Understanding Texts As Graphs

Bleeker, Elli; Buitendijk, Bram; Haentjens Dekker, R.; Kulsdom, Astrid

2018

Understanding Texts as Graphs: An Inclusive Approach to Text Modeling

Elli Bleeker
Research and Development
Humanities Cluster
Royal Dutch Academy of Sciences
elli.bleeker@di.huc.knaw.nl

Bram Buitendijk
Research and Development
Humanities Cluster
Royal Dutch Academy of Sciences
bram.buitendijk@di.huc.knaw.nl

Ronald Haentjens Dekker
Research and Development
Humanities Cluster
Royal Dutch Academy of Sciences
ronald.dekker@di.huc.knaw.nl

Abstract

The paper introduces TAG, a hypergraph data model for the modeling and processing of text. The features of a hypergraph allow for an inclusive and idiomatic approach to humanities text, and support a wide range of research perspectives. Furthermore, editing texts as hypergraphs gives touches upon pivotal issues regarding our understanding of text.

1 Introduction

It is a given that the complex nature of textual studies poses a set of interesting challenges for modeling, processing, and representation. In and by themselves texts constitute a "complicated web of interwoven and overlapping relationships of elements and structures" (Vanhoutte, 2006) and the information within this web is often implicit. Moreover, within the humanities text is rarely taken to be straightforward or linear: modeling textual information results in multi-layered and non-linear objects. Elsewhere we discussed the advantages of a particular type of graphs – property hypergraphs – for the modeling of text, introducing the new "Text as Graph" (TAG) model of text (2017) and demonstrating how to model textual variation in TAG (2018). The present paper discusses an implementation of TAG, the *Alexandria* repository, that supports the editing of multi-layered and non-linear documents in an idiomatic way. *Alexandria*'s potential gives rise to a number of questions that are crucial for the field of computational humanities. Can we, in fact, represent our knowledge of a text for others to benefit from and interact with? What does it mean when our textual models are, conceptually, no longer limited by a particular format or structure? A description of *Alexandria*'s workflow allows us to address these and similar questions and leads to a reconsideration of our understanding of modeling and examining texts, in the humanities and beyond.

2 Modeling Text as Graphs

Since graph structures by definition impose not one single hierarchy on the data they contain, graphs address well-known issues like overlapping hierarchies that often arise when aspects of text and document are structured as hierarchical trees. They thus seem a logical choice to model non-hierarchical textual features like discontinuity, nonlinearity, and (self)overlap. These functionalities are supported even better in a hypergraph structure as it builds on the characteristics of a directed acyclic graph (DAG), adding some qualities that are specifically valuable for the modeling of unstructured data like humanities texts. The advantages of graphs for text modeling have been discussed before and have led to alternative data

models[1] These graph data models are primarily concerned with overlap, one of the white whales of markup theory and practice. TAG, making use of a property hypergraph, deals with overlapping structures in a natural manner and is able to deal with discontinuous and non-linear aspects of text as well. Hypergraphs are used extensively in mathematics and computer science, but as of yet they have not been applied to the domain of text modeling. In short, the TAG model consists of Text Nodes, Markup Nodes and Annotation Nodes. A node may be connected to one or more nodes with hyperedges. Currently, TAG has two implementations: the collation engine *HyperCollate* and the repository *Alexandria*. Below we give a brief outline of *HyperCollate*, to illustrate the value of an inclusive approach to examining textual variation[2], followed by a description of *Alexandria*'s editorial workflow. The main goal of the paper, however, is not to present these tools but rather to assess the conceptual implications of TAG's inclusive and advanced approach to text modeling.

2.1 HyperCollate

The collation engine *HyperCollate* makes use of a hypergraph model for textual variation. *HyperCollate* can thus natively process both markup and text characters, as well as more than one hierarchical structure. Most existing collation tools do take TEI-XML encoded transcriptions as input, but they collate the witnesses on a plain-text level (string characters) only. Transforming TEI-XML files into character strings conveniently removes the need to deal with issues like overlap on a programmatic level, but removing the markup inevitably entails information loss. That is, intradocumentary variation[3] and structural variation (paragraphs, chapters, etc.) are generally ignored even though they are valuable aspects of a text's development.[4] *HyperCollate*, in contrast, uses the valuable intelligence that is expressed in markup to improve the analysis of textual variation. It results in an exhaustive representation of the variance within and between different versions of a literary work, thus allowing scholars to better analyze the dynamic nature of literary text. Furthermore, *HyperCollate*'s technology of comparing documents on the level of text and markup is similar to the way TAG documents are managed in the *Alexandria* repository.

2.2 Alexandria

The design of the repository *Alexandria* addresses an important issue for modeling in the humanities, which is identified in the workshop's call for papers as "the particular challenges posed by humanities research, e.g., [...] different positions (points of view, values, criteria, perspectives, approaches, readings, etc.)?"[5] The repository stores multiple TAG documents, each of them a hypergraph consisting of Text Nodes, Markup Nodes, and Annotation Nodes. Since a TAG document in its full hypergraph glory contains more information than can be visualized in any informative way, Alexandria allows users to *check out* a *view* on the TAG document. A view is defined as a version of a TAG document with one or more layers of markup. Assuming that users are (almost) never interested to see every aspect of a text, we provide them with the possibility to focus on specific aspects and ignore others. Simply put, users can identify the markup layer(s) they are interested in, *check out* from the Alexandria repository a version of the TAG document with this specified set of markup (the view), editing this view, and *check in* the edited view back into the repository.[6] The edited view is merged with the TAG master file in the repository which thus contains a wealth of information and knowledge about the textual object. It can be continuously enriched with new information from various scholarly perspectives. In other words, a single TAG document can be studied from a wide range of research perspectives and used by scholars from different disciplines, from history to linguistics and from textual genetics to

[1] See GODDAG (Sperberg-McQueen and Huitfeldt, 2000), GrAF (Ide and Suderman, 2007), and Extended Annotation Graphs (Barrellon et al., 2017).

[2] For a more extensive discussion of HyperCollate, see (Bleeker et al., 2018)

[3] Intradocumentary variation can be defined as in-line or in-text variation, e.g., the authorial revisions on one manuscript document. It can be contrasted with *interdocumentary* variation which manifests itself only by comparing two or more documents

[4] Although a number of tools retain certain markup elements in order to visualize revisions in the collation result, e.g. the Beckett Digital Manuscript Project's implementation of CollateX (see https://collatex.net/doc/ or Juxta Commons http://www.juxtasoftware.org/juxta-commons/), these elements play no (analytical) role for the alignment of the tokens.

[5] See http://wp.unil.ch/llist/event/comhum2018/

[6] The repository's workflow is inspired by Git, an open source and distributed version control system used in the software development community (see https://git-scm.com/).

paleography.

3 Modeling Perspectives on Text

A closer look at workflow of editing documents in *Alexandria* may clarify matters. Let us assume, for instance, that user C ("Claire") wants to focus on the material aspects of a manuscript and user D ("Dirk") only cares for the linguistic properties of the text on that manuscript. Claire creates a transcription and uploads her TAG document in *Alexandria*. Dirk subsequently wants to work on the same document but as he's not interested the materiality of the document, he checks out a view that contains only a small amount of Claire's markup. Dirk adds his own markup, possibly also altering some textual content, and commits his document in *Alexandria*. Dirk's view, then, is merged with the master file which now has several layers of markup, containing information about the materiality of the source document as well as the linguistic aspects of the source text. The technical implications of storing multiple perspectives on the same text are twofold: first, it inevitably leads to overlapping structures. Secondly, merging two "views" means dealing with document changes on the level of both text and markup. Both issues constitute important research endeavors in and by themselves that have captivated the field of computational humanities for some time now. Yet the TAG hypergraph model of text and textual variation addresses these technical challenges to a large extent. More interestingly, therefore, are the conceptual implications of this approach as they provide an opportunity to scrutinize our very definition of text modeling. If we can store a theoretically infinite amount of layers of information on a text, our very definition of textual editing may very well change. What is more, removing the need to separate text and markup before processing a file sheds new light on the age-old question what text really is.

References

Vincent Barrellon, Pierre-Edouard Portier, Sylvie Calabretto, and Olivier Ferret. 2017. Linear extended annotation graphs. In *Proceedings of the 2017 ACM Symposium on Document Engineering*. ACM, pages 9–18.

Elli Bleeker, Bram Buitendijk, Ronald Haentjens Dekker, and Astrid Kulsdom. 2018. Including xml markup in the automated collation of literary texts. In *Proceedings of the XML Prague Conference 2018*. XML Prague, pages 77–97.

Ronald Haentjens Dekker and David Birnbaum. 2017. It's more than just overlap: Text as graph. In *Proceedings of Balisage: The Markup Conference 2017. Balisage Series on Markup Technologies, vol. 19*. Balisage.

Nancy Ide and Keith Suderman. 2007. Graf: A graph-based format for linguistic annotations. In *proceedings of the Linguistic Annotation Workshop*. Association for Computational Linguistics, pages 1–8.

C Michael Sperberg-McQueen and Claus Huitfeldt. 2000. Goddag: A data structure for overlapping hierarchies. In *International Workshop on Principles of Digital Document Processing*. Springer, pages 139–160.

Edward Vanhoutte. 2006. Traditional editorial standards and the digital edition. In *Learned Love. Proceedings of the Emblem Project, Utrecht Conference on Dutch Love Emblems and the Internet*. pages 157–174.