Vytváření, porovnávání a párování tektogramatických stromů různých jazyků

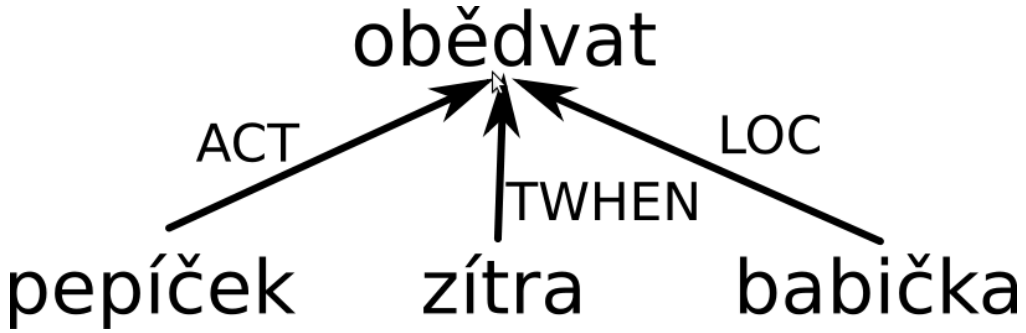
David Mareček

marecek@ufal.mff.cuni.cz

10. dubna 2009

Co je tektorgramatický strom?

- orientovaný zakořeněný strom
- zachyzuje význam věty
- uzly stromu odpovídají plnovýznamovým slovům ve větě
- hrany odpovídají jejich vzájemným vztahům
- každý uzel má množství atributů

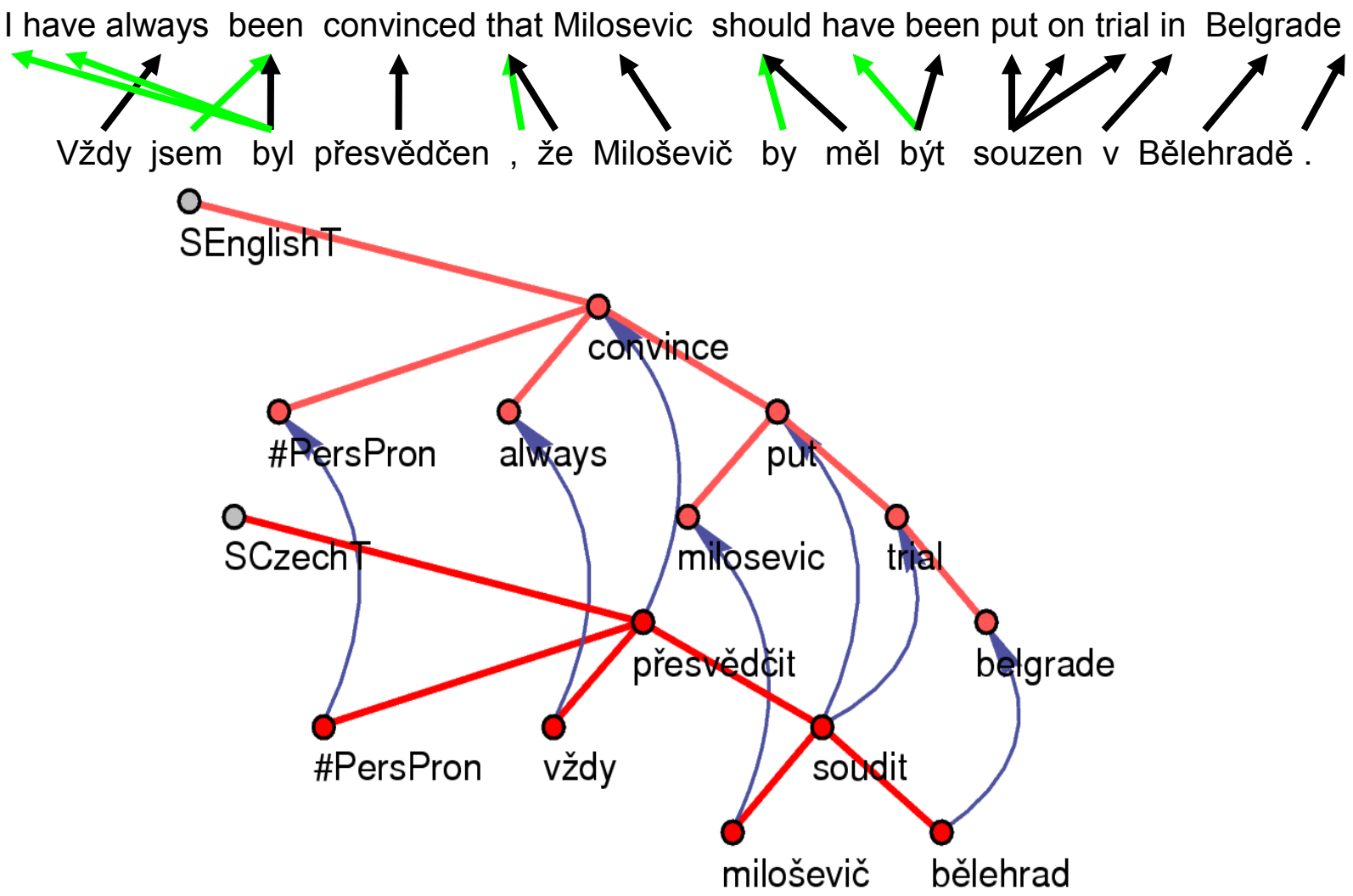


Pepíček bude zítra obědvat u babičky.

K čemu jsou t-stromy dobré?

- Reprezentace věty srozumitelnější pro stroj
  - extrakce informací z textu
  - QA systémy
- Tektogramatické stromy ekvivalentních vět v různých jazycích mají podobnou strukturu
  - Automatický překlad přes tektogramatickou rovinu
    - Projekt TectoMT (primárně čeština-angličtina)
  - Extrakce dvojjazyčného slovníku z paralelního jazykového korpusu
    - Tektogramatická analýza + alignment + statistika

Tecto-alignment x word-alignment



Moje práce

- párování uzlů tektogramatických stromů různých jazyků
  - do teď CZ-EN, v budoucnu chci přidat další jazyky (DE, RU)
  - hladový algoritmus, maximální párování
  - použití perceptronu, několik desítek rysů
- automatické vytváření tektogramatických stromů
  - zatím především pravidlově, jazykově závisle
  - snaha o co největší podobnost stromů mezi různými jazyky
  - v budoucnu automaticky, jazykově nezávisle, učení bez učitele
- porovnávání tektogramatických stromů pro ekvivalentní věty v různých jazycích
  - podobnější stromy => lepší párování uzlů => lepší slovník => lepší překlad