Published in: Digital Humanities, Lausanne - Switzerland '14. Lausanne, 2014, pp. 112-114.

An XML annotation schema for speech, thought and writing representation

Brunner, Annelen
annelen_brunner@gmx.de
Institut für deutsche Sprache

This contribution presents an XML schema for annotating a high level narratological category: speech, thought and writing representation (ST&WR). It focusses on two aspects: Firstly, the original schema is presented as an example for the challenge to encode a narrative feature in a structured and flexible way and secondly, ways of adapting this schema to TEI are considered, in order to make it usable for other, TEI-based projects.

The phenomenon ST&WR

ST&WR refers to the way the voice of a character is embedded in the narrator's text and is a feature that is present in most works of fiction. It has been widely studied in narratology, as it contributes to the construction of a fictional character, the narrator-character relationship and fictional world-building in general. Though ST&WR is partly defined by formal features like punctuation, verb mode, and sentence structure, narrative function is what is of interest in literary studies (cf. [1] for an overview). The challenge is to develop an annotation schema which is sufficiently structured to allow consistent annotation (especially with multiple annotators) and still captures nuances that are relevant for literary scholars.

The schema presented here – called ST&WR schema (ST&WR-S) – ties into literature studies as it uses categories agreed upon by most scholars and is similar to categorial systems proposed by narratologists Genette and Leech/Short (cf. [2], [3]). The main influence was a project of Semino and Short, who annotated a corpus of English fictional, newspaper and

(auto)biographical texts for ST&WR with an SGML-conformant schema (cf. [4]).

ST&WR schema

ST&WR-S was developed for manual annotation of a corpus of 13 German narrative texts written between 1786 and 1917 (about 57 000 tokens). This corpus was then used as a reference for the development and evaluation of automatic methods for ST&WR recognition (cf. [3]). The purpose of ST&WR-S was twofold: It allows for a very fine-grained classification of ST&WR instances which is helpful in order to study the phenomenon and to do statistical studies on manually annotated data, like in Semino/Short's project. On the other hand it was designed to be modular and easily simplified to accommodate for the rougher classifications of automatic recognizers. Experiences during corpus annotation strongly influenced the design of the annotation schema.

ST&WR-S has three levels of specificity: Main categories, attributes and in some cases different values for further specifications of certain attributes. These are modelled as XML tags with attributes and values.

The manual annotation was done in the GATE framework for natural language processing (cf. [1], http://gate.ac.uk). ST&WR-S is specified in XML schema files used by the plugin Schema_Annotation_Editor. Primarily, it is designed for inline XML, but GATE internally manages annotations as nodes and can convert them to a standoff XML format.

The main categories can be described with two axes: One axis represents the medium – speech, thought or written text (e.g. a quote from a character's letter). The second axis represents the four most common techniques of ST&WR: direct representation ("He said 'I am hungry.'"), free indirect representation ("Well, where would he get something to eat now?"), indirect representation ("He said that he was hungry."), and reported representation, which can be a mere mentioning of a speech, thought or writing act ("They talked about lunch."). This results in twelve main categories which are modelled as XML tags (*direct_speech*, *direct_thought*, etc.).

However, such a set of categories is necessarily rigid. When annotating a narrative phenomenon in a real corpus you will find many instances which are not clear-cut realisations of a predefined category. To deal with this fact, rather than just adding a confidence marker to the annotation, attributes are used to classify the type of deviation, so that the cases may be further studied and contrasted. As all attributes are optional and can be added to any main category, ST&WR-S allows for different levels of detail very easily. It is also possible to filter your annotation results afterwards by ignoring instances that carry a certain attribute.

Structurally, there is one numerical attribute (*level*), three attributes which are binary and just indicating whether the feature is present or not (*narr*, *prag*, *metaph*), two with optional further specification (*border*, *non-fact*) and one with mandatory further specification (*ambig*). All lists of attribute values are closed sets. Table 1 gives an overview.

Attribute name	Description	Values
level	level of embedment	numeric (default: 1)
ambig	ambiguity of the main category	Name of an alternative main category
non-fact	non-factual (eg. negated or hypothetical ST&WR) ("He did not admit that he loved her.")	*neg*, *hyp*, *fut*, *ques*, *imp*, *plan*, *unspec* (default: *unspec*)
border	borderline case of ST&WR("He knew that he had lost.")	*percept*, *feel*, *state*, *unspec*(default: *unspec*)
narr	Ambiguity between ST&WR and non-verbal action("She greeted her friends.")	binary (dummy value: *yes*)
prag	ST&WR, but with non-representional intent (e.g. politeness("I suggest you leave now.")	binary (dummy value: *yes*)
metaph	metaphorical use("His conscience told him to go.")	binary (dummy value: *yes*)

Functionally,*level* stands alone in the group as it does not mark a non-prototypical instance but is rather a 'monitor attribute'. It captures the level of embedment of a ST&WR instance, e.g. an instance of indirect thought that appears as part of an instance of direct speech would be tagged as *level="2"*. This marker can then be used to study the behaviour of such embedded instances and compare their behaviour to non-embedded ones.

All other attributes deal with instances that deviate from the prototypical idea of ST&WR in relation to the definition of the main categories.

Ambig and *narr* both mark ambiguity. While *ambig* indicates that there is uncertainty as to which main category should be applied, *narr* signals that it is uncertain whether the instance is a case of ST&WR at all.

Border deals with uncertainty in regard to what is considered speech, thought and writing respectively. Especially thought representation is extremely tricky, as you have to decide what constitutes a thought. For example, the sentence "He knew he had lost." would be marked as < *indirect_thought border="state"*>, as "to know" expresses a state of knowledge rather than a clear-cut thought. *Border* can also be applied to speech representation, e.g. if it is unclear whether there is a true verbalization like in the sentence "He screamed bloody murder."

Non-fact deals with instances where the ST&WR is non-factual and thus not a real 'representation' in the story world. Similarly, *prag* marks instances where ST&WR forms are used for non-representational purposes, especially politeness, and *metaph* represents metaphorical use of ST&WR.

In addition to that, the ST&WR-S contains two special categories modelled as XML tags. One is *frame*, which marks the framing clause of a direct representation which is not part of the representation itself but still interesting in the context of ST&WR. The other is called *embedded*. It can be used to mark embedded narratives which appear in direct representation (usually direct speech), e.g. if a character tells a story. Marking such cases with *embedded* essentially shifts the whole annotation level into a new narrative frame and gives it a different status than *direct_speech*. The use of *embedded* is optional and the tag can be easily transformed to *direct_speech* if this effect is not desired.

Adaptation to TEI Guidelines

ST&WR-S is a valid XML schema but not compliant to TEI Guidelines. For sustainability it would be desirable to adapt it, as this would allow its usage in TEI-conformant documents without compromising their validity.

However, such an adaptation is not straightforward. The logical starting point is <*said*>, a tag from the quotation context which is defined for passages thought or spoken by real people or fictional characters (cf. [7]). Though <*said*> is clearly intended to capture instances of ST&WR, its scope is narrower than the instances covered by ST&WR-S. In its core form, it only carries the attributes *aloud*and *direct*, both specified by truth values. *Aloud* is designed to distinguish between silent thought and passages spoken aloud (speech), but does not accomodate writing representation. *Direct* does not allow for any distinction between the ST&WR categories free indirect, indirect and reported. Of course, the rich attribute system of ST&WR-S does not have a predefined equivalent in TEI, either.

Several possibilities are considered how to adapt ST&WR-S while conserving its power as well as its modularity as much as possible. Ideas include use of standoff markup, possibly via the <*span*> tag, modelling of the complex categorizations via feature structures, referenced by the @*ana* attribute, or extentions of exisiting TEI-tags (most likely <*said*>).

References

1. **McHale, B** (2013). *Speech Representation*, in: Hühn, Peter et al. (eds.): The living handbook of narratology, Hamburg: Hamburg University Press, 2013 URL: www.lhn.uni-hamburg.de/article/speech-representation (last checked 17.10.2013).
2. **Genette, G.**(1980). *Narrative discourse*. An Essay in Method, Oxford: Blackwell.
3. **Leech, G. and Short, M.** (2007). *Style in fiction*. A Linguistic Introduction to English Fictional Prose 2. ed., London: Pearson Education Limited.
4. **Semino, E. and Short, M.** (2004). *Corpus stylistics*. Speech, writing and thought presentation in a corpus of English writing, London/New York: Routledge.
5. **Brunner, A** (2013). *Automatic recognition of speech, thought, and writing representation in German narrative texts*. Literary and Linguistic Computing 2013; doi: 10.1093/llc/fqt024
6. **Cunningham, H. et al.** (2011). *Text Processing with GATE (Version 6)*: www.tinyurl/gatebook (last checked 01.11.2013).
7. **Text Encoding Initiative** (2013). *P5: Guidelines for Electronic Text Encoding and Interchange*, URL: www.tei-c.org/Guidelines/P5 (last checked 31.10.2013)