Annotating Foreign Learners' Czech

Alexandr Rosen[1] Svatava Škodová[2]
Barbora Štindlová[2] Jirka Hana[1]

[1]Charles University, Prague, Czech Republic

[2]Technical University, Liberec, Czech Republic

FDSL 8.5
Formal Description of Slavic Languages 8.5
Faculty of Arts, Masaryk University, Brno
26–27 November 2010

Outline of the talk

Outline of the talk

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

A learner corpus of Czech

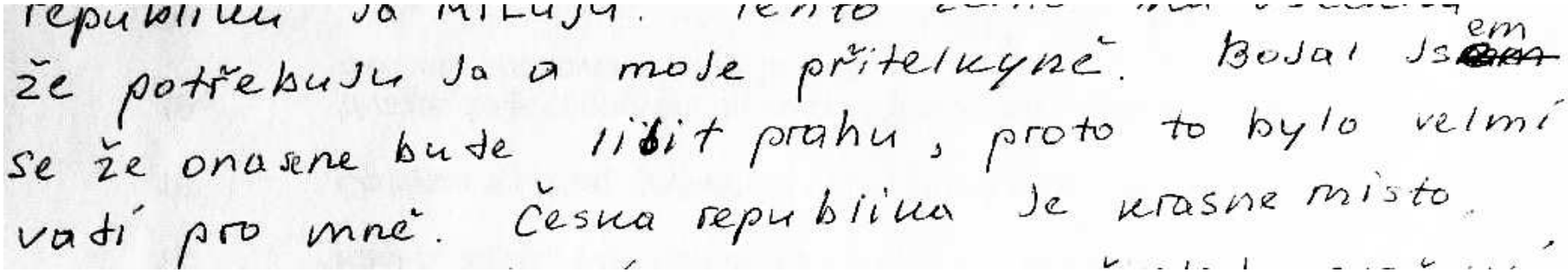
- The *CzeSL* Project: Czech as a Second Language
- Czech: rich inflection, derivation, complex agreement rules and information-structure-driven constituent order
- 2 million words to be transcribed, corrected and annotated within 3 years
- L1: Slavic (Russian, Ukrainian), Vietnamese, Romani, Chinese,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

Outline of the talk

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

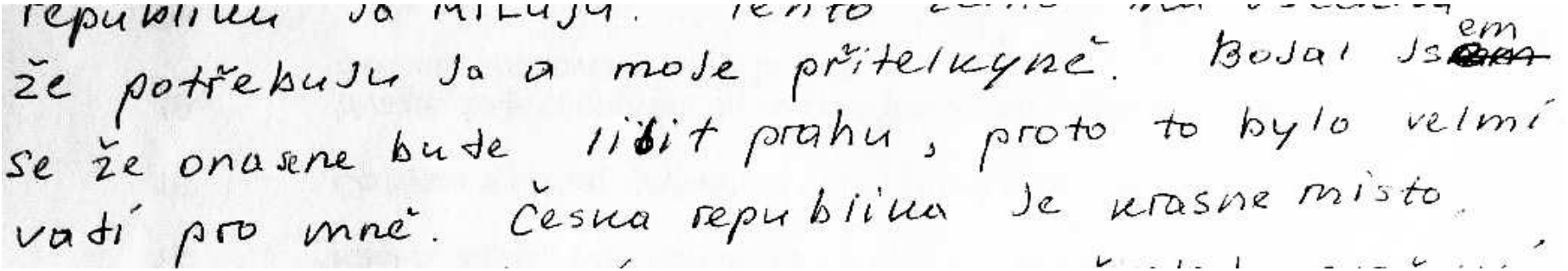


Bojal jsme
*feared AUX-*PL
I was afraid

že ona se ne bude libila slavnou prahu,
that she RFL not will *like *famous *prague,
that she would not like the famous city of Prague,

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

A sample sentence

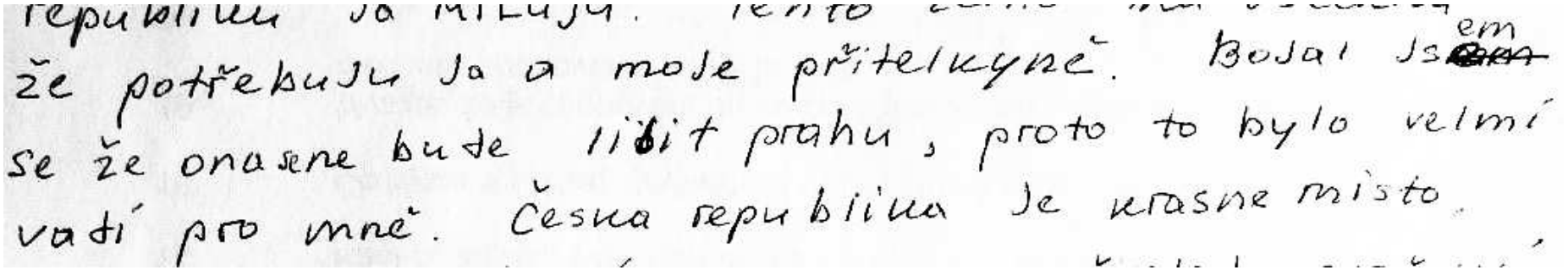


B~~ojaá~~l jsme
feared AUX-*PL
I was afraid

že ona se ne bude libila slavnou prahu,
that she RFL not will *like *famous *prague,
that she would not like the famous city of Prague,

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

A sample sentence

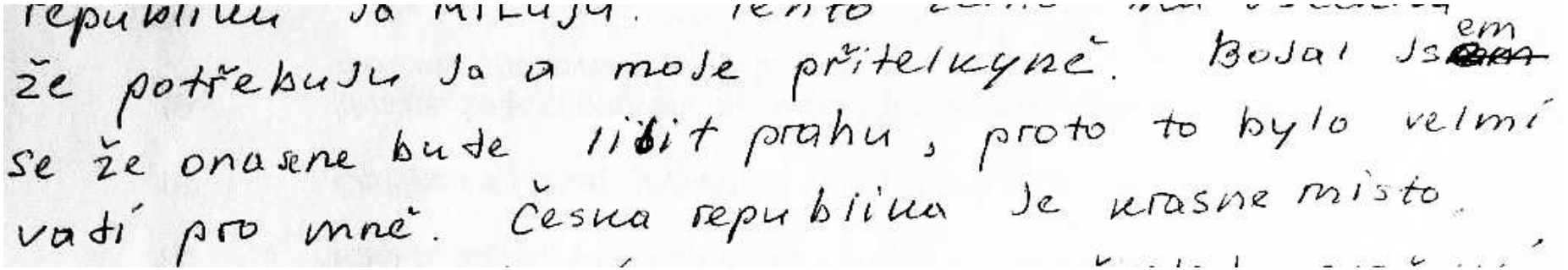


B~~ojaá~~l js~~me~~em
feared AUX-SG
I was afraid

že ona se ne bude libila slavnou prahu,
that she RFL not will *like *famous *prague,
that she would not like the famous city of Prague,

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

A sample sentence

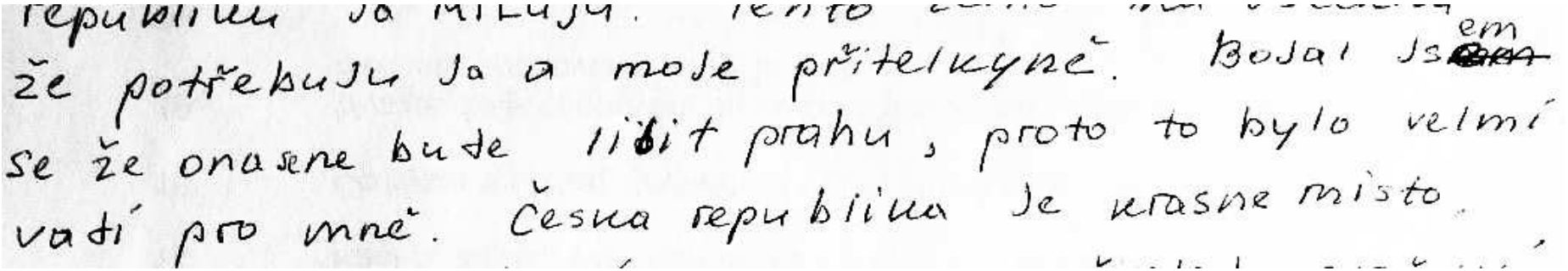


Bojáál jsmeem se
feared AUX-SG RFL
I was afraid

že ona se ne bude libila slavnou prahu,
that she RFL not will *like *famous *prague,
that she would not like the famous city of Prague,

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

A sample sentence

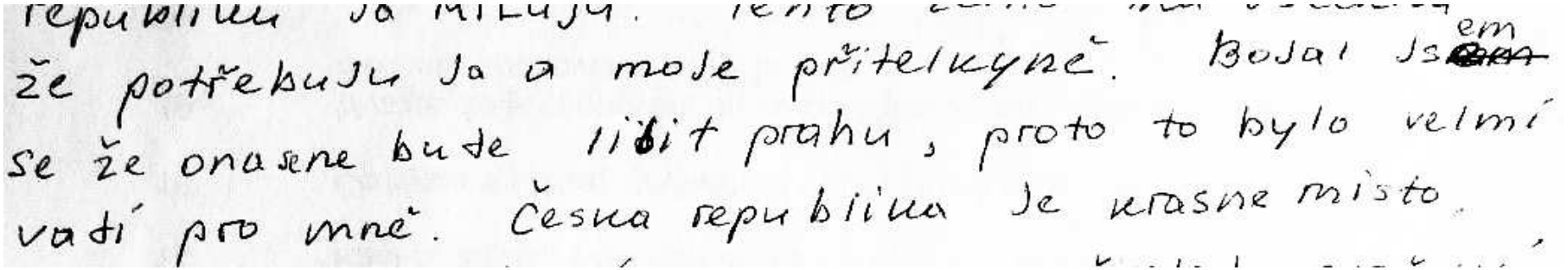


B~~ojaá~~l js~~me~~em se,
feared AUX-SG RFL
I was afraid

že ona se ne bude libila slavnou prahu,
that she RFL not will *like *famous *prague,
that she would not like the famous city of Prague,

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

A sample sentence

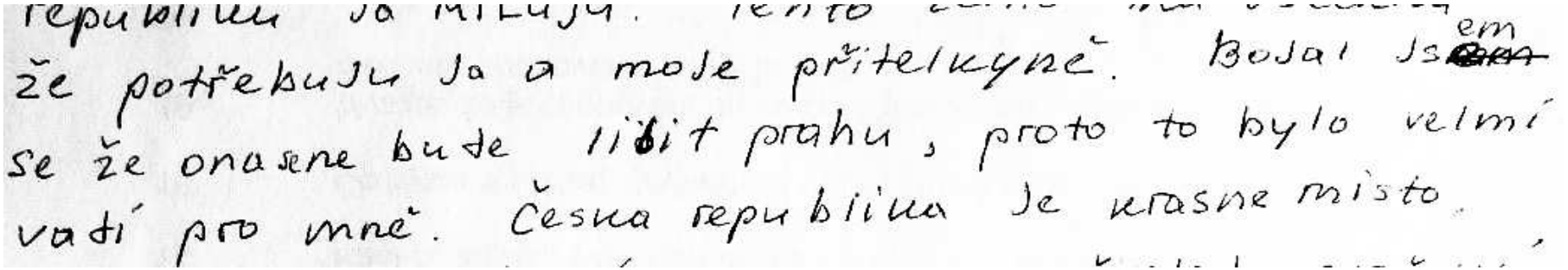


B~~ojaá~~l js~~me~~em se,
feared AUX-SG RFL
I was afraid

že ~~ona~~jí se ne bude libila slavnou prahu,
that her RFL not will *like *famous *prague,
that she would not like the famous city of Prague,

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

A sample sentence

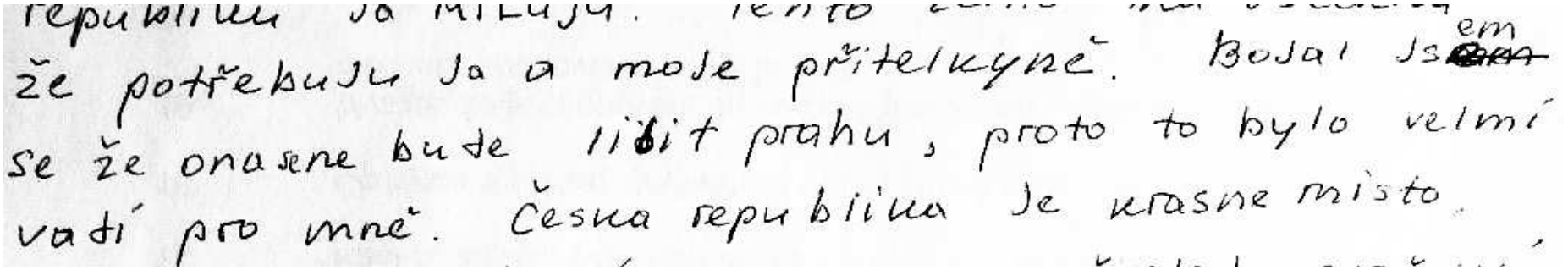


B~~oja~~á~~l~~ js~~me~~em se,
feared AUX-SG RFL
I was afraid

že ~~ona~~jí se ne bude libila slavnou prahu,
that her RFL not will *like *famous *prague,
that she would not like the famous city of Prague,

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

A sample sentence

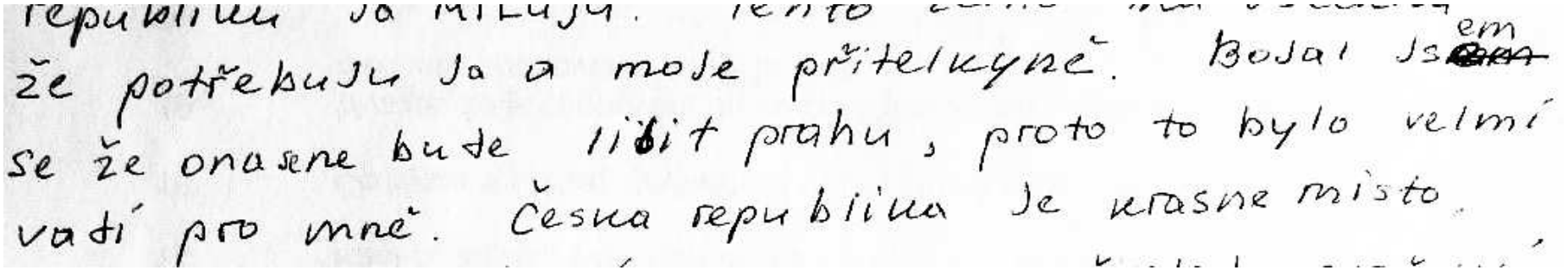


B~~oja~~á~~l~~ js~~me~~em se,
feared AUX-SG RFL
I was afraid

že ~~ona~~jí se nebude libila slavnou prahu,
that her RFL not will *like *famous *prague,
that she would not like the famous city of Prague,

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

A sample sentence

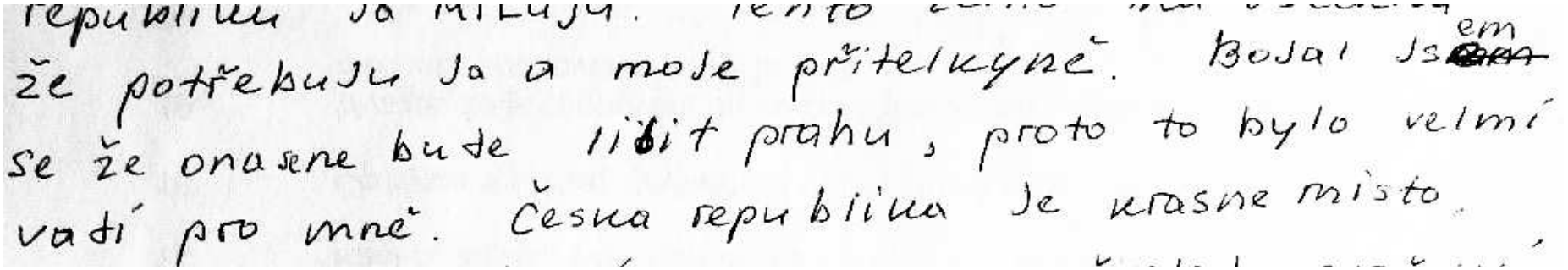


B~~oja~~á~~l~~ js~~me~~em se,
feared AUX-SG RFL
I was afraid

že ~~ona~~jí se nebude l~~í~~bi~~la~~t slavnou prahu,
that her RFL not will like *famous *prague,
that she would not like the famous city of Prague,

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

A sample sentence

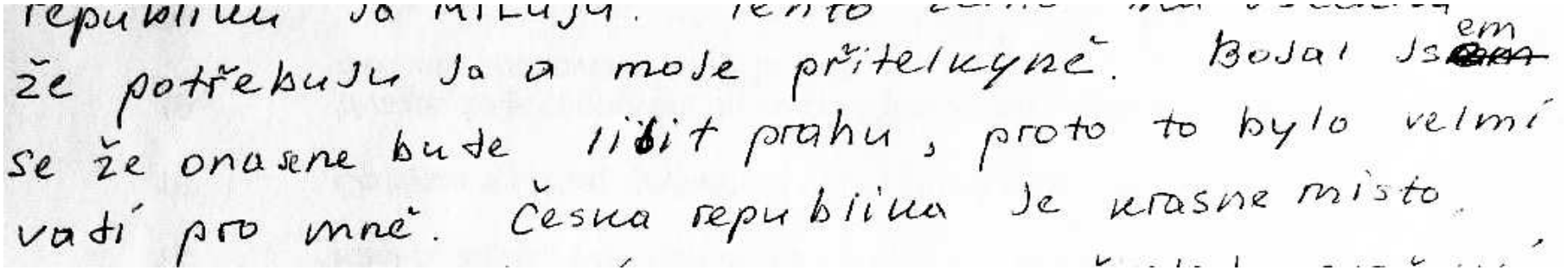


B~~oja~~á~~l~~ js~~me~~em se,
feared AUX-SG RFL
I was afraid

že ~~ona~~jí se nebude l~~í~~bi~~la~~t slavn~~ou~~á prahu,
that her RFL not will like famous *prague,
that she would not like the famous city of Prague,

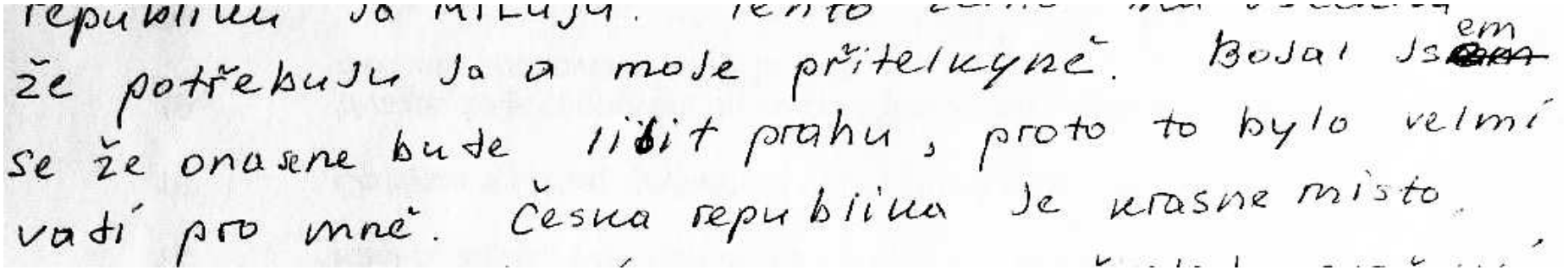
proto to bylo velmí vadí pro mně .
therefore it was very resent for me.
because I would be very unhappy about it.

A sample sentence



B~~oja~~á[l]	js~~me~~em	se,					
feared	AUX-SG	RFL					
I was afraid							
že	~~ona~~jí	se	nebude	l~~i~~íbi~~la~~t	slavn~~ou~~á	~~p~~Prah~~u~~a,	
that	her	RFL	not will	like	famous	prague,	
that she would not like the famous city of Prague,							
protože	to	bylo	velmí	vadí	pro mně .		
therefore	it	was	very	resent	for me.		
because I would be very unhappy about it.							

A sample sentence

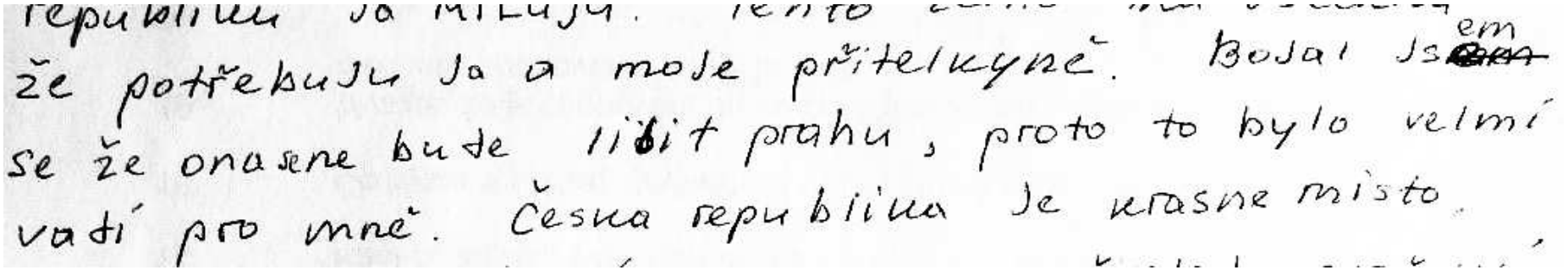


B~~oja~~á~~l~~ js~~me~~em se,
feared AUX-SG RFL
I was afraid

že ~~ona~~jí se nebude l~~í~~bi~~la~~t slavn~~ou~~á ~~p~~Prah~~u~~a,
that her RFL not will like famous prague,
that she would not like the famous city of Prague,

protože to by~~lo~~ velmí vadí pro mně .
therefore it was very resent me.
because I would be very unhappy about it.

A sample sentence

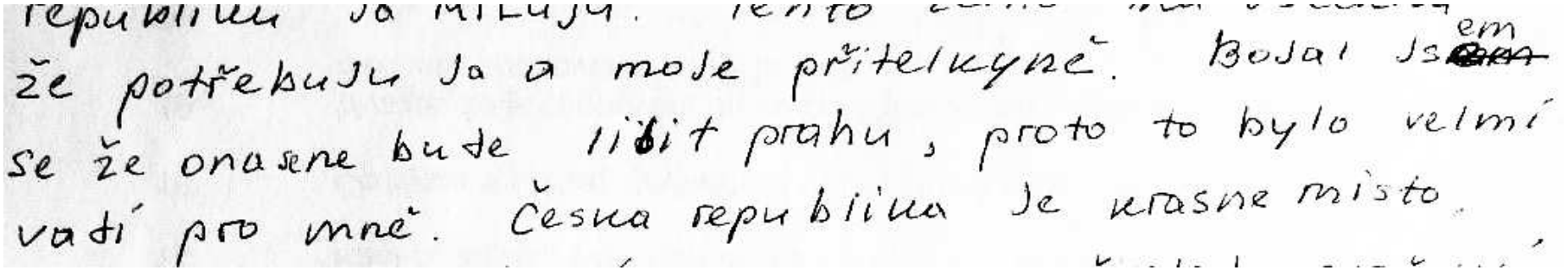


B~~ojaá~~l js~~me~~em se,
feared AUX-SG RFL
I was afraid

že ~~ona~~jí se nebude l~~í~~bi~~la~~t slavn~~ou~~á ~~p~~Prah~~u~~a,
that her RFL not will like famous prague,
that she would not like the famous city of Prague,

protože to by~~lo~~ velm~~í~~i vadí pro mně .
therefore it would very resented me.
because I would be very unhappy about it.

A sample sentence

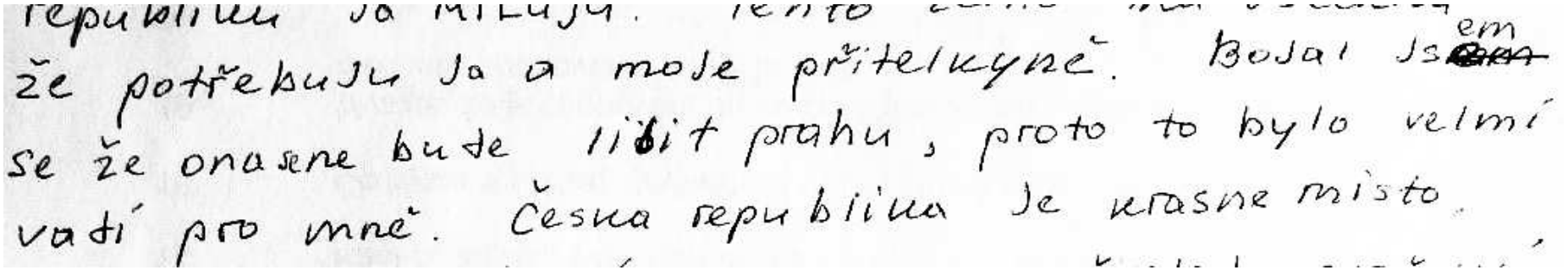


B~~oja~~á~~m~~ js~~me~~em se,
feared AUX-SG RFL
I was afraid

že ~~ona~~jí se nebude l~~í~~bi~~la~~t slavn~~ou~~á ~~p~~Prah~~u~~a,
that her RFL not will like famous prague,
that she would not like the famous city of Prague,

protože to by~~lo~~ velm~~í~~i vad~~í~~ilo pro mně .
therefore it would very resent me.
because I would be very unhappy about it.

A sample sentence

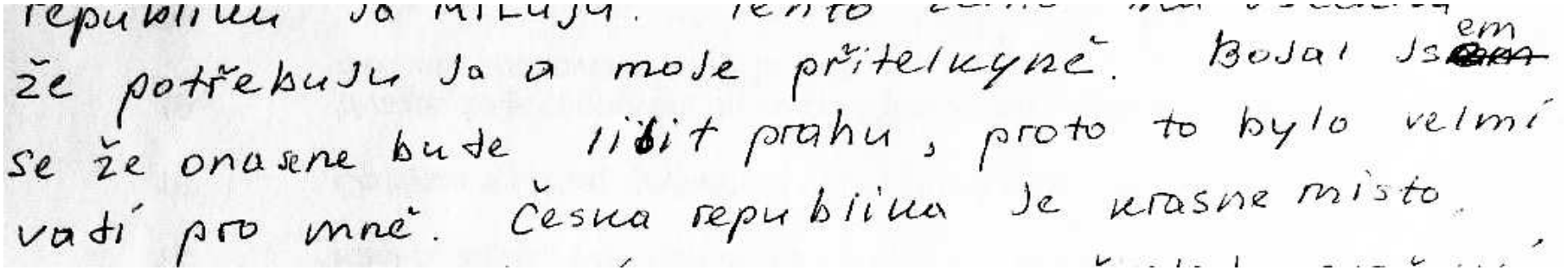


B~~oja~~á~~m~~ js~~me~~em se,
feared AUX-SG RFL
I was afraid

že ~~ona~~jí se nebude l~~í~~bi~~la~~t slavn~~ou~~á ~~p~~Prah~~u~~a,
that her RFL not will like famous prague,
that she would not like the famous city of Prague,

protože to by~~lo~~ velm~~í~~i vad~~í~~ilo pro m~~n~~ě .
therefore it would very resent me.
because I would be very unhappy about it.

A sample sentence



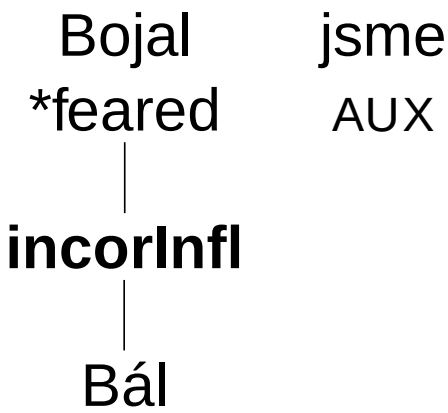
B~~oja~~á<u>á</u>l	js~~me~~<u>em</u>	<u>se,</u>					
feared	AUX-SG	RFL					
I was afraid							
že	~~ona~~<u>jí</u>	se	<u>nebude</u>	l~~í~~<u>í</u>bi~~la~~<u>t</u>	slavn~~ou~~<u>á</u>	~~p~~<u>P</u>rah~~u~~<u>a</u>,	
that	her	RFL	not will	like	famous	prague,	
that she would not like the famous city of Prague,							
proto<u>že</u>	to	by~~lo~~	velm~~í~~<u>í</u>	vad~~í~~<u>ílo</u>	mi.		
therefore	it	would	very	resent	me.		
because I would be very unhappy about it.							

Annotation of a sample sentence, part I

Bojal	jsme
*feared	AUX

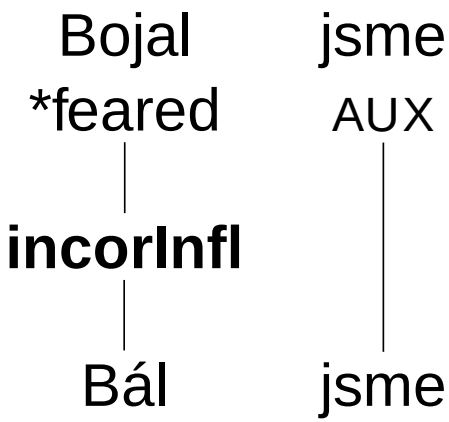
I was afraid

Annotation of a sample sentence, part I



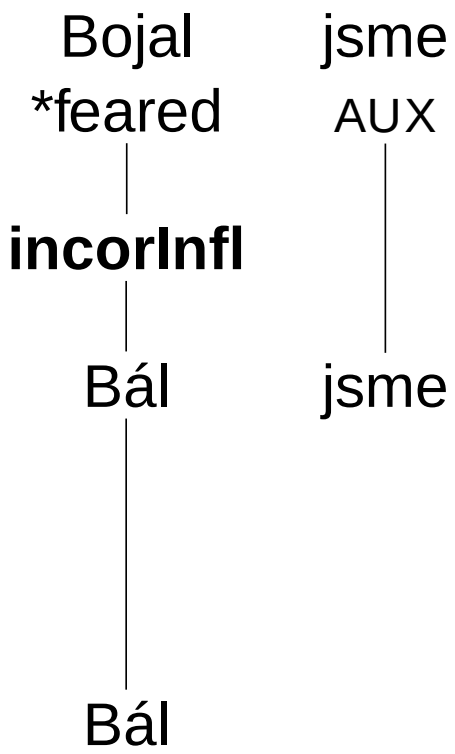
I was afraid

Annotation of a sample sentence, part I



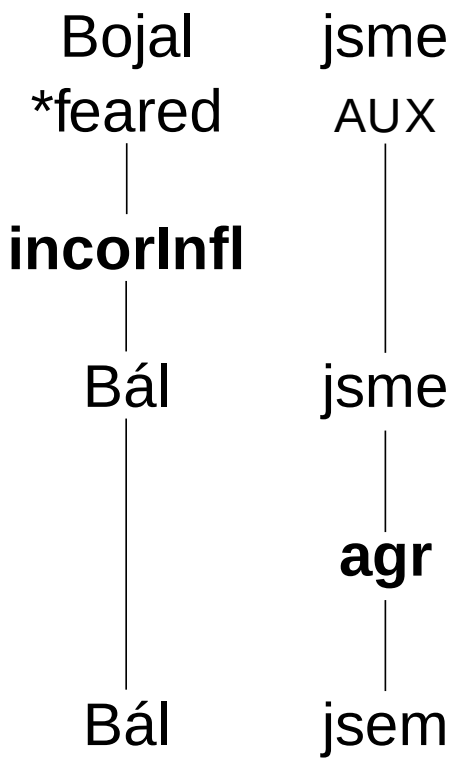
I was afraid

Annotation of a sample sentence, part I



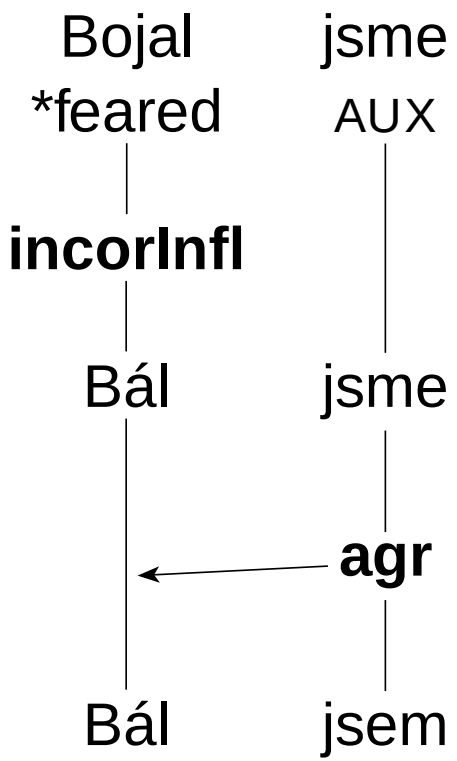
I was afraid

Annotation of a sample sentence, part I



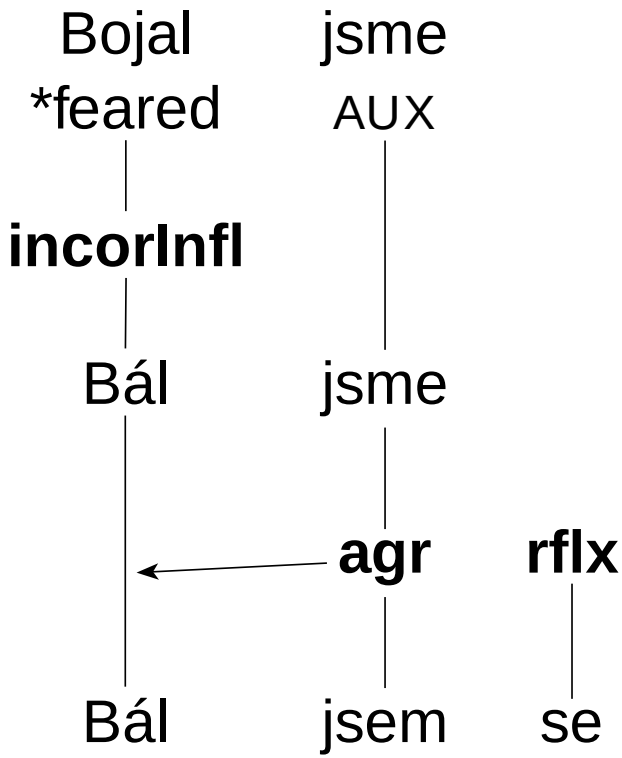
I was afraid

Annotation of a sample sentence, part I



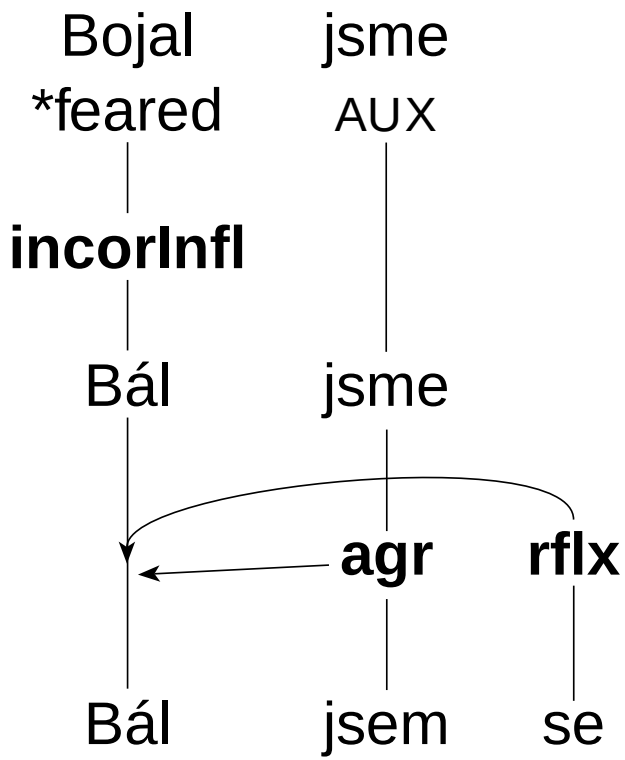
I was afraid

Annotation of a sample sentence, part I



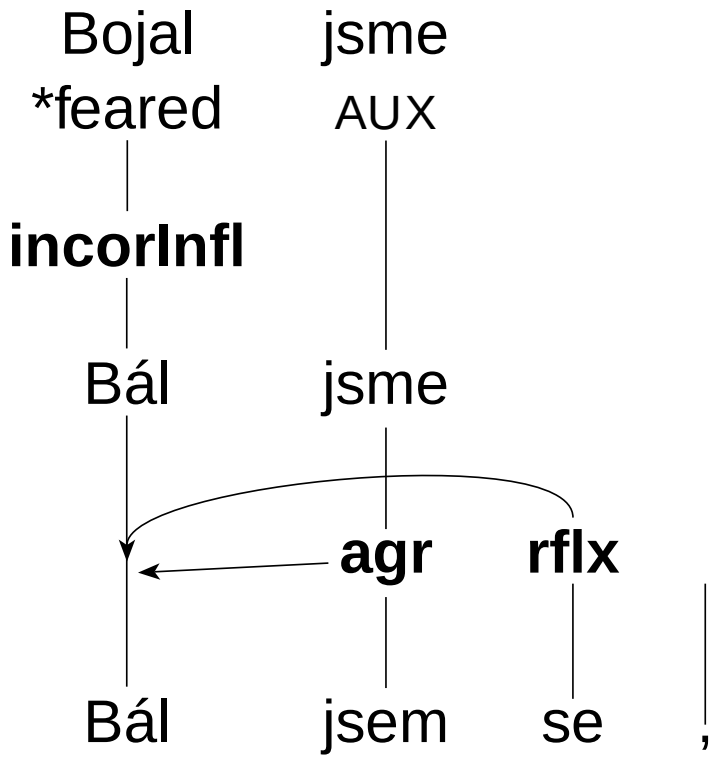
I was afraid

Annotation of a sample sentence, part I



I was afraid

Annotation of a sample sentence, part I



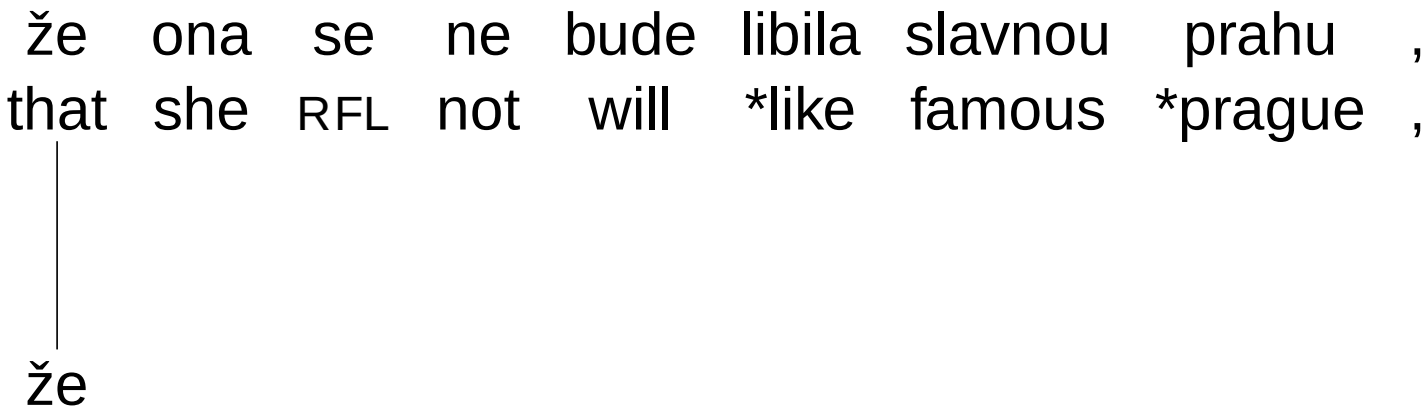
I was afraid

Annotation of a sample sentence, part II

že	ona	se	ne	bude	libila	slavnou	prahu	,
that	she	RFL	not	will	*like	famous	*prague	,

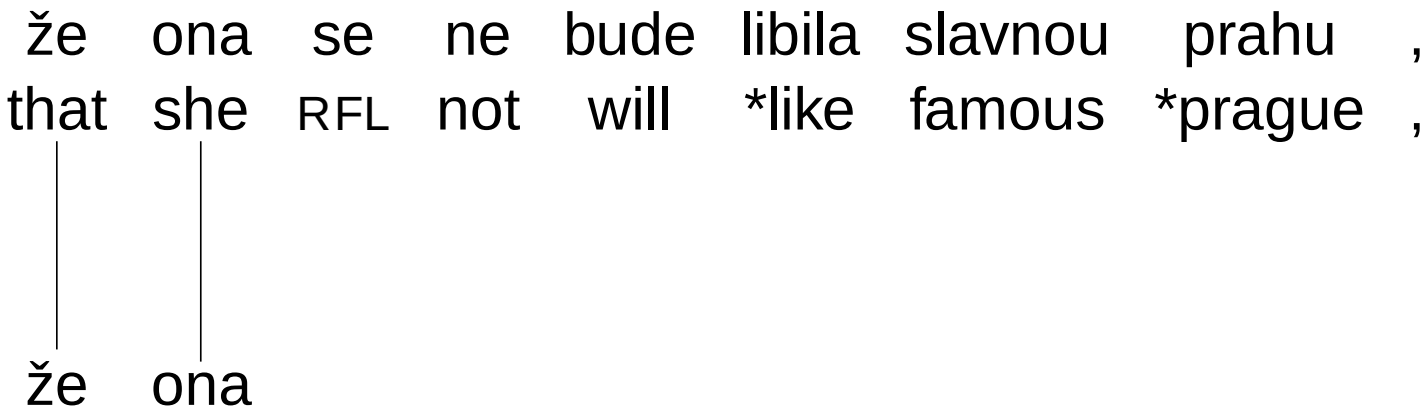
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



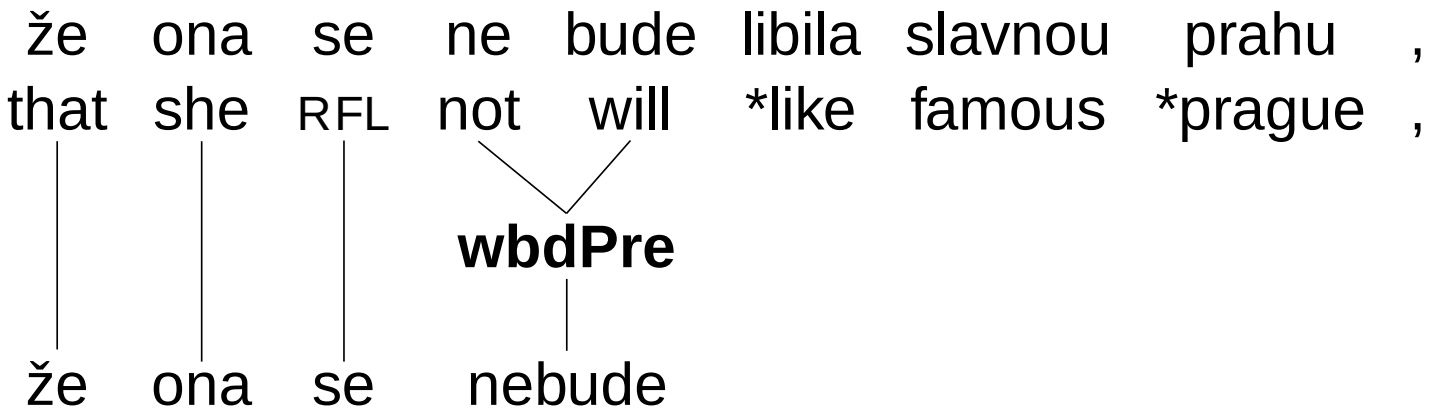
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II

že	ona	se	ne	bude	libila	slavnou	prahu	,
that	she	RFL	not	will	*like	famous	*prague	,
že	ona	se						

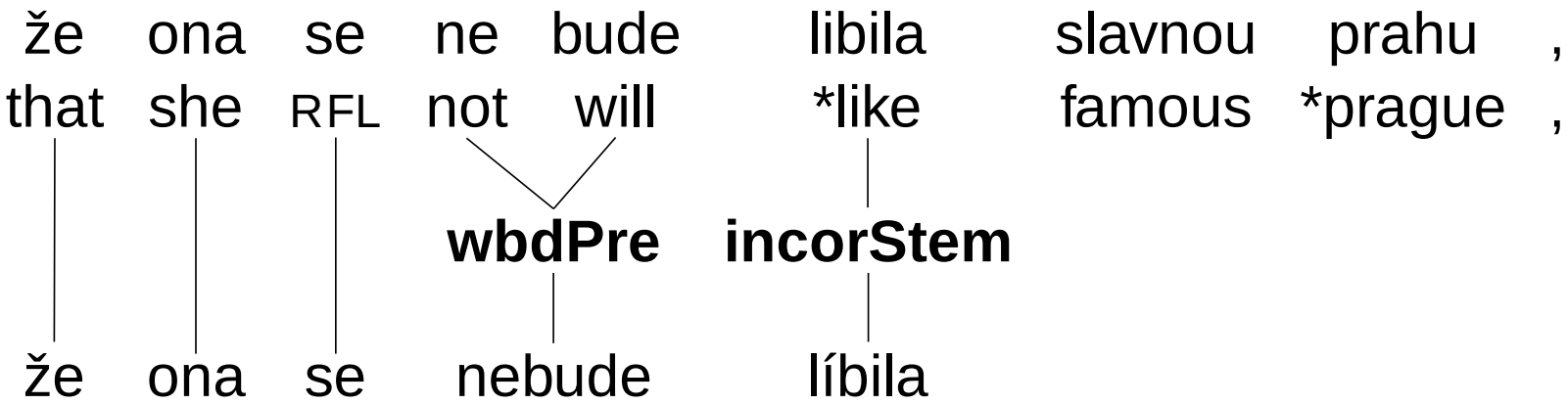
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



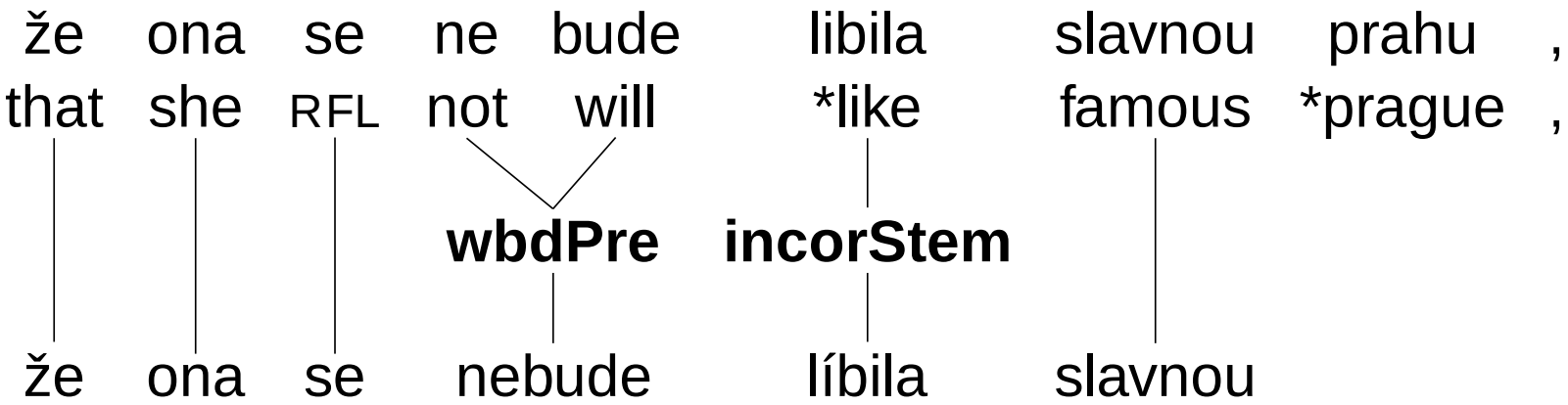
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



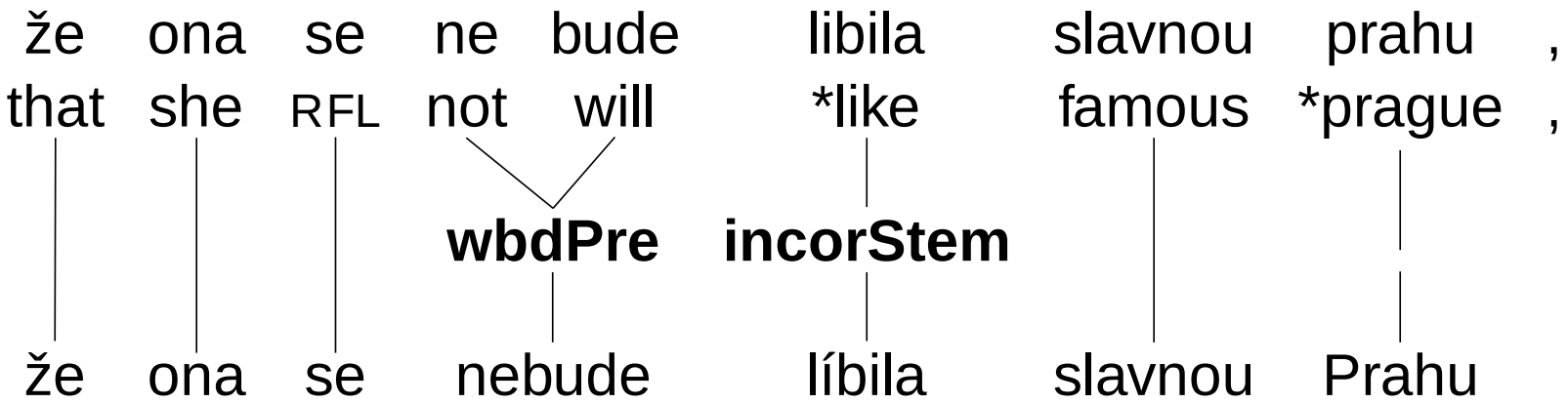
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



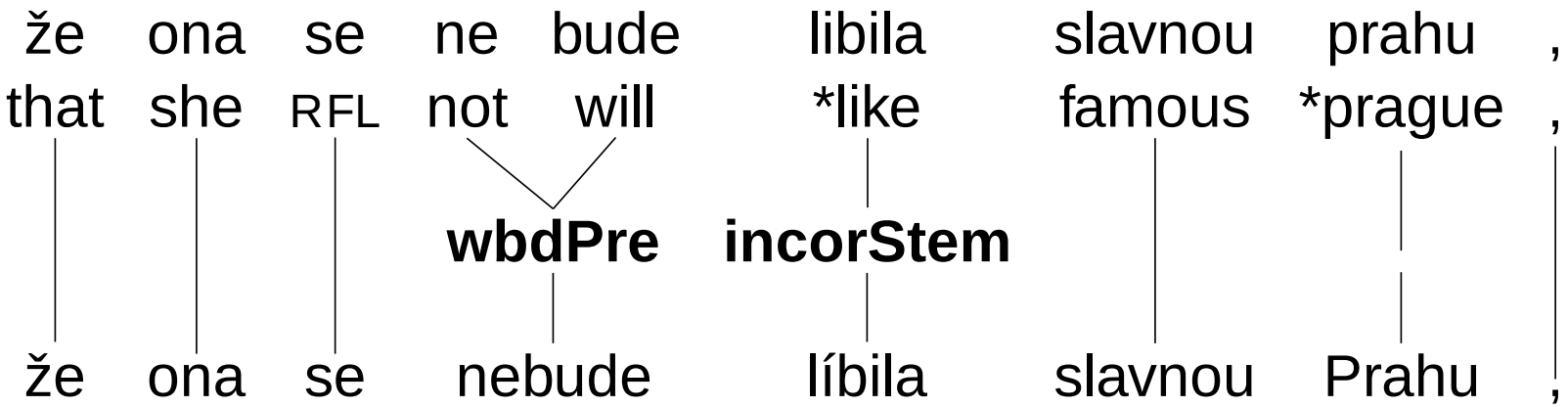
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



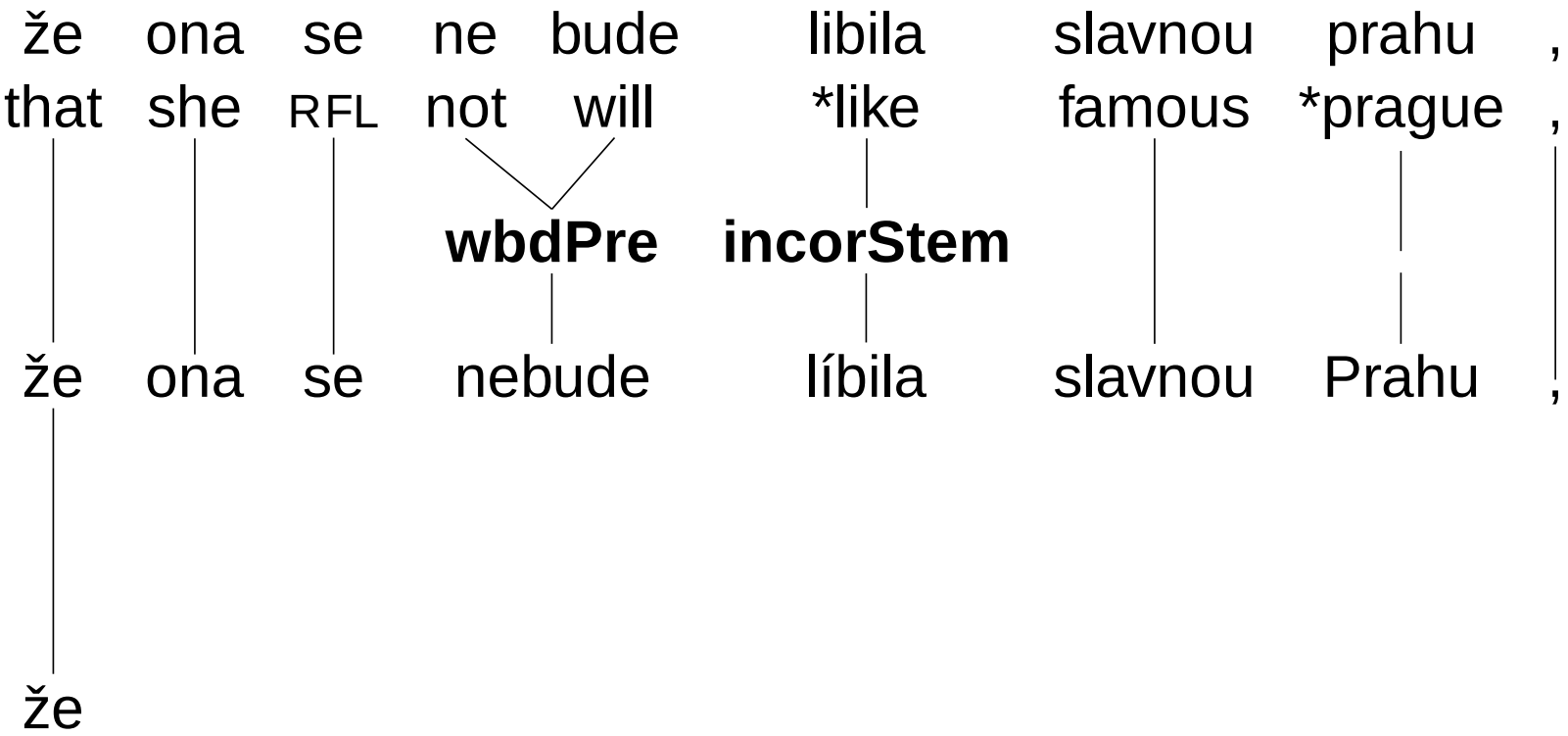
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



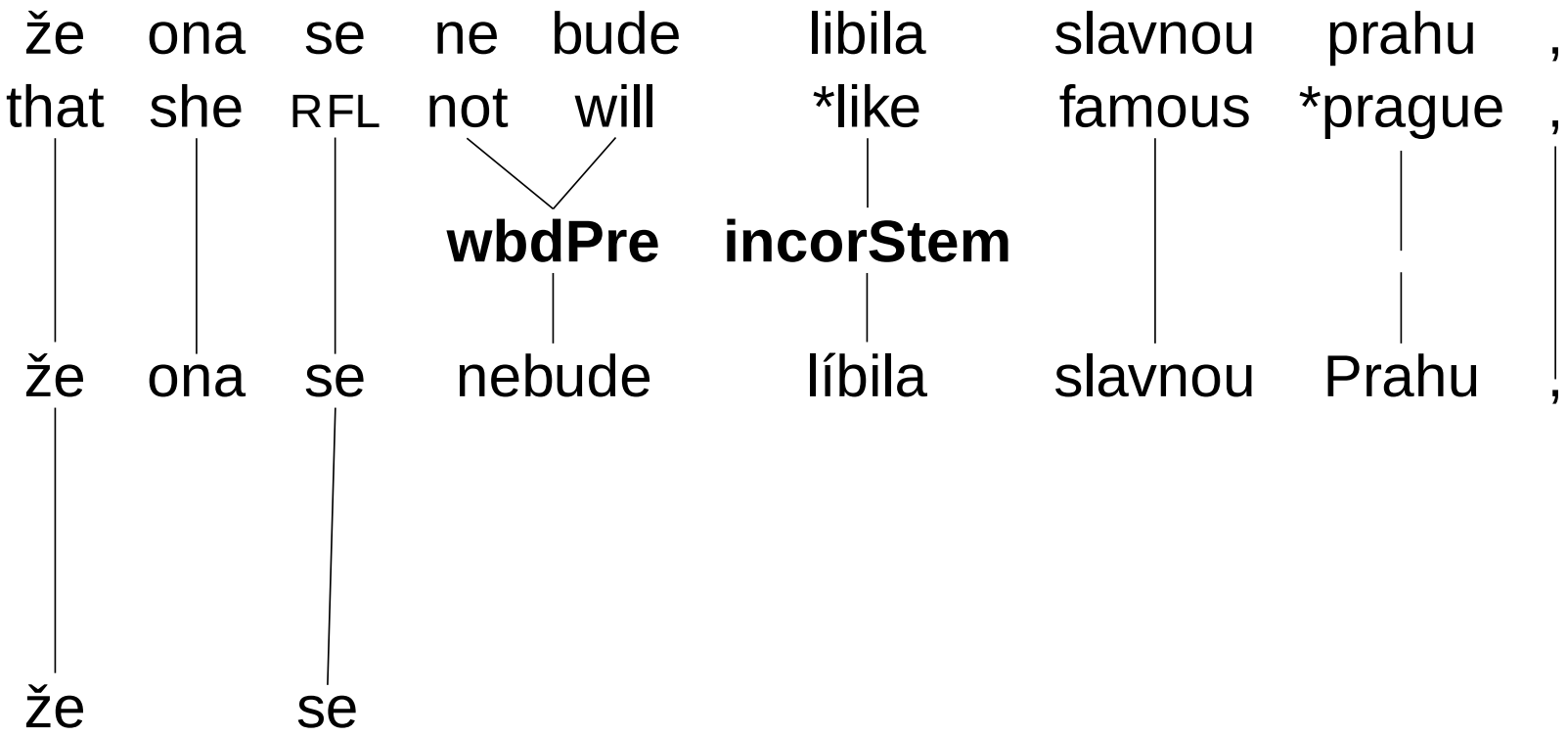
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



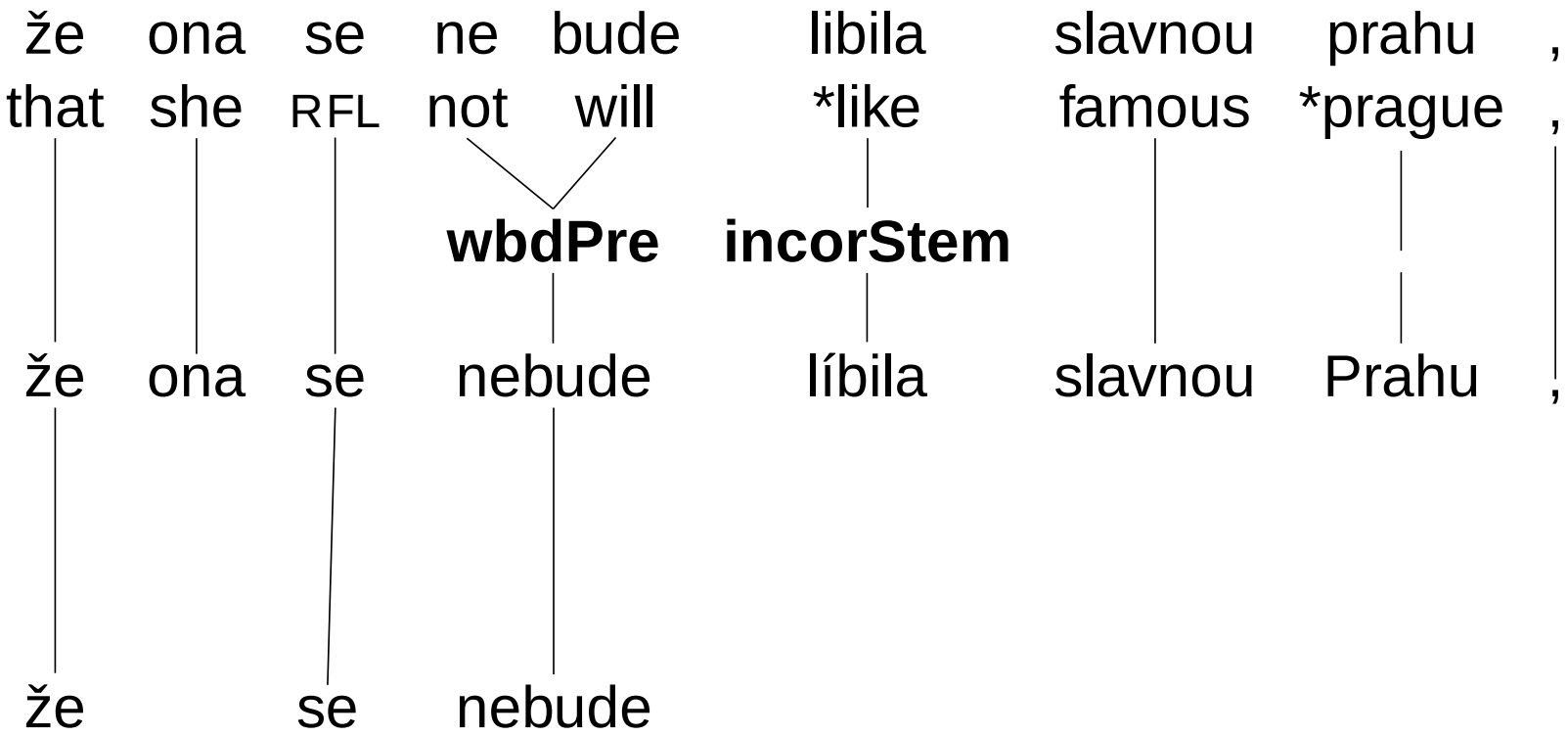
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



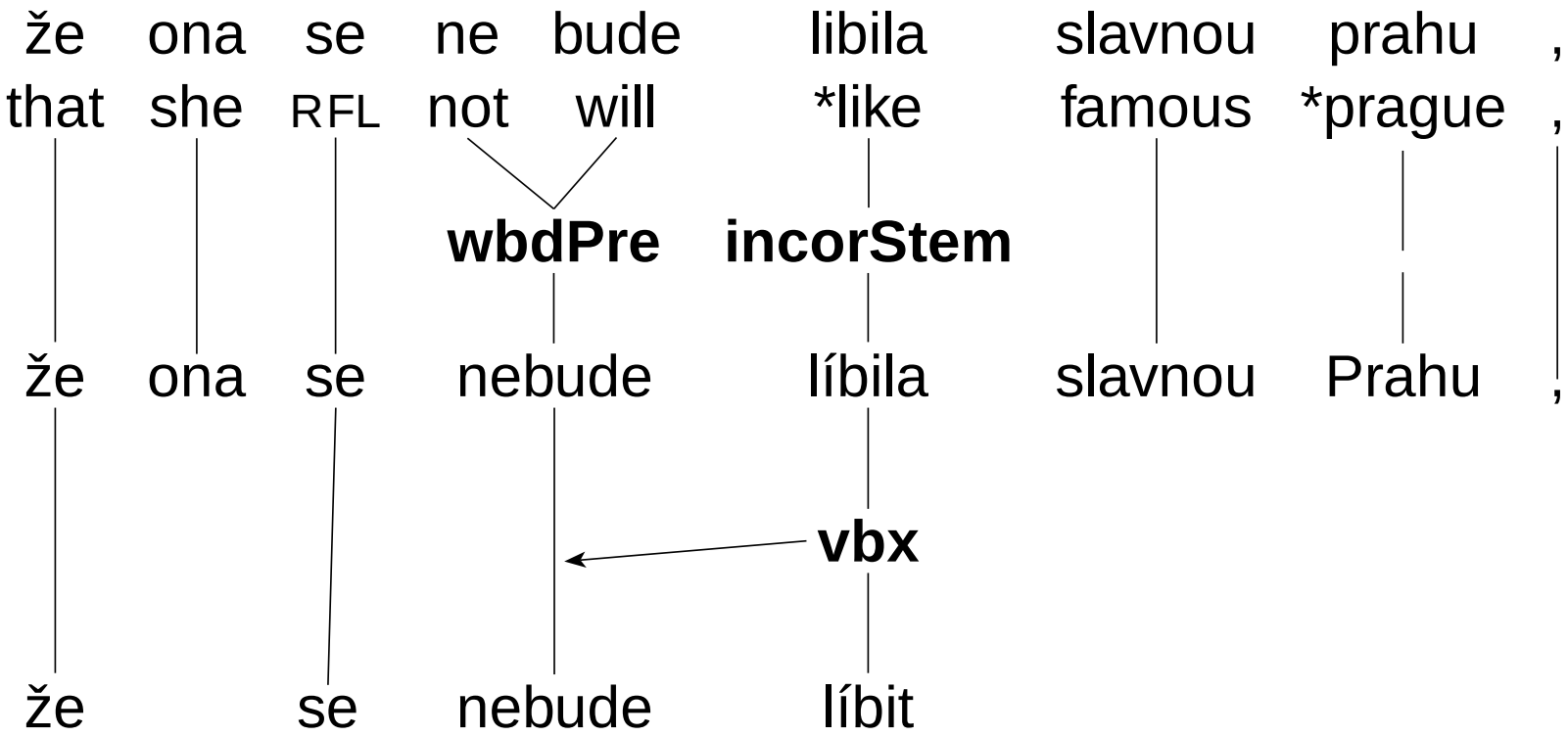
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



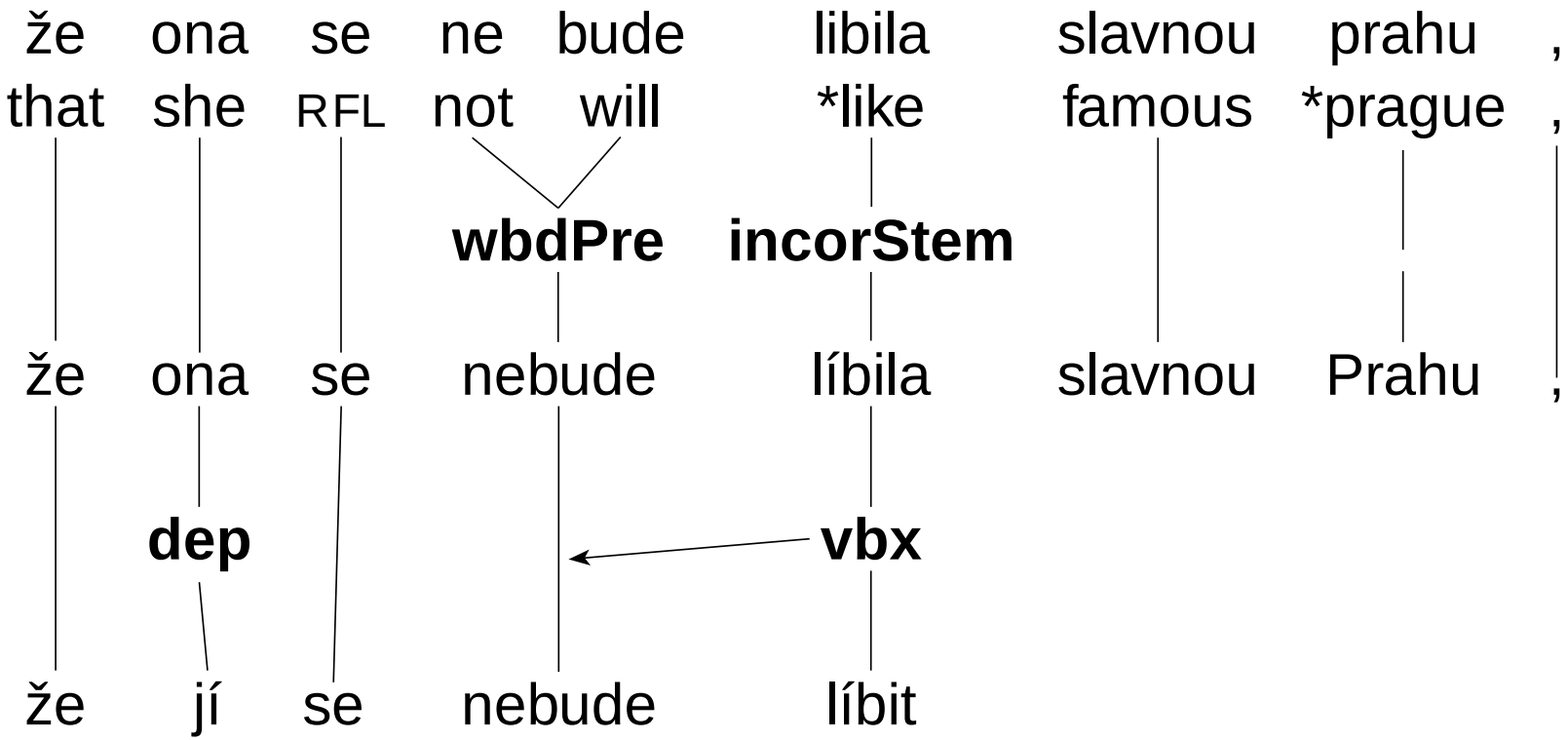
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



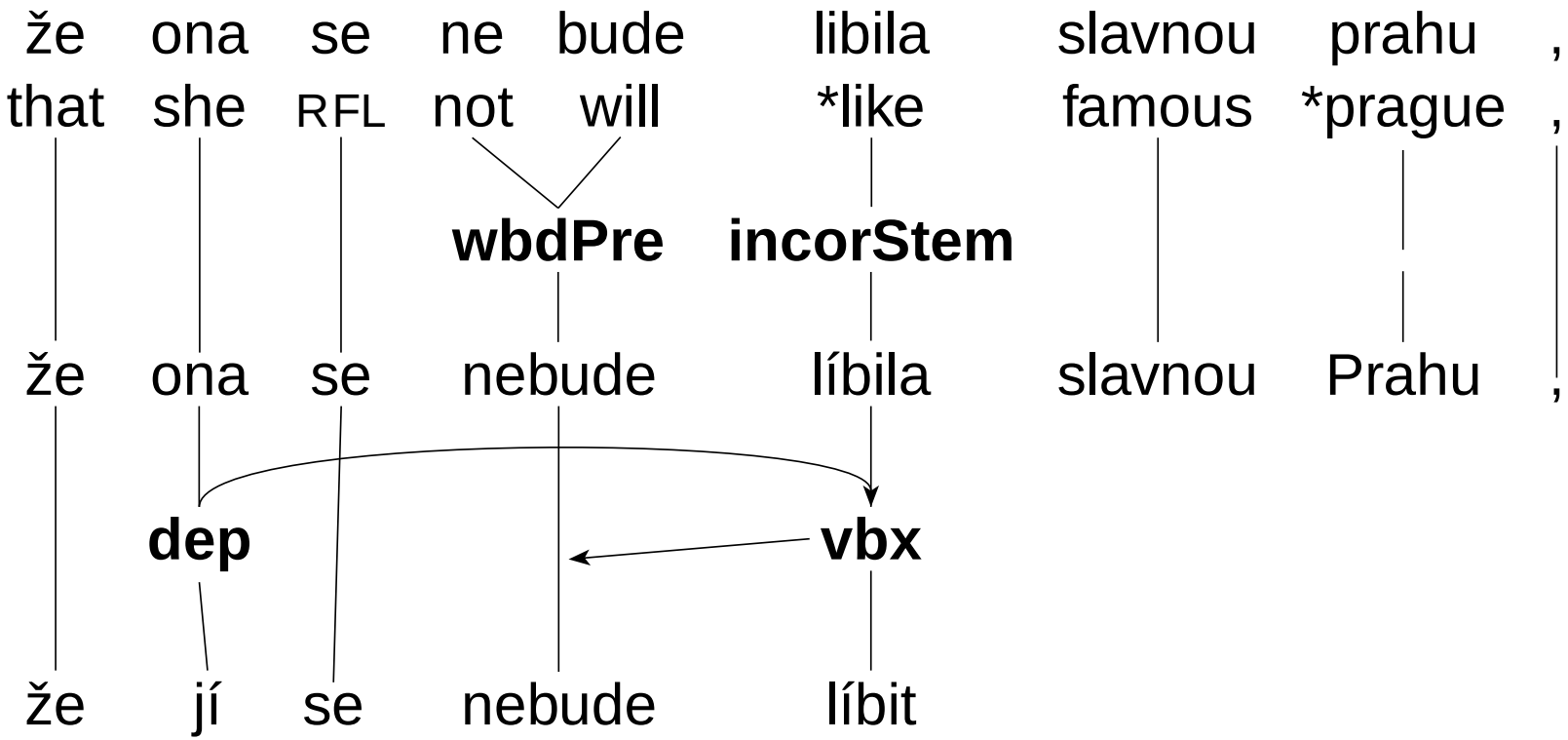
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



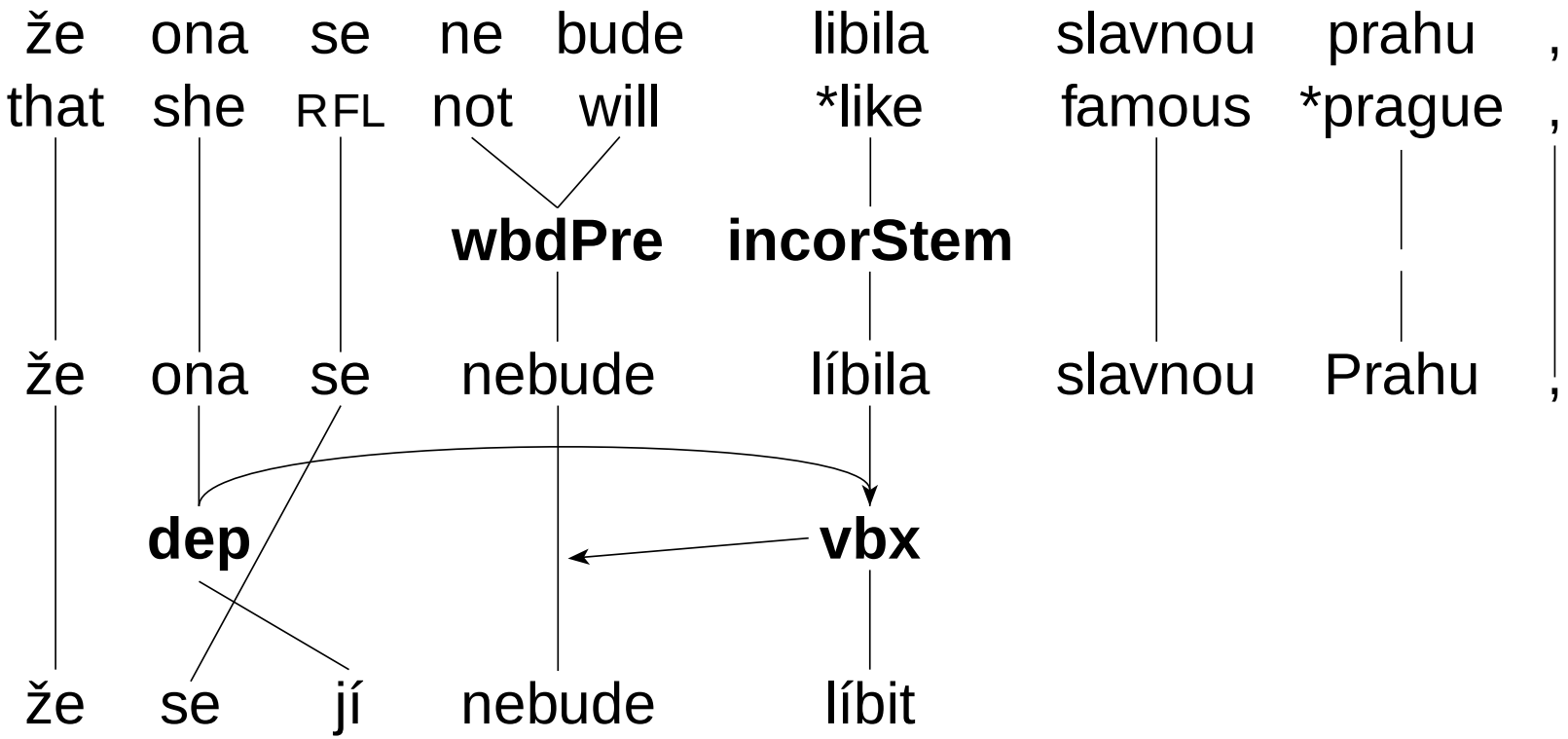
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



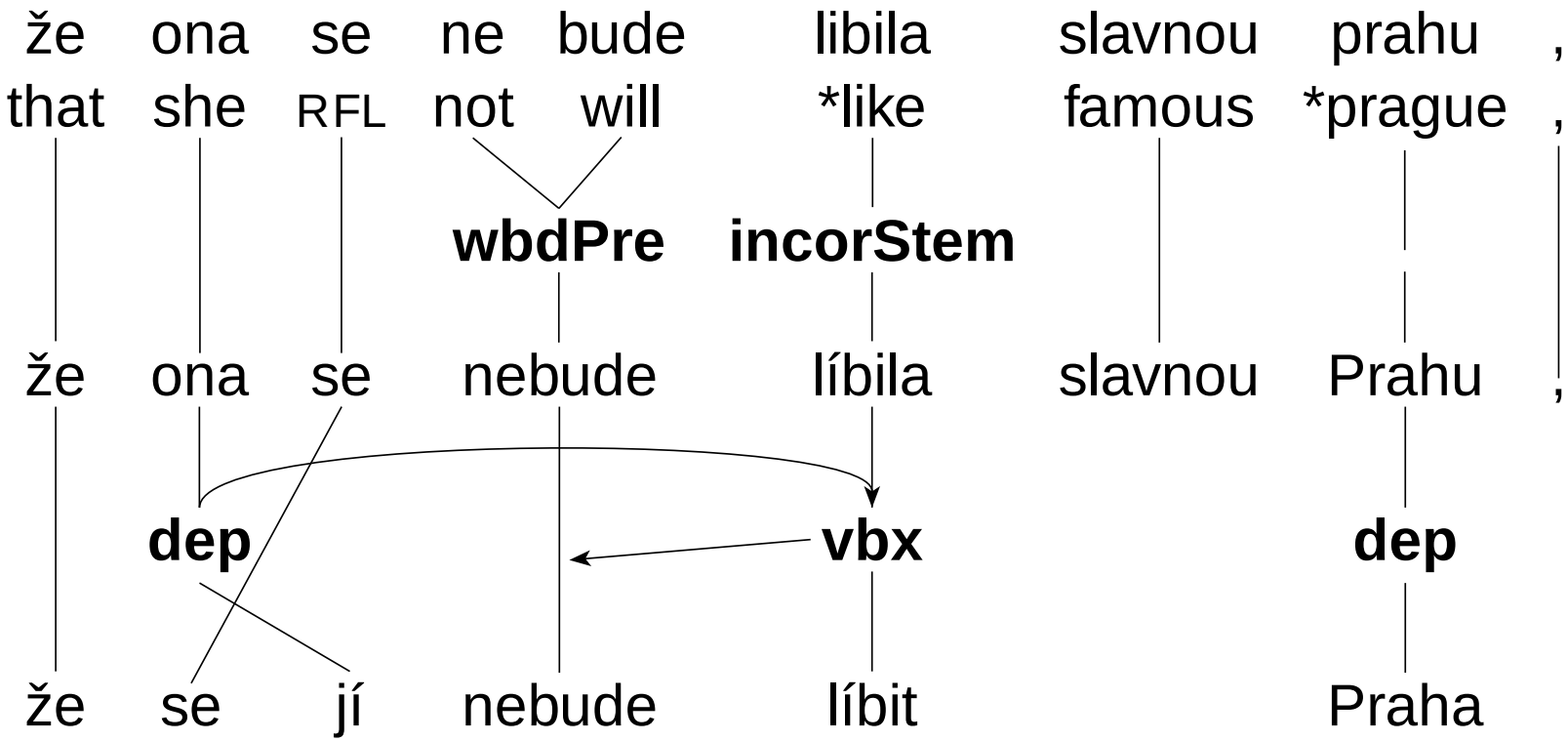
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



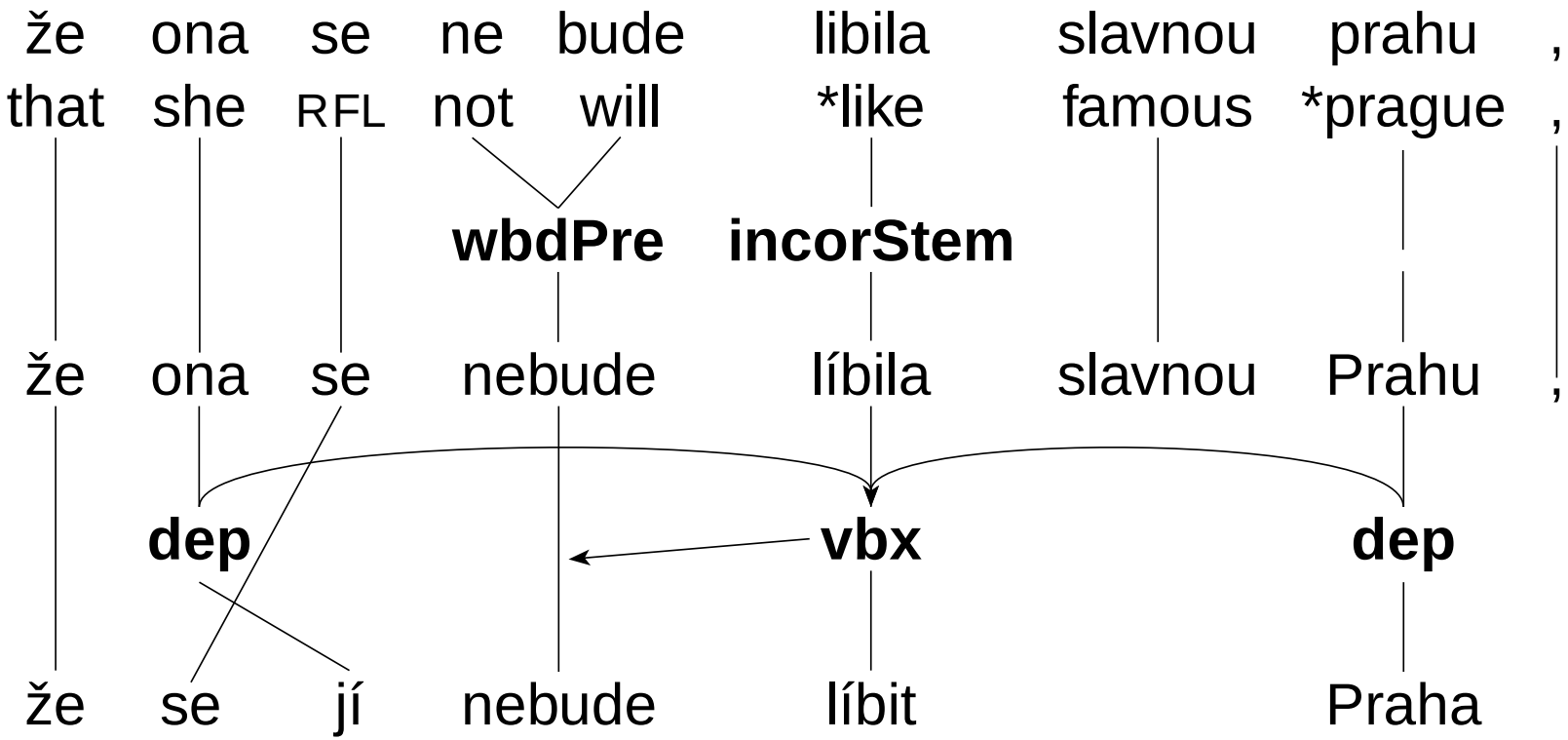
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



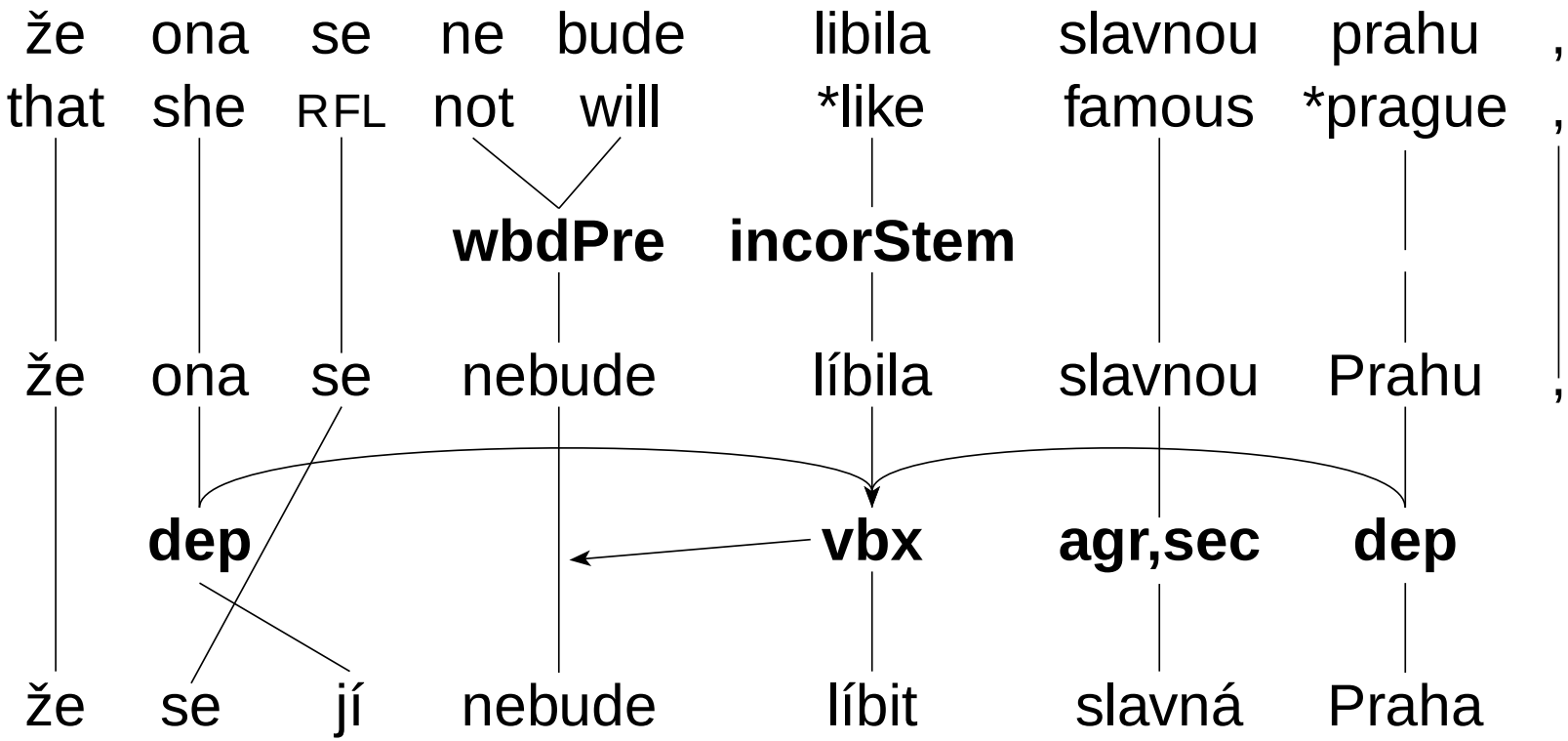
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



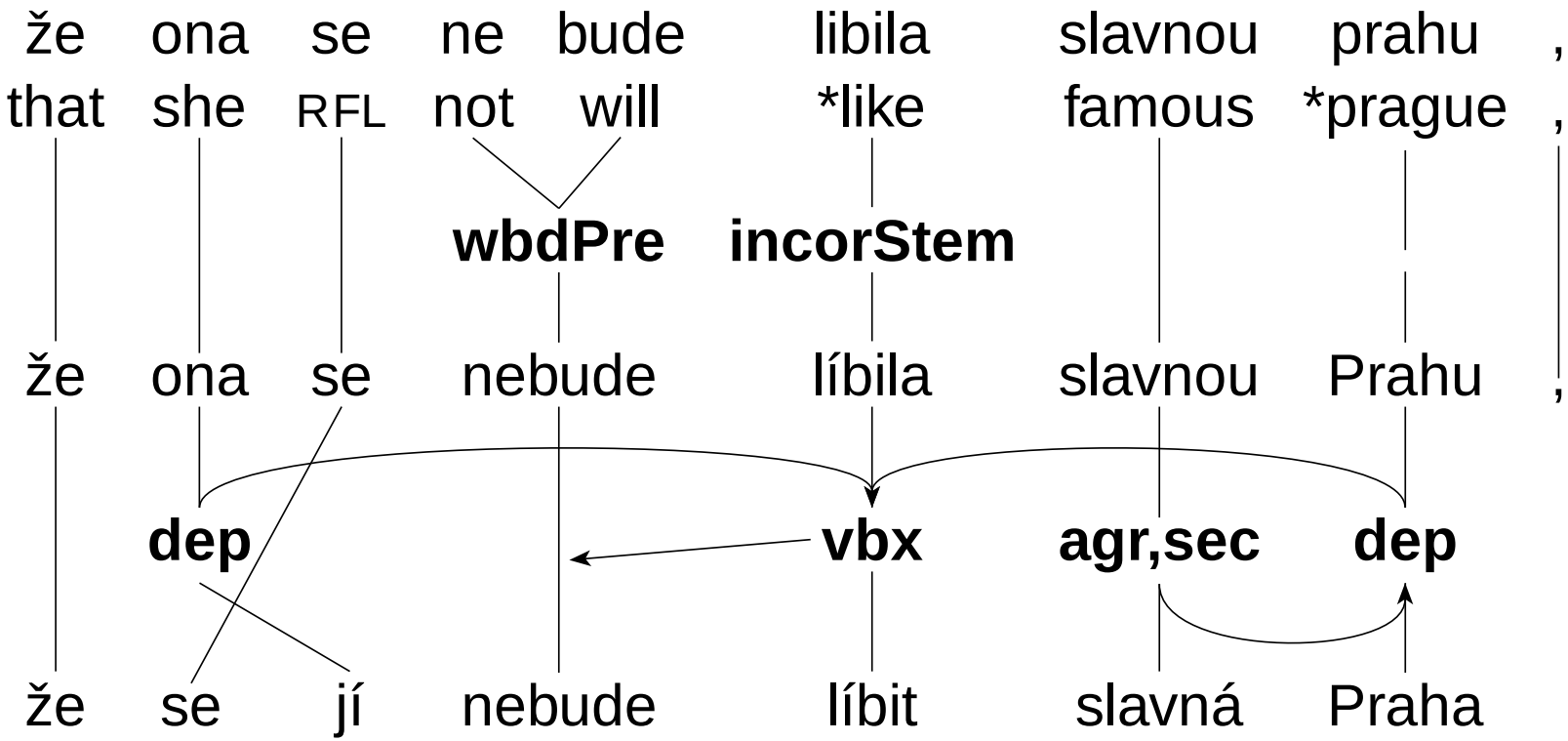
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



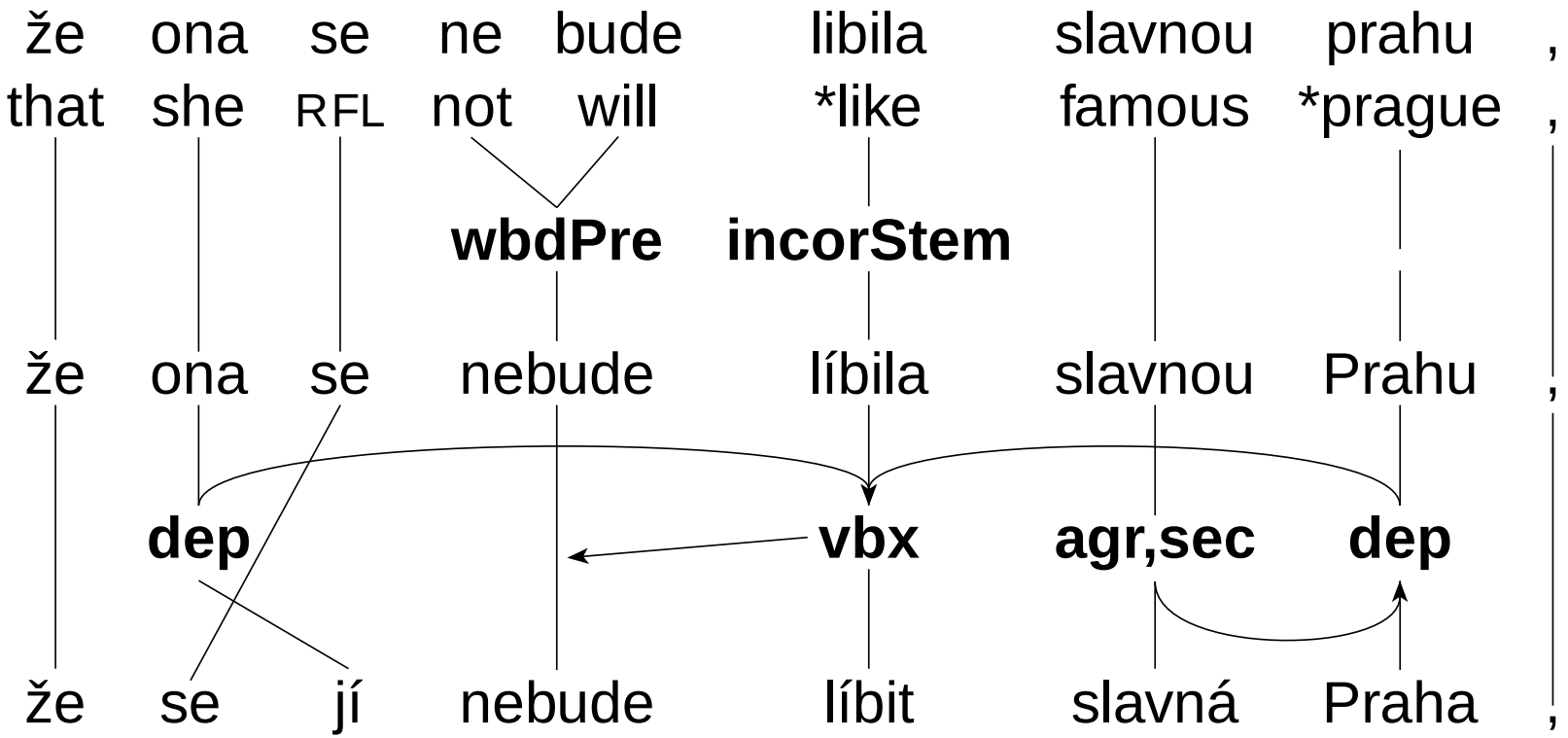
that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



that she would not like the famous city of Prague,

Annotation of a sample sentence, part II



that she would not like the famous city of Prague,

Annotation of a sample sentence, part II

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

proto

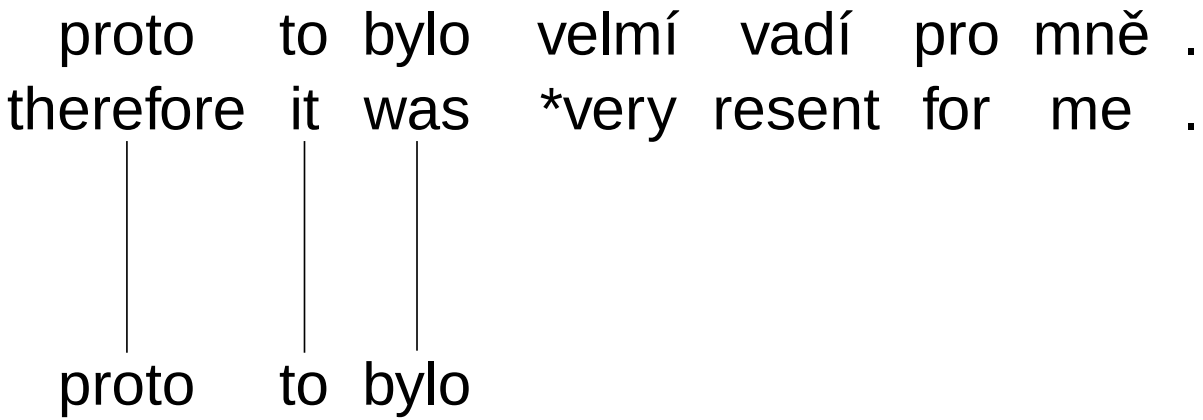
because I would be very unhappy about it.

Annotation of a sample sentence, part II

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

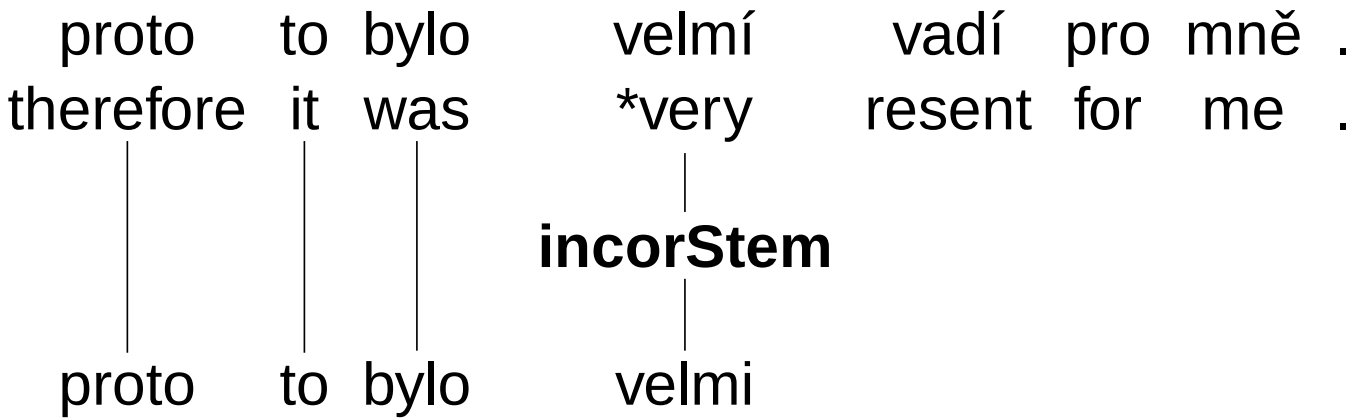
because I would be very unhappy about it.

Annotation of a sample sentence, part II



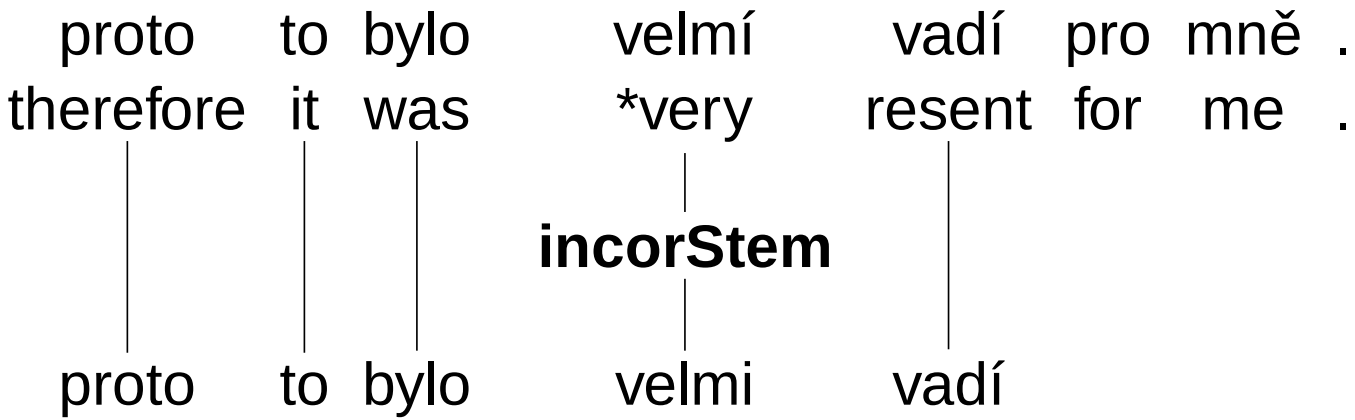
because I would be very unhappy about it.

Annotation of a sample sentence, part II



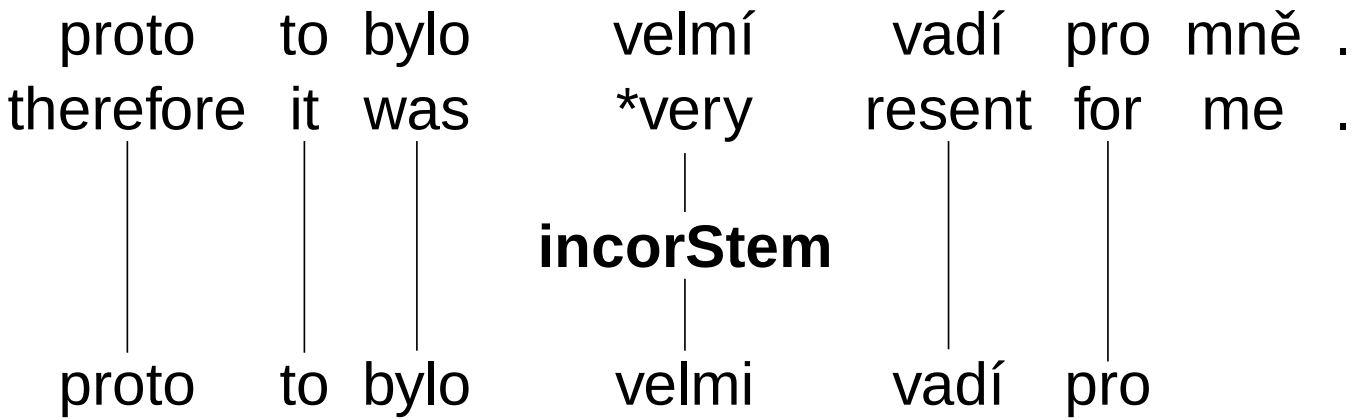
because I would be very unhappy about it.

Annotation of a sample sentence, part II



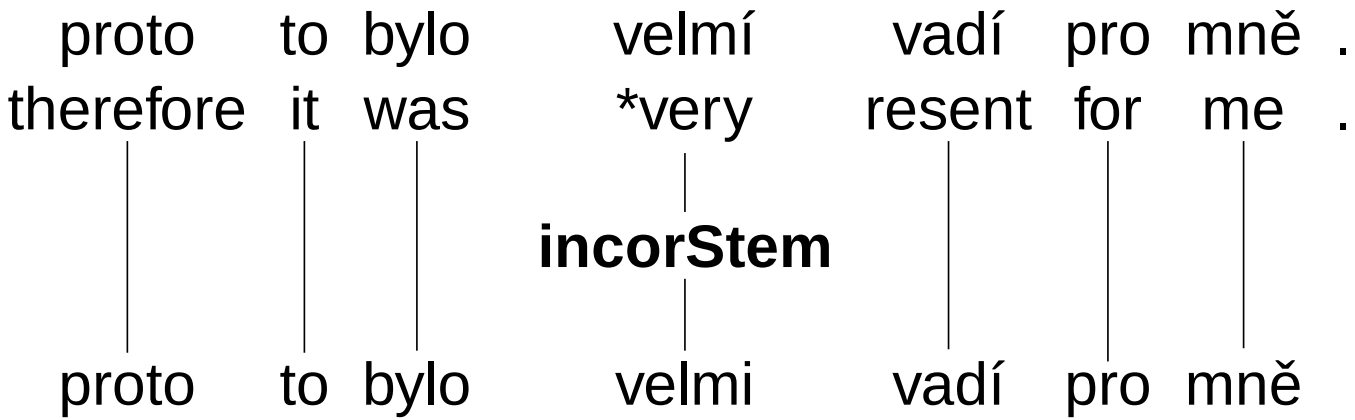
because I would be very unhappy about it.

Annotation of a sample sentence, part II



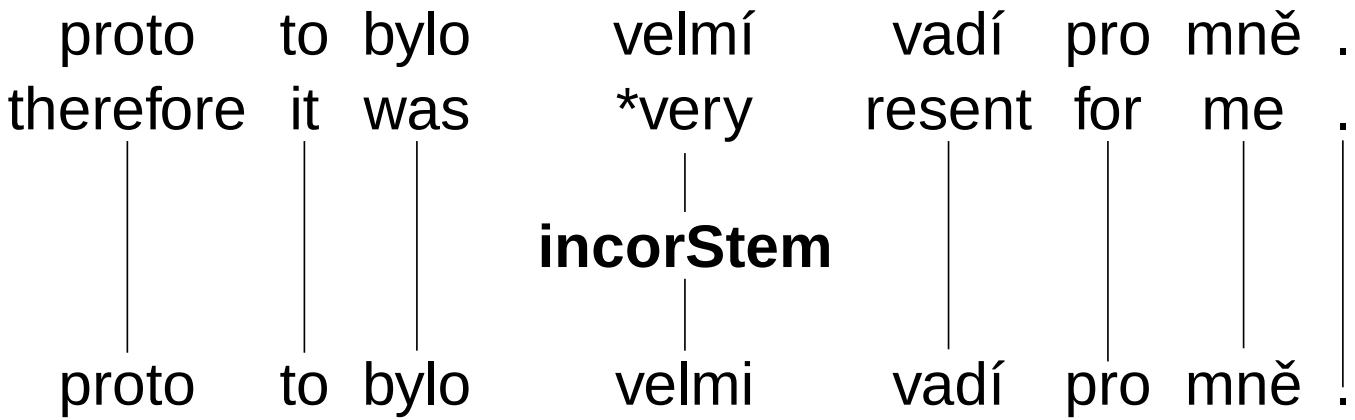
because I would be very unhappy about it.

Annotation of a sample sentence, part II



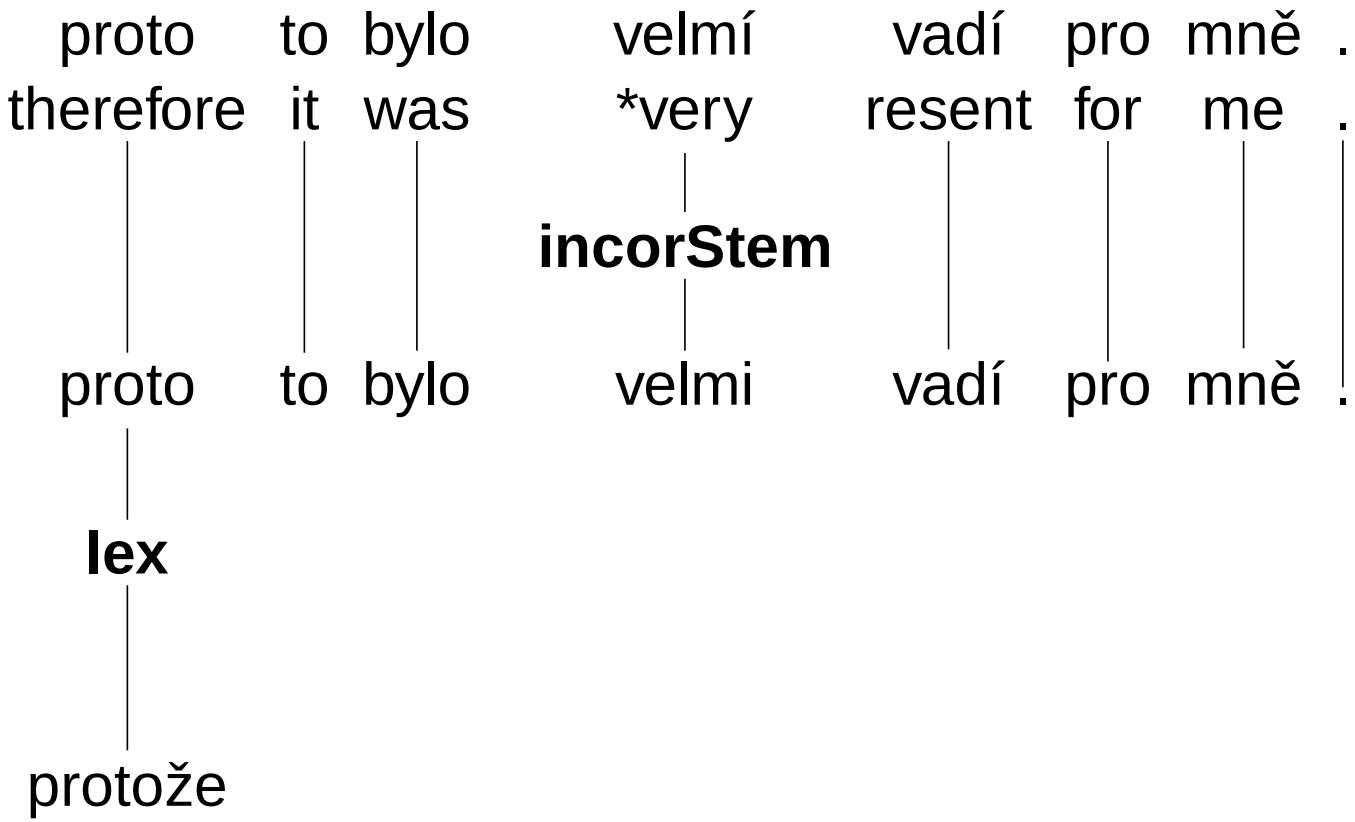
because I would be very unhappy about it.

Annotation of a sample sentence, part II



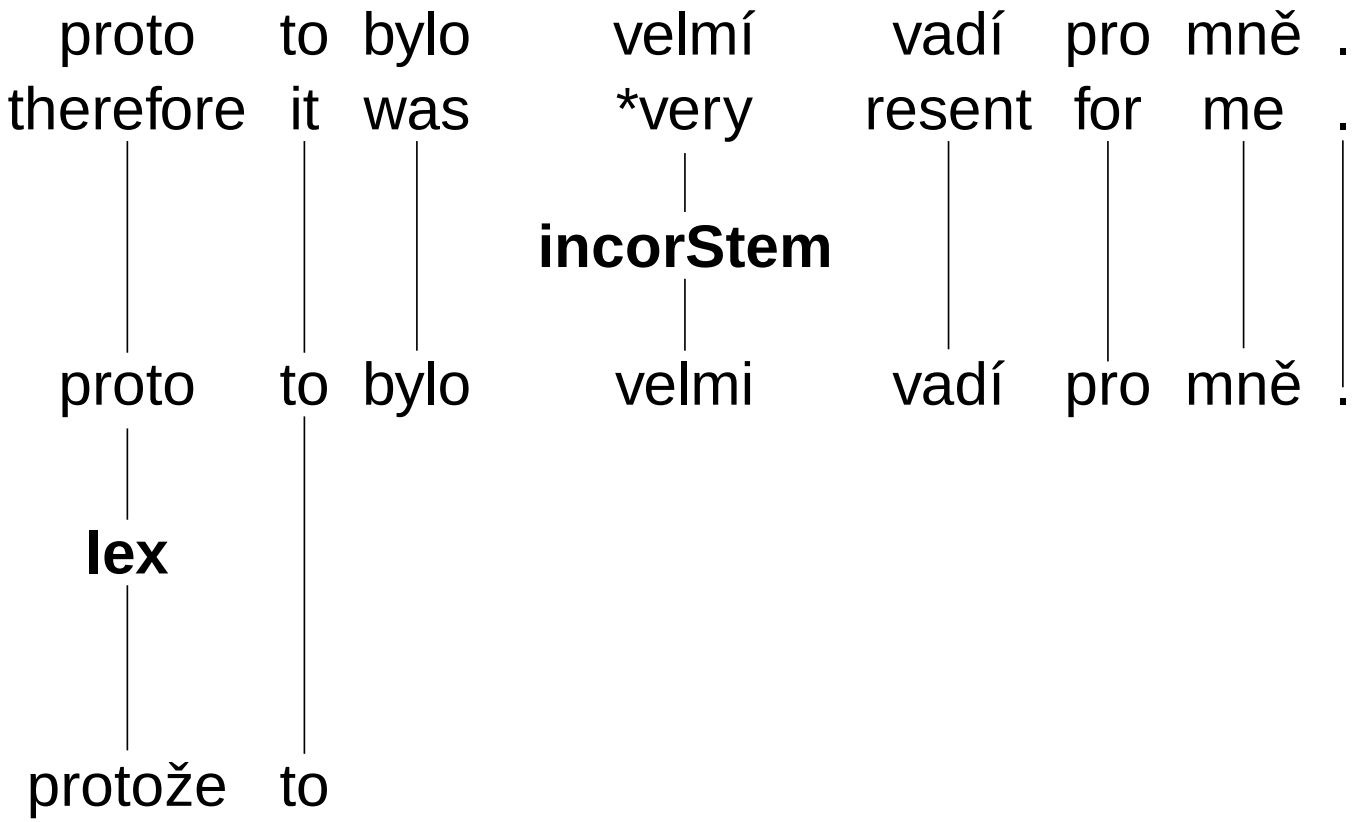
because I would be very unhappy about it.

Annotation of a sample sentence, part II



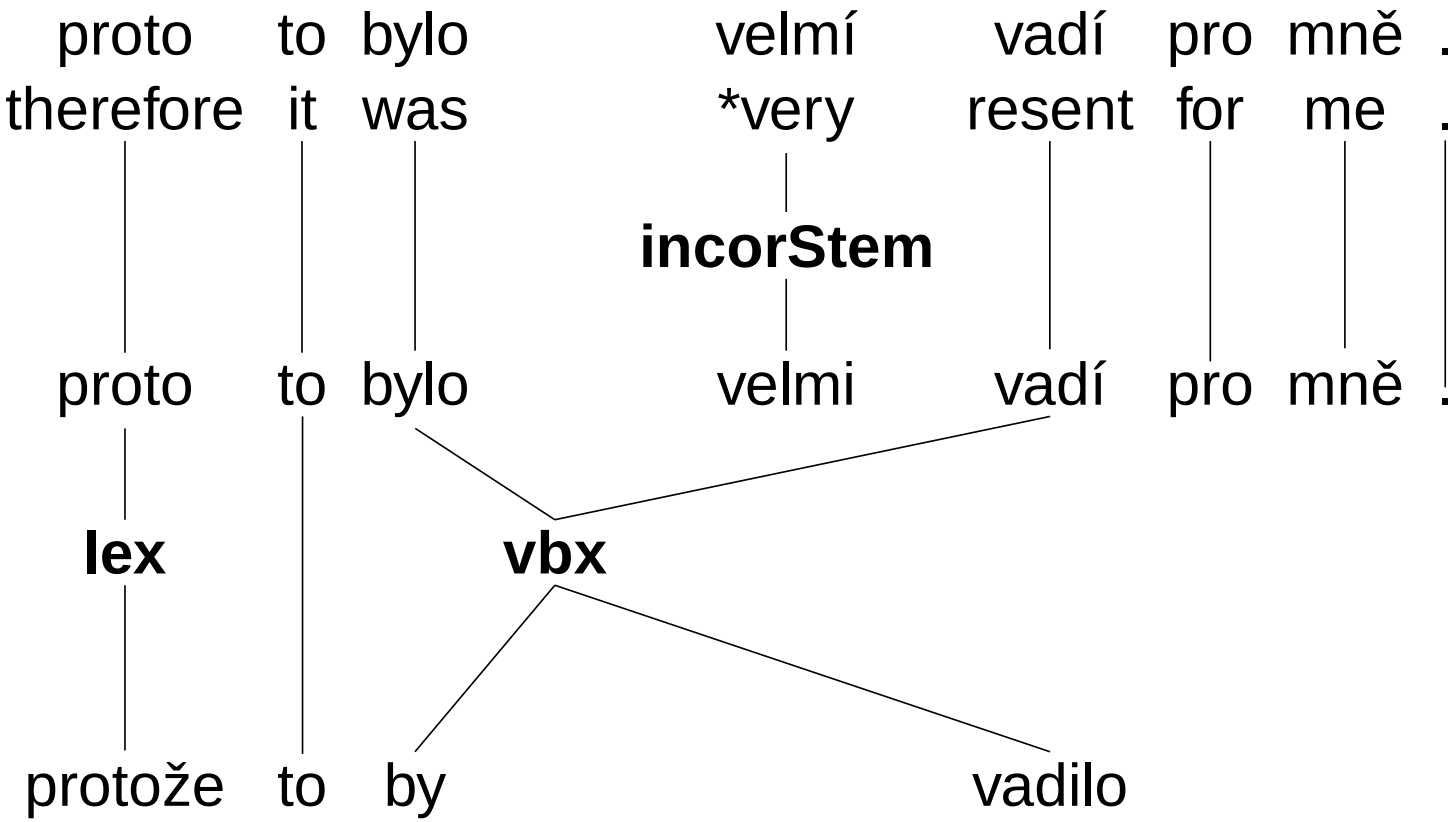
because I would be very unhappy about it.

Annotation of a sample sentence, part II



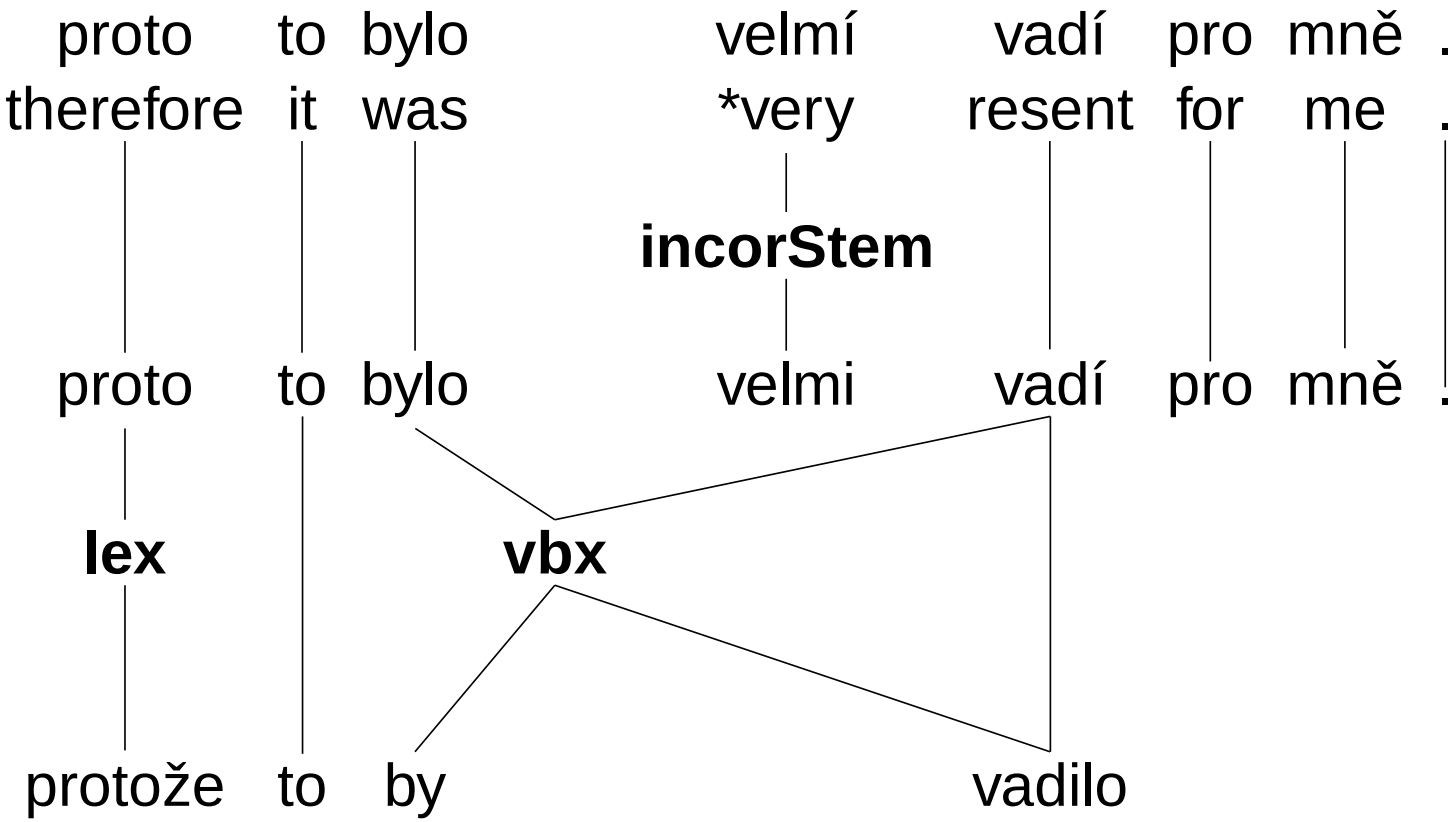
because I would be very unhappy about it.

Annotation of a sample sentence, part II



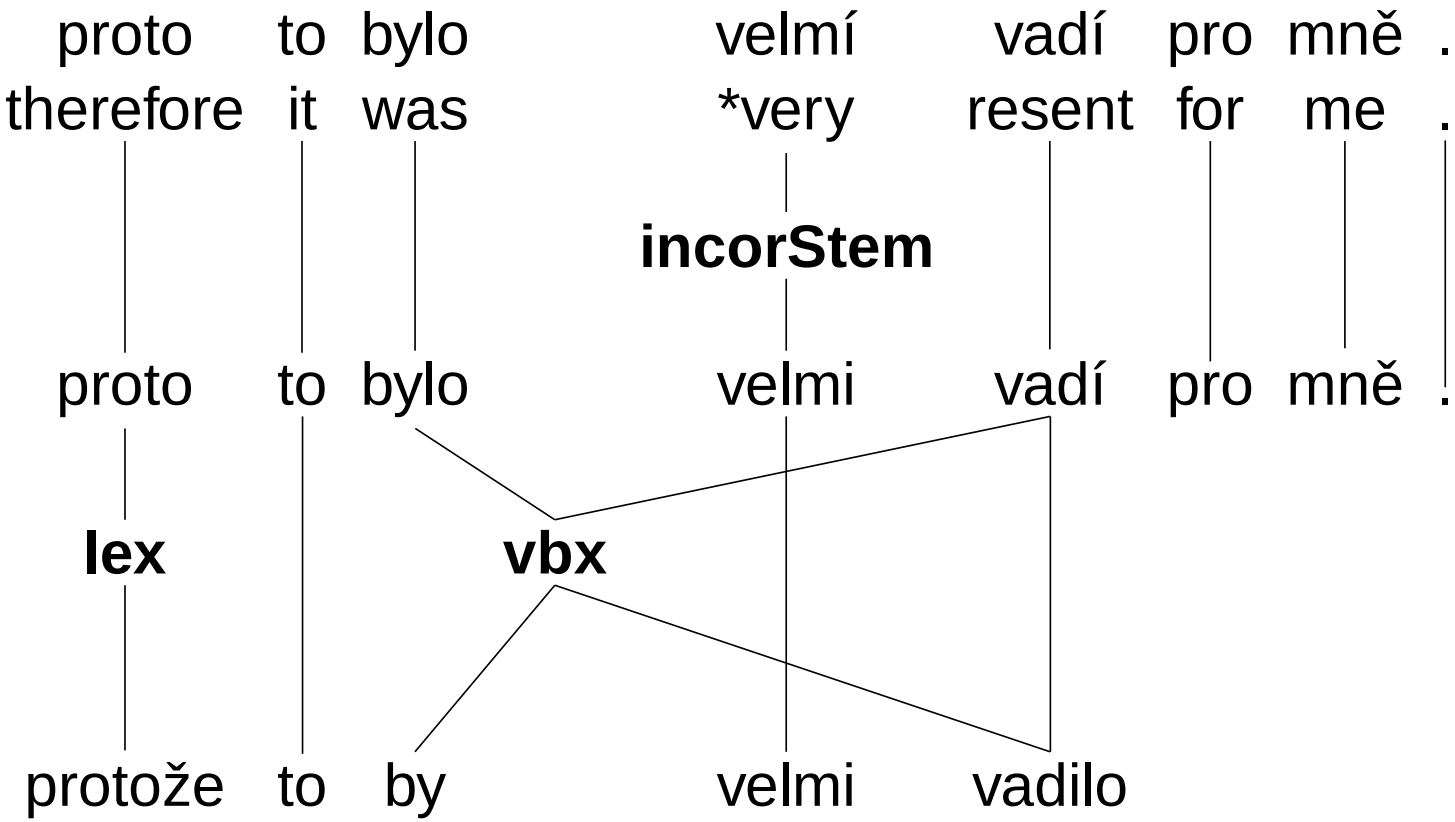
because I would be very unhappy about it.

Annotation of a sample sentence, part II



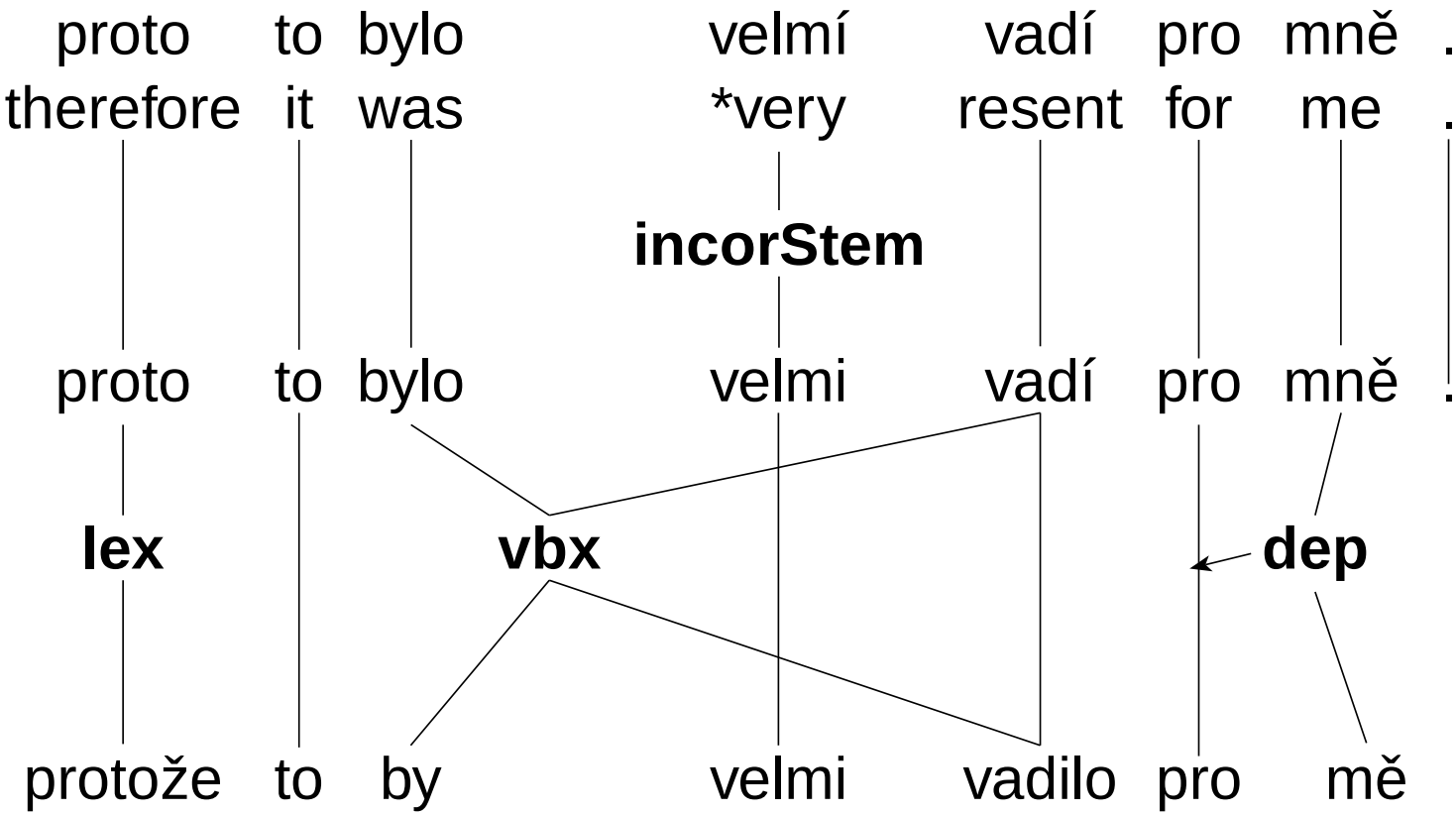
because I would be very unhappy about it.

Annotation of a sample sentence, part II



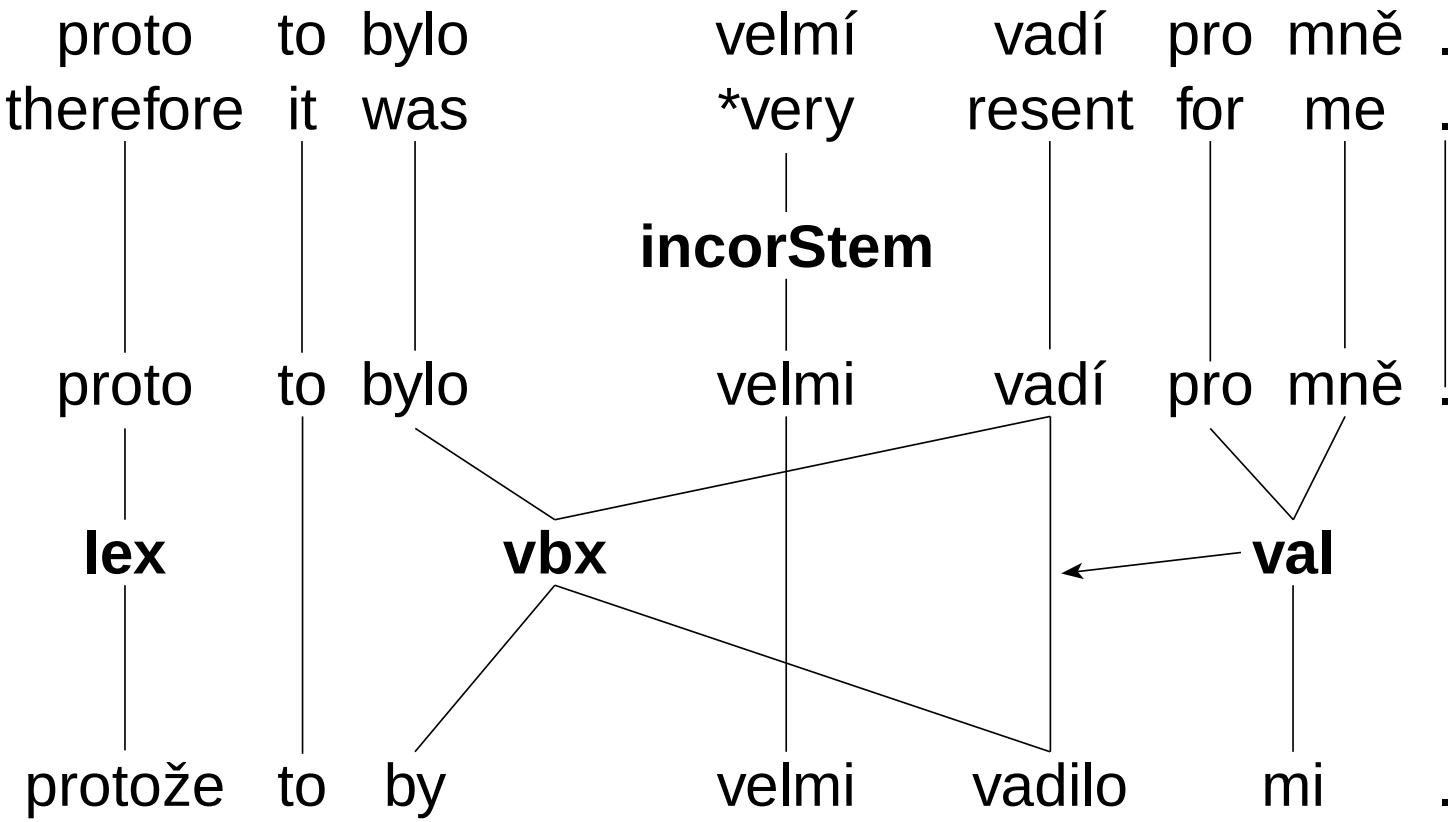
because I would be very unhappy about it.

Annotation of a sample sentence, part II



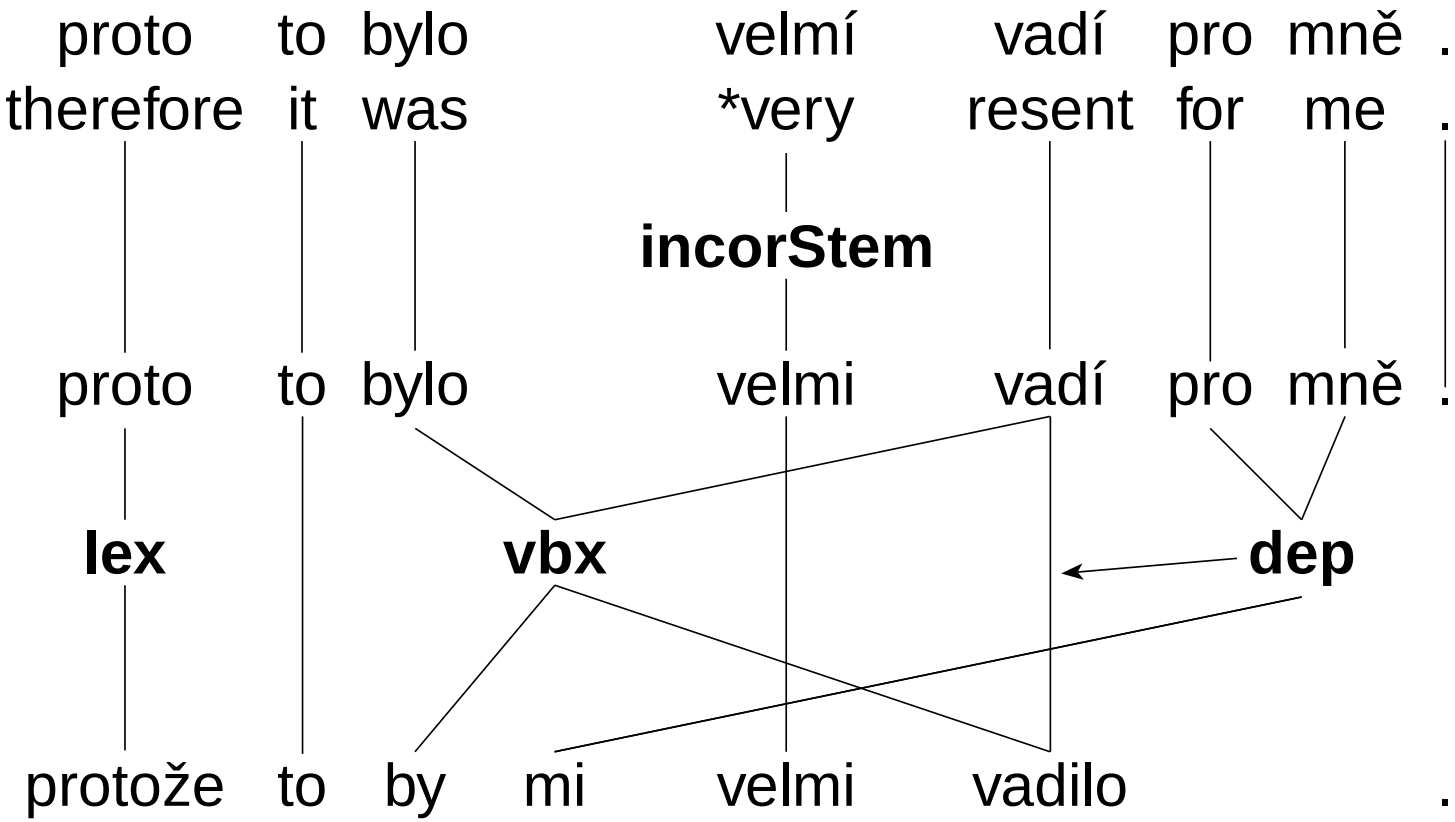
because I would be very unhappy about it.

Annotation of a sample sentence, part II



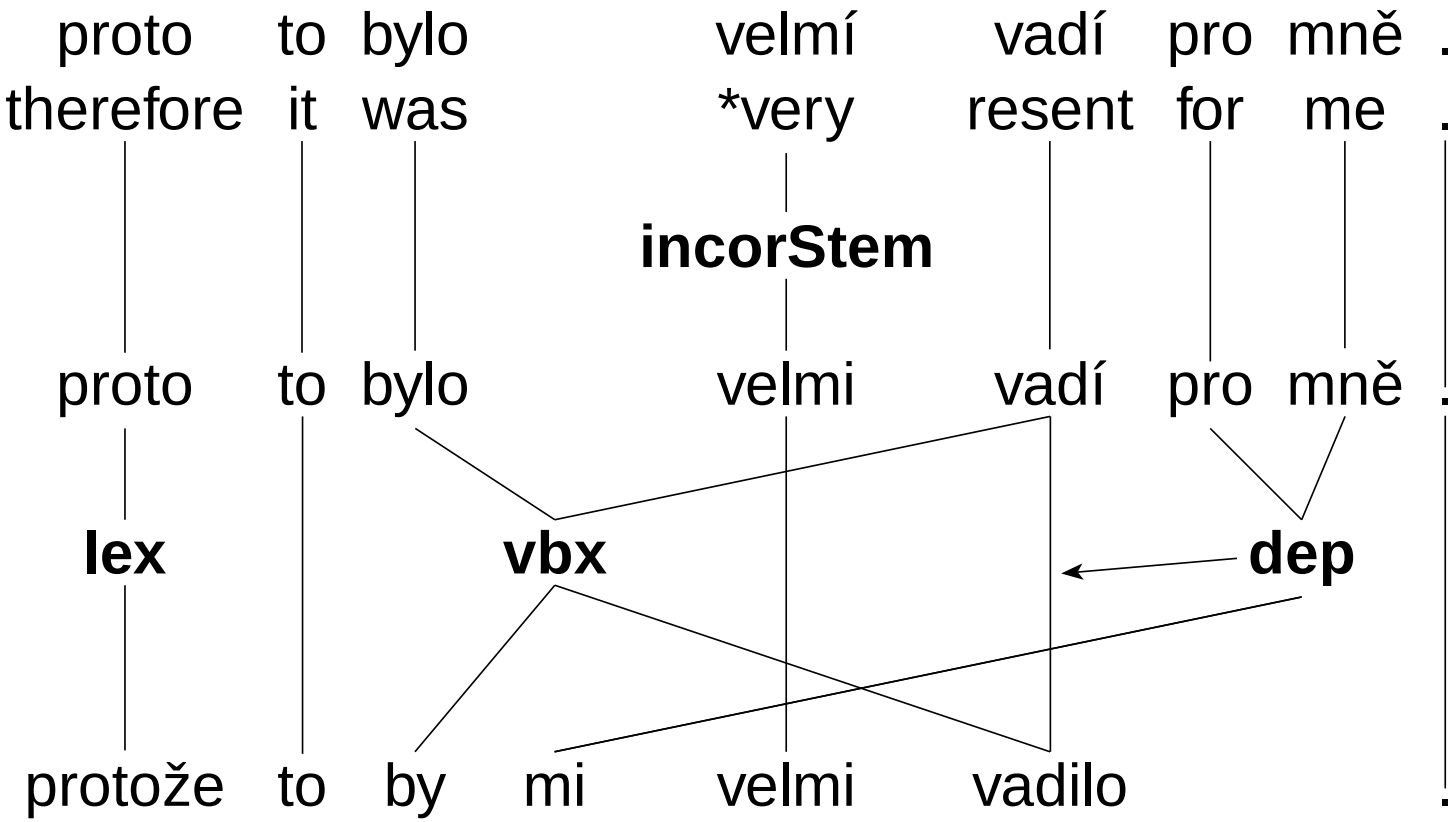
because I would be very unhappy about it.

Annotation of a sample sentence, part II



because I would be very unhappy about it.

Annotation of a sample sentence, part II



because I would be very unhappy about it.

Types of errors at Level 1

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

Types of errors at Level 2

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

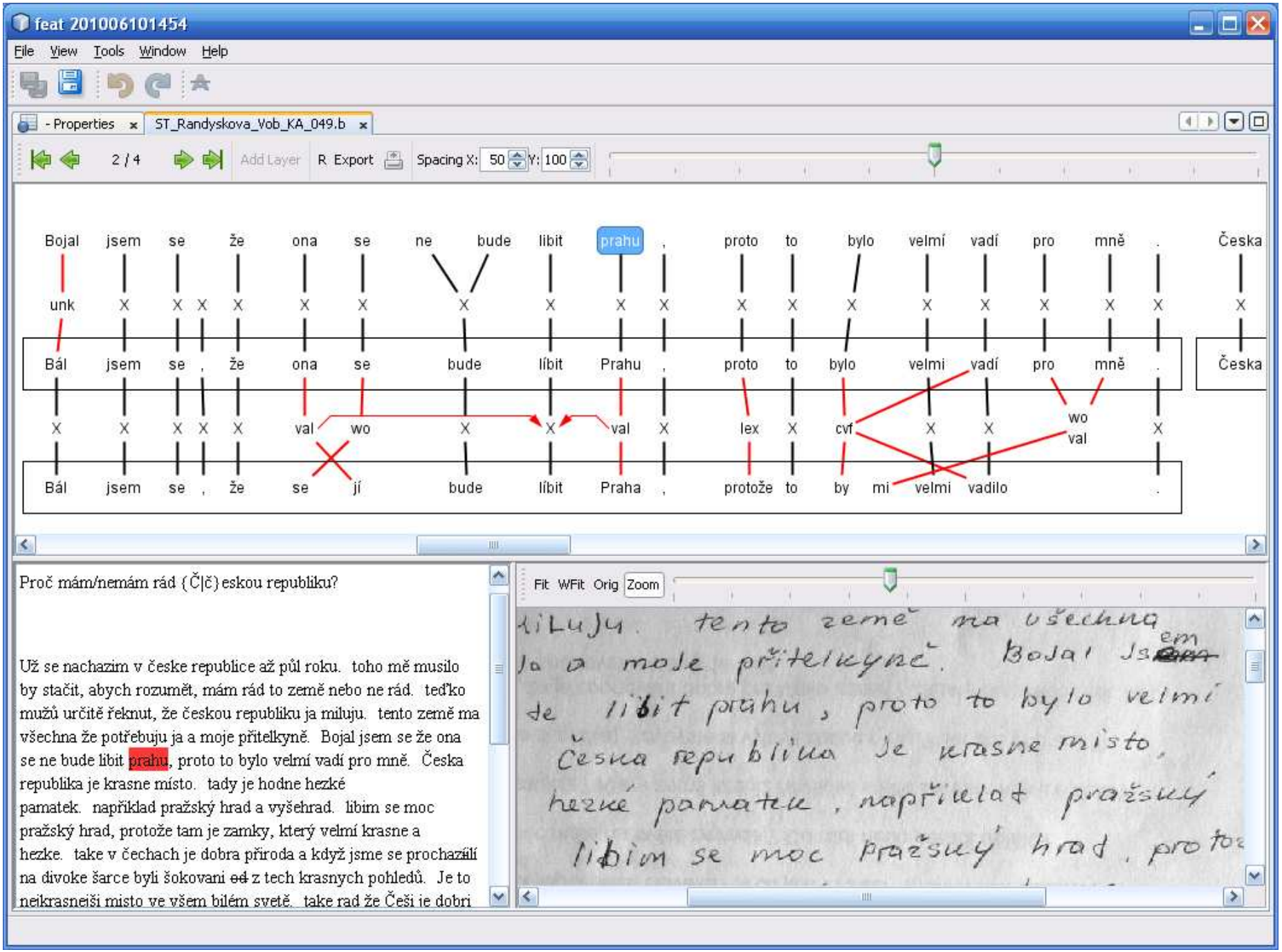
- Prague Markup Language
  (PML, used in *Prague Dependency Treebank*)
- Generic, XML-based, for rich layered annotation
- A higher level contains information about words on that level, about errors and about relations to tokens on lower levels
- Portion of Level 1 of the sample sentence encoded in the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

```
<?xml version="1.0" encoding="UTF-8"?>
<adata xmlns="http://utkl.cuni.cz/czesl/">
  <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

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
ST_Randyskova_Vob_KA_049.b
2 / 4
Add Layer
R Export
Spacing X: 50 Y: 100
Bojal jsem se že ona se ne bude libit prahu , proto to bylo velmí vadí pro mně . Česka
unk 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 cvf X X wo X
val
Bál jsem se , že se jí bude líbit Praha , protože to by mi velmi vadilo .
Proč mám/nemám rád (Č|č)eskou republiku?
Už se nachazim v ceske republice až půl roku. toho mě musilo by stačit, abych rozumět, mám rád to země nebo ne rád. teďko mužů určitě řeknut, že českou republiku ja miluju. tento země ma všechna že potřebuju ja a moje přitelkyně. Bojal jsem se že ona se ne bude libit prahu, proto to bylo velmí vadí pro mně. Česka republika je krasne místo. tady je hodne hezké pamatek. například pražský hrad a vyšehrad. libim se moc pražský hrad, protože tam je zamky, který velmí krasne a hezke. take v čechach je dobra přiroda a když jsme se prochazíli na divoke šarce byli šokovani ~~od~~ z tech krasnych pohledů. Je to neikrasneiši misto ve všem bilém svetě. take rad že Češi je dobri
Fit WFit Orig Zoom


EXMARaLDA Partitur-Editor 1.4.5 [D:\Downloads\uppsala2.exb]
File Edit View Transcription Tier Event Timeline Format Help
0 1 2 3 4 5 6 7
L0 proto to bylo velmí vadí pro mně .
gloss therefore it was *very resent for me .
errid dia
L1 proto to bylo velmi vadí pro mně .
errid lex cvf 2 val 4
L2 protože to by velmi vadilo mi .
errid wo
L3 protože to by mi velmi vadilo .
Done.
Transcription D:\Downloads\uppsala2.exb saved

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

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