Blanka Sedlačíková

Matematická lingvistika (3)

Učitel matematiky, Vol. 10 (2002), No. 3, 174–181

Persistent URL: http://dml.cz/dmlcz/150484

MATEMATICKÁ LINGVISTIKA (3)

BLANKA SEDLAČÍKOVÁ

Algebraická lingvistika

Celou dobu si tu povídáme o aplikaci matematických metod v lingvistice. Nematematikové (a tedy i řada lingvistů) se často dopouštějí jednoho omylu – matematiku chápou jako vědu pouze o kvantitativních vztazích. Od druhé poloviny 19. století se však v matematice objevuje celá řada takových teorií a metod, které mají výrazně nekvantitativní charakter. Patří sem například algebra, teorie grafů, matematická (formální) logika, teorie množin či kombinatorika. Díky vysokému stupni abstrakce lze uvedené teorie úspěšně použít k obecnému studiu systémů, a tedy i ke studiu přirozeného jazyka. Právě tento abstraktní přístup je podstatou algebraické lingvistiky, která je dalším odvětvím matematické lingvistiky (a to spolu s lingvistikou kvantitativní, kterou jsme se již zabývali, a lingvistikou strojovou).

Algebraická lingvistika se začala formovat ve druhé polovině 50. let 20. století zejména v souvislosti s potřebami strojového překladu, neboť bylo nutno odstranit vágnost některých lingvistických teorií. Jediným možným způsobem se ukázala být důsledná formalizace, která byla úspěšně používána v matematice. Za dnes běžně užívaný termín **algebraická lingvistika** vděčíme Y. Bar-Hillelovi. U některých matematiků a lingvistů bývalého Sovětského svazu se můžeme setkat také s označením **teorie jazykových modelů**.

Základ algebraické lingvistiky tvoří zejména tyto teorie: generativní a transformační mluvnice N. Chomského, kategoriální gramatika Y. Bar-Hillela, aplikačně generativní model jazyka S. K. Šaumjana, teorie analytických modelů O. S. Kulaginové, I. I. Revzina a S. Markuse, u nás pak funkční generativní popis jazyka P. Sgalla.

Přesný popis každé z těchto teorií by byl poměrně rozsáhlý, proto si zde ukažme alespoň základní principy dvou z nich, a to **generativní a transformační mluvnice** a dále **kategoriální gramatiky**.

1. Generativní a transformační mluvnice

Zakladatelem generativní nebo transformační gramatiky (popřípadě se označuje oběma názvy) je americký jazykovědec Noam Chomsky, který vypracoval v letech 1957 a 1965 dvě poměrně rozdílné verze této mluvnice. Protože si obě varianty generativní a transformační gramatiky získaly řadu příznivců a pokračovatelů, mluvíme někdy ještě o tzv. třetí variantě generativní mluvnice. My si tu představíme principy Chomského první verze z roku 1957.

Slovo generativní nám říká, že jazyk je chápán jako tvůrčí proces, v němž se všechny věty jazyka tvoří (generují) podle předem daných pravidel. Souhrn těchto pravidel tvoří gramatiku jazyka. Aby byl popis gramatiky co nejjednodušší, zavádí Chomsky tzv. **jádrové věty**, tj. základní jednoduché věty, z kterých pomocí transformačních pravidel dostaneme všechny ostatní věty a souvětí (odtud tedy pojem transformační).

Chomského teorii bychom mohli rozdělit na tři části: 1) generování jádrových vět pomocí přepisovacích pravidel; 2) odvození zbývajících vět užitím transformačních pravidel; 3) aplikace fonologických pravidel.

Nejprve Chomsky rozebírá tzv. **mluvnici frázové struktury**. Vyjdeme z tzv. výchozího symbolu (respektive řetězce symbolů), který je možno chápat jako označení věty. Pravidla gramatiky určují, jakým symbolem či řetězcem symbolů lze dřívější symbol (řetězec symbolů) nahradit. Postupnou aplikací těchto pravidel se dostaneme od výchozího symbolu (řetězce symbolů) až k řetězcům koncovým, správně tvořeným větám jazyka. Dále zavádí pojem **frázového ukazatele**, který má podobu stromu a zachycuje, v jakých vzájemných vztazích jednotlivé složky věty jsou.

Uveďme si pro názornost jeden příklad. Velkými počátečními písmeny budeme označovat pomocné symboly (tzn. ty, které se objevují v gramatických pravidlech, ale ne ve větách jazyka) a

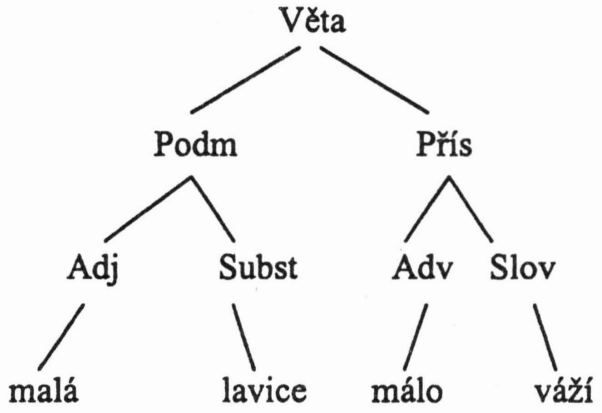
písmeny malými symboly koncové (tzn. symboly abecedy jazyka, které už nelze měnit). Symbol → čteme *nahraď*. Pořadí pravidel je závazné, postupujeme tedy od pravidla 1. k 16. Nechť je dána tato gramatika (frázová struktura):

1. Věta → Podm + Přís
2. Podm → Subst
3. Subst → Adj + Subst
4. Přís → Slov
5. Slov → Adv + Slov
6. Adj → Adv + Adj
7. Subst → pán
   dívka
   lavice ... (a další česká substantiva)
8. Adv → mnoho
   málo
   špatně ... (a další česká adverbia)
9. Slov → zpívá
   jí
   váží ... (a další česká slovesa)
10. Adj → Km Konc
11. Km → mlad
    mal
    menš
12. mlad Konc + pán → mlad-ý + pán
13. mlad Konc + dívka → mlad-á + dívka
14. mal Konc + pán → mal-ý + pán
15. mal Konc + lavice → mal-á + lavice

16. menš Konc → menš-í

Podle ní vytvoříme například věty: *Malá lavice málo váží. Malý pán málo váží. Mladá žena mnoho jí.* Ale můžeme odvodit větu: *Mladá mladá dívka zpívá*, která jistě není gramaticky správnou českou větou. Dále lze získat rovněž věty: *Mladá lavice málo váží. Malá lavice špatně jí.* I ty totiž naše gramatika vymezuje jako správně gramaticky tvořené. Otázkou zůstává, mají-li být takové věty považovány za špatně gramaticky utvořené, nebo zda se jich nepoužívá pouze ze sémantických důvodů, tj. že jsou gramaticky utvořeny správně, ale nejsou smysluplné.

Frázový ukazatel například věty *Malá lavice málo váží* by vypadal takto:



Protože může být frázový ukazatel a frázová struktura v některých případech značně nepřehledná a navíc nemůže tato gramatika zachytit vztahy mezi různými typy vět, navrhuje Chomsky, aby se tato frázová struktura aplikovala pouze u tzv. jádrových vět (ty považuje za axiomy celého systému). Všechny ostatní věty se pak získají z těchto jádrových vět pomocí transformačních pravidel.

Je vidět, že generativní a transformační gramatika je jakousi obdobou matematického deduktivního systému, který je složen: a) z axiomů; b) ze souboru znaků a symbolů logické povahy; c) z pravidel, na jejichž základě odvodíme z axiomů další pravidla a teze a vytvoříme tak celý systém.

Na dalším příkladu si objasněme pojem **transformační pravidlo**. Podle mluvnice frázové struktury podobné té naší odvodíme dejme tomu větu *Jana čte báseň.* Tato věta by patřila do jádrových vět. Zcela jistě existuje vztah mezi větami *Báseň je čtena Janou. Jano, čti báseň! Jana nečte báseň* (převedení do pasiva, převedení do podoby rozkazovací věty a do záporu). Pasivní transformaci základní věty Jana čte báseň na větu Báseň je čtena Janou bychom mohli popsat takto:

$$\mathrm{Subst}^{1}_{\mathrm{Nom}}+\mathrm{Slov}_{\mathrm{Přech}}+\mathrm{Subst}^{2}_{\mathrm{Akuz}} \rightarrow \mathrm{Subst}^{2}_{\mathrm{Nom}}+\mathrm{Slov}_{\mathrm{Pas}}+\mathrm{Subst}^{1}_{\mathrm{Instr}},$$

což obecně znamená, že z každé věty skládající se ze jména (jmenného syntagmatu $\mathrm{Subst}^{1}_{\mathrm{Nom}}$), přechodného slovesa $\mathrm{Slov}_{\mathrm{Přech}}$ a dalšího jména (jmenného syntagmatu $\mathrm{Subst}^{2}_{\mathrm{Akuz}}$) získáme větu pasivní tak, že obě jména si prohodí místo, nominativ se změní na instrumentál, akuzativ na nominativ a sloveso se převede do pasivní podoby. Současně by muselo být ještě pomocnými symboly formulováno, že druhé substantivum je předmětem slovesa, neboť například větu *Jana čte celou noc* (tady je druhé substantivum příslovečným určením) nemůžeme převést do pasivní podoby.

Protože generováním jádrových vět pomocí přepisovacích pravidel a derivací ostatních vět pomocí transformačních pravidel ještě nezískáme skutečné věty daného jazyka, jen posloupnosti určitých symbolů, rozlišuje Chomsky ještě tzv. **fonologická pravidla**, která získané symboly přemění na fonetickou podobu věty, tj. například sled symbolů $\mathrm{Subst}^{2}_{\mathrm{Nom}}+\mathrm{Slov}_{\mathrm{Pas}}+\mathrm{Subst}^{1}_{\mathrm{Instr}}$ na větu *Báseň je čtena Janou.* Tato třetí složka zůstává pouze naznačena, autor ji blíže nerozebírá.

2. Kategoriální gramatika

Generativní gramatika je tedy soubor pravidel, jejichž aplikací získáme z výchozího symbolu všechny gramaticky správné věty daného jazyka (a jejich strukturní popisy). Lze ale postupovat opačně, tzn. vycházet od konkrétní věty, převést ji na řetěz symbolů a zjišťovat, zda je to gramaticky správná věta. Tento

typ gramatiky nazýváme gramatikou rekognoskativní. Příkladem takové rekognoskativní gramatiky může být tzv. kategoriální gramatika, kterou v 50. letech minulého století vypracoval Yehoshua Bar-Hillel. Podstata je následující:

1. Každému slovnímu tvaru zadané věty můžeme přiřadit určitou kategorii (odtud kategoriální gramatika) a tu pak nahradit odpovídajícím symbolem. Celou větu tak lze převést na řetěz symbolů.

2. Rozlišujeme dva druhy kategorií:

   - tzv. **základní kategorie** (sem řadíme věty se symbolem S a názvy pojmů, tj. substantiva, která mohou být holým podmětem – symbol N)
   - tzv. **operátory** (všechny ostatní slovní tvary, které nejsou základními kategoriemi, ale nějak se k nim vztahují)

3. Vedle jednoduchých kategorií S a N rozlišujeme složené kategorie – např. N/N, (N/N)/(N/N)

4. Zavádíme symboly připomínající zlomkové čáry, a to / a \:

   - N/N („N nad N“) – odpovídá např. adjektivu, které ve větě předchází před příslušným substantivem, na němž je závislé
   - N\S („N pod S“) – odpovídá zpravidla slovesu, jehož tvar je řízen podmětem, který před slovesem předchází

5. Definujeme operace krácení, které nám umožňují dvojice sousedních symbolů v řetězu symbolů nahradit symbolem jediným. Rozlišujeme dva druhy krácení:

   - **krácení zprava** (posloupnost symbolů N/N,N zkrátíme na N, tak jako zlomek $1/5 \cdot 5 = 1$)
   - **krácení zleva** (N,N\S lze krátit na S, podobně jako $5 \cdot 1/5 = 1$)

6. Každou větu můžeme převést na odpovídající posloupnost symbolů (takových možností bývá často několik, podobně jako některým slovním tvarům lze přiřadit několik kategorií). Na každou z těchto posloupností lze aplikovat krácení zleva a zprava (opět v různém možném pořadí). Podaří-li se nám alespoň jednou získat jako výsledek krácení jednoduchý symbol, jedná se o gramaticky správnou větu daného jazyka. Opět si demonstrujme tuto teorii na jednoduchém příkladu. Nechť je dána tato věta:

Pěkně napsaná slohová práce vždy potěší.

1. Tvar práce může plnit funkci holého podmětu, proto mu přiřadíme symbol N. Adjektiva *napsaná* a *slohová* se váží k substantivu *práce*, předchází jej a spolu s ním plní funkci podmětu, proto jim náleží symbol N/N. Tvar *potěší* je závislý na tvaru *práce*, který mu ve větě předchází, proto mu odpovídá symbol N\S. Tvar *pěkně* se vztahuje k adjektivu *napsaná*, které následuje až po něm, proto jej nahradíme symbolem (N/N)/(N/N). A konečně tvar *vždy* závisí na tvaru *potěší*, ve větě stojí před ním, přiřadíme mu proto symbol (N§)\(N§). Celou větu můžeme tedy převést na tento řetěz šesti symbolů:

(N/N)/(N/N),N/N,N/A,N,(N\S)/(N\S),N\S.

2. Získaný řetěz symbolů můžeme dále krátit. Neboť se zde vyskytují oba druhy zlomkových čar, můžeme použít krácení zleva i krácení zprava (samozřejmě v různých pořadích). Jedno z těchto pořadí si zde ukažme:

a) Krácením páté a šesté kategorie zprava získáme výsledný symbol N\S. Celý řetěz se nám tedy zkrátí na sled těchto pěti symbolů: (N/N)/(N/N), N/N, N/N, N, N\S, což by odpovídalo větě *Pěkně napsaná slohová práce potěší.*

b) V tomto řetězu symbolů můžeme krátit opět zprava třetí a čtvrtou kategorii na symbol N, čímž obdržíme

řetěz čtyř symbolů: (N/N)/(N/N), N/N, N, N\S, odpovídající větě *Dobře napsaná práce potěší.*

c) Dále můžeme krátit zprava první a druhý symbol na podobu N/N a dostaneme tento tříčlenný řetěz symbolů: N/N,N,N\S (*Napsaná práce potěší*).

d) Zkrácením zprava prvního a druhého symbolu na symbol N získáme dvoučlenný řetěz symbolů: N,N\S, příslušný větě *Práce potěší.*

e) A konečně krácením zleva těchto dvou zbývajících článků řetězu obdržíme výsledný jednoduchý symbol S, jako kdybychom předcházející větu zkrátili na větu *Potěší.*

Vidíme, že po převedení věty *Dobře napsaná slohová práce vždy potěší* na řetěz symbolů a po příslušném krácení jsme získali jediný jednoduchý symbol (S). Znamená to tedy, že daná věta je gramaticky správnou českou větou.

Využití takovéto gramatiky můžeme najít zejména při strojovém překladu, kde je nezbytné zjistit, zda jsou tvořeny správné věty daného jazyka. Nevýhodou je to, že je použitelná pouze pro jazyky s málo rozvinutou morfologií a pevným slovosledem, zejména tedy pro angličtinu.

Je zřejmé, že v lingvistice nevystačíme s jednoduchými typy gramatik. Otázkou ovšem zůstává, zda je vůbec možno tak složitý systém, jakým přirozený jazyk je, matematickými prostředky popsat. Zcela jistě takovému popisu odolává sémantika. Mluvnici jazyka (tzn. syntax, morfologii a fonologii) víceméně popsat lze. A i když je to úkol velmi náročný, vyžaduje si jej sama praxe (zejména počítačová lingvistika). V této souvislosti můžeme vyzvednout ten fakt, že má-li matematika problémy s popisem některých jazykových jevů, není tento jev dokonale zpracován ani jinými lingvistickými metodami.

Mgr. Blanka Sedlačíková
doktorandka Katedry matematiky PřF MU
Janáčkovo nám. 2a, 662 95 Brno
e-mail: hvezdova@math.muni.cz