Corpus2 - Błąd #6578
Problem z obsługą tagów <prop> i encji HTML

05 Jan 2015 22:12 - Jan Kocoń

Status:	Rozwiązany	**Start date:**	05 Jan 2015
Priority:	Normalny	**Due date:**	12 Jan 2015
Assignee:		**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:			

Description

Problem leży najprawdopodobniej w piśniku ccl. Przykładowy fragment pliku, dla którego występuje błąd:

```
<tok>
 <orth>firma</orth>
 <lex disamb="1"><base>firma</base><ctag>subst:sg:nom:f</ctag></lex>
 <ann chan="nam_eve">0</ann>
 <ann chan="nam_loc">0</ann>
 <ann chan="nam_org">0</ann>
</tok>
<tok>
 <orth>Bögl</orth>
 <lex disamb="1"><base>Bögl</base><ctag>ign</ctag></lex>
 <ann chan="nam_eve">0</ann>
 <ann chan="nam_loc">0</ann>
 <ann chan="nam_org">1</ann>
 <prop key="nam_org:lemma">Bögl &amp;amp; Krýsl</prop>
</tok>
<tok>
 <orth>&amp;amp;</orth>
 <lex disamb="1"><base>&amp;amp;</base><ctag>interp</ctag></lex>
 <ann chan="nam_eve">0</ann>
 <ann chan="nam_loc">0</ann>
 <ann chan="nam_org">1</ann>
</tok>
```

Chodzi o linijkę:

```
<prop key="nam_org:lemma">Bögl &amp;amp; Krýsl</prop>
```

Po przetworzeniu (np. przy pomocy corpus-get) pliku zawierającego tę zawartość poleceniem:

```
corpus-get -i ccl -o ccl -t nkjp agora-1.1.0-names-disamb/articles/00107679.xml > out2.xml
```

linijka wygląda tak:

```
<prop key="nam_org:lemma">Bögl & Krýsl</prop>
```

Na wyjściu powinniśmy otrzymać to samo, co trafiło na wejście. Ponowna próba przetworzenia wyjścia poleceniem:

```
corpus-get -i ccl -o ccl -t nkjp out2.xml > out3.xml
```

Powoduje wyświetlenie błędu:

```
XML Error: xmlParseEntityRef: no name

terminate called after throwing an instance of 'xmlpp::parse_error'
  what():  Document not well-formed.
Line 239, column 21 (fatal):
xmlParseEntityRef: no name

Aborted (core dumped)
```

Linijka na którą wskazuje wyjątek zawiera niewyeskejpowany ampersand. Podobnie jest po podwójnym przetwarzaniu iobberem oraz nawet najnowszym wcrft-app. Do poprawki!

History

#1 - 05 Jan 2015 22:14 - Jan Kocoń

- Description updated

#2 - 12 Jan 2015 16:58 - Radosław Warzocha

- Due date set to 12 Jan 2015

- Status changed from Nowy to Rozwiązany

- % Done changed from 0 to 100

Wrzucone do repo