Propp Revisited: Integration of Linguistic Markup into Structured Content Descriptors of Tales

Piroska Lendvai¹, Thierry Declerck², Sándor Darányi³, Scott Malec⁴

1Research Institute for Linguistics, Hungarian Academy of Sciences
2Language Technology Lab, DFKI GmbH
3Swedish School of Library and Information Sciences, University College Boras/Göteborg University
4Carnegie Mellon University, Pittsburgh

1 Introduction

Metadata that serve as semantic markup, such as conceptual categories that describe the macrostructure of a plot in terms of actors and their mutual relationships, actions, and their ingredients annotated in folk narratives, are important additional resources of digital humanities research. Traditionally originating in structural analysis, in fairy tales, they are called functions (Propp, 1968), whereas in myths – mythemes (Lévi-Strauss, 1955); a related, overarching type of content metadata is a folklore motif (Uther, 2004; Jason, 2000).

In his influential study, Propp treated a corpus of tales in Afanas’ev’s collection (Afanas’ev, 1945), establishing basic recurrent units of the plot (‘functions’), such as Villainy, Liquidation of misfortune, Reward, or Test of Hero, and the combinations and sequences of elements employed to arrange them into moves¹. His aim was to describe the DNA-like structure of the magic tale sub-genre as a novel way to provide comparisons. As a start along the way to developing a story grammar, the Proppian model is relatively straightforward to formalize for computational semantic annotation, analysis, and generation of fairy tales. Our study describes an effort towards creating a comprehensive XML markup of fairly tales following Propp’s functions, by an approach that integrates functional text annotation with grammatical markup in order to be used across text types, genres and languages.

The Proppian fairy tale Markup Language (PftML) (Malec, 2001) is an annotation scheme that enables narrative function segmentation, based on hierarchically ordered textual content objects. We propose to extend PftML so that the scheme would additionally rely on linguistic information for the segmentation of texts into Proppian functions. Textual variation is an important phenomenon in folklore, it is thus beneficial to explicitly represent linguistic elements in computational resources that draw on this genre; current international initiatives also actively promote and aim to technically facilitate such integrated and standardized linguistic resources. We describe why and how explicit representation of grammatical phenomena in literary models can provide interdisciplinary benefits for the digital humanities research community.

¹The full list of functions is available at http://clover.slavic.pitt.edu/sam/propp/praxis/features.html
In two related fields of activities, we address the above as part of our ongoing activities in the CLARIN\(^2\) and AMICUS\(^3\) projects. CLARIN aims to contribute to humanities research by creating and recommending effective workflows using natural language processing tools and digital resources in scenarios where text-based research is conducted by humanities or social sciences scholars. AMICUS is interested in motif identification, in order to gain insight into higher-order correlations of functions and other content units in texts from the cultural heritage and scientific discourse domains. We expect significant synergies from their interaction with the PftML prototype.

2 Proppian fairy tale Markup Language (PftML)

Creating PftML was based on the insight that Propp’s functions – organized in tables to categorize his observations – were analogous to metadata, and as such renderable by hierarchically arranged elements in eXtensible Markup Language (XML) documents. A tale in the corpus consists of one or more moves and on a lower level of functions which are modeled as elements. Function elements themselves have XML attributes that allow for the efficient extraction of data from the text using XQuery from within a native XML database. The embedded structure of Proppian functions as represented by PftML markup is illustrated in Fig. 1 by an excerpt from the English translation of the Russian fairy tale *The Swan-Geese*.

Note that Proppian functions are applied to relatively long, semantically coarse-grained textual chunks, i.e. sentences, but linguistic elements that convey a function actually encompass a shorter sequence of words; e.g. contrary to the markup in the example, both *Command* and *Execution* only pertain to linguistic units smaller than full sentences.

3 Integration of PftML with linguistic annotation

We propose to combine PftML with a stand-off, multi-layered linguistic markup scheme to ensure modularity and reusability of linguistic information associated with textual elements, supporting interoperability of fairy tales annotation in different languages and versions. As seen in Fig. 1, PftML is interleaving the Proppian annotation with the text. This in-line annotation strategy has some drawbacks: a text can hardly be annotated in fine-grained manner without losing readability, or with information originating from different sources e.g. indicating different views on narrative functions.

Stand-off annotation strategy, following the standardization initiatives for language resources conducted within ISO\(^4\), stores annotation separately from the original text, linking these by referencing mechanisms. We adopt the ISO multi-layered annotation strategy, representing linguistic information on the following levels: segmentation of the text in tokens; morpho-syntactic properties of the tokens; phrasal constituencies; grammatical dependencies; semantic relations (e.g. temporal, co-referential), cf. (Ide and Romary, 2006).

\(^2\)http://www.clarin.eu
\(^3\)http://ifl.k.uv.nl/amicus
\(^4\)http://www.tc3?sc4.org
We illustrate how the linguistic annotation layers can be combined with the PftML annotation in one stand-off annotation file, showing here only the morphosyntactic and constituency annotation, as they are applied to the first five tokens of the sub-sentence annotated with the 'Violated Execution' function in Fig. 1. In the morphosyntactic annotation, the value of the TokenID of the 12th word is pointing to the original data (e.g., daughter is the 12th token in the text).\(^5\)

\(^5\)Note that the original string is normally not present in these layers but are displayed in annotation examples for readability's sake.
In the constituency annotation level displayed below, words are grouped into syntactic constituents (e.g., the nominal phrase the daughter). The span of constituents is marked by the value of the features from and to, which are pointing to the previous morpho-syntactic annotation layer.

```xml
<phrases>
  <phrase id="p4" from="w11" to="w12" type="NP">the daughter</phrase>
  <phrase id="p5" from="w13" to="w14" type="ADV">soon enough</phrase>
  <phrase id="p6" from="w15" to="w15" type="VC">forgot</phrase>
  <phrase id="p7" from="w16" to="w20" type="REL_COMP">what they had told her</phrase>
  ...
</phrases>
```

PtML and (for example) word-level annotation can be combined in one stand-off XML element, where each specific PtML annotation receives a span of textual segments associated with it:

```xml
<Execution subtype="Violated" id="e1" inv_id="Command1" from="w11" to="w21">
</Execution>
```

The values w11 and w21 are used for defining a region of the text for which the Propp function holds; Command1 refers to the Interdiction function label used earlier in the text.\(^6\)

## 4 Benefits for humanities research

The integrated annotation scheme enables narrative segmentation enhanced by additional information about the linguistic entities that constitute a given function. A folklore researcher might be interested in which natural language expressions correspond to which narrative function: in the `<Execution subtype="Violated">` example, forgot can be an indicator of this function. In fact, it is also relevant to signal that forgot is a verb, and to reduce the strings forgets, forgot, forgotten to one lemma (i.e. base form), so that all morphological forms are retrieved when any of these variants is queried.

Navigating through the different types of IDs included in the multilayered annotation, a researcher can obtain statistics over linguistic properties of fairy tales. For example, the grammatical subject of a function can be extracted, e.g. to see which characters participate in commands and their violation. Note that if – according to the current scheme – the narrative function boundaries are imprecise, the `<Execution subtype="Violated">` function in our example sentence would incorrectly contain two grammatical – and three semantic – subjects (father and mother, and daughter).

Linguistic information will enable detecting functions that refer to each other, as syntax and semantics of sentence pairs in such relations mirror – at least partly – each other, e.g. Don’t go out of the yard and ran onto the street. Detecting cross-reference in turn contributes to identifying a function’s core elements, which is a crucial step in understanding the linguistic vehicles by which motifs operate and the degree of variation and optionality they allow.

---

\(^6\) We started to implement this work within the D-SPIN project (see http://www.sfs.uni-tuebingen.de/dspin), which is the German complementary project to CLARIN.
5 Concluding remarks

Since the content descriptors in PftML might pertain to textual material on the supra- or sub-sentential level, there is a need to investigate the mechanisms underlying the assignment of a function to a span of words. We propose to tackle this issue based on linguistic analysis, hypothesizing that boundaries of certain linguistic objects overlap with boundaries of Proppian functions. A direct consequence of more precise segmentation of functions is that linguistic characterization, retrieval, and further computational processing of texts from the folktale genre will improve, and facilitate detecting higher-level, domain-specific cognitive phenomena. It would also become feasible to detect from corpus evidence if there exist additional functions beyond Propp’s scheme.

Integration along the above lines with ontological resources of fairy tales is described in a separate study by us (Lendvai et al., 2010). We expect from our strategy – applied to tales in different versions in different languages – to lead to the generation of a multilingual ontology of folktale content descriptors, which would be extending the efforts of the MONNET project7, originally focusing on financial and governmental issues. In future work we plan to address embedding our annotation work into the TEI framework8, and extend the ISO strategy on using well-defined data categories for linguistic annotation labels9 to those of functions corresponding to PftML labels, to facilitate porting our approach to other literary genres.

References


7Multilingual ONtologies for NETworked Knowledge, see http://cordis.europa.eu/fp7/ict/language-technologies/project-monnet_en.html

8http://www.tei-c.org/index.xml

9see http://www.isocat.org/