„Handling the Implicit – Three Perspectives on Markup", 9 July 2014, Digital Humanities Conference 2014

Paper Session 2
Chair: Julianne Nyhan

XML-Print. Typeserng arbitrary XML documents in high quality
Georgieff, Lukas; Küster, Marc Wilhelm; Selig, Thomas; Sievers, Martin
Link zum Paper: http://dharchive.org/paper/DH2014/Paper-205.xml

An XML annota2on schema for speech, thought and wri2ng representa2on
Brunner, Annelen
Link zum Paper: http://dharchive.org/paper/DH2014/Paper-374.xml

Transcriptonal implicature: a contribution to markup semantics
Sperberg-McQueen, C. M.; Marcoux, Yves; Huigeldt, Claus
Link zum Paper: http://dharchive.org/paper/DH2014/Paper-61.xml

Report by: Peter Dängeli

The development of Digital Humanities, it could be argued, is inseparable from the development of markup languages.[1] The long paper session held on 9th of July 2014 chaired by **JULIANNE NYHAN** (University College London) was composed of three rather different perspectives on markup. Whilst the third contribution was concerned with the meaning of markup and thus relatively abstract in nature, the focus of the first two presentations was laid on the application of specific approaches for the rendering of encodings in print and for the annotation of speech, thought and writing as narratological entities respectively.

The session was opened with a paper by **LUKAS GEORGIEFF**, **MARC WILHELM KÜSTER**, **THOMAS SELIG** (University of Applied Science Worms) and **MARTIN SIEVERS** (University of Trier); the latter, a contributor well-known in the realm of (scholarly) typesetting, predominantly presented and demonstrated the overall results of the research project ‚XML-Print'[2] (2009 – May 2014) which is funded by the Deutsche Forschungsgemeinschaft (DFG). Arguing that paper remains the scholarly format that is accepted in most circles and that the production of printed output should neatly tie in with the source data, this project aimed to create an open source application that facilitates the typesetting of relatively complex (e.g. multilingual, critical) editions straight from XML markup – not necessarily but presumably in many cases based on TEI-XML. In order to specify the exact layout the user records a set of instructions in the XML-Print format/style editor that is available as a graphical user interface. These formatting instructions are formalised in an XSL-FO+ stylesheet, from which the typesetting engine creates the desired output. This workflow does not necessitate any alteration of the source documents, which is a great advantage over many existing typesetting methods. In order to satisfy the often arcane typesetting requirements of scholarly texts, the project team extended the common XSL-FO typesetting functions in order to (better) support multi-column layouts, cross-referencing/counting, and various apparatuses. The developers also devised a new line-breaking algorithm that, while retaining the high-quality typography

[1] For a number of early humanities research that built on markup technologies (and in turn influenced their development) cf. Boris Bosančić (2011). *Uloga opisnih označiteljskih jezika u razvoju digitalne humanistike* [*Descriptive markup languages and the development of digital humanities*]. Libellarium, IV, 1 (2011): 65 - 82 – http://ozk.unizd.hr/libellarium/index.php/libellarium/article/view/154/153; Desmond Schmidt (2012). *The Role of Markup in the Digital Humanities*. Historical Social Research Vol. 37, No. 3 (Controversies around the Digital Humanities): 125–146. – http://nbn-resolving.de/urn:nbn:de:0168-ssoar–37836

[2] http://www.xmlprint.de; zu einigen Illustrationen der Software siehe https://sites.google.com/a/budabe.eu/xmlprint_de/screenshots

output, excels the predecessor (devised in the early 1980s by D. E. Knuth and M. F. Plass[3], two pioneers in digital typography) in terms of efficiency.
The style editor (GUI) utilises a number of categories and dialogs in order to simplify the definition of formatting rules. The allocations of formats and mappings can be dragged and dropped in the order of their importance; higher entries take priority over lower ones in the course of the transformation. While the overall complexity of the publishing task does not vanish, the provision of the user with a graphical interface that allows setting up and testing transformation rules that e.g. apply to specific elements in defined contexts is a very valuable contribution. The learning curve for novice users indeed appears less challenging than for TeX[4]-based applications, TUSTEP[5] or commercial software such as Adobe InDesign[6].
XML-Print is designed to run as a standalone application, a TextGrid component, in batch mode on the CLI or on a webserver. The presentation could not remove the doubts entirely whether XML-Print „is taking the fight with the page model and winning" (M. Sperberg-McQueen), but the examples shown indicate that it can go a long way even when the requirements are very demanding.

The focus of the second contribution was laid on the rules on how to produce XML encodings in order to allow for a fruitful analysis of the contents. **ANNELEN BRUNNER** (IDS Mannheim) presented an XML annotation schema for narratological phenomena, specifically the annotation of the representation of speech, thought and writing as a narrative function in texts (shorthanded ST&WR by Brunner). Put differently, this schema should allow to formally describe how the voice of a character is realised by a narrator, referring to the four modes of direct, free indirect, indirect, and reported speech/thought/writing as they are agreed upon by most narratologists. Her attempt to schematically formalise ST&WR phenomena follows the path laid out by Semino and Short who devised a similar model using SGML in 2004[7] that indeed served as the main influence of her work.
Brunner's schema ought at the same time to be specific enough to conduct a very fine-grained manual annotation and yet of limited complexity to allow for automated recognition of ST&WR instances. Whereas Brunner had illustrated the problems and rates of success for the latter method at DH 2012 in Hamburg[8] (and published on the matter in LLC[9]), she now elaborated on the specific properties of the suggested XML data model, that was of course influenced by the findings of her work with automated recognition. Pairing the three narratological categories of speech, thought and writing with the four modes of (in)directness, Brunner arrived at twelve basic elements. In the event of difficulty in adding an instance to a category, seven optional attributes with a closed set of values are used to specify the ambiguity or the type of deviation, in order to treat occurrences that are not prototypical. Using this model, information on the factuality, usage of metaphors, alternative categorisations, or the level of embedded ST&WR representations can be recorded in a structured manner and subsequently undergo computational analysis.
While this model serves its purpose and has successfully been tested on a (German) corpus, it currently requires an additional layer of abstraction when annotating TEI documents as the proposed schema is not compliant with the TEI Guidelines. Given the wide adherence to TEI in the encoding of literary texts, compliance is however desirable. Using *<tei:said>* as the salient value of reference, Brunner exemplified the severe restrictions that the guidelines impose, although the element is clearly intended to capture the narratological phenomena in question. The ways towards compliance pointed out by Brunner (use of

[3] D. E. Knuth, M. F. Plass (1981). *Breaking Paragraphs Into Lines*. Software: Practice and Experience 11/11: 1119–1184 - http://dx.doi.org/10.1002/spe.4380111102.
[4] http://tug.org
[5] http://www.tustep.uni-tuebingen.de
[6] http://www.adobe.com/products/indesign
[7] E. Semino, M. Short (2004). *Corpus stylistics. Speech, writing and thought presentation in a corpus of English writing*. London/New York: Routledge.
[8] http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/automatic-recognition-of-speech-thought-and-writing-representation-in-german-narrative-texts/
[9] A. Brunner (2013). *Automatic recognition of speech, thought, and writing representation in German narrative texts*. Lit Linguist Computing 28 (4): 563–575. - http://dx.doi.org/10.1093/llc/fqt024

standoff markup, categorisation modelling using *feature structures*, extension of the guidelines, possibly in form of a module) were briefly discussed, yet their adaptation is not straightforward by any means, leaving TEI-compliant annotation of speech, thought and writing representation a desideratum for the time being.

These two accounts on the creation and application of (XML) markup starkly contrasted with the third presentation, delivered by **MICHAEL SPERBERG-MCQUEEN** (Black Mesa Technologies LLC), **YVES MARCOUX** (Università de Montréal) and **CLAUS HUITFELDT** (Universitetet i Bergen). Markup semantics have been a concern to Huitfeld, Sperberg-McQueen, Marcoux (and other scholars, e.g. Paul Caton, who gave a related talk at DH 2014[10]) for a number of years, and various publications and contributions to conferences give an account of this endeavour, that can be described as an attempt to comprehensively formalise the activity of transcription (using markup, i.e. artificial languages) and perhaps even as an approximation to „formalizing the meaning of arbitrary natural-language utterances"[11]. Over time Huitfeld, Sperberg-McQueen and Marcoux developed a logical model that encompasses notions such as *surfaces* (perceptible, measurable, non-ephemeral carrier of marks), *marks* (meaningful signs on a surface), *tokens* (reading-instantiated marks of a particular type), *types* (abstract token-instantiated objects), *readings* (token-type mappings), *type-sequences*, *token-sequences*, *documents* (mark-bearing physical objects), *exemplars* (specific token-sequences with respect to transcriptions) or *transcriptions* (transcribed token-sequences with respect to exemplars)[12].
Whereas earlier work on markup semantics found that the degree of similarity between transcripts and exemplars is not a useful measure due to the difficulties of operationalisation, the focus has more recently been directed to the transcription conventions that are adopted in communities of practice. For many projects, editorial statements and encoding descriptions detail the decisions that were taken with regard to the encoding of specific phenomena, e.g. editorial deletions, extensions or line breaks. Yet such statements typically only describe how encoding decisions vary from practices regarded as usual (within the community), and what is taken as common practice is not further elaborated. Following this observation, Huitfeld, Sperberg-McQueen and Marcoux suggested the extension of their logical model by the notion of *transcriptional implicatures*. A *conversational implicature* (after Herbert Paul Grice) denominates a meaningful predicate that remains unspoken, but is still successfully communicated due to underlying conversational maxims to which communicating individuals (normally) adhere, transcriptional implicature denotes a set of transcriptional rules that apply by default, but are usually not made explicit.
What constitutes this implicature may vary between communities of scholarship, and whether there is an overarching subset that covers several or all of these communities would need to be investigated empirically. Huitfeld, Sperberg-McQueen and Marcoux postulated however, that there is a default set of rules for transcriptional implicature, in relation to which any transcriptional implicatures of the various communities of practice can be described. The attempt to outline this hypothetical default transcriptional implicature led to observations such as top-level identity of types, reciprocity, completeness, purity and thorough type-similarity (between exemplar and transcript, respectively). First tests on the cogency of these basic rules with randomly chosen editions indicated that community-specific rules indeed specify deviations from these tacitly assumed common grounds. By formalising more statements of encoding practice, the

[10] Cf. P. Caton (2014). *Six terms fundamental to modelling transcription*. Digital Humanities 2014. Book of Abstracts: 125–126. - http://dharchive.org/paper/DH2014/Paper–780.xml
[11] M. Sperberg-McQueen, Y. Marcoux, C. Huitfeld (2014). Transcriptional implicature: *a contribution to markup semantics*. Digital Humanities 2014. Book of Abstracts: 360. - http://dharchive.org/paper/DH2014/Paper–61.xml
[12] Cf. C. Huitfeldt, C. M. Sperberg-McQueen (2008). *What is transcription?* Literary & Linguistic Computing 23.2: 295–310; C. M. Sperberg-McQueen, C. Huitfeldt, Y. Marcoux (2009). *What is transcription? Part 2*. Talk given at Digital Humanities 2009, College Park, Maryland; P. Caton (2009). *Lost in Transcription: Types, Tokens, and Modality in Document Representation*. Presented at Digital Humanities 2009, College Park, Maryland; C. Huitfeldt, Y. Marcoux, C. M. Sperberg-McQueen (2010). *Extension of the type/token distinction to document structure*. Presented at Balisage 2010, Montréal, Canada; P. Caton (2013). *Pure Transcriptional Markup*. Presented at Digital Humanities 2013, University of Nebraska, Lincoln as well as the foundational concepts of token and type by Charles S. Peirce (1909).

scholars plan to learn more about the implicatures of various communities, to fine-tune their logical model with regard to them, and to propel the formulation of a concise logical account of transcription.

The interest for markup related topics – be it theoretical considerations, methodological questions or newly crafted applications – has obviously not diminished in the community. The arrangement of presentations for this session, related yet contrastive, drew a large audience and an auditorium twice the actual size would have been a better fit, considering the many persons standing along the walls or in front of the room entrance. Interestingly, many of the topics touched upon and questions raised during the sessions (cost of TEI compliance, markup or rather markdown semantics) were, besides many other challenging problems related to the Bengali writing system or the sheer scope of the edition, reiterated in Sukanta Chaudhuri's plenary lecture on the online variorum edition of the complete works of Rabindranath Tagore in English and Bengali[13] that concluded the DH 2014 conference on Friday, 11th July 2014.

Peter Dängeli
Universität Köln
p.daengeli@uni-koeln.de

[13] http://bichitra.jdvu.ac.in