

# Error-tagged Learner Corpus of Czech

Jirka Hana<sup>1</sup>    Alexandr Rosen<sup>1</sup>  
Svatava Škodová<sup>2</sup>    Barbora Štindlová<sup>2</sup>

<sup>1</sup>Charles University, Prague, Czech Republic

<sup>2</sup>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

Size	L1	TL	TL proficiency	Error annotation
<b>ICLE</b> – <i>Internat'l Corpus of Learner English</i>				
3M	21	English	advanced	yes
<b>CLC</b> – <i>Cambridge Learner Corpus</i>				
30M	130	English	all levels	yes
<b>USE</b> – <i>Uppsala Student English Corpus</i>				
1.2M	Swedish	English	advanced	no
<b>HKUST</b> – <i>Hong Kong UST Corpus of Learner English</i>				
25M	Chinese	English	advanced	yes
<b>CLEC</b> – <i>Chinese Learner English Corpus</i>				
1M	Chinese	English	5 levels	yes
<b>JEFL</b> – <i>Japanese EFL Learner Corpus</i>				
0.7M	Japanese	English	advanced	yes
<b>FALKO</b> – <i>Fehlerannotiertes Lernerkorpus</i>				
1.2M	various	German	advanced	yes
<b>FRIDA</b> – <i>French Interlanguage Database</i>				
0.2M	various	French	intermediate	yes
<b>CIC</b> – <i>Chinese Interlanguage Corpus</i>				
2M	96	Chinese	intermediate	?
<b>ASK</b> – <i>Andersspråkskorpus</i>				
?	10	Norwegian	two levels	yes

## 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

## A sample sentence

republicu a miluju. tento ...  
 že potřebuju ja a moje přítelkyně. Bojal jsem  
 se že ona se ne bude líbit prahu, proto to bylo velmi  
 vadí pro mně. Česká republika je krásné místo.

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.*

## A sample sentence

republicu va miluju. tento ...  
 že potřebuju ja a moje přítelkyně. ~~Bojal~~ <sup>em</sup> ~~js~~  
 se že ona se ne bude líbit prahu, proto to bylo velmi  
 vadí pro mně. Česká republika je krásné místo.

**Boja**ál 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.*

## A sample sentence

republicu va miluju. tento ...  
 že potřebuju ja a moje přítelkyně. ~~Bojal~~ <sup>em</sup> ~~js~~  
 se že ona se ne bude líbit prahu, proto to bylo velmi  
 vadí pro mně. Česká republika je krásné místo.

**Boja**ál 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.*

## A sample sentence

republicu a miluju. tento ...  
 že potřebuju ja a moje přítelkyně. ~~Bojál~~ <sup>em</sup> jsem  
 se že ona se bude líbit prahu, proto to bylo velmi  
 vadí pro mě. Česká republika je krásné místo.

**Boja**ál 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 mě.  
 therefore it was \*very resent for me.  
*because I would be very unhappy about it.*

## A sample sentence

republicu a miluju. tento ...  
 že potřebuju ja a moje přítelkyně. ~~Bojál~~ <sup>em</sup> jsem  
 se že ona se nebude líbit prahu, proto to bylo velmi  
 vadí pro mně. Česká republika je krásné místo.

**Boja**ál 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.*

# A sample sentence

republicu a miluju. tento ...  
 že potřebuju ja a moje přítelkyně. ~~Boj~~ <sup>em</sup> jsem se  
 že ona se bude líbit prahu, proto to bylo velmi  
 vadí pro mě. Česká republika je krásné místo.

**Boj**ál 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 mě.  
 therefore it was \*very resent for me.  
*because I would be very unhappy about it.*

# A sample sentence

republicu a miluju. tento ...  
 že potřebuju ja a moje přítelkyně. ~~Bojal~~ <sup>em</sup> jsem  
 se že ona se bude líbit prahu, proto to bylo velmi  
 vadí pro mně. Česká republika je krásné místo.

Bojaál jsem se, že ona jí se nebude líbit praha,  
 feared AUX RFL that her 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.*

# A sample sentence

republicu va miluju. tento ...  
 že potřebuju ja a moje přítelkyně. ~~Bojal~~ <sup>em</sup> ~~js~~  
 se že ona se bude líbit prahu, proto to bylo velmi  
 vadí pro mně. Česká republika je krásné místo.

Bojaál jsem se, že ona jí se nebude líbit řPrahu,  
 feared AUX RFL that her 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.*

## A sample sentence

republicu a miluju. tento ...  
 že potřebuju ja a moje přítelkyně. ~~Bojal~~ <sup>em</sup> ~~jsa~~  
 se že ona se bude líbit prahu, proto to bylo velmi  
 vadí pro mně. Česká republika je krásné místo.

Bojaál jsem se, že ona jí se nebude líbit řPrahu,  
 feared AUX RFL that her RFL not will like prague,  
*I was afraid that she would not like Prague,*

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

## A sample sentence

republicu a miluju. tento ...  
 že potřebuju ja a moje přítelkyně. ~~Bojál~~ <sup>em</sup> ~~js~~  
 se že ona se bude líbit prahu, proto to bylo velmi  
 vadí pro mně. Česká republika je krásné místo.

Bojaál jsem se, že ona jí se nebude líbit řPrahu,  
 feared AUX RFL that her RFL not will like prague,  
*I was afraid that she would not like Prague,*

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

## A sample sentence

republiku a miluju. Tento ...  
 že potřebuju ja a moje přítelkyně. ~~Bojal~~ <sup>em</sup> ~~jsa~~  
 se že ona se bude líbit prahu, proto to bylo velmi  
 vadí pro mně. Česká republika je krásné místo.

Bojaál jsem se, že ona jí se nebude líbit řPrahu,  
 feared AUX RFL that her RFL not will like prague,  
*I was afraid that she would not like Prague,*

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

## A sample sentence

republiku va miluju. tento ...  
 že potřebuju ja a moje přítelkyně. ~~Bojál~~ <sup>em</sup> ~~js~~  
 se že ona se bude líbit prahu, proto to bylo velmi  
 vadí pro mně. Česká republika je krásné místo.

Bojaál jsem se, že ona jí se nebude líbit řPrahu,  
 feared AUX RFL that her RFL not will like prague,  
*I was afraid that she would not like Prague,*

protože to bylo velmi vadí mi.  
 therefore it was very resent me.  
*because I would be very unhappy about it.*

## A sample sentence

republiku va miluju. tento ...  
 že potřebuju ja a moje přítelkyně. ~~Bojal~~ <sup>em</sup> ~~js~~  
 se že ona se bude líbit prahu, proto to bylo velmi  
 vadí pro mně. Česká republika je krásné místo.

Bojaál jsem se, že ona jí se nebude líibit řPrahu,  
 feared AUX RFL that her RFL not will like prague,  
*I was afraid that she would not like Prague,*

proto že to by lo velmi vadilo mi.  
 therefore it would very resented me.  
*because I would be very unhappy about it.*

## 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 ,

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

## 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

Bál

Bál

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

# 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

Bál jsem se

Bál jsem se

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

# 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 ,

Bál jsem se ,

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

# 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**

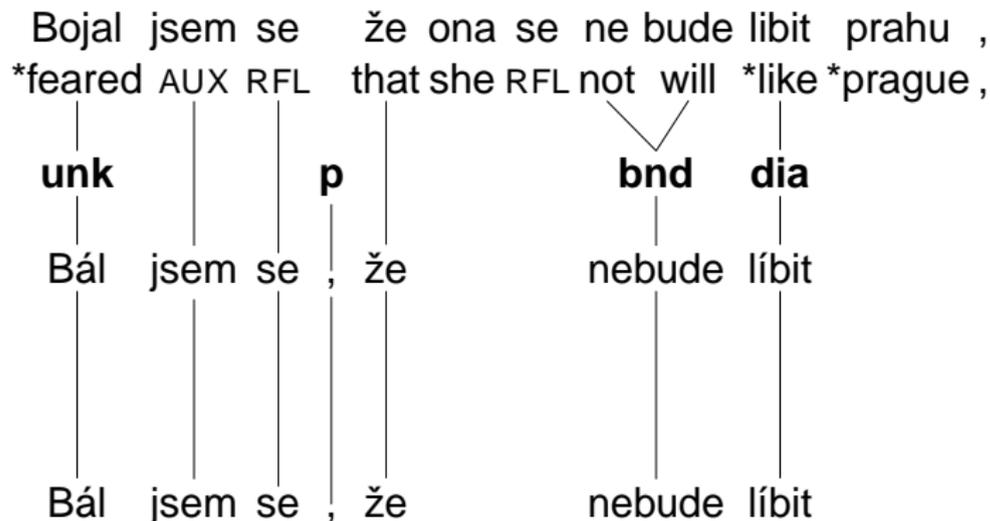
Bál jsem se , že

Bál jsem se , že

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

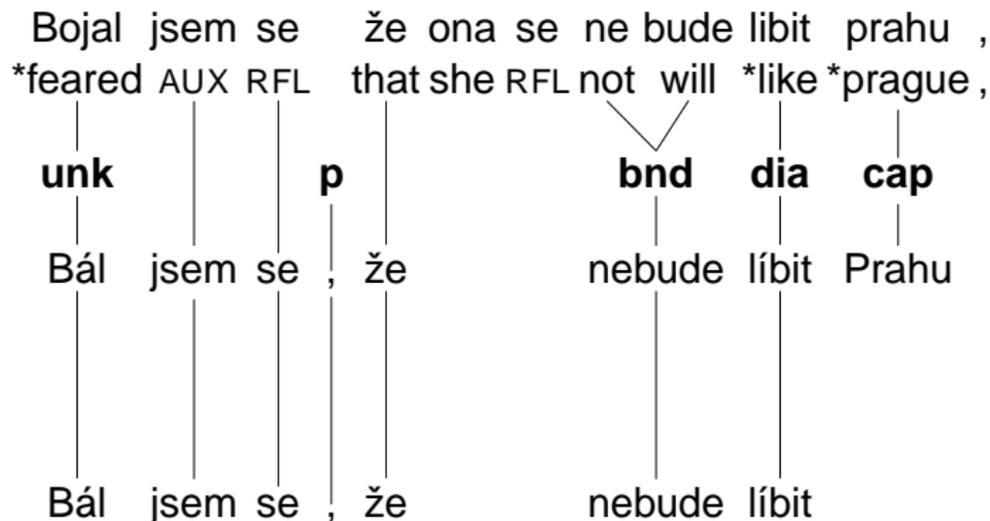


# Annotation of a sample sentence, part I



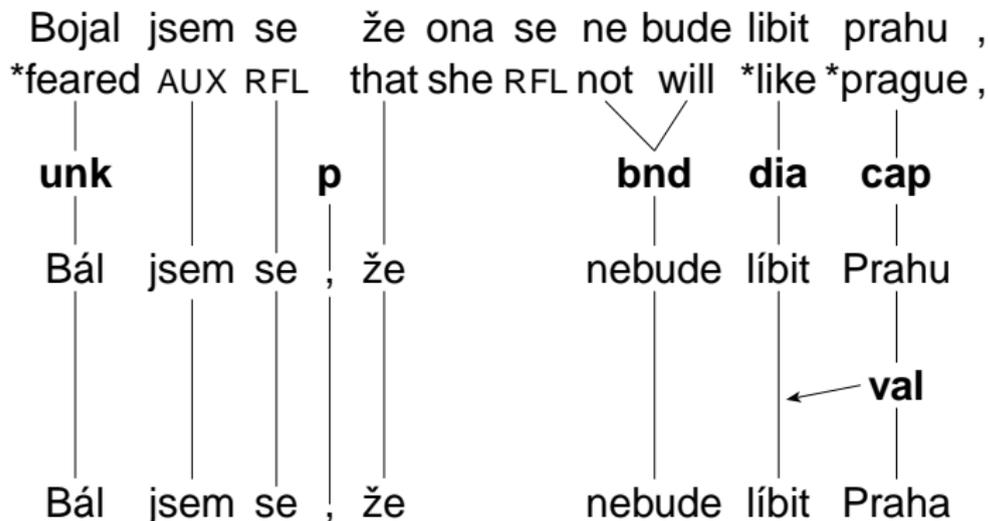
*I was afraid that she would not like Prague,*

# Annotation of a sample sentence, part I



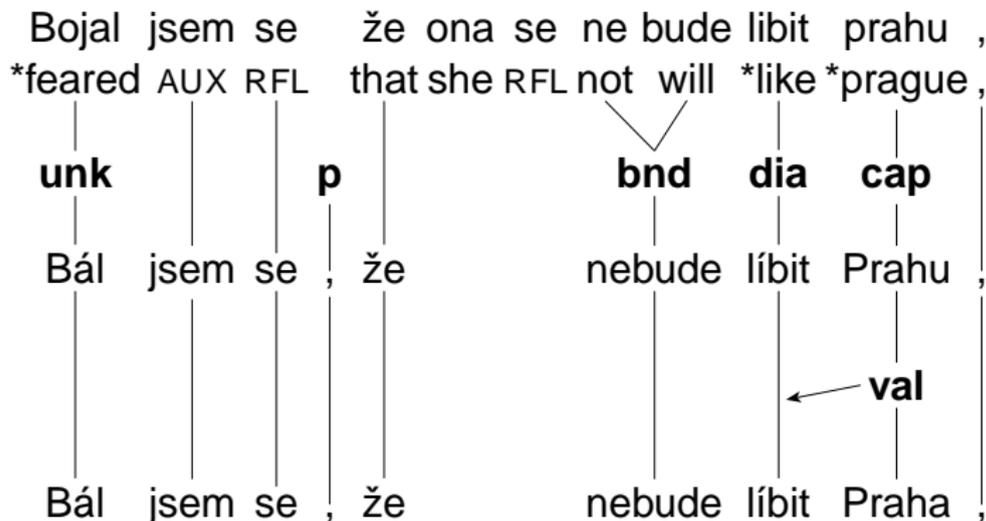
*I was afraid that she would not like Prague,*

# Annotation of a sample sentence, part I



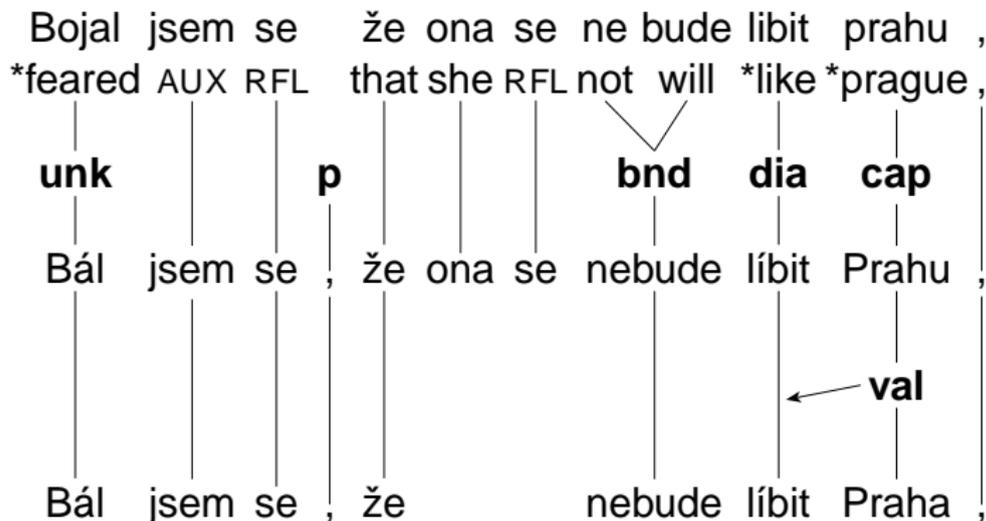
*I was afraid that she would not like Prague,*

# Annotation of a sample sentence, part I



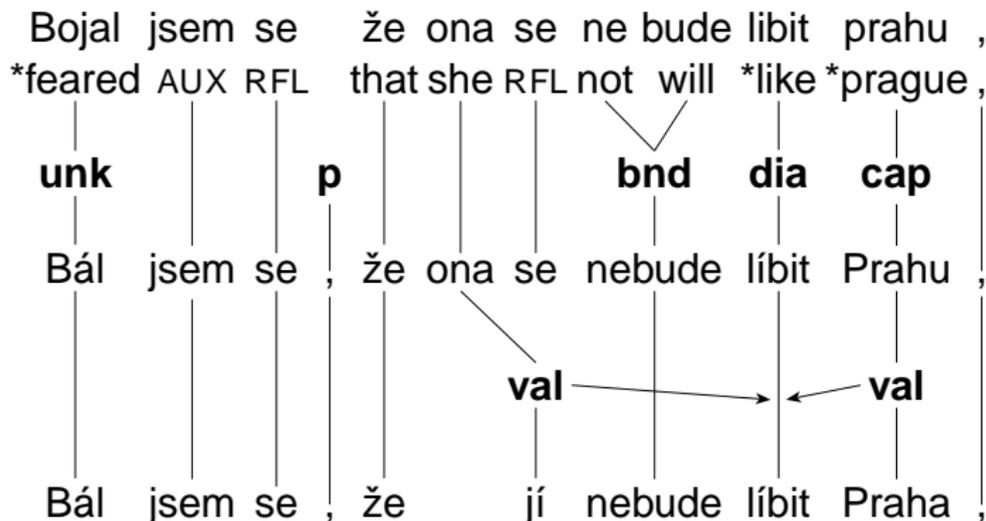
*I was afraid that she would not like Prague,*

# Annotation of a sample sentence, part I



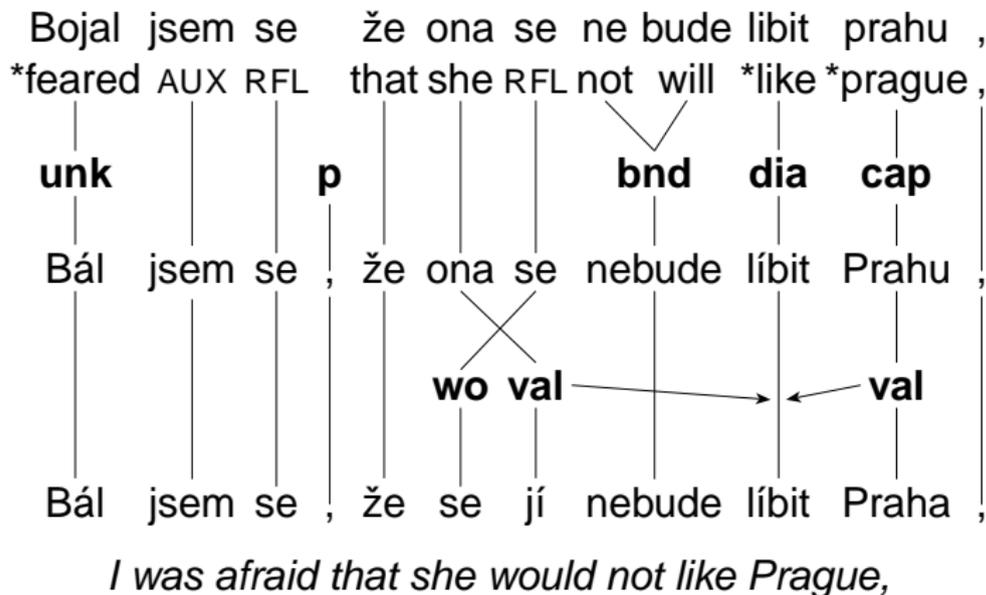
*I was afraid that she would not like Prague,*

# Annotation of a sample sentence, part I



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

# Annotation of a sample sentence, part I



## 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.*

## Annotation of a sample sentence, part II

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

—  
**dia**

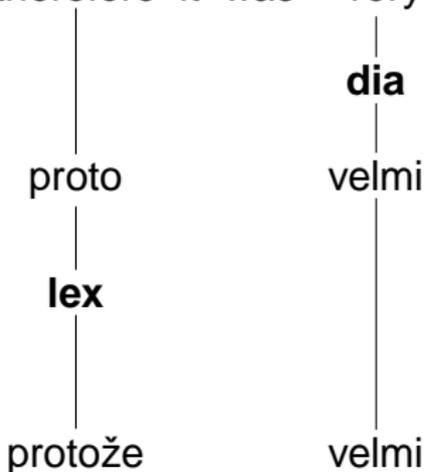
—  
 velmi

—  
 velmi

*because I would be very unhappy about it.*

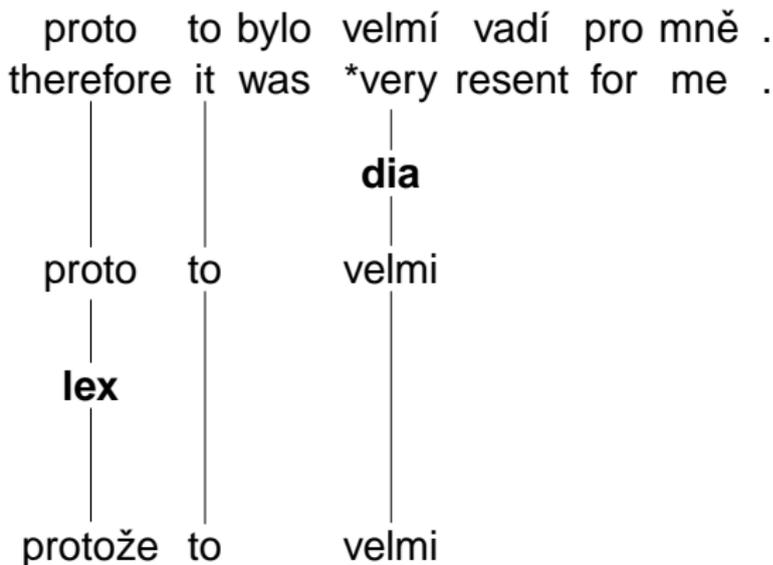
## Annotation of a sample sentence, part II

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



*because I would be very unhappy about it.*

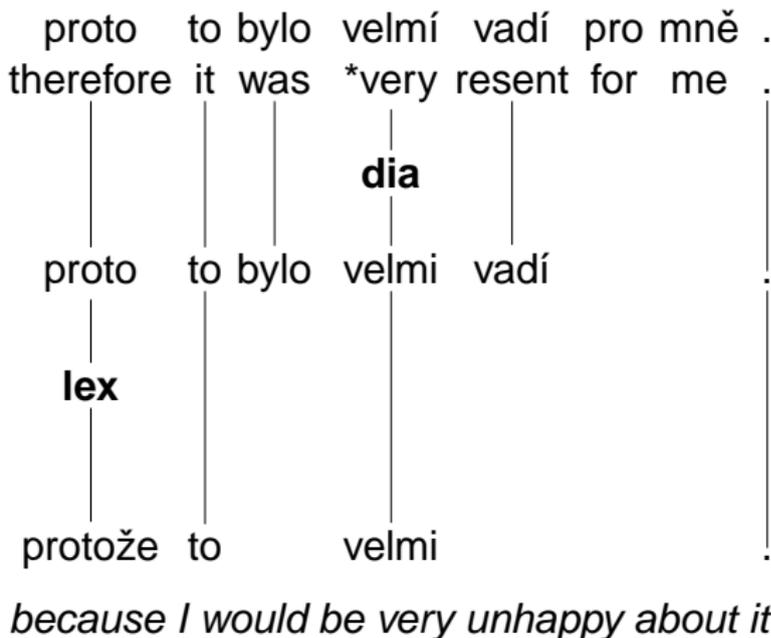
## Annotation of a sample sentence, part II



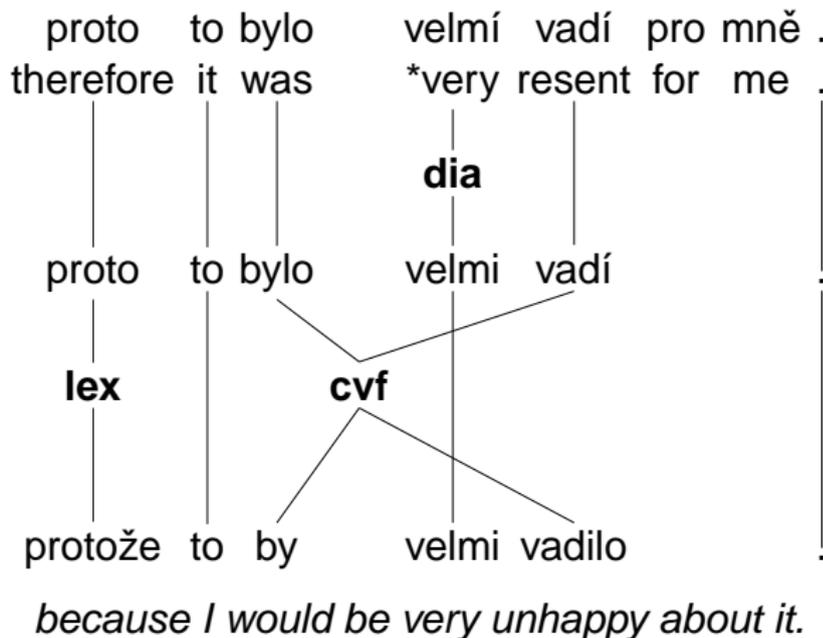
*because I would be very unhappy about it.*



## Annotation of a sample sentence, part II



# Annotation of a sample sentence, part II

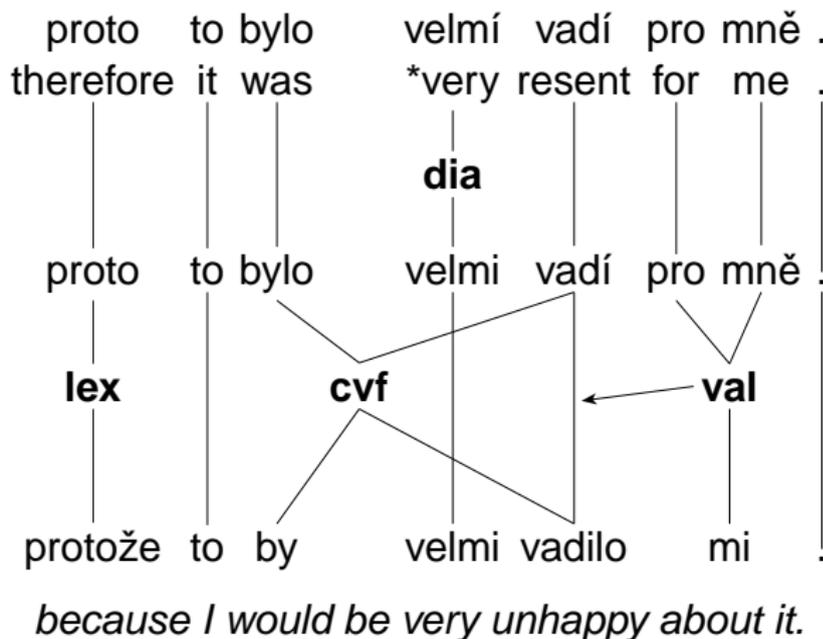




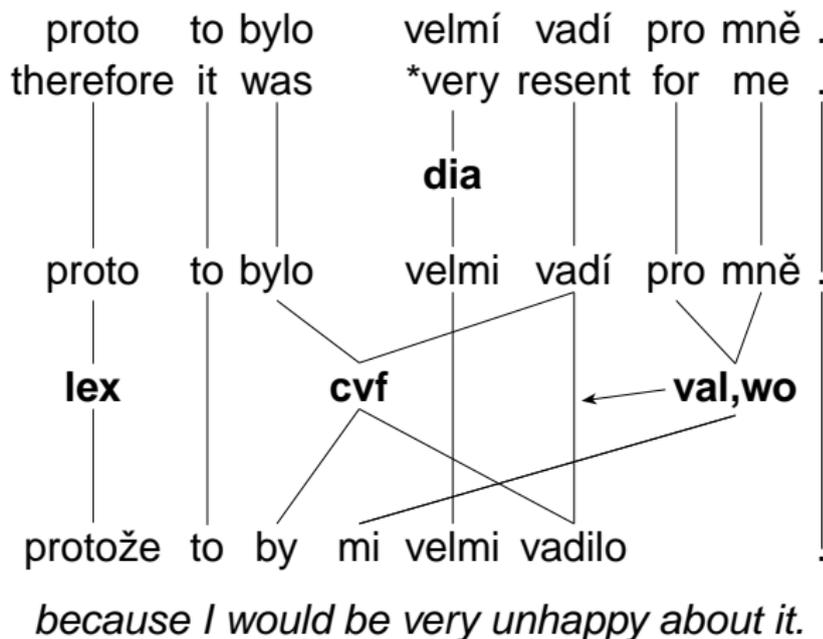




# Annotation of a sample sentence, part II



# Annotation of a sample sentence, part II



# Types of errors at Level 1

<b>Error type</b>	<b>Tag</b>	<b>Links</b>	<b>Assignment</b>
Word boundary	<b>bnd</b>	m:n	Auto
Punctuation	<b>p</b>	0:1, 1:0	Auto
Capitalisation	<b>cap</b>	1:1	Auto
Diacritics	<b>dia</b>	1:1	Auto
Character(s)	<b>char</b>	1:1	Auto
Inflection	<b>infl</b>	1:1	Auto
Unknown lexeme	<b>unk</b>	1:1	Manual
Foreign word	<b>fw</b>	1:1	Manual

## Types of errors at Level 2

Error type	Tag	Links	Ref	Assignment
Agreement	<b>agr</b>	1:1	1	Manual
Valency	<b>val</b>	1:1	1	Manual
Pronominal reference	<b>ref</b>	1:1	1	Manual
Complex verb forms	<b>cvf</b>	m:n	0,1	Manual
Negation	<b>neg</b>	m:n	0,1	Manual
Missing constituent	<b>miss</b>	0:1	0	Manual
Odd constituent	<b>odd</b>	1:0	0	Manual
Modality	<b>mod</b>	1:1	0	Manual
Word order	<b>wo</b>	m:n	0	Manual
Lexis & phraseology	<b>lex</b>	m:n	0,1	Manual

## 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

```
<?xml version="1.0" encoding="UTF-8"?>
<adata xmlns="http://utkl.cuni.cz/czes1/">
  <head>
    <schema href="adata_schema.xml" />
    <references>
      <reffile id="w" name="wdata" href="r049.w.xml" />
    </references>
  </head>
  <doc id="a-r049-d1" lowerdoc.rf="w#w-r049-d1">
    ...
    <para id="a-r049-d1p2" lowerpara.rf="w#w-r049-d1p2">
      ...
      <s id="a-r049-d1p2s5">
        <w id="a-r049-d1p2w50">
          <token>Bál</token>
        </w>
        <w id="a-r049-d1p2w51">
          <token>jsem</token>
        </w>
        ...
      </s>
      ...
      <edge id="a-r049-d1p2e54">
        <from>w#w-r049-d1p2w46</from>
        <to>a-r049-d1p2w50</to>
        <error> <tag>unk</tag> </error>
      </edge>
      <edge id="a-r049-d1p2e55">
        <from>w#w-r049-d1p2w47</from>
        <to>a-r049-d1p2w51</to>
      </edge>
      ...
    </para>
    ...
  </doc>
</adata>
```

# 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.

feat 201006101454

File View Tools Window Help

- Properties x ST\_Randyskova\_Vob\_KA\_049.b x

2 / 4 Add Layer R Export Spacing X: 50 Y: 100

Bojal	jsem	se	že	ona	se	ne	bude	líbit	prahu	,	proto	to	bylo	velmi	vadí	pro	mně	.	Česka
unk	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Bál	jsem	se	že	ona	se	bude	líbit	Prahu	,	proto	to	bylo	velmi	vadí	pro	mně	.	Česka	
X	X	X	X	X	val	wo	X	X	val	X	lex	X	cvř	X	X	wo	val	X	
Bál	jsem	se	že	se	ji	bude	líbit	Praha	,	protože	to	by	mi	velmi	vadilo	.			

Proč mám/nemám rád (Č)ěskou republiku?

Už se nacházím v České republice až půl roku. toho mě musilo by stačit, abych rozuměl, mám rád to země nebo ne rád. teďko můžu určitě říct, že Českou republiku já miluju. tento země má všechna že potřebuju ja a moje přítelkyně. Bojal jsem se že ona se ne bude líbit **prahu**, proto to bylo velmi vadí pro mně. Česka republika je krásne místo. tady je hodně hezké pamatek. například pražský hrad a výšehrad. líbim se moc pražský hrad, protože tam je zámky, který velmi krásne a hezke. take v čechach je dobra příroda a když jsme se procházili na divoke šarce byli šokovani o4 z tech krásnych pohledů. Je to nekrásnější místo ve všem bílém světě. take rad že Češi ie dobri

Fit WFit Orig Zoom

miluju. tento země má všechna  
ja a moje přítelkyně. Bojal jsem se  
že líbit prahu, proto to bylo velmi  
Česka republika je krásne místo,  
hezke pamatek, například pražský  
líbim se moc pražský hrad, proto

	0	1	2	3	4	5	6	7	8
L0	proto	to	bylo	velmí	vadí	promně	.		
gloss	<i>therefore</i>	<i>it</i>	<i>was</i>	<i>*very</i>	<i>resent</i>	<i>for</i>	<i>me</i>	.	
errid				<b>dia</b>					
L1	proto	to	bylo	velmi	vadí	promně	.		
errid	<b>lex</b>		<b>cvf</b>		<b>2</b>	<b>val 4</b>			
L2	protože	to	by	velmi	vadilo	mi	.		
errid				<b>w0</b>					
L3	protože	to	by	mi	velmi	vadilo	.		

Done.

# 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**

## 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

