Some Computational Experiments with Czech

Ondřej Bojar
obo@cuni.cz

December 7, 2006

Outline

- Background: Computer Science at Charles University in Prague
  - Student software project: Simulated family house
  - My master's: Picking nice examples
- Properties of Czech, analysis of Czech, available data
- Some of my previous experiments
- PhD research (ongoing): Constructing verb valency frames
- Experiments towards MT
  - This year's JHU summer workshop: Moses
- My task here: tree-based machine translation
- Summary of keywords

Background: Computer Science

Master Study at Charles University culminates with two (separate) tasks:

- Software Project
  Joint work of 3–6 students.
  Should take 1 year, never takes less than 1.5 or 2.
  The goal: experience team work on a large scale project, submit a usable piece of software.

- Master Thesis: Picking nice examples of linguistic phenomena

Our Project: The Ents (2000–2002)

The Goal: A simulation of human-like environment (a family house) with user- and computer-controlled inhabitants (ents).

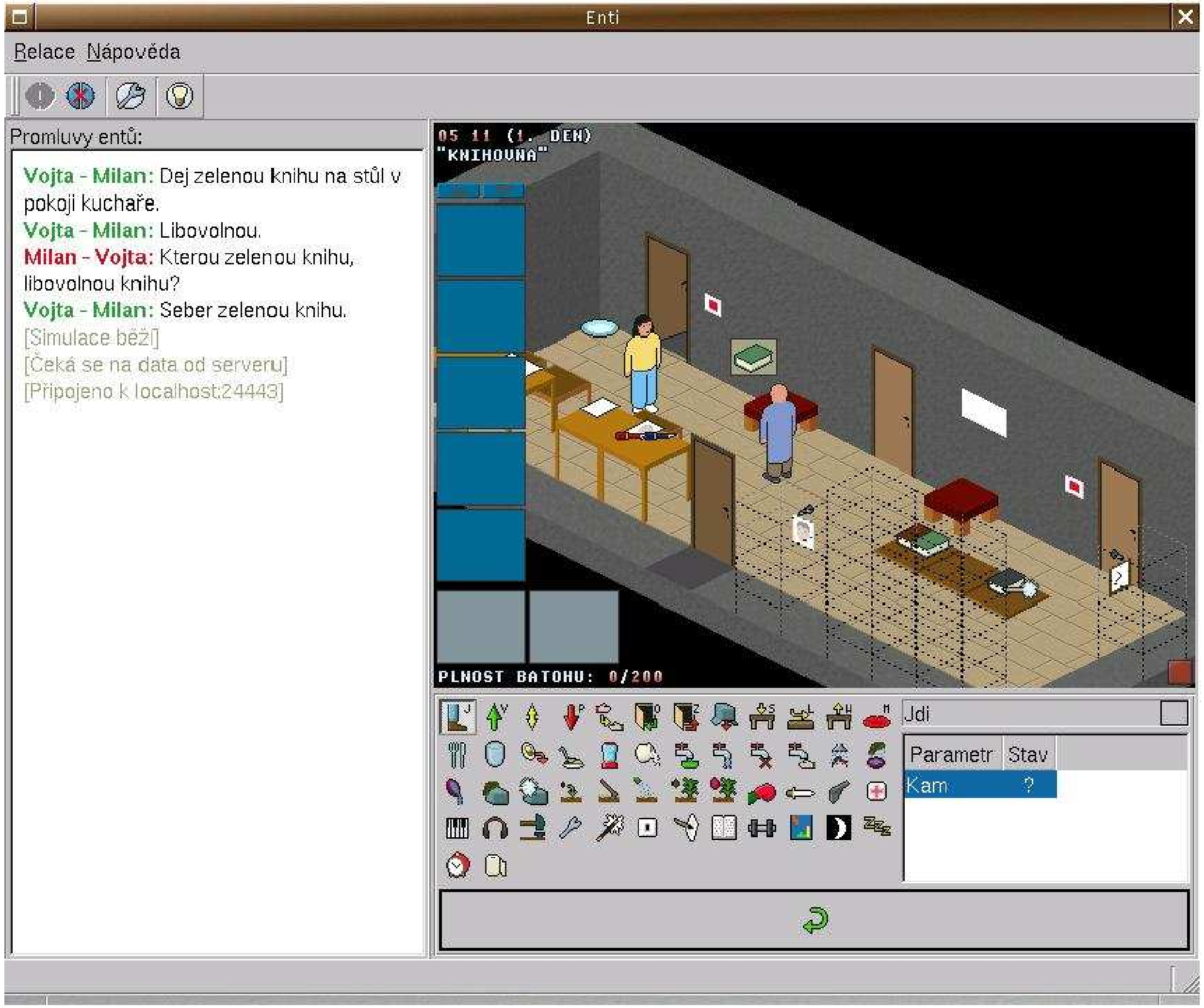
The Result:

- 6 students, 2 years (student style of intensive work)
- a distributed (client-server) unix application
- $>$ 100,000 lines of code in C, C++, Pascal, Mercury, Perl
- 5000 lines of code in a new scripting language E
- 500 pages of documentation in Czech

My contribution: E scripts + NLP module implemented in Mercury:

- understanding definite descriptions of objects in the environment
- concretization – a process of further communication to identify an object uniquely

⇒ ents respond to commands in Czech



Vojta - Milan: Dej zelenou knihu na stůl v pokoji kuchaře.
Vojta - Milan: Libovolnou.
Milan - Vojta: Kterou zelenou knihu, libovolnou knihu?
Vojta - Milan: Seber zelenou knihu.
[Simulace běží]
[Čeká se na data od serveru]
[Připojeno k localhost:24443]

My Master's: Picking Nice Examples (2002/3)

Motivation:

- Accuracy of parsing Czech is limited, especially around the verbs.
- Valency of verbs is (supposedly) crucial for many NLP tasks.

⇒ Goal: Automatically extract nice examples, i.e. sentences easy to parse.

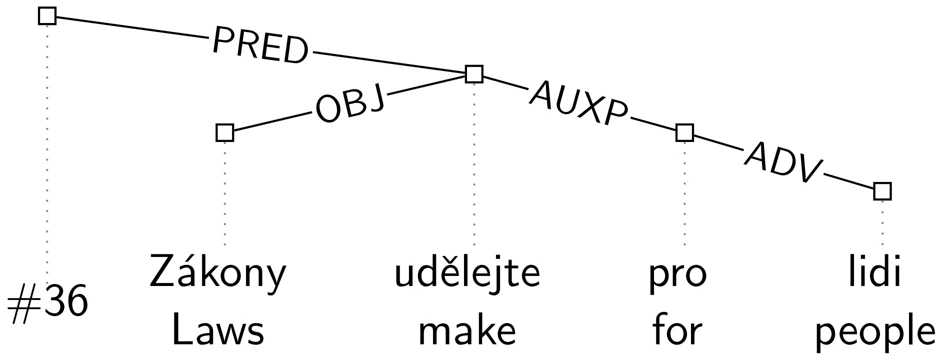
The result:

- a scripting language for partial parsing and filtering sentences

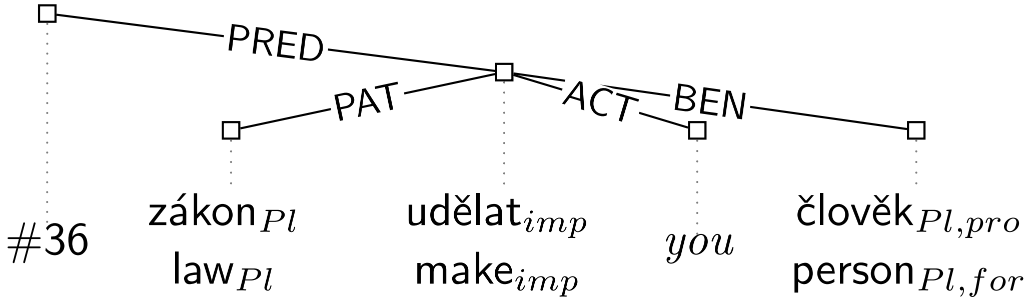
  Engine in Mercury, regular expressions over untyped feature structures.
- a script of 15 filters and 21 rules for Czech:
  - selects 10–15% of sentences
  - improves parsing accuracy by 5–10% absolute (correct dependencies) or 10–15% absolute (correct verb modifications)

Analysis of Czech

Analytic (surface syntactic):



Tectogrammatical (deep syntactic):



Morphological (ambig.):

Form	Lemma	Morphological tag
zákony	zákon	NNIP1-----A----
zákony	zákon	NNIP4-----A----
zákony	zákon	NNIP5-----A----
zákony	zákon	NNIP7-----A----
udělejte	udělat	Vi-P---2--A----
udělejte	udělat	Vi-P---3--A---4
pro	pro-1	RR--4----------
lidi	člověk	NNMP1-----A----
lidi	člověk	NNMP4-----A----
lidi	člověk	NNMP5-----A----

Properties of Czech language

	Czech	English
Rich morphology	$\geq$ 4,000 tags possible, $\geq$ 2,300 seen	50 used
Word order	free	rigid

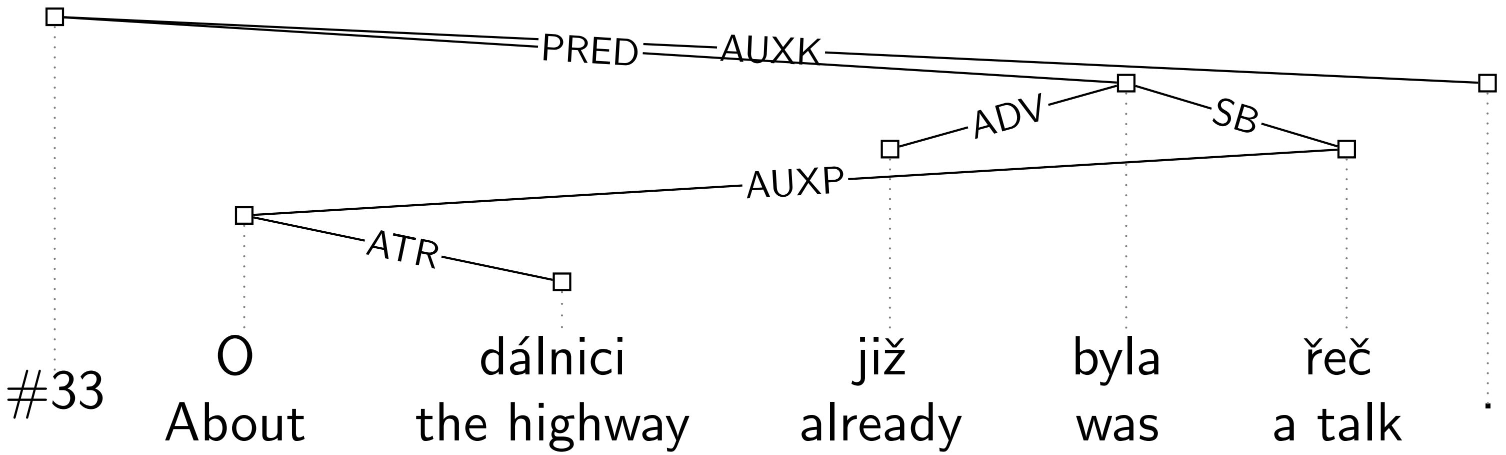
- rigid global word order phenomena: clitics
- rigid local word order phenomena: coordination, clitics mutual order

Nonprojective sentences	16,920	23.3%
Nonprojective edges	23,691	1.9%

Known parsing results	Czech	English
Edge accuracy	69.2–82.5–86%	91%
Sentence correctness	15.0–30.9%	43%

Data by (Collins et al., 1999), (Holan, 2003), Zeman (http://ckl.mff.cuni.cz/˜zeman/ /projekty/neproj/index.html) and (Bojar, 2003). Consult (Kruijff, 2003) for measuring word order freeness.

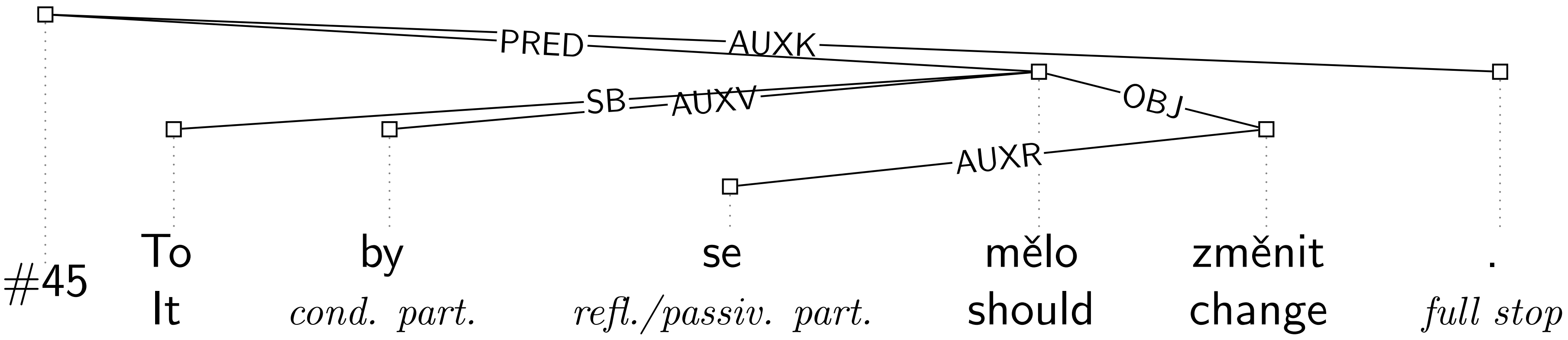
Nonprojectivity



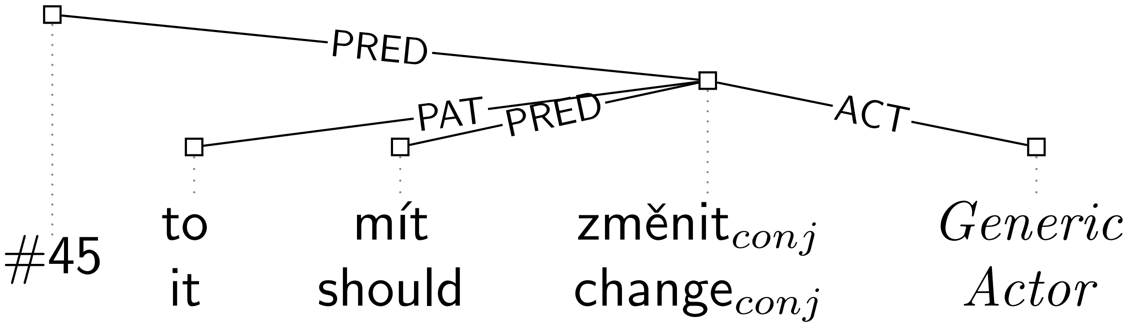
Non-projectivity:

- does not seem to cause delays in reading experiments (Bojar et al., 2004)
- disappears at the deep syntactic level (Veselá, Havelka, and Hajičová, 2004)
- parsing ($O(n^2)$) solved only recently (McDonald et al., 2005)

Analytic vs. Tectogrammatical



#45	To	by	se	mělo	změnit	.
	It	*cond. part.*	*refl./passiv. part.*	should	change	*full stop*

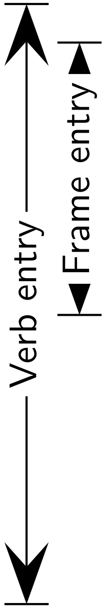


#45	to	mít	změnit$_{conj}$	*Generic*
	it	should	change$_{conj}$	*Actor*

- hide auxiliary words, add nodes for "deleted" participants
- resolve e.g. active/passive voice, analytical verbs etc.
- "full" tecto resolves much more, e.g. topic-focus articulation or anaphora

Czech Verb Valency Lexicon VALLEX

Key components: Frames, functors, obligatoriness, morphemic form(s)



odpovídat (imperfective)

$\boxed{1}$ odpovídat$_1$ $\sim$ odvětit [answer; respond]

- frame: ACT_{1}^{obl} ADDR_{3}^{obl} $\mathrm{PAT}_{na+4,4}^{opt}$ $\mathrm{EFF}_{4,aby,ať,zda,že}^{obl}$ MANN^{typ}
- example: *odpovídal mu na jeho dotaz pravdu / že . . .* [he responded to his question truthfully / that . . .]
- asp.counterpart: odpovědět$_1$ pf.
- class: communication

$\boxed{2}$ odpovídat$_2$ $\sim$ reagovat [react]

- frame: ACT_{1}^{obl} $\mathrm{PAT}_{na+4}^{obl}$ MEANS_{7}^{typ}
- example: *pokožka odpovídala na včelí bodnutí zarudnutím* [the skin reacted to a bee sting by turning red]
- asp.counterpart: odpovědět$_2$ pf.

. . .

odpovídat se (imperfective)

$\boxed{1}$ odpovídat se$_1$ $\sim$ být zodpovědný [be responsible]

- frame: $\mathrm{ACT}_{1}^{obl}\mathrm{ADDR}_{3}^{obl}\mathrm{PAT}_{z+2}^{obl}$
- example: *odpovídá se ze ztrát* [he answers for the losses]

An abbreviated example for the base lemma "odpovídat".

Available Czech Data (not exhaustive!)

Monolingual Corpora

Name and version	Sents.	Tokens	Annotation
Czech National Corpus (SYN2000d)	6.8M	114M	automatic lemmas+tags
Prague Dep Tbk (PDT 2.0)	50k–115k	0.8M–2.0M	manual tecto–manual morph

Parallel Czech-English

Name and version	Sents.	Tokens	Annotation
Prague Cz-En Dep Tbk (PCEDT 1.0)	22k/49k	0.5M/1.2M	automatic tecto trees
CzEng 0.5	1.4M/1.2M	19M/21M	automatic sent. ali, tokenized

Dictionaries

VALLEX 1.5	verbs: 2.4k entries (1.8k lemmas); covers 6% of types, 65% of tokens
PDT-VALLEX	verbs, nouns, adjs: part of PDT 2.0, only items occurring in PDT 2.0
ENG-VALLEX	PropBank→VALLEX-like, for PCEDT 2.0
BEAST	an ugly compilation of web dictionaries (400k pairs, 235k cs, 225k en entries)

Some of My Recent Experiments (2003–2005)

Constraint-based parsing of Czech didn't work out (Bojar, 2004):

- XDG (Debusmann, 2006), constr.-based dep. parser implemented in Mozart-Oz
- Local constraints on tree structure induced from a treebank were too weak ⇒ exponentially many analyses remained possible (though not correct).
- Disregarding probabilities *is* harmful.

Inter-annotator agreement of verb-frame disam. (Lopatková et al., 2005):

- Allowed to check quality of VALLEX.
- Results comparable with others (PropBank etc.), best for Czech so far.
  Better than e.g. agreement of Czech WordNet annotation.

PhD. studies: Constructing Verb Valency Frames

Motivation:

- VALLEX development time-consuming, entries very complex.
- 93% of verb types make only 10% of verb tokens ⇒ human labour hardly justifiable.

Necessary steps given a verb lemma:

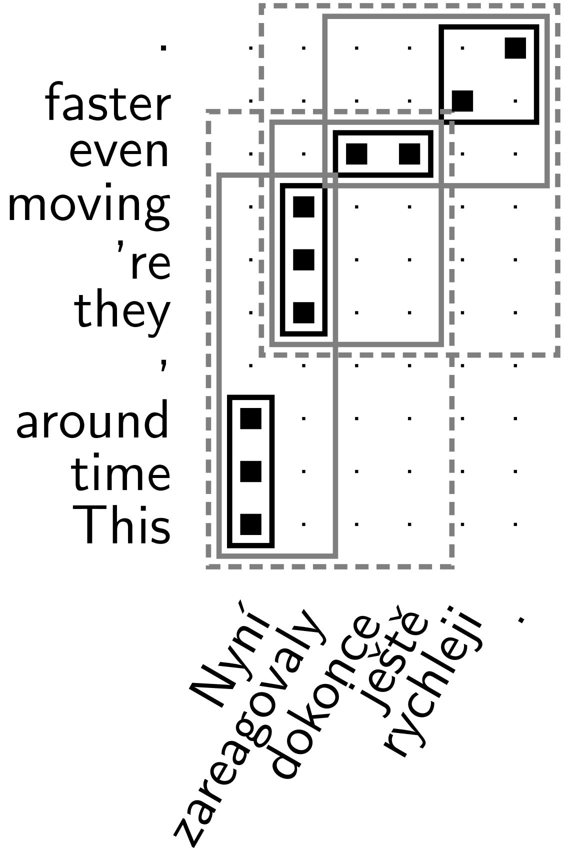
- Find (nice) examples of verbs usage.
- Classify verb occurrences wrt. to reflexivity.
- Cluster (not classify) verb+refl occs into groups with the same (hidden) frame.
- Derive frame description from the set of grouped examples:
  - Cluster/classify verb modifications into groups with the same (hidden) functor.
  - Decide obligatoriness for all observed functors.

Metric: Verb Entry Similarity (Benešová and Bojar, 2006)
∼ Edit distance necessary to convert suggested frames to golden frames.

Experiments Towards Machine Translation

- Augmenting machine-readable dicts. with syntactic information (Bojar, 2005)
- (Rather unsuccessful) attempts at reusing an old rule-based MT system (Bojar, Homola, and Kuboň, 2005)
- Preliminary experiments with extracting parallel verb frames (Bojar and Hajič, 2005)
- Experiments with Czech-English word alignment (Bojar and Prokopová, 2006)
  ⇒where GIZA++ fails, humans often (38% of tokens) disagree as well

Alignments, Phrases and Phrase-Based MT



This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
...	=	...
This time around, they 're moving	=	Nyní zareagovaly
even faster	=	dokonce ještě rychleji
...	=	...

Phrase-based MT: choose such segmentation of input string and such phrase "replacements" to make the output sequence "coherent" (3-grams most probable).

My Phrase-Based Cs→En MT Impressions

lemmatization for alignment	+2.0*
handling numbers	+0.9*
fixing clear BLEU errors	+0.5
dependency-based corpus expansion	+0.3
more out-of-domain parallel texts, also in LM	+0.4
bigged in-domain LM	+1.7*
more out-of-domain parallel texts, bigger in-domain LM	+5.0*

Given BLEU as “the” MT metric:

- Phrase-based system from Czech better than expected (BLEU up to 37%)
  (But the setting was easy, the MT was translating *back* to English.)
- With small data (20k s), focus on alignments, corpus specifics and clear errors.
- With more data (20k+80k s), in-domain language model is vital.

The asterisk (*) denotes stat. signif. More details in (Bojar, Matusov, and Ney, 2006).

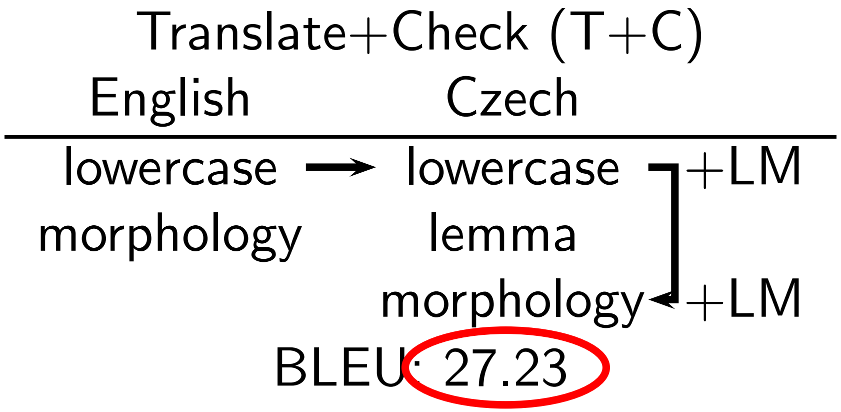
“Moses”

Summer 2006: MT workshop at JHU: En→Cs

Motivation: (phrase-based) MT to morphologically rich languages performs worse.

Room for improvement: En→Cs baseline BLEU 25%, BLEU disregarding word forms 33%.

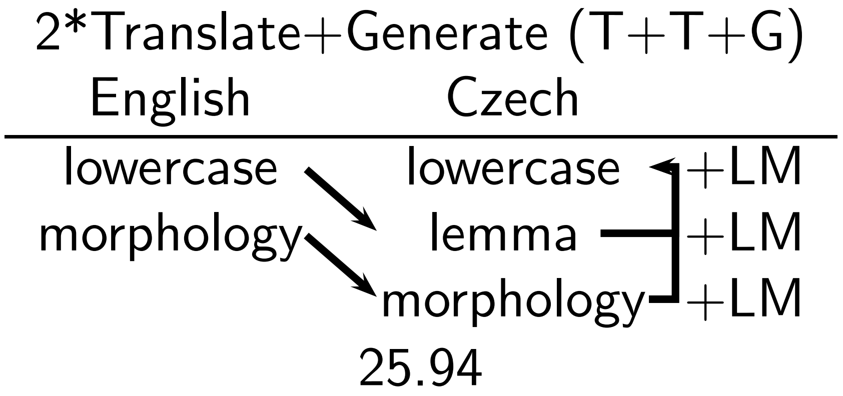
⇒ Keep track of morphology (or other “hidden variables”) explicitly.



Translate+Check (T+C)

English	Czech	
lowercase →	lowercase	+LM
morphology	lemma	
	morphology	+LM

BLEU: 27.23



2*Translate+Generate (T+T+G)

English	Czech	
lowercase	lowercase	+LM
morphology	lemma	+LM
	morphology	+LM

25.94

- The simplest factored model (T+C) improves MT to Czech, German, Spanish.
- MT output locally coherent, but sentence as a whole usually garbled.

  E.g. verbs often missing (21%) or mis-translated (14%).

My Current Main Topic: Tree-based MT

Syntax-based MT becomes fashionable, various approaches possible.
See (Čmejrek, 2006) for a partial survey.

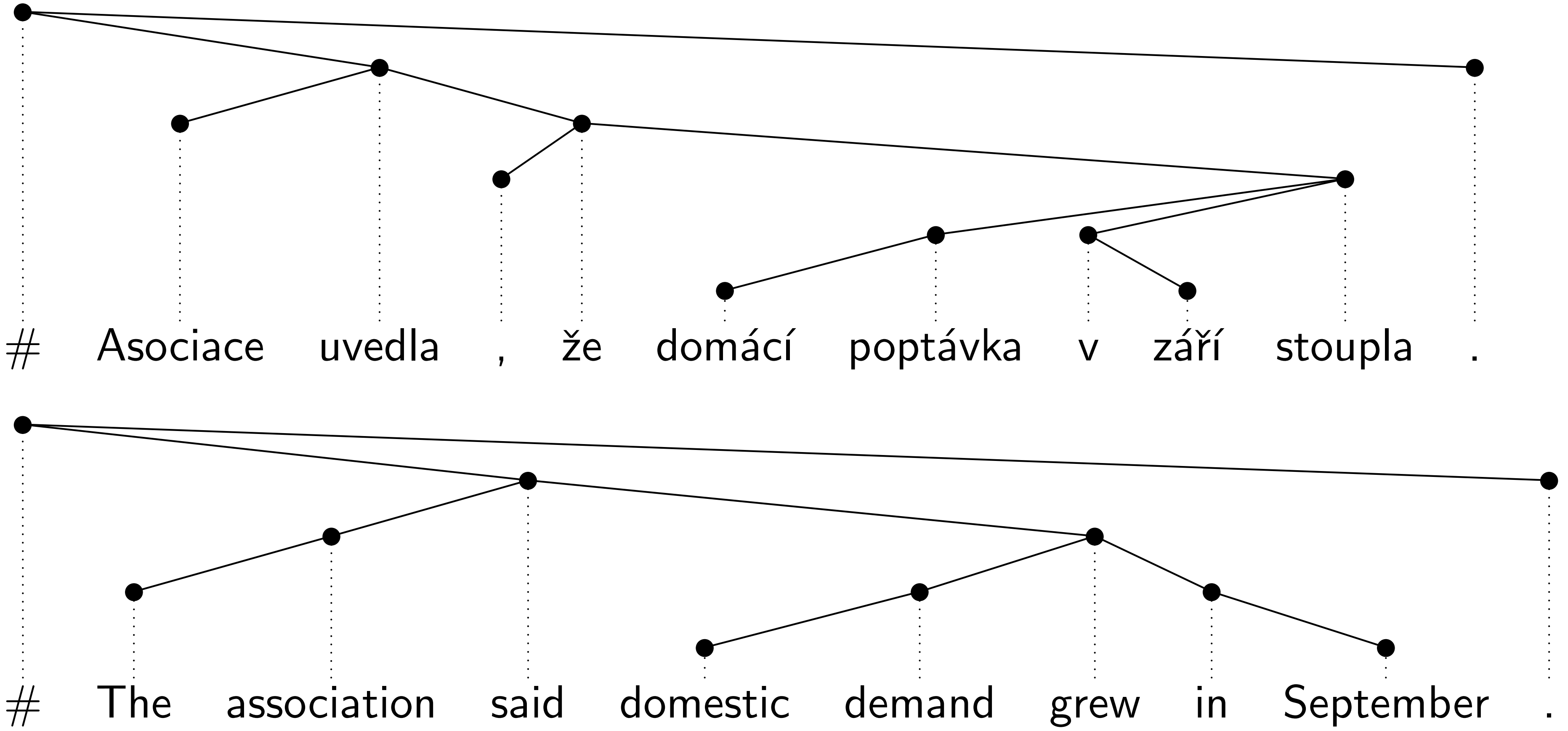
Synchronous Tree Substitution Grammar (Čmejrek, 2006):

- training (treelet alignment) implemented by Martin Čmejrek.
- decoding (search for translation) given a source tree needed.

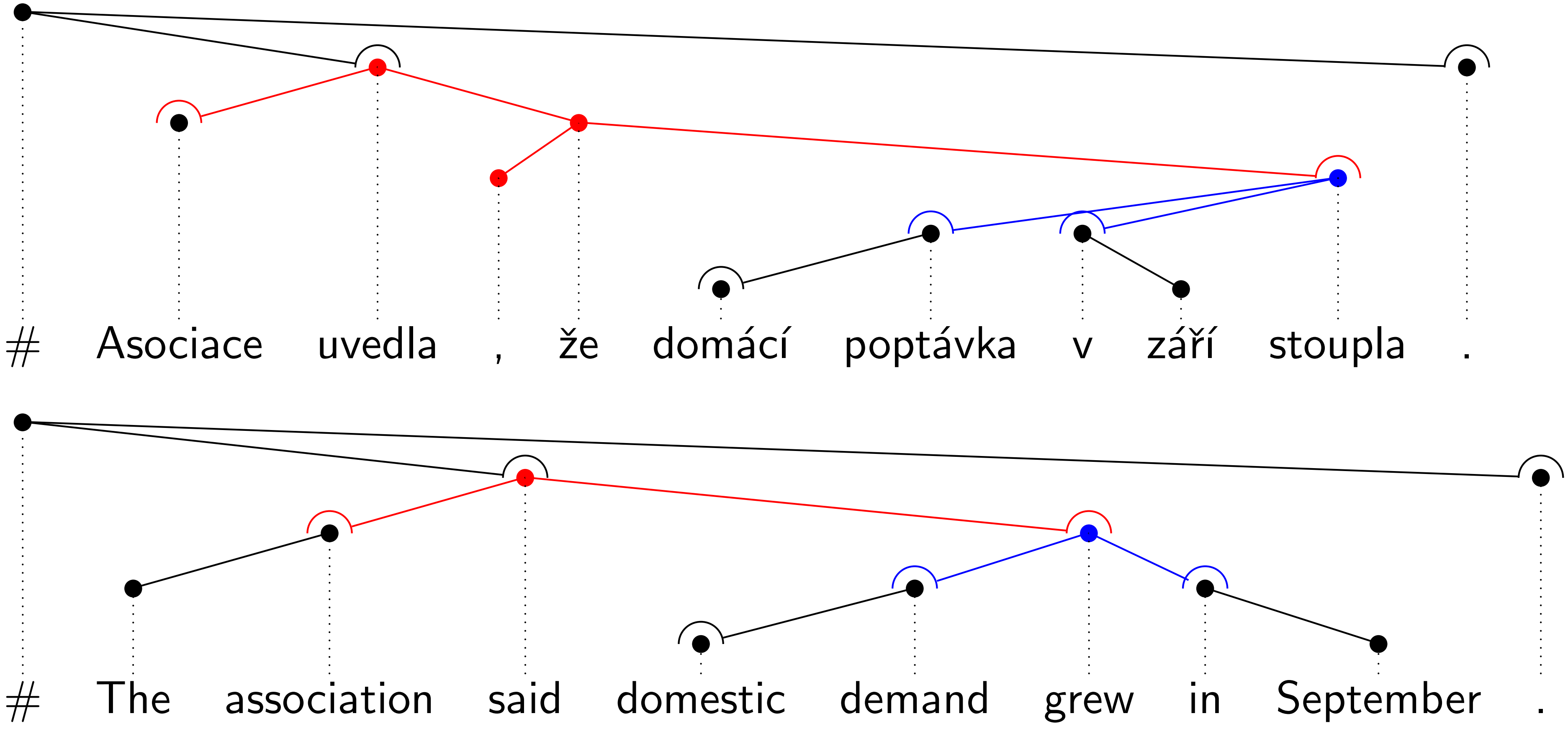
Model generic enough to allow various scenarios:

- Czech analytical → English analytical
- Czech tecto → English tecto (tecto-trees are much more similar!)
- Czech tecto → English analytical

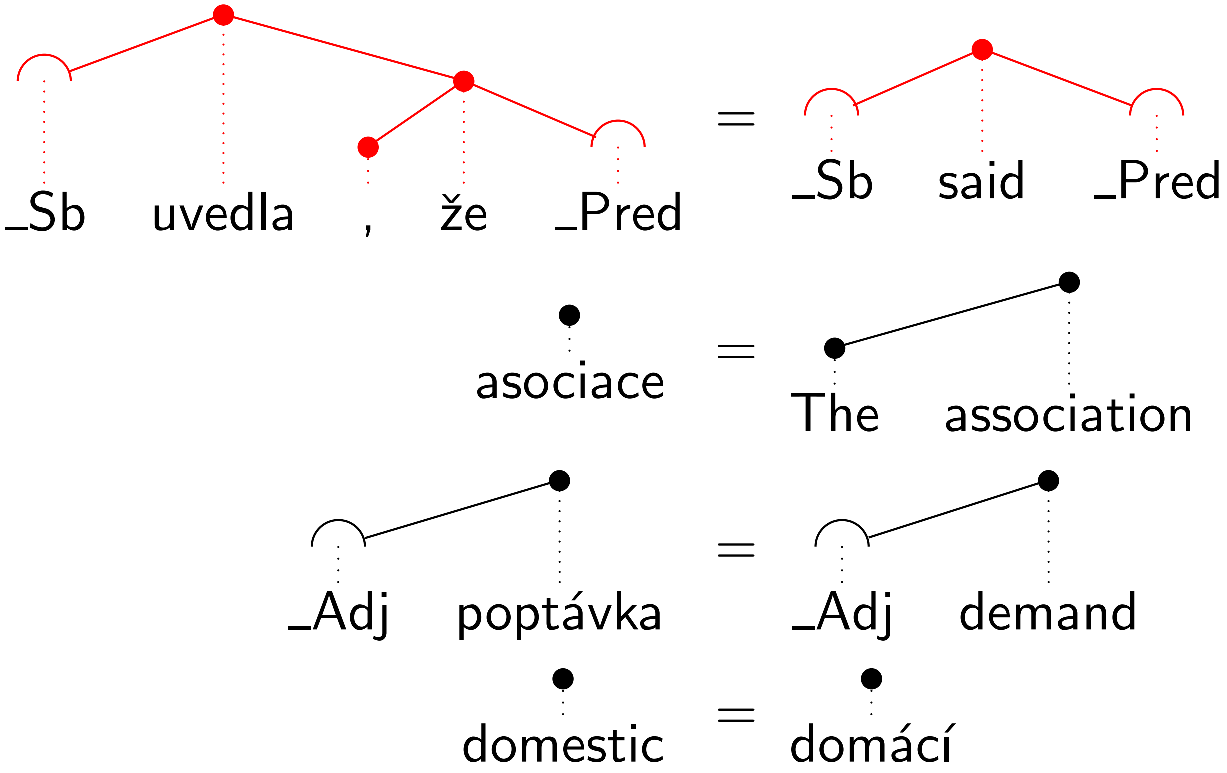
Training: Observe a Pair of Dependency Trees



Training: Decompose Trees into Treelets

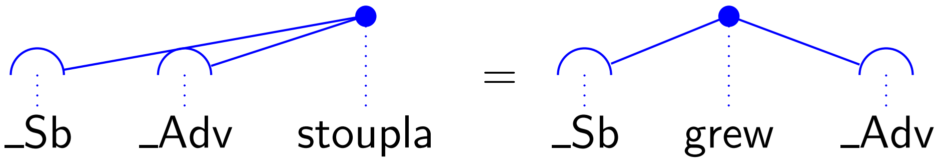


Training: Collect Dictionary of Treelet Pairs

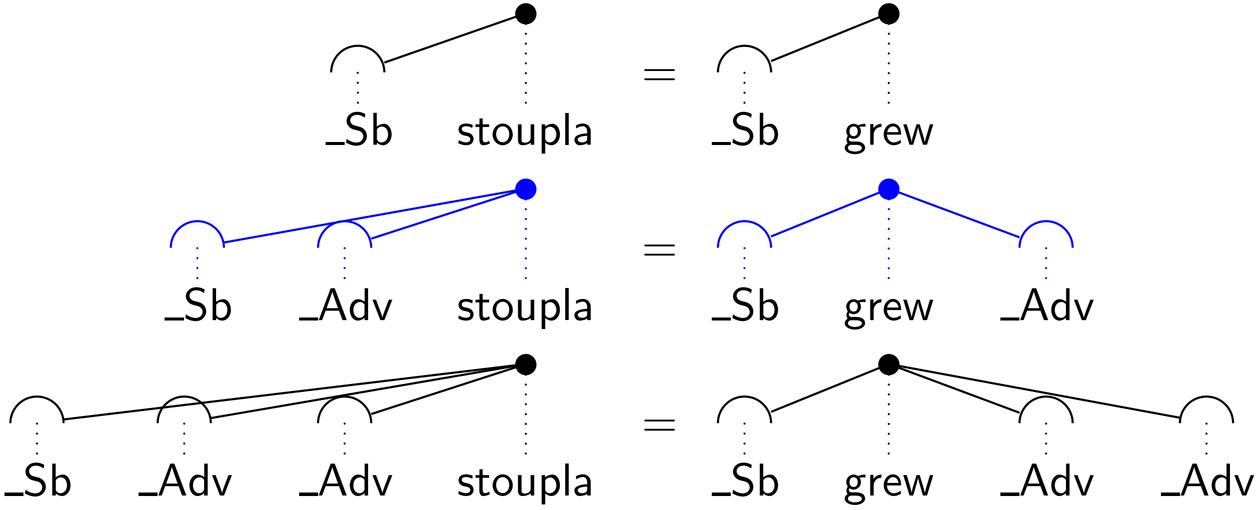


Training: Collect Dictionary of Treelet Pairs (2)

Treelets can be used to encode reordering (or we may force canonic ordering):


_Sb _Adv stoupla = _Sb grew _Adv

But are prone to sparse data problem (they explicitly encode the number of the sons):


_Sb stoupla = _Sb grew

_Sb _Adv stoupla = _Sb grew _Adv

_Sb _Adv _Adv stoupla = _Sb grew _Adv _Adv

Decoding STSG

Given an input dependency tree:

- decompose it into known treelets,
- replace treelets by their translations,
- join output treelets and produce output final tree (or string).

Decoder design:

- beam-search similar to Moses,
- top-down output generation (not left-to-right),
- built-in support for plain string language model (MT is scored by BLEU).

Current main concern:

- combining various back-off schemes correctly
  (Looking for someone experienced to help me.)

Summary of Keywords

Keywords describing my research:

- Czech, Czech-English MT
- syntactic analysis, machine translation
- extraction of (parallel) syntactic information about words; dictionaries

Keywords important for Prague (as far as I know):

- deep syntax, tectogrammatical layer
- valency, information structure (topic-focus articulation, coreference)
- PDT, PCEDT, PADT (Arabic!), TrEd (tree editor)

Important links:

- PDT 2.0 and a tutorial: `http://ufal.mff.cuni.cz/pdt.html`
- Moses decoder: `http://www.statmt.org/moses/`

References

Benešová, Václava and Ondřej Bojar. 2006. Czech Verbs of Communication and the Extraction of their Frames. In *Text, Speech and Dialogue: 9th International Conference, TSD 2006*, volume LNAI 3658, pages 29–36. Springer Verlag, September.

Bojar, Ondřej. 2003. Towards Automatic Extraction of Verb Frames. *Prague Bulletin of Mathematical Linguistics*, 79–80:101–120.

Bojar, Ondřej. 2004. Problems of Inducing Large Coverage Constraint-Based Dependency Grammar for Czech. In *Constraint Solving and Language Processing, CSLP 2004*, volume LNAI 3438, pages 90–103, Roskilde University, September. Springer.

Bojar, Ondřej. 2005. Budování česko-anglického slovníku pro strojový překlad. In Peter Vojtáš, editor, *ITAT 2005 Information Technologies – Applications and Theory*, pages 201–211, Košice, Slovakia, September. University of P. J. Šafařík.

Bojar, Ondřej and Jan Hajič. 2005. Extracting Translation Verb Frames. In Walther von Hahn, John Hutchins, and Christina Vertan, editors, *Proceedings of Modern Approaches in Translation Technologies, workshop in conjunction with Recent Advances in Natural Language Processing (RANLP 2005)*, pages 2–6. Bulgarian Academy of Sciencies, September.

Bojar, Ondřej, Petr Homola, and Vladislav Kuboň. 2005. Problems Of Reusing An Existing MT System. In *IJCNLP*

2005 - Companion Volume to the Proceedings of Conference including Posters/Demos and Tutorial Abstracts, pages 181–186, October.

Bojar, Ondřej, Evgeny Matusov, and Hermann Ney. 2006. Czech-English Phrase-Based Machine Translation. In *FinTAL 2006*, volume LNAI 4139, pages 214–224, Turku, Finland, August. Springer.

Bojar, Ondřej and Magdalena Prokopová. 2006. Czech-English Word Alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1236–1239. ELRA.

Bojar, Ondřej, Jiří Semecký, Shravan Vasishth, and Ivana Kruijff-Korbayová. 2004. Processing noncanonical word order in Czech. In *Proceedings of Architectures and Mechanisms for Language Processing, AMLaP 2004*, pages 91–91, Université de Provence, September 16-18.

Čmejrek, Martin. 2006. *Using Dependency Tree Structure for Czech-English Machine Translation*. Ph.D. thesis, ÚFAL, MFF UK, Prague, Czech Republic.

Collins, Michael. 1996. A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 184–191.

Collins, Michael, Jan Hajič, Eric Brill, Lance Ramshaw, and Christoph Tillmann. 1999. A Statistical Parser of Czech. In *Proceedings of 37th ACL Conference*, pages 505–512, University of Maryland, College Park, USA.

Debusmann, Ralph. 2006. *Extensible Dependency Grammar: A Modular Grammar Formalism Based On Multigraph Description*. Ph.D. thesis, Saarland University, 4.

Holan, Tomáš. 2003. K syntaktické analýze českých(!) vět. In *MIS 2003*. MATFYZPRESS, January 18–25, 2003.

Kruijff, Geert-Jan M. 2003. 3-Phase Grammar Learning. In *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Development*.

Lopatková, Markéta, Ondřej Bojar, Jiří Semecký, Václava Benešová, and Zdeněk Žabokrtský. 2005. Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation. In Václav Matoušek, Pavel Mautner, and Tomáš Pavelka, editors, *Text, Speech and Dialogue: 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12-15, 2005. Proceedings*, volume LNAI 3658, pages 99–106. Springer Verlag, September.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of HLT/EMNLP 2005*, October.

Veselá, Kateřina, Jiří Havelka, and Eva Hajičová. 2004. Condition of Projectivity in the Underlying Dependency Structures. In *Proceedings of Coling 2004*, pages 289–295, Geneva, Switzerland, August. COLING.

Detailed Numbers on Non-Projectivity

Edge length	1	≤ 2	≤ 5
English [%]	74.2	86.3	95.6
Czech [%]	51.8	72.1	90.2

[1]

Number of gaps	0	1	2
Sentences [%]	76.9	22.7	0.42

[2]

Climbing steps	1	2	3	4	5
Nodes [%]	90.3	8.0	1.3	0.3	0.1

[3]

[1] Data for English by (Collins, 1996). Data for Czech by (Holan, 2003).

[2] Data by (Holan, 2003).

[3] Data by (Holan, 2003).

Data Sparseness

After having seen	20,000	75,000	sentences
a new lemma comes every	1.6	1.8	test sentences
a new full morphological tag comes every	110	290	test sentences
a new simplified tag comes every	280	870	test sentences

Simplified morphological tag = POS, SUBPOS, CASE, NUMBER and GENDER.

Where GIZA Fails, Humans Have Troubles, Too

Percentage of running words where the alignment matches (Ok) or mismatches (With Problems):

- Humans against each other
- GIZA++ againts golden set derived by joining the human annotations

		Baseline		Improved	
Humans	GIZA++	en	cs	en	cs
With Problems	With Problems	14.3	15.5	14.3	15.5
With Problems	OK	0.1	0.1	0.2	0.1
OK	With Problems	38.6	35.7	25.2	25.0
OK	OK	46.9	48.7	60.4	59.4

Sample Cs→En Phrase-Based MT Output

System Output:

We 'll see whether the campaigns work .

Immediately after Friday 's 190 14-point stock market and a consequent uncertainty excretes several big brokerage firms new ads UNKNOWN_vytrubující usual message : Go on in investing , the market is in order .

Their business is persuade clients from escaping from the market , which individual investors masse fact , after plunging in October .

Source:

Uvidíme , zda reklama funguje .

Okamžitě po pátečním 190 bodovém propadu akciového trhu a následné nejistotě vypouští několik velkých brokerských firem nové inzeráty vytrubující obvyklé poselství : Pokračujte v investování , trh je v pořádku .

Jejich úkolem je odradit klienty od útěku z trhu , což jednotliví investoři hromadně činili po propadu v říjnu .