007) MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. Natural Language Engineering 13(1), 1–41.
- Przepiórkowski, A., Bańko, M., Górski, R. L., i Lewandowska-Tomaszczyk, B., [redaktorzy] (2012). Narodowy Korpus Języka Polskiego. Wydawnictwo Naukowe PWN, Warszawa.
- Sgall, P., Hajičova, E. i Panevova, J. (1986) The Meaning of the Sentence in Its Pragmatic Aspects, Reidel.
- D. Sleator and D. Temperley (1991) Parsing English with a link grammar. Technical Report CMU-CS-91-196, Carnegie Mellon University, Computer Science.
- Świdziński, M. (1992). Gramatyka formalna języka polskiego, Rozprawy Uniwersytetu Warszawskiego, tom 349. Wydawnictwa Uniwersytetu Warszawskiego, Warszawa.
- Woliński, M. (2004). Komputerowa weryfikacja gramatyki Świdzińskiego, rozprawa doktorska, Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warszawa.
- Woliński, M., Głowińska, K. i Świdziński, M. (2011) A preliminary version of Składnica—a treebank of Polish. In Zygmunt Vetulani, editor, Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, pages 299–303, Poznań, Poland.
- Wróblewska, A. (2014) *Polish Dependency Parser Trained on an Automatically Induced Dependency Bank*, rozprawa doktorska. Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warszawa.