Multilingual Ontologies for the Representation and Processing of Folktales

Thierry Declerck

DFKI GmbH,
Multilingual Technologies
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
declerck@dfki.de

Anastasija Aman, Martin Banzer, Dominik Macháček, Lisa Schäfer, Natalia Skachkova
Saarland University
Computational Linguistics Department
P.O. Box 15 11 50
D-66041 Saarbrücken, Germany

Abstract

We describe work done in the field of folkloristics and consisting in creating ontologies based on well-established studies proposed by "classical" folklorists. This work is supporting the availability of a huge amount of digital and structured knowledge on folktales to digital humanists. The ontological encoding of past and current motif-indexation and classification systems for folktales was in the first step limited to English language data. This led us to focus on making those newly generated formal knowledge sources available in a few more languages, like German, Russian and Bulgarian. We stress the importance of achieving this multilingual extension of our ontologies at a larger scale, in order for example to support the automated analysis and classification of such narratives in a large variety of languages, as those are getting more and more accessible on the Web.

1 Introduction

A final goal of our work is to make high quality content in the larger field of folkloristics available to a broad range of applications involving language technologies. The content, which is based on well-established and widely used classical motif-indexation, description and classification systems for folklore and related narratives, has been ported to standard W3C[1] formal representation languages, like OWL[2], RDF(s)[3] and RDF[4].

The result of this work is an integrated ontology containing about 60,000 classes and instances that are interlinked and searchable by means of semantic query engines. The integrated ontology, encoding distinct types of descriptives for folktales, offers a unique combination of generic classifications of narrative types and very fine-grained motifs (or patterns) occurring in tales.

This ontology can be optimally re-used in the context of narrative engines supporting the re-organization and adaptation of plots and characters in different and new contexts of narration to be displayed in various environments, like automated text generation or audio story telling.

The original folktale knowledge sources that are now available in this rich semantic environment are the Thompson's Motif-Index of Folk-Literature (Thompson, 1955-1958), the Aarne-Thompson-Uther Classification System of International Folktales (Uther, 2004), the so-called Proppian functions and also the typical characters of (Russian) folktales (Propp, 1968).

The work by Propp is very relevant, as his approach was aiming not only at characterizing the elements of a tale, but also and mainly at describing the organization of such elements in a tale for building consistent stories. The influence

[1] https://www.w3.org/
[2] https://www.w3.org/OWL/
[3] https://www.w3.org/TR/rdf-schema/
[4] https://www.w3.org/RDF/

Proceedings of the LT4DHCSEE in conjunction with RANLP 2017, pages 20–23,
Varna, Bulgaria, 8 September 2017.
http://doi.org/10.26615/978-954-452-046-5_003

of the Proppian approach on the production of cinematic narratives has been studied and discussed in many publications (see for example (Fell, 1997)).

The ontologization of the Proppian elements, in combination with the Motifs Index and the Tales Types offers a very closely woven set of fined-grained modules that can serve as an (extensible) catalogue of basic plot elements. Instances for the Aarne-Thompson-Uther Classification System of International Folktales have been collected from various Web resources, like the ``Multilingual Folk Tale Database[5] and the "Ashliman collection"[6], so that we have now more than 4,000 stories included as instances in the ontology.

2 Towards Multilingual Ontologies for Folktales

In a first step we created an ontology based on one story, the Russian tale "Гуси-лебеди" (*The Magic Swan Geese*), including a large family ontology, as families are playing an important role in many tales. The ontology is described in (Koleva et al., 2012). This was our first development of a multilingual ontology in the field of folktales. Languages covered were English, German, Russian and Bulgarian.

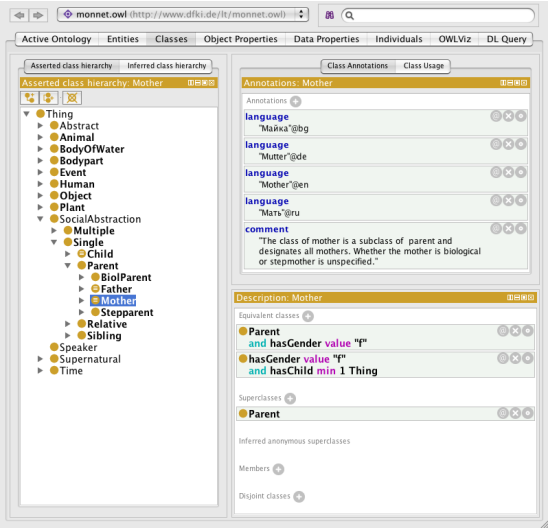

Figure 1: Screenshot from the Protégé tools, showing an element of the multilingual family class hierarchy in our first multilingual folktale ontology.

[5] http://www.mftd.org/index.php?action=atu

[6] http://www.pitt.edu/~dash/folktexts.html

In a second step, we proposed a multilingual ontologization of the Proppian system. A version is available now for English, German and Russian. This work was done together with 2 Russian native speakers, who could read the original version of the work by Propp, and could also extract the relevant terms from the German translation (published in 1972) and from the English version (Propp, 1968).

Table 1 gives an example on the way this multilingual information is encoded in our ontology, for the Proppian function "Interdiction".

```
:Gamma\~Interdiction
  rdf:type owl:Class ;
  rdf:type owl:NamedIndividual ;
  rdf:type
    SoPro:Instances-Propp-Function ;
  rdfs:comment "Dem Helden wird ein
    Verbot erteilt."@de ;
  rdfs:label "interdiction"@en ;
  rdfs:comment "An interdiction is
    addressed to the hero."@en ;
  rdfs:label "словозапрет"@ru ;
  rdfs:comment "К герою обращаются с
    запретом."@ru ;
  rdfs:label "Verbot"@de .
  SoPro:inHierarchyProppFunctionOf
    SoPro:Gamma1 ;
  SoPro:inHierarchyProppFunctionOf
    SoPro:Gamma2 ;
  rdfs:subClassOf
    SoPro:Propp-Function ;
.
```

Table 1: The Propp Function "Interdiction" in the ontology, including multilingual rdf(s) "label" and "comment" properties.

In Table 1, the reader can see how we encode the multilingual data using the RDF(s) annotation properties "label" and "comment". The texts encoded by those properties are taken directly from the Russian text and its official translations in English and German.

In a following step, we will provide for a translation of our ontology version of the Aarne-Thompson-Uther Classification System of International Folktales[7]. For adding German terms and descriptions of tale types we will use (Uther,

[7] See (Declerck et. al., 2017) for details on this ontology.

2015), a classification system for German folktale types covering to a great extent (Uther, 2004), but with German language data. Figure 2[8] is giving an idea of the type of hierarchical structure we are dealing with in this resource, which we abbreviate to "ATU".

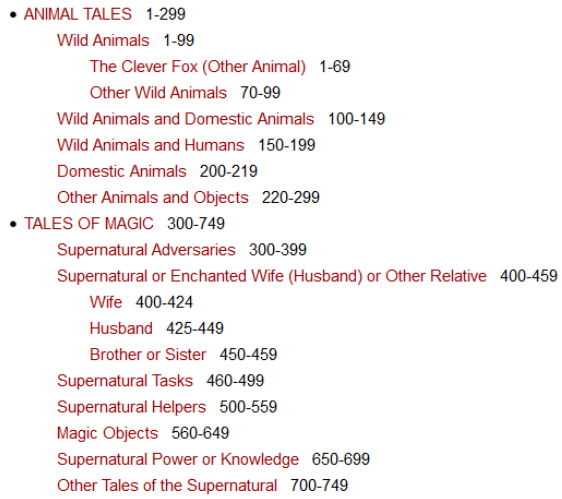

Figure 2: Example of the classification levels of ATU. ANIMAL TALES and TALES OF MAGIC are in the first level, Wild Animals, Wild Animals and Domestic Animals and Wild Animals and Humans are on the second level and The Clever Fox and Other Wild Animals are on the third level. The ranges behind each line are the so called ATU numbers which belong to each class.

A way to start to expand the language coverage of the ontological ATU resource lies for example in taking term equivalents that are available in corresponding Wikipedia articles, but this is limited by now to French, Estonian, Spanish, Hebrew, Chinese and Portuguese, which are the Wikipedia articles that contain a relevant number of ATU types in their language. There is a need thus to identify textual resources in other languages that are covering the ATU types.

The same strategy can be followed for obtaining the term equivalents for the Proppian functions and characters. We notice for example that the 31 function names are available in the Croatian Wikipedia article on Propp. Also the work (Propp, 1968) has been translated in the Bosnian-Croatian-Montenegrin-Serbian language (in 1982). And as (Propp, 1968) has been translated in quite some languages, it should be easy to extract from those translations the list of function and character terms.

3 Multilingual Folktale Resources

One goal of having different indexation, classification or other schemes for folktales standardized in one formal representation is to support NLP tasks applied to such types of texts and to allow their automated mark-up or classification along the lines of a well-established knowledge source.

There is one very interesting resource on the Web, called "Multilingual Folk Tale Database"[9], which is offering for some of the uploaded tale texts, in many languages, a corresponding type number from the Aarne-Thompson-Uther classification. The "Multilingual Folk Tale Database" is also displaying some folktales in a parallel fashion, in various languages, offering to a certain extent a comparable multilingual corpus of folktales.

This data can thus provide for a kind of gold standard for classification tasks, at least for tales in English, which are in the majority in the database. We wrote a crawler for accessing the content of the Multilingual Folktale Database and extracted folktales with available metadata. This gave us an access to information about the language, the ATU class number and its label.

4 Issues for a multilingual Classification Task

The Multilingual Folk Tale Database contains stories in 11 languages. We crawled all of them.

Figure 3 summarizes the number of stories by language. In total we have 901 stories in English and between roughly 500 and 600 stories in other 4 languages, French, Spanish, Hungarian and Russian[10]. We realized that we have too few stories in German, Danish, Polish, Italian, Czech and Dutch to implement a strong supervised classifier for them. We also realized that many of the stories in the Multilingual Folk Tale Database do not have an ATU number assigned.

[8] Image taken from http://www.mftd.org/index.php?action=atu

[9] See again: http://www.mftd.org/index.php?action=atu

[10] Those figures are valid for May 2017, when we crawled the database.

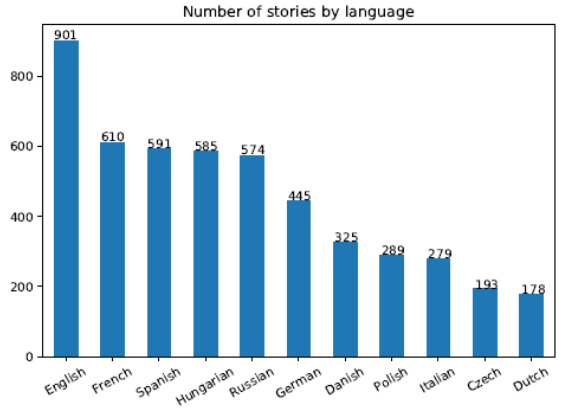

Figure 3: Number of all stories in the Multilingual Folk Tale Database by language (May 2017)

Figure 4 shows the numbers of stories with assigned ATU level 1 labels, for each language. The language with most labelled stories is English, with 342 labelled stories. The second is German, with 227 stories. For other languages we have too few resources. So there is a need to expand this kind of resources with tales from the corresponding languages.

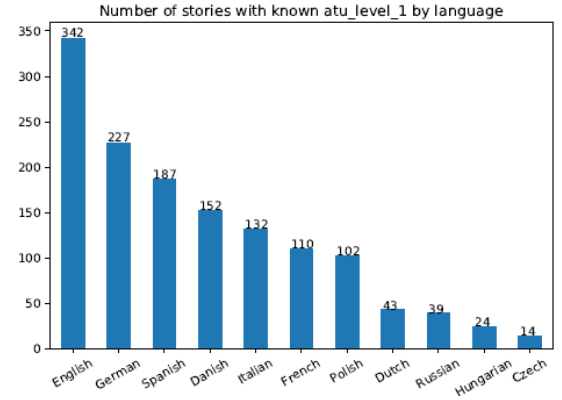

Figure 4: Number of stories in the Multilingual Folk Tale Database with an ATU type tag of level 1 by language (May 2017)

As the database of multilingual folktales is in continuous expansion, we are quite confident that in a limited amount of time a collection of tales in the languages of concerns for the LTDHCSEE workshop can be reached, whereas additional focus should be given on the correct assignment of ATU types to each tales.

5 Conclusions

In this short paper, we presented work done in generating multilingual ontologies that are encoding past and current "classical" (meaning non-digital-born) knowledge sources in the field of folklore, more specifically folktales. We described the current level of multilingualism in some of the ontologies we developed and proposed a way to extend this to more languages. Additionally we described a multilingual source of tales, which are partly labelled with the type numbers of the Aarne-Thompson-Uther classification scheme, and which can be used by now for an automated classification task for English folktales. We note that the current collection of labelled tales in other languages has to be considerably extended for allowing supervised training.

Acknowledgments

We thank the anonymous reviewers for their comments.

References

Thierry Declerck, Antónia Kostová and Lisa Schäfer. 2017. Towards a Linked Data Access to Folktales classified by Thompson's Motifs and Aarne-Thompson-Uther's Types. Proceedings of Digital Humanities 2017. Montréal.

John L. Fell. 1977. Vladimir Propp in Hollywood.Film Quarterly 30(3):19–28. https://doi.org/10.2307/1211770.

Nikolina Koleva, Thierry Declerck, and Hans-Ulrich Krieger. 2012. An ontology-based iterative text processing strategy for detecting and recognizing characters in folktales. In Jan Christoph Meister, editor, Digital Humanities 2012 Conference Abstracts, Hamburg University Press.

Vladimir Propp. 1968. Morphology of the folktale. Trans., Laurence Scott. 2nd ed., University of Texas Press.

Stith Thompson. 1955-1958. 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, Indiana University Press.

Hans-Jörg Uther. 2004. The Types of International Folktales: A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson.Suomalainen Tiedeakatemia.