Software Projects for developing Digital Humanities Resources

Thierry Declerck
DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg, 3
D-66123 Saarbrücken
declerck@dfki.de

Abstract

In this short paper we report on experiences gained from bachelor and master theses, and from a series of software projects conducted in cooperation with the Department of Computational Linguistics of the Saarland University. Those bachelor/master theses and software projects were dealing with the application of Natural Language Processing and Semantic Web technologies to the representation and analysis of folktales. Data, codes and results of the software projects have been made available in various repository management services, like GitLab, GitHub or Bitbucket. We think that it will be important to discuss the design of such openly accessible repositories in order to ensure their re-usability and further extensions across various educational institutions.

1 Introduction

In the past 3-4 years we proposed in cooperation with the Computational Linguistics (CL) department of the Saarland University a series of bachelor/master theses and software projects, which were dealing with various aspects related to the wider field of folktales and therefore introducing Digital Humanities (DH) topics to students trained primarily to learn and apply computational methods of language technologies.

Our diagnosis was that the approach building on software projects for introducing CL students, and some few students from other departments, to Digital Humanities topics has been very successful. It is also the case that some of the projects we conducted have gained the interest of a broader public, including press coverage[1] and a broadcast programme[2]. We think that a main aspect of this success story lies in the fact that the students had to work together, building teams for working on modules and meeting for integrating the work done so far.

In all the 4 different software projects conducted until now, we could observe that the folktale topic was a driver calling for participation of a larger group of students (they can choose between different software projects). We describe in the following sections the types of approaches we followed and the results that the students generated and made available on various repository management services, like GitLab, GitHub or Bitbucket. The idea of having software projects as a platform followed the work done by two students in their master and bachelor theses, which were written in the context of their Research Assistant appointments within a larger national project[3]. We describe briefly the results of all those endeavours in the following sections.

2 Annotations

In the context of cooperation between the past D-SPIN[4] and AMICUS projects[5] a master thesis was written by the student Antonia Scheidel on the

[1] http://derstandard.at/2000004368363/Wenn-der-Computer-zum-Maerchenonkel-wird or http://www.abitur-und-studium.de/Bilder/Jana-Ott-Christian-Eisenreich-und-Christian-Willms-Studenten-von-Thierry-Declerck-haben-ein-Programm-entwickelt-das-Maerchen-vorlesen-kann.aspx

[2] See http://kulturellebildung.de/fa/user/Fachbereiche/Literatur_Sprache/Aktuelles/141121_PRESSE_Erzaehlen.pdf

[3] We do think that involvement of students as Research Assistant in projects is an important aspect to be considered.

[4] D-SPIN was a predecessor of CLARIN-D. See https://weblicht.sfs.uni-tuebingen.de/englisch/index.shtml

[5] AMICUS: Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts, was a Dutch project dealing partly with the annotation of folktales with recurrent Motifs. See https://ilk.uvt.nl/amicus/

annotation of fairy tales with Propp's functions[6]. Vladimir Propp "was a Soviet folklorist and scholar who analysed the basic plot components of Russian folk tales to identify their simplest irreducible narrative elements."[7]. Those basic plot elements are called by Propp "functions" and he identified 31 such functions, like "Interdiction", "Delivery" or "Rescue", etc. Propp also introduced circa 150 sub-functions that are specialisations of the 31 top-level functions. Complementary to the functions, Propp identified 7 broad characters, like "the villain", "the donor" or the "hero". The "morphology of the tale" described by Vladimir Propp was based on a subset of the so-called Afanasyev collection of Russian Folktales[8].

Antonia Scheidel developed a new annotation scheme according to which fairy tales can be queried for texts, temporal structures, characters, dialogues, and Propp's functions[9]. The annotation scheme has been named APftML, standing for "Augmented Propp fairy tale Mark-up Language". Antonia Scheidels' work is documented in (Declerck and Scheidel, 2010) and (Declerck et al., 2011). Annotated fairytale textual data is important in that automated systems have a data set against which they can map their results (see, for example (Scheidel and Declerck, 2010), describing an information extraction application in the folktale domain)[10]. If fairy tales are manually annotated with the annotation scheme, the results of the automatic processing can be compared with the human annotation.

3 Syntactic Analysis and a first Ontology

Based on the annotation framework mentioned in the previous section, Nikolina Koleva has worked for her bachelor thesis on an automated system for processing fairy tale texts. She considered for her work two tales, "The Magic Swan Geese", an English version of the Russian fairy tale "*Gusi-lebedi*", and "Väterchen Frost", a German version of the Russian fairy tale "*Djed Moros*". She has written a program that analyzes the text according to linguistic criteria, with the aim of recognizing the (main) characters in it, and storing those in a database. This database is of the "Ontology" type, on the base of which logical operations can be performed. The background is a formal description of what can be found in these fairy tales, including an ontology about family relations. Thus, the system can recognize that in the text "the daughter" is the same person as the "sister" when this is suggested by the context. This way, recognized characters in fairy tales are semantically annotated with more general categories, like "Woman". And we then know in which contexts (or situations) a specific family member (for example the "daughter") is involved (see (Declerck et al., 2012) and (Koleva et al., 2012) for more details on the results of her work.).

Once we had those resources, i.e an annotation framework for folktales, based in a first instance of the mark-up of Proppian functions, and an ontology framework in which characters playing a role in folktales are stored as instances of domain-specific classes, the idea was to extend those to a larger framework supporting DH application scenarios.

4 Approaches to Story Segmentation

In a first software project which was building on the top of the two resources mentioned in the previous sections, a division of work could be established between the four members of the project team. One task consisted in offering a meaningful segmentation of the tales. The approach for this consisted in automatically segmenting the tales along the lines of the dialogue structure. This had one motivation: to offer a base for the integration of a text-to-speech system supporting the "read aloud" of a tale, in which voices are associated to each contributors to the dialogues (and for sure one voice for the narrator). This application is described in more details in the next section.

The students worked in this project mainly on the English version of the "Froschkönig" tale (*The Frog Prince*)[11]. Following those new steps, the initial annotation format has been augmented with detailed dialogue descriptions. And the ontology has also been extended, including now a description of dialogues (questions, answers, monologues etc.), including the encodings of the participants

[6] See (Propp, 1968)

[7] https://en.wikipedia.org/wiki/Vladimir_Propp

[8] See https://en.wikipedia.org/wiki/Alexander_Afanasyev

[9] The annotation scheme can be downloaded at http://www.coli.uni-saarland.de/~ascheidel/APftML.xsd

[10] Examples of such annotated data can be downloaded at http://www.coli.uni-saarland.de/~ascheidel/APftML.xml

[11] See https://en.wikipedia.org/wiki/The_Frog_Prince

and the dialogue turns. In the two most recent and currently still running software projects the students are implementing a strategy on additionally segmenting a tale by the locations in which events are occurring. There is an interesting correlation between the segmentation by dialogues and the one by locations, as in this kind of narratives the participants to a dialogue are often sharing a location.

5 Emotions Detection and Text-to-Speech Modules

One student had the task to implement a program able to detect emotions. For this the original annotation scheme has been extended, supporting the mark-up of 6 basic emotions (fear, grief, joy, etc.), which are also encoded in the ontology. The automatic processing of the text (based in this case on the NLTK package[12]) was then marking the emotion detected in one sentence, on the base of a emotion lexicon build from annotated examples that served a as seed that was completed by consulting the WordNet[13] module implemented in NLTK[14].

A major extension of the past work in this software project was that synthetic voices also play a role. Once a character has been recognized, for example the princess (in the fairy tale "Frog King") additional features are coded (for example age, gender, emotion, etc.). Then a previously defined synthetic voice is automatically added to the character. And when the text is processed by the system, the story can be "told" by the voices. If there is no detected character in a dialogue situation, it is assumed that the narrator is the speaker and the reader the receiver. A demo can be heard in the corresponding Bitbucket repository[15]. In this software project we made use of the "Mary" Text-To-Speech System[16]. The overall results of the projects are described also in (Eisenreich et al., 2014).

[12]NLTK stands for "Natural Language Toolkit" and is written in Python, including a lot of corpus processing and statistical libraries. See http://www.nltk.org/

[13]See https://wordnet.princeton.edu/ for more details.

[14]See http://www.nltk.org/howto/wordnet.html for more details.

[15]The data, algorithms and results of the projects are stored in https://bitbucket.org/ceisen/apftml2repo. A demo of the TTS application is available at: https://bitbucket.org/ceisen/apftml2repo/src/cbf4d71de7f96146d17c4c84572ceb9a99cd300f/example%20output/audio_output.mp3?at=master&fileviewer=file-view-default

[16]See http://mary.dfki.de/ for more details.

6 Iterative Ontology Developments

We described in sections 4 and 5 how the original ontology has been enriched with additional features. In a second software project, work was dedicated in the ontologisation of classical knowledge – indexation and classification – resources in the field of folklore. We were considering in this software project two such resources: The "Motif-index of folk-literature" (Thompson, 1955 1958) and the "Types of International Folktales" (Uther, 2004). The first resource, which we abbreviate as TMI, is available as an on-line resource[17]. A folktale motif can be defined as a "repeated story element, e.g., a character, an object, an action, or an event that can be found in several stories"[18]. In TMI all motifs are organized in a tree structure, so that each motif has a more abstract class that describes a span of subordinated motifs. One motif entry consists of a motif-id, motif name, motif description (optional), and references to literature where it occurs.

The second resource builds on former work by Antti Aarne (Aarne, 1961) and Stith Thompson. This classification system was extended by Hans-Jörg Uther (see (Uther, 2004)), and in the following we are using the acronym ATU for referring to this resource. A folktale type can be described as a main story line that can be found in several cultures. The parts of this story line can refer to specific story elements also known as motifs. A folktale type is therefore a bigger unit than a motif.

Our approach consisted in extracting from those knowledge resources, which are stored in different formats, classification relevant information and to re-organize them in two interrelated ontologies, using for this the W3C standards OWL[19], RDF(s)[20] and RDF[21].

The integrated ontology resulting from the software project, also after curation done in the context of an internship at DFKI, contains 46,950 motifs for the TMI domain and 2802 elements for the ATU domain, most of them interrelated by corresponding properties. Results of this software project are available in a

[17]https://sites.ualberta.ca/~urban/Projects/English/Motif_Index.htm.

[18]https://en.wikipedia.org/wiki/Motif_(folkloristics)

[19]See http://www.w3.org/TR/owl-semantics/.

[20]See http://www.w3.org/TR/rdf-schema/formoredetails.

[21]See https://www.w3.org/RDF/ for more details.

GitLab repository: `https://gitlab.com/folktaleclassification/`.

An application of this new integrated ontology for the classification of characters in folktales has been presented in (Declerck et al., 2016) and more recent developments related to this integrated ontology are described in (Declerck et al., 2017).

7 Conclusion

We did report on specific teaching activities in the field of the representation and processing of folktales by students (mainly) in the field of computational linguistics. The specificity of the experiences we are reporting is that those activities took place in the context of software projects or internships, thus with a focus on practical implementation and development works. We noticed that this kind of team work, or also compact work done in the context of an internship, is delivering a very large amount of resources that are potentially very relevant for being reused in other type of teaching activities. Maybe also a coordinated action between universities and other educational institutions toward the organization of such software projects could be an idea to discuss and implement. Last but not least, many of the results presented in this short paper have been submitted to and accepted at relevant workshops and conferences, bringing the students thus also closer to this type of academic achievements.

References

Antti Aarne. 1961. *The Types of the Folktale: A Classification and Bibliography*. The Finnish Academy of Science and Letters. Translated and Enlarged by S. Thompson. Second Revision (FFC 184).

Thierry Declerck and Antonia Scheidel. 2010. An information extraction approach to the semantic annotation of folktales. In Sándor Darányi and Piroska Lendvai, editors, *First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*. University of Szeged, Hungary.

Thierry Declerck, Antonia Scheidel, and Piroska Lendvai. 2011. Proppian content descriptors in an integrated annotation schema for fairy tales. In *Language Technology for Cultural Heritage. Selected Papers from the LaTeCH Workshop Series*, Theory and Applications of Natural Language Processing, pages 155–169. Springer.

Thierry Declerck, Nikolina Koleva, and Hans-Ulrich Krieger. 2012. Ontology-based incremental annotation of characters in folktales. In Kalliopi Zervanou and Antal van den Bosch and, editors, *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2012)*, pages 30–35, 209 N. Eighth Street Stroudsburg, PA 18360 USA, 4. Association for Computational Linguistics (ACL), ACL.

Thierry Declerck, Tyler Klement, and Antónia Kostová. 2016. Towards a wordnet based classification of actors in folktales. In Verginica Barbu Mititelu, Corina Forascu, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Eighth Global WordNet Conference*. Global WordNet Association, GWA, 1.

Thierry Declerck, Antónia Kostová, and Lisa Schäfer. 2017. Towards a linked data access to folktales classified by thompsons motifs and aarne-thompson-uthers types. In *Proceedings of Digital Humanities 2017*. ADHO, 8.

Christian Eisenreich, Jana Ott, Tonio Sdorf, Christian Willms, and Thierry Declerck. 2014. From tale to speech: Ontology-based emotion and dialogue annotation of fairy tales with a tts output. In *Proceedings of ISWC 2014*. Springer.

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*, pages 467–470, Hamburg, 7. University of Hamburg, Hamburg University Press.

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

Antonia Scheidel and Thierry Declerck. 2010. Apftml - augmented proppian fairy tale markup language. In Sándor Darányi and Piroska Lendvai, editors, *First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*. Szeged University.

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.