Towards Linked Language Data for Digital Humanities

Thierry Declerck, Piroska Lendvai, Karlheinz Mörth, and Gerhard Budin and Tamás Váradi

Abstract We investigate the extension of classification schemes in the Humanities into semantic data repositories, the benefits of which could be the automation of so far manually conducted processes, such as detecting motifs in folktale texts. In parallel, we propose linguistic analysis of the textual labels used in these repositories. The resulting resource, which we propose to publish in the Linked Open Data (LOD) framework, will explicitly interlink domain knowledge and linguistically enriched language data, which can be used for knowledge-driven content analysis of literary works.

1 Introduction

We discuss strategies of porting semi-structured resources in the field of folk literature into the expanding linked open data (LOD) framework.[1] Prominent examples of such resources are the "Thompson Motif-Index of folk-literature" (Thompson,

Thierry Declerck
DFKI GmbH, Stuhlsatzenhausweg 3, D-66123 Germany & ICLTT, Sonnenfelsgasse 19/8, 1010 Wien, Austria. e-mail: declerck@dfki.de

Piroska Lendvai
HASRIL, 1068 Budapest, Benczúr u. 33, Hungary. e-mail: piroska@nytud.hu

Karlheinz Mörth
ICLTT, Sonnenfelsgasse 19/8, 1010 Wien, Austria. e-mail: Karlheinz.moerth@oeaw.ac.at

Gerhard Budin
ICLTT, Sonnenfelsgasse 19/8, 1010 Wien, Austria. e-mail: gerhard.budin@univie.ac.at

Tamás Váradi
HASRIL, 1068 Budapest VI., Benczúr u. 33, Hungary. e-mail: varadi@nytud.hu

[1] See http://linkeddata.org

1955-58, TMI), which is now also available, in English, on the Web,[2] as well as the Aarne-Thompson-Uther classification of folk tales (ATU, see Uther, 2004), of which excerpts in various languages are available in Wikipedia.[3] The longer term goal of our work is to obtain a LOD-compliant representation not only for the (hierarchy of) classes used in those resources, but also for the language data associated with the classes, and so to establish semantic links between the language data and the classes, across languages and classification systems. A pre-requisite for achieving this stage is the linguistic processing of the content of the labels, and the representation of this analysis in terms of standards compliant with the LOD.

The publication of the resulting domain knowledge and enriched language resources in the LOD can support the automated analysis of literary works by advanced knowledge-driven Natural Language Processing, putting at the disposal of scholars a large set of linguistically and semantically annotated multilingual text segments. Not only researchers in the Humanities will benefit from such a transfer of Humanities resources onto the LOD, but also the general public can be offered extended search possibilities. Folk-literature in general is a very popular topic: as the online presence of the French National Library states, the fables of Jean de la Fontaine are the most consulted literary works in their catalogues;[4] the maintainers of the online Dutch database of folk tales at the Meertens Institute[5] likewise report a high number of visitors.

Such institutions are highly interested in getting machine readable and processable versions of the type of classification systems we mentioned above in order to semi-automatically improve their own indexing, across versions of literary works in different languages and cultures. This would for example allow to detect the differences in the various "national" versions of classical tales over time, and how those are leading to different types of interpretation of the story.

The goal of our work is to provide a resource type that is potentially complementary to the ones already published by national libraries in the LOD.[6] By these, mainly the bibliographical metadata, including the structured part of classification systems, have been ported to the LOD, and we are interested in making all the linguistically and semantically analysed language data included in those classification

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

[3] See http://en.wikipedia.org/wiki/Aarne-Thompson_classification_system for the English version, http://de.wikipedia.org/wiki/Aarne-Thompson-Index for the German one, and for the French version http://fr.wikipedia.org/wiki/Classification_Aarne-Thompson We note that the online ATU data do not reflect the original catalogue, as this was the case for TMI. Therefore it would be highly appreciated to dispose over an electronic version of ATU.

[4] See http://data.bnf.fr/

[5] See http://www.verhalenbank.nl/

[6] See for example the press release of the Hungarian National Library: http://lists.w3.org/Archives/Public/public-lod/2010Apr/0155.html, or for the German National Library (DNB): http://www.eco4r.org/workshop2010/eco4r_workshop2010_mirjam_kessler.pdf

systems LOD compliant.[7] At the end we will generate a kind of matrix of classification items and the natural language expressions that are typically associated to the classes. This will improve the automatic classification of literary works, on the basis of the automatically combined linguistic and semantic analysis of the texts, detecting variants of the textual expressions used in the labels of the classes.

In doing so we hope to contribute to the "Global Cultural Graph" that is emerging in the LOD.

2 Language Data in the Linked Open Data Framework

Our vision of Linked Language Data (LLD) is to aim at making a better organisation and use of the language data that is (to be) incorporated in the linked open data (LOD) framework. While the LOD initiative was primarily conceived as a way of interconnecting structured knowledge resources in a standardized way using specific links between dereferenceable URIs, the wealth of language data, existing as the content of labels or comments, intuitively associated with those knowledge resources has not been exploited till now, although the LOD framework offers a configuration that can be used and extended for knowledge-driven organization of language data in a set of NLP applications. Additionally, the corresponding linguistic information (e.g. lemma, part-of-speech, morphology, constituency, argument structure, etc.) for those language data is missing, which poses a serious obstacle to direct re-usability in NLP applications.

Adding linguistic information to the language data in LOD will lead to the generation of a very large and dynamically increasing set of enriched language data being structured with both (domain) semantics and linguistic information. This integrated resource can positively impact a range of applications that build on combinations of world and linguistic knowledge, e.g. in Cross-Lingual Information Extraction and Summarization, Localization, Machine Translation, etc.: The knowledge represented in the LOD data sets will become associated with the language data according to linguistic parameters, and thus ready to be directly included in Natural Language Processing modules.

The focus of LLD is thus on bridging the gap between raw language data and knowledge descriptions, by solving representation issues and inclusion of multilingual knowledge-based terminological resources in the LOD. We started to experiment on those ideas with resources in the Humanities, like *Thompson's Motif Index* (TMI) mentioned above, that are not yet in the LOD, but which will be directly published in this framework, including the combination of linguistic and semantic information, as a result of our work.

[7] We note for example that the transformation of the index-terms of the DNB onto SKOS is not proposing any additional linguistic categorisation of the terms. See `http://www.kim-forum.org/material/pdf/BPG_Repraesentation_von_KOS_im_Semantic_Web.pdf`

3 Two Classification Systems for Folk-Literature: TMI and ATU

As mentioned in the introduction section, we are for now dealing with two existing extended catalogues that hold conceptual schemes for classifying narrative elements in folktales, ballads, myths, and related genres: the *Thompson's Motif Index* (TMI, see Thompson, 1955-58) and *The Types of International Folktales* (ATU, see Uther, 2004). ATU and TMI come along with extensive terminologies, and our idea is that these can be linguistically processed, enriched, linked, and represented for language technology mechanisms to identify the semantic classes that can be associated with a text. Pursuing this line of research, we soon discovered that digital catalogues require important terminological and semantic pre-processing in order to be successfully used as a knowledge base to be matched with textual data. Furthermore, catalogues have to be made interoperable with each other, and we need to map/transform them into a semantically harmonized representation, using standards such as XML-TEI or RDF.

We focus therefore on porting TMI and ATU into adequate semantic resources, under consideration of linguistic and terminological aspects. A strong wish in the Digital Humanities community lies in linking ATU and TMI, and we think that this can be carried out only by providing for multilingual and semantic extensions of those resources.

3.1 Work on the Thompson Motif-Index

The Thompson Motif-Index is an hierarchical structure of motifs descriptions consisting in a label associated with an alphanumeric class index, so for example: "A21.1 :: Woman who fell from the sky". Higher in the hierarchy are:"A0 :: Creator", "A20 :: Origin of the Creator" and "A21 :: Creator from above".

We propose both a lexical and a syntactic analysis of all those labels, and for our example A21.1. we get[8]:

```
woman,N+Nb=s+Distribution=Hum
who,PRO+Distribution=RelQ
fell,fall,V+Tense=PT+Pers=3+Nb=s
from,PREP
the,DET
sky,N+Nb=s
```

Lexical Analysis of the label: Woman who fell from the sky.

```
<NP>
   <NP>
      <HEAD><REFOF XREF="396.2">Woman</HEAD>
   </NP>
```

[8] This analysis is the result of a specialized grammar we wrote using the NooJ platform, see www.nooj4nlp.net/.

```
   <SENT>
      <RELCLAUSE>
         <SUBJ><XREF>who</XREF></SUBJ>
         <PRED>fell</PRED>
         <PP><PPOBJ>from
            <NP>
               <SPEC>the</SPEC>
               <HEAD>sky</HEAD>
            </NP>
         </PPOBJ></PP>
      </RELCLAUSE>
   </SENT>
</NP>
```

Analysis of the label: Woman who fell from the sky.

In our analysis, we also make use of the Conceptual, Terminological and Linguistic Objects in Ontologies (CTL) approach, described in Declerck and Lendvai (2010), and which consists in keeping track of the relations between annotated labels of knowledge systems and the classes or properties they are associated with. For the the linguistic level, we provide syntactic information on both the constituency (phrasal grouping of word forms) and the dependency relations between (groups of) word forms. For example, *woman* and *the sky* are marked as a constituent of type NP. At the dependency level, *woman* is marked as the "Subj(ect)" of the "Pred(icate)" *fell*. The detection of this Subj-Pred relation is possible only on the basis of an earlier round of computing the co-reference relation between *woman* and *who*, which we marked with the XML element "XREF". We also mark the dependency relation between the word forms within a phrasal constituent. This allow us to create a "lexicalized ontology" stating for example that a *woman* can be a *creator*, that the *sky* can be related to the *above*, on the base of the lexical and syntactic realisations of concepts included in the hierarchy of TMI.

But we note that the NooJ encoding of the CTL mechanisms is lacking interoperability with other resources. Therefore actual work is dedicated in linking the NooJ annotation to the ISO data categories,[9] and to use the *lemon* model developed in the Monnet project.[10] for representing the different kinds of data concerned – conceptual, terminological and lexical , and with this to test if we can publish the result of our work directly as a LOD data set. Further we are aiming at linking results of the CTL-*lemon* driven semantic annotation of folktale text to actual LOD data sets published by libraries, relating thus semantic annotation of literary text to bibliographical metadata.

[9] See `http://www.isocat.org/`, also Windhouwer and Wright (this vol.).

[10] See `http://www.monnet-project.eu/lemon`, also McCrae et al (this vol.).

3.2 Towards a Multilingual and Semantic Extension of TMI

Analysing the on-line version of TMI, we discovered very soon that it would be beneficial for further automatic processing to turn the basic classification of TMI into a real taxonomy. The actual alphanumeric organization of TMI, which simulates the class hierarchy of motifs does not allow to properly express the hierarchy and inheritance properties of motifs. Furthermore, it is not made explicit in TMI which elements introduce pure classification information (“A0-A99. Creator”, “A20. Origin of the creator”), which ones are abstractions over concrete motifs (“A21. /Creator from above./”), and which ones are in fact the actual motifs, i.e. their possible concrete realisation in text (“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 therefore wrote a script that transforms the digital version of TMI onto an XML representation that marks this kind of information explicitly by using designated tags:

```
<CLASS ID="0" SPAN="0-99" LABEL="Creator">
<CLASS ID="20" LABEL="Origin of the creator" SubClassOf="0">
<CLASS ID="21" MOTIF="Creator from above" TYPE="Abstract"
       SubClassOf ="0">
<CLASS ID="21.1" MOTIF="Woman who fell from the sky"
       TYPE="Ref" PartOf="21">
```

As a next step, ongoing work is dedicated to upgrading the XML representation to RDF and OWL, so that we have the adequate means for differentiating between hierarchical realisations and real properties associated with classes, and the possibility to compute the transitive closure of the subclass hierarchy. In parallel, we targeted the extension of motifs listed in TMI in English into a multilingual version. This is carried out by accessing the multilingual Wiktionary[11] lexicon, and suggesting multilingual equivalents to the motifs formulated in in English. Using the lexvo[12] service (available in the LOD framework), one is getting access to Wiktionary and other sources. Lexvo suggests for example for the English word ”creator” more than 20 translations. Note that this approach is limited to word-based translations. Our previous study in Mörth et al (2011) analyses some shortcomings of the use of Wiktionary in its actual state, and proposes a conversion of the Wiktionary lexicon into a TEI compliant format[13]

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

[12] See http://www.lexvo.org

[13] See http://corpus3.aac.ac.at/showcase/index.php/wiktionary001 for a demo of the transformation of the German wiktionary.

3.3 Towards a Multilingual Combination of ATU and TMI

The complete TMI is available on the web. It is available only in English, as we already mentioned. ATU presents a different situation: only segments of ATU are available on-line in Wikipedia, but in different languages. Looking at the English, French and German Wikipedia pages some discrepancies in the presentation become evident, as illustrated below:

```
(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 (EN) version links the German tale *Rapunzel* to four tale versions, in different languages. The original German tale is reached from the English Wikipedia page if the reader clicks on the name “Rapunzel”. The German (DE) version links additionally to a German classification system (KHM = Kinder- und Hausmärchen – Children’s and Household Tales–, used for the classification of Grimm’s Fairy Tales). Interesting enough is the fact that only the French (FR) Wikipedia page introduces the English name for the tale. None of the three Wikipedia pages is making use of the same abbreviation of the ATU index (”‘310”’, ”‘AaTh”’, or ”‘AT310”’). Therefore we suggest a restructuration of the information availble in the three pages, merging it in one XML format:

```
<ATU ID="310">
  <LABEL lang="EN">Rapunzel</LABEL>
  <LABEL lang="DE">Jungfrau im Turm</LABEL>
  <LABEL lang="FR">La Fille dans la tour</LABEL>
  <ALT lang="DE">Rapunzel</ALT>
  <ALT lang="EN">The Maiden in the Tower</ALT>
</ATU>
```

This representation will also be ported to RDF, using the SKOS standard for encoding the preferred and alternative forms. On the basis of a small fragment of correctly aligned Wikipedia pages, a representative multilingual terminology of ATU terms can be aggregated, and this terminology can be re-used for supporting the translation of motifs used in TMI, overcoming the shortcomings of the word-based translation approach we discuss in Sect. 3.2. We started to utilize terminology alignment techniques used in Machine Translation (see for example Federmann et al, 2011), adapting them to the short terms that are employed in the catalogues we are focusing on.

As mentioned already, a promising approach of this project is the design of the RDF-based *lemon* representation model for lexicon entries used in ontologies. We are starting to investigate in which way the Monnet project and its *lemon* model can help in translating the English labels of TMI/ATU on the one side and to publish the combination of linguistic information included in the labels of TMI/ATU and their corresponding classes in a LOD compliant way, see also McCrae et al (this vol.).

4 Conclusion

We have presented actual work dealing with classification systems in the field of eHumanities, with the goal of "upgrading" those resources into interoperable multilingual systems, taking also into consideration linguistic and terminological issues. Next step of our work will be in proposing a LOD format for the resulting combined linguistically and semantically analysed classification data.

Acknowledgements The work described in this paper is done in the context of a collaboration between the Academy of Sciences of Budapest and Vienna. The contribution by Thierry Declerck is partly supported by the R&D project "Monnet", which is co-funded by the European Union under Grant No. 248458.

References

Declerck T, Lendvai P (2010) Towards a standardized linguistic annotation of the textual content of labels in knowledge representation systems. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10), Valetta, Malta

Federmann C, Hunsicker S, Wolf P, Bernardi U (2011) From statistical term extraction to hybrid machine translation. In: Proceedings of the 15th Annual Conference of the European Association for Machine Translation

McCrae J, Montiel-Ponsoda E, Cimiano P (this vol.) Integrating WordNet and Wiktionary with *lemon*. P. 25-34

Mörth K, Declerck T, Lendvai P, Várdi 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

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

Uther HJ (2004) The Types of International Folktales: A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson. Suomalainen Tiedeakatemia, Helsinki

Windhouwer M, Wright SE (this vol.) Linking to linguistic data categories in ISOcat. P. 99-107