Accessing and Standardizing Wiktionary Lexical Entries for Supporting the Translation of Labels in Taxonomies for Digital Humanities

Thierry Declerck[1, 2], Karlheinz Mörth[2], Piroska Lendvai[3]
[1]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)
Stuhlsatzenhausweg, 3, D-66123 Saarbrücken, Germany
[2]ICLTT, Austrian Academy of Sciences
Sonnenfelsgasse 19/8, A-1010 Wien, Austria
[3]Research Institute for Linguistics, Hungarian Academy of Sciences
Benczúr u. 33. H-1068 Budapest, Hungary
E-mail: declerck@dfki.de, karlheinz.moerth@oeaw.ac.at, lendvai.piroska@nytud.mta.hu

Abstract

We describe the usefulness of Wiktionary, the freely available web-based lexical resource, in providing multilingual extensions to catalogues that serve content-based indexing of folktales and related narratives. We develop conversion tools between Wiktionary and TEI, using ISO standards (LMF, MAF), to make such resources available to both the Digital Humanities community and the Language Resources community. The converted data can be queried via a web interface, while the tools of the workflow are to be released with an open source license. We report on the actual state and functionality of our tools and analyse some shortcomings of Wiktionary, as well as potential domains of application.

Keywords: multilingual lexical resources, standards, digital humanities

1. Introduction

In (Moerth et al., 2011) we explained the scenario that motivated us to develop the suite of tools this submission describes: some of the existing canonical folk literature[1] catalogues only exist in English versions, while extending them to other languages would increase their value considerably. We exploit web-based multilingual resources for supporting the translation of the taxonomically organized labels in such resources; currently, we investigate German and Hungarian. The potential use of Wiktionary dictionaries[2] have been described in detail in (Zesch et al. 2008a) for several NLP applications. Wiktionary is the lexicographic counterpart of the encyclopaedic project Wikipedia. It is currently[3] available in 158 languages[4], although only a small number of these versions are sufficiently large to be useful.

As the result of collaborative work pursued by enthusiastic volunteers, Wiktionary dictionaries are not edited by professional lexicographers and may thus lack basic information or be of poor quality otherwise. Another drawback of the Wiktionary project is that the contents in its database are formatted in a lightweight mark-up system commonly used in Wiki applications. This system is neither standardised nor sufficiently structure-oriented. On top of it, the Wiki format is often applied in an inconsistent manner within one dictionary or across different language versions of Wiktionary, which makes the extraction of structured lexical information a challenging task.

A Wiktionary in a certain language typically contains not only entries in that specific language, but entries in other languages too (see here again Zesch et al., 2008a)[5], for which seemingly no editorial coordination exists. E.g. the entry for the German word "Schöpfer" (*creator*) in the English Wiktionary is very restricted compared to the entry for the same word in the German Wiktionary[6]. Moreover, the different entries for the same word are not explicitly linked.

Nonetheless, the steadily growing language resources in Wiktionary are being used in various experiments both by computational linguists and lexicographers to pursue monolingual as well as multilingual studies (Krizhanovsky & Lin, 2009; Meyer & Gurevych, 2010); many of the larger versions of Wiktionary turn out to provide valuable lexical information. In sum, to develop programs that transform the Wiktionary formats into a more structured representation is a worthy enterprise, which can in turn be offered to the Wiktionary project, resulting in improvement of the templates that are used for the submission of lexical entries.

[1] Our main taxonomic resource in this experiment is Stith Thompson's Motif-index of folk-literature, as available online at http://www.ruthenia.ru/folklore/thompson/index.htm

[2] See http://www.wiktionary.org/

[3] February 2012

[4] A language-specific Wiktionary is a dictionary in which all the descriptions and explanations are given in one language. Nonetheless, each dictionary may contain entries belonging to other languages.

[5] So for example the English Wiktionary edition contains entries for more then 400 languages, so that out of this source more language-specific Wiktionaries could be created as there are actually officially listed.

[6] The reader can just compare the English entry at http://en.wiktionary.org/wiki/Sch%C3%B6pfer with the German one at http://de.wiktionary.org/wiki/Sch%C3%B6pfer

In the current study, we focus on describing the development of tools that support the conversion of Wiktionary onto a standardised representation, i.e. format of the Text Encoding Initiative (TEI)[7]. The choice in favour of this format is partly due to TEI being widely used in the Digital Humanities community, and partly to the strong cooperation established between TEI and ISO committees in the field of language resources, for example on the definition and use of feature structures[8]. This enables us to include in our representation morpho-syntactic information available in Wiktionary using in part the MAF format[9].

We also briefly present ongoing work on possible solutions to some problematic issues by (a) addressing the topic of semantic interoperability in the treatment of entries for a specific language across different Wiktionaries, and (b) building a machine-readable multi- and cross-lingual semantic network, based on Wiktionary categories.

In the following sections we describe the different tools developed and used for the conversion from the English and the German Wiktionary onto TEI/MAF, and the role this conversion can have for eLexicography or translation.

As an additional comment, we consider our use of Wiktionary as potentially complementary to the use of other lexical resources, like the well-known WordNet resources, e.g. for example GermaNet and EuroWordNet. Henrich et al (2011) propose a comparative study of GermaNet and Wiktionary, and show how the Wiktionary data can enrich ca 30% of the GermaNet senses with definitions. John McCrae et al. (2012) show a similar percentage (ca 25%) of overlapping between lexical data encoded in Wiktionary and WordNet, stressing thus the potential complementarities of these resources in NLP applications.

We concentrate in this paper on the Wiktionary data, since these include not only lexical semantic information, but we will extend our work to the type of approach suggested by Henrich et al. (2011) and consider the use of augmented GermaNet senses in our work.

2. Tools for processing Wiktionary Data

Attempts at enabling Wiktionary for use in NLP applications have been made before (see the reference section in this paper). What makes our tool different is that it is geared towards the needs of users beyond the narrow circle of corpus linguists, addressing also the needs of the Digital Humanities community, using for example for the representation of lexical and linguistic information the TEI format, which is widely accepted in the Digital Humanities The freely available converter tool is therefore equipped with a graphical user interface which allows users to follow the conversion steps.

Four pieces of software have been developed by us so far to make the data provided by the Wiktionary project machine-readable and accessible on the Internet. The first tool, *Wiktionary converter*, is used to translate the non-standardized wiki-format into a reusable XML format. The current target format is TEI (P5). The second tool is the *Viennese Lexicographic Editor*, which can be used to manually adjust the output of the converter tool. The third tool is a web-interface and the server scripts implementing a service offering access to the TEI encoded resources[10]. The fourth tool is still in a prototypical state: it is supposed to semantically organize lexical entries across languages on the basis of the Wiktionary "categories"[11]. This effort is particularly relevant for providing lexical candidates for the translation of labels[12] in folktale taxonomies, which is the ultimate goal of our work.

2.1 Wiktionary converter

The starting point for the conversion experiments are the XML dumps[13] of the English, German and Hungarian language versions of Wiktionary. Having a closer look at the data, it becomes apparent that XML unfortunately does not deliver what it promises, namely structured information. A very simple wrapper is put around the content, which is thus formatted in a lightweight markup system used in many Wiki applications, designed in a format-oriented manner to be transformed into HTML. As a result, the content is neither standardized (various applications use considerably divergent forms of the Wikitext language) nor truly structure-oriented. The actual conversion process is carried out in three main steps. Each of these steps can be performed separately:

(1) The comparatively large database dumps (the English dump is particularly unwieldy with its of 2 Gigabytes of data) are split into manageable smaller chunks. In this process, the currently available Wiktionary dump is split into roughly 85000 entries.

(2) The top-level constituents contained in these entries are identified and transformed into XML elements. This task is straightforward as the *entries* 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 entries (from 3KB up to 338KB) to be analysed. In the resulting sets, there are chunks containing *linguistic* information, e.g. part of speech. There are chunks containing *etymological* information and/or *usage* information. Many entries contain *morphological* data, in numerous cases complete

[7] See www.tei-c.org

[8] See www.tei-c.org/Activities/Workgroups/FS/

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

[10] For the time being this interface only offers German language data.

[11] See for a list of (nearly) all the categories: (http://en.wiktionary.org/wiki/Category:All_topics)

[12] See here a related study by (Nguyen & Ock, 2012), who are using Wiktionary categories for the lexical disambiguation in English, Korean and Vietnamese.

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

inflectional paradigms. The files also hold data concerning *hyphenation* and *pronunciation*. Semantic information is stored in sections describing the various *senses* of words. These, in turn, are linked to *translations*, as well as related *taxonomical* items such as synonyms, antonyms, hyperonyms, hyponyms.
(3) The last step in the transformation process is the conversion of the constituents into TEI P5. Iterating through the first untyped chunks, the program attempts to identify the right category and subsequently to translate it into the target TEI format. At this point, the main programming challenge was to merge the data coming from different chunks (e.g. senses and translations) into neatly nested XML structures.

2.2 Viennese Lexicographic Editor

Additional modifications of the output of the conversion process can be performed by making use of a software application that was developed at ICLTT. This dictionary writing client, called *Viennese Lexicographic Editor (VLE)*[14], is a standalone application that supports web-based dictionary editing; it was initially designed for collaborative glossary editing projects, but an enhanced version is now put to use by lexicographers. Basically, the system is meant to work with any kind of XML encoded data that can be organised in dictionary-like structures. It relies heavily on XML and cognate technologies such as XSLT and XPath.

2.3 Web Interface

The result data set of the conversion process described above has been put online on the ICLTT's showcase website[15]. It is used as an example dataset in a proof-of-concept implementation of our so-called *resource viewers*, visualisation tools for heterogeneous resources that can be easily deployed. This component of the system is implemented as a REST service which can be accessed both by applications and by human users. The web interface allows search and display of data by means of XSLT stylesheets or in form of the TEI P5 data.

2.4 Towards a cross-lingual Wiktionary Semantic Network

Ongoing work is dedicated to offering a semantic network across the entries in various languages, making use of the *categories* provided by Wiktionary, a concept for consistent labeling of domain semantics. For example, the entries "mother" (English), "Mutter" (German) or "anya" (Hungarian) in the English Wiktionary are respectively marked with the semantic labels *en:Family*, *de:Family* and *hu:Family*, so that one can easily link the TEI standardized representation of entries in the three languages that share the semantic category *Family*[16], and harvest related lexical items. An example: in the Hungarian page for "anya"[17] there are three synonyms, but in the English page for "anya"[18] there are five. These synonyms can be merged into a single list. As a result, one can boost the network of words that correspond to the *Family* sense.

3. Digital Humanities Use Case

Our main motivation for conducting the work of mapping Wiktionary lexical data into a standardized representation was to enable the multilingual extension of online resources in the field of folktales. One representative of such a resource is the Thompson Motif Index (TMI)[19].
TMI is organized along the lines of an alphanumeric order, which simulates a taxonomy of motifs, but does not express explicitly an 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.")*, some are abstractions of motifs *("A21. /Creator from above./")*, and some are summaries of a 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.")*. The decimal structure of the digital version of TMI can serve as a starting point to make such information explicit. We prepared a program that converts TMI to an XML representation and marks properties explicitly by designated tags:

```
<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>
```

Ongoing work is dedicated to upgrading the XML representation to 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[20].
The TMI catalogue is available only in English, but it would be desirable to use this catalogue for indexing German and Hungarian texts too. There is thus an obvious need to translate the descriptors of the motifs. In

[14] http://corpus3.aac.ac.at/showcase/index.php/vle
[15] http://corpus3.aac.oeaw.ac.at/showcase/index.php/wiktionary The tools and the data will also be freely downloadable.
[16] http://en.wiktionary.org/wiki/Category:Family
[17] http://hu.wiktionary.org/wiki/anya
[18] http://en.wiktionary.org/wiki/anya
[19] Cf. footnote 1.
[20] This work is documented in more details in (Declerck et a., 2012)

the case of the "Woman who fell from the sky", we can extract from Wiktionary German and Hungarian lexical equivalents for "woman" (*Frau* resp. *nő/asszony*), "sky" (*Himmel* resp. *ég/égbolt*) and the lemma of the inflected word form "fell" (*fallen* resp. *esik*).

Certainly, finding equivalent word-by-word translations is not enough to ensure an accurate translation, but navigating through the language-specific dumps allows to propose a semantic "reconciliation" of the words and to extend the lexical base. For example, the word "sky" is semantically related to several categories in the various Wiktionaries (Theology, Astronomy, Religion, Hereafter, Nature, etc.), so that entries from different languages around those categories can be grouped. Via this procedure, we also find new entries, e.g. the Hungarian "menny", corresponding to "heaven" (in turn related to "sky"), which is encoded in the Hungarian dictionary as a word related to Religion. We are thus working towards generating a machine-readable lexical semantics network for the folktale domain on the basis of Wiktionary data.

4. Conclusion and further Work

We have presented results of work dealing with the access to and transformation of Wiktionary language data, making them machine readable, in compliance with existing standards. Although our first goal was to obtain multilingual lexical entries that can help offering translations of Digital Humanities taxonomies, it turns out that the extraction and transformation of Wiktionary data can be valuable for the Language Technology community at large, as other studies have equally shown. Results of our work are being made available on the basis of open source licenses.

Our current activities include converting the Italian and the Persian Wiktionaries. We are also planning to port the TEI representation into RDF; and more specifically into the lemon model for lexicons in ontologies[21], which is developed within the EU-funded project Monnet[22]. In cooperation with the Monnet project, we are running extended experiments on the Wiktionary-based translation of labels of taxonomies in the field of folktales. Our resources will be contributed to the Linguistic Linked Open Data (LLOD) initiative[23].

5. 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. The work has also been accomplished as part of the ICLTT's CLARIN-AT/DARIAH-AT digital humanities initiatives.

[21] http://www.monnet-project.eu/Monnet/resource/Monnet-Website/0000%20-%20Library/0500%20-%20Files/lemon-cookbook.pdf

[22] See http://www.monnet-project.eu/ While the Monnet project focuses on the financial domain for localizing ontologies, we conduct our experiments in the field of folk literature.

[23] http://wiki.okfn.org/Wg/linguistics/llod#How_to_contribute

6. References

Budin, G., Majewski, S., Karlheinz Mörth, K. (2012). Creating lexical resources in TEI P5. A schema for multi-purpose digital dictionaries. In *jTEI* 3 (forthcoming).

Budin, G., Moerth, K. (2012). Hooking up to the corpus: The Viennese Lexicographic Editor's corpus interface. In *Electronic lexicography in the 21st century: New applications for new users* (eLEX2011)

Declerck, T., Lendvai, P., Darányi, S. (2012). Multilingual and Semantic Extensions of Folktale Catalogues. In: Proceedings of Digital Humanties (DH2012)

Henrich, V., Hinrichs, E., Vodolazova, T. (2011). Semi-Automatic Extension of GermaNet with Sense Definitions from Wiktionary. In: *Proceedings of the 5th Language & Technology Conference*. (LTC 2011)

Krizhanovsky, A. (2010). The comparison of Wiktionary thesauri transformed into the machine-readable format. (http://arxiv.org/abs/1006.5040)

Krizhanovsky, A., Lin F. (2009). Related terms search based on WordNet / Wiktionary and its application in ontology matching. In *Proceedings of the 11th Russian conference on Digital Libraries* (RCDL 2009).

John McCrae, J., E. Montiel-Ponsoda, E., Cimiano, P. (2012) Integrating WordNet and Wiktionary with lemon. In *Proceedings of the DgfS Workshop on Linked Data in Linguistic* (LDL 2012)

Meyer, C.M., Gurevych, I. (2010). Worth its Weight in Gold or yet another resource – A comparative study of Wiktionary, OpenThesaurus and Germanet. In *Proceedings of the 11th International conference on Intelligent Text Processing and Computational Linguistics*. Iasi (Romania) 2010: pp. 38-49

Moerth, K., Declerck, T., Lendvai, P., Váradi, T. (2011). Accessing Multilingual Data on the Web for the Semantic Annotation of Cultural Heritage Texts. In E. Montiel-Ponsoda, J. McCrae, P. Buitelaar, P. Cimiano (eds.): *Proceedings of the 2nd International Workshop on the Multilingual Semantic Web, Bonn, Germany, Springer, 10/2011*

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.

Nguyen, K-H., Ock. C-Y. (2012). Using Wiktionary to Improve Lexical Disambiguation in Multiple Languages. In *Proceedings of CICling 2012*.

Zesch, T., Mueller, C., Gurevych, I. (2008a). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation*. LREC 2008.

Zesch, T., Mueller, C., Gurevych, I. (2008b). Using Wiktionary for computing semantic relatedness. In *Proc. of AAAI conference on Artificial Intelligence*.