2 Teoretická lingvistika a strojový překlad I

2.1 Gramatika

- Gramatika v tradičním evropském (a americkém) pojetí popisuje kompetenci (jazyk jako abstraktní systém), narozdíl od performance (užívání jazyka).
- V teoretické lingvistice se klade důraz na formální metody, aby byl popis jednoznačný a ověřitelný, případně implementovatelný. Hlavním cílem je formulace popisu na co nejvyšší úrovni obecnosti, tedy minimalizace počtu specifických pravidel gramatiky.
- Gramatika by dále měla být:
  - empiricky ('observačně') adekvátní (t.j. měla by umět rozhodnout, zda určitý řetězec znaků a slov je správně utvořený výraz daného jazyka)
  - deskriptivně adekvátní (t.j. měla by umět přisoudit správně utvořeným výrazům a větám strukturní popis, odpovídající intuici mluvčích daného jazyka)
  - explanatorně adekvátní (t.j. měla by umět popsat daný jazyk v souvislostech s obecnými principy postulovanými pro všechny přirozené jazyky)
- Universální gramatika — zobecnění gramatik všech jazyků.
- Měla by být formální gramatika psychologicky realistická?

2.2 Komponenty (formálního) popisu jazyka

- fonologie
  - hlavní problém při analýze: variabilita fonémů podle kontextu, mluvčího, okolností promluvy; při řešení se často hledá nejpravděpodobnější interpretace posloupností zvukových signálů pomocí statistických metod
  - hlavní problém při syntéze: intonace
- ortografie
  - viz morfologie a slovník
  - interpunkce (oddělování volných a těsných přívlastků, uvozovky), zdůraznění (kurzívou, proloženým písmem), velká nebo malá počáteční písmena (adjektiva národností, ap.), grafická úprava typů textů
- morfologie
  - derivační (přípony a předpony, skládání slov)
  - flektivní nebo aglutinační; analytické tvary se obvykle řeší v rámci syntaxe
- slovník
  - 1 heslo : 1 zákl.tvar nebo 1 heslo : 1 tvar
  - údaje morfologické: morfologický vzor; údaje syntaktické: slovní druh, podtřída (N: (ne)počitatelnost, V: (ne)přechodnost, valence); údaje sémantické (sémantická třída, selekční omezení)
- syntax
- sémantika
- pragmatika

V systémech strojového překladu je podstatné vyjádření ekvivalencí mezi jednotkami popisu (morfémy, lexémy, kolokacemi, syntagmaty) různých jazyků. V systémech s transferem to obvykle znamená, že kromě částí (slovník, pravidla gramatiky) popisujících určitý jazyk nezávisle na jiném jazyku jsou přítomny také dvou- nebo vícejazyčné komponenty:

- překladový slovník
- transfer struktur

Podmínky pro uplatnění určité operace v transferu se mohou odvolávat na informace morfologické, syntaktické, sémantické či pragmatické (*ženou* – V/N, *půjčit, učit, kolej, ahoj*, téma/réma)

2.3 Syntax

- pravidla a/nebo principy, podle nichž se slova kombinují do větších celků
- pravidla popisují chování nikoli jednotlivých slov, ale kategorií
- kategorie mohou být: lexikální (autosémantické/synsémantické) nebo frázové; atomické nebo strukturované
- syntax se zabývá popisem vztahů mezi kategoriemi (větnými členy): pořadí, závislost, těsnost spojení

Strukturu věty lze zachytit grafem v podobě stromu. Existují dva základní typy takového grafu (stromu, struktury, reprezentace):

- Frázový (složkový) graf postupně člení větu na menší (obvykle souvislé) části (fráze, složky, konstituenty), až dojde nakonec k jednotlivým slovům.
- Závislostní graf representuje závislostní vztahy mezi slovy.

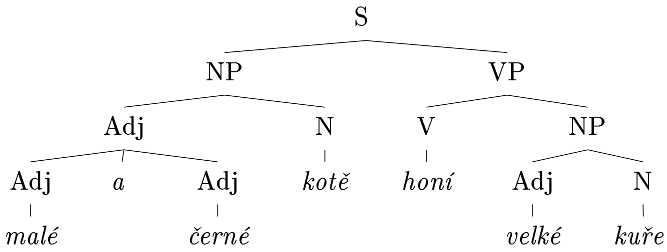
Podoba representace je závislá na gramatice, která byla užita. Gramatika vyjadřuje vztah mezi větou jakožto řetězcem (fonémů/písmen, morfémů, slov) a representací této věty.
Při formalisovaném popisu jazyka se používají **generativní gramatiky**, které obsahují **přepisovací pravidla**. Tato pravidla *generují* representaci[1] věty a jejích částí. Generováním se zde myslí statický vztah mezi gramatikou a representací. Přepisovací pravidlo se použije tehdy, když lze najít v dosud vygenerovaném zápisu symbol(y) odpovídající levé straně pravidla.
Generativních gramatik je více typů, nejčastěji se při popisu syntaxe setkáme s gramatikami **nekontextovými**, které spolu s gramatikami **kontextovými** patří do třídy gramatik **frázových**. Níže je uveden příklad nekontextové gramatiky a jedné ze struktur, kterou tato gramatika umí generovat.

[1] Někdy se rozlišuje derivační zápis a syntaktická representace.

Repertoár symbolů tvoří startovní symbol S, neterminální symboly (A, B a další s velkým počátečním písmenem), terminální symboly (a, b, ... w a slova – řetězce v uvozovkách “ ... ”), řetězce terminálních i neterminálních symbolů (x, y).

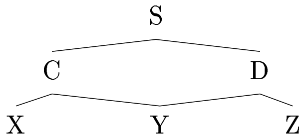
$S \longrightarrow NP\ VP$
$NP \longrightarrow Adj\ N$
$NP \longrightarrow N$
$VP \longrightarrow V\ NP$
$Adj \longrightarrow$ “malé, velké, černé”
$N \longrightarrow$ “štěně, kotě, kuře”
$V \longrightarrow$ “honí, trápí, krmí”
$X \longrightarrow X$ “a” X

“Metaproměnná” $X = S \vee NP \vee VP \vee V \vee Adj$

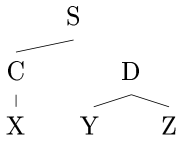


Reprezentace věty podle frázové gramatiky (phrase structure grammar – PSG) musí splňovat tyto podmínky:

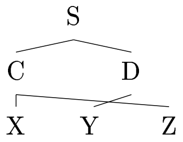
1. Žádný prvek nesmí být bezprostřední složkou dvou různých frází.



2. Každý prvek musí být bezprostřední složkou nějaké fráze, kromě kořene S.



3. Fráze se skládá z prvků, které spolu sousedí.



Pravidla frázové gramatiky musí splňovat podmínky:

1. Na levé straně pravidla musí být jediný přepisovaný symbol. *(U kontextových gramatik může levá strana obsahovat i symboly týkající se okolí přepisovaného symbolu.)*

2. Pravá strana nesmí být prázdná. *(Tato podmínka je obsažena v původní formulaci N. Chomského a většinou se od ní upouští.)*

Pravidla, která nejsou pravidly frázové gramatiky:

a. $\epsilon \longrightarrow rst$

b. $Q \longrightarrow \epsilon$ *(jen podle Chomského formulace)*

c. $CDG \longrightarrow W$

d. $GHI \longrightarrow mnop$

Frázová gramatika může být nekontextová (context-free, CF-PSG) nebo kontextová (context-sensitive, CS-PSG). Kromě frázových gramatik mezi generativní gramatiky patří ještě gramatiky **regulární** a **obecné přepisovací** systémy. Tyto třídy gramatik lze uspořádat podle *generativní síly.*
Stupnice generativních gramatik (Chomsky 1959):

- Typ 0: obecný přepisovací systém – na levé straně pravidla může být jeden nebo více symbolů, na pravé cokoli, včetně prázdného řetězce.
- Typ 1: kontextová frázová gramatika – $xAy \longrightarrow xay$
- Typ 2: nekontextová frázová gramatika – $A \longrightarrow a$
- Typ 3: regulární gramatika – $A \longrightarrow a$, nebo $A \longrightarrow aB$

Čím vyšší typ, tím slabší generativní síla. Cílem je popsat jazyk gramatikou co nejslabší. Přirozené jazyky přesahují generativní sílu tradičně chápané nekontextové gramatiky.
Příklady: shoda, volný slovosled (rozdělení na podmětovou a přísudkovou část neplatí), “topologické” uspořádání věty (němčina, příklonky v češtině), neprojektivní konstrukce (crossing dependencies, long-distance dependencies, unbounded dependencies) *Ríšu nemohu pozvat. Ríšu jsem nemohl pozvat. Ríšu jsem Božence slíbil pozvat. Ríšu jsem myslel, že Boženka nepozve. Ríšu jsem už Marii slíbit pozvat nestihl. Koho sis myslel, že pozvu? Koho sis myslel, že slíbím Marii, že pozvu?*
Možnosti, jak to napravit:

- transformace
- nahrazení atomických symbolů strukturovanými kategoriemi

Anotace symbolů dalšími informacemi a strukturované kategorie umožňují zmenšit počet pravidel a tím dosáhnout lepší ‘explanatornosti’.
Jak vytvořit takový popis jazyka?

- symbolický popis
- stochastické modelování

2.4 Frázová a závislostní gramatika

PSG a závislostní gramatika (alespoň podle Gaifmana) jsou *slabě ekvivalentní.*
Příklad: jazyk $a^n b a^n$
PSG:
$X \longrightarrow aXa$
$X \longrightarrow b$

2.4.1 Závislostní gramatika – Gaifman

(podmínka: projektivita)

- Pravidla:
    1. $X(*, Y)$
    2. $Y(X, *)$
    3. $Y(Z, *)$
    4. $Z(*)$
- Kategorie:
    1. $X : a$
    2. $Y : a$
    3. $Z : b$
- Startovní symbol: X
- Problém: více členů závislých na jednom řídícím uzlu

Závislostní gramatika – FGP

- Pravidla:
  $X \longrightarrow \langle a\rangle X \langle a\rangle$
  $X \longrightarrow b$
- Problémy:
    1. $X \longrightarrow X\langle Y_1\rangle$
    2. $X \longrightarrow X\langle Y_2\rangle$
    3. $X \longrightarrow \langle Y_3\rangle X$

  1,2: pořadí Y_1, Y_2
  1,3: spurious ambiguity
  dvě pravidla pro generování závislého členu zleva a zprava

LITERATURA:

P. Sgall a kol.: Úvod do algebraické lingvistiky, SPN 1974