PMLandCSTS

DOKUMENTÁCIA ROČNÍKOVÉHO PROJEKTU

Michal Vachna

Zimný semester 2007/08 Letný semester 2008/09
PRG033, PRG034

UŽÍVATEĽSKÁ ČASŤ

1. Úvod

PMLandCSTS je samostatná aplikácia určená na konverziu súborov vo formátoch CSTS a PML oboma smermi. Aplikácia sa spúšťa z príkazovej riadky s použitím trojice predpísaných parametrov.

Aplikácia PMLandCSTS bola vytvorená ako alternatíva k procedúre, ktorá tiež slúži na konverziu medzi uvedenými formátmi a je súčasťou anotačného editoru [b]TrEd.

Aplikácia PMLandCSTS je vo verzii pre operačné systémy Windows a Linux. Obe verzie sú distribuované ako jeden balíček PMLandCSTS.rar. Po rozbalení balíčka sú dostupné nasledujúce podadresáre: **doc**, ktorý obsahuje programátorskú a užívateľskú dokumentáciu; **source**, ktorý obsahuje zdrojové kódy aplikácie pre verziu Windows a Linux; **sample**, ktorý obsahuje vstupné a výstupné súbory ilustratívnych behov aplikácie; **Windows**, ktorý obsahuje spustiteľný binárny súbor pre operačný systém Windows; **Linux**, ktorý obsahuje spustiteľný binárny súbor pre operačný systém Linux.

1.1 Anotácia textu a roviny anotácie

Pojem anotácia označuje priradenie údajov slovným jednotkám textu. Tieto údaje charakterizujú napríklad syntaktické alebo morfologické vlastnosti slovných jednotiek. Podľa toho o aké vlastnosti z jazykovedného pohľadu ide sa rozlišujú roviny anotácie.

Anotácia Českého akademického korpusu 2.0 (CAK 2.0) pokrýva 2 roviny – morfologickú a analytickú. Operuje sa ešte s jednou rovinou a to s rovinou slovnou. Slovná rovina je v skutočnosti rovinou neanotačnou, obsahuje len pôvodný text rozdelený na slovné jednotky. Slovná rovina sa označuje ako ***w***-rovina (z anglického *word*). Morfologická rovina sa označuje ako ***m***-rovina a anotácia na tejto rovine znamená, že slovným jednotkám textu sú priradzované údaje, ktoré charakterizujú ich morfologické vlastnosti. Analytická rovina sa označuje ako ***a***-rovina a anotácia na tejto rovine znamená, že slovným jednotkám textu sú priradzované údaje, ktoré charakterizujú ich syntaktické vlastnosti.

Anotácia Pražského závislostného korpusu 2.0 oproti CAK 2.0 pokrýva naviac rovinu tektogramatickú.

1.2 Formát CSTS

Formát je založený na formáte SGML. Je lepšie čitateľný pre človeka a je možné ho spracovať jednoduchými nástrojmi. Vo formáte CSTS sú všetky roviny anotácie uchované v jednom súbore.

Súbor CSTS začíná (nepovinnou) hlavičkou (element ***h***) a ďalej obsahuje aspoň jeden element ***doc***. Element ***doc*** pozostáva z hlavičky (element ***a***) a obsahu (element ***c***). Element ***c*** potom pozostáva z postupnosti odstavcov (element ***p***) a viet v týchto odstavcoch (elementy ***s***).

Každá slovná jednotka vety je na samostatnom riadku súboru (element ***f***, resp. ***d*** pre interpunkciu), ďalej na tomto riadku nasleduje anotácia tejto slovnej jednotky na všetkých rovinách.

Ukážka anotácie vety vo formáte CSTS: Súhlasíme s ním.

```
<csts lang=cs>
<doc file="sample.1" id=000>
<a>
<mod>s
<txtype>inf
<genre>x
<med>x
<temp>x
<authname>x
<opus>REC1X
<id>001
</a>
<c>
<p n=0>
<s id=sample1-s1>
<f id=sample1-s1W1>Súhlasíme<l>súhlasiť_:T<t>VB-P---1P-AA---<r>1<g>0<A>Pred
<f id=sample1-s1W2>s<l>s<t>RR--7----------<r>2<g>1<A>AuxP
<f id=sample1-s1W3>ním<l>on<t>P5ZS7--3-------<r>3<g>2<A>Obj
<D>
<d id=sample1-s1W4>.<l>.<t>Z:-------------<r>4<g>0<A>AuxK
</c>
</doc>
</csts>
```

1.3 Formát PML

Formát *Prague Markup Language* (PML) je založený na XML a navrhnutý pre bohatú reprezentáciu lingvistickej anotácie textov. Každá rovina anotácie je popísaná v soubore *PML schéma*, ktorý je akousi formalizáciou abstraktnej anotačnej schémy pre tú ktorú rovinu anotácie. Rozlišujú sa tri roviny anotácie: ***w***, ***m***, ***a*** . Rovina ***w*** obsahuje slovné jednotky. Rovina ***m*** obsahuje morfologickú anotáciu a hranice viet. Rovina ***a*** anotáciu stromov závislosti.

Každý PML súbor začína hlavičkou odkazujúcou na PML schéma súboru. V hlavičke sú uvedené všetky externé zdroje. Zbytok súboru obsahuje vlastnú anotáciu.

Anotácia je vyjádrená pomocou XML elementov a atributov pomenovaných a použitých v súlade s príslušnou PML schémou. XML elementy všetkých súborov patria do vyhradeného menného priestoru http://ufal.mff.cuni.cz/pdt/pml/.

Ukážka anotácie vety vo formáte PML: Súhlasíme s ním.

V rovine a:

```
<?xml version="1.0" encoding="utf-8"?>
<adata xmlns="http://ufal.mff.cuni.cz/pdt/pml/">
 <head>
  <schema href="adata_schema.xml" />
  <references>
   <reffile id="m" href="sample1.m" name="mdata" />
   <reffile id="w" href="sample1.w" name="wdata" />
  </references>
 </head>
 <trees>
  <LM id="a-sample1-s1">
   <s.rf>m#m-sample1-s1</s.rf>
   <afun>AuxS</afun>
   <ord>0</ord>
   <children>
    <LM id="a-sample1-s1W1">
     <afun>Pred</afun>
     <m.rf>m#m-sample1-s1W1</m.rf>
     <is_member>0</is_member>
     <is_parenthesis_root>0</is_parenthesis_root>
     <ord>1</ord>
     <children id="a-sample1-s1W2">
      <afun>AuxP</afun>
      <m.rf>m#m-sample1-s1W2</m.rf>
      <is_member>0</is_member>
      <is_parenthesis_root>0</is_parenthesis_root>
      <ord>2</ord>
      <children id="a-sample1-s1W3">
       <afun>Obj</afun>
       <m.rf>m#m-sample1-s1W3</m.rf>
       <is_member>0</is_member>
       <is_parenthesis_root>0</is_parenthesis_root>
       <ord>3</ord>
      </children>
     </children>
    </LM>
    <LM id="a-sample1-s1W4">
     <afun>AuxK</afun>
     <m.rf>m#m-sample1-s1W4</m.rf>
     <is_member>0</is_member>
     <is_parenthesis_root>0</is_parenthesis_root>
     <ord>4</ord>
    </LM>
   </children>
  </LM>
 </trees>
</adata>
```

V rovine m:

```
<?xml version="1.0" encoding="utf-8"?>
<mdata xmlns="http://ufal.mff.cuni.cz/pdt/pml/">
  <head>
    <schema href="mdata_schema.xml"/>
    <references>
      <reffile id="w" name="wdata" href="sample1.w"/>
    </references>
  </head>
  <meta>
    <lang>cs</lang>
  </meta>
  <s id="m-sample1-s1">
    <m id="m-sample1-s1W1">
      <src.rf>manual</src.rf>
      <w.rf>w#w-sample1-s1W1</w.rf>
      <form>Súhlasíme</form>
      <lemma>súhlasiť_:T</lemma>
      <tag>VB-P---1P-AA---</tag>
    </m>
    <m id="m-sample1-s1W2">
      <src.rf>manual</src.rf>
      <w.rf>w#w-sample1-s1W2</w.rf>
      <form>s</form>
      <lemma>s</lemma>
      <tag>RR--7----------</tag>
    </m>
    <m id="m-sample1-s1W3">
      <src.rf>manual</src.rf>
      <w.rf>w#w-sample1-s1W3</w.rf>
      <form>ním</form>
      <lemma>on</lemma>
      <tag>P5ZS7--3-------</tag>
    </m>
    <m id="m-sample1-s1W4">
      <src.rf>manual</src.rf>
      <form_change>insert</form_change>
      <form>.</form>
      <lemma>.</lemma>
      <tag>Z:-------------</tag>
    </m>
  </s>
</mdata>
```

V rovine w:

```
<?xml version="1.0" encoding="utf-8"?>
<wdata xmlns="http://ufal.mff.cuni.cz/pdt/pml/">
  <meta>
```

```
    <original_format>csts</original_format>
    <lang>cs</lang>
  </meta>
  <doc id="w-sample1-001" source_id="REC1X.DAT-001">
    <docmeta>
      <othermeta origin="csts/doc/a"><![CDATA[
<mod>s
<txtype>inf
<genre>x
<med>x
<temp>x
<authname>x
<opus>REC1X
<id>001
]]></othermeta>
    </docmeta>
    <para>
      <othermarkup origin="csts/doc/p/@n">1</othermarkup>
      <w id="w-sample1-s1W1">
        <token>Súhlasíme</token>
      </w>
      <w id="w-sample1-s1W2">
        <token>s</token>
      </w>
      <w id="w-sample1-s1W3">
        <token>ním</token>
        <no_space_after>1</no_space_after>
      </w>
    </para>
  </doc>
</wdata>
```

1.4 TrEd a bTrEd

TrEd (Tree Editor) je integrovaným prostredím primárne navrhnutým pre syntaktické anotovanie viet, pri ktorom je vete priradená stromová štruktúra. Zároveň môže byť použitý i k prehliadaniu dát. TrEd podporuje formáty PML a CSTS.

Súčasťou TrEdu je bTrEd, nástroj bez grafického rozhrania pre dávkové spracovanie dát (Batch-mode Tree Editor). bTrEd obsahuje perlovské makro, ktoré slúži ako procedúra na prevod medzi PML a CSTS. Použitie makra si však vyžaduje inštaláciu celého TrEdu.

2. Parametre

Prvý parameter určuje smer konverzie, respektíve formát výstupných súborov. Parameter **–c** určuje smer konverzie z PML do CSTS, výstupný súbor bude vo formáte CSTS. Parameter **-p** určuje smer konverzie z CSTS do PML, výstupné súbory budú vo formáte PML.

Po parametri **–c** nasleduje druhý parameter, ktorý určuje súbory akých hladín sa majú pri konverzii použiť. Možné hodnoty parametru sú **–a** a **–m.** Hodnota parametru **–a** znamená,

že pre konverziu z PML do CSTS sa majú použiť súbory všetkých troch hladín ***a***, ***m***, ***w***. Hodnota parametru **–m** znamená, že pre prevod z PML do CSTS sa majú použiť súbory hladín ***m***, ***w***. Tretím parametrom je absolútna alebo relatívna cesta s názvom súboru najvyššej hladiny, ktorá sa má pri konverzii použiť. Teda pre hodnotu druhého parametru **–a** bude tretí parameter cesta s názvom súboru hladiny a pre hodnotu **–m** cesta s názvom súboru hladiny m. Súbory nižších hladín sa musia nachádzať v tom istom adresári ako súbor v treťom parametri.

Po parametri **–p** nasleduje druhý parameter, ktorý určuje súbory hladín, ktoré sa majú pri konverzii vytvoriť. Možné hodnoty parametru sú **–all, -w, -m** alebo **–a.** Hodnota **–all** znamená, že pri konverzii z CSTS do PML sa majú vytvoriť súbory všetkých hladín. Jednotlivé hodnota **–w, –m, –a** znamenajú, že sa má vytvoriť len súbor danej hladiny. Tretím parametrom je absolútna alebo relatívna cesta s názvom vstupného súboru vo formáte CSTS, ktorý sa má pri konverzii použiť.

2.1 Ukážky použitia aplikácie

Všetky výstupy v adresári sample sú v tvare *_out.*. To je len kvôli prehľadnosti, PMLandCSTS však výstupy vždy pomenúva podľa vstupu a mení len koncovky formátov.

Príklad spustení aplikácie je uvádzaný najprv vo verzii Windows a následne ekvivalentné spustenie vo verzii Linux.

Prevod *sample1* z PML do CSTS s použitím súborov všetkých troch rovín, výstupom je *sample1_out.csts*:

```
Windows/PMLandCSTS.exe –c –a sample/sample1.a

Linux/PMLandCSTS –c –a sample/sample1.a
```

Prevod *sample2* z PML do CSTS s použitím súborov rovín ***m*** a ***w*,** výstupom je *sample2_out.csts*:

```
Windows/PMLandCSTS.exe –c –m sample/sample2.m

Linux/PMLandCSTS –c –m sample/sample2.m
```

Prevod *sample3* z CSTS do PML s vytvorením súborov všetkých troch rovín, výstupom sú *sample3_out.a, sample3_out.m, sample3_out.w* :

```
Windows/PMLandCSTS.exe –p –all sample/sample3.csts

Linux/PMLandCSTS –p –all sample/sample3.csts
```

Prevod *sample4* z CSTS do PML s vytvorením súboru roviny ***a***, výstupom je *sample4_out.a*:

```
Windows/PMLandCSTS.exe –p –a sample/sample4.csts

Linux/PMLandCSTS –p –a sample/sample4.csts
```

Prevod *sample5* z CSTS do PML s vytvorením súboru roviny ***m***, výstupom je *sample5_out.m*:

```
Windows/PMLandCSTS.exe -p -m sample/sample5.csts

Linux/PMLandCSTS -p -m sample/sample5.csts
```

Prevod *sample6* z CSTS do PML s vytvorením súboru roviny ***w***, výstupom je *sample6_out.w*:

```
Windows/PMLandCSTS.exe -p -w sample/sample6.csts

Linux/PMLandCSTS -p -w sample/sample6.csts
```

3. Vstup

Pri konverzii z CSTS do PML musí vstupný súbor splňať syntaktické pravidlá formátu CSTS a musí byť v kódovaní ISO-8859-2. Iné kódovania pre formát CSTS nie sú v aktuálnej verzii aplikácie podporované. V prípade, že vstupný súbor bude v inom kódovaní, konverzia prebehne, no kódovanie výstupných súborov nemusí byť korektné a aplikácie, ktoré slúžia na prehliadanie a editáciu súborov vo formáte PML, nemusia byť schopné tieto súbory spracovať. V prípade zlej syntaxe konverzia neprebehne.

Pri konverzii z PML do CSTS musia vstupné súbory splňať syntaktické pravidlá formátu PML a musia byť v kódovaní UTF-8. Iné kódovania pre formát PML nie sú v aktuálnej verzii aplikácie podporované. V prípade, že vstupné súbory budú v inom kódovaní, konverzia prebehne, no kódovanie výstupného súboru nemusí byť korektné a aplikácie, ktoré slúžia na prehliadanie a editáciu súborov vo formáte CSTS, nemusia byť schopné tento súbor spracovať. V prípade zlej syntaxe konverzia neprebehne.

4. Výstup

Výstupné súbory sa po úspešnej konverzii ukladajú do adresára, v ktorom sa nachádzajú vstupné súbory. Súbory vo formáte PML sa ukladajú s kódovaní UTF-8. Súbory vo formáte CSTS sa ukladajú v kódovaní ISO-8859-2. Ako názvy výstupných súborov sa použije názov súboru, ktorý bol uvedený v treťom parametri, a pridá sa príslušná koncovka výstupných súborov.

5. Priebeh konverzie

Po spustení aplikácie z príkazovej riadky s paramtrami prebieha kontrola zadaných parametrov. V prípade, že počet parametrov nie je tri, bol zadaný neznámy parameter alebo

vstupné súbory nie je možné otvoriť na čítanie, aplikácie vypíše príslušné chybové hlásenie a ukončí sa, konverzia neprebehne.

V opačnom prípade pokračuje v načítaní vstupných súborov. Ak počas načítavania aplikácia narazí na syntaktickú chybu v súbore, po vypísaní chybového hlásenia sa aplikácia ukončí, konverzia je neuspešná.

Ak boli vstupné súbory úspešne načítané, prebieha konverzia, výstupné súbory sú uložené do adresára, v ktorom sa nachádzajú vstupné súbory.

6. Literatúra

Formát PML
- · stručný popis - http://ufal.mff.cuni.cz/morce/cac/?chapter=3#data-format
- · podrobná dokumentace - http://ufal.mff.cuni.cz/pdt2.0/doc/data-formats/pml/

Formát CSTS
- · stručný popisy - http://ufal.mff.cuni.cz/morce/cac/?chapter=3#data-format
- · podrobná dokumentace - http://ufal.mff.cuni.cz/pdt2.0/doc/data-formats/csts/html/c_s_t_s_.html

nástroj TrEd
- · domovská stránka - http://ufal.mff.cuni.cz/~pajas/tred/index.html

Pražský závislostný korpus 2.0 (PDT 2.0)
- · domovská stránka - http://ufal.mff.cuni.cz/pdt2.0

Český akademický korpus 2.0 (CAK 2.0)
- · domovská stránka - http://ufal.mff.cuni.cz/rest/CAC/cac_20.html