Využití korpusu InterCorp při vytváření ručních pravidel pro automatickou detekci pleonastického *it* a jeho českých ekvivalentů v závislostních datech

Kateřina Veselovská
ÚFAL MFF UK
veselovska@ufal.mff.cuni.cz

Workshop InterCorp
6. září 2013

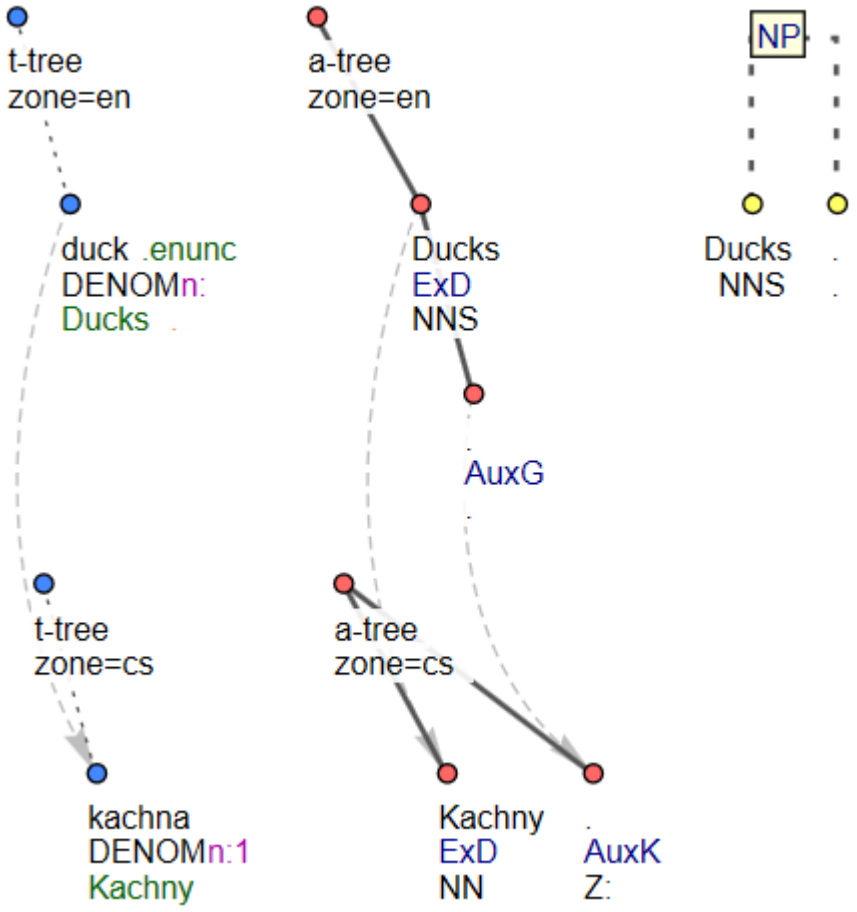
Paralelní závislostní data

- zatím pouze cs – en
- Prague Czech-English Dependency Treebank 2.0 (PCEDT)
- 2312 paralelních česko-anglických dokumentů, přes 1, 2 mil. tokenů / 49 208 vět v každé paralele
- Penn Treebank – Wall Street Journal

Paralelní závislostní data

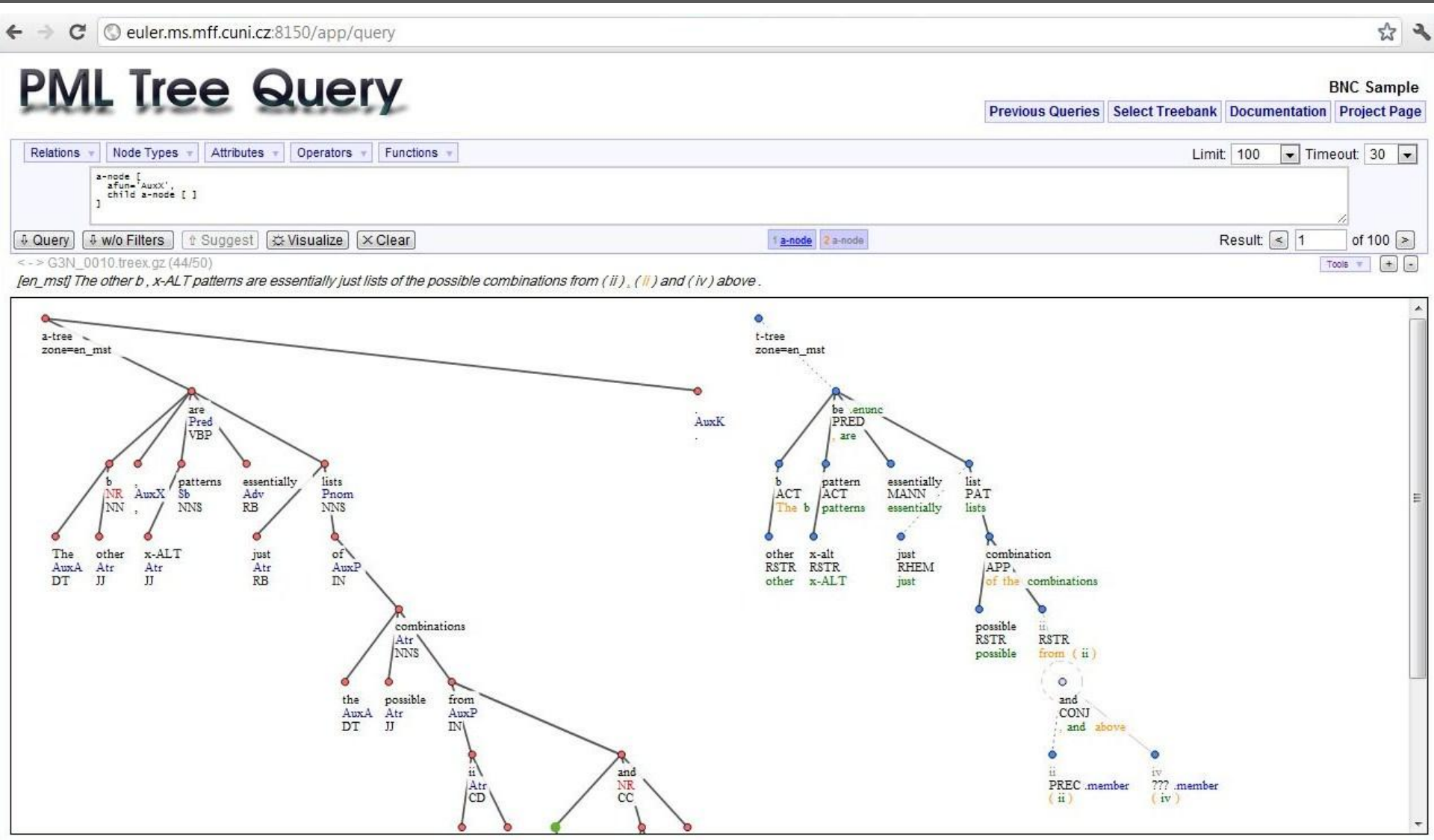
- manuální parsing
- anglické *it* a zrekonstruované chybějící uzly pro *to/ono* v češtině v první fázi anotovány ručně
- hloubková syntax – sémantické vztahy
- automatické propojení uzlů
- praktická aplikace ve strojovém překladu

Paralelní závislostní data

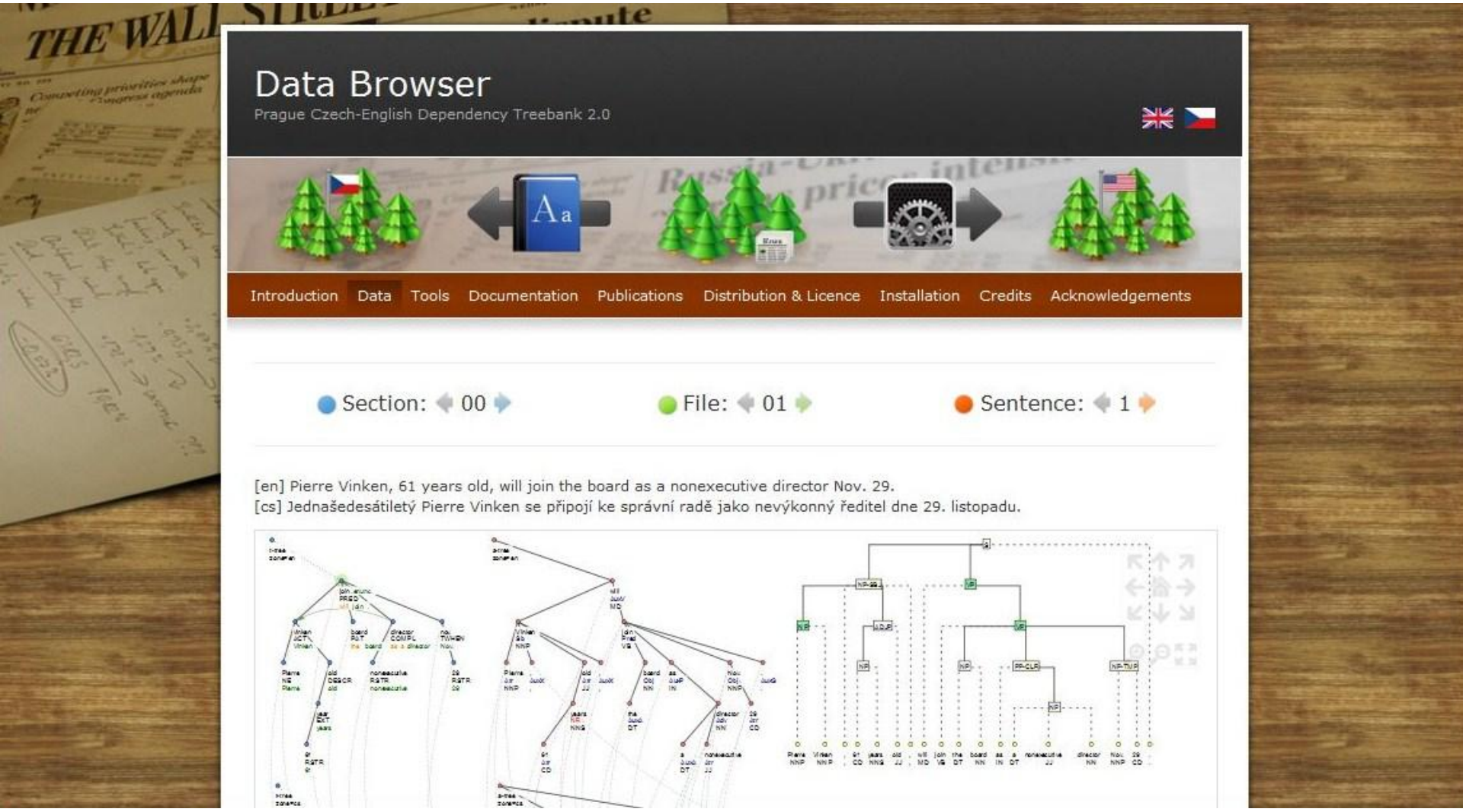


http://ufal.mff.cuni.cz/pcedt2.0/

Paralelní závislostní data



Paralelní závislostní data



Cíle projektu (Veselovská et al., 2012)

- zlepšení anotačního schématu PCEDT 2.0:
  jak nakládat se zájmenem *it*?
- experimenty s automatickou identifikací *it*
- pravidlový klasifikátor pro každou stranu korpusu
- kontrola správnosti zarovnání protějšků

Cíle projektu (Veselovská et al. 2012)

- zlepšení anotačního schématu PCEDT 2.0:
  jak nakládat se zájmenem *it*?
- experimenty s automatickou identifikací *it*
- pravidlový klasifikátor pro každou stranu korpusu
- kontrola správnosti zarovnání protějšků

Angličtina (Quirk et al. 1985, Sinclair 1995 atd.)

a) anaforické: *I bought a new hat but my husband did not like it.*

b) anticipační: *It is no good bothering about it.*

c) deiktické: *Is it your suitcase (over there)?*

d) exklamativní: *"It' s me, open the* door!"

e) prop: *It is 5 o'clock.*

Čeština: nevyjádřený subjekt (Nguy a Ševčíková 2011)

a) implicitní subjekt — *Slunce zezlátlo. (Ono) pomalu zašlo za obzor.*

b) všeobecný subjekt — *S rizikem se počítá.*

c) blíže nespecifikovaný subjekt — *Hlásili to v rádiu.*

d) nulový subjekt — *Zítra bude oblačno.*

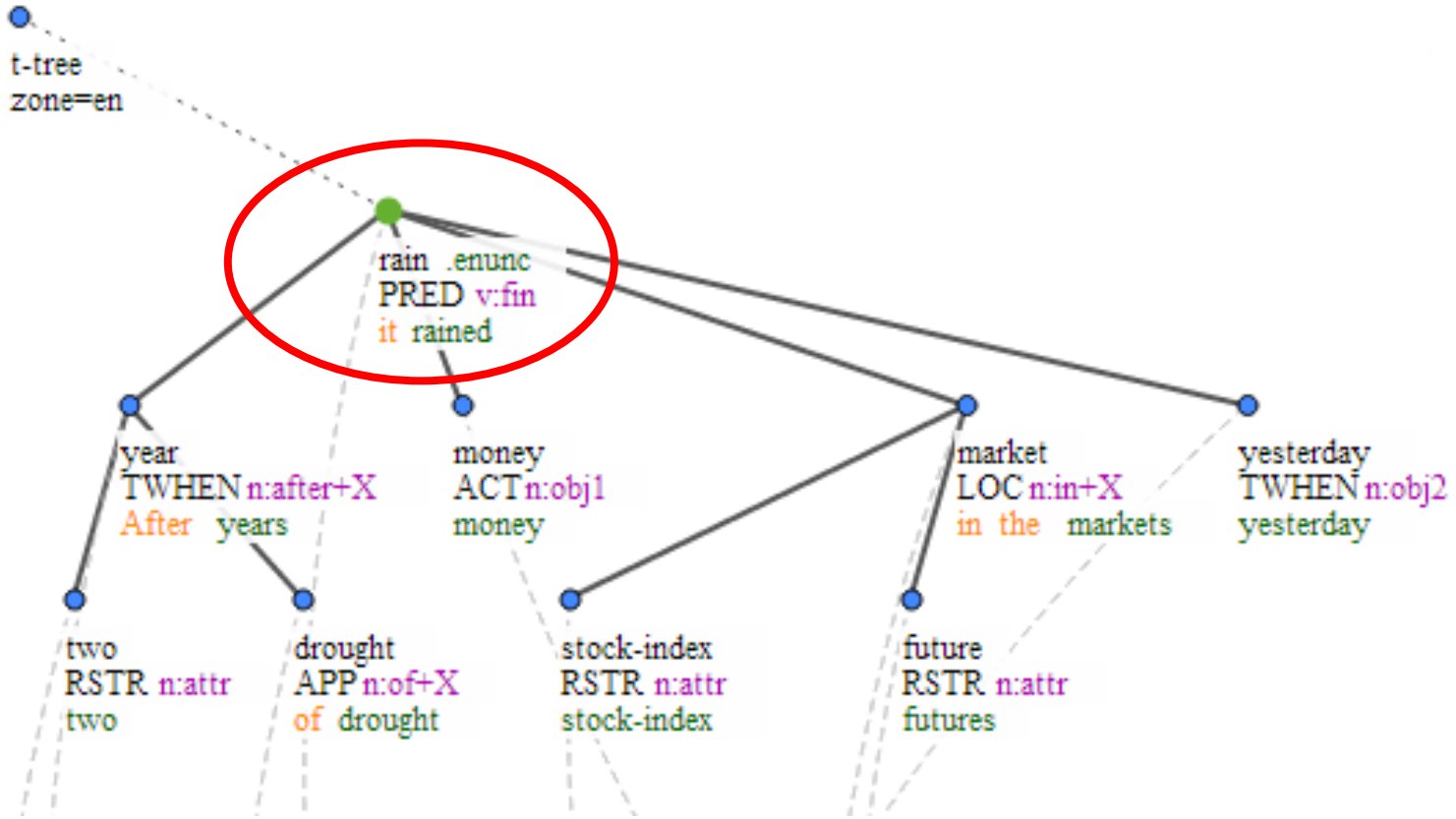
Zjednodušení úlohy

Anaforické: anaforické a anticipační v angl. + nevyjádřená 3.os sg. t-lemma #Perspron (a-lemma it), šipka k antecedentu

Neanaforické: deiktické a exklamativní v angl. + nevyjádřená 3. os. sg. , t-lemma #Perspron (a-lemma it), bez šipky

Pleonastické: prop v angl. + nulový subjekt nemá vlastní t-uzel, v češtině #Gen

Pleonastické *it*



Pleonastické *it*

- „prop it“ (Quirk et al., 1985)
- neodkazuje k žádné specifické entitě
- je považováno za sémanticky vyprázdněné
- v jednočlenných větách
- v angličtině tyto věty vyžadují predikát se sponovým slovesem

Pleonastické *it* v InterCorpu

- dotazy na vyhledávání specifických struktur s *it,* omezení i na české protějšky

- na základě vyhledávek ruční pravidla pro klasifikátor (kombinace závislostních vlastností a zjištění z InterCorpu)

Pleonastické *it* v InterCorpu

- časové — *It is 5 o'clock.*
- místní — *It is not far to New York.*
- atmosférické — *It is raining.*

Pleonastické *it* v InterCorpu

- lemma hlavního predikátu je *make* a predikát vedlejší věty je v pozici pacientu
  make it easy/hard
- ve sponě je infinitiv slovesa smyslového vnímání
  It is heard/seen/felt atd.
- identifikace na základě českých protějšků
  je možné/je nutné/jde o…

Pravidla

- pokud je sloveso *mak*e a podřízený subjekt je pacientem, pak je *it* pleonastické)
- pokud nejde o nulový subjekt (součást fráze *zdá se* atp.), přidej vygenerovaný uzel pro osobní zájmeno…
- kombinace obou částí: pokud má paralelní uzel subjekt, musí být na české straně #PersPron

Závěry

- opravy *it*, která visela v PCEDT 2.0 špatně
- využití informace o českém protějšku
- obohacení systému o ruční pravidla
- úpravy manuálu
- vylepšení automatické detekce pleonastického it

Závěry

	NON-ANAPH+PLEO				PLEO			
	A	P	R	F	A	P	R	F
EN: Majority class	70.30	–	–	–	85.75	–	–	–
EN: Rules-gold	83.76	99.31	39.15	56.16	94.67	90.31	68.68	78.03
EN: Rules-autom	76.31	73.24	31.90	44.44	87.54	56.90	51.66	54.16
EN: NADA	83.86	81.10	59.51	68.65	86.19	51.00	78.01	61.68
EN: NADA + Rules-autom	84.44	78.61	65.40	71.40	89.83	71.88	47.06	56.88

=> zvýšení přesnosti automatické detekce pleonastického it v anglické části paralelního závislostního korpusu

Závěry

- na základě správně detekovaného pleonastického *it* úpravy české části korpusu

	ANAPH+NON-ANAPH			
	A	P	R	F
CZ: Majority class	86.58	–	–	–
CZ: Rules-gold	98.79	92.89	98.39	95.56
CZ: Rules-autom	87.68	52.97	73.34	61.51
CZ: Rules-autom+EN	91.08	64.20	75.87	69.55

=> zvýšení přesnosti detekce o 3,5%

Literatura

Hajič, J., Hajičová, E., Panevová J., Sgall, P., Cinková S., Fučíková E., Mikulová M., Pajas P., Popelka J., Semecký, J., Šindlerová, J., Štěpánek J., Toman, J., Urešová, Z., Žabokrtský, Z.: *Prague Czech-English Dependency Treebank 2.0.* Data/software, Institute of Formal and Applied Linguistics, Prague, Czech republic, http://ufal.mff.cuni.cz/pcedt2.0/, 2011.

Quirk, R., Greenbaum, S., Leech, G. And J. Svartvik. *A Comprehensive Grammar of the English Language*. Longman, 1985.

Sinclair, J. M. *English Grammar*. Harper Collins Publisher, UK, 1995.

Veselovská K., Nguy Giang L., Novák M.: Using Czech-English Parallel Corpora in Automatic Identification of It. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012),* European Language Resources Association, İstanbul, Turkey, 2012.