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

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
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.):

<table>
<thead>
<tr>
<th>Form</th>
<th>Lemma</th>
<th>Morphological tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>zákony</td>
<td>zákony</td>
<td>NNIP1-----A----</td>
</tr>
<tr>
<td>zákony</td>
<td>zákony</td>
<td>NNIP4-----A----</td>
</tr>
<tr>
<td>zákony</td>
<td>zákony</td>
<td>NNIP5-----A----</td>
</tr>
<tr>
<td>zákony</td>
<td>zákony</td>
<td>NNIP7-----A----</td>
</tr>
<tr>
<td>udělejte</td>
<td>udělat</td>
<td>Vi-P---2--A----</td>
</tr>
<tr>
<td>udělejte</td>
<td>udělat</td>
<td>Vi-P---3--A---4</td>
</tr>
<tr>
<td>pro</td>
<td>pro-1</td>
<td>RR--4----------</td>
</tr>
<tr>
<td>lidi</td>
<td>člověk</td>
<td>NNMP1-----A----</td>
</tr>
<tr>
<td>lidi</td>
<td>člověk</td>
<td>NNMP4-----A----</td>
</tr>
<tr>
<td>lidi</td>
<td>člověk</td>
<td>NNMP5-----A----</td>
</tr>
</tbody>
</table>
# Properties of Czech language

<table>
<thead>
<tr>
<th></th>
<th>Czech</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rich morphology</td>
<td>$\geq 4,000$ tags possible, $\geq 2,300$ seen</td>
<td>50 used</td>
</tr>
<tr>
<td>Word order</td>
<td>free</td>
<td>rigid</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Czech</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonprojective sentences</td>
<td>16,920</td>
<td>23.3%</td>
</tr>
<tr>
<td>Nonprojective edges</td>
<td>23,691</td>
<td>1.9%</td>
</tr>
</tbody>
</table>

**Known parsing results**

<table>
<thead>
<tr>
<th></th>
<th>Czech</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge accuracy</td>
<td>69.2–82.5–86%</td>
<td>91%</td>
</tr>
<tr>
<td>Sentence correctness</td>
<td>15.0–30.9%</td>
<td>43%</td>
</tr>
</tbody>
</table>

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:
- 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 it by cond. part. se refl./passiv. part. mělo should změnit change full stop

#45 to mít změnit<sub>conj</sub> Generic 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)

**odpovídá**t (imperfective)

1. odpovídá1 ~ odvětit [answer; respond]
   - frame: \(ACT_1^{obl} \ ADRR_3^{obl} \ PAT_{na+4,4}^{opt} \ EFF_4^{obl} \ aby,at,zda,ze \ MANN_{typ}^{typ}\)
   - example: odpovídal mu na jeho dotaz pravdu / že ... [he responded to his question truthfully / that ...]
   - asp.counterpart: odpověďt1 pf.
   - class: communication

2. odpovídá2 ~ reagovat [react]
   - frame: \(ACT_1^{obl} \ PAT_{na+4}^{obl} \ MEANS_{7}^{typ}\)
   - example: pokožka odpovídala na včelí bodnutí zarudnutím [the skin reacted to a bee sting by turning red]
   - asp.counterpart: odpověďt2 pf.

... odpovídá se (imperfective)

1. odpovídá se1 ~ být zodpovědný [be responsible]
   - frame: \(ACT_1^{obl} \ ADRR_3^{obl} \ PAT_{z+2}^{obl}\)
   - example: odpovídá se ze ztrát [he answers for the losses]

An abbreviated example for the base lemma “odpovídá”.

Ondřej Bojar
Some Computational Experiments with Czech

December 7, 2006
### Available Czech Data (not exhaustive!)

#### Monolingual Corpora

<table>
<thead>
<tr>
<th>Name and version</th>
<th>Sents.</th>
<th>Tokens</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech National Corpus (SYN2000d)</td>
<td>6.8M</td>
<td>114M</td>
<td>automatic lemmas+tags</td>
</tr>
<tr>
<td>Prague Dep Tbk (PDT 2.0)</td>
<td>50k–115k</td>
<td>0.8M–2.0M</td>
<td>manual tecto–manual morph</td>
</tr>
</tbody>
</table>

#### Parallel Czech-English

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

#### 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)

**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  
  \[ \Rightarrow \text{exponentially many analyses remained possible (though not correct).} \]
- 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 \(\Rightarrow\) 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)
\(\sim\) 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)
  \[\Rightarrow\text{where GIZA++ fails, humans often (38\% of tokens) disagree as well}\]
Alignments, Phrases and Phrase-Based MT

This time around, they’re moving even faster.

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**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>lemmatization for alignment</td>
<td>+2.0*</td>
</tr>
<tr>
<td>handling numbers</td>
<td>+0.9*</td>
</tr>
<tr>
<td>fixing clear BLEU errors</td>
<td>+0.5</td>
</tr>
<tr>
<td>dependency-based corpus expansion</td>
<td>+0.3</td>
</tr>
<tr>
<td>more out-of-domain parallel texts, also in LM</td>
<td>+0.4</td>
</tr>
<tr>
<td>bigged in-domain LM</td>
<td>+1.7*</td>
</tr>
<tr>
<td>more out-of-domain parallel texts, bigger in-domain LM</td>
<td>+5.0*</td>
</tr>
</tbody>
</table>

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).
**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.

<table>
<thead>
<tr>
<th>Translate+Check (T+C)</th>
<th>2*Translate+Generate (T+T+G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Czech</td>
</tr>
<tr>
<td>lowercase</td>
<td>lowercase</td>
</tr>
<tr>
<td>morphology</td>
<td>lemma</td>
</tr>
<tr>
<td></td>
<td>morphology</td>
</tr>
<tr>
<td>BLEU: 27.23</td>
<td>BLEU: 25.94</td>
</tr>
</tbody>
</table>

- 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

# Asociace uvedla, že domácí poptávka v září stoupla.

# The association said domestic demand grew in September.
Training: Decompose Trees into Treelets

# Asociace uvedla, že domácí poptávka v září stoupla.

# The association said domestic demand grew in September.
Training: Collect Dictionary of Treelet Pairs

\[ \text{Sb uvedla , že Pred} = \text{Sb said Pred} \]
\[ \text{asociace} = \text{The association} \]
\[ \text{Adj poptávka} = \text{Adj demand} \]
\[ \text{domestic} = \text{domácí} \]
Training: Collect Dictionary of Treelet Pairs (2)

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

\[
\text{Sb} \quad \text{Adv} \quad \text{stoupla} = \text{Sb} \quad \text{grew} \quad \text{Adv}
\]

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

\[
\text{Sb} \quad \text{Adv} \quad \text{stoupla} = \text{Sb} \quad \text{grew}
\]

\[
\text{Sb} \quad \text{Adv} \quad \text{stoupla} = \text{Sb} \quad \text{grew} \quad \text{Adv}
\]

\[
\text{Sb} \quad \text{Adv} \quad \text{Adv} \quad \text{stoupla} = \text{Sb} \quad \text{grew} \quad \text{Adv} \quad \text{Adv}
\]
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


---

Ondřej Bojar  
**Some Computational Experiments with Czech**  
December 7, 2006


## Detailed Numbers on Non-Projectivity

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge length</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English [%]</td>
<td>74.2</td>
<td>86.3</td>
<td>95.6</td>
</tr>
<tr>
<td>Czech [%]</td>
<td>51.8</td>
<td>72.1</td>
<td>90.2</td>
</tr>
<tr>
<td>Number of gaps</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Sentences [%]</td>
<td>76.9</td>
<td>22.7</td>
<td>0.42</td>
</tr>
<tr>
<td>Climbing steps</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Nodes [%]</td>
<td>90.3</td>
<td>8.0</td>
<td>1.3</td>
</tr>
</tbody>
</table>

2. Data by (Holan, 2003).
### Data Sparseness

<table>
<thead>
<tr>
<th>After having seen</th>
<th>20,000</th>
<th>75,000</th>
<th>sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>a new lemma comes every</td>
<td>1.6</td>
<td>1.8</td>
<td>test sentences</td>
</tr>
<tr>
<td>a new full morphological tag comes every</td>
<td>110</td>
<td>290</td>
<td>test sentences</td>
</tr>
<tr>
<td>a new simplified tag comes every</td>
<td>280</td>
<td>870</td>
<td>test sentences</td>
</tr>
</tbody>
</table>

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++ againsts golden set derived by joining the human annotations

<table>
<thead>
<tr>
<th>Humans</th>
<th>GIZA++</th>
<th>Baseline</th>
<th>Improved</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>en</td>
<td>cs</td>
</tr>
<tr>
<td>With Problems</td>
<td>With Problems</td>
<td>14.3</td>
<td>15.5</td>
</tr>
<tr>
<td>With Problems</td>
<td>OK</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>OK</td>
<td>With Problems</td>
<td>38.6</td>
<td>35.7</td>
</tr>
<tr>
<td>OK</td>
<td>OK</td>
<td>46.9</td>
<td>48.7</td>
</tr>
</tbody>
</table>
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_ytubující 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 .