XML and multimodal corpus design: experiences with layered stand-off annotations in the GeM corpus

John Bateman[1], Judy Delin[2,3], Renate Henschel[2]

1 University of Bremen, Bremen, Germany
2 University of Stirling, Stirling, FK9 4LA, Scotland
3 Enterprise IDU, 79 High St., Newport Pagnell, Bucks MK16 8AB, England.
Contact author:
Judy Delin
(j.l.delin@stir.ac.uk)

www.purl.org/net/gem

Abstract

Current views of multimodal language resources have not yet sufficiently captured the complex interrelationships within page-based information delivery. This is critical for development of multimodal corpora and language resources suitable for large-scale empirical investigation. Serious attempts to interrogate the nature of multimodal meaning-making in professionally-produced documents, both paper and electronic, require a clear understanding of the organisation of the layers into which meaning is organised. In this paper, we present the first multi-layered XML annotation scheme that meets these requirements, developed using a combination of expertise from computational linguists and designers from various sectors of the publishing industry.

The corpus design we present has been developed within the ongoing GeM (Genre and Multimodality) project (see, e.g., Delin, Bateman, Allen (forthcoming), Delin and Bateman (in press)) and encompasses a scheme for corpus annotation that permits the collection of multimodal documents accompanied by close classifications of their forms, functions and contexts. The corpus consists of a single base-level that identifies all document elements (sentences and sentence fragments, graphics, diagrams, connecting arrows, frames, etc.), combined with several levels of stand-off annotation that impose differing and non-isomorphic structures over the units of the base level. Each level is represented fully in XML and appropriate DTDs are specified. Contact is made with currently emerging standards in document content and layout specification including the extended style language transformation (XSLT) and formatting layers (XSL:FO). Relations between the layers of description are given in the standard XML linking language; we are evaluating a variety of query languages for picking out and testing regularities across the corpus.

The purpose of the corpus development and the GeM project in general is to investigate the systematic connections that can be drawn between a rich characterisation of the context of use of multimodal documents and their linguistic, graphical, and layout realisations. Within the GeM project itself, four broad document genres have been selected for initial treatment: traditional paper-based newspapers, online web-based newspaper sites, instructional documents, and wildlife books; in each area we have secured a collection of documents and have established contact with designers either expert in these respective fields or, in several cases, actually responsible for the documents gathered. The layers that are represented in the corpus are constructed in terms of:

- content structured according to standard computational knowledge representation techniques
- rhetorical description drawing on an extended form of Mann and Thompson's Rhetorical Structure Theory (1988)
- a hierarchical layout structure
- a navigational structure for guiding document consumption by the reader

Following and extending the initial ground-breaking work in document design of Waller (1987), we claim that not only is it possible to find systematic correspondences between these layers, but also that those correspondences themselves will depend on specifiable aspects of their context of use. In particular, they will depend on `canvas constraints' set by the nature of the realizational medium (paper, screen-based browser, palmtop, screen resolution) and `production constraints' imposed by available technology and design choices (allowable cost, number of pages, available printing or rendering techniques, etc.). Our provision of a corpus of multimodal documents serves as the empirical basis for more thorough investigations of this claim. So far our work has identified widespread mismatches between rhetorical purposes and layout structures even among professionally produced documents; this offers a useful basis for constructive critique.

The final goal of the GeM project is to incorporate the empirically revealed generic constraints between document type and document form in a prototype computational system for both fully and semi-automatic multimodal document generation. For this we are investigating conditionalized translation processes written largely in XSLT that can produce XSL:FO formatting objects documents as output. Such output can then be fed to industry-standard XSL:FO renderers for producing Postscript, Acrobat, etc. versions of the final documents for evaluation by professional designers. We suggest that this approach is not only a fruitful one for multimodal document analysis, but for extension into representation and interrogation of other media. An adequate understanding of the complexities of dealing seriously with the language, typography and layout of 2D documents is a pre-requisite for the specification of any useful roadmap for multimodal language resources and evaluation.

Delin, J., Bateman, J. and Allen, P. (forthcoming) A model of genre in document layout. *Information Design Journal.*

Delin, J and Bateman, J (in press). Describing and critiquing multimodal documents. *Document Design*, **3**(2). Amsterdam: John Benjamins.

Mann, W. and Thompson, S. (1988) Rhetorical Structure Theory: toward a functional theory of text organisation. *Text*, **8**, 243-281.

Waller, R. (1987) *The typographical contribution to genre*. Unpublished PhD dissertation, Department of Typography and Graphic Communication, University of Reading.