Proppian Content Descriptors in an Augmented Annotation Schema for Fairy Tales

Thierry Declerck[1], Antonia Scheidel[2], Piroska Lendvai[3]

Abstract. This paper describes a proposal for combining linguistic and domain specific annotation for supporting Cultural Heritage and Digital Humanities research, exemplified in the fairy tale domain. Our goal is to semi-automatically annotate fairy tales, in particular to locate and mark up fairy tale characters and the actions they are involved in, which can be subsequently queried in a corpus by both linguists and specialists in the field. The characters and actions are defined in Propp's structural analysis to folk tales, which we aim to implement in a fully fledged way, contrary to existing resources. We argue that the approach devises a means for linguistic processing of folk tale texts in order to support their automated semantic annotation in terms of narrative units and functions.

1 INTRODUCTION

Various theories in narratology research may assign properties to characters in different ways. For example, in actant theory (cf. the actant model, developed by [2], or [3] for more details) actants are positions, kind of behavioral patterns, in a story situation. Importantly, one and the same actor can serve as a different actant in different situations – as opposed to the classical view of possessing consistent roles throughout a story, advocated for example by folklorist Vladimir Propp (see [6]). According to Propp, main characters (or *dramatis personae*) that are occuring in a fairy tale may be the following[4]:

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

Additionally to the characters, Propp introduces the following concepts or units for the interpretation of Russian fairy tales:

31 Functions At the heart of the *Morphology of the Folktale* (see [6]) lies the description of actions that can be performed by the dramatis personae of a folktale. These so-called *functions* are the prototypical invariant features of fairy tales such as "Conflict", "Call for Help", "Kidnapping", "Test of Hero", and so on. Functions are frequently divided into sub-functions: in the case of function *A: Villainy*, they range from *A*^{1}*: The villain abducts a person* to *A*^{19}*: The villain declares war*. Functions and subfunctions are described in detail and illustrated with examples from Russian folktales in [6].

A sequence of all the functions from one folktale is called a *scheme* and can be used as a formal representation of the tale (see Fig. 1 for an example).

$$\gamma^1\beta^1\delta^1A^1C\uparrow\{[DE^neg.Fneg.]^3d^7E^7F^9\}G^4K^1\downarrow \\ [Pr^1D^1E^1F^9 = Rs^4]^3$$

Figure 1. Functional scheme for *The Magic Swan-Geese*

150 Elements. In Appendix I of *Morphology of the Folktale*, Propp provides what he calls a "list of all the elements of the fairy tale". The list contains 150 elements, distributed over six tables:

1. Initial Situation
2. Preparatory Section
3. Complication
4. Donors
5. From the Entry of the Helper to the End of the First Move
6. Beginning of the Second Move

Some of the 150 elements appear alone, others are grouped under a descriptive heading. If these "element clusters" (as shown in Fig. 2) are counted as one, the appendix contains 56 - as they shall tentatively be called in the following - narratemes. About a third of the narratemes can be mapped directly to functions, such as the aforementioned *30-32. Violation of an interdiction*. Other narratemes can be combined to form an equivalent to a function (together, narratemes *71-77: Donors* and *78: Preparation for the transmission of a magical agent* can presumably be considered as a superset to the information expressed by function *D: First Function of the donor*.

Another group of narratemes, however, goes beyond the 31 functions: *70. Journey from home to the donor*, for example, can be seen as filling the gap between the functions ↑*: Departure* and *D: First Function of the donor*. The first table (*Initial Situation*[5]) contains a

[1] DFKI GmbH, Language Technology Lab, Germany, declerck@dfki.de
[2] DFKI GmbH, Language Technology Lab, Germany, Antonia.Scheidel@dfki.de
[3] Research Institute for Linguistics, Hungarian Academy of Sciences, piroska@nytud.hu
[4] Source: http://www.adamranson.plus.com/Propp.htm

[5] Propp makes use of the symbol α*: Initial Situation* to refer to everything that happens before the hero's parents announce their departure, but it is not a function as such.

```
30-32. Violation of an interdiction
      30. person performing
      31. form of violation
      32. motivation
```

Figure 2. Example for a narrateme

multitude of narratemes dedicated to the circumstances of the hero's birth and other events/situations which precede the actual adventure.

Furthermore, Table 1 (Initial Situation) includes two "element-clusters" describing the hero and false hero, respectively, in term of 'future hero' (see Fig. 3). .

```
10-15. The future hero
      10. nomenclature; sex
      11. rapid growth
      12. connection with hearth, ashes
      13. spiritual qualities
      14. mischievousness
      15. other qualities
```

Figure 3. Example for an element cluster serving as profile

A closer examination of the appendix reveals such profiles for each of the dramatis personae, although sometimes spread over two clusters or narratemes.

The longer term objective of our work consists in devising a means for linguistic processing of folk tale texts in order to support their automated semantic annotation in terms of Propp's theory, using an appropriate encoding schema, but which could be adapted to other theories of fairy tales or literary genres. As a starting point for our work, we analysed available resources that could be used or re-used in our work, and make a among others a first proposal for an augmented XML annotation schema.

2 RESOURCES

Similarly to the well-developed application of NLP technologies to certain domains (e.g. financial news, biomedicine), it is equally important albeit less trivial in Digital Humanities to identify a set of textual as well as domain-specific semantic resources based on which complex information can be gained and modeled.

2.1 Textual Resources

About 200 fairy tales are available from the Afanas'ev collection (see [1]) that is English translations of the Russian originals, on which Propp based his model. Additionally, popular tales such as "Little Red Riding Hood" exist in many versions in many languages. We focus in this paper on one of the German versions of this tale, to illustrate our proposed combination of linguistic annotation and Proppian character functions. This annotation exercise is part of the D-SPIN project[6] and a first prototype of an automated annotation of a multilingual version of this tale is part of a use case of the CLARIN project[7]. The annotation exercise is planned to be extended to all Grimm tales, as they become available within the Gutenberg project[8].

[6] http://weblicht.sfs.uni-tuebingen.de/englisch/index.shtml
[7] http://www.clarin.eu/external/
[8] See http://www.gutenberg.org/

2.2 Semantic resources

There exist a number of computational models for processing, annotation, and generation of fairy tales, in the form of semantic resources or annotation schemas. These describe narration in terms of moves and their ingredients, often drawing on the work of Propp. We consider two ontologies that model certain aspects of Propp: ProppOnto ([5]) and [8][9], built for generation and control purposes in interactive digital storytelling and games. It is notable that such resources typically do not specify or type which linguistic elements need to be associated with the model's constituents (e.g. concepts, relations) in order to express a domain-specific function in natural language. In other words, the semantic information is exclusively encoded in the ontology classes, with no link to potential linguistic realizations, allowing little or no flexibility of reusing the resource across languages or across domains.

A central part of the MONNET project[10] consists of modeling linguistic information in ontologies; we are extending the default domains of application of MONNET (i.e. financial reporting, eGovernment) to Digital Humanities.

2.3 Annotation schemas

Additionally, there are a few XML-based annotation schemas for Proppian functions we are aware of: the Proppian Fairy Tale Markup Language (PftML)[11], and one supporting the generation of animated movies, based on Proppian functions [7].

While the first schema is used for analysis purposes, it remains at a very coarse-grained level, allowing inline textual markup, which is typically assigned at the sentence or paragraph level. Association of specific linguistic expressions with the functions is not supported. The second schema pertains to word level, but only of (unstandardized) semantic features that integrate generic semantic roles (agent, location, etc.) and Proppian functions; linguistic information is not considered at all. Neither of the schemas support cross-referencing of objects (i.e., roles and actions) in the text, which would be a desirable property of resources in Humanities research. More detailed description on our former work with PftML and ProppOnto is given in [4].

3 AUGMENTING THE PftML SCHEMA AND THE SEMANTIC RESOURCES

In our investigation of the available resources for our work, we noticed that many of these incorporate only a subset of the elements described by Propp, depending on the application at hand. So for example PftML is concentrating on the functions but does not adress the elements. Therefore, one of our first steps consisted of creating an augmented annotation schema involving all elements of fairy tales that were discussed by Propp, as described above.

Below we give a preliminary instantation of the augmented annotation schema, which we call APftML (Augmented Proppian fairy-tale Markup Language), currently under development. APftML is intended to afford a fine-grained, stand-off annotation of folk (or fairy)

[9] See http://www.fdi.ucm.es/profesor/fpeinado/projects/kiids/apps/protopropp/ respectively http://eprints.aktors.org/440/01/tuffieldetal.pdf
[10] MONNET – Multilingual Ontologies for Networked Knowledge – is an R&D project co-funded by the European Commission, with grant 248458. See http://cordis.europa.eu/fp7/ict/language-technologies/project-monnet_en.html
[11] developed by S. Malec, see http://clover.slavic.pitt.edu/sam/propp/theory/propp.html

tales in accordance with concepts introduced in Vladimir Propps Morphology of the Folktale. It is loosely based on PftML. APftML will integrate / interact with ProppOnto.

```
<annotation>
  <InitialSituation>
  <Content>Es war einmal eine kleine
  suesse Dirne, die hatte jedermann lieb,
  der sie nur ansah, am allerliebsten
  aber ihre Grossmutter, die wusste gar nicht,
  was sie alles dem Kinde geben sollte.
  Einmal schenkte sie ihm ein Kaeppchen
  von rotem Sammet, und weil ihm das
  so wohl stand und es nichts anders
  mehr tragen wollte, hiess es nur
  das Rotkaeppchen.</Content>
  </InitialSituation>
...
</annotation>
```

In the given example just above, we introduce with ”Content” element an explicit way of encoding the exact portion of the text that is interpreted as describing the ”InitialSituation”[12].

```
<Narrateme>
  <Command subtype="command" id="i0">
     Eines Tages sprach seine Mutter zu ihm:
     Komm, Rotkaeppchen, da hast du ein Stueck
     Kuchen und eine Flasche Wein,
     bring das der Grossmutter hinaus;
     sie ist krank und schwach und
     wird sich daran laben.</Command>
  <Agent id="p0">seine Mutter</Agent>
  <Patient id="p1">Rotkaeppchen</Patient>
  <Content>bring das der Grossmutter hinaus
  </Content>
  <Motivation>sie ist krank und schwach und
    wird sich daran laben</Motivation>
</Narrateme>
```

In this second example, the segment of the text containing the ”Command” is given again by the ”Content” element. This annotation also introduces semantic information on ”agent” and ”patient” and we provide for an index for the function ”Command”, so that we can refer to it if the text we encounter a textual segment. Those new elements and features are major additions to PftML, and include also Proppian elements that were not consistently used in the previously existing ontologies.

4 TOWARDS A STANDARDIZED TEXTUAL AND LINGUISTIC ANNOTATION

For the annotation of the textual and linguistic information we suggest to use the family of standards developed within TEI (Text Encoding Initiatitive)[13] and ISO TC 37/SC4[14], also in order to verify the potential of those standards for serving as pivot format for the representation of textual and linguistic information. 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 itself, as the two examples below show:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0"
 xmlns:ht="http://www.w3.org/1999/xhtml">
 <teiHeader>
 <fileDesc>
- <titleStmt>
  <title>Rotkaeppchen</title>
- <respStmt>
  <resp>collector</resp>
  <persName>Gebrueder Grimm</persName>
  </respStmt>
  </titleStmt>
- <publicationStmt>
  <p>"http://gutenberg.spiegel.de/?i
     d=5&xid=969&kapitel=230&cHash=
     b2042df08brotk#gb_found"</p>
  </publicationStmt>
  <sourceDesc />
  </fileDesc>
- <revisionDesc>
  <change when="2010-06-16">Tentative
    Annotation</change>
  </revisionDesc>
  </teiHeader>

<body>
- <p>
  ...
  <w xml:id="t3">Es</w>
  <w xml:id="t4">war</w>
  <w xml:id="t5">einmal</w>
  <w xml:id="t6">eine</w>
  <w xml:id="t7">kleine</w>
  <w xml:id="t8">suesse</w>
  <w xml:id="t9">Dirne</w>
  <w xml:id="t10">,</w>
  <w xml:id="t11">die</w>
  <w xml:id="t12">hatte</w>
  <w xml:id="t13">jedermann</w>
  <w xml:id="t14">lieb</w>
  <w xml:id="t15">,</w>
  <w xml:id="t16">der</w>
  <w xml:id="t17">sie</w>
  <w xml:id="t18">nur</w>
  <w xml:id="t19">ansah</w>
  <w xml:id="t20">,</w>
  ...
 </body>
```

The TEI encoding of the body is including the information about the tokenization of the text, and this is an anchor point for the subsequent annotation levels, using a stand-off strategy: all the annotation levels can point back to the original text on the base of the numbering of the tokens.

4.1 Morpho-syntactic Annotation

On the top of TEI annotation we are applying the ISO-MAF standard for morpho-syntactic annotation [15], linking its elements to the words

[12] We do not provide for a translation here, since we assume the story to be well known.
[13] http://www.tei-c.org/index.xml
[14] http://www.tc37sc4.org/
[15] see http://pauillac.inria.fr/ĉlerger/MAF/html/body.1_div.5.html

as they are marked by TEI, using the token IDs we introduced into the 'w' elements:

```
<?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>
...
```

This specific morpho-syntactic annotation, using an XML representation of feature structures, is pointing to the token number 135, which in the text is the verb "see" in a particular grammatical form (i.e. the imperfect). We are currently working on the syntactic annotation following the guidelines of ISO SynAF[16]. This annotation is building on the top of MAF and is annotating (groups of) words with constituency (e.g. nominal or verbal phrases, etc.) and dependency information (subject, object, predicate etc.). The idea is that an identified subject of a sentence can usually be mapped onto the "Agent" element of a Proppian function. But here we have to take into account also the modus of the sentence (Active vs Passive modus).

5 INTEGRATION WITH THE APftML ANNOTATION SCHEMA

This step is straightforward: we take the functional annotation and add the proper indexing information, so that all the descriptors of the functional annotation are linked to the available levels of linguistic annotation. 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 multi-layered linguistic annotation, the user can extract all kind of (possibly) relevant linguistic information and combine it with the functional annotation in terms of Propp.

On the basis of this combination of distinct types of annotation – linguistic and (Proppian) functional–, we expect that the work of finding variations in fairy tales can be enhanced, but most significant is probably the fact that it is getting much easier to pursue text-based research on tales.

6 CONCLUSIONS

We described ongoing work in adapating and augmenting existing annotation schemas of fairy tales. We described also a strategy for using natural language processing in order to support the automated markup of character roles and functions. Generalizing the results will shed light on the computational applicability of a manually created Humanities resource, in our case of Propp's method for narratives, in Digital Humanities research. If we can detect genre-specific narrative units on evidence based on a linguistically annotated corpus, we plan to take this research further and analyse higher level motifs (such as narratemes), as well as other types of narratives.

[16] http://www.iso.org/iso/catalogue_detail.htm?csnumber=7329

ACKNOWLEDGEMENTS

The ongoing work described in this paper has been partially supported by the European FP7 Project "MONNET" (Multilingual Ontologies for Networked Knowledge), with Grant 248458, and by the BMBF project "D-SPIN"[17]. Investigating higher-order content units such as motifs is the focus of the AMICUS project[18], which is supported by The Netherlands Organisation for Scientific Research (NWO).

REFERENCES

[1] A. Afanas'ev, *Russian fairy tales*, Pantheon Books, New York, 1945.
[2] A. J. Greimas, *Sémantique structurale*, Larousse, Paris, 1966.
[3] D. Herman, 'Pragmatic constraints on narrative processing: Actants and anaphora resolution in a corpus of North Carolina ghost stories', *Journal of Pragmatics*, **32(7)**, 959–1001, (2000).
[4] Piroska Lendvai, Thierry Declerck, Sándor Darányi, Pablo Gervás, Raquel Hervás, Scott Malec, and Federico Peinado, 'Integration of linguistic markup into semantic models of folk narratives: The fairy tale use case', in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, ed., Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Mike Rosner Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, Valletta, Malta, (May 2010). European Language Resources Association (ELRA).
[5] F. Peinado, P. Gervas, and B. Diaz-Agudo, 'A description logic ontology for fairy tale generation', in *Proceedings of LREC*. ELRA, (5 2004).
[6] V.J. Propp, *Morphology of the folktale*, University of Texas Press:, Austin, 1968.
[7] N. Takahashi, D. Ramamonjisoa, and O. Takashi, 'A tool for supporting an animated movie making based on writing stories in xml', in *Proceedings of IADIS International Conference Applied Computing*, (2007).
[8] M. M. Tuffield, D. E. Millard, and N. R. Shadbolt, 'Ontological approaches to modelling narrative.', in *Proceedings of the 2nd AKT DTA Symposium*. Aberdenn University, (1 2006).

[17] http://weblicht.sfs.uni-tuebingen.de/englisch/index.shtml
[18] http://amicus.uvt.nl