Problémy recyklování systému automatického překladu

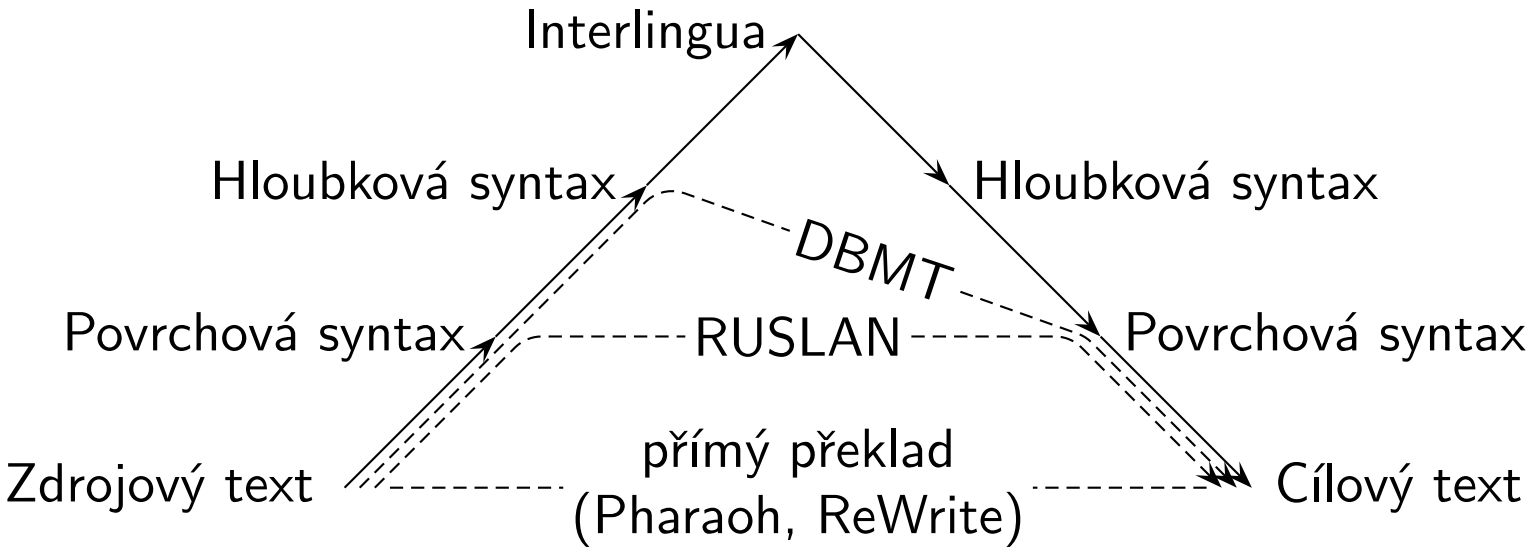
Ondřej Bojar, Petr Homola, Vladislav Kuboň
{bojar,homola,vk}@ufal.mff.cuni.cz

21. září 2005

Osnova

- Motivace: proč recyklovat systémy strojového překladu
- RUSLAN jako základ
- Syntaktická analýza v RUSLANu
- Nové moduly:
  - Rozpoznávání pojmenovaných entit
  - Česko-anglický slovník
- Problémy stávající gramatiky

Trojúhelník strojového překladu (MT)



DBMT a ReWrite viz **?** a citované, Pharaoh viz **?**

RUSLAN: Automatický překlad čeština→ruština

- ⋆1985 †1990; MFF UK + VÚMS; **?**
- Automatický překlad manuálů k operačním systémům sálových počítačů
- Překlad jedné věty trval asi 4 minuty na IBM PC 286

Nejcennější části systému:

- Syntaktický slovník klíčovou součástí systému
  Obsahoval cca 8500 kmenů
- Gramatika češtiny založená na ručně psaných pravidlech (**?**)
  Doposud jediná svého druhu

⇒Nový experiment:

- Je možné zachránit a znovu použít znalosti investované do systému RUSLAN v kombinaci s novými moduly?
- Odpověď: Možná ano, ale je to výjimečně komplikované

RUSLAN do angličtiny

Cíl:

- překládat ekonomické texty Wall Street Journalu
  Prague Czech-English Dependency Treebank, PCEDT, **?**

Zachovat:

- prostředí Q-systémů (bylo reimplementováno)
- gramatiku pro analýzu češtiny

Nově:

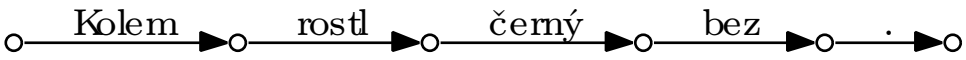
- česko-anglický slovník (jiný jazyk, jiná doména!)
- modul pro pojmenované entity

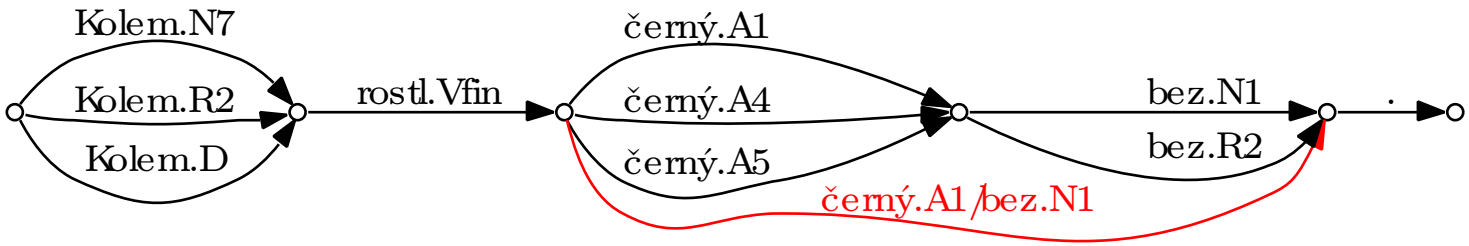
Systémy Q (?)

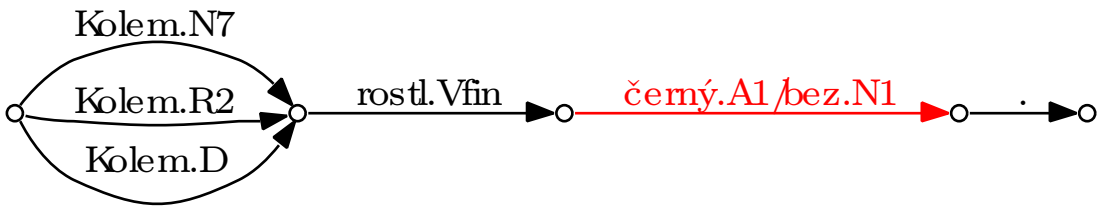
- Grafový analyzátor (chart parser).
- Jednotlivá pravidla se uplatňují nedeterministicky, dokud lze nějaké pravidlo uplatnit.
- Poté následují dvě fáze čištění výsledného grafu:
  1. odstraněny použité vstupní hrany
     tj. hrany, které se objevily na levé straně kteréhokoli uplatněného pravidla
  2. odstraněny vybočující hrany
     tj. hrany, které neleží na cestě od počátečního uzlu ke koncovému

  Pokud se čištěním odstranil celý graf, použije se původní graf z počátku fáze.

Příklad syntaktické analýzy






Kolem.N7
Kolem.R2
Kolem.D
rostl.Vfin
černý.A1/bez.N1
.

Příklad pravidla

```
  A*(U*1,#(U*2,<,U*3,/,U*4,E*(F*,X*9,J*),U*5,/,U*6,>,U*7),U*8,
     1(F*1(Y*),X*1,#(X*),X*2))
+ 7(W*,E*)
+ 1(B*,Z*1,F*(C*),Z*2,@(V*),#(Z*),Z*3)
     ==  A*(U*1,#(U*2,<,/,U*4,U*5,/,>,U*7),U*8,1(F*1(Y*),X*1,#(X*),X*2),
            1(E*(F*),B*(C*),@(V*),#(Z*),T(J*),SCF,Z*3))
               /      -NON- (. + -DANS- X*9 -ET- +(V*) -HORS- X*9,+(VZT,OSOB) .)
                    -ET-       -(V*,CDSUS) -HORS- X*9,*
                    -ET-       VZT -HORS- V*,*
                    -ET-        /  -HORS- U*4,U*5,*
                    -ET-       <,> -HORS- U*3,U*6,*
                    -ET- AD(+(SP)) -HORS- Z*3
                    -ET- (. FM -HORS- Z*3 -OU- 5 -DANS- Z*3 .)
                    -ET- (. -NON- 1(FM,LMCN) -DANS- Z*3 -OU- 1(5) -DANS- Z*3 .)
                    -ET- (.      F* -HORS- Z*1,Z*2,*
                           -OU- C* = S -ET- -NON- *A,*C -HORS- V*
                           -OU- C* = P -ET-       *A,*C -HORS- V* .)
                    -ET- (.      A* -DANS- 1,2 -ET- 2(@),5,6 -HORS- U*8
                           -OU- A*  -DANS- V,6                          .)
                    -ET- (.               E*(F*) -HORS- X*
                           -OU-       E*(+(V*,*)) -HORS- X*
                           -OU- -NON- E*(-(V*)) -HORS- X* .)
```

Předložkový pád vyplňuje pozici ve valenčním rámci zprava.

Charakter překládaného textu

Vedení Chrysleru oznámilo, že závod opět zahájí výrobu 20. listopadu, což se bude týkat 3300 dělníků placených od hodiny.
Chrysler officials said the plant is scheduled to resume production on Nov. 20, and 3,300 hourly workers will be affected.

Modely Corsica a Beretta představují největší vozovou řadu Chevroletu, ale tržby pro tento rok jsou o 9,6 % nižší a na počátku tohoto měsíce prudce spadly až o 34,2 %.
The Corsica and Beretta make up the highest-volume car line at Chevrolet, but sales of the cars are off 9.6% for the year, and fell a steep 34.2% early this month.

- obrovské množství číslovek v nejrůznějších formátech, kalendářových dat, měnových výrazů, procent, zlomků apod.
- velkém množství pojmenovaných entit (názvy firem a institucí, jména osob, geografické názvy apod.)
- idiomy

Pojmenované entity

- Pojmenované entity: souhrnné označení pro sémanticky atomické, ale často víceslovné jednotky v textu.
- Vnitřní strukturou často odlišné od obecných částí gramatiky:
  - časové údaje (dlouhé a krátké datum, čas, kombinace)
  - číselné údaje s jednotkami (měna, metrické ap.; dle gazeteeru[1])
  - vlastní jména osob, organizací, geografické názvy (dle gazeteeru)

první dáma Laura Bushová	*first lady Laura Bush*	⇒stačí transliterace jmen + překlad titulů
150 milionů dolarů	*$ 150 millions*	⇒specifický transfer jmenné skupiny

[1] Gazeteer = specializovaný, typicky strojově plněný slovník s vysokým pokrytím.

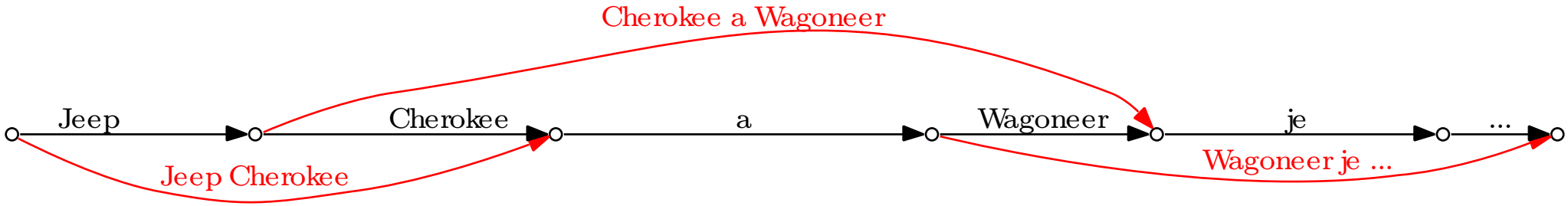
Budování slovníku

- Původní slovník obsahuje cenné informace: valenční rámce, sémantické rysy. . .
- Průnik s novou doménou je však poměrně malý a stará data se nevyplatí použít.

Budování nového slovníku je věnován samostatný příspěvek na této konferenci.
?

Ukázka problémů stávající gramatiky

Kvůli klesajícímu prodeji svých lukrativních modelů sportovních aut Jeep Cherokee a Wagoneer je společnost Chrysler Corp. nucena k dočasné odstávce svého montážního závodu v Toledu ve státě Ohio, a to poprvé od dubna 1986.



- Konstrukce typu *Petr a Pavel šli* je gramatikou řešena, vede však k tomu, že skupina *Cherokee a Wagoneer* dostane množné číslo.
- Gramatika nepodporuje rozvití *Jeep* (koordinovanou) skupinou v mn. č.

$\Rightarrow$ graf z nových hran přestane být souvislý $\Rightarrow$ celý zapomenut.

Problémy jsou sice řešitelné, ale je nutné hluboko zasáhnout do gramatiky.

Shrnutí

- Popsali jsme starý systém strojového překladu.
- Ilustrovali specifika odlišné domény překládaných textů.
- Nastínili nový modul pro rozpoznávání jmenných entit.
- Popsali jsme problémy stávající gramatiky.

Závěr: Recyklace starých systémů se spíše nevyplatí.

References