Error-tagged Learner Corpus of Czech

Jirka Hana\textsuperscript{1} \hspace{1em} Alexandr Rosen\textsuperscript{1} \\
Svatava Škodová\textsuperscript{2} \hspace{1em} Barbora Štindlová\textsuperscript{2}

\textsuperscript{1}Charles University, Prague, Czech Republic \hspace{1em} \textsuperscript{2}Technical University, Liberec, Czech Republic

ACL 2010 \\
Fourth Linguistic Annotation Workshop \\
Uppsala, 15–16 July 2010
Outline of the talk

1. Introduction
2. Annotation scheme
3. Annotation process
4. Conclusion
Outline of the talk

1. Introduction
2. Annotation scheme
3. Annotation process
4. Conclusion
Learner Corpora

- Include texts produced by learners of a foreign language
- Early 1990s: used to compile learners’ dictionaries (e.g., *Longman Learner Corpus*)
- Used by authors of textbooks and researchers in *2nd Language Acquisition*
- Deviant forms can be corrected and their error type identified
- There can be simultaneous deviations on multiple levels
Some currently available learner corpora

<table>
<thead>
<tr>
<th>Size</th>
<th>L1</th>
<th>TL</th>
<th>TL proficiency</th>
<th>Error annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ICLE</strong> – Internat’l Corpus of Learner English</td>
<td>3M</td>
<td>21</td>
<td>English</td>
<td>advanced</td>
</tr>
<tr>
<td><strong>CLC</strong> – Cambridge Learner Corpus</td>
<td>30M</td>
<td>130</td>
<td>English</td>
<td>all levels</td>
</tr>
<tr>
<td><strong>USE</strong> – Uppsala Student English Corpus</td>
<td>1.2M</td>
<td>Swedish</td>
<td>English</td>
<td>advanced</td>
</tr>
<tr>
<td><strong>HKUST</strong> – Hong Kong UST Corpus of Learner English</td>
<td>25M</td>
<td>Chinese</td>
<td>English</td>
<td>advanced</td>
</tr>
<tr>
<td><strong>CLEC</strong> – Chinese Learner English Corpus</td>
<td>1M</td>
<td>Chinese</td>
<td>English</td>
<td>5 levels</td>
</tr>
<tr>
<td><strong>JEFLL</strong> – Japanese EFL Learner Corpus</td>
<td>0.7M</td>
<td>Japanese</td>
<td>English</td>
<td>advanced</td>
</tr>
<tr>
<td><strong>FALKO</strong> – Fehlerannotiertes Lernerkorpus</td>
<td>1.2M</td>
<td>various</td>
<td>German</td>
<td>advanced</td>
</tr>
<tr>
<td><strong>FRIDA</strong> – French Interlanguage Database</td>
<td>0.2M</td>
<td>various</td>
<td>French</td>
<td>intermediate</td>
</tr>
<tr>
<td><strong>CIC</strong> – Chinese Interlanguage Corpus</td>
<td>2M</td>
<td>96</td>
<td>Chinese</td>
<td>intermediate</td>
</tr>
<tr>
<td><strong>ASK</strong> – Andersspråkskorpus</td>
<td>?</td>
<td>10</td>
<td>Norwegian</td>
<td>two levels</td>
</tr>
</tbody>
</table>
A learner corpus of Czech

- The CzeSL Project: Czech as a Second Language
- Czech: rich inflection, derivation, complex agreement rules and information-structure-driven constituent order
- 2 million words to be transcribed, corrected and annotated within 3 years
- L1: Slavic (Russian, Ukrainian), Vietnamese, Romani, Chinese, ...
- Beginners to advanced learners
- Hand-written texts, elicited on various occasions in the class
Outline of the talk

1. Introduction
2. Annotation scheme
3. Annotation process
4. Conclusion
### Three-level format

- **Level 0** for the original
- **Successive corrections:**
  - Level 1 – graphemics and morphology.
  - Level 2 – agreement, valency, complex verb forms, lexicon, word order and negative concord
- Able to capture errors in multi-word discontinuous expressions
- Errors due to missed agreement, valency and pronominal reference have links to the words responsible for the proper form
- Automatic assignment of error tags wherever possible, based on comparing faulty and corrected forms, sometimes using morphosyntactic tags, assigned by a tagger
Bojal jsem se že ona se ne bude líbit prahu, *feared AUX RFL that she RFL not will *like *prague, 
I was afraid that she would not like Prague,

proto to bylo velmi vadí pro mně. therefore it was *very resent for me.
because I would be very unhappy about it.
Bojál jsem se že ona se ne bude líbit praahu,
feared AUX RFL that she RFL not will *like *prague,

*I was afraid that she would not like Prague,*

proto to bylo velmí vadí pro mně.
therefore it was *very resent for me.

*because I would be very unhappy about it.*
Bojaál jsem se, že ona se ne bude líbit prahu, feared AUX RFL that she RFL not will *like *prague,
I was afraid that she would not like Prague,

proto to bylo velmi vadí pro mně.
therefore it was *very resent for me.
because I would be very unhappy about it.
I was afraid that she would not like Prague,

because I would be very unhappy about it.
Bojaž jsem se, že ona se nebude líbit prahu, feared AUX RFL that she RFL not will like *prague, 
I was afraid that she would not like Prague,

proto to bylo velmi vadí pro mně. therefore it was *very resent for me. 
because I would be very unhappy about it.
Bojaál jsem se, že ona se nebude líubit pPrahu, feared AUX RFL that she RFL not will like prague,

I was afraid that she would not like Prague,

proto to bylo velmi vadí pro mně.

therefore it was *very resent for me.

because I would be very unhappy about it.
feared that she would not like Prague,

because I would be very unhappy about it.

Therefore it was very resent for me.
I was afraid that she would not like Prague,

because I would be very unhappy about it.
I was afraid that she would not like Prague,

because I would be very unhappy about it.
Bojál jsem se, že oná jí se nebude lí́bit práh̆u, proto to bylo velmi vádí pro mně. Česka republika je krásné místo.

I was afraid that she would not like Prague,

protože to bylo velmií vadi pro mně. because I would be very unhappy about it.
Bojál jsem se, že oná ji se nebude líbit práhu, proto to bylo velmi vadí pro mné. Česka republiku je u rasne místo.

I was afraid that she would not like Prague,

because I would be very unhappy about it.

protože to bylo velmií vadí pro mně́.

therefore it was very resent for me.
I was afraid that she would not like Prague,

because I would be very unhappy about it.
I was afraid that she would not like Prague, because I would be very unhappy about it.
Bojal jsem se že ona se ne bude libit prahu,
*feared AUX RFL that she RFL not will *like *prague,

I was afraid that she would not like Prague,
An annotation scheme of a sample sentence, part I

Bojal jsem se že ona se ne bude libit prahu,
*feared AUX RFL that she RFL not will *like *prague,

I was afraid that she would not like Prague,
I was afraid that she would not like Prague,
Annotation scheme

Annotation of a sample sentence, part I

Bojal jsem se že ona se ne bude libit prahu,
*feared AUX RFL that she RFL not will *like *prague,

Bál jsem se,

Bál jsem se,

I was afraid that she would not like Prague,
Annotation scheme

Annotation of a sample sentence, part I

Bojal jsem se že ona se ne bude libit prahu,
*feared AUX RFL that she RFL not will *like *prague,

unk  p

Bál jsem se, že

Bál jsem se, že

I was afraid that she would not like Prague,
I was afraid that she would not like Prague,
Annotation of a sample sentence, part I

Bojal jsem se že ona se ne bude líbit prahu , *

I was afraid that she would not like Prague,
Annotation of a sample sentence, part I

Bojal jsem se že ona se ne bude líbit Prahu,

*I feared that she will not like Prague,*

Bál jsem se, že nebude líbit Prahu

*I was afraid that she would not like Prague,*
I was afraid that she would not like Prague,
I was afraid that she would not like Prague,
Bojal jsem se že ona se ne bude líbit prahu,

I was afraid that she would not like Prague,
Bojal jsem se že ona se ne bude líbit Prahu,

Bál jsem se, že jí nebude líbit Praha,

*I was afraid that she would not like Prague,*

---

I. Hana et al. (CU Prague & TU Liberec)  
Error-tagged Learner Corpus of Czech  
ACL 2010 – LAW IV 10 / 25
I was afraid that she would not like Prague,
Annotation of a sample sentence, part II

proto to bylo velmi vadí pro mně.
therefore it was *very resent for me.

because I would be very unhappy about it.
proto to bylo velmi vadí pro mně.
therefore it was *very resent for me.

dia

velmi

velmi

*because I would be very unhappy about it.*
proto to bylo velmi vadí pro mně.
therefore it was *very resent for me.

proto

velmi

dia

protože

velmi

because I would be very unhappy about it.
proto to bylo velmi vadí pro mně.
therefore it was *very resent for me.

proto to velmi
dia

proto to velmi
lex

protože to velmi

because I would be very unhappy about it.
Annotation of a sample sentence, part II

proto to bylo velmi vadí pro mně.
therefore it was *very resent for me.

dia

proto to velmi.

lex

protože to velmi.

because I would be very unhappy about it.
Annotation of a sample sentence, part II

proto to bylo velmí vadí pro mně .
therefore it was *very resent for me .
dia
proto to bylo velmi vadí .
lex
protože to velmi .

because I would be very unhappy about it.
proto to bylo velmi vadí pro mně.
therefore it was *very resent for me.

dia

proto to bylo velmi vadí.

lex cvf

protože to by velmi vadilo.

because I would be very unhappy about it.
proto to bylo velmi vadí pro mně.

therefore it was *very resent for me.

because I would be very unhappy about it.
proto to bylo velmi vadí pro mně.
therefore it was *very resent for me .

dia

proto to bylo velmi vadí pro mně .

lex cvf

protožе to by velmi vadilo

because I would be very unhappy about it.
Annotation scheme

Annotation of a sample sentence, part II

proto to bylo velmí vadí pro mně .
therefore it was *very resent for me .
dia
proto to bylo velmi vadí pro mně .

lex cvf val
protože to by velmi vadilo pro mě .

because I would be very unhappy about it.
proto to bylo velmi vadí pro mně.
therefore it was *very resent for me.
dia
proto to bylo velmi vadí pro mně.
lex cvf val
protože to by velmi vadilo mi.

because I would be very unhappy about it.
Annotation of a sample sentence, part II

proto to bylo velmi vadí pro mně.
therefore it was *very resent for me.

dia

proto to bylo velmi vadí pro mně.

lex cvf

protože to by mi velmi vadilo

val,wo

because I would be very unhappy about it.
# Types of errors at Level 1

<table>
<thead>
<tr>
<th>Error type</th>
<th>Tag</th>
<th>Links</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word boundary</td>
<td>bnd</td>
<td>m:n</td>
<td>Auto</td>
</tr>
<tr>
<td>Punctuation</td>
<td>p</td>
<td>0:1, 1:0</td>
<td>Auto</td>
</tr>
<tr>
<td>Capitalisation</td>
<td>cap</td>
<td>1:1</td>
<td>Auto</td>
</tr>
<tr>
<td>Diacritics</td>
<td>dia</td>
<td>1:1</td>
<td>Auto</td>
</tr>
<tr>
<td>Character(s)</td>
<td>char</td>
<td>1:1</td>
<td>Auto</td>
</tr>
<tr>
<td>Inflection</td>
<td>infl</td>
<td>1:1</td>
<td>Auto</td>
</tr>
<tr>
<td>Unknown lexeme</td>
<td>unk</td>
<td>1:1</td>
<td>Manual</td>
</tr>
<tr>
<td>Foreign word</td>
<td>fw</td>
<td>1:1</td>
<td>Manual</td>
</tr>
</tbody>
</table>
## Types of errors at Level 2

<table>
<thead>
<tr>
<th>Error type</th>
<th>Tag</th>
<th>Links</th>
<th>Ref</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement</td>
<td>agr</td>
<td>1:1</td>
<td>1</td>
<td>Manual</td>
</tr>
<tr>
<td>Valency</td>
<td>val</td>
<td>1:1</td>
<td>1</td>
<td>Manual</td>
</tr>
<tr>
<td>Pronominal reference</td>
<td>ref</td>
<td>1:1</td>
<td>1</td>
<td>Manual</td>
</tr>
<tr>
<td>Complex verb forms</td>
<td>cvf</td>
<td>m:n</td>
<td>0,1</td>
<td>Manual</td>
</tr>
<tr>
<td>Negation</td>
<td>neg</td>
<td>m:n</td>
<td>0,1</td>
<td>Manual</td>
</tr>
<tr>
<td>Missing constituent</td>
<td>miss</td>
<td>0:1</td>
<td>0</td>
<td>Manual</td>
</tr>
<tr>
<td>Odd constituent</td>
<td>odd</td>
<td>1:0</td>
<td>0</td>
<td>Manual</td>
</tr>
<tr>
<td>Modality</td>
<td>mod</td>
<td>1:1</td>
<td>0</td>
<td>Manual</td>
</tr>
<tr>
<td>Word order</td>
<td>wo</td>
<td>m:n</td>
<td>0</td>
<td>Manual</td>
</tr>
<tr>
<td>Lexis &amp; phraseology</td>
<td>lex</td>
<td>m:n</td>
<td>0,1</td>
<td>Manual</td>
</tr>
</tbody>
</table>
Annotation policy

Minimal intervention: corrected text need not be perfect, grammatical is enough

To do

We still need to provide annotators with guidelines on how to:

- handle uncertainty about the author’s intended meaning,
- identify false-friends errors,
- handle colloquial language.
Data format

- Prague Markup Language (PML, used in *Prague Dependency Treebank*)
- Generic, XML-based, for rich layered annotation
- A higher level contains information about words on that level, about errors and about relations to tokens on lower levels
- Portion of Level 1 of the sample sentence encoded in the PML data format – see next slide
Bál jsem
Outline of the talk

1. Introduction
2. Annotation scheme
3. Annotation process
4. Conclusion
The annotation workflow

1. A handwritten document is transcribed into HTML using off-the-shelf tools.
2. The information in the html document is used to generate Level 0 and a default Level 1 encoded in the PML format.
3. An annotator manually corrects the document and provides some information about errors using our annotation tool.
4. Error information that can be inferred automatically is added.
5. See next slide for a sample sentence in the annotation tool.
<table>
<thead>
<tr>
<th>Boj</th>
<th>jsem</th>
<th>se</th>
<th>že</th>
<th>ona</th>
<th>se</th>
<th>ne</th>
<th>bude</th>
<th>lít</th>
<th>Prahu</th>
<th>proto</th>
<th>to</th>
<th>bylo</th>
<th>velmi</th>
<th>vadí</th>
<th>pro</th>
<th>mně</th>
<th>Česka</th>
</tr>
</thead>
<tbody>
<tr>
<td>unk</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Proč mám/nemám rád Českou republiku?**

Už se nacházím v česko-české republice až půl roku. Tohoto mi muselo by stačit, alech rozumí, mám rád to země nebo ne rád. Teďko muži určitě řeknet, že českou republiku má mějí. Ten to země na všech potočebují a mnohé přízkyně. Boj jsem se že ona se ne bude lít Prahu proto to bylo velmi vadí pro mně. Česka republika je krásné místo, tady je hodné hezké památky, například pražský hrad a vyšehrad. Líbím se moc pražský hrad, protože tam je zámky, který velmi krásné a hezky. Také v Čechách je dobra příroda a když jsme se procházeli na divocí šarže byli šokováni s některých krásných pohledů. Je to nejkrásnější místo ve všem světě.
<table>
<thead>
<tr>
<th>errid</th>
<th>L0</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>proto</td>
<td>to bylo velmi</td>
<td>vadí</td>
<td>promně</td>
</tr>
<tr>
<td>gloss</td>
<td>therefore it was</td>
<td>*very resent for me</td>
<td></td>
<td></td>
</tr>
<tr>
<td>errid</td>
<td>dia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>lex</td>
<td>cvf</td>
<td>2</td>
<td>val 4</td>
</tr>
<tr>
<td></td>
<td>protože</td>
<td>to by</td>
<td>velmi</td>
<td>vadilo</td>
</tr>
<tr>
<td>errid</td>
<td>wo</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>protože</td>
<td>to by</td>
<td>mi</td>
<td>velmi</td>
</tr>
</tbody>
</table>
Postprocessing

Manual annotation is followed by automatic post-processing, providing the corpus with additional information:

1. Level 1: lemma and morphosyntactic tags (not disambiguated)
2. Level 2: lemma and morphosyntactic tags (disambiguated)
3. Level 1: type of error (by comparing the original and corrected strings) (e.g. *libit – líbit ‘like’ – error in diacritics)
4. Level 2: type of morphosyntactic errors caused by agreement or subcategorisation error (by comparing morphosyntactic tags at Level 1 and 2)
5. Formal error description: missing/extra expression, wrong order
6. In the future, we plan to automatically tag errors in verb prefixes, inflectional endings, spelling, palatalisation, metathesis, etc.
Outline of the talk

1. Introduction
2. Annotation scheme
3. Annotation process
4. Conclusion
Error annotation is a very resource-intensive task, but an error-tagged corpus is an invaluable tool:
- to obtain a reliable picture of the learners’ interlanguage and
- to adapt teaching methods and learning materials.
Acknowledgments

Thanks to

other members of the project team, namely Karel Šebesta, Milena Hnátková, Tomáš Jelínek, Vladimír Petkevič, and Hana Skoumalová