Budování česko-anglického slovníku pro strojový překlad

Ondřej Bojar
bojar@ufal.mff.cuni.cz

21. září 2005

Osnova

- Motivace
- Problémy dostupných česko-anglických slovníků
- Ruční čištění: metainformace, slovní druhy, částečně závislostní struktura
- Kde se opřít o korpus: morfologická omezení, paralelní slovesné rámce
- Podrobněji o metodě extrakce paralelních slovesných rámců
- Shrnutí a plány do budoucna
- Pomíjené syntaktické problémy

Motivace

- Strukturální strojový překlad (MT) nezbytně vyžaduje podrobné překladové slovníky

  Konkrétně potřebujeme připravit česko-anglický slovník pro systém RUSLAN (**?**).

- Pro češtinu a angličtinu neexistují dostatečně podrobné slovníky pro strojový překlad.

  Dostupné strojově čitelné slovníky, např. **?** nebo WinGED, neobsahují potřebnou syntaktickou informaci buď vůbec, nebo v nepoužitelné podobě.

  I další výzkumníci (**?**) se museli spokojit se zjednodušeným slovníkem obsahujícím jen jednoslovná hesla.

Problémy dostupných česko-anglických slovníků

- Nekonzistentní nebo nejednoznačná anotace metainformací.

V textu hesel najdeme	Příklad:
poznámku o stylu či dialektu	*am.*, *mat.*
oddělovač více variant výrazu	*be liable/subordinate to*
označení nepovinné části	*(auto)stop*, *(na)líčení tváře*
tvaroslovné příznaky	*průdušky pl.*, *pyramida f*
náznak syntaktických vlastností	*adjudge sb. to be guilty = uznat vinným koho*

- Chybějící informace o slovním druhu a další morfologické vlastnosti hesel.
- Chybějící syntaktické vlastnosti víceslovných hesel.

Určení slovních druhů: ručně!

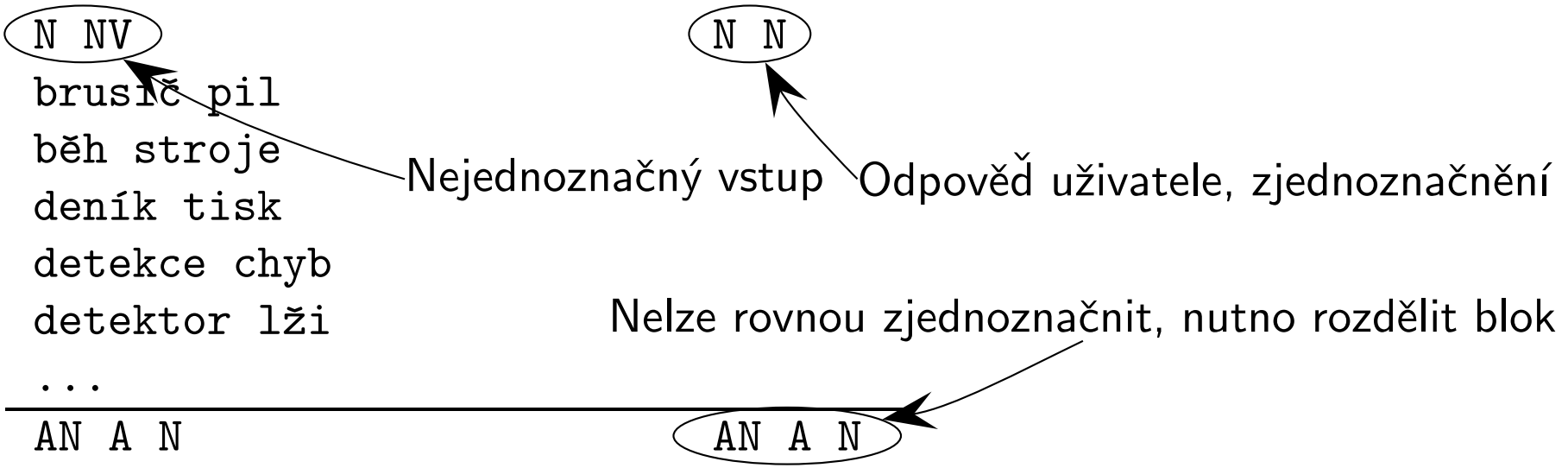
Subst. a subst./adj.	**Správná interpretace**	**Anglický ekvivalent**
husa divoká	subst. adj.	grey goose
kniha účetní†	subst. adj.	account book
napětí dovolené†	subst. adj.	permissible stress
chyba měření	subst. subst.	measurement error
plán prací†	subst. subst.	schedule of operation
rozsah měření	subst. subst.	range of measurement
Čísl./sloveso a subst.	**Správná interpretace**	**Anglický ekvivalent**
tři prdele†	čísl. subst.	shitloads
pět švestek	čísl. subst.	one's duds*
pět chválu	sloveso subst.	sing someone's praises

† Hesla dovolují i druhou interpretaci, většinou poněkud směšnou.

* Součást idiomu *pick up one's duds*.

Ukázka rychlé ruční práce

Anotátor ať pracuje vždy najednou s celými bloky hesel se stejným problémem:



Další morfologická omezení automaticky

1. Najdi věty obsahující všechna lemata z hesla bez ohledu na pořadí.
2. Pro každý z předem definovaných morfologických rysů zjisti, zda ve většině případů není nalezenými slovy splněn.

Český výraz	**Unární omezení**	**Binární omezení**
za nízkou cenu	RR–4 AAF** NNF*4	pčr:2=3
v jistém smyslu	RR–6 AAIS6 NNIS6	pád:1=2 pád:1=3 pčr:2=3
získané informace	AAFP* NNFP*	pčr:1=2
první světová válka	CrFS* AAFS* NNFS*	pčr:1=2 pčr:1=3 pčr:2=3
v jediném dnu	RR–6 AA*S* NNIS*	číslo:2=3
bohatý člověk	AA*** NNM**	číslo:1=2
v jiném stavu	RR–6 AA*S* NNIS*	

- Pracuje dobře v případě typických kolokací.
- Nezvládne hesla složená z velmi četných slov.

Doplnění syntaktických informací

Ukázka různých struktur i při stejných slovních druzích:

Subst. adj. subst.	**Syntaktická struktura**	**Anglický ekvivalent**
Komise Evropské unie		Commission of the European Community
náhrada způsobené škody		dilapidation
látkou potažené sedadlo		fabric-covered seat
nevolnost způsobená pohybem		kinesia

- Podobně jako u morfologie lze postupovat:
  - ručně po blocích
  - automaticky dohledáváním častých struktur nad hesly
- Až bude syntaktická informace dostupná, bude možné:
  - Přidávat morfologická omezení automaticky pro dané typy hran (adj.→subst.)
  - Hledání příkladů hesel bude přesnější.

Překladové slovesné rámce

Cíl: Ze dvou (povrchově) syntaktických analýz + automatického zarovnání po slovech potřebujeme získat překladové slovesné rámce.

Např. *dělit=divide na+4.pád=into*