

Automatické párování uzlů českých a anglických tektogramatických stromů

David Mareček

pondělní seminář

13. 10. 2008

Osnova

- rozdíly mezi párováním na morfologické a na tektogramatické rovině
- ruční párování slov na paralelním korpusu
- nástroj pro automatické párování uzlů tektogramatických stromů
- evaluace, porovnání výsledků
- plány do budoucna

Motivace

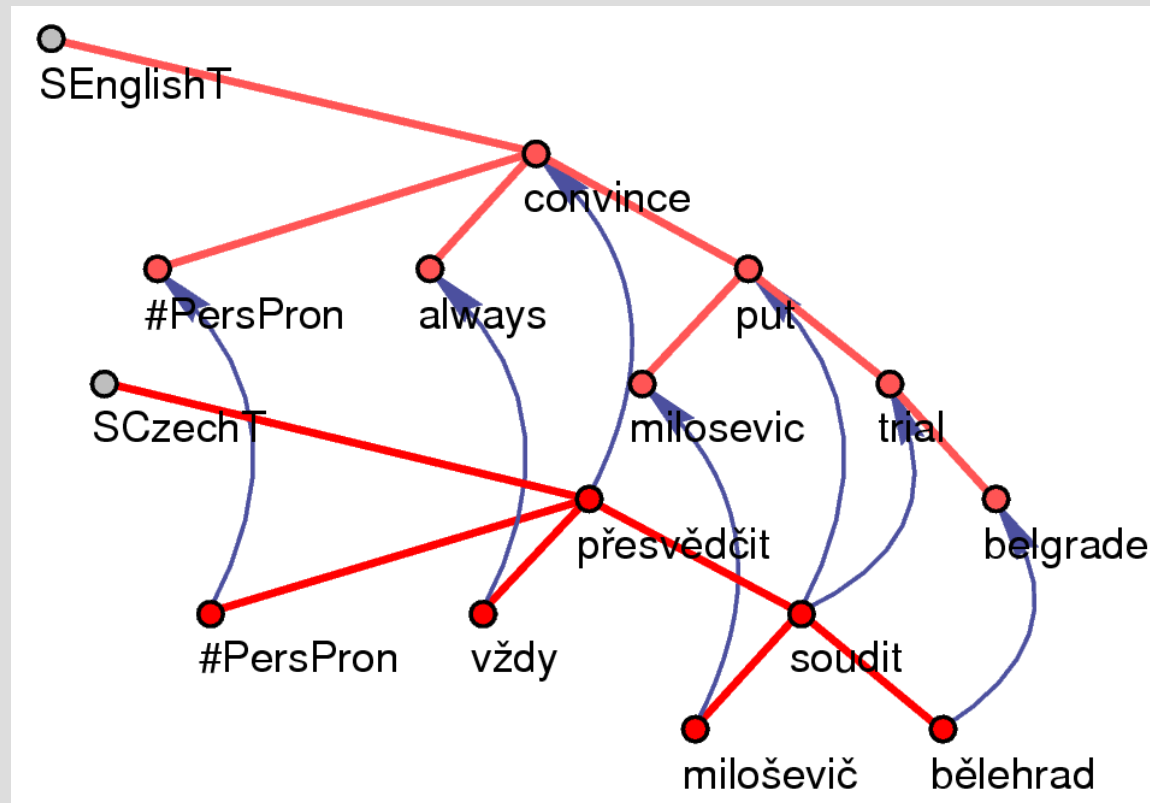
- Na t-rovině jsou si jazyky podobnější \Rightarrow alignment by zde měl být „jednodušší“ než na m-rovině.
- Zlepšení kvality automatického překladu založeného na transferu přes tektogramatickou rovinu (TectoMT)
- Vyhledávání překladových ekvivalentů mezi dvěma jazyky

Příklad

I have always been convinced that Milosevic should have been put on trial in Belgrade .

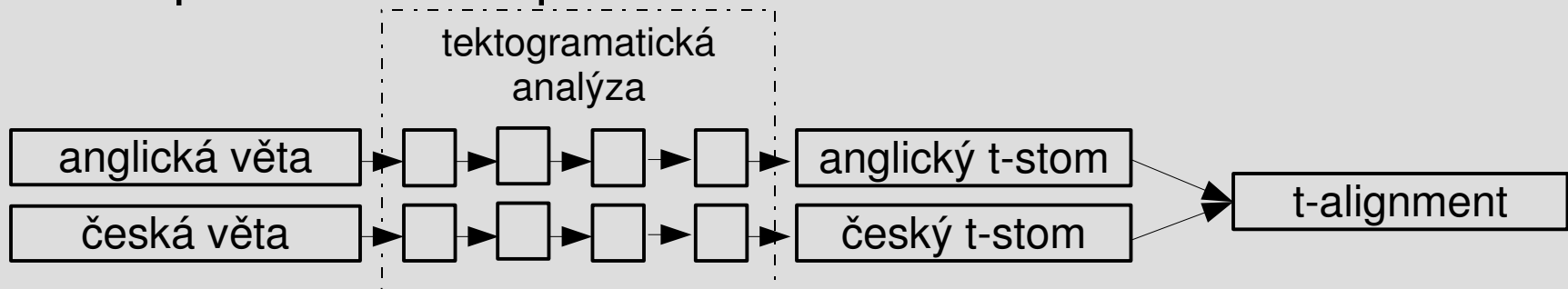


Vždy jsem byl přesvědčen , že Milošević by měl být souzen v Bělehradě .



Úkol

- Implementovat nástroj, který automaticky spáruje uzly českého a anglického tektogramatického stromu
- Měl by využívat stromových struktur, ale zároveň by měl být dostatečně odolný proti chybám vzniklým během parsingu
- Implementace v prostředí TectoMT



Zjednodušená t-rovina

- Struktura t-roviny převzatá z TectoMT
 - Chybí aktuální členění (pořadí uzlů v t-stromu odpovídá pořadí slov ve větě)
 - Chybí „kopírované uzly“
 - Chybí odkazy do valenčního slovníku
 - Chybí funktory (kromě koordinačních)

Test data pro t-alignment

- ruční párování automaticky vygenerovaných stromů **není** vhodné
 - po změně nástrojů pro automatickou analýzu by se stala tato data nepoužitelnými
- **Řešení:**
 - ruční spárování slov na morfologické rovině
 - následný převod na t-rovinu pomocí odkazů **lex.rf**

Manuální word-alignment

- K dispozici 515 ručně anotovaných vět z PCEDT (Ondřej Bojar, Magda Prokopová)
- Tři typy hran:
 - ♦ ***Sure*** – spojuje dvě ekvivalentní slova
 - ♦ ***Phrasal*** – fráze si navzájem odpovídají, samostatná slova však nikoli, spojuje se každé slovo s každým
 - ♦ ***Possible*** – možná spojení, typicky připojuje slovo, které ve druhém jazyce nemá svůj protějšek, ale k některému syntakticky patří (např. členy v angličtině)

Příklady word-alignmentu

Not long before the visit of the Chinese premier , India hosted US Secretary of State Condoleezza Rice .
Nedlouho před návštěvou čínského premiéra hostila Indie ministryni zahraničí USA Condoleezu Riceovou .

I stuck it out as far as ever it would go , and I shut one eye , and try to examine it with the other .
Vyplázl jsem ho tak daleko , jak to jen šlo , zavřel jsem jedno oko a snažil se druhým ho prohlédnout .

- > sure link
- > possible link
- > phrasal link

Anotace

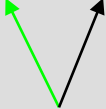
- anotace dalších 1985 vět z paralelního corpusu CzEng
- 4 anotátoři (MFF, FF, 2x ÚJČ)
- každá věta anotována nezávisle dvěma anotátory
- vytvořena další pravidla pro anotaci některých častých jazykových jevů

typ dat	vět	EN tokenů	CZ tokenů
Právní texty (Acquis Communautaire)	501	13 512	10 752
Beletrie (Reader's Digest, E-Books)	500	10 097	8 978
Novinové články (Project Syndicate)	484	10 714	9 990
Nov. Čl. s pojmenovanými entitami (Project Syndicate)	500	12 799	11 052

Pravidla

- týkají se především **possible** hran
- popisují, kam připojit dané slovo, pokud nemá ve druhém jazyce svůj protějšek nebo zda-li ho vůbec připojovat
- členy a předložky

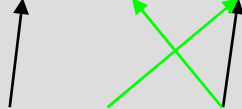
the dog
pes




depends on
závisí na



play the violin
hrát na housle



equivalent to
stejný jako



Pravidla

- zájmena

she laughed
smála se

help him
pomoz mu

I remember
pamatuji si

- pomocná slovesa

he is n't sleeping
nespí

we worked
pracovali jsme

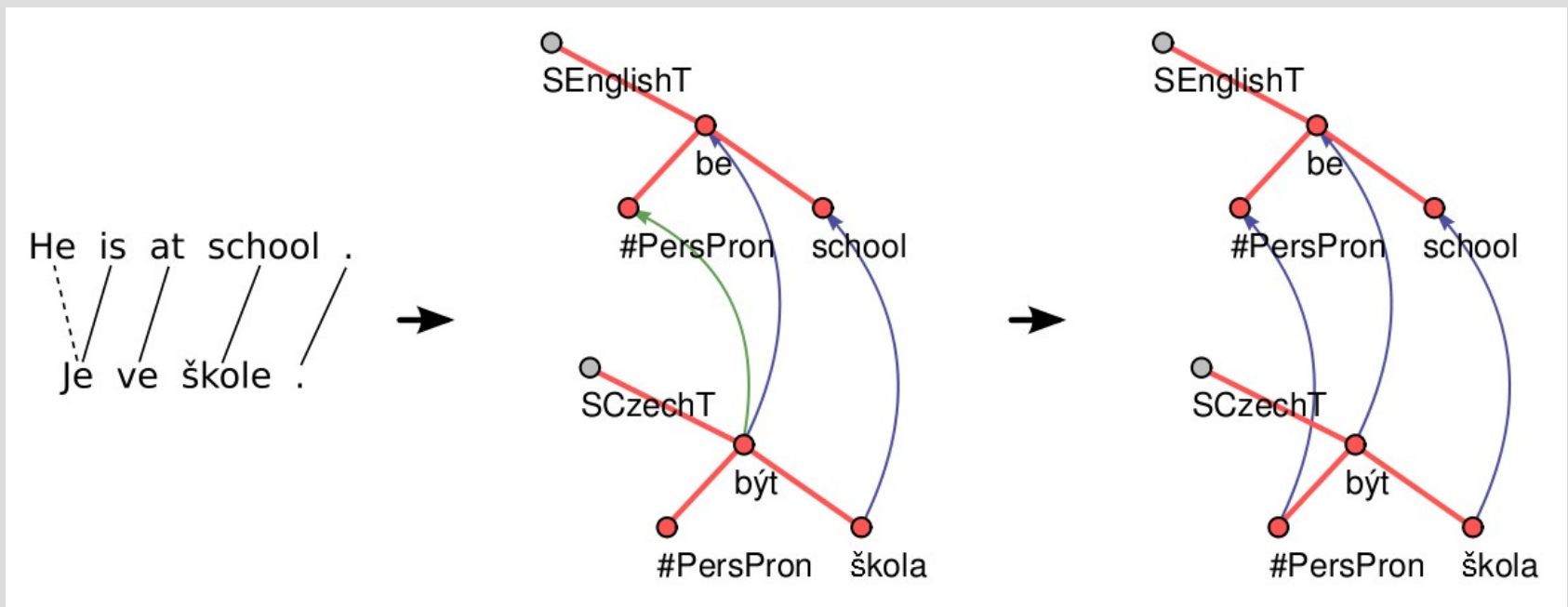
I have never been
nikdy jsem nebyl

Převod do t-roviny

- využití atributu **lex.rf**
- dva uzly na t-rovině jsou spojeny hranou \Leftrightarrow jejich odpovídající tokeny na m-rovině jsou spojeny hranou
- typy hran jsou zachovány
- problém: přidané uzly, např. **#PersPron** reprezentující nevyjádřený podmět

Nevyjádřený podmět

- heuristický algoritmus zajistí alespoň správné připojování českého nevyjádřeného podmětu
 - zásah do ruční anotace

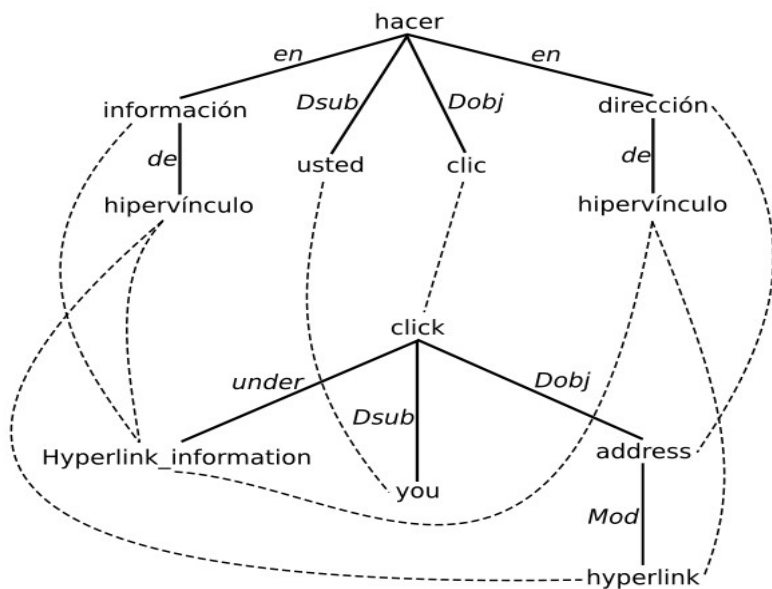


Mezianotátorská shoda

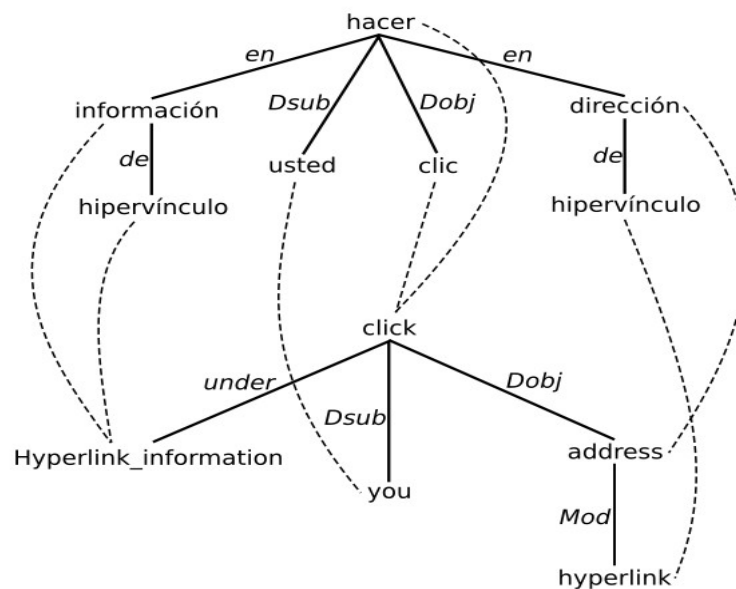
typ dat	typy hran rozlišovány		typy hran nerozlišovány		pouze hrany „sure“	
	m	t	m	t	m	t
právní texty	86,7	92,0	92,1	96,1	94,0	95,6
beletrie	76,6	82,9	85,8	90,5	90,1	91,2
novinové články	80,8	87,9	87,8	93,3	92,0	93,7
pojmenované entity	86,5	94,3	91,5	96,5	93,6	96,4
PCEDT	84,1	94,3	89,8	95,1	93,8	95,3

Automatický t-alignment

- Haruno and Yamazaki (1996) - myšlenka párovat pouze autosémantické uzly
- Menezes and Richardson (2001) – párování uzlů sémantických stromů



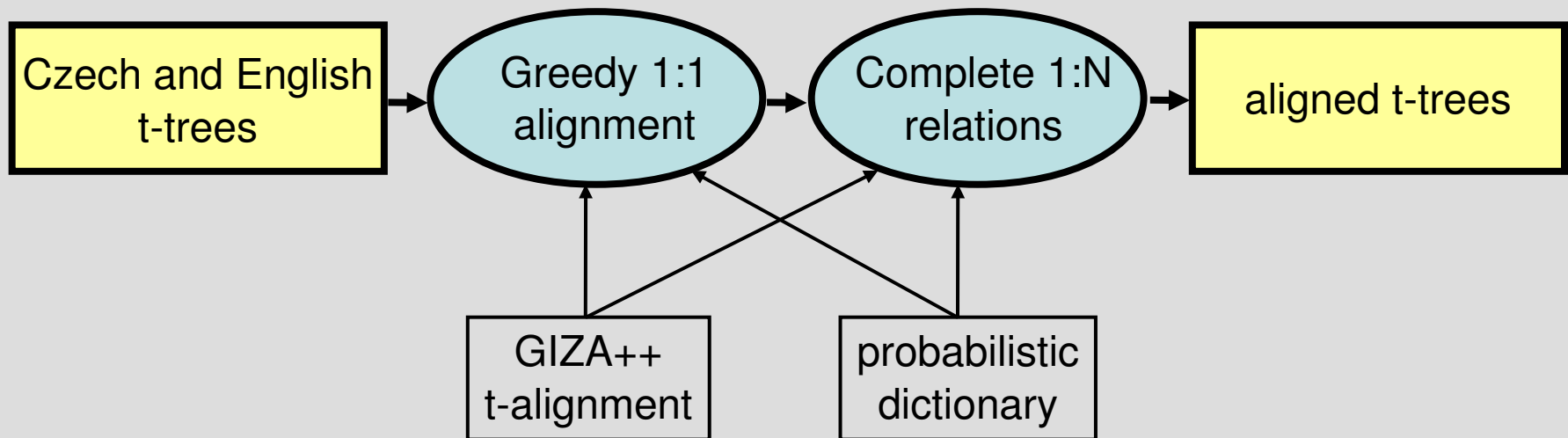
a)



b)

T-aligner

- Inspirace článkem Menezese a Richardsona
- Implementace v prostředí TectoMT
- 2 fáze



Fáze 1 – Greedy 1:1 alignment

- hladový algoritmus založený na rysech (features)

- skóre potenciálního páru t-uzlů:

$$\text{score}(\text{cnode}, \text{enode}) = \vec{w} \cdot \vec{f}(\text{cnode}, \text{enode})$$

váha hodnota

- algoritmus vybere vždy ten pár s nejvyšším skóre a příslušné uzly spojí
- to opakuje, dokud je skóre větší než daná prahová hodnota

Rysy (features)

- vlastnosti dvojice českého a anglického t-uzlu
- celkem 15 rysů
- jejich váhy byly optimalizovány pomocí perceptronu
- příklady rysů:
 - Pravděpodobnost překladu dvojice t-lemmat (z pravděpodobnostního slovníku) (hodnota 0 až 1)
 - Tato dvojice rovněž byla/nebyla spárována nástrojem GIZA++ (0 nebo 1)
 - T-lemmata jsou shodná (0 nebo 1)
 - Shoda t-lemmat v prvních třech písmenech (0, 1)

Rysy (features) (2)

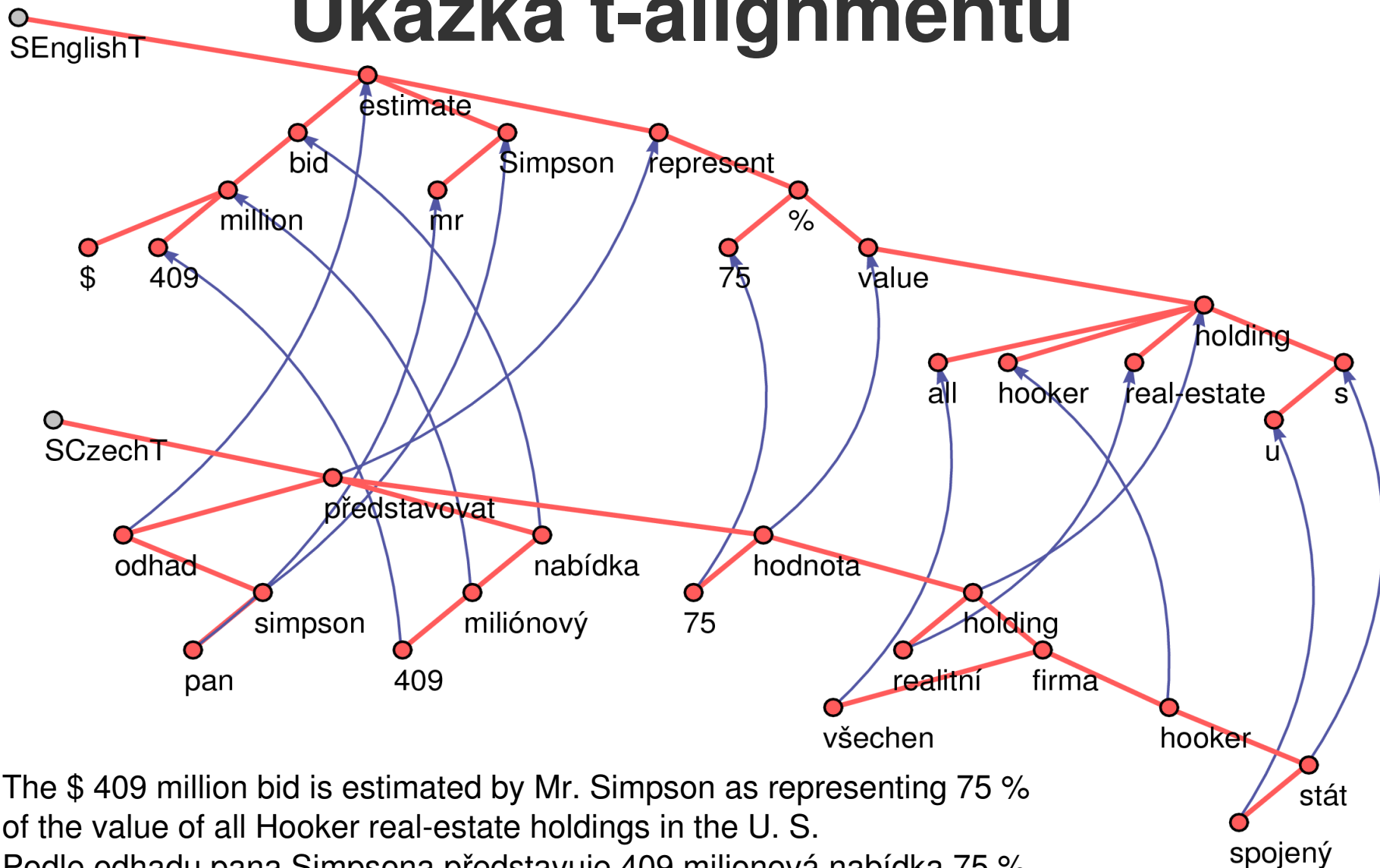
- Rodičovské uzly obou zkoumaných uzlů jsou již spárovány (hodnoty 0, 1)
- Někteří potomci zkoumaných uzlů jsou již spárováni (hodnoty 0, 1, 2, 3 ...)
- Shodný sémantický slovní druh (hodnoty 0, 1)
- Podobná pozice uzlů ve stromě (real 0 až 1)
 - Rozdíl relativních pozic českého a anglického uzlu

Fáze 2 – Completing 1:N relations

Spojíme navíc ty dvojice uzlů (K, L) , pro které jsou splněny následující podmínky:

- K ještě není spojen s žádným protějším uzlem, ale jeho synovský nebo rodičovský uzel je spojen s L .
- Dvojice (K, L) byla navíc spojena systémem GIZA++.
- Dvojice (K, L) se vyskytuje v pravděpodobnostním slovníku.

Ukázka t-alignmentu



The \$ 409 million bid is estimated by Mr. Simpson as representing 75 % of the value of all Hooker real-estate holdings in the U. S.

Podle odhadu pana Simpsona představuje 409 milionová nabídka 75 % hodnoty všech realitních holdingů firmy Hooker ve Spojených státech.

Vyhodnocení úspěšnosti

- pro porovnání výsledků využita pouze spojení typu „sure“.
- 10-násobná cross-validace
- 9/10 dat pro natrénování optimálních vah jednotlivých rysů (features) perceptronem
- 1/10 dat pro samotnou evaluaci

T-aligner – evaluace

typ dat	precision	recall	f-measure
právní texty	92,8 %	91,4 %	92,1 %
beletrie	86,3 %	82,7 %	84,4 %
novinové články	90,7 %	91,9 %	91,3 %
pojmenované entity	94,8 %	92,8 %	93,8 %
PCEDT	94,9 %	88,7 %	91,7 %
Dohromady	92,5 %	89,6 %	91,0 %

GIZA++ na m-lemmatech

- GIZA++ byla spuštěna na sekvence m-lemmat.
- Průniková symetrizace
- Výsledný word-alignment byl pak převeden na t-alignment pomocí odkazů **lex.rf**.

precision	recall	f-measure
95,5 %	77,8 %	85,7 %

GIZA++ na t-lemmatech

- Z tektogramatických stromů byla vyextrahována t-lemmata a seřazena podle jejich pozice ve stromě.
- Každý uzel reprezentuje jedno t-lemma.
- Na takovéto sekvence t-lemmat byla spuštěna GIZA++ s průnikovou symetrizací.

precision	recall	f-measure
93,1 %	75,9 %	83,6 %

Závěr

- zlepšení f-measure o 5% oproti baseline (GIZA++)

	f-measure
GIZA++ t-alignment	85,7 %
feature based t-aligner	91,0 %
mezianotátorská shoda	94,8 %

Děkuji za pozornost.