Towards Czech-Russian Parallel Treebank

AEPC, Tartu, Estonia 2010

Natalia Klyueva and David Mareček
Institute of Formal and Applied Linguistics
Charles University in Prague

Outline

- Where it all started
- Treebanks PDT and SynTagRus
- Treebank compilement

Where it all started

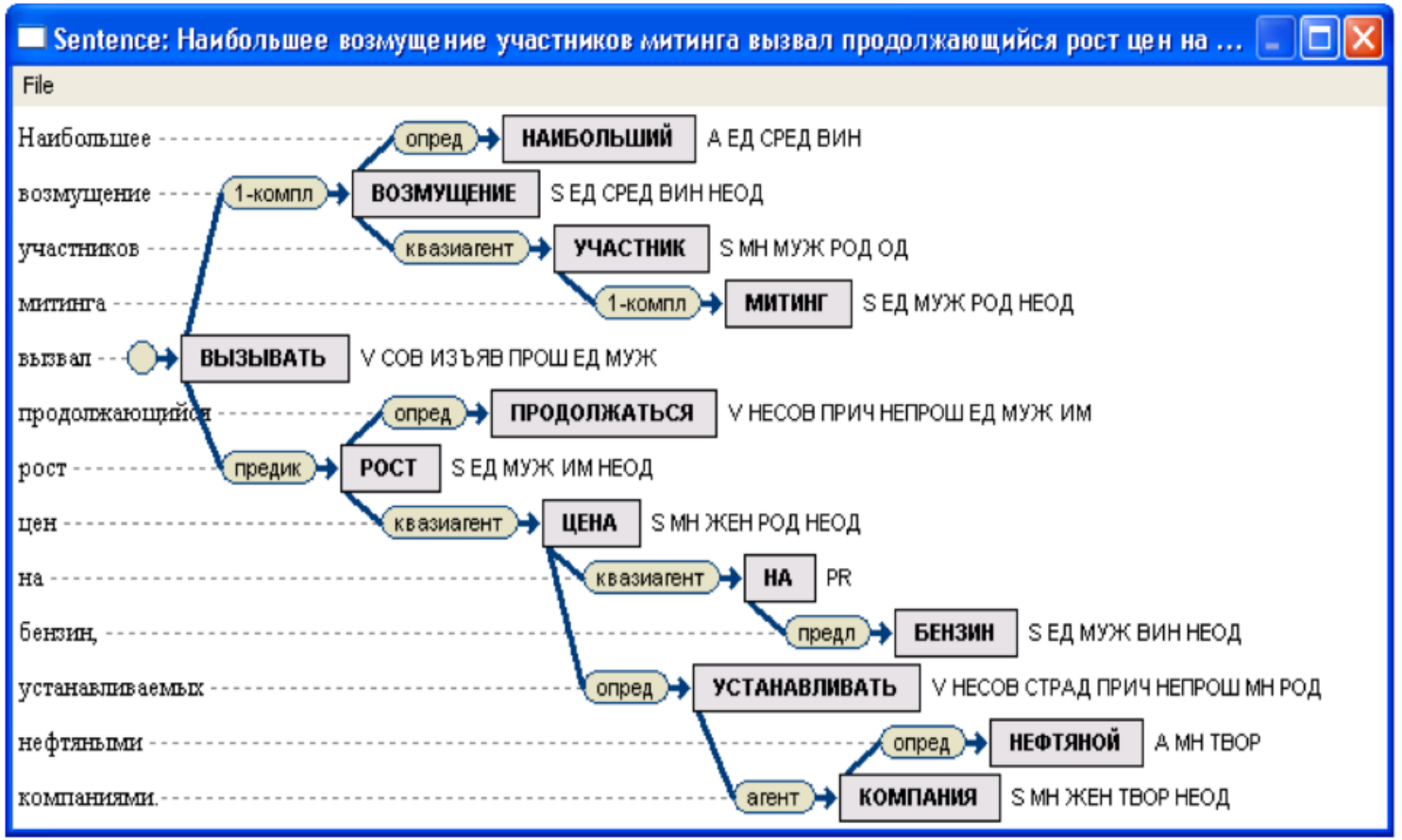
- Rule-Based MT system between Czech and Russian “Česílko”:
  - We had dictionary, parallel corpus, taggers
  - We wanted to have: syntactic transfer module
  - Create rules out of a head or use syntactically annotated treebanks?
- Related projects: PCEDT, SMULTRON
- Used annotated Russian data form the SynTagRus and generated dependency trees for a Czech text with the help of PDT tools

Two giants of syntactic information

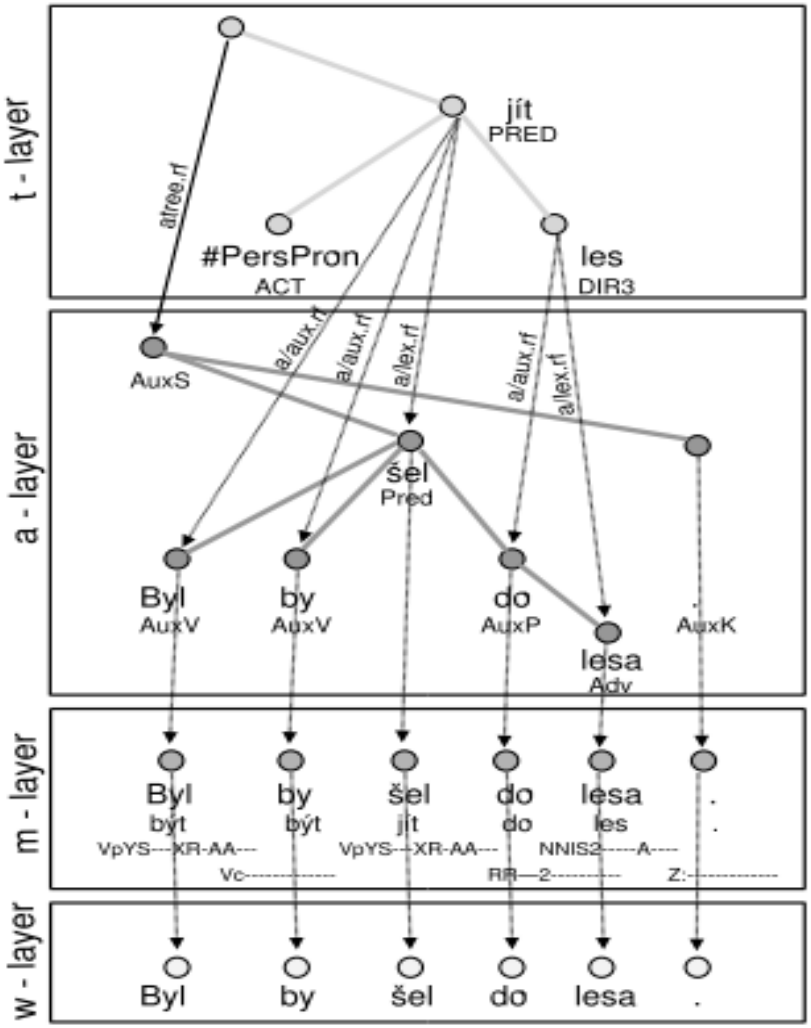
- PDT for Czech
  - 115,884 sentences from newspapers and journals
  - Morphological, analytical and tectogrammatical levels of annotation
  - 1,500,000 words annotated on the analytical level
  - tools for automatic processing of "raw" texts available
- SynTagRus for Russian
  - 32,000 sentences from newspaper articles, prose.
  - 460,000 words with deep syntactic annotation
  - SynTagRus is not an open-source

SynTagRus:a sentence visualized in sTred

greatest
indignation
participants
meeting
caused
continuing
growth
prices
for
petrol
set
oil
companies



PDT:a sentence visualized in Tred



Process of a Treebank Compilement

- Choose a portion of data from SynTagRus that is translated into Czech (Novel “The Faculty” by I. Grekova, 460 sentences annotated)
- Sentence and Word alignment
- Process the raw Czech text
- Convert SynTagRus format into PDT style
  - XML -> PML
  - Syntactic functions -> Afuns or functors from PDT

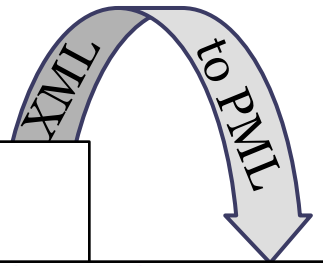
Processing the Czech text

- tokenization
- tagging and lemmatization using Morce tagger
- parsing with McDonald's MST parser
- automatic conversion to tectogrammatical trees using mainly rule-based scripts, which are included in a TectoMT framework: https://ufal.mff.cuni.cz/tectomt/

Processing the Russian text

- Format transfer of SynTagRus XML-based to Prague Markup Language (PML)
- Adapting annotation:
  - Russian tagset was left as it was
  - Transformation of Russian syntactic functions into Czech analytical/tectogrammatical functors

Format Transfer




```
Суть</W>
<W DOM="1" FEAT="S ЕД МУЖ РОД НЕОД" ID="2" KSNAME="ДОКУМЕНТ" LEMMA="ДОКУМЕНТ" LINK="квазиагент">
<HOM FEAT="S ЕД МУЖ РОД НЕОД" KSNAME="ДОКУМЕНТ" LEMMA="ДОКУМЕНТ"/>
документа</W>
<W DOM="_root" FEAT="A КР ЕД ЖЕН" ID="3" KSNAME="ПРОСТОЙ1" LEMMA="ПРОСТОЙ">
<HOM FEAT="A КР ЕД ЖЕН" KSNAME="ПРОСТОЙ1" LEMMA="ПРОСТОЙ"/>
проста</W>:
<W DOM="3" FEAT="V НЕСОВ ИЗЪЯВ НЕПРОШ ЕД 3-Л" ID="4" KSNAME="СОЗДАВАТЬСЯ" LEMMA="СОЗДАВАТЬСЯ" L
<HOM FEAT="V НЕСОВ ИЗЪЯВ НЕПРОШ ЕД 3-Л" KSNAME="СОЗДАВАТЬСЯ" LEMMA="СОЗДАВАТЬСЯ"/>
<HOM FEAT="V НЕСОВ СТРАД ИЗЪЯВ НЕПРОШ ЕД 3-Л" KSNAME="СОЗДАВАТЬ" LEMMA="СОЗДАВАТЬ"/>
создается</W>
<W DOM="6" FEAT="ADV" ID="5" KSNAME="СОВЕРШЕННО" LEMMA="СОВЕРШЕННО" LINK="огранич">
<HOM FEAT="A КР ЕД СРЕД" KSNAME="СОВЕРШЕННЫЙ" LEMMA="СОВЕРШЕННЫЙ"/>
<HOM FEAT="ADV" KSNAME="СОВЕРШЕННО" LEMMA="СОВЕРШЕННО"/>
совершенно</W>
<W DOM="8" FEAT="A ЕД ЖЕН ИМ" ID="6" KSNAME="НОВЫЙ" LEMMA="НОВЫЙ" LINK="опред">
<HOM FEAT="A ЕД ЖЕН ИМ" KSNAME="НОВЫЙ" LEMMA="НОВЫЙ"/>
новая</W>
<W DOM="8" FEAT="A ЕД ЖЕН ИМ" ID="7" KSNAME="ГОСУДАРСТВЕННЫЙ" LEMMA="ГОСУДАРСТВЕННЫЙ" LINK="опр
<HOM FEAT="A ЕД ЖЕН ИМ" KSNAME="ГОСУДАРСТВЕННЫЙ" LEMMA="ГОСУДАРСТВЕННЫЙ"/>
государственная</W>
<W DOM="4" FEAT="S ЕД ЖЕН ИМ НЕОД" ID="8" KSNAME="СТРУКТУРА" LEMMA="СТРУКТУРА" LINK="предик">
<HOM FEAT="S ЕД ЖЕН ИМ НЕОД" KSNAME="СТРУКТУРА" LEMMA="СТРУКТУРА"/>
структура</W>
<W DOM="10" FEAT="ADV" ID="9" KSNAME="ФОРМАЛЬНО" LEMMA="ФОРМАЛЬНО" LINK="обст">
<HOM FEAT="A КР ЕД СРЕД" KSNAME="ФОРМАЛЬНЫЙ" LEMMA="ФОРМАЛЬНЫЙ"/>
<HOM FEAT="ADV" KSNAME="ФОРМАЛЬНО" LEMMA="ФОРМАЛЬНО"/>
```

```
<LM id="SRussianA-syntagrus-s5-w4">
 <m>
  <form>создается</form>
  <lemma>СОЗДАВАТЬСЯ</lemma>
  <tag>V НЕСОВ ИЗЪЯВ НЕПРОШ ЕД 3-Л</tag>
 </m>
 <ord>4</ord>
 <children>
  <LM id="SRussianA-syntagrus-s5-w8">
   <m>
    <form>структура</form>
    <lemma>СТРУКТУРА</lemma>
    <tag>S ЕД ЖЕН ИМ НЕОД</tag>
   </m>
   <ord>8</ord>
   <children>
    <LM id="SRussianA-syntagrus-s5-w6">
     <m>
      <form>новая</form>
      <lemma>НОВЫЙ</lemma>
      <tag>A ЕД ЖЕН ИМ</tag>
     </m>
     <ord>6</ord>
```

Morphological layer

- Morphological systems of Czech and Russian are very similar
- Czech: lemma + morphological tag, which has 15 positions filled with a morphological category:

před	brankou	stál	pokroucený	dub
před-1	branka	stát-5_^(sníh)	pokroucený_^(*4tit)	dub
RR--7----------	NNFS7-----A----	VpYS---XR-AA---	AAIS1----1A----	NNIS1-----A----

- Russian: lemma + semi-positional tag:

Никто	не	хочет	делиться	с	соседом
НИКТО	НЕ	ХОТЕТЬ	ДЕЛИТЬСЯ	С	СОСЕД
S ЕД МУЖ ИМ ОД	PART	V НЕСОВ ИЗЪЯВ НЕПРОШ ЕД 3-Л	V НЕСОВ ИНФ	PR	S ЕД МУЖ ТВОР ОД

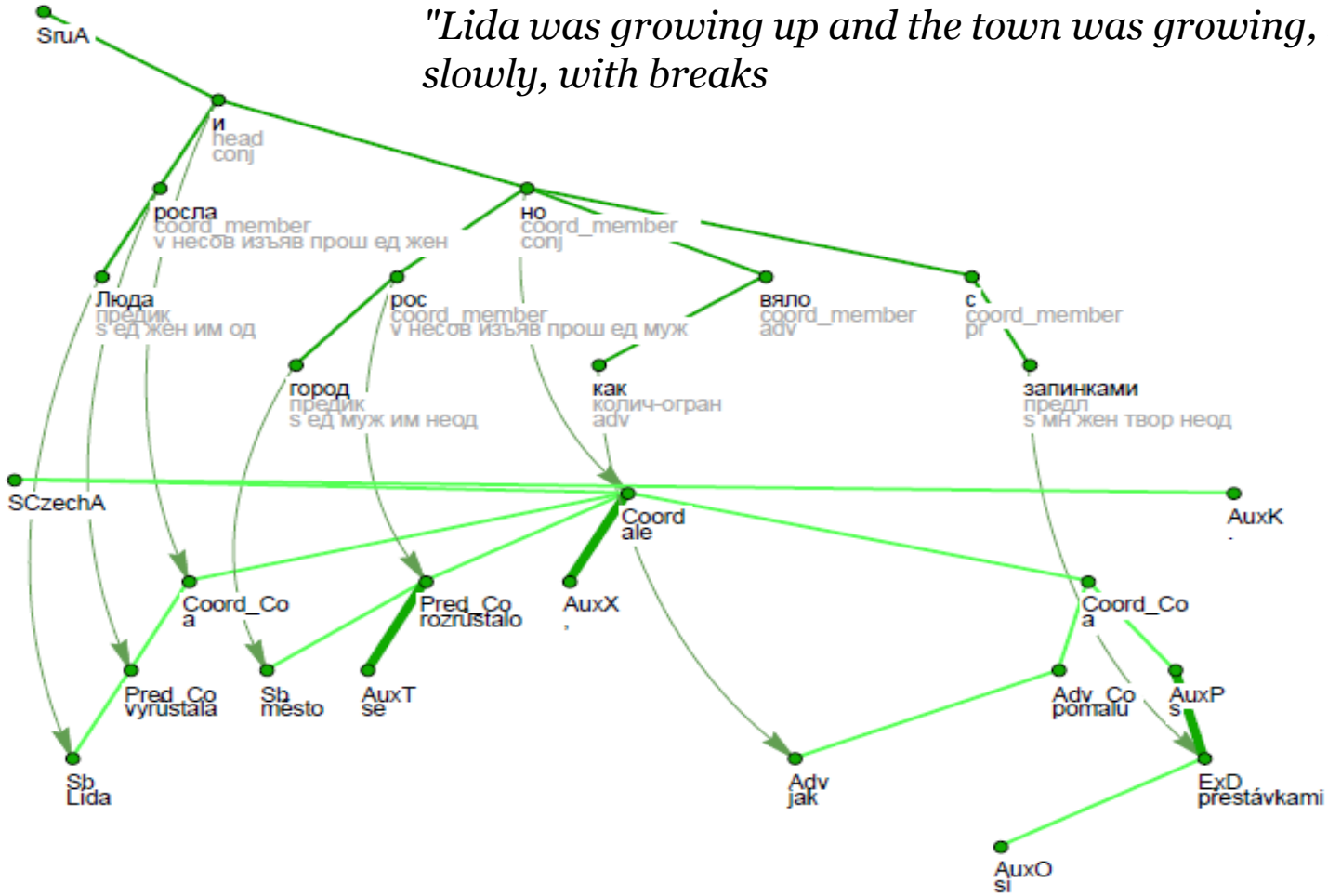
Word Alignment

- Words in Czech and Russian sentences are automatically aligned with GIZA++(1-to-1)
- Word alignment was run on this corpus and a parallel Czech-Russian corpus (almost 100,000 sentences)
- 100 sentences from the treebank evaluated: precision = 85%

Analytical layer(1)

- For Russian:
  - Syntactic functions(ru) are referred to corresponding analytical functions(cz):
    - Predicative *he reads* — Pred
    - 1-compl *translate a book* — Object
    - Atributive *house we leave in* — Atv
    - Adverbial *to be at home* — Adv
    - Coord *milk and cream* — Coord
    - Auxiliary *will buy* — Aux
- 78 syntactic functions in SynTagRus, 23 afuns in PDT
- Incorrespondences: intersection with tectogrammatical layer

Analytical layer(2)

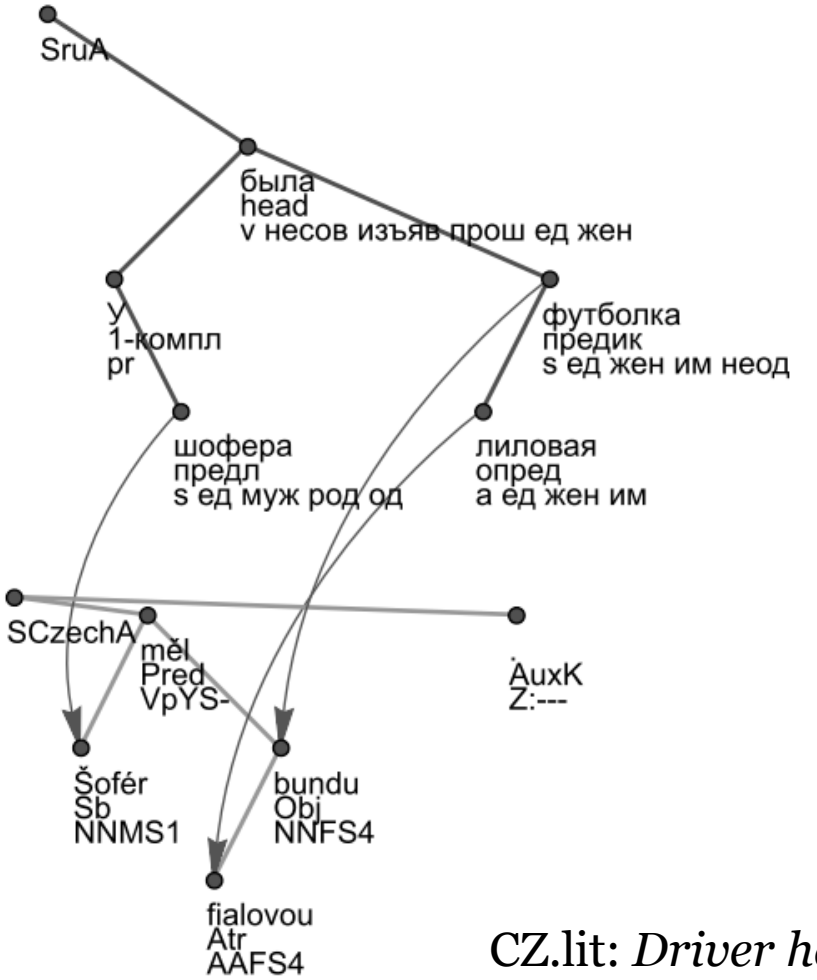


"Lida was growing up and the town was growing, but somehow slowly, with breaks

Tectogrammatical layer(1)

- Differences of annotation schemes:
  - The syntactic layer of annotation for Russian is more deep and semanticalized, and it is one layer.
  - PDT distincts shallow and deep syntax. Syntactic features belong to the analytical layer and more semantics ones to the tectogrammatical layer.
- Decision for the unmatching functors: rules
  - Ru: 1-compl in Acc. → Cz: Patient,
  - Ru: 1-compl in Ins. → Cz: Means.

Analytical and tectogrammatical layers



CZ.lit: *Driver had a lilac coat*
RU.lit: *For driver was a lilac coat*

Comparison with PCEDT

- People:
  - PCEDT : many people involved into a project
  - Czech-Russian Treebank – only 2.
- Corpus size: 53,000 vs. 460 sentences
- Translations:
  - PCEDT: as close to the original as possible
  - Czech-Russian: Novel translation
- What did help us: dependency based approach for both Czech and Russian Treebanks, languages' relatedness.

Conclusion and Plans for Future

- 460 sentences – only a start. The treebank is suitable for comparative linguistic studies, not as the data for the Machine Translation
- A lot to improve:
  - Quality. Develop rules for tectogrammatical annotation
  - Quantity. Add new texts and even experiment with the automatic annotation of Czech-Russian corpus

Thank you