Maximising the Power of Semantic Textual Data: CASTEMO Data Collection and the InkVisitor Application

Zbíral, David
david.zbiral@mail.muni.cz
Masaryk University, Czech Republic

Shaw, Robert L. J.
robert.shaw@mail.muni.cz
Masaryk University, Czech Republic

Hampejs, Tomáš
tomas.hampejs@mail.muni.cz
Masaryk University, Czech Republic

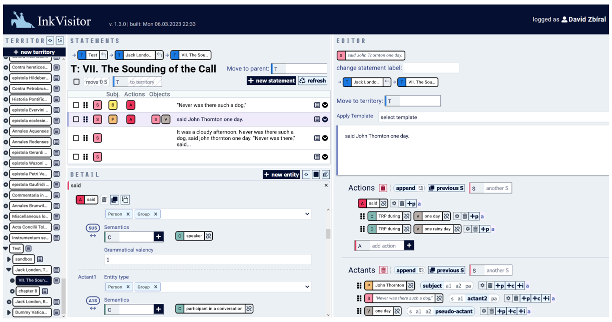
Mertel, Adam
mertel.adam@mail.muni.cz
Masaryk University, Czech Republic

Outline

In this paper, we present Computer-Assisted Semantic Text Modelling (CASTEMO), a novel but now well-developed approach to transformation of textual resources into rich structured data, CASTEMO knowledge graphs, stored in JSON-based document databases (Zbíral et al. 2022). We also present the open-source InkVisitor research environment which assists in CASTEMO data collection workflow (Zbíral et al. 2023). Both the workflow and the environment were developed within the ERC-funded Dissident Networks Project (DISSINET, https://dissinet.cz), but are now made available to use by other researchers and projects.

The CASTEMO data collection approach aims to preserve the rich qualitative texture of texts and at the same time produce structured data suitable for computational analysis. It preserves the contextual embeddedness of knowledge and the natural features of human knowledge, such as conflicting evidence and information given in a non-indicative modality, e.g. questions and conditional sentences. It thus answers a significant challenge in the digital study of texts, where a decision must often be taken to prefer extracting content or analysing discursive features, as well as whether to focus on distant or close reading. With CASTEMO, these levels can be readily interwoven into “scalable reading”.

This presentation introduces the essential data modelling principles of CASTEMO, as well as its use cases and advantages for certain types of study. It also introduces the InkVisitor research environment.



"Never was there such a dog,"

said John Thornton one day.

It was a cloudy afternoon. Never was there such a dog, said john thornton one day. "Never was there," said...

said

said

speaker

1

participant in a conversation

said John Thornton one day.

said John Thornton one day.

said

TRP during — one day

TRP during — one rainy day

John Thornton — subject

"Never was there such a dog," — actant2

one day — pseudo-actant

CASTEMO principles and comparisons

CASTEMO is based on widespread ideas, such as knowledge graphs, semantic data (e.g. Semantic Web), Linked Data, and the syntactic structure of natural-language sentences. It sits within a wider family of statement-based approaches to data collection, and we acknowledge convergent developments within that field, most prominently Roberto Franzosi’s Quantitative Narrative Analysis (Franzosi 2009).

Nevertheless, the DISSINET project has followed its own path in the development of CASTEMO to model textual data in a way that captures not only information from the text, but also the precise discursive context via which this information is conveyed. It preserves the order and syntactic embeddedness of information; the textual embeddedness of information (i.e. who is speaking, to whom, and in what context); the original language and expression; and the distinction between epistemic levels, allowing for editorial annotation beyond the text. At its core, written clauses are captured via statements following the natural syntactic “subject-predicate-object1-object2” structure, which relate entities of two basic kinds: types (Action types and Concepts) and individual entities (Persons, Groups, Events, Objects, Beings, Locations, Resources, Territories/Texts, and Statements themselves). Individual entities are extended through properties (e.g. to capture place, time, and other adverbs and adverbials; adjectives; appositions; etc.), while the whole data structure is interwoven into a knowledge graph through a network of semantic and ontological relations (e.g. synonym, superclass, classification, identification, nominal equivalent of a verb, etc.) (Zbíral et al. 2022). With CASTEMO, it is possible to capture most of what natural language can express, and do so in the form of structured data.

For digital humanists, CASTEMO offers a time-intensive but powerful alternative to (1) text mining (Jockers, 2013; Jockers and Underwood, 2015), which often fails to answer fine-grained questions, and (2) data collection and modelling approach enacted by Computer-Assisted Qualitative Data Analysis Software (CAQDAS) such as ATLAS.ti (Friese 2019). CASTEMO can be considered especially for research projects which are comparatively data-driven (source-driven), and language and narrative perspective matters a lot in them. CASTEMO offers far more ontological depth and flexibility for precisely capturing textual semantics and the contextual features of discourse than existing statement-based data modelling strategies known to us, even those, like Quantitative Narrative Analysis (Franzosi 2009), which were explicitly designed for the purpose.

InkVisitor as a platform for CASTEMO data collection

As an approach, CASTEMO is not tied to any particular software: it can even be pursued using spreadsheets. However, to allow users take full advantage of all the features provided within the CASTEMO data model in a user-friendly manner, the DISSINET team have launched InkVisitor (Zbíral et al. 2023), an open-source web-based research environment for the manual entry of complex structured data from textual resources following CASTEMO principles. InkVisitor serves as a data-entry front-end for RethinkDB JSON-based research databases.

Discussion and conclusions

CASTEMO is of course not the ideal workflow for every digital humanist studying texts. For instance, if one is only focusing on an exact set of hypotheses, is less interested in discursive aspects of the text, and/or is happy to decide on the reliability of data and the impact of confounding factors at the point of collection, then CASTEMO's nuance and tendency towards maximalism may lead to inefficiency. To others, however, CASTEMO and its implementation via InkVisitor will provide a suitable ontology and workflow for modelling the complexity of written language. Digital humanists interested in quantitatively analysing textual information in the precise context of its production and looking at the discursive framing of content will find a highly relevant feature set adapted to their needs.

Bibliography

Franzosi, Roberto (2009): *Quantitative Narrative Analysis*. Thousand Oaks: Sage.

Friese, Susanne (2019): *Qualitative Data Analysis with ATLAS.ti*. Third Edition. Thousand Oaks: Sage.

Jockers, Matthew Lee (2013): *Macroanalysis: Digital Methods and Literary History*. Champaign: University of Illinois Press.

Jockers, Matthew Lee / **Underwood, Ted** (2015): Text-Mining the Humanities, in: Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.): *A New Companion to Digital Humanities*. Second Edition. Chichester: Wiley & Blackwell, 291–306.

Zbíral, David / **Mertel, Adam** / **Hanák, Petr** / **Mertel, Ján** / **Ondrejka, Peter** / **Hampejs, Tomáš** / **Shaw, Robert L. J.** (2023): InkVisitor 1.3. GitHub. Accessed April 28, 2023. https://github.com/DISSINET/InkVisitor/.

Zbíral, David / **Shaw, Robert L. J.** / **Hampejs, Tomáš** / **Mertel, Adam** (2022): Model the source first! Towards Computer-Assisted Semantic Text Modelling and source criticism 2.0. Zenodo. Last modified August 6, 2022. Accessed April 28, 2023. https://doi.org/10.5281/zenodo.6963579.