Text Encoding and Ontology – Enlarging an Ontology by Semi-Automatic Generated Instances

Amélie Zöllner-Weber
amelie.zoellnerweber@gmail.com
University of Bergen, Norway

In this contribution, we present an application that supports users when working with ontologies in literary studies. Thereby, semi-automatic suggestions for including information in an ontology are generated. This application is meant for users, who are familiar with annotation and markup and are interested in topic of literary studies.

When reading literature we can identify literary phenomena but we cannot prove them directly in the text. Our ability is to puzzle sentences together so that they form a meaning. But this process happens in our mind not in texts. However, these interpretations are individual and can differ from reader to reader since they are influenced by our cultural and social background. It is therefore a challenge to create a model of these interpretations to be able to have a more general and formal description, e.g. of a character.

In computer philology, one can detect several applications when modeling texts: 1) by using mark up languages like XML (meta) information can be marked in texts (e.g. Jannidis et al. 2006, Meister 2003), 2) one can model theories in literary studies that try to represent mental representations (Jannidis 2004, Schneider 2000). However, text structures and mental representations can differ from each other so that we are not able to model them in the same way.

In Zöllner-Weber 2007, mental representations have been modelled by an ontology. It tried to regard a character as a complex cognitive entity in the reader's mind. Here, the description of literary characters has been realised as an ontology. For manipulating this ontology, users have to extract information manually about characters from literary texts and add them to the ontology. This process might be time-consuming, and users who are not familiar with the structure of an ontology might need even more time to become familiar with the application. We want to solve this problem by combining text encoding and the ontology. Therefore, an annotation system has been developed, which takes the mark up from the text and generates semi-automatically suggestions of instances be included into the ontology.

1. Methods

For the description of literary characters, an ontology that models characters by their mental representations was used (Zöllner-Weber 2006). Briefly, an ontology is a hierarchy of classes. In addition, the classes contain instances that represent individuals. Properties, which contain additional information, are attached to the individuals (Noy et al. 2001). By using this kind of structure, information is described formally. We chose an ontology because its structure corresponds to the mostly hierarchical structures of proposed theories to describe or analyse literary characters (Jannidis 2004, Lotman 1977). Several theories of literary characters are combined to create a base of a formal description using an OWL ontology (Grigoris and Harmelen 2003, Jannidis 2004, Nieragden 1995). The frame of mental representations is presented by the main classes of the ontology, e.g. inner and outer features, actions on other characters and objects. The sub classes contain characteristics of special characters (special features or groups of characters). We decided to include single pieces of information gained from literary texts into instances of the ontology. In addition, so-called instances of the classes represent individual and explicit objects of the domain of literary characters. Here, direct information about a character given in a text is assigned to an instance. Properties contain additional information, e.g. type of narrator, author, annotation information or reference to literature. Together with the information of the class hierarchy, instances and their properties, a single mental representation of a character is modelled (cf. Figure 1). In this

approach, individual description, the pre-step of interpretation, is focused. The main description categories secure a general classification so that it is also possible to compare two different interpretations of one character, which might be spread over different categories of the ontology.

In order to fill the aforementioned ontology of literary characters in a more automated fashion an encoding scheme has been developed. For the annotation, we selected tags of the TEI-DTD (Text Encoding Initiative 2003, `http://www.tei-c.org/`), which were developed for marking interpretation sections in texts. Thereby, the encoding scheme had to be exploited and rearranged so that it is usable for literary studies. This means that the usage of elements was enlarged. By using this special markup, a user can directly add interpretive pieces of information about a literary character to a text. Here, the annotation scheme is based on four main categories, *description*, *statement*, *action*, and *speech*, which classify pieces of information. All descriptions about a character that are stated by a narrator are subsumed under *description*. The category *statement* depicts commentaries of a character about another character. To mark non-verbal and verbal actions of a character, the categories *action* and *speech* should be used. In addition, a user should add e.g. information about the type of narrator, the name of a character and depending on the chosen category additional information to complete the annotation. After the process of annotation, a user sends the marked texts via a web form to a server where the annotations are evaluated by an in-house developed programming algorithm (cf. Fig 1). The pre-sorting of encoded information about a character is based on the four categories, which match the main classes of the ontology. If further encoded information is given by the markup, the algorithm tries to generate a further sub-classification. Figure 1 depicts an example of this process. After successful processing, a user is presented with a list for all processed annotations that probably form instances. Additionally, for all of these suggestions a class assignment is also given.

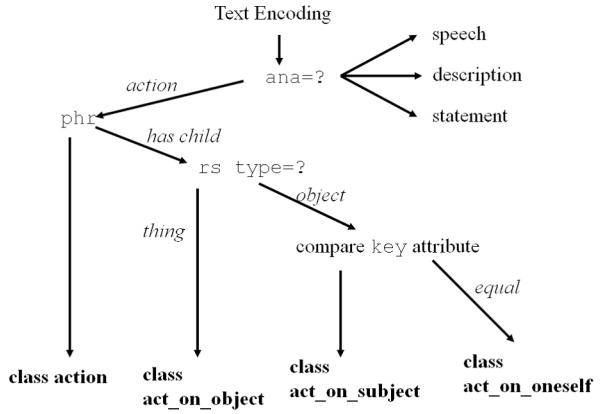

Figure 1

In addition, we present surrounding classes by showing an extracted list of classes of the ontology so that a user is able to inspect the environment of the new instance and its class. Whether a class that should include a new instance does not exist yet, a user can also add a new class. Afterwards, (s)he can include the instance in the new class.

2. Results

The application has been tested by using an extract of the novel "Melmoth the Wanderer" (1820), written by Charles Robert Maturin. We encoded the text with the mentioned TEI-DTD and afterwards, by using the programming algorithm, we obtained suggestions for new instances. In figure 2, the process of generating an instance from a text passage is shown as an example. For the main character Melmoth 72 instances were generated and assigned to the ontology.

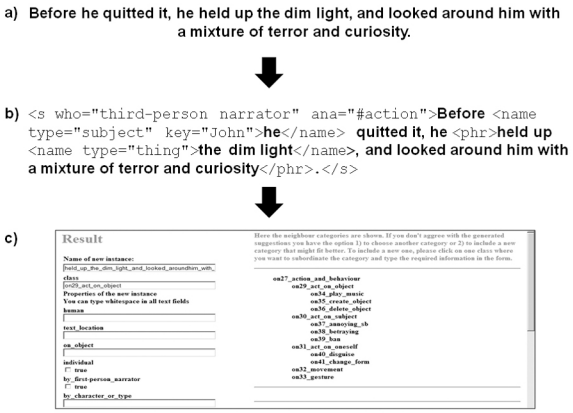

Figure 2

3. Conclusion

In this contribution, a system has been presented that includes information into an ontology, which is generated from markup. We tested this application by using an ontology for literary characters. In comparison to the manual manipulation of the ontology, the application comprises a semi-automatic generation of ontology instances and supports the user when assigning this information about a character to classes of the ontology. In addition, it is not only possible to add information about a single character to the ontology, but the application can simultaneously process annotations of several characters. Thereby, time and work can be saved, as the whole text can be annotated at once and will then be transferred to the ontology. There is no need to go back and forth between text and ontology as for the pure manual insertion of character information into an ontology.

Ontologies and their applications are often linked to logical reasoning. However, incorporating such techniques into the present application might be difficult, especially for untrained users, as shown elsewhere (Zöllner-Weber 2009).

References

Grigoris, A., Harmelen, F. V. (2003). 'Web ontology Language: OWL'. *Handbook on Ontologies*. Staab, S., Studer, R. (eds.). Berlin: Springer, pp. 67-92.

Jannidis, F. (2004). *Figur und Person - Beitrag zur historischen Narratologie*. Berlin: Gruyter.

Jannidis, F., Lauer, G., Rapp, A. (2006) (2006). 'Hohe Romane und blaue Bibliotheken. Zum Forschungsprogramm einer computergestützten Buch- und Narratologiegeschichte des Romans in Deutschland (1500-1900)'. *Literatur und Literaturwissenschaft auf dem Weg zu den neuen Medien*. Lucas, M. G., Loop, J., Stolz, M. (eds.). Bern.

Lotman, J. M. (1977). *The Structure of the Artistic Text*. Michigan: University of Michigan Press.

Meister, J. C. (2003). *Computing Action. A Narratological Approach*. Berlin/New York: Gruyter.

Noy, N. F., McGuinness, D. L. (2001). 'Ontology Development 101: A Guide to Creating Your First ontology'. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

Schneider, R. (2000). *Grundriß zur kognitiven Theorie der Figurenrezeption am Beispiel des viktorianischen Romans.*. Tübingen: Stauffenburg.

Zöllner-Weber, A. (2006). 'Formale Repräsentation und Beschreibung von literarischen Figuren'. *Jahrbuch für Computerphilologie*. **7**: 187–203.

Zöllner-Weber, A. (2007). 'Noctua literaria - A System for a Formal Description of Literary Characters'. *Data Structures for Linguistic Resources and Applications*. Rehm, G., Witt, A., Lemnitzer, L. (eds.). Tübingen: Narr, pp. 113-121.

Zöllner-Weber, A. (2009). 'Ontologies and Logic Reasoning as Tools in Humanities?'. *Digital Humanities Quarterly*. **3(4)**.