Towards a Standardised Linguistic Annotation of Fairy Tales

Thierry Declerck[1], Kerstin Eckart[2], Piroska Lendvai[3], Laurent Romary[4], Thomas Zastrow[5]
[1] DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
[2] Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
Azenbergstraße 12, D-70174 Stuttgart, Germany
[3] Research Institute for Linguistics, Hungarian Academy of Science
Benczúr u. 33, H-1068 Budapest, Hungary
[4] INRIA, France & HUB-ISDL, Berlin, Germany
[5] Seminar für Sprachwissenschaft, Universität Tübingen
Wilhelmstr. 19-23, D-72074 Tübingen, Germany
E-mail: declerck@dfki.de, eckartkn@ims.uni-stuttgart.de, piroska@nytud.hu, laurent.romary@loria.fr, thomas.zastrow@uni-tuebingen.de

Abstract

In our contribution to this workshop we propose incorporating standardized linguistic annotation in semantic resources of the cultural heritage domain, more specifically in the field of fairy tales. Although there are computational resources relevant for research in this area, these currently do not include linguistic annotation. We think here in particular to the The Proppian fairy tale Markup Language (PftML, see Malec, 2001), which is an annotation scheme that enables narrative function segmentation, based on hierarchically ordered textual content objects, but lacking linguistic information. We propose an approach to enrich PftML with standardized linguistic annotation, and so to support interoperability of linguistic information when it comes to combine it with annotation structures used in the eHumanites studies.

1. Introduction

In the context of both the CLARIN[1] and the D-SPIN[2] projects (http://www.sfs.uni-tuebingen.de/dspin), we are working towards the goal of making available language resources and technologies that could be supporting research in the field of eHumanties. As a specific case of this endeavour we present a strategy (that is by now partially implemented) for the integration of linguistic annotation and annotation of character roles and typed action descriptors in the literary genre of fairytales. For the latter, our departure point is the work by Vladimir Propp (Propp, 1968) and a XML schema, called PftML, for the annotation of fairy tales suggested by (Malec, 2004) We give here just some examples of Proppian functions[3]:

- Hero: a character that seeks something
- Villain: who opposes or actively blocks the hero's quest
- Donor: who provides an object with magical properties
- Dispatcher: who sends the hero on his/her quest via a message
- False Hero: who disrupts the hero's success by making false claims
- Helper: who aids the hero
- Princess: acts as the reward for the hero and the object of the villain's plots
- Her Father: who acts to reward the hero for his effort

Table 1: Some examples of Proppian functions

[1] http://www.clarin.eu
[2] http://www.sfs.uni-tuebingen.de/dspin
[3] http://www.adamranson.plus.com/Propp.htm

Looking at the concrete XML representation proposed by Scott Alexander Malec of Vladimir Propp's Morphology of the Folk Tale, one can notice that the text of the tale itself is annotated in a coarse-grained manner and following an inline annotation strategy. Below we can see an example:

```
<Folktale Title="The Swan-Geese" AT="480" NewAfanasievEditionNumber="113" ProppConformity="Yes"> ….
    <CommandExecution>
      <Command subtype="Interdiction">
"Dearest daughter," said the mother, "we are going to work. Look after your brother! Don't go out of the yard, be a good girl, and we'll buy you a handkerchief."
      </Command>
    <Execution subtype="Violated">
The father and mother went off to work, and the daughter soon enough forgot what they had told her. She put her little brother on the grass under a window and ran into the yard, where she played and got completely carried away having fun.
    </Execution>
    </CommandExecution>
```

Figure 1. A part of a tale annotated with Propp's functions

While in a closely related paper (Lendvai et al., 2010), we describe the whole integration chain, also introducing ontological resources modelling character roles and action descriptors in the fairy tale domain, we could not address the issue of the standardization of linguistic annotation we integrate with PftML or the ontologies. At the actual stage of work we use a configuration of natural language processing tools a supported by the WebLicht[4] web services, as they are implemented in the D-SPIN project. WebLicht makes use of (but is not restricted to) of TextCorpus format (*TCF*), which has been chosen for efficiency reasons for the internal process of the various levels of linguistic annotation that can be supported by WebLicht. Our aim is to map this format to the family of standards developed within TEI (Text Encoding Initiativce)[5] and ISO TC 37/SC4[6], also in order to verify the potential of those standards for serving as pivot format in the representation of textual and linguistic information. In the following we just present examples of the actual mapping of the TCF format, when applied to the text "Rotkäppchen" (*Little Red Riding Hood*), as it is stored in the Gutenberg project[7].

2. TEI Annotation

Figure 4: List of possible linguistic annotation for an ontology label

As a first step we apply the TEI encoding standard, so that we get clearly marked textual content objects. We distinguish here between the TEI header and the text properly speaking:

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0"
xmlns:ht="http://www.w3.org/1999/xhtml">
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title>Das RotkÃ€ppchen</title>
    <author>Charles Perrault</author>
    <respStmt>
     <resp>translator</resp>
     <persName>nacherzÃ€hlt von Moritz
Hartmann</persName>
    </respStmt>
    <respStmt>
     <resp>sender</resp>
     <persName>reuters@abc.de</persName>
    </respStmt>
   </titleStmt>
   <publicationStmt>
    <p>http://projekt.gutenberg.de; created in
20040916</p>
   </publicationStmt>
   <sourceDesc>
    <biblStruct>
     <monogr>
      <imprint>
       <publisher>Der Kinderbuchverlag
Berlin</publisher>
       <pubPlace>Berlin</pubPlace>
       <date when="1987"/>
      </imprint>
     </monogr>
     <idno type="isbn">3-358-00163-6</idno>
    </biblStruct>
   </sourceDesc>
  </fileDesc>
  <revisionDesc>
   <change when="2010-3-18"> Tokenised </change>
  </revisionDesc>
 </teiHeader>
```

In the following TEI example, the text properly speaking is encoded in markup that describes embedded textual content objects (<p> for paragraphs, <w> for words etc.:

```
<text>
  <front>
   <docAuthor>Charles Perrault</docAuthor>
   <docTitle>
    <titlePart>Das RotkÃ€ppchen</titlePart>
   </docTitle>
  </front>
  <body>
   <p>
    <w xml:id="t0">Es</w>
    <w xml:id="t1">war</w>
    <w xml:id="t2">einmal</w>
    <w xml:id="t3">ein</w>
    <w xml:id="t4">kleines</w>
    <w xml:id="t5">MÃ€dchen</w>
    <c xml:id="c0">,</c>
    <w xml:id="t6">ein</w>
    <w xml:id="t7">herziges</w>
    <w xml:id="t8">Ding</w>
    <c xml:id="c1">,</c>
    <w xml:id="t9">das</w>
    <w xml:id="t10">alle</w>
    <w xml:id="t11">Welt</w>
    <w xml:id="t12">liebhatte</w>
    <c xml:id="c2">.</c>
    <w xml:id="t13">Am</w>
    <w xml:id="t14">liebsten</w>
    <w xml:id="t15">hatte</w>
    <w xml:id="t16">es</w>
    <w xml:id="t17">die</w>
    <w xml:id="t18">GroÃ€mutter</w>
    <c xml:id="c3">,</c>
    <w xml:id="t19">die</w>
    <w xml:id="t20">kaufte</w>
    <w xml:id="t21">ihm</w>
    <w xml:id="t22">ein</w>
    <w xml:id="t23">MÃ€ntelchen</w>
    <w xml:id="t24">mit</w>
    <w xml:id="t25">einer</w>
    <w xml:id="t26">roten</w>
    <w xml:id="t27">Kapuze</w>
    <w xml:id="t28">daran</w>
    <c xml:id="c4">,</c>
```

[4] Details on the implementation of WebLicht, is given in http://weblicht.sfs.uni-tuebingen.de/englisch/weblicht.shtml

[5] http://www.tei-c.org/index.xml

[6] http://www.tc37sc4.org/

[7] http://www.gutenberg.org/wiki/Main_Page

```
<w xml:id="t29">und</w>
<w xml:id="t30">danach</w>
<w xml:id="t31">hieÃ</w>
<w xml:id="t32">es</w>
<w xml:id="t33">RotkÃ€ppchen</w>
<c xml:id="c5">.</c>
```

… .

```
<p>
 …..
<w xml:id="t135">sieh</w>
<w xml:id="t136">nicht</w>
<w xml:id="t137">rechts</w>
<c xml:id="c31">,</c>
<w xml:id="t138">nicht</w>
<w xml:id="t139">links</w>
<c xml:id="c32">,</c>
<w xml:id="t140">und</w>
<w xml:id="t141">lasse</w>
<w xml:id="t142">dich</w>
<w xml:id="t143">durch</w>
<w xml:id="t144">niemanden</w>
<w xml:id="t145">vom</w>
<w xml:id="t146">geraden</w>
<w xml:id="t147">Weg</w>
<w xml:id="t148">ablocken</w>
<c xml:id="c33">!</c>
<c xml:id="c34">Â«</c>
</p>
```

3. Morpho-Syntactic Annotation

On the top of TEI we are the MAF standard for morpho-syntactic annotation (http://pauillac.inria.fr/~clerger/MAF/html/body.1_div.5.html) , and link those to the words as they are marked by the TEI annotation (whereas still some alignment work is to be done, and some incertitudes in the mapping are still to be solved): The MAF notation refers to the “tokens” identified as <w> elements in the TEI annotation-

```
<?xml version="1.0" encoding="UTF-8"?>
<maf:MAF xmlns:maf="__">

<maf:tagset>
        <dcs local="KON" registered=
"http://www.isocat.org/datcat/DC-1262" rel="eq"/>
        <!-- __ -->
</maf:tagset>

<maf:wordForm tokens="t135">
 <fs>
  <f name="lemma"><symbol value="sehen"/></f>
  <f name="partOfSpeech"><symbol value="VVIMP"/></f>
  <f  name="grammaticalNumber"><symbol  value="singular"/></f>

 </fs>
</maf:wordForm>
<maf:wordForm tokens="t136">
 <fs>
  <f name="lemma"><symbol value="nicht"/></f>
  <f name="partOfSpeech"><symbol value="PTKNEG"/></f>
 </fs>
</maf:wordForm>
<maf:wordForm tokens="t137">
 <fs>
  <f name="lemma"><symbol value="rechts"/></f>
  <f name="partOfSpeech"><symbol value="ADV"/></f>
 </fs>
</maf:wordForm>
<maf:wordForm tokens="t138">
 <fs>
  <f name="lemma"><symbol value="nicht"/></f>
  <f name="partOfSpeech"><symbol value="PTKNEG"/></f>
 </fs>
</maf:wordForm>
<maf:wordForm tokens="t139">
 <fs>
  <f name="lemma"><symbol value="links"/></f>
  <f name="partOfSpeech"><symbol value="ADV"/></f>
 </fs>
</maf:wordForm>
<maf:wordForm tokens="t140">
 <fs>
  <f name="lemma"><symbol value="und"/></f>
  <f name="partOfSpeech"><symbol value="KON"/></f>
 </fs>
</maf:wordForm>
<maf:wordForm tokens="t141">
 <fs>
  <f name="lemma"><symbol value="lassen"/></f>
  <f name="partOfSpeech"><symbol value="VVIMP"/></f>
  <f name="grammaticalNumber"><symbol value="singular"/></f>
 </fs>
 </maf:wordForm><maf:wordForm tokens="t142">
 <fs>
  <f name="lemma"><symbol value="__"/></f>
  <f name="partOfSpeech"><symbol value="__"/></f>
  <f name="grammaticalNumber"><symbol value="singular"/></f>
  <f name="case"><symbol value="accusativeCase"/></f>
 </fs>
</maf:wordForm>
<maf:wordForm tokens="t143">
 <fs>
  <f name="lemma"><symbol value="durch"/></f>
  <f name="partOfSpeech"><symbol value="PREP"/></f>
 </fs>
</maf:wordForm>
<maf:wordForm tokens="t144">
 <fs>
  <f name="lemma"><symbol value="niemand"/></f>
  <f name="partOfSpeech"><symbol value="PIS"/></f>
  <f name="case"><symbol value="accusativeCase"/></f>
 </fs>
</maf:wordForm>
<maf:wordForm tokens="t145">
 <fs>
  <f name="lemma"><symbol value="vom"/></f>
  <f name="partOfSpeech"><symbol value="APPRART"/></f>
 </fs>
</maf:wordForm>
<maf:wordForm tokens="t146">
 <fs>
  <f name="lemma"><symbol value="gerade"/></f>
  <f name="partOfSpeech"><symbol value="ADJA"/></f>
  <f name="grammaticalNumber"><symbol value="singular"/></f>
  <f name="case"><symbol value="dativeCase"/></f>
  <f name="grammaticalGender"><symbol value="masculine"/></f>
 </fs>
```

```
</maf:wordForm>
<maf:wordForm tokens="t147">
 <fs>
  <f name="lemma"><symbol value="Weg"/></f>
  <f name="partOfSpeech"><symbol value="NN"/></f>
  <f name="grammaticalNumber"><symbol value="singular"/></f>
  <f name="case"><symbol value="dativeCase"/></f>
  <f name="grammaticalGender"><symbol value="masculine"/></f>
 </fs>
</maf:wordForm>
<maf:wordForm tokens="t148">
 <fs>
  <f name="lemma"><symbol value="ablocken"/></f>
  <f name="partOfSpeech"><symbol value="VVINF"/></f>
 </fs>
</maf:wordForm>

</maf:MAF>
```

We can not go into the details of the annotation here, but just to stress that in this way we have all the morpho-syntactic annotation attached to the TEI <w> elements.
We are currently working on mapping the syntactic annotation provided by the used configuration of WebLicht to the ISO SynAF model (http://www.iso.org/iso/catalogue_detail.htm?csnumber=37329).

The reader can see how the linguistic objects are pointing to the tokenized terms, and how the terms point then to the classes. On the basis of this model, we can obtain a matrix of linguistic objects, terms, and classes (including attributes and relations). This matrix can then deliver interesting insights on the use of natural language in knowledge representation systems. In the longer term, this can lead to proposal for a normalization of natural language expressions that fit best for building a terminology representing most adequately a formal representation of a domain.

4. Integration with the PftML annotation scheme

This step is straightforward: we take the functional annotation proposed by Scott A. Malec out of the document and include as an attribute the span of words that is in fact concerned by the Propp's function. This can look like:

```
<semantic_propp>
      <Command subtype="Interdiction" id="Command1"
     inv_id="Violated1" from="t135" to="t148">
</semantic_propp>
```

T135 and t148 are used here as defining a region of the text for which the Propp function holds. Navigating through the different types of IDs included in the multilayered annotation, the user can extract all kind of (possibly) relevant information.

We can also add to the functional annotation an additional ID which refers to a related detected function (here we point to the violation of the command that happens later in the text).
We plan also to use the ISO data category registry for entering the "labels" of Proppian functions (as for example shown in Table 1), with an adequate definition of those.

5. Acknowledgements

The research presented in this paper is partially funded by the European Commission in the context of the FP7 project CLARIN MONNET - Common Language Resources and Technology Infrastructure, with grant agreement number Grant Agreement Number 212230, and by the AMICUS network, which is sponsored by a grant from the Netherlands Organization for Scientific Research, NWO Humanities, as part of the Internationalization in the Humanities programme.

6. References

Afanas'ev, A. 1945. Russian fairy tales. Pantheon Books: New York.

Boas, H. 2005. From Theory to Practice: Frame Semantics and the Design of FrameNet. In: Semantisches Wissen im Lexikon, pp.129-160. Tübingen: Narr.

Jason, H. 1977. Precursors of Propp: Formalist Theories of Narrative in Early Russian Ethnopoetics. Poetics and Theory of Literature, 3, pp. 477-485.

Levi-Strauss, S. 1955. The structural study of myth. Journal of American Folklore, 68, pp. 428-444.

Lendvai, P, Declerck ,T., Darányi, S., Hervás R., Malec, S. and Peinado, F. 2010. Integration of Linguistic Markup into Semantic Models of Folk Narratives: The Fairy Tale Use Case. Proceedings of LREC 2010.

Malec, Scott A, 2004. Proppian structural analysis and XML modeling. http://clover.slavic.pitt.edu/sam/propp/theory/propp.html.

Peinado, F., Gervás, P., Diaz-Agudo, B. 2004. A Description Logic Ontology for Fairy Tale Generation. Proceedings of LREC.

Propp, V.J. 1968. Morphology of the folktale. University of Texas Press: Austin.

Takahashi, N, Ramamonjisoa, D., and Takashi, O. 2007. A tool for supporting an animated movie making based on writing stories in XML. IADIS International Conference Applied Computing

Tuffield, M. M., Millard, D. E. and Shadbolt, N. R. 2006. Ontological Approaches to Modelling Narrative. In: 2nd AKT DTA Symposium, January 2006, AKT, Aberdeen University.)