Accessing Multilingual Data on the Web for the Semantic Annotation of Cultural Heritage Texts

Karlheinz Moerth[1,], Thierry Declerck[1,2], Piroska Lendvai[3], Tamás Váradi[3]

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

[2] DFKI GmbH, Stuhlsatzenhausweg, 3
66123 Saarbrücken, Germany

[3] HASRIL, Hungarian Academy of Sciences, Benczúr u. 33.
H-1068 Budapest, Hungary
Karlheinz.moerth@oeaw.ac.at,
declerck@dfki.de,
piroska@nytud.hu,varadi@nytud.hu

Abstract. Our study targets interoperable semantic annotation of Cultural Heritage or eHumanities texts in German and Hungarian. A semantic resource we focus on is the Thompson Motif-index of folk-literature (TMI), the labels of which are available only in English. We investigate the use lexical data on the Web in German and Hungarian for supporting semi-automatic translation of TMI: lexical resources offered by Wiktionary accessed via the Lexvo service, and discuss shortcomings of those resources. An approach for mapping the XML dump of Wiktionary onto a TEI and MAF compliant data is presented, whereby we discuss improvements in the representation of Wiktionary data for exploiting its multilingual value within the LOD framework.

Keywords: Multilinguality, LOD, Cultural Heritage, Semantic Annotation

1 Introduction

In the context of a cooperation between the Austrian and the Hungarian Academies of Sciences we investigate the possibility to generate interoperable and multilingual semantic annotation of Cultural Heritage or eHumanities texts. One of the semantic resources we consider for this task is the Thompson Motif-index of folk-literature (TMI)[1] [5], which contains around 36,000 terms, cataloguing typical narrative content of folk tales and myths from around the world. The terms, or 'labels', of the classification system are available only in English.

Our general hypothesis is that converting resources such as TMI into a LOD compliant combination of multi-layered linguistic annotation and their taxonomic classes can support the automatic detection and semantic annotation of motifs in literary work, across genres and languages.

[1] An electronic version of TMI is available at: http://www.ruthenia.ru/folklore/thompson/

A motif is an element conveying an idea or theme e.g. in film or music, but also in folklore or scientific texts[2]. Motifs are cognitively complex notions expressed in lexically and syntactically highly variable but compact structures. Linguistic features of motifs have so far not been systematically investigated, but these have been exposed and aim to be worked out by the authors of this paper, in collaboration with the international AMICUS network[3], with a clear motivation for enhanced indexing and modelling of cultural heritage data (cf. [1], [3] and [4]).

The TMI catalog focuses on motifs that emphasize ideas or themes. For example, "*K3. Substitute in contest*" is one motif in TMI (its parent node being "*K0-K99. Contests won by deception*", subsumed under *"K. Deceptions*"). Dozens of subtypes are assigned to this single motif; these catalogue descriptions, or labels, are short phrases such as "*Supernatural substitute in tournament for pious warrior*", "*Wise man disguised as monk beats learned heretic in debate*". The TMI lists 23 main categories[4] and provides a deep hierarchical structure of motifs.

To semantically annotate texts in German and Hungarian with this resource, we aim to enrich TMI with German and Hungarian labels. Our strategy consists in providing first for the linguistic annotation of the phrasal heads detected in the English labels[5], and to try to find equivalent lexical entries in German and Hungarian retrieved from online multilingual lexical resources.

2 Access to Online Lexical Resources in the LOD

The scarcity of freely available professional on-line multilingual lexical data made us turn to the lexical resources offered by the collaborative dictionary project Wiktionary, and the access provided to within the Lexvo service[6], which has been deployed within the Linked (Open) Data (LOD) framework[7]. Some observations we could make on this combination of resources are described in this section.

We noted first that in Wiktionary, variants of an entry (e.g. singular or plural form), often do not feature identical sense or translation information.[8] It is necessary to link those entries into a consistent unit, and to use an appropriate model for this. Two candidates can be considered for this modeling: ISO-LMF[9] and *lemon*

[2] Some random examples for motifs in folk tales are e.g. the cruel stepmother, the poor girl who was chosen as wife in preference to a rich one, or a supernatural who substitutes the hero in a tournament.

[3] http://amicus.uvt.nl

[4] E.g. Animal Motifs, Magic, the Dead, Marvels, Tests, the Wise and the Foolish, Deceptions, Reversals of Fortune

[5] The details of this linguistic analysis are described in a submission currently under review.

[6] http://www.lexvo.org/

[7] http://linkeddata.org/

[8] One example is the English Wiktionary entry "creator" (http://en.wiktionary.org/wiki/creator), which lists the basic morpho-syntactic information, associated senses and translations whereas the entry "creators" (http://en.wiktionary.org/wiki/creators#English) only states that it is the plural of "creator".

[9] http://en.wikipedia.org/wiki/Lexical_Markup_Framework

(developed in the Monnet project and related to the W3C community)[10]. An advantage of the *lemon* approach would be that one could represent the Wiktionary data in the RDF format, making Wiktionary data available in the Linked Data framework. Nevertheless, as a first step we ported the XML dump of Wiktionary into a TEI[11] and MAF[12] compliant format (see Section 3).

Lexvo is a service that "brings information about languages, words, characters, and other human language-related entities to the Linked Data Web and Semantic Web"[13]. Lexvo points to Wiktionary entries, displaying for each word that can be queried (in a variety of languages) a link to senses that are encoded either in the LOD version of WordNet[14] or/and of OpenCyc[15], but in those versions the senses are available only for English entries. Since the the Wiktionary data is not yet available in a machine-readable format, Lexvo cannot display the senses available in the resource. This is an additional argument for porting Wiktionary to RDF. Due to the same reason, linguistic information associated to each word in WIktionary cannot be made available in Lexvo. A Lexvo specific shortcoming is the fact that it refers only to the English version of Wiktionary, regardless of entries that are in fact written in other languages, ignoring as a consequence several pieces of language-specific information.

3 Porting Wiktionary to a Standardised Representation

Our starting point is the XML dump[16] of Wiktionary. Nevertheless, the data do not really deliver what one might expect from xml data, namely well-formed structured information. The content is formatted making use of a lightweight markup system which is used in different Wiki applications, and is neither standardized (various applications use considerably divergent forms of the wikitext language) nor truly structure-oriented. It is designed in a format-oriented manner to be transformed into HTML.

Our initial goal was to transfer these data into an XML format suitable for further processing. Although, as mentioned above, we consider ISO-LMF and *lemon* as the final candidates, for pragmatic reasons, we eventually opted for TEI p5[17] as our

[10] http://greententacle.techfak.uni-bielefeld.de/drupal/sites/default/files/lemon-cookbook.pdf and [8].

[11] http://www.tei-c.org/index.xml

[12] http://lirics.loria.fr/doc_pub/maf.pdf

[13] http://www.lexvo.org

[14] http://semanticweb.cs.vu.nl/lod/wn30

[15] http://sw.opencyc.org

[16] http://dumps.wikimedia.org

[17] As the TEI p5 dictionary module was conceptualized as the digital representation of printed dictionaries, it appears not to be the most natural candidate for the task at hand. However, the main motive behind adopting the dictionary module of this “de facto” text encoding standard was that ongoing lexicographic projects of the ICLTT had yielded tools to process this kind of data. Besides an online dictionary editor geared towards the particular needs of TEI, there are also a number of thoroughly tested XSLT stylesheets to visualize the particular kind of data. A second reason, equally important, is the fact that the ICLTT’s dictionary working group has been working recently on a TEI dictionary schema suitable for use in NLP applications.

starting point·. While several attempts at preparing Wiktionary for use in NLP applications have been made before [2, 5, 7], the tool we present here is – to our knowledge – the first such application targeting TEI p5, and the first such tool provided with a graphical user interface.

The actual conversion process is carried out in three main steps. Each of these steps can be performed separately, which allows the interested user to pursue the transformation process in detail.

First, the comparatively large database dump (287 MB) was split into manageable smaller chunks. This process resulted in a collection of roughly 85000 entries.

In the second phase of the conversion, the top-level constituents of these entries were identified and transformed into XML elements. This task turned out to be pretty straightforward as the *entries* (we stick to traditional lexicographic nomenclature here) display a rather flat hierarchical structure. The resulting chunks each contain a particular type of data, the main constituents of the dictionary entries. The number of constituent parts varies with the size of the individual entries (from 3KB up to 338KB). In the result sets, there are chunks containing grammatical data such as for instance part of speech. There are chunks containing etymological information and/or usage information. Many entries contain morphological data, in numerous cases complete inflectional paradigms. The files also hold data concerning hyphenations of word forms and their pronunciation. However, the central concern of our work here has been semantic data. This kind of information is stored in sections describing the various meanings of words. These, in turn, are linked to translations, synonyms, antonyms, hyperonyms, hyponyms, and often to examples.

The last step in the transformation process has been the conversion of the above described constituents into TEI p5. Iterating through all the untyped chunks, the program attempts to identify the right category and subsequently to translate it into TEI p5. At this point, the main challenge for the programmer was the merging of data on the same hierarchical level (e.g. meanings and translations) into neatly nested XML structures. Successful data conversion depends largely on the quality of the underlying markup. While many errors can be compensated by some trickery in the program, inconsistencies remain.

The actual tag set applied in our project can also be seen as a contribution aiming at developing the TEI guidelines towards an encoding system suitable to be used in NLP applications.[18] We will not go into the gory details of modeling TEI documents here, just one small digression: one particularly useful module of the TEI p5 guidelines was the chapter on *feature structures*. This mechanism allowed us to model the representation of the morpho-syntactic data in accordance with the MAF standard (Morpho-syntactic Annotation Framework, ISO TC 37). Canonical TEI for inflected word forms such as *gingst* "(you) went" usually look like this:

[18] An initiative towards this end was the workshop *Tightening the representation of lexical data, a TEI perspective* at the TEI's members meeting this year in Würzburg (Germany).

```
<form>
    <orth>gingst</orth>
    <gramGrp>
      <gram type="pos">verb</gram>
      <gram type="number">plural</gram>
      <gram type="person">2</gram>
      <gram type="tense">preterite</gram>
      <gram type="mood">indicative</gram>
    </gramGrp>
</form>
```

We tried to encode such structures in a more MAF-like manner, which is still TEI conformant:

```
<form ana="#v_pret_ind_pl_p2"><orth>gingst</orth></form>
```

In this encoding scheme, the morpho-syntactic identifiers used in the *ana* atribute of the form element is defined as a set of TEI conformant feature structures. The values used here refer to a feature value library, which is also linked to the ISO data categories.

Although the conversion tool already works quite nicely, a number of issues registered in its requirement specification remain to be solved. It goes without saying that the first thing that comes to mind, is the issue of other languages, which is on top of our agenda. First candidates for this are English and French.

The second issue is moving on to LMF which is a project reaching far beyond our Wiktionary tool. Creating LMF data from TEI is something apparently non-trivial.

One other important task to be achieved in the near future is setting up a service delivering the data. First steps towards implementing a restful server have been taken. We hope that by the time this paper is presented, our TEI version of the German-language Wiktionary will be up and running.

4 Further work on porting Wiktionary to the Semantic Web

Although our work represents a step in making the full Wiktionary information available for NLP applications, it is not sufficient to represent links between entries (for example, one entry being the plural of the other, etc), or to make this information available in the Web or in the LOD and so to establish links between entries and senses in Wiktionary, WordNet or OpenCyc, on the one, but also between TMI and LOD data sets on the other hand. Just to name an example: In TMI the concept "A0: Creator" is the upper class of a large number of (hierarchically ordered) terms. We collected all the head nouns of those terms, and can build so a kind of domain specific "WordNet". This list of nouns is for sure very different and more complex than what we find in WordNet or OpenCyc. We need a way to relate the semantic organization of TMI and WordNet/OpenCyc (or other data sets), also on the base of linguistic information we can find in the (Semantic) Web. There is therefore a need to port both Wiktionary and the analyzed labels of TMI onto a LOD compliant RDF. For this we are getting also advices from the Monnet project[19].

Acknowledgments

Part of the described in this paper is supported by the R&D project "Monnet", which is co-funded by the European Union under Grant No. 248458, and by the AMICUS network, which is sponsored by a grant from the Netherlands Organization for Scientific Research, NWO Humanities, as part of the Internationalization in the Humanities programme

References

1. Declerck, T., K. Eckart, Lendvai, P., L. Romary, T. Zastrow (2010a). Towards a Standardised Linguistic Annotation of Fairy Tales. In: Proc. of the LRT standards workshop at LREC-2010.
2. Krizhanovsky, A. (2010). The comparison of Wiktionary thesauri transformed into the machine-readable format. (http://arxiv.org/abs/1006.5040)
3. Lendvai, P., Declerck, T., S. Darányi, P. Gervás, R. Hervás, S. Malec, F. Peinado (2010a). 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).
4. Lendvai, P. (2010). Granularity Perspectives on Modeling Humanities Concepts. In: S. Darányi, P. Lendvai, (eds.). First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts, Vienna, Austria. University of Szeged, Hungary.
5. Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S.-K., Kuo, T.-Y., Magistry, P., Huang, C.-R. (2009). Wiktionary and NLP: Improving synonymy networks. In: Proceedings of the 2009 Workshop on Peoples's Web Meets NLP, ACL-IJCNLP. Singapore: pp. 19-27.
6. 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. Revised and enlarged edition. Bloomington, Indiana University Press.
7. Zesch T., Mueller C., Gurevych I. (2008). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In: Proceedings of the Conference on Language Resources and Evaluation. LREC 2008.
8. McCrae, J, Aguado-de-Cea G, Buitelaar P, Cimiano P, Declerck T, Gomez-Perez A, Gracia J, Hollink L, Montiel-Ponsoda E, Spohr D et al.. In Press. Interchanging lexical resources on the Semantic Web. Language Resources and Evaluation 2011.

[19] http://www.monnet-project.eu