Lingwistyka I – wykład 7

Adam Przepiórkowski

Kognitywistyka UW

4 kwietnia 2017

Gramatyki formalne

Na przykładzie:

(1) S $\longrightarrow$ NP VP

(2) NP $\longrightarrow$ *John*

(3) VP $\longrightarrow$ *sneezed*

Gramatyka formalna to czwórka uporządkowana $\langle N, \Sigma, P, S \rangle$, gdzie:

- N jest skończonym zbiorem *symboli nieterminalnych*, np. NP, VP i S;
- Σ jest skończonym zbiorem *symboli terminalnych*, np. *John* i *sneezed*; $N \cap \Sigma = \varnothing$;
- P jest skończonym zbiorem *produkcji*, tj. reguł postaci: $a_1 a_2 \ldots a_n \longrightarrow b_1 b_2 \ldots b_k$, gdzie $n > 0$, $k \geqslant 0$, oraz każde a_i i każde b_i jest terminalem lub nieterminalem (zob. przykład);
- $S \in N$ jest *symbolem początkowym*, np. S.

Hierarchia Chomskiego (1)

Cztery główne klasy gramatyk formalnych $\langle N, \Sigma, P, S \rangle$, różniące się dozwoloną formą produkcji P.

Według definicji na poprzednim slajdzie produkcje mają postać $\alpha \longrightarrow \beta$, gdzie α i β to dowolne ciągi terminali i nieterminali, przy czym $\alpha \neq \epsilon$, tj. $\alpha \in (N \cup \Sigma)^+$ i $\beta \in (N \cup \Sigma)^*$.

Gramatyki typu 0 to dowolne gramatyki formalne w tym sensie.

Przykłady dozwolonych reguł produkcji:

- $a\ \mathrm{N} \longrightarrow b$
- $a\ \mathrm{N}\ b \longrightarrow a\ \mathrm{M}\ b$
- $\mathrm{N} \longrightarrow a\ \mathrm{M}\ b$
- $\mathrm{N} \longrightarrow a\ \mathrm{M}$
- $\mathrm{N} \longrightarrow \epsilon$

Mocne ale nieobliczalne.

Hierarchia Chomskiego (2)

Gramatyki typu 1 (gramatyki **kontekstowe**; ang. *Context-Sensitive Grammars*) dopuszczają wyłącznie produkcje następującej postaci (gdzie α i β należą do $(N \cup \Sigma)^*$, zaś $\gamma \in (N \cup \Sigma)^+$; $A \in N$):

- $\alpha\ A\ \beta \longrightarrow \alpha\ \gamma\ \beta$
- $A \longrightarrow \epsilon$

Inna notacja używana np. w Chomsky i Halle (1968):

- $A \longrightarrow \gamma\ /\ \alpha\_\_\beta$

Np. *flapping rule* w amerykańskim angielskim (np. amerykańska wymowa *notable* ze spółgłoską uderzeniową [dx] w miejscu [t] po akcentowanej samogłosce):

- [t] $\longrightarrow$ [dx] / V́__V

Wystarczająco mocne lingwistycznie, z trudem obliczalne (w ogólnym wypadku). Niektóre formalizmy lingwistyczne są (łagodnie) kontekstowe.

Hierarchia Chomskiego (3)

Gramatyki typu 2 (gramatyki **bezkontekstowe**; ang. *Context-Free Grammars*) dopuszczają produkcje następującej postaci (gdzie $\gamma \in (N \cup \Sigma)^*$):

- $A \longrightarrow \gamma$

Przykłady:

- $S \longrightarrow NP\ VP$
- $NP \longrightarrow Det\ N$
- $Det \longrightarrow \textit{the}$
- $N \longrightarrow \textit{boy}$
- $VP \longrightarrow VDT\ NP\ NP$
- $VDT \longrightarrow \textit{gives}$
- $Trace \longrightarrow \epsilon$

Podstawa znacznej większości teorii lingwistycznych.

Hierarchia Chomskiego (4)

Gramatyki typu 3 dopuszczają produkcje dwóch postaci (gdzie A i B należą do N, zaś a należy do Σ):

- $A \longrightarrow a\ B$
- $A \longrightarrow \epsilon$

Gramatyki takie nazywane są **ściśle prawostronnie liniowymi** (ang. *Strictly Right-Linear Grammars*) lub po prostu **regularnymi** (ang. *Regular Grammars*).

Popularne w przetwarzaniu języka naturalnego, bo bardzo szybkie (i często wystarczająco mocne).

Czemu **hierarchia Chomskiego**?

- Chomsky 1959,
- każdy język zdefiniowany przez gramatykę typu n można także zdefiniować przez gramatykę typu $n-1$ (ale nie zawsze na odwrót).

Hierarchia Chomskiego (5)

Hierarchia klas gramatyk odpowiada hierarchii klas języków przez nie generowanych:

- **języki typu 0**: języki **rekurencyjnie przeliczalne**, dowolny język, który może zostać zdefiniowany algorytmicznie;
- **języki typu 1**: języki **kontekstowe**;
- **języki typu 2**: języki **bezkontekstowe**;
- **języki typu 3**: języki **regularne**.

Skoro każda gramatyka regularna jest jednocześnie gramatyką bezkontekstową, to oczywiście każdy język regularny jest jednocześnie językiem bezkontekstowym itd.:

- języki regularne $\subseteq$ języki bezkontekstowe;
- języki bezkontekstowe $\subseteq$ języki kontekstowe;
- języki kontekstowe $\subseteq$ języki rekurencyjnie przeliczalne.

Hierarchia Chomskiego (6)

Ale czy mogą zachodzić równości? Nie, nie ma równości – zawsze **zawieranie właściwe** ($\subsetneq$).

Język **bezkontekstowy**, który nie jest regularny, to np.: $a^n b^n$.

Język **kontekstowy**, który nie jest bezkontekstowy, to np.: $a^n b^n c^n$.

Trudniej opisać języki **rekurencyjnie przeliczalne**, które nie są kontekstowe, ale takie istnieją.

Do których klas należą **języki naturalne**? Trudne pytanie i nie do końca jasne, czy sensowne.

Obliczalność: dla każdego typu gramatyk posiadamy najbardziej efektywny program, który dostaje na wejściu gramatykę tego typu i ciąg, o którym program ma orzec, czy należy do języka generowanego przez tę gramatykę. Jaki jest czas tego orzekania?

Hierarchia Chomskiego (7)

To zależy od typu gramatyki i długości ciągu. Załóżmy, że mamy ciąg A o pewnej długości (np. 10 słów), B – dwa razy dłuższy niż A (czyli np. 20 słów), oraz C – pięć razy dłuższy (czyli np. 50 słów).

- **Gramatyki regularne**: orzekanie w czasie liniowym ($O(n)$), tj. dla B 2 razy dłużej niż dla A, dla C – 5 razy dłużej, np.:
  - A: 1 ms
  - B: 2 ms
  - C: 5 ms
- **Gramatyki bezkontekstowe**: orzekanie w czasie (pesymistycznie) wielomianowym (kubicznym; $O(n^3)$), tj. np. dla B – $2^3 = 8$ razy dłużej, a dla C – $5^3 = 125$ razy dłużej, np.:
  - A: 1 ms
  - B: 8 ms
  - C: 125 ms

Hierarchia Chomskiego (8)

- **Gramatyki kontekstowe**: orzekanie w czasie (pesymistycznie) wykładniczym (rzędu $O(2^n)$); jeżeli A jest długości 10, to odpowiedź dla B może zająć $2^{20}/2^{10} = 2^{10} = 1024$ razy dłużej, a dla C – $2^{50}/2^{10} = 2^{40} = 1\,099\,511\,627\,776$ razy dłużej, np.:
  - A: 1 ms
  - B: ponad sekundę (vs 8 ms dla bezkontekstowych vs 2 ms dla regularnych)
  - C: prawie 35 lat (vs 125 ms dla bezkontekstowych vs 5 ms dla regularnych)

- **Gramatyki rekurencyjnie przeliczalne**: problem nierozstrzygalny (nie ma gwarancji, że uzyskamy odpowiedź w skończonym czasie).

Hierarchia Chomskiego (9)

Ale możliwe jest definiowanie klas pośrednich; najważniejsze to różne klasy gramatyk umiarkowanie kontekstowych (ang. **Mildly Context-Sensitive Grammars**; Joshi 1985).

Dana **klasa gramatyk** jest **umiarkowanie kontekstowa** jeżeli:

- klasa języków przez nie definiowanych zawiera w sposób właściwy klasę języków bezkontekstowych,
- problem rozstrzygania przynależności ciągu do języka jest obliczalny w czasie (pesymistycznie) wielomianowym,
- każdy język danej klasy ma własność *przyrostu liniowego* (ang. *linear growth property*):
  - sortujemy ciągi danego języka wg długości,
  - nie musi być tak, że dla każdej długości jest jakiś ciąg (np. $a^n b^n c^n$ to ciągi o długości 0, 3, 6, 9, 12 itd.),
  - ale musi istnieć pewna liczba naturalna taka, że nie ma „przeskoków" długości większych niż ta liczba (dla przykładu w poprzednim punkcie taką liczbą jest np. 3).

(To jedna z możliwych definicji...)

Hierarchia Chomskiego (10)

Język, który nie ma takiej własności przyrostu liniowego: a^{2^n}, tj.:

a
aa
aaaa
aaaaaaaa
aaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
...

Inny taki język: a^p, gdzie p jest liczbą pierwszą.

Hierarchia Chomskiego (11)

Znane **formalizmy lingwistyczne**, które są **umiarkowanie kontekstowe**:

- Tree Adjoining Grammar (Joshi 1985, 1987)
- Head Grammar (Pollard 1984)
- Combinatory Categorial Grammar (Steedman 1987, 1996, 2000)

Co najmniej **całkowicie kontekstowe**:

- Lexical Functional Grammar (Kaplan i Bresnan 1982, Bresnan 2001, Dalrymple 2001)

Rekurencyjnie przeliczalne:

- Head-driven Phrase Structure Grammar (Pollard i Sag 1987, 1994)

Bresnan J., 2001, Lexical-Functional Syntax, Blackwell, Malden, MA.

Chomsky N., 1959, On certain formal properties of grammars, Information and Control 2, s. 137–167.

Chomsky N., Halle M., 1968, The Sound Pattern of English, Harper and Row, Nowy Jork.

Dalrymple M., 2001, Lexical Functional Grammar, Academic Press, San Diego, CA.

Joshi A. K., 1985, Tree adjoining grammars: How much context sensitivity is required to provide reasonable structural descriptions?, [w:] Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives, red. D. Dowty, L. Karttunen, A. M. Zwicky, Cambridge University Press, Cambridge.

Joshi A. K., 1987, An introduction to Tree Adjoining Grammars, [w:] Mathematics of Language, red. A. Manaster-Ramer, John Benjamins, Amsterdam.

Kaplan R. M., Bresnan J., 1982, Lexical-Functional Grammar: A formal system for grammatical representation, [w:] The Mental Representation of Grammatical Relations, red. J. Bresnan, MIT Press Series on Cognitive Theory and Mental Representation, The MIT Press, Cambridge, MA, s. 173–281.

Pollard C., 1984, Generalized Phrase Structure Grammars, Head Grammars, and Natural Languages, Rozprawa doktorska, Stanford University, Stanford, CA.

Pollard C., Sag I. A., 1987, Information-Based Syntax and Semantics, Volume 1: Fundamentals, CSLI Publications, Stanford, CA.

Pollard C., Sag I. A., 1994, Head-driven Phrase Structure Grammar, Chicago University Press / CSLI Publications, Chicago, IL.

Steedman M., 1987, Combinatory grammars and parasitic gaps, Natural Language and Linguistic Theory 5, s. 403–439.

Steedman M., 1996, Surface Structure and Interpretation, The MIT Press, Cambridge, MA.

Steedman M., 2000, The Syntactic Process, The MIT Press, Cambridge, MA.