Multilingual and Semantic Extension of Folktale Catalogues

Thierry Declerck[1,2], Piroska Lendvai[3], Sándor Darányi[4]

[1] DFKI GmbH, Stuhlsatzenhausweg, 3
66123 Saarbrücken, Germany
declerck@dfki.de

[2] ICLTT, Austrian Academy of Sciences,
Sonnenfelsgasse 19/8, 1010 Wien, Austria

[3] Research Institute for Linguistics
Benczúr u. 33. H-1068 Budapest, Hungary
lendvai.piroska@nytud.mta.hu

[4] Swedish School of Library and Information Science
50190 Borås, Allégatan 1, Sweden
sandor.daranyi@hb.se

Abstract We address the multilingual and semantic upgrades of two digital catalogues of motifs and types in folk-literature: the Thompson's Motif-Index of Folk-Literature (TMI) and the Aarne-Thompson-Uther classification system (ATU). The methods convert, translate, and represent their digitized content in terms of various (so far often implicit) structural and linguistic components. The results will enable (i) utilizing these resources for semi-automatic analysis and indexing of texts of relevant genres, in a multilingual setting, and (ii) pre-processing the data, for analysing motif sequences in folktale plots. We plan to publish the resulting data, which can be made available in the Linked Open Data (LOD) framework.

1. Introduction

The modelling of phenomena related to higher-order knowledge mechanisms poses significant challenges to research in any domain, this being also valid for narratives, which "are an important form of knowledge representation", as (Tuffield et al, 2006) state. In the context of the AMICUS network[1] we are more specifically interested in the representation of motifs in narratives. Motifs are recurring conceptual, textual, audio or visual units appearing in artefacts — in folk tales, they can be seen as cognitively complex notions (e.g. "Rescue of princess", "Helpful animal", "Cruel stepmother", etc.), expressed by lexically and syntactically variable narrative structures. The modelling of such motifs, including their typical realizations in natural language form, can help in supporting the automatic motif detection in large (multilingual) text collections[2]. But this kind of formal representation of motifs is so far unresolved, and thus a large amount of cultural heritage collections of which motifs are typical constructive units can still only be manually indexed, which significantly limits access to these resources and to this type of knowledge. Recently, we started to investigate the utility of linguistic and semantic analysis and mark-up of motifs (Lendvai et al., 2010), which we would like to extend and apply to studies on motif sequencing (Darányi, Wittek and Forró, 2012).

In the current study, we address two priorities of the Digital Humanities discipline: devising procedures that integrate semantic enhancement of legacy folk tale indexes, classification systems and taxonomies (Declerck and Lendvai, 2011) and their automatized translation (Mörth et al., 2011). For this study we are dealing with two extended catalogues that hold conceptual schemes of narrative elements from folktales, ballads, myths, and related genres: the *Thompson's Motif-Index of Folk-Literature*, TMI (Thompson, 1955) and *The Types of International Folktales*, ATU (Uther, 2004). TMI has an available on-line version[3], only in English language, but the digitized ATU is not yet available on-line in its entirety, only some of its subsections are reproduced in Wikipedia, in various languages[4].

TMI indexes and ATU types are both combined with extensive labels, and a novel approach to those resources is that this combination can be linguistically processed (Declerck and Lendvai, 2011), semantically represented, and, ultimately, turned into domain-specific ontology classes that can be interlinked. The upgrade leads to semantically enriched catalogues that qualify as interoperable language technology resources that can be harnessed to assign text units to the folktale domain classes, creating in an automatized way indexed folktale corpora. As part of such a normalization process, catalogues have to be made interoperable with each other, and stored in a semantically harmonized representation, like the SKOS[5] standard, which is using RDF[6] as its formal representation language.

2. Towards Semantic and Multilingual Extensions of TMI

The Thompson's Motif-Index of Folktale-Literature is organized by alphanumeric indexes, which resembles a taxonomy structure of motifs, but does not express hierarchy or inheritance properties. i.e. it is not made explicit that some elements of the taxonomy introduce mere classification information over a span of labels *("A0-A99. Creator"*, split into finer-grained subclasses, e.g. *"A20. Origin of the creator.")*, that some elements are abstractions of motifs *("A21. /Creator from above./")*, and that some elements are summaries of a concrete motif, supplying source information as well *("A21.1. /Woman who fell from the sky./--Daughter of the sky-chief falls from the sky, is caught by birds, and lowered to the surface of the water. She becomes the creator.--*Iroquois: Thompson Tales n.27.--Cf. Finnish: Kalevala rune 1.")*.

We prepared a program that converts TMI to an XML representation and marks such properties explicitly by designated tags, as exemplified below:

```
<label class="TMI_A0" span="A0-A99" type="abstract" lang="en">Creator</label>
<label class="TMI_A20" span="A21-A27" type="abstract" Property_Of="A0" lang="en">Origin of the Creator</label>
<label class="TMI_A21" span="A21.1-A21.2" type="abstract" SubClassOf="A20" lang="en">Creator from Above</label>
<label class="TMI_A21.1" span="A21.1-A21.1" type="concrete" SubClassOf="A21" lang="en">Woman who fell from the sky</label>
```

This representation makes explicit the fact that the natural language expressions (like "Creator from Above") are labels of classes that we explicitly organise in a class hierarchy. The feature "type" can take two values: "abstract" or "concrete". The latter is indicating that the index (for example: A21.1) is pointing to concrete examples in selected tales. The "span" feature is indicating the number (from 1 to many) of subsumed indexes. We noticed that indexes ("classes" in our terminology), which are in fact leaves (spanning only over their own number) point all to concrete tales, and can thus be considered as first level motifs, whereas the other classes have more a classification role.

Ongoing work is dedicated to upgrading the XML representation to SKOS-RDF, to provide adequate means for differentiating between hierarchical realizations and properties associated with classes, and the possibility to compute inheritance properties of the class hierarchy. SKOS-RDF is also appropriated for publishing the enriched TMI resource on the LOD. We display below a simplified example (where "TMI_A*" are shortcuts for URIs):

```
<TMI_A0> rdf:type skos:Concept ;
    skos:prefLabel "Creator"@en.
<TMI_A20> rdf:type skos:Concept ;
    skos:prefLabel "Origin of the Creator"@en
    skos:related <TMI_A0>.
<TMI_A21> rdf:type skos:Concept ;
```

```
    skos:prefLabel "Creator from Above"@en.
<TMI_A21.1> rdf:type skos:Concept ;
    skos:prefLabel "Woman who fell from the sky"@en ;
    skos:broader <TMI_A21>
```

In parallel, we target the extension of motifs listed in TMI in English into a multilingual version. This is carried out by accessing the *Wiktionary* lexicon[7], via the LOD-compliant *lexvo*[8] service. Actual work (Declerck et al. 2012) is aiming at extracting from Wiktionary a multilingual lexical semantics network for the labels included in TMI.

3. Towards the Harmonization of Multilingual ATU On-line Data

As we mentioned above, only segments of ATU are available on-line, in the context of Wikipedia articles, in different languages. We note the following discrepancies in these:

(EN) Rapunzel 310 (Italian, Italian, Greek, Italian)
(DE) AaTh 310 Jungfrau im Turm KHM 12 *Rapunzel*
(FR) AT 310 : « La Fille dans la tour » (*The Maiden in the Tower*) : version allemande

The English Wikipedia links *Rapunzel* to four Wikipedia pages on tales belonging to the same ATU type. Rhe Wikipedia page for the original Rapunzel tale is reached if the reader clicks on “Rapunzel”. The German page links additionally to the German classification KHM (*Kinder- und Hausmärchen*). The French Wikipedia page gives an English translation of the French naming, while the French title of the Rapunzel tale is accessible only if the user clicks on the link “version allemande”, leading to the French Wikipedia page “Raiponce”). There is a clear need to structure this disparate information in one representation format. We turned the basic information from the Wikipedia pages into an integrated SKOS representation:

```
<ATU_310> rdf:type skos:Concept;
  skos:prefLabel "Rapunzel"@en;
  skos:altLabel "The Maiden in the Tower"@en;
  skos:prefLabel "Jungfrau im Turm"@de;
 skos:altLabel "Rapunzel"@de;
  skos:prefLabel "La Fille dans la Tour"@fr;
  skos:altLabel "Raiponce"@fr.
```

On the basis of a small fragment of such aligned information from Wikipedia pages, a representative multilingual terminology of ATU terms can be aggregated, and this terminology can be re-used for supporting the translation of other labels of ATU or of TMI. We investigate for this terminology alignment techniques used in the Machine Translation field, adapting them to the short terms that are employed in the catalogues.

4. Relating the Upgraded Representations of TMI and ATU

The SKOS vocabulary allows to establish matching relations between the upgraded TMI and ATU catalogues, so that for example the motif TMI_A2223.1. “Cat helps man build house: may occupy chimney corner.” can be linked to ATU_545 “The Cat as Helper":

```
<TMI_A2223.1> rdf:type skos:Concept ;
    skos:prefLabel "Cat helps man build house: may occupy chimney corner"@en.
<ATU_545> rdf:type skos:Concept;
  skos:prefLabel "The Cat as Helper"@en.
<TMI_AA2223.1> skos:relatedMatch <ATU_545>
```

Whereas we still do not take decisions on the type of matching relation, which could be "broader", "narrower" or "close". Additionally we currently investigate the concrete linguistic and semantic markup of the tokens used in labels, using another RDF-based formalism: the *lemon*[9] representation scheme. Precise linguistic mark-up allows establishing in an automated way this kind of SKOS matching, since the noun "Helper" can be marked as a derivation of the verb "help". This work also supports the building of a lexical semantics network for folktales, based on the labels used in TMI and ATU.

5. Towards the Automated Analysis of Motif Sequences in Folktale Plots

The semantically upgraded TMI and ATU resources incorporate information about class hierarchy, which allows potential users to query for motifs that are connected on the same or on different levels of the taxonomy. This is feature that could be exploited for the analysis or generation of storytelling scenarios, since plot graphs of folktales are like watersheds, progressing by connecting motifs at different hierarchical levels as points at various heights. Thus the semantically upgraded TMI could help to learn about the concatenation patterns of motif categories used in tales In turn, one could use such category information as a source of probabilities for network construction. Such probabilistic networks, prominently Hidden Markov Models, are a highly significant research topic in bioinformatics, a field as much utilizing the concept of motifs as narratology, therefore methodological cross-pollination is a totally realistic option. Darányi *et al*. (2012) have recently shown that ATU tale types as motif strings exemplify the basic recombination patterns of chromosome mutations already on a limited sample of types, and more new findings can be expected by an extension of the quest for "narrative DNA". Further, the manual tagging of tale types by motifs in ATU is known to be incomplete, so to repeat this tagging by automatic means by annealing motif definitions and NLP-based term or phrase extraction is an ultimately fruitful endeavour.

6. Conclusion

In this study, we focus on converting TMI and ATU into adequate semantic resources, compliant with linguistic and terminological requirements, to enable multilingual, content-level indexing of folktale texts. TMI and ATU in their paper form have mainly been used for manual assignment of English-language metadata. To upgrade these established classification systems of folk narratives, we specify a development program that enhances and links them using language processing and semantic technologies. Recent work was dedicated to establish terminological relationships between both catalogues, using for this the SKOS framework establishing semantic links between specific labels of both semantically enriched catalogues. Our work will lead to the deployment of such resources in the LOD framework, complementing the work of national libraries that already ported their bibliographic data to the LOD. This may not only facilitate automated text indexing, but also allow for more detailed analysis of motif sequencing.

Notes

[1] AMICUS (Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts), http://ilk.uvt.nl/amicus/.

[2] The interconnection of Knowledge Representation systems and automated textual analysis is being successfully applied to various domains. One of our goal is to adapt and apply such technologies to the automated analysis of folk narratives, using for example the approach presented in (Declerck and Koleva, 2012), to be extended among others by ontology elements described in (Zöllner-Weber, 2008).

[3] See http://www.ruthenia.ru/folklore/thompson/index.htm.

[4] See http://en.wikipedia.org/wiki/Aarne%E2%80%93Thompson_classification_system for the English version, http://de.wikipedia.org/wiki/Aarne-Thompson-Index for the German version, and for the French version http://fr.wikipedia.org/wiki/Classification_Aarne-Thompson. Note that the online ATU data do not reflect the layout of the original catalogue, as was the case for TMI.

[5] SKOS stands for "Simple Knowledge Organization System". See for more details: http://www.w3.org/2004/02/skos/.

[6] RDF stands for "Resource Description Framework". See http://www.w3.org/RDF.

[7] See http://en.wiktionary.org/wiki/Wiktionary:Main_Page.

[8] "Lexvo.org brings information about languages, words, characters, and other human language-related entities to the Linked Data Web and Semantic Web." See http://www.lexvo.org.

[9] "lemon (LExicon Model for ONtologies) is an RDF model that allows to specify lexica for ontologies and allows to publish these lexica on the Web" (see McCrae etal., 2012)). This model was developed within the European R&D project "Monnet" (see www.monnet-project.eu)

Acknowledgements

The contribution of DFKI to the work reported in this paper has been partly supported by the R&D project "Monnet", which is co-funded by the European Union under Grant No. 248458.

References

Daránvi, S., Wittek, P., and Forró, L. (2012). Toward Sequencing "Narrative DNA": Tale Types, Motif Strings and Memetic Pathways. In: Proceedings of the Workshop on Computational Models of Narrative, Istanbul, May 2012.

Declerck, T. and Lendvai, P. (2011). Linguistic and Semantic Representation of the Thompson's Motif-Index of Folk-Literature. In: *Proceedings of the International Conference on Theory and Practice of Digital Libraries*, Berlin, Germany, Springer, 9/2011

Declerck, T. and Koleva, N. (2012). An Iterative Ontology-Based Text Processing Strategy for Detecting and Recognizing Characters in Folktales. In: *Proceedings of Digital Humanities -2012*.

Declerck, T., Mörth, K., and Lendvai, P. (2012). Accessing and standardizing Wiktionary Lexical Entries for supporting the Translation of Labels in Taxonomies for Digital Humanities. In: *Proceedings of LREC-2012*.

Lendvai, P., Declerck, T., Darányi, S., Gervás, P., Hervás, R., Malec, S., and Peinado, F. (2010). Integration of Linguistic Markup into Semantic Models of Folk Narratives: The Fairy Tale Use Case. In: *Proceedings of the Seventh International conference on Language Resources and Evaluation*, Pages 1996-2001, Valetta, Malta, European Language Resources Association (ELRA).

McCrae, J., Montiel-Ponsoda, E. and Cimiano, P. (2012) Integrating WordNet and Wiktionary with lemon. In: *Proceedings of LDL 2012*.

Mörth, K., Declerck, T., Lendvai, P., and Váradi, T. (2011). Accessing Multilingual Data on the Web for the Semantic Annotation of Cultural Heritage Texts. In: *Proceedings of the 2nd International Workshop on the Multilingual Semantic Web*, Bonn, Germany, Springer, 10/2011

Thompson, S. (1955). *Motif-index of folk-literature: A classification of narrative elements in folktales, ballads, myths, fables, medieval romances, exempla, fabliaux, jest-books, and local legends. Revised and enlarged edition*. Bloomington, Indiana University Press, 1955-58

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.

Uther, H.-J. (2004). *The Types of International Folktales: A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson*. FF Communications no. 284–286. Helsinki: Suomalainen Tiedeakatemia, 2004

Zöllner-Weber, A. 2008. *Noctua literaria : a computer-aided approach for the formal description of literary characters using an ontology*. Ph.D. Thesis, Bielefeld University.