Narrative Markup with TEI

Ben Miller, Jennifer Olive, Ayush Shrestha, Jin Zhao, Nicholas Subtirelu, Yanjun Zhao

Georgia State University; jolive1@gsu.edu

Capturing rich human experiences, narrative is an important site of inquiry in various disciplines across the humanities. The formalized study of narrative finds its roots in the seminal works of Russian formalist linguists and literary theorists. One of the earlier works from this tradition that articulates the kind of methodology employed for formalist study is Vladimir Propp's Morphology of the Folktale (1962) in which he compared the narrative structures from various stories in a corpus of Russian fairy tales by systematically classifying the functions of the dramatis personae. Following Propp, others, such as Gérard Genette (1972), Tzvetan Todorov (1973), and Mieke Bal (1985), have further developed methods to identify and classify narrative elements and to construct and analyze the structure of narrative text for the purposes of determining a definition for narrative. These early Russian formalists' works, especially that of Propp, laid the foundation of narratological analysis, which, well received by influential Western structuralists such as literary critic Roland Barthes (1975) and anthropologist Claude Lévi-Strauss (1976), has become an important paradigm of inquiry in humanities and social sciences.

Traditionally in the humanities, theories and research paradigms that situate narrative as the site of inquiry, including narratological analysis, rely on close reading of text and qualitative analysis, e.g., famously, Kenneth Burke's "dramaticism" in rhetorical studies (1969) and Roland Barthes's (1970) close reading of literary texts. Narratological analysis derived from the Russian formalist tradition, however, offers a more quantitative approach precisely because of its focus on linguistic structure of the text instead of semantic interpretation, which creates opportunities for computer automation and distance reading in larger corpora. In social sciences, automated narrative analysis methodologies informed by this tradition of narratological analysis have been developed and applied in research dealing with large corpora. Notably, sociologist Roberto Franzosi (2010) developed a method of quantitative narrative analysis (QNA) by applying the methodology of the structuralist narrative analysis tradition to the functions of computer science. The application of such methods to analyzing large corpora of narrative texts, such as Sudhahar, Franzosi, and Cristianini's (2011) analysis of 100,000 crime articles from the New York Times, can yield insights impossible to achieve with qualitative methods. The application of such automated narratological analysis methods demonstrates the potential for new research opportunities in humanities; however, in order to seamlessly share data across research projects and disciplines, there is an urgent need for the development of a standardized markup schema reflective of such methods.

Recent efforts in segmenting and classifying narrative elements involve applying aspects of the Russian formalist theoretical tradition and modifying them for use in computational processes. One of the most promising applications of a markup suitable for computational processes is presented in Inderjeet Mani's Computational Modelling of Narrative (2013). Mani's solution involved integrating the elements most useful for analyzing aspects of narrative discourse as well as fabulaic elements into an easily understood XML markup schema. The primary issue with this schema, however, lies in the issue of transferability between projects. Without referencing a centralized taxonomy, it becomes difficult to use, augment, expand this schema to similar projects in throughout the humanities. Another issue prevalent in decentralized markup schema is that of definition. This issue can be seen in Bod et al.'s (2012) attempts to recreate a Proppinian analysis in which they asked individuals to annotate a selection of the same fairy tales using the methodology outlined in Propp's Morphology. During their experiment to determine the issues of automating a Proppinian methodology, the researchers found that the string length and identification of dramatis personae improved in relation to Propp's results when participants were given shorter definitions, limitations to their options regarding the exclusion of sub-functions, and an example. While not their purpose, this experiment speaks to the need for an easier reference source for annotators to reference as well as the need for that reference source to include more intuitive or, at least, flexible explanations. The Text Encoding Initiative (TEI) offers a solution to these types of reference problems by providing a large schema organized into modules for use in a variety of encoding projects using XML as well as a robust system of documentation that aids humanitarians in a variety of disciplines in their respective applications of the larger TEI schema. By noting the availability of such a reference as TEI, we would like to argue for the creation of a standardized narrative schema in TEI to further aid in the connection and interaction of narrative projects throughout the humanities.

To support and help further explain the necessity for developing a standardized narrative schema using TEI, we will describe portions of our project, an NSF-funded Digging into Data Challenge project entitled "Digging into Human Rights Violations," in which we found that the incorporation of an established narrative schema would be helpful to the final goals of the project and, ultimately, make the project's output more comprehensible to other humanities scholars working on similar projects. One of the goals for this project is to automatically tag fabula-level information for individual narrative texts within a larger corpus, which, when visualized, represent existing and emergent narratives related to a collectively traumatic event. For this purpose, we have incorporated a variety of extraction tools including NLTK for entity recognition, SUTime for time stamping, Google Map API for geotagging, and TARSQI for event recognition. The issue with this automatic approach, however, regards the output of the documents using the various annotations for their respective programs. To generate a working version of each individual text that is usable for our purposes as well as understandable to a human annotator, all of these outputs must again be combined into a singular text after their initial processing. This integration, however, presents a challenging step in the further implementation of the project because the output of the tools, while in XML format, might consist of (1) overlapping tags and (2) different taxonomies as each of software utilizes a specific library. These issues prove issues not only in the pipeline portion of our project but also to the use of XML as a message-passing paradigm. To rectify such issues, we propose the integration of a TEI-compliant narrative schema which would assist in (1) the standardization of the taxonomy for the purposes of the project and other projects using similar data, and (2) the hierarchical tagging structure to alleviate the overlap issues produced in the taxonomies.

Bibliography

Bal, Mieke. Narratology: Introduction to the Theory of Narrative. 2nd ed. Toronto: U of Toronto P (1985). Print.

Barthes, Roland and Lionel Duisit. "An Introduction to the Structural Analysis of Narrative." New Literary History 6.2 (1975): 237-72.

Barthes, Roland. S/Z. New York: Hill and Wang (1970).

Bod, Rans, Bernhard Fisseni, Aadil Kurji, and Benedikt Löwe. “Objectivity and Reproducibility of Proppian Narrative Annotations.” The Third Workshop on Computational Models of Narrative (CMN ‘12). Lütfi Kirdar Istanbul Exhibition and Congress Centre. Istanbul, Turkey. 26-27 May 2012. Print. 17-21.

Burke, Kenneth. A grammar of motives. Los Angeles: University of California Press (1969).

Franzosi, Roberto. “Narrative Analysis-Or Why (And How) Sociologists Should be Interested in Narrative.” Annual Review of Sociology 24 (1998): 517-54.

Franzosi, Roberto. Quantitative Narrative Analysis. Los Angeles: Sage (2010).

Genette, Gerard. Narrative Discourse: An Essay in Method. Trans. Jane E. Lewin. Ithaca, New York: Cornell UP (1980).

Lévi-Strauss, Claude. “Structure and Form: Reflections on a Work by Vladimir Propp.” In Structural Anthropology, Vol 2. Chicago: U of Chicago Press (1976): 115-45.

Mani, Inderjeet. Computational Modelling of Narrative. Lecture #18 in Synthesis Lectures on Human Language Technologies series. Ed. Graeme Hirst. Morgan & Claypool Publishers (2013). Ebook.

Propp, Vladamír. Morphology of the Folktale. 2nd Ed. Trans. Laurence Scott. Austin, Texas: U of Texas P (1968).

Sudhahar, Saatviga, Roberto Franzosi, and Nello Cristianini. “Automating Quantitative Narrative Analysis of News Data.” JMLR: Workshop and Conference Proceedings 17 (2011): 63-71.