

# Workshop on Computational Methods in the Humanities 2018 (COMHUM 2018)

---

Université de Lausanne, June 4–5, 2018

The logo for the University of Lausanne (Unil), featuring the word "Unil" in a blue, cursive script.

UNIL | Université de Lausanne

Section des sciences  
du langage  
et de l'information

COMHUM  
2018



## Introduction

It is often said that the digital humanities are “situated at the intersection of computer science and the humanities,” but what does this mean? We believe that the point of using computers in the humanities is not just to automatically analyze larger amounts of data or to accelerate research. We therefore prefer to understand digital humanities as (1) the study of means and methods of constructing formal models in the humanities and (2) as the application of these means and methods for the construction of concrete models in particular humanities disciplines.

The central research questions are thus correspondingly (1) which computational methods are most appropriate for dealing with the particular challenges posed by humanities research, e.g., uncertainty, vagueness, incompleteness, but also with different positions (points of view, values, criteria, perspectives, approaches, readings, etc.)? And (2) how can such computational methods be applied to concrete research questions in the humanities?

The goal of this workshop is to bring together researchers involved with computational approaches in the humanities with the objective of stimulating the research and exchange around innovative, methodologically explicit approaches, to encourage discussion among researchers and developers from different communities, and to help bridging the divide that still exists between the different disciplines involved in this field.

The workshop is organized by members of the Department of Language and Information Sciences (SLI) at the University of Lausanne, with the support of the Faculty of Arts: François Bavaud, Raphaël Ceré, Isaac Pante, Davide Picca, Stéphanie Pichot, Michael Piotrowski, Yannick Rochat, and Aris Xanthos. It underlines the commitment of the Department of Language and Information Sciences (SLI) at the University of Lausanne to the computational dimension of the digital humanities, including formal and mathematical methods.

The organizers would like to thank everybody involved with COMHUM 2018: the members of the program committee for their rigorous and timely reviews, the speakers for their contributions, Maristella Agosti, Bruno Cornelis, and Manfred Thaller for accepting our invitation to Lausanne and sharing their thoughts with us, and—of course—all who are attending the workshop!

**Workshop Web Site** <https://unil.ch/llist/comhum2018/>

## Program Committee

- François Bavaud (UNIL, SLI and IGD)
- Raphaël Ceré (UNIL, IGD)
- Giovanni Colavizza (Turing Institute, London)
- Leonardo Impett (EPFL, Image and Visual Representation Lab)
- Maria Kraxenberger (Max Planck Institute for Empirical Aesthetics)
- Cerstin Mahlow (Bern University of Applied Sciences)
- Barbara McGillivray (Turing Institute, London)
- Isaac Pante (UNIL, SLI)
- Davide Picca (UNIL, SLI)
- Michael Piotrowski (UNIL, SLI) – Chair
- Yannick Rochat (UNIL, SLI)
- Elena Spadini (UNIL, Centre de recherches sur les lettres romandes)
- Aris Xanthos (UNIL, SLI)

## Program

All talks take place in room ANT-2064, building Anthropole, stop *UNIL-Chamberonne* of the m1 metro line.

### Monday, June 4, 2018

---

|             |   |
|-------------|---|
| 10:30       | Registration  |
| 11:00–11:30 | Welcome   |
| 11:30–12:30 | Invited talk: Bruno Cornelis (Vrije Universiteit Brussel): <i>Image Processing for Art Investigation</i>  |
| 12:30–14:00 | Lunch (Caf  teria Unith  que)   |
| 14:00–15:00 | Invited talk: Maristella Agosti (Universit   di Padova): <i>The Confluence in Digital Humanities: the Computer Scientist, the Digital Humanist, and the Final User</i>  |
| 15:00–15:30 | Coffee  |
| 15:30–17:00 | Contributed talks: <ul style="list-style-type: none"><li>• Mats Dahll  f: <i>Clustering Writing Components from Medieval Manuscripts</i></li><li>• Elli Bleeker, Ronald Haentjens Dekker, and Bram Buitendijk: <i>Understanding Texts as Graphs: An Inclusive Approach to Text Modeling</i></li><li>• Jean-Baptiste Camps and Julien Randon-Furling: <i>A Dynamic Model of Manuscript Transmission</i></li><li>• Elena Spadini: <i>Exercises in Modelling: Textual Variants</i></li></ul> |

---

**Dinner** Participants are cordially invited to join us for an informal dinner (self-paid) at 19:00 upstairs at the restaurant *Le Milan*, Boulevard de Grancy 54 (south-west of the main train station, at the intersection of Avenue William-Fraisse and Boulevard de Grancy).



## Tuesday, June 5, 2018

---

08:30 Welcome desk opens

---

09:15–10:00 Contributed talks:

- Christelle Cocco, Raphaël Ceré, and Pierre-Yves Brandt: *Quantification of Drawing Colours in Human Sciences*
  - Mattia Egloff and François Bavaud: *Taking Into Account Semantic Similarities in Correspondence Analysis*
- 

10:00–10:30 Coffee

---

10:30–11:30 Invited talk: Manfred Thaller (emeritus, Universität zu Köln): *Decoding What the Sender Did Not Want to Transmit. Information Technology and Historical Data; or Something*

---

11:30–13:00 Contributed talks:

- Barbara McGillivray, Giovanni Colavizza, and Tobias Blanke: *Towards a Quantitative Research Framework for Historical Disciplines*
  - Franziska Diehr, Maximilian Brodhun, Sven Gronemeyer, Christian Prager, Elisabeth Wagner, Katja Diederichs, and Nikolai Grube: *Modelling Vagueness – A Criteria-Based System for the Qualitative Assessment of Reading Proposals for the Deciphering of Classic Mayan Hieroglyphs*
  - Gary Munnely, Annalina Caputo, and Séamus Lawless: *Linking Historical Sources to Established Knowledge Bases in Order to Inform Entity Linkers in Cultural Heritage*
  - Cristina Vertan: *Supporting Hermeneutic Interpretation of Historical Documents by Computational Methods*
- 

13:00–14:30 Lunch (Cafétéria Géopolis)

14:30–16:00 Contributed talks:

- Susan Leavy, Karen Wade, Gerardine Meaney, and Derek Greene: *Navigating Literary Text Using Word Embeddings and Semantic Lexicons*
  - Jose Luis Losada: *Map Visualization and Quantification of Literary Places in a Spanish Corpus*
  - Thomas Schmidt and Manuel Burghardt: *Toward a Tool for Sentiment Analysis for German Historic Plays*
  - Kyoko Sugisaki, Nicolas Wiedmer, Marcel Naef, Heiko Hausendorf: *Modeling Thematic Structure in Holiday Postcards*
-



## Image Processing for Art Investigation

**Bruno Cornelis**

Vrije Universiteit Brussel (VUB)  
Department of Electronics and Informatics (ETRO)  
Pleinlaan 2, 1050 Brussels, Belgium  
bcorneli@etrovub.be

Advances in digital image acquisition methods and the wide range of imaging modalities currently available have lowered the threshold for museums to digitize their painting collections. This is not only crucial for archival or dissemination purposes but it also enables the digital analysis of the painting through its digital image counterpart. It also set in motion a cross-disciplinary collaboration between image analysis specialists, mathematicians, statisticians and art historians that have the common goal to develop algorithms and build a digital toolbox in support of art scholarship. Computer processing of digital images of paintings has become a fast growing and challenging field of research during the last few years.

This talk will highlight some of the contributions of the international joint initiative on big data, encompassing researchers from the Vrije Universiteit Brussel, Duke University, Ghent University and University College London, to this research domain. Since paintings are complex structures the analysis of all pictorial layers and the support requires a multimodal set of high-resolution image acquisitions.

The developed tools that are used to process these vast amounts of multimodal data are based on dimensionality reduction methods, sparse representations and dictionary learning techniques. These tools are designed to be used in art related matters such as restoration, conservation, art history, material and structure characterization, authentication, dating and even style analysis.

The presented research can broadly be subdivided into three main fields. The first one is the digital enhancement of painting acquisitions in order to assist art experts in their professional assessment of the painting. The second main field of research is the automated detection of cracks within the Ghent Altarpiece, which is meant to help in the delicate matter of the conservation of this exceptional masterpiece but also as guidance during its current campaign of restoration. The last field consists of a set of methods that can be deployed in art forensics. These methods consist of the characterization of canvas, the analysis of multispectral imagery of a painting and even the objective quantification of the style of a particular artist.

# **The Confluence in Digital Humanities: the Computer Scientist, the Digital Humanist, and the Final User**

**Maristella Agosti**

Department of Information Engineering  
University of Padua  
Via Gradenigo 6/a, 35131 Padova, Italy  
maristella.agosti@unipd.it

## **1 The Confluence of Competences in Digital Humanities**

Some of the computational methods and techniques that have been proposed in the diversified area of digital humanities have contributed to the creation and development of different types of information management systems that manage and preserve digital resources of cultural heritage.

Issues related to the conception and implementation of these types of information management systems concern the need to create new models for the automation of processes of representation and processing of specific cultural heritage resources that we want to represent and manage in digital form. Depending on the type of the cultural resources of interest, represented in digital form, and on the operations that we want to have available on them, it can be necessary to envisage a new solution of information management; this new solution can result only from an effective collaboration established between the experts of the specific domain of cultural heritage, the experts of computer science and the users that are going to use the solution. In fact, the experts of the specific domain of cultural heritage – to name just a few of these domains: archives, art history, library science, archeology, linguistics, history – know the characteristics and peculiarities of the resources of the specific domain, while the experts in computer science know methods of digital representation and automatic management so they can imagine new solutions that make available the innovative functions requested by the final users on the digital resources of interest. It is the synergic cooperation among the computer scientist, the digital humanist, and the final user that produces effective new methodological solutions. Once created and formalized the new resource representation and the management model, it is possible to devise a cor-

responding new information management system. Computer science is only one of the necessary cultures to envisage and design new systems.

We could ask us why is it necessary to devise new models and systems? Because we want to consider aspects of reality that are different and more complex than those that were previously addressed. As we increase the diversification and the complexity of the aspects of reality that we want to address and manage, we need new methods and systems capable to deal with and manage them. Bearing in mind the greater complexity of the aspects of reality that we want to address, we need to devise methods to match them and systems able to manage them.

The functions, that a new system provides, are presented to final and professional users through a user interface, that is the external level of the system and that is what the users know of it. The intermediate level implements a method or the methods useful for supporting necessary functions on digital resources of interest. The innermost level serves to represent and manage the data that correspond to the representations of the digital resources of interest together with tools to assist in the storage of data (e.g. indexes and tools for efficient and effective data management).

## **2 Presentation and Critical Analysis of Relevant Case Studies**

Some relevant case studies are presented and critically analyzed to show that when the cooperation between the necessary and different skills is lacking, mistakes can be made that make one lose the possibility of having available innovative digital humanities solutions; when, on the other hand, cooperation is effective, then the solutions that are made contribute to advancing the sector.

## **Decoding What the Sender Did Not Want to Transmit. Information Technology and Historical Data; or Something.**

**Manfred Thaller**

Formerly University at Cologne

In 1978 I was hired by the then Max Planck Institute for History at Göttingen, to support a number of research projects in the field of micro-history by the provision of appropriate IT technologies. The projects planned to use an approach, which was based on “extended family reconstitutions”, even if that precise term was coined only a few years later. A “family reconstitution” traditionally is employed in historical demography. It starts with the marriage registers of a historical community (a village or small city) over at least two hundred years, identifies all brides and groom of the marriages in the birth and death registers, assigns all children in the birth registers to the marriages of their parents and identifies their death entries. To this network of all demographic relationships within a community an “extended family reconstitution” adds all mentions of every person in taxation registers, testaments, local court protocols – and basically every other surviving source.

It was clear from the very beginning, that such a project would take time – and it was impossible to predict at the time of data entry, what part of the source would be needed for analysis the years later. The decision was therefore, to preserve “all information” contained in the source – even if such information was vague, unclear or contradictory. A short impression of the rough solutions provided to come to grips with these properties of the data will be given.

Handling massive (for the time) data bases, quickly leads to the understanding, that while one may in the long run understand, what information is conveyed by a particular chunk of data in the source, one certainly does not immediately. This raises the question, how far the kind of information processing to be supported actually follows the classical paradigm of Shannon, where receivers are able to decode cognitively a message transmitted to them immediately. It gets worse,

when one hopes to apply the usual model of information science, where a common understanding of the context is supposed to allow such a cognitive understanding.

We propose, therefore, to replace the sender-receiver metaphor in information systems dealing with historical data with an observer metaphor, where observers use observed messages to understand the context in which they have been encoded – the understanding of the observed message itself being a welcome side benefit. If one tries to implement this metaphor determinedly and without compromise, one soon discovers, that quite a few technologies of current IT systems become awkward soon – embedded markup, e.g., loses its charms, when a clear-cut separation between the (mainly) static representation of the data and the (always) dynamic interpretation of these data, a.k.a. the information assumed provisionally, is required.

While, as just mentioned, a number of technological assumptions become problematic with this new metaphor, one of the most obvious bundles of problems deals with the inherent vagueness and uncertainty of the information derived from the data.

In order of increasing complexity we will in the second half of our presentation with three example problems. For the sake of generality, we will handle these on the levels of concepts to be supported by programming languages, not on the level of application systems. While many of the approaches discussed owe much to Zadeh’s concept of Fuzzy Sets, we use fuzziness in a broader sense, leaving it uncapitalized therefore.

### **1 Fuzzy numbers**

In many historical sources – or descriptions of their assumed content – numerical data are not

points in a continuum, but ranges, or sets of ranges. This is particularly obvious in the case of temporal information, where the handling of intervals has a long tradition in IT applications for historical sources, therefore, it is a more general problem however. We will briefly describe, how a datatype would look like, which can integrate the handling of such data smoothly into existing programming paradigms. We will use the examples presented earlier from the work of the late seventies and early eighties, to show how mathematical developments since then can overcome limitations of the earlier approaches and where major barriers still exist.

## 2 Fuzzy terms and structures

The greatest successes of computational approaches which are based on alternatives to Boolean logic are visible in the fuzzy control structures of industrial applications described as “computing with words”. The classical examples in this field, as “the truth value of ‘Lausanne is more or less close to Geneva’ is *more or less true*”, seem at first look to be extremely close to the kind of reasoning historians – or, indeed, humanists – frequently employ. We will briefly examine reasons, why that kind of approach has, nevertheless, only very rarely been applied in historical research.

We will concentrate, however, on two broader problems.

(a) As it stands, computing with words is currently almost always employed as a fuzzy pocket in an otherwise crisp information system, where the uncertainty of the decision is hidden from the main stream of the program. This would require a more general concept of a fuzzy term which could be seamlessly integrated into a program in such a way that it coexists with variables of traditional datatypes.

(b) In the semantic technologies, which are making much headway in the Humanities currently, ontologies organize terms in graphs currently. In graphs, where two nodes are either connected or unconnected by a node. Applying the logic of computing with words, we have to consider graphs, where some nodes are connected by edges which connect them with a truth value other than ‘true’ or ‘false’.

## 3 Fuzzy control structures

The thorniest problem seems at first look to be the most simple. To support a logic with any kind of truth values other than ‘true’ or ‘false’ is of course no problem, as long as it is restricted to situations, where a decision about the combined truth value of a decision problem has to be made. As soon, as we intend to employ such a truth value in the parts of a programming language controlling the flow of the program, we encounter quite serious situations, where we briefly describe to what sort of larger framework a solution would require.

## References

- Bernard Favre-Bull: *Information und Zusammenhang. Informationsfluß in Prozessen der Wahrnehmung, des Denkens und der Kommunikation*, Springer, 2001.
- Sifeng Liu and Yi Lin: *Grey Systems. Theory and Practical Applications*, London, 2011.
- Claude E. Shannon and Warren Weaver: *The Mathematical Theory of Communication*, 1949.
- Lotfi A. Zadeh and Janusz Kacprzyk (Eds.): *Computing with Words in Information / Intelligent Systems I and II* (= Studies in Fuzziness and Soft Computing 33 and 34 (1999)).
- Lotfi A. Zadeh: “Toward a Generalized Theory of Uncertainty (GTU) – an outline”, *Information Sciences* 172 (2005), 1–40.

# Clustering Writing Components from Medieval Manuscripts

Mats Dahllöf

Department of Linguistics and Philology

Uppsala University

mats.dahllöf@lingfil.uu.se

## 1 Introduction

The present work explores unsupervised extraction and clustering of writing components from historical manuscripts. The primary purpose is to locate letters and to group them into classes capturing graphemic equivalence. The method will also find ligatures, scribal abbreviations, parts of letters, and letter sequences. The output will provide cheap, but partial, manuscript transcriptions in combination with human annotation of the clusters. The system can be used to curate data for further training of handwritten text recognition systems or as a tool for presenting manuscript data for qualitative palaeographic analysis. Related proposals include the work of Rath and Manmatha (2007), who use clustering of words from historical manuscripts (18th C.) as a means for obtaining labelled data for word spotting. Vuurpijl and Schomaker (1997) use clustering to find allography in on-line handwriting. Another application is proposed by Stutzmann (2016), who is interested in the use of clustering for the categorization of medieval script types.

## 2 Component Extraction and Clustering

The first main module of the present system performs component extraction. It is based on binarization (ink-background separation), using a version of the Otsu (1979) algorithm, and connected component labelling. The pieces of connected writing identified in this way are then segmented further, at the positions where the pixel column sum of ink is thinnest, but not too thick.

In order to adapt the component extraction to the actual size and scale of the writing, the processing is guided by the typical stroke width,  $w_s$  (for the manuscript images being analysed). The system estimates  $w_s$  by determining the most common width of sequences of continuous horizon-

tal foreground (ink) pixels separated by at least two pixels of background. After that, the system rescales the images to make sure all manuscripts are processed at roughly the same writing-relative resolution. In the experiments we discuss here,  $w_s = 7$  pixels.

The component extraction process is guided by five parameters,  $(t_i, w_{mn}, w_{mx}, h_{mn}, h_{mx})$ . Connected components whose width and height are in the intervals  $[w_{mn}, w_{mx}]$  and  $[h_{mn}, h_{mx}]$ , respectively, are extracted, while those wider than  $w_{mx}$  are fed to a segmentation module. Loosely speaking,  $t_i$  is the thickest amount of ink that allows a cut to be made. The segmentation process operates on the column sum of foreground (ink)  $I(x)$ , as computed with reference to the bounding box. It scans the component pixel by pixel,  $x_L$  being the current position. When  $x_L = 0$  or  $I(x_L) \leq t_i$ , the system looks for a  $x_R \in [x_L + w_{mn}, x_L + w_{mx}]$  where  $I(x_R) \leq t_i$  and  $I(x_R)$  is the smallest value. If it is not unique, the leftmost (smallest)  $x_R$  is preferred. A component spanning  $x_L$  to  $x_R$  is then proposed, and scanning resumes with  $x_L = x_R$ . If no  $x_R$  meets the condition, scanning resumes with  $x_L = x_L + 1$ . If the height of a proposed component spanning  $x_L$  to  $x_R$  is in the interval  $[h_{mn}, h_{mx}]$ , it is added to the set of extracted components. In the experiments reported here, we used  $(t_i, w_{mn}, w_{mx}, h_{mn}, h_{mx}) = (1.0w_s, 3.0w_s, 8.0w_s, 3.0w_s, 15.0w_s)$ . These parameters represent a heuristic assumption about the appearance of the handwriting. We have tailored them to medieval book hands, aiming for a “wide-spectrum” retrieval of letter-size elements, but more or less excluding the letter ⟨i⟩ and other “minims”.

The components are characterized by a feature vector, which quantifies the distribution of foreground pixels as captured by a grid of  $11 \times 11$  equal subrectangles over the bounding box. Each value

Table 1: The manuscript page sequences used as data. The UUB images cover spreads.

|   |
|---|
| Gen. 1. Schaffhausen Stadtbibliothek, pp. 6 ff. [“Irische Halbunziale”, 7th/8th C.]<br><a href="http://dx.doi.org/10.5076/e-codices-sbs-0001">http://dx.doi.org/10.5076/e-codices-sbs-0001</a>  |
| CS 60. Cod. Sang. 60, St. Gallen, Stiftsbibliothek, pp. 6 ff. [“irischer Schrift”, 8th C.]<br><a href="http://dx.doi.org/10.5076/e-codices-csg-0060">http://dx.doi.org/10.5076/e-codices-csg-0060</a>   |
| CS 557. Cod. Sang. 557, St. Gallen, Stiftsbibliothek, pp. 13 ff. [“Qualifizierte St. Galler Carolina”, late 9th C.]<br><a href="http://dx.doi.org/10.5076/e-codices-csg-0557">http://dx.doi.org/10.5076/e-codices-csg-0557</a>                                      |
| CS 564. Cod. Sang. 564, St. Gallen, Stiftsbibliothek, pp. 16 ff. [“Grosse, sorgfältige Spätcarolina”, late 12th C.]<br><a href="http://dx.doi.org/10.5076/e-codices-csg-0564">http://dx.doi.org/10.5076/e-codices-csg-0564</a>                                      |
| B 59. National Library of Sweden, pp. 3 recto ff. (image 9 ff.) [Textualis, late 13th C.]<br><a href="https://data.kb.se/datasets/2015/01/fornsvenska/B_59.002611384">https://data.kb.se/datasets/2015/01/fornsvenska/B_59.002611384</a>                            |
| B 10. Uppsala University Library, pp. 24 verso ff. [Textualis, 1350–1399.]<br><a href="http://urn.kb.se/resolve?urn=urn:nbn:se:alvin:portal:record-90664">http://urn.kb.se/resolve?urn=urn:nbn:se:alvin:portal:record-90664</a>                                     |
| C 61(a). C 61, Uppsala University Library, pp. 138 ff. (spread image 74 ff.) [Cursiva recentior, late 15th C.]<br><a href="http://urn.kb.se/resolve?urn=urn:nbn:se:alvin:portal:record-55762">http://urn.kb.se/resolve?urn=urn:nbn:se:alvin:portal:record-55762</a> |
| C 61(b). The same codex, pp. 540 ff. (spread image 275 ff.) [Cursiva recentior, late 15th C. A different hand.]   |

is the ratio of the number of foreground pixels to the size of the subrectangle region. The clustering relies on Euclidean distance operating on these vectors.

The system uses the density-based DBSCAN algorithm (Ester et al., 1996) to obtain a “core clustering”. It was proposed for applications, like the present one, where a fair amount of noise data points are present. DBSCAN is guided by two parameters:  $Eps$ , the largest distance between two points which are to be counted as neighbours, and  $minPts$ , the minimal number of neighbouring points required for the formation of a same-cluster dense region. We estimate  $Eps$  from another value,  $p_{Eps}$ , which is the probability that two randomly selected image components be at most  $Eps$  distant from each other (for the manuscript being processed). We used  $minPts = 11$  and  $p_{Eps} = 0.0007$  as a baseline setting. After the DBSCAN step, small clusters (size  $< 40$  here) are removed. In the last step of the clustering, clusters are extended in a “nearest neighbour” (to centroids) classification step, which assigns some of the not clustered components to the remaining core clusters.

### 3 Evaluation

In the evaluation, we applied the system to eight 7th–15th C. book manuscripts (see Table 1), representing four scripts, Irish and Carolin-

Table 2: Precision and recall (both in %) for the baseline and  $p_{Eps} = 0.0014$  setups for three categories and four manuscripts.

| Manus.  | Baseline            |                     |                     | $p_{Eps} = 0.0014$  |                     |                     |
|---------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
|         | $\langle e \rangle$ | $\langle m \rangle$ | $\langle o \rangle$ | $\langle e \rangle$ | $\langle m \rangle$ | $\langle o \rangle$ |
| Gen. 1  | 100 14              | 100 54              | 100 58              | 100 46              | 100 68              | 100 83              |
| CS 557  | – 0                 | 100 44              | 100 61              | 100 8               | 100 71              | 100 76              |
| CS 564  | 100 4               | 65 58               | 100 17              | 100 7               | 57 61               | 100 36              |
| C 61(b) | 100 28              | 97 63               | 100 28              | 100 31              | 96 65               | 29 36               |



Figure 1: Clusters (excerpts) from CS 60, representing  $\langle d \rangle$ ,  $\langle \& \rangle$  ( $\langle et \rangle$  ligature),  $\langle m \rangle$ , and  $\langle t \rangle$ .

gian minuscule, textualis, and cursiva, each in two instances. The data are high-resolution images (JPEG or TIFF) published open access by the libraries. From each sequence of pages, the system extracted exactly 20 000 components. Between 25 and 44 images had to be read. (All experiments used the same extraction settings.) We report which clusters were established, their precision, and (for a few cases) their recall. Output for all combinations of data and set-ups discussed here are available at URL <http://stp.lingfil.uu.se/~matsd/ch2018/>.

The baseline settings (specified above) led the system to assign between 5000 and 14 800 components (of the 20 000 ones for each manuscript) to between 16 and 27 clusters. Between 5 and 15 of the clusters corresponded to letter categories with a precision  $> 60\%$ . The majority had a precision  $> 98\%$ . Letters like  $\langle a \rangle$ ,  $\langle d \rangle$ ,  $\langle m \rangle$ , and  $\langle q \rangle$  had a strong tendency to appear. In a few cases, the system generated two different clusters for the same letter. For the Irish script of Gen. 1 it correctly distinguished two allographs of  $\langle d \rangle$  (289 and 251 instances) with 100% precision. The number of clusters for ligatures and two-letter sequences were higher for textualis and cursiva, in which letters typically are connected.

The manuscript CS 60, to take one specific example, yielded 16 clusters containing 14 556 components. These had a precision  $> 98\%$ :  $\langle c \rangle$ : 601,  $\langle d \rangle$ : 593,  $\langle e_1 \rangle$ : 317,  $\langle e_2 \rangle$ : 181 ( $\langle e_2 \rangle$  split ligatures),  $\langle m \rangle$ : 734,  $\langle o \rangle$ : 851,  $\langle q \rangle$ : 371,  $\langle s \rangle$ : 755,  $\langle t \rangle$ : 361,



and ⟨&⟩: 248. There were also clusters of lower precision (> 60%): ⟨a⟩: 1138 and ⟨er⟩: 62. Four clusters were mixtures of several categories (sizes: 4300, 3274, 718, and 52).

We saw two kinds of outcome: For some manuscripts (e.g. CS 557), components were generally assigned to high-precision one-category clusters. In the other cases (e.g. CS 60), large “useless” clusters were established, along with categorically precise ones. This suggests that more generous clustering settings should be used in the former cases, while the opposite kinds of outcome invite us to explore more reluctant settings. We made experiments in which one parameter of the baseline setting was modified. By setting  $p_{Eps} = 0.0014$  (i.e. doubling it), we made the clustering more generous. As can be expected, the clusters became larger and less pure, but the system also discovered new useful clusters. To some extent categories discerned in the baseline setup merged. Two more restrictive modifications were also tried,  $p_{Eps} = 0.00035$  and  $minPts = 22$ . Both settings made the clusters fewer and smaller across the board. They also caused some letter categories to detach themselves in precise clusters, e.g. ⟨h⟩ and ⟨k⟩ from B 59.

We estimated recall for a few cases, categories, and two setups, by annotating the three or four first pages of the sequences for four manuscripts as regards three letter categories, see Table 2. In comparison with the baseline setup, the (more generous)  $p_{Eps} = 0.0014$  setting leads to a clear increase in recall in most cases. There is a loss in precision in two cases. For C 61(b) we get a cluster merging the ⟨a⟩ and ⟨o⟩ categories (separated in the baseline output) with an ⟨o⟩-precision of 29%.

## 4 Conclusions

The current study has shown that simple component extraction and clustering in combination with limited human intervention can be used to produce partial transcriptions of medieval manuscripts in a range of styles. The pipeline provides a low-cost method for initial annotation, which is potentially useful in many contexts of digital palaeography. The method could, for instance, be used to extract features for manuscript classification, e.g. dating and scribe attribution, as well as to present data for manual palaeographic analysis.

The feature model seems to work quite well for the styles studied here, because letter distinctions

generally correspond to marked contrasts in how ink is distributed in the bounding box. Admittedly, we have only studied fairly regular book styles. The  $11 \times 11$  “resolution” is reasonable for letters, but will blur larger components, such as sequences of several letters. The model is insensitive to the absolute or relative size of the components. This is an advantage when there is linguistically insignificant size variation. A system like this one could probably benefit from also looking at the contexts in which the components occur. The current system only “sees” the foreground components as framed by the bounding box.

The basic modules are simple and invite exploration of more sophisticated mechanisms. Our experiments with different parameter settings suggest that a system like this should be tuned separately for different categories, rather than rely on one-pass application of algorithms partitioning the same components into non-overlapping clusters.

## Acknowledgments

The author is grateful for the support of two projects, funded by the Swedish Research Council (Vetenskapsrådet, Dnr 2012-5743) and Riksbankens Jubileumsfond (Dnr NHS14-2068:1), and led by Anders Brun and Lasse Mårtensson, respectively.

## References

- Martin Ester, Hans-Peter Kriegel, Jiirg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. AAAI Press, pages 226–231.
- Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms 9:62–66.
- Tony M. Rath and R. Manmatha. 2007. Word spotting for historical documents. *International Journal of Document Analysis and Recognition (IJ DAR)* 9:139–152.
- Dominique Stutzmann. 2016. Clustering of medieval scripts through computer image analysis: Towards an evaluation protocol. *Digital Medievalist* 10.
- Louis Vuurpijl and Lambert Schomaker. 1997. Finding structure in diversity: a hierarchical clustering method for the categorization of allographs in handwriting. In *Proceedings of the Fourth International Conference on Document Analysis and Recognition*. IEEE, pages 387–393.

# 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<sup>1</sup> 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<sup>2</sup>, 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<sup>3</sup> and structural variation (paragraphs, chapters, etc.) are generally ignored even though they are valuable aspects of a text's development.<sup>4</sup> *HyperCollate*, in contrast, uses the

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

<sup>2</sup>For a more extensive discussion of HyperCollate, see (Bleeker et al., 2018)

<sup>3</sup>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

<sup>4</sup>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/>

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.)?"<sup>5</sup> 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.<sup>6</sup> 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

[juxta-commons/](http://www.juxta-commons.org/)), these elements play no (analytical) role for the alignment of the tokens.

<sup>5</sup>See <http://wp.unil.ch/llist/event/comhum2018/>

<sup>6</sup>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.

# A Dynamic Model of Manuscript Transmission

**Jean-Baptiste Camps**

Centre Jean-Mabillon

École nationale des chartes

Université Paris Sciences & Lettres

jean-baptiste.camps@chartes.psl.eu

**Julien Randon-Furling**

SAMM (EA4543)

Université Panthéon Sorbonne

j.randon-furling@cantab.net

## Abstract

Trees and tree-like graphs are of wide application in the natural and human sciences, and are particularly apt to represent genealogical information. Inherited from the XIX<sup>th</sup> century scientific method, from Darwin’s theories to comparative philology and textual criticism (Timpanaro, 2003), they are still widely used to reflect progressive divergences in the data (Moretti, 2005), be it genes, linguistic features or textual variants.

The predominance of root bifurcation (bifidity) in text genealogical trees (*stemmata*) has been discussed at least since Bédier (1928). It is an important issue in philology, because it does not allow to use a simple majority rule when reconstructing the text of the archetype. It is also of broader interest for manuscript and text studies.

Estimating manuscripts loss rates is another long standing question, that has occasionally been envisioned in terms of population dynamics (Cisne, 2005).

In this study, we examine the question of node furcation through a dynamic model of manuscript transmission, taking into account both “birth” and “death” rates. We use stochastic modelling to represent a variety of processes: diffusion of texts, reproduction of manuscripts, and extinction of branches.

## 1 Introduction

During Antiquity and the Middle Ages, manual copying was the only way to disseminate a written text, resulting in the introduction of variants and

errors. Since the XIX<sup>th</sup> century, philologists have used common errors to reconstruct the genealogy of surviving copies (witnesses), a field known today as stemmatology. It is customary to represent text genealogies as trees, where nodes stand both for the extant witnesses or the intermediary lost manuscripts that can be inferred from the examination of common errors.

In a famous work, Bédier (1928) shed light on the over-representation of root bifurcation, a phenomenon he termed “*bifidité*” (bifidity). Bédier made it an argument against the use of the common error methods by his fellow philologists as a mean to reconstruct the text, because root bifurcation prevents from using majority rule in the evaluation of variants. Knowing whether this ubiquitous bifidity is an artefact of the practice of the common errors method or a legitimate phenomenon resulting from text transmission has been a subject of heated debates ever since.

Even though the question of bifidity has been studied, among other approaches, by estimating the proportion of root bifurcation for a given number of witnesses, many previous studies have been based on the assumption that all configurations are equiprobable (Maas, 1937; Castellani, 1957; Hoenen et al., 2017).

Independently, some studies have examined manuscript transmission and loss in the light of population dynamics (Cisne, 2005), or tried to assess loss rates using, for instance, regression (Buringh, 2010).

Some works have even linked bifidity to the population dynamics of manuscript transmission. Tentative modelling have been suggested (Canettieri et al., 2008), sometimes drawing a parallel between extinction of manuscript branches and of biological genres, as well as with the famous “gambler’s ruin” problem (Raup, 1992), or the process known as “genetic drift”.



The work of Weitzman (1982, 1987) pioneered the use of birth and death models and computer simulations in the field of manuscript transmission. Some recent studies have explored further the relation between bifidity and manuscript loss, through the use of probabilistic models and simulation of textual genealogies (Guidi et al., 2004; Hoenen, 2016).

## 2 Stochastic modelling of genealogies

Mathematical models of manuscript genealogies have thus so far often been close, if not identical, to models appearing in phylogenetics and other lineage studies in biology. They often consist in trees satisfying a number of constraints, and algorithms have been conceived in order to reconstruct genealogies, for instance using dissimilarity coefficients between manuscripts (Buneman, 1971). Algorithms based on textual criticism principles or on compression based methods have also shown to perform quite well (Roos and Heikkilä, 2009).

Like networks, trees are a specific breed of the mathematical objects called *graphs*. There exists a substantial body of mathematical literature on trees, within the broader field of combinatorics – indeed, trees and their properties are often amenable to exact counting. A number of such exact results have found direct applications in stemmatology (Hoenen et al., 2017).

We favour here an approach akin to stochastic modelling, not altogether absent from previous studies in stemmatology (Weitzman, 1987). More specifically, we introduce a model that belongs to a group of stochastic processes called *birth and death* processes. These are processes whereby individual agents appear (“are born”) and disappear (“die”) at certain rates, engendering offsprings in the meanwhile.

As special cases of Markov chains (Norris, 1998), birth and death processes have aroused much interest *per se* among mathematicians and probabilists. Such a general process may be adapted to describe the birth and death of manuscripts. This not only allows one to simulate *in silico* possible genealogies (Hoenen, 2016), but also to examine long-standing questions such as the over-representation of bifid trees (Bédier, 1928). Specifically, we explore the variety of manuscript tree patterns that emerge when key parameters are varied across their ranges, viz. *fecundity* and *decimation* rates. In terms of a birth-

and-death process, the former corresponds to the intensity of the offspring distribution and the latter to the mortality rate – how do they influence the proportion of bifurcations in a tree? are there fine-tuned combinations of parameter values for which one observes a greater number of bifurcations?

We aim at finding and classifying tree patterns obtained for all possible combinations of parameter values, thus effectively producing what physicists call a phase diagram (Yeomans, 1992). We show in this paper how such a diagram may be obtained by computational methods as well as by analytical methods. We also compare exemplar genealogies produced through our numerical simulations to real-data genealogies.

## References

- Joseph Bédier. 1928. La tradition manuscrite du lai de l’ombre. réflexions sur l’art d’éditer les anciens textes (premier article). *Romania* 54(214):161–196.
- Peter Buneman. 1971. The recovery of trees from measures of dissimilarity. *Mathematics in the archaeological and historical sciences* .
- Eltjo Buringh. 2010. *Medieval Manuscript Production in the Latin West*. Brill. <https://doi.org/10.1163/9789047428640>.
- Paolo Canettieri, Vittorio Loreto, Marta Rovetta, and Giovanna Santini. 2008. *Philology and Information Theory*. *Cognitive Philology* 1. <http://ojs.uniroma1.it/index.php/cogphil/article/view/8816>.
- Arrigo Castellani. 1957. *Bédier avait-il raison?: La méthode de Lachmann dans les éditions de textes du Moyen Age. Leçon inaugurale donnée à l’université de Fribourg le 2 juin 1954*. Number 20 in Discours universitaires, Nouvelle série = Freiburger Universitätsreden, Neue Folge. Éditions Universitaires, Fribourg.
- John L. Cisne. 2005. How Science Survived: Medieval Manuscripts’ “Demography” and Classic Texts’ Extinction. *Science* 307(5713):1305–1307. <https://doi.org/10.1126/science.1104718>.
- Vincenzo Guidi, Paolo Trovato, et al. 2004. Sugli stemmi bipartiti. decimazione, asimmetria e calcolo delle probabilità. *Filologia italiana* 1:9–48.
- Armin Hoenen. 2016. *Silva Portentosissima – Computer-Assisted Reflections on Bifurcativity in Stemmas*. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University. DH 2016, pages 557–560. <http://dh2016.adho.org/abstracts/311>.

- Armin Hoenen, Steffen Eger, and Ralf Gehrke. 2017. How many stemmata with root degree  $k$ ? In *Proceedings of the 15th Meeting on the Mathematics of Language*. pages 11–21.
- Paul Maas. 1937. Leitfehler und stemmatische Typen. *Byzantinische Zeitschrift* 37(2):289–294. <https://doi.org/10.1515/byzs.1937.37.2.289>.
- Franco Moretti. 2005. *Graphs, maps, trees: abstract models for a literary history*. Verso.
- James R Norris. 1998. *Markov chains*. Cambridge university press.
- David M Raup. 1992. *Extinction: bad genes or bad luck?*. WW Norton & Company.
- Teemu Roos and Tuomas Heikkilä. 2009. Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets. *Literary and Linguistic Computing* 24(4):417–433.
- Sebastiano Timpanaro. 2003. *La genesi del metodo del Lachmann*. UTET Libreria, Torino, 4 edition.
- Michael P Weitzman. 1982. Computer simulation of the development of manuscript traditions. *Bulletin of the Association for Literary and Linguistic Computing* 10(2):55–59.
- Michael P Weitzman. 1987. The evolution of manuscript traditions. *Journal of the Royal Statistical Society. Series A (General)* pages 287–308.
- Julia M Yeomans. 1992. *Statistical mechanics of phase transitions*. Clarendon Press.

# Exercises in modelling: Textual variants

**Elena Spadini**

Centre de recherche sur les lettres romandes

Université de Lausanne

Elena.spadini@unil.ch

## Abstract

This paper presents a model for annotating textual variants. The annotations made can be queried in order to analyse and find patterns in textual variation. The model is flexible, allowing scholars to set the boundaries of the readings, to nest or concatenate variation sites, and to annotate each pair of readings; furthermore, it organizes the characteristics of the variants in features of the readings and features of the variation. After presenting the conceptual model and its applications in a number of case studies, the paper introduces two implementations in logical models, namely a relational database schema and an OWL 2 ontology. While the scope of this paper is a specific issue in textual criticism, its broader focus is on how data are structured and visualized in digital scholarly editing.

## 1 Introduction

Textual variation is a central object of study for textual criticism, *philologie*, scholarly editing. The variation takes place when there are competing readings of a portion of a work. It might occur in different locations and its nature is variegated. Finding patterns in the moving universe of textual variation is one of the scholar's goal. Patterns indicate direction of changes, tracing precious paths for exploring the work and its *mouvance*; they help making sense out of a shapeless set of variants and shed light on textual dynamics.

This paper introduces a model for annotating textual variants. Querying the annotations made allows to find patterns in textual variations.

The practice of “modelling” is used in this paper taking into account the studies of McCarty (2004), Eide (2014), Ciula and Eide (2017); in

particular, we refer to data modelling (Flanders and Jannidis, 2015 and 2016), applied to textual criticism (Unsworth, 2002; Pierazzo, 2015).

## 2 Conceptual model

A reading is the atomic unit of the model. The model describes two main aspects of the elements involved in the variation: the features of the single reading and those of the variation (Figure 1).

For each single reading, two general features must be set: the witness to which the reading belongs, and the location of the reading in the witness. Each single reading can also be annotated using customized categories, which might vary greatly; for example, one might want to record the writing tool in use.

The features of the variation express what kind of difference exists between the competing readings. Two categories are used to record the general features of the variation: the category of change and, in the case of substitution, the linguistic aspect involved. Specific categories can also be used to describe precise features of the variation, such as the direction of change or the nature of the intervention.

When a variation site includes more than two readings, annotations are created for each pair of readings, in order to obtain the maximum of expressiveness. The variation sites can be nested and concatenated.

The model outlined here allows:

- to distinguish between the features of the reading and those of the variation between the readings;
- to append more than one feature to each reading and variation;
- not to set a base witness to orient the variation;



- to annotate each pair of readings or a combination of them for each variation site;
- to nest and concatenate variation sites.

namely in articulating relationships. XML, on the contrary, is less suitable for conveying the information gathered using the model because of its overlapping structure, even if XML solutions can eventually be implemented.

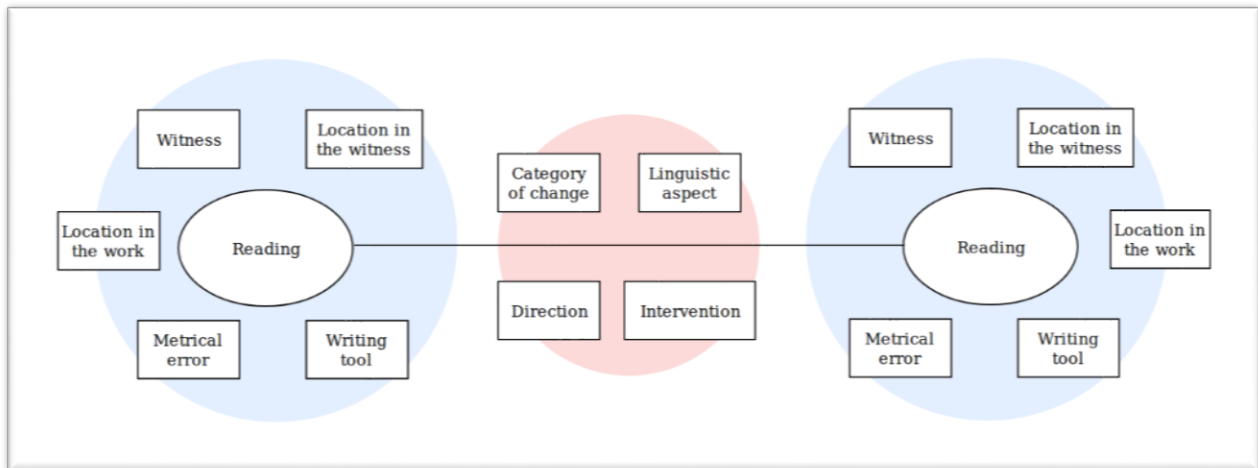


Figure 1. General and specific features of the readings and of the variation.

### 3 Case studies

The model has been implemented in Tempestini-Spadini (2015-2018). In this paper, use cases from Old French *pastourelles* and Giacomo Leopardi's lyrics are presented.

### 4 Logical model

The conceptual model is transformed into a relational database schema and an OWL ontology. An XML solution is also taken into account, but will not be presented in details.

### 5 Conclusions

This article presents a model for annotating textual variants. Once the annotations are made and conveniently stored, they can be queried, in order to find patterns and analyse the *mouvance* of the work. Possible queries depend on the categories of the readings and of the variation in use.

Adopting the model implies a cumbersome work. On the other side, it provides detailed and organized information, which is fundamental for certain projects of scholarly editing. Annotating variations following the model could benefit from a dedicated GUI. Also, some of the categories might be identified automatically (category of change, linguistic aspects).

The implementation in different data structures proves that the relational DB schema and the OWL ontology have the same expressiveness,

Ongoing experiences prove that there is an interest in the digital scholarly editing community to explore solutions others than the tree formalism of XML. In particular, the graph structure is emerging, as a conceptual model to be implemented in different ways [Haentjens Dekker and Birnbaum, 2017; Eide, 2014; Ciotti and Tomasi, 2016; Tomasi, Daquino and Giovannetti, 2018; <<http://knora.org>>]. The adoption of graphs raises a number of technical and theoretical challenges. Among the technical ones, there might be the need to integrate the information stored in graphs within the XML (or HTML) representation of the text; stand-off solutions can peer out here, for overcoming the limitation of XML and for filling the gap with other data structures. Among the theoretical challenges, there is the possibility to call into question the way texts are employed and consumed, which is not unrelated to the way they are visualized. This means, for instance, that scholarly editing can produce various outputs: diplomatic or critical texts; but also SVG objects and, more in general, graphics and dynamic visualizations, which might represent some of the features of the texts better than typographical devices reproduced by HTML [Andrews and van Zundert, 2016; Cummings, Hadley and Noble, 2017]. The term visualization recalls that what is represented is data, and not only words or sentences. In this scenario, it is easier to take advantage of data structured in graphs or in relational tables.

## References

- Andrews, Tara L. 2016. 'Analysis of Variation Significance in Artificial Traditions Using Stemmaweb'. *Digital Scholarship in the Humanities* 31 (3): 523–39. <https://doi.org/10.1093/llc/fqu072>.
- Brandoli, Cristina. 2007. 'Due Canoni a Confronto: I Luoghi Di Barbi e Lo Scrutinio Di Petrocchi'. In *Nuove Prospettive Sulla Tradizione Della Commedia. Una Guida Filologico Linguistica Al Poema Dantesco*, edited by Paolo Trovato, 99–214. Firenze: Cesati.
- Camps, Jean-Baptiste. n.d. 'Louis Havet, Cesare Segre, critique verbale et diasystème'. Billet. *Sacré Gr@@l* (blog). Accessed 8 March 2018. <https://graal.hypotheses.org/550>.
- Ciotti, Fabio, and Francesca Tomasi. 2016. 'Formal Ontologies, Linked Data, and TEI Semantics'. *Journal of the Text Encoding Initiative*, no. Issue 9 (September). <https://doi.org/10.4000/jtei.1480>.
- Ciula, Arianna, and Øyvind Eide. 2017. 'Modelling in Digital Humanities: Signs in Context'. *Digital Scholarship in the Humanities* 32 (suppl\_1): i33–46. <https://doi.org/10.1093/llc/fqw045>.
- Colwell, E. C., and E. W. Tune. 1964. 'Variant Readings: Classification and Use'. *Journal of Biblical Literature* 83 (3): 253–61. <https://doi.org/10.2307/3264283>.
- Cummings, James, Martin Hadley, and Howard Noble. 2017. 'It Has Moving Parts! Interactive Visualisations in Digital Publications'. In *DiXiT Workshop. The Educational and Social Impact of Digital Scholarly Editions*. <http://dixit.uni-koeln.de/programme/materials/#aiucd2017>.
- Eide, Øyvind. 2014. 'Ontologies, Data Modeling, and TEI'. *Journal of the Text Encoding Initiative*, no. Issue 8-PREVIEW (December). <https://jtei.revues.org/1191>.
- Haentjens Dekker, Ronald, and David J. 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. <https://doi.org/10.4242/balisagevol19.dekker01>.
- Italia, Paola. 2010. *Che cosa è la filologia d'autore*. Roma: Carocci.
- Italia, Paola, Fabio Vitali, and Angelo Di Iorio. 2015. 'Variants and Versioning Between Textual Bibliography and Computer Science'. In *Proceedings of the Third AIUCD Annual Conference on Humanities and Their Methods in the Digital Ecosystem*, 2:1–2:5. AIUCD '14. New York, NY, USA: ACM. <https://doi.org/10.1145/2802612.2802614>.
- McCarty, Willard. 2004. 'Modeling: A Study in Words and Meanings'. In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth. Oxford: Blackwell. <http://www.digitalhumanities.org/companion/>.
- Micha, Alexandre. 1979-1983. *Lancelot: roman en prose du XIIIe siècle*. Genève: Droz.
- Pierazzo, Elena. 2015. *Digital Scholarly Editing: Theories, Models and Methods*. Ashgate.
- Riviere, Jean-Claude. 1974. *Pastourelles*. Genève: Droz.
- Schauweker, Yela. 2013. 'Variantes « significatives » et Variantes « récurrentes ». Repenser l'appareil Critique'. In *Actes Du XXVIIe Congrès International de Linguistique et de Philologie Romanes (Nancy, 15-20 Juillet 2013)*. ATILF. <http://www.atilf.fr/cilpr2013/actes/section-13/CILPR-2013-13-Schauwecker.pdf>.
- Spadini, Elena. 2017. 'The Role of the Base Manuscript in the Collation of Medieval Texts'. In *Advances in Digital Scholarly Editing. Papers Presented at the DiXiT Conferences in The Hague, Cologne, and Antwerp*, edited by Peter Boot and alii, 345–50. Leiden: Sidestone Press.
- Stussi, Alfredo. 2011. *Introduzione Agli Studi Di Filologia Italiana*. Bologna: Il Mulino.
- Tempestini, Sonia, and Elena Spadini. 2015-2018. 'La "Commedia" Di Boccaccio. Un Apparato in Movimento'. <http://boccacciocommedia.it>.
- Tomasi, Francesca, Fr, Marilena Daquino, and Francesca Giovannetti. 2018. 'Linked Data Ed Edizioni Scientifiche Digitali. Esperimenti Di Trasformazione Di Un Quaderno Di Appunti'. In *7th AIUCD Conference. Cultural Heritage in the Digital Age*. Bari. <http://www.aiucd2018.uniba.it>.
- Unsworth, John. 2002. 'What Is Humanities Computing and What Is Not?' *Jahrbuch Für Computerphilologie* 4: 71–84.
- Van Hulle, Dirk, and Vincent Neyt, eds. 2011. *The Beckett Digital Manuscript Project*. Brussels: University Press Antwerp (ASP/UPA).
- Vanhoutte, Edward. 2007. 'Traditional Editorial Standards and the Digital Edition'. In *Learned Love. Proceedings of the Emblem Project Utrecht Conference on Dutch Love Emblems and the Internet (November 2006)*, edited by Els Stronks and Peter Boot, 157–74. The Hague: DANS Symposium Publications.
- Zundert, Joris van, and Tara Andrews. 2014. *Apparatus vs. Graph – an Interface as Scholarly Argument*. Interface Critique. <https://vimeo.com/114242362>.

# Colours quantification of children’s drawings

**Christelle Cocco**

Institute for Social Sciences of Religions (ISSR)  
University of Lausanne  
Christelle.Cocco@unil.ch

**Raphaël Céré**

Institute of Geography and Sustainability  
University of Lausanne  
Raphael.Cere@unil.ch

**Pierre-Yves Brandt**

Institute for Social Sciences of Religions (ISSR)  
University of Lausanne  
Pierre-Yves.Brandt@unil.ch

## Abstract

Children’s drawings reveal that developing colour retrieval technics must consider various aspects of what is a colour from the flow of energy (for physics) to discrete information (for humanities and human sciences). Hence, the proposed method deals with color variety measures and colours quantifications upon the drawing to respond at specific queries e.g. “Which colours appear?”, or “How much yellow appears?”.

## Introduction

In computer vision, there are well developed techniques to analyse natural images (e.g. pictures or videos); whereas the status of the image in humanities or in human sciences depends more on the perception of the image, not necessarily natural (e.g. paintings or drawings), and needs specific techniques according to precise research questions.

Perception of an image by a human includes colour recognition which is the central point of this contribution using a children’s drawings dataset. The particularity of this kind of image depends on the drawer’s colour choices. It can be intentional according to his/her own perception and his/her human singularity or constraints according to colour availability. Subsequently, the resulting object analysis is technically constrained: the final colours are altered by the digitalisation encoding, as well as the quantification of colours.

Indeed, two aspects of the quantification of a drawing’s colours are developed:

1. the gap between the human perception (continuous colours) and the colour categorisation (discrete colours) (see e.g. [Alejandro and Akbarinia, 2016](#); [Benavente et al., 2008](#); [Berlin and Kay, 1969](#)),

2. the gap between the human categories of colours (ideally universal) and the exact colour of each pixel in the image (continuous colours) ([Khan et al., 2012, 2013](#)).

More precisely, this research is based on a case study whose aim is to extract the colours of  $n = 1212$  children’s drawings of gods, drawing  $i = 1, \dots, n$ , which are mainly studied from the psychology of religion perspective (see [Brandt, 2018](#)). Several research questions in the project are related to colours, such as “Which colours are used to draw god?”, “Are the same colours utilised in all countries?”, “Did older children use more or less colours than younger ones?”

## Colour perception

The human colour perception occurs through the radiance incident upon the retina on which three photoreceptor cells, named cones, are located. Only these three photoreceptors are necessary, corresponding to the *Red*, *Green* and *Blue* numerical components (RGB) of the pixel  $\vec{s} = \{s^R, s^G, s^B\}$ , to describe a colour using appropriate spectral functions. Standard curves have been adopted in 1931 by the Commission Internationale de l’Eclairage (CIE) to specify the colour by those three numbers from spectral power distribution transformation.

Although the CIE standard exists, only computers are able to assign those three numbers to a colour in a precise manner. Therefore, any human assignment is basically obsolete due to the influence of an individual’s interpretation of the colour (e.g. “Where can I put a threshold between red and orange?”, “Is it not already brown?”). Moreover, the human perception of colour distribution of an image can be drastically different in terms of vision ([Jobson et al., 1997](#)) and, respectively, in terms of perception depending on the context e.g. *Chubb illusion* or *Checker shadow illusion*.

A number of previous researches in computer vision proposed promising descriptors such as colour histograms (Sun et al., 2006) or colour names (van de Weijer et al., 2009; Lindner and Süsstrunk, 2013) through mapping learned from images. However, their aims differed from the one found in the human sciences and described above. For instance, when colour is used in image retrieval, the aim is to find which images of a dataset correspond to the one stated as the query. Thus, the main point is to determine if the colours of the query and those of the dataset are similar, no matter which colour it is. More specifically, the paper of (Konyushkova et al., 2015) proposes a solution closer to the aim of finding the set of colours displayed in children’s drawings with the well-known method of *K-means*. However, it allows to extract mean colours for a set of drawings which is not the aim here.

For human sciences questions, it is necessary to develop specific techniques, since the aim is not to precisely find the nuance of the colour in the drawing, but to determine the diversity of colours on the one hand (gap 1) ; and to figure whether there is one colour that stands out from a set of colours on the other hand (gap 2).

## Method

Concerning the first gap, through the *state-of-art* of colour retrieval, the most fruitful quantitative approach to translate information between humans (drawer  $\leftrightarrow$  observer) is the average amount of information provided by a stochastic source, namely the Shannon information entropy. Using linear-light conversion, namely greyscale,  $s^{\text{CRT}} = 0.2125s^{\text{R}} + 0.7154s^{\text{G}} + 0.0721s^{\text{B}}$  which reveals the intensity of light, the colour diversity of the drawing is measured by the entropy

$$H(S^{\text{CRT}}) = \frac{-\sum_s p(s^{\text{CRT}}) \log(p(s^{\text{CRT}}))}{\log \delta^{\text{CRT}}} .$$

In the same way, the number of unique grey levels  $\delta^{\text{CRT}}$ , namely type, and the mean intensity  $\mu^{\text{CRT}}$  for each image are computed (see table 1). Those three quantities provide fundamental quantitative information about human colour perception based on the pixel intensity : the higher the entropy value the lighter the colour of the drawing, otherwise coloured drawing, whereas the higher the number of types the coloured drawing (see figure 1) and the higher the mean intensity the lighter the drawing.

In order to fill the gap 2 between the categories of colour and its perception, a method based on the image in the RGB colour space is proposed. With a similar approach of the one developed by (Kim et al., 2007), the dissimilarity between each pixel and a set of colours is computed. More precisely, for each image, a squared




|                       |  |   |   |
|-----------------------|--|---|---|
|                       |  |  |  |
| $\mu^{\text{CRT}}$    | 0.696  | 0.876   | 0.999   |
| $\delta^{\text{CRT}}$ | 457068   | 37982   | 41  |
| $H(S^{\text{CRT}})$   | 1.538  | 1.903   | 5.200   |

Table 1: Illustration of variety measures.

Euclidean dissimilarity is computed between each pixel of the image, in RGB, and a reference set of 117 colours defined in the same colour space. For each pixel its colour will be determined according to its least dissimilarity to the reference set. Finally, the 117 colours are grouped into 11 main colours (red, orange, yellow, green, cyan, blue, purple, pink, white, grey and black) and therefore each pixel belongs to one of these groups.

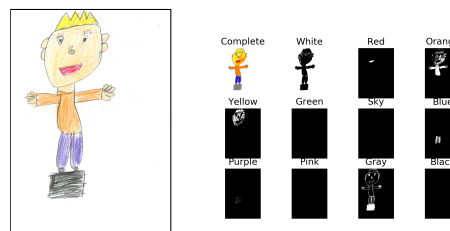


Table 2: Colours quantification from the left image.

Then we can extract the proportion of each colour used in each drawing (see table 2), as well as the presence or absence of each colour. Thus, we can answer the psychological research questions mentioned above. These methods are applicable to other data and research, e.g. studying the main colours used by a painter according to various periods of his/her life.

## Further issues

colquant As a next step, since humans are more sensitive to colour patches than to isolated pixels, a filter should be applied at the beginning of the process, such as the Mumford-Shah Regulariser proposed by Erdem and Tari (2009). Indeed, when children (or adults) fill in an area of the sheet with colour, the application is not regular and consequently not all pixels of the zone are coloured. Thus, in order to avoid underestimating the proportion of one particular colour, standardizing colours by zone could help work around the possible issue.

## References

- P. C. Alejandro and A. Akbarinia. 2016. Nice: A computational solution to close the gap from colour perception to colour categorization. *PLOS ONE* 11(3):1–32.
- R. Benavente, M. Vanrell, and R. Baldrich. 2008. Parametric fuzzy sets for automatic color naming. *J. Opt. Soc. Am. A* 25(10):2582–2593.

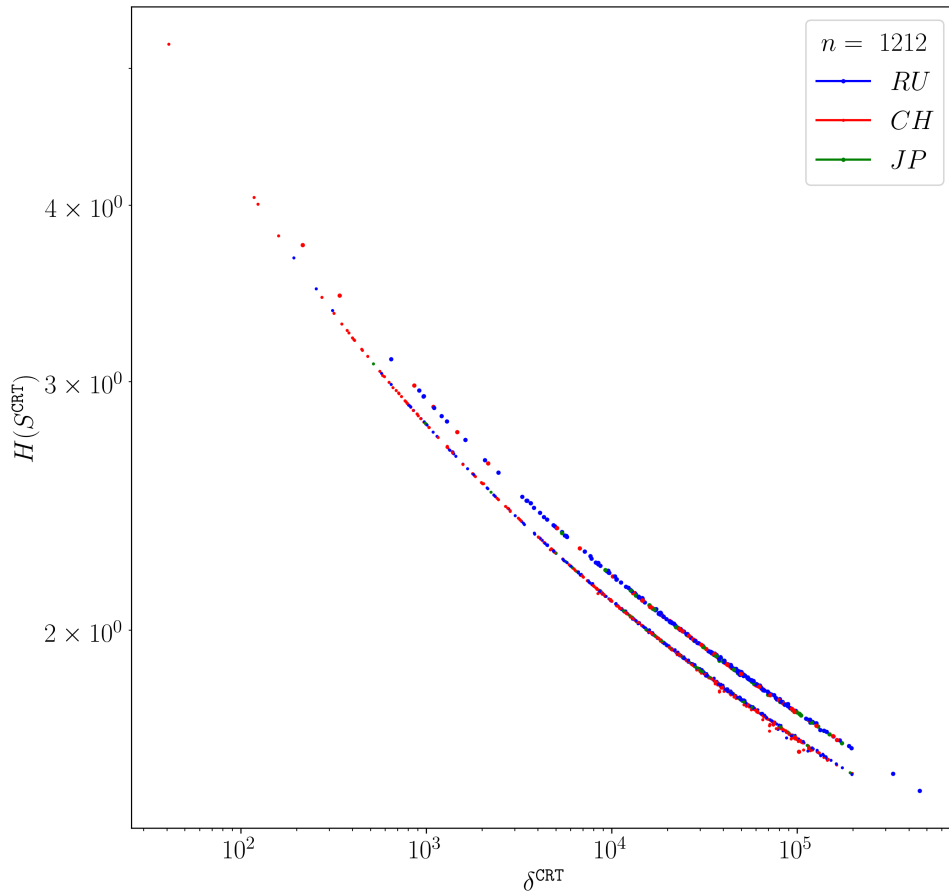


Figure 1: Entropy and Type scatter plot with the country.

- B. Berlin and P. Kay. 1969. *Basic Color Terms: their Universality and Evolution*. University of California Press, Berkeley and Los Angeles.
- P.-Y. Brandt. 2018. [Dessins de dieux](http://ddd.unil.ch/). <http://ddd.unil.ch/>.
- E. Erdem and S. Tari. 2009. Mumford-shah regularizer with contextual feedback. *Journal of Mathematical Imaging and Vision* 33(1):67–84.
- D. J. Jobson, Z. Rahman, and G. A. Woodell. 1997. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image Processing* 6(7):965–976.
- F. S. Khan, R. M. Anwer, J. van de Weijer, A. D. Bagdanov, M. Vanrell, and A. M. Lopez. 2012. Color attributes for object detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3306–3313.
- R. Khan, J. van de Weijer, F. S. Khan, D. Muselet, C. Ducottet, and C. Barat. 2013. Discriminative color descriptors. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2866–2873.
- S. Kim, J. Bae, and Y. Lee. 2007. A computer system to rate the color-related formal elements in art therapy assessments. *The Arts in Psychotherapy* 34(3):223 – 237.
- K. Konyushkova, N. Arvanitopoulos, Z. Dandarova Robert, P.-Y Brandt, and S. Süssstrunk. 2015. God(s) know(s): Developmental and cross-cultural patterns in children drawings. *CoRR* abs/1511.03466. <http://arxiv.org/abs/1511.03466>.
- A. Lindner and S. Süssstrunk. 2013. Automatic color palette creation from words. *Color and Imaging Conference* 2013(1):69–74.
- J. Sun, X. Zhang, J. Cui, and L. Zhou. 2006. Image retrieval based on color distribution entropy. *Pattern Recognition Letters* 27(10):1122 – 1126.
- J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus. 2009. Learning color names for real-world applications. *IEEE Transactions on Image Processing* 18(7):1512–1523.





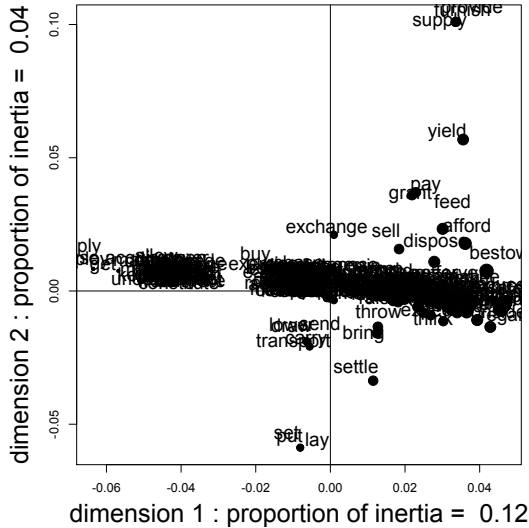
of *correspondence analysis* (CA), where terms are represented by circles and documents by triangles.

As noted, the  $n = 234$  verbs can be *semantically similar* in part, as quantified by a *similarity matrix*  $\mathbb{S} = (s_{ij})$  or a *dissimilarity matrix*  $\mathbb{D} = (d_{ij})$  between pairs of terms, and taking this (arguably substantial) circumstance into account should *reduce* the *distributional dissimilarity* (1) between documents, and, consequently, *lower* the corresponding term-document chi2 statistic, measuring the total dispersion or *inertia*  $\Delta = \frac{1}{2} \sum_{kl} \rho_k \rho_l D_{kl}^X$ , where  $\rho_k = \frac{n \bullet k}{n \bullet \bullet}$ , in the above figure.

Lists of synonyms<sup>1</sup>, yield binary similarity matrices  $s_{ij} = 0$  or 1. More generally,  $\mathbb{S}$  can be defined as a convex combination of binary synonymy relations, insuring its non-negativity, symmetry, positive definiteness, with  $s_{ii} = 1$  for all terms  $i$ . A family of such semantic similarities indexed by the *bandwith parameter*  $\beta > 0$  obtains as

$$s_{ij} = \exp(-\beta d_{ij}/\Delta) \quad \text{where} \quad \Delta = \frac{1}{2} \sum_{ij} f_i f_j d_{ij}$$

$\Delta$  is the *semantic inertia*, and  $d_{ij}$  is the ultrametric, squared Euclidean semantic dissimilarity, here obtained as the *path distance between first senses* in WordNet<sup>2</sup> as mentioned above. Weighted MDS on  $\mathbb{D}$  returns:



Of course, the *distributional* versus *semantic* configuration of verbs depicted in the previous figures differ, and an original proposal aimed at combining both state of affairs consists in replacing the

<sup>1</sup>e.g. <http://www.crisco.unicaen.fr/des/>

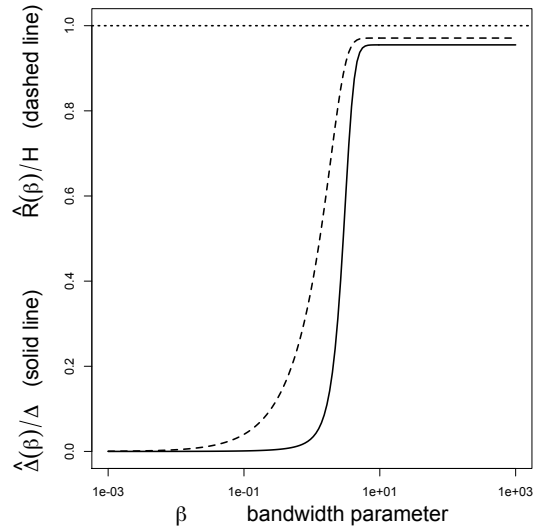
<sup>2</sup><http://search.cpan.org/dist/WordNet-Similarity/lib/WordNet/Similarity/path.pm>

chi2 dissimilarity (1) by the *reduced* squared Euclidean distance between documents

$$\hat{D}_{kl} = \sum_{ij} \mathbb{t}_{ij} (q_{ik} - q_{il})(q_{jk} - q_{jl}) \quad (2)$$

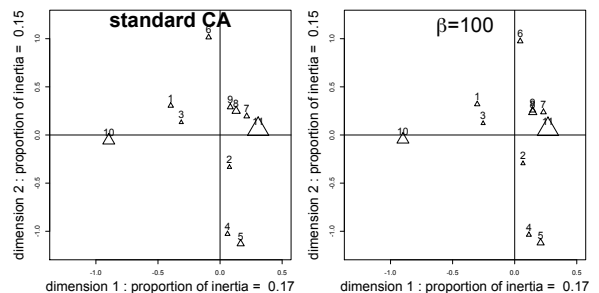
$$\text{where } \mathbb{t}_{ij} = \frac{f_i f_j s_{ij}}{\sqrt{b_i b_j}} \quad \text{and} \quad b_i = \sum_j s_{ij} f_j = (\mathbb{S}f)_i$$

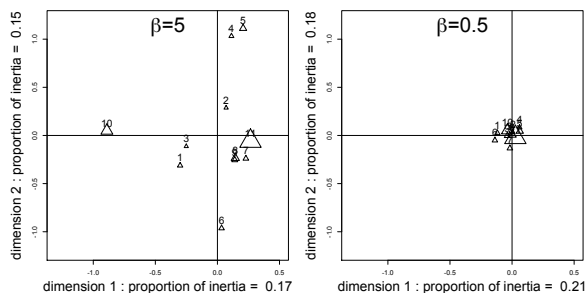
Quantity  $b_i \in [f_i, 1]$  is identified as a measure of *ordinariness* or *banality* in ecology (e.g. [Leinster and Cobbold, 2012](#); [Marcon et al., 2014](#)). As a matter of fact, it can be shown that a binary  $\mathbb{S}$  makes  $\hat{D}_{kl}$  identical to the chi2 dissimilarity (1), with the exception that the sum now runs on *cliques of synonyms* rather than terms. Also, the limit  $\beta \rightarrow 0$  makes  $\hat{D}_{kl} \rightarrow 0$  with a reduced inertia  $\hat{\Delta}(\beta) = \frac{1}{2} \sum_{kl} \rho_k \rho_l \hat{D}_{kl}$  tending to zero. In the opposite direction,  $\beta \rightarrow \infty$  makes  $\hat{D}_{kl} \rightarrow D_{kl}^X$  provided  $d_{ij} > 0$  for  $i \neq j$ , a circumstance violated in the case study, where the  $n = 234$  verbs display, accordingly to their first sense in WordNet, 15 cliques of size 2 and 3 cliques of size 3 (namely, employ-apply-use, set-lay-put and supply-furnish-provide). In any case, the *relative reduced inertia*  $\hat{\Delta}(\beta)/\Delta$  is increasing in  $\beta$ :



and so is its *reduced diversity*  $\hat{R}(\beta) = -\sum_i f_i \ln b_i \leq H = -\sum_i f_i \ln f_i$ , where  $H$  is Shannon entropy.

The resulting MDS on reduced dissimilarities (2) among the 11 documents yields a new, *semantically-reduced correspondance analysis*:





The bandwidth parameter  $\beta$  controls the *paradigmatic sensitivity* of the linguistic subject: the higher  $\beta$ , the larger the distances between the semantic of documents, and the larger the spread of the factorial cloud as measured by reduced inertia  $\hat{\Delta}(\beta)$ . On the other direction, a low  $\beta$  can model an illiterate person, sadly unable to discriminate between documents, which look all alike.

**Conclusion and further issues** Despite the technicality of its exposition, the idea of this contribution is straightforward, namely to propose a way to take semantic similarity explicitly into account, within the classical distributional similarity framework provided by correspondence analysis. Alternative approaches and variants are obvious: further analysis on non-verbs should be investigated; other definitions of  $\hat{D}$  are worth investigating; other choices of  $\mathbb{S}$  are possible (in particular the original  $\hat{\mathbb{S}}$  extracted from Wordnet), and, in particular, alternatives to WordNet path similarities (e.g., for languages in which WordNet is not defined) are required.

On the document side, and despite its numerous achievements, the term-document matrix still relies on a rudimentary approach to textual context, modeled as  $p$  documents consisting of *bag of words*. Much finer *syntagmatic* descriptions are possible, captured by the general concept of *exchange matrix*  $E$ , giving the joint probability to select a *pair of textual positions* through textual navigation (by reading, hyperlinks or bibliographic zapping, etc.).  $E$  defines a weighted network whose nodes are the textual positions occupied by terms (Bavaud et al., 2015).

The parallel with *spatial issues* (quantitative geography, image analysis), where  $E$  defines the “where”, and the features dissimilarities between positions  $\mathbb{D}$  defines the “what”, is immediate (see e.g. Egloff and Ceré, 2017). In all likelihood, developing both axes, that is taking into account semantic similarities on generalized textual networks, should provide a fruitful extension and renewal of the venerable term-document matrix

paradigm.

## References

- François Bavaud. 2011. On the Schoenberg transformations in data analysis: Theory and illustrations. *Journal of Classification* 28(3):297–314. <https://doi.org/10.1007/s00357-011-9092-x>.
- François Bavaud, Christelle Cocco, and Aris Xanthos. 2015. Textual navigation and autocorrelation. In G. Mirkros and J. Macutek, editors, *Sequences in Language and Text*. De Gruyter Mouton, pages 35–56.
- Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, page 31.
- David M Blei. 2012. Probabilistic topic models. *Communications of the ACM* 55(4):77–84.
- Jinho D Choi. 2016. Dynamic feature induction: The last gist to the state-of-the-art. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 271–281.
- Mattia Egloff and Raphaël Ceré. 2017. Soft textual cartography based on topic modeling and clustering of irregular, multivariate marked networks. In *International Workshop on Complex Networks and their Applications*. Springer, pages 731–743.
- Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788.
- Tom Leinster and Christina A Cobbold. 2012. Measuring diversity: the importance of species similarity. *Ecology* 93(3):477–489.
- Eric Marcon, Zhiyi Zhang, and Bruno Héroult. 2014. The decomposition of similarity-based diversity and its bias correction .
- Barbara McGillivray, Christer Johansson, and Daniel Apollon. 2008. Semantic structure from correspondence analysis. In *Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing*. Association for Computational Linguistics, pages 49–52.
- George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.
- Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Disability Studies* 20:33–53.
- Adam Smith. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations; Book I*. Project Gutenberg, Urbana, Illinois. Also known as: Wealth of Nations. <http://www.gutenberg.org/ebooks/3300>.



# Towards a quantitative research framework for historical disciplines

**Barbara McGillivray**

The Alan Turing Institute  
University of Cambridge

bmcgillivray@turing.ac.uk

**Giovanni Colavizza**

The Alan Turing Institute  
gcolavizza@turing.ac.uk

**Tobias Blanke**

King's College London  
tobias.blanke@kcl.ac.uk

## 1 Background and motivation

The ever-expanding wealth of digital material that researchers have at their disposal today, coupled with growing computing power, makes the use of quantitative methods in historical disciplines increasingly more viable. However, applying existing techniques and tools to historical datasets is a non-trivial enterprise (Piotrowski, 2012; McGillivray, 2014). Moreover, scholarly communities react differently to the idea that new research questions and insights can arise from quantitative explorations that could not be made using purely qualitative approaches. Some of them, such as linguistics (Jenset and McGillivray, 2017), have been acquainted with quantitative methods for a longer time, while others have attempted and largely rejected them in the past, thus assuming nowadays a more conservative attitude, as is the case for history (Hitchcock, 2013).

## 2 Towards a historical research framework

Historical disciplines, i.e. those focusing on study of the past, possess at least three characteristics, which set them apart and require careful consideration in this context: the need to work with closed corpora which can only be expanded working on past records (Mayrhofer, 1980), the focus on phenomena that change over time, and the frequent need to combine quantitative and qualitative methods. For these reasons, we notice the need for a general methodological reflection that can help in the process of conducting research in historical disciplines, by taking full advantage of quantification. In this contribution, we start from a framework proposed by Jenset and McGillivray (2017) for quantitative historical linguistics and illustrate it with a case study focusing on semantic change in a corpus of UK government texts after

1945. We then apply the framework on a different case study related to economic history: a statistical analysis of the conditions of contracts of apprenticeship in early modern Venice. This comparison will allow us to highlight the points of alignment or friction that a framework developed for historical linguistics displays when applied to history, in view of proposing a more general one. Following Andersen and Hepburn (2015), we focus on the relationship between evidence, modelling and research practice in historical disciplines.

Jenset and McGillivray (2017)'s framework is the only general framework available for quantitative historical linguistics. A comparable framework, but more limited in scope, can be found in Köhler (2012). Jenset and McGillivray (2017)'s framework starts from the assumption that linguistic historical reality is lost and the aim of quantitative research is to arrive at models of and claims on such reality, which are quantitatively driven from evidence and lead to consensus among the scholarly community. The scope is delimited to the cases where quantifiable evidence (such as n-grams or numerical data) can be gathered from primary sources. Claims are defined as statements based on evidence (Carrier, 2012) and annotated datasets such as corpora are considered as distributional evidence to study phenomena in historical linguistics. Claims possess a strength proportional to that of the evidence supporting them. In this context, "model" means a formalized representation of a phenomenon, be it statistical or symbolic (Zuidema and de Boer, 2014). Models (including those deriving from hypotheses tested quantitatively against evidence) are research tools embedding claims or hypotheses, useful in order to produce novel claims and hypotheses in turn via "a continual process of coming to know by manipulating representations" (McCarty, 2004).

### 3 Case studies

The first case study where we apply [Jenset and McGillivray \(2017\)](#)'s framework considers a recent collaboration between Digital Humanities and History at Kings College London ([Blanke and Wilson, 2017](#)), to develop a “materialist sociology of political texts” following Moretti’s ideas of distant reading ([Moretti, 2013](#)). The project worked on a corpus of post-1945 UK government white papers to map connections and similarities in political communications from 1945 to 2010. As the corpus is time-indexed, a quantitative analysis allowed to trace the changing shape of political language, by tracking clusters of terms relating to particular concepts and charting the changing meaning of words. Temporal information was added as annotations to the corpus using a dictionary-based approach. Creating the distributional quantitative evidence involved text pre-processing to create a term-document matrix. Compared to earlier attempts, the project relied on models for historical texts not only to read the texts themselves but also to develop ways of classifying them into time intervals. More advanced modelling was applied to trace changes of meaning in key political concepts across time intervals, using topic models and word embeddings, allowing to test historiographical and linguistic hypotheses.

The second case study focuses on apprenticeship contracts registered in Venice between the end of the 16<sup>th</sup> century and the beginning of the 18<sup>th</sup> ([Ehrmann et al., 2018](#)). These archival records constitute primary sources often used qualitatively by historians. The source was annotated and transcribed into a structured database. The annotation schema was developed bottom-up via incremental refinements, and included controlled vocabularies and heuristics. The quantitative evidence consists of key and recurring information contents from each contract. This data can be textual, categorical or numerical. An example of quantitative analysis on this dataset is on the historical use of contracts of apprenticeship: to hire cheap workforce, for actual training, or both? The working hypothesis was that contracts could be used flexibly for both ends, with fine-grained variability at the level of professions, guilds and even masters. A statistical model (linear regression) provided support for this claim, and results were interpreted within a broader scope of primary and secondary evidence ([Bellavitis et al., forth-](#)

coming).

### 4 Conclusion and future work

This comparison leads us to the conclusion that, despite the broad applicability of [Jenset and McGillivray \(2017\)](#)'s framework in both cases, some important differences emerge between historical linguistics and history. We discuss two. First of all, the scope of primary source and its quantitative representation is broader in history, including not only distributional but also categorical, ordinal, and numerical evidence. Secondly, the scope for a purely quantitative approach is less broad: quantitative evidence and models can often only contribute to inform hypotheses and claims which rely on qualitative evidence and methods. While the framework can extend to a variety of primary sources and different quantitative evidence, it does not yet integrate qualitative results and methods. We thus conclude with the following question for debate and future work: how can quantitative and qualitative methods be combined into a single methodological framework?

### Acknowledgments

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. BM is supported by the Turing award TU/A/000010 (RG88751).

### References

- Hanne Andersen and Brian Hepburn. 2015. Scientific method. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University. Winter 2015 edition.
- Anna Bellavitis, Riccardo Cella, and Giovanni Colavizza. forthcoming. Apprenticeship in early modern venice. In Maarten Prak and Patrick Wallis, editors, *Apprenticeship in Early Modern Europe*, Cambridge University Press, Cambridge.
- Tobias Blanke and Jon Wilson. 2017. Identifying epochs in text archives. In *2017 IEEE International Conference on Big Data (Big Data)*. pages 2219–2224.
- Richard Carrier. 2012. *Proving history: Bayes’s theorem and the quest for the historical Jesus*. Prometheus Books, Amherst, N.Y.
- Maud Ehrmann, Giovanni Colavizza, Orlin Topalov, Riccardo Cella, Davide Drago, Andrea Erbo, Francesca Zugno, Anna Bellavitis, Valentina Sapienza, and Frédéric Kaplan.

2018. From documents to structured data: First milestones of the ‘Garzoni’. *DHCOMMONS* <http://dhcommons.org/journal/documents-structured-data-first-milestones-garzoni>.
- Tim Hitchcock. 2013. Confronting the digital: Or how academic history writing lost the plot. *Cultural and Social History* 10(1):9–23.
- Gard B. Jensen and Barbara McGillivray. 2017. *Quantitative Historical Linguistics. A Corpus Framework*. Oxford University Press, Oxford.
- Reinhard Köhler. 2012. *Quantitative Syntax Analysis*. de Gruyter Mouton.
- Manfred Mayrhofer. 1980. *Zur Gestaltung des etymologischen Wörterbuchs einer “Großcorpus-Sprache”*. Akademie der Wissenschaften. Phil-Hist. Klasse., Wien: Österr.
- Willard McCarty. 2004. Modeling: A study in words and meanings. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A Companion to Digital Humanities*, Blackwell Publishing Ltd., Malden, MA, USA, pages 254–270.
- Barbara McGillivray. 2014. *Methods in Latin Computational Linguistics*. Brill, Leiden.
- Franco Moretti. 2013. *Distant Reading*. Verso, London.
- Michael Piotrowski. 2012. *Natural language processing for historical texts*. Morgan & Claypool, San Rafael, CA.
- Willem Zuidema and Bart de Boer. 2014. Modeling in the language sciences. In Robert J. Podesva and Devyani Sharma, editors, *Research Methods in Linguistics*, Cambridge University Press, Cambridge, pages 428–445.

# Modelling vagueness - A criteria-based system for the qualitative assessment of reading proposals for the deciphering of Classic Mayan hieroglyphs

**Franziska Diehr**  
**Maximilian Brodhun**  
Niedersächsische Staats- und  
Unveristätsbibliothek Göttingen  
Platz der Göttinger Sieben 1  
37073 Göttingen

**Sven Gronemeyer**  
**Christian Prager**  
**Elisabeth Wagner**  
**Katja Diederichs**  
**Nikolai Grube**  
Rheinische Friedrich-Wilhelms-  
Universität Bonn  
Abteilung für Altamerikanistik  
Oxfordstr. 15  
53111 Bonn

## Abstract

In this paper we present an ontology-based modelling approach that deals with vague information in the process of the decipherment of the Classic Mayan script. We introduce the challenges of deciphering this script and the resulting requirements for the development of a digital Sign Catalogue. Subsequently, we consider the process of modelling as a method of Digital Humanities research and describe how we applied it to the development of the Sign Catalogue. Further we present our concept for the systematisation and classification of signs and how we developed a system for the qualitative assessment of reading hypotheses. Finally, we highlight the advantages of the developed model, which adapts flexibly to the ongoing decipherment process.

## 1 Dealing with vague information in the process of deciphering

The aim of our project 'Text Database and Dictionary of Classic Mayan' is to compile a corpus of all known inscriptions in order to develop a dictionary for the not yet fully deciphered language and script of Classic Mayan.<sup>1</sup> One of the challenges we have to deal with is the low status of decipherment of the logo-syllabic script. In the short history of deciphering the Maya script, researchers presented various hypotheses about the reading of the glyphs. The first results did not appear until the

<sup>1</sup>see <http://mayadictionary.de> for further information

1950s. But it was not until the 1980s that a number of breakthroughs, such as that of Stuart (1987), significantly influenced the deciphering process, resulting also in the publication of eleven glyph inventories. Nevertheless, the exact number of signs and their graphic variants is still unknown. In addition, for approximately a quarter to a third of the characters there are only vague or no deciphering proposals at all. Many of those proposals are competing with each other, because the readings may only be valid in selected contexts, but they do not have to exclude each other because of the possible polyvalence of the signs (Gronemeyer et al., 2018). In our project we have to face the challenges of working with those vague and uncertain information. To deal with this situation, we developed a digital Sign Catalogue aiming to establish a new concept for the systematisation and classification of signs and a system for the qualitative assessment of reading hypotheses. The Sign Catalogue is also used as basis for creating a machine-readable text for the corpus.

## 2 Modelling as a method of Digital Humanities research

We understand modelling as a method of Digital Humanities research, which aims to represent objects and the knowledge about them in a computational model. In our belief, the process of knowledge representation is a hermeneutic method that is used to construct a machine-readable model. In the sense of Sowa (2000) this means making the semantics of knowledge objects explicit and to transferring it into a data model. In order to determine which domain-specific requirements exist for the classification of Maya hieroglyphs, we used an explorative-hermeneutic method. This

process presupposes that the modeller is familiar with the domain and can describe the subject area and its knowledge base from a disciplinary point of view. Further this approach forces domain experts to question how their objects are defined and which methods were used to gain knowledge about them in the first place. This process of conceptual modelling is used to define what Sowa (2000) calls 'ontological categories'. They "determine everything that can be represented in a computer application". This concludes that the creation of an ontological model aims to explicitly describe the objects, their relation to each other, and to their domain. Defining these categories is especially challenging when dealing with vague and uncertain information, because "any incompleteness, distortions, or restrictions in the framework of categories must inevitably limit the generality of every program and database that uses those categories" (Sowa, 2000). If 'knowledge' about objects can be questioned or interpreted differently, it is necessary to present the different states of knowledge in the model in order to counteract such distortions while limiting the knowledge base exactly for the purpose of the defined ontological categories. The following sections outline how we dealt with this by modelling a Sign Catalogue that can handle complex sign classification in a flexible way. Further we explain the system for the qualitative assessment of reading hypotheses and how it supports the ongoing research on the decipherment of the Classic Mayan script.

### 3 Defining concepts for the systematisation and classification of signs

To find suitable concepts for describing signs, we have examined existing classification systems and linguistic terminologies (GOL, 2010) (Chiarcos and Sukhareva, 2015). We have found that most concepts are not suitable for the classification of Mayan glyphs because they focus too much on applicability in a particular linguistic context and therefore cannot be applied to a writing system with a low degree of decipherment. For this reason, we created a model that uses linguistic categories only on a meta level and does not take further analysis levels and grammars into account. For the development of the Sign Catalogue we chose an ontological modelling approach, which uses the CIDOC CRM (Crofts et al., 2011) as base

ontology. Despite its focus on documentation processes of cultural heritage objects, the ontology contains a lot of meta-concepts that are suitable for our catalogue.<sup>2</sup> In our catalogue, we define the sign as an entity consisting of a functional and phonemic level (*Sign*) and a graphical representation (*Graph*). The class *Graph* represents all variants of a grapheme (allographs). By the separately recording discrete graphs, we enable an exact method for their identification. This relation from *Graph* to the corresponding *Sign* is optional, so that even graphs can be recorded that could not have been assigned to any functional-phonemic level yet (Diehr et al., 2017). The class *Sign* is determined by its *SignFunction*: the use of the sign as a logogram, syllabogram, numeral or diacritical sign. The phonemic level of the sign is recorded as *transliterationValue* at the respective *SignFunction*. To represent the polyvalency of signs only one value is allowed per function, but one sign can have several sign functions and therefore readings (Diehr et al., 2018).

The developed concept for the digital Sign Catalogue requires a data structure that allows to create semantic relations between uniquely referenceable entities. Therefore we chose to implement the model in RDF. For the management, creation and presentation of the data generated in the project we use the virtual research environment (VRE) TextGrid (Neuroth et al., 2015). In order to record the signs in the VRE, we have adapted the RDF input mask of the TextGrid Lab to project-specific requirements.

### 4 A system for the qualitative assessment of reading hypotheses

The system for the formal evaluation of reading proposals arose from the requirement to explicitly describe those aspects that led to the formulation of a specific reading hypothesis. What factors must be taken into account for a proposed reading to be plausible? For this purpose we defined formal criteria with which reading hypotheses can be described. For assigning a level of confidence to a transliteration value, we modelled the class *ConfidenceLevel*, which is related to the *SignFunction* and therefore to the *translationValue* of the sign. For each *SignFunction* a separate set of criteria

---

<sup>2</sup>for a deeper insight have a look on the documentation of the Sign Catalogue: <http://idiom-projekt.de/idiommask/schema.html>



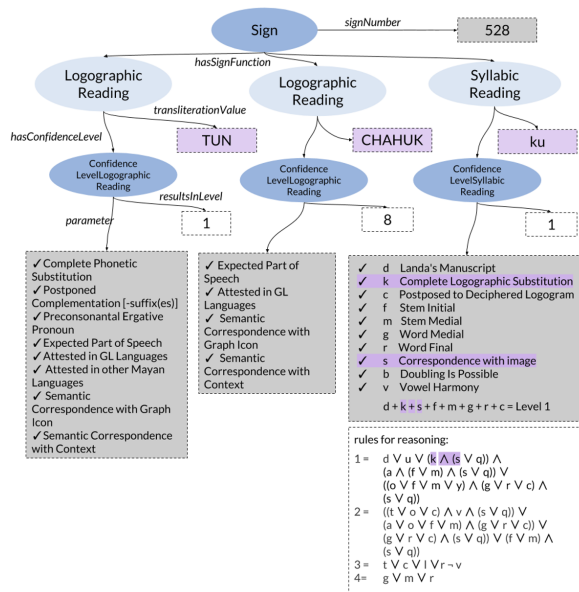


Figure 1: Modelling Confidence of Reading Hypotheses

based on Kelley (1962) and Houston (2001) was developed. The criteria are primarily oriented towards the context of graphematic and linguistic use (e. g. plausible text-image-reference) or the substantiation in modern Mayan languages. The criteria are related by means of propositional logic so that, depending on their combination, a quality level is inferred, see Fig. 1 (Diehr et al., 2017). The qualitative evaluation is particularly relevant for examine the plausibility of the reading hypotheses in the corpus. Readings with a high level can be compared with those with a low level. For the latter, new criteria for their plausibility could also be found in the context of the text, which can then be added to the Sign Catalogue. This may also increase the confidence level and therefore the quality of the reading proposal (Diehr et al., 2018).

## 5 Conclusion and prospect

The ontological modelling approach and the implementation in a RDF data model offers a flexibility that redefines the classification of signs and allows precise identification on the basis of distinguishing characteristics. By incorporating known and adapting new results, the digital Sign Catalogue is specifically designed to deal with vagueness in research processes. This approach can also be applied to other (complex) writing systems. It would also be interesting to apply the criteria-based approach to other applications and to investigate to what extent it offers a suitable method for dealing with vagueness.

investigate to what extent it offers a suitable method for dealing with vagueness.

## References

2010. General ontology for linguistic description (gold). Department of Linguistics (The LINGUIST List) <http://linguistics-ontology.org/>.
- C. Chiarcos and M. Sukhareva. 2015. Olia - ontologies of linguistic annotation. *Semantic Web Journal* 6(4):379–386.
- N. Crofts, M. Doerr, T. Gill, S. Stead, and M. Stiff, editors. 2011. *CIDOC Conceptual Reference Model, Version 5.0.4*. <http://www.cidoc-crm.org/cidoc-crm/>.
- F. Diehr, S. Gronemeyer, C. Prager, M. AND Wagner E. Brodhun, K. Diederichs, and N. Grube. 2018. Ein digitaler zeichenkatalog als organisationssystem für die noch nicht entzifferte schrift der klassischen maya. In C. Wartena, M. Franke-Maier, and E. De Luca, editors, *Knowledge Organization for Digital Humanities - Proceedings of the 15th Conference on Knowledge Organization WissOrg'17 of the German Chapter of the International Society for Knowledge Organization (ISKO)*. pages 37–43. <https://doi.org/10.17169/FUDOCs.document.000000028863>.
- F. Diehr, S. Gronemeyer, C. Prager, M. Brodhun, E. Wagner, K. Diederichs, and N. Grube. 2017. Modellierung eines digitalen zeichenkatalogs fr die hieroglyphen des klassischen maya. In M. Eibl and M. Gaedke, editors, *INFORMATIK 2017*. Gesellschaft für Informatik, Bonn, pages 1185–1196. [https://doi.org/10.18420/in2017\\_120](https://doi.org/10.18420/in2017_120).
- S. Gronemeyer, F. Diehr, C. Prager, M. Brodhun, E. Wagner, K. Diederichs, and N. Grube. 2018. Vagheit hoch zweifel plus kritik! die bewertung von widersprüchen in einer digitalen entzifferungsarbeit der maya-hieroglyphen. In *DHd 2018 Kritik der digitalen Vernunft - Konferenzabstracts*. <http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf>.
- Stephen Houston. 2001. The decipherment of ancient maya writing pages 3–19.
- David H. Kelley. 1962. Review of a catalog of maya hieroglyphs, by j. eric s. thompson. *American Journal of Archaeology* 66:436–438.
- H. Neuroth, A. Rapp, and S. Söring, editors. 2015. *TextGrid: Von der Community fr die Community*. <https://doi.org/https://doi.org/10.3249/webdoc-3947>.
- John F. Sowa. 2000. *Knowledge Representation: logical, philosophical, and computational foundations*. Brooks/Cole.
- David Stuart. 1987. Ten phonetic syllables. *Research Reports on Ancient Maya Writing* 14 .

# Linking Historical Sources to Established Knowledge Bases in Order to Inform Entity Linkers in Cultural Heritage

Gary Munnely and Annalina Caputo and Séamus Lawless

Adapt Centre

Trinity College Dublin

firstname.lastname@adaptcentre.ie

## 1 Introduction

While Entity Linking (EL) has seen much development over the years (Bunescu and Pasca, 2006; Milne and Witten, 2008; Ratnov et al., 2011; Yosef et al., 2011; Usbeck et al., 2014; Waitelonis and Sack, 2016), it is hindered by several limitations when applied to Cultural Heritage (CH) collections. Most notable is a significant underrepresentation of entities in common Knowledge Bases (KB) such as DBpedia (Agirre et al., 2012; Van Hooland et al., 2015; Munnely and Lawless, In Press). A possible solution is to construct more specific KBs from resources used by scholars investigating CH material.

The overarching research related to this paper investigates methods of performing EL on primary source Irish historical archives. Two resources used by historians in this domain are the Oxford Dictionary of National Biography (ODNB)<sup>1</sup> and the Dictionary of Irish Biography (DIB)<sup>2</sup>. Both are collections of biographies written by historians with a single entity usually being the focus of the text. Titles contain the subject’s forename, surname and variant names, and links between related biographies exist in the text of each article. Hence they exhibit structural properties similar to those that originally made Wikipedia a useful KB for EL. They are of greater specificity to the history of the British Isles than other more general resources and thus may help to fill some of the gaps in DBpedia, or at the very least limit the scope of the linker’s search to entities that are relevant to this geographic region.

Research by Brando et al. (Brando et al., 2016) has shown that it is beneficial to EL in CH when a specialised KB can be integrated with a more general one. Hence the goal of this work is to con-

nect entries in ODNB and DIB with their corresponding entries in DBpedia, such that a new KB built on these resources would be linked with their counterparts in a larger, more established semantic resource where such counterparts exist. This also helps to identify entities in ODNB and DIB which are not yet documented in DBpedia, showing where an EL system that is informed by a KB based on ODNB and DIB may be better equipped for linking in Irish historical archives.

## 2 Method

In order to facilitate the integration of a KB derived from ODNB or DIB with DBpedia, an approach for linking biographies to their DBpedia counterparts was developed. First, all DBpedia entities belonging to the class `dbo:Person` are indexed using Solr. The name of each entity, the full text of the Wikipedia article from which they are derived, and anchor text on incoming links to the article were indexed. Anchor text indicates alternative names which may refer to an entity. For example, the DIB biography for the 7<sup>th</sup> Earl of Mayo uses his full name and excludes his title, “Dermot Robert Wyndham-Bourke” while his name in DBpedia is given as “Dermot Bourke 7<sup>th</sup> Earl of Mayo”. Indexing anchor text loosely captures the equivalence of these two references.

For each biography entry in ODNB and DIB  $b \in \mathcal{B}$ , the title  $b_{title}$  is executed as a query against Solr. Matches on the title field and anchor text are boosted over matches in the article’s content. A list of up to ten top-ranked candidates  $\mathcal{P}_b$  is returned. The best matching DBpedia referent  $p_b^* \in \mathcal{P}_b$  for a given biography is the one that maximises the expression:

$$p_b^* = \operatorname{argmax}_p \Psi(b, p), \forall p \in \mathcal{P}_b \quad (1)$$

Where  $\Psi(b, p)$  is computed as a linear combi-

<sup>1</sup><http://www.oxforddnb.com/>

<sup>2</sup><http://dib.cambridge.org/>

nation of content similarity and name similarity.

For a given candidate  $p \in \mathcal{P}_b$ , content similarity  $\Omega$  between the biography  $b_{content}$  and the candidate’s Wikipedia article  $p_{article}$  is computed using negative Word Mover’s Distance (WMD) (Kusner et al., 2015) as implemented in gensim (Řehůřek and Sojka, 2010). This method establishes a vector representation of documents using word embeddings and then computes the distance between points in the two representations. Similarity is the negation of the normalised distance. Word embeddings are computed using a Word2Vec model (Mikolov et al., 2013) trained on a Wikipedia dump excluding redirects, disambiguation pages etc.

The name similarity function  $\Phi$  is based on the Monge-Elkan Method (Monge and Elkan, 1996). The biography title  $b_{title}$  and name of a candidate  $p_{name}$  are lower-cased and tokenized. Stop words are removed yielding two sets of tokens  $\mathcal{T}_b$  and  $\mathcal{T}_p$ . The sets are added to a bipartite graph with edge weights computed using Jaro-Winkler similarity (Winkler, 1990). An optimal mapping  $\mathcal{T}_b \mapsto \mathcal{T}_p$  is found using Edmond’s blossom algorithm (Edmonds, 1965) giving  $\mathcal{W}$ , the set of weighted edges which comprise the mapping. Name similarity is the generalised mean of the edge weights in  $\mathcal{W}$  as described by Jimenez et al. (Jimenez et al., 2009) where  $m = 2$  in this experiment:

$$\Phi(b, p) = \left( \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} w^m \right)^{\frac{1}{m}} \quad (2)$$

This yields the final formulation of  $\Psi$  as a function of the form:

$$\Psi(b, p) = \alpha \Phi(b_{title}, p_{name}) + \beta \Omega(b_{content}, p_{article}) \quad (3)$$

Where  $\alpha$  and  $\beta$  are tuning parameters chosen such that  $\alpha + \beta = 1$ .

A hard threshold  $\tau$  is applied to  $p_b^*$ , enforcing a minimum similarity between a biography and its final chosen referent  $\bar{p}_b^*$ :

$$\bar{p}_b^* = \begin{cases} p_b^*, & \text{if } \Psi(b, p_b^*) > \tau \\ NIL, & \text{otherwise} \end{cases} \quad (4)$$

NIL indicates that a biography does not have a DBpedia counterpart.

### 3 Evaluation

The approach described is essentially an EL solution, hence the BAT Framework (Cornolti et al., 2013) was used for evaluation. Two ground truths were derived from a random sample of 200 biographies obtained from both DIB and ODNB (400 samples in total). Samples were manually assigned a DBpedia URI. Where no URI could be established, a NIL label was applied. Ultimately 64 of the ODNB samples and 72 of the DIB samples were labelled as NIL.

A threshold similarity of  $\tau = 0.55$  was found to give the best results when  $\alpha = 0.1$  and  $\beta = 0.9$ , but the disparity in performance between the two collections is wide. Based on the evaluation, this approach achieves a score of 81.5% on DIB, but only 67.5% on ODNB. Some of the imprecision stems from Solr as 43.1% of incorrect labels on ODNB and 45.9% of incorrect labels on DIB can be ascribed to the correct referent not being among the results returned by the search engine. It is likely that the remaining errors with the approach are due to problems with document length normalisation as WMD is a measure of distance rather than relevance.

### 4 Conclusion

The approach documented seems promising given the scores achieved on DIB, but the wide variance in performance across the two collections indicates underlying problems. Future work will focus on making the approach more robust in an effort to eliminate this unreliability.

An objective of this work is to construct a KB which will be useful for EL on Irish historical archives. Future work will also investigate the applications of a KB, which is integrated with other semantic web resources, for EL, using the approach presented in this paper.

### Acknowledgments

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

### References

Eneko Agirre, Ander Barrena, Oier Lopez De Lacalle, Aitor Soroa, Samuel Fern, and Mark Steven-



- son. 2012. *Matching Cultural Heritage items to Wikipedia*.
- Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia. 2016. REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets. *Complex Systems Informatics and Modeling Quarterly* (7):60 – 80.
- Razvan C Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Eacl*. volume 6, pages 9–16.
- Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. 2013. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, pages 249–260.
- Jack Edmonds. 1965. Paths, trees, and flowers. *Canadian Journal of mathematics* 17(3):449–467.
- Sergio Jimenez, Claudia Becerra, Alexander Gelbukh, and Fabio Gonzalez. 2009. Generalized monguelkan method for approximate text string comparison. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pages 559–570.
- Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*. pages 957–966.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.
- David Milne and Ian H. Witten. 2008. Learning to Link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. ACM, New York, NY, USA, CIKM '08, pages 509–518.
- Alvaro Monge and Charles Elkan. 1996. The field matching problem: Algorithms and applications. In *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. pages 267–270.
- Gary Munnely and Séamus Lawless. In Press. Investigating entity linking in early english legal documents. In *Digital Libraries (JCDL), ACM/IEEE Joint Conference on*.
- Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 1375–1384.
- Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, pages 45–50. <http://is.muni.cz/publication/884893/en>.
- Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Michael Röder, Daniel Gerber, Sandro Athaide Coelho, Sören Auer, and Andreas Both. 2014. Agdistis-graph-based disambiguation of named entities using linked data. In *International Semantic Web Conference*. Springer, pages 457–471.
- Seth Van Hooland, Max De Wilde, Ruben Verborgh, Thomas Steiner, and Rik Van de Walle. 2015. Exploring entity recognition and disambiguation for cultural heritage collections. *Digital Scholarship in the Humanities* 30(2):262–279.
- Jörg Waitelonis and Harald Sack. 2016. Named entity linking in# tweets with kea. In *# Microposts*. pages 61–63.
- William E. Winkler. 1990. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. .
- Mohamed Amir Yosef, Johannes Hoffart, Ilaria Bordino, Marc Spaniol, and Gerhard Weikum. 2011. Aida: An online tool for accurate disambiguation of named entities in text and tables. *Proceedings of the VLDB Endowment* 4(12):1450–1453.

# Supporting hermeneutic interpretation of historical documents by computational methods

**Cristina Vertan,**  
University of Hamburg  
[cristina.vertan@uni-hamburg.de](mailto:cristina.vertan@uni-hamburg.de)

## Abstract

Digitalization campaigns during the last ten years made available a considerable number of historical texts. The first digitalization phase concentrated on archiving purposes; thus the annotation was focused on layout and editorial information. The TEI standard developed dedicated modules for this purpose. However, the next phase of digital humanities implies active involvement of computational methods for interpretation and fact discovery within digital historical collections, i.e. active computational support for the hermeneutic interpretation.

We argue that interpretation of historical documents cannot be realised by simple black-box algorithms which rely just on the graphical representation of words (bag of strings - BoW) but by:

1. Considering semantics, which implies a deep annotation of text at several layers and
2. Explicitly annotating vague information
3. Making use of non-crisp reasoning (fuzzy logic, rough sets)

For any high-level content analysis, the deep annotation (manual semi-automatic or even automatic) is an unavoidable process.

For modern languages there are meanwhile established standards and rich tools which ensure an easy and error-prone annotation process. In this contribution we want to illustrate the challenges and special requirements connected with the annotation of historical texts, and argue that in many cases the data-model is so complex that corpus, respectively language tailored tools have still to be developed.

The annotation of historical texts has to consider following criteria:

- The text to be annotated may change during the annotation. Several scenarios may converge to this situation:
  - o Original text is damaged and only the deep annotation and interpretation of neighbouring context can provide a possible reconstruction;
  - o The text is a transliteration from another alphabet. In this case transliterations are rarely standardised (also because historical language was not standardised and phonetical changes like insertion of vowels, doubling of consonants are subject of the interpretation of the annotator and assignment of one or other part-of-speech;
  - o The documents are a mixture of several languages and OCR performs low.
- The annotation has to be done at several layers: text structure, linguistic, domain-specific. Annotations from different levels may overlap.
- All annotation should consider a degree of imprecision and vague assertions have to be marked. Otherwise interpretations of doubtful events are falsified by crisp yes/no decisions. Vagueness and uncertainty may lead to different branches of the same annotation base.
- Original text and transliteration have to be both kept and synchronised.
- Historical texts lack digital resources. Historical language requires more features for annotation than modern ones. Thus a fully automatic (linguistic) annotation is in many cases impossible. Manual annotation is time con-

suming, so that functions allowing a controlled semi-automatisation of the annotation process is more than desirable.

- The annotation tool has to be user-friendly as annotators do not have often deep IT-skills

As none of the current widespread annotation tools (Bollman & Petran & Dipper 2014), (de Castilho et. Al. 2016) fulfils all criteria above, many projects alter the data model, i.e. features of language or of the text respectively domain are not included in the annotation model. This has consequences on the analysis and interpretation process.

In this paper we will introduce a novel framework for data modelling which allows implementation of tailored annotation tools for the specific DH-project. We will illustrate the generic framework model by mean of three examples from completely different domains each treating another language: the construction of a diachronic corpus for classical Ethiopic texts (Vertan & Ellwardt & Hummel 2016); the annotation of classical Maya database of inscriptions and texts and the computer-based analysis of original and translation in three languages of historical documents from the 18<sup>th</sup> century (Vertan & v. Hahn & Dinu 2017). We will present the generic model and show the derived data model for each of the 3 examples and we will discuss the challenges implied by the development of a new software. We will illustrate also how interchangeability with other digital resourced is secured.

Furthermore, we will show how this framework can be used as well for the annotation of linguistic and factual vagueness in texts.

The aim of such annotations is not to develop an expert system in the classical way from Artificial Intelligence. Such expert system assumes that the computer is reasoning and presents its interpretation to the user. We consider that for interpretation of historical facts such system is not reliable. The background knowledge necessary for producing reliable result is huge and relies often either on materials which are not in digital form. Thus our goal is more to make the user aware that:

- There is a bunch of possible answers to one query and
- These possible answers may have different degree of reliability (i.e .they are not for sure true).

The interpretation and the final decision is left entirely to the user.

## Acknowledgements

This article presents work performed within the project HerCoRe (Hermeneutic and Computer – based Analysis of Reliability, Consistency and Vagueness in historical Texts) funded by the Volkswagen Foundation within the framework “Mixed Methods in Humanities). Works reported in this paper was done in collaboration with Walther v. Hahn and Alptug Güney.

## References

- Bollmann, Marcel and Petran, Florian and Dipper, Stefanie and Krasselt, Julia, 2014: „*CorA: A web-based annotation tool for historical and other non-standard language data*“, in: Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH). Gothenburg, Sweden, 86-90.
- Eckart de Castilho and Richard Mújdricza-Maydt, Éva and Yimam Seid Muhie and Hartmann, Silvana and Gurevych, Iryna and Frank, Anette and Biemann, Chris, 2016: „*A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures*“, in: Proceedings of the LT4DH workshop at COLING 2016, Osaka, Japan: 76-84.
- Vertan, Cristina and Ellwardt, Andreas / Hummel, Susanne, 2016: "Ein Mehrebenen-Tagging-Modell für die Annotation altäthiopischer Texte", in: Proceedings der DHd-Konferenz 2016 <http://www.dhd2016.de/abstracts/vortr%C3%A4ge-061.html> [last access 25.09.2017].
- Vertan, Cristina and Hahn, Walther von and Dinu, Anca, 2017: „*On the annotation of vague expressions: a case study on Romanian historical texts*“, Proceedings of the first Workshop on Language Technology for Digital Humanities in Central and (South-) Eastern Europe, in association with RANLP 2017, Varna, Bulgaria: 24-31
- Alfred. V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling, volume 1*. Prentice-Hall, Englewood Cliffs, NJ.
- American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

- Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. *Alternation*. *Journal of the Association for Computing Machinery*, 28(1):114-133. <https://doi.org/10.1145/322234.32224>. Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503-512.
- James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. *Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing*. In *Proceedings of the 54<sup>th</sup> Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1–11. <https://doi.org/10.18653/v1/P16-1001>.
- Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.
- Mary Harper. 2014. *Learning from 26 languages: Program management and science in the babel program*. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, page 1. <http://aclweb.org/anthology/C14-1001>.
- Alexander V. Mamishev and Murray Sargent. 2013. *Creating Research and Scientific Documents Using Microsoft Word*. Microsoft Press, Redmond, WA.
- Alexander V. Mamishev and Sean D. Williams. 2010. *Technical Writing for Teams: The STREAM Tools Handbook*. Wiley-IEEE Press, Hoboken, NJ.

# Navigating Literary Text with Word Embeddings and Semantic Lexicons

Susan Leavy<sup>1</sup>, Karen Wade<sup>2</sup>, Gerardine Meaney<sup>2</sup>, and Derek Greene<sup>1</sup>

<sup>2</sup>Humanities Institute, University College Dublin, Ireland

<sup>1</sup>Insight Centre for Data Analytics, University College Dublin, Ireland  
{*susan.leavy,karen.wade,gerardine.meaney,derek.greene*}@ucd.ie

## Abstract

Word embeddings represent a powerful tool for mining the vocabularies of literary and historical text. However, there is little research demonstrating appropriate strategies for representing text and setting parameters, when constructing embedding models within a digital humanities context. In this paper we examine the effects of these choices using a case study involving 18th and 19th century texts from the British Library. The study demonstrates the importance of examining implicit assumptions around default strategies, when using embeddings with literary texts and highlights the potential of quantitative analysis to inform critical analysis.

## 1 Introduction

This research is part of a digital humanities project exploring attitudes towards disease and illness in the 18th and 19th centuries. The associated corpus contains a large, diverse selection of digitised texts. Lexicons generated using word embeddings are part of a suite of big data approaches which are applied in order to navigate this corpus, which consists of over 46,000 texts dedicated to a range of subjects. It is hoped that these techniques will allow the identification of key texts and thematic trends concerned with illness and disease, so that these can be interpreted with reference to current and historical debates surrounding biopolitics, medical culture, and migration.

Word embeddings are increasingly being used to generate semantic lexicons for a variety of tasks (Mikolov et al., 2013). This includes uncovering changes in the sense of terms over time (Hamilton et al., 2016), extracting social networks from literary texts (Wohlgenannt et al., 2016), and text clas-

sification (Leavy et al., 2017). However, there is a lack of research demonstrating optimal strategies for setting parameters, when constructing these models on literary and historical texts. There has also been little study on the effect of text preprocessing decisions on the resulting models (Lapesa and Evert, 2014; Camacho-Collados and Pilehvar, 2017). Given that the assumptions behind preprocessing can have particular significance within a digital humanities context, it is important to explore the impact of these decisions, which is often not reported or considered (Sculley and Pasanek, 2008).

This research evaluates the setting of parameters in word embedding along with standard preprocessing approaches including conversion of all letters to lowercase and removal of stop-words. Evaluation of lexicons generated using word embedding is commonly conducted using an intrinsic approach whereby the resulting lexicons are evaluated against existing standard lexical databases such as WordNet (Miller, 1995). However, given the domain specificity and historical nature of this corpus, extrinsic evaluation was conducted based on the effectiveness of each lexicon in identifying texts that relate to medical topics.

## 2 Methodology

The corpus used in this research is comprised of a diverse collection of digital texts from the British Library. In the analysis described here, we focus on a subset of 35,916 English language texts, dating from 1700 to 1899. Word2vec Continuous Bag-of-Words (CBOW) and Skip-Gram (SG) embedding models were generated from the English corpus using 30 combinations of parameters and text processing strategies (see Table 1).

A set of 10 seed terms was derived from a 19th century medical reference book (Guy, 1856),

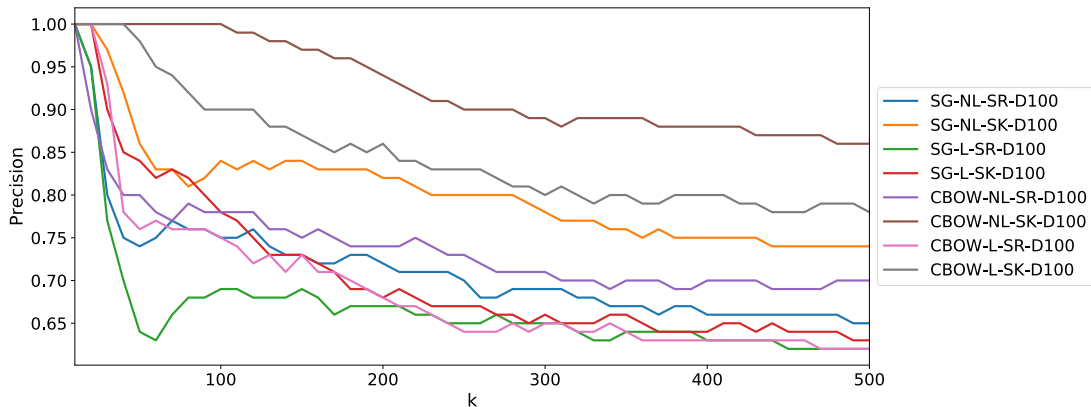


Figure 1: Precision@ $k$  results for models with  $D = 100$  dimensions. Model labels are indicated as: Model (SG/CBOW), Lower / Not Lowercase (L/NL), Stop-word Remove/Keep (SR/SK).

| Parameters    | Values                |
|---------------|-----------------------|
| Model         | Skip-gram / CBOW      |
| Dimension     | 25 / 100 / 400 / 800  |
| Preprocessing | Lowercase / Stopwords |

Table 1: Text preprocessing strategies and parameters evaluated.

along with initial close reading of the corpus:

health, disease, physic, physiology, pathology, therapeutics, remedies, medicine, physician, medical

For each embedding model, we extracted the top 20 terms that were most similar to the seed words used, and used these to build lexicons.

The evaluation of the lexicons involved using them as a basis for ranked document retrieval. We use a sample of 19,290 documents, each representing a labelled fixed-length excerpt from a text in the overall English corpus. Of these, 20% were labelled as medical texts. Texts were ranked according to the frequency of occurrence of terms from each generated lexicon. The quality of the lexicon was evaluated based on whether this ranking of documents surfaced texts related to medical topics. In this project, given the objective of enabling close reading of the retrieved texts within an exploratory interface, the precision of the returned results was of prime importance and evaluation was based on the level of precision relative to the top- $k$  ranked texts (i.e. precision@ $k$ ).

### 3 Findings and Analysis

Before considering document retrieval, we looked at the overall level of agreement between the lexicons generated by the models, by measuring their Jaccard set similarity for all 10 seed terms. We see

a surprisingly low level of agreement between the lexicons – mean 0.31 and median 0.29.

Next, we measured the precision of the retrieval of medical documents for rankings of size  $k \in [10, 500]$ . The subset of results shown in Figure 1 reveal patterns indicating the importance of the choice of parameters and settings when generating word embeddings for literary and historical texts. Contrary to standard practice, not converting all text to lowercase and retaining stop-words resulted in better performance. This demonstrates that established standards for preprocessing modern texts may not produce the best results in a digital humanities context.

Error analysis, in the form of close reading, was conducted where strategies resulted in significantly lower precision (e.g. see SG-L-SR-D100 in Figure 1). These results appear to be due to the retrieval of country reports that, while they were not medical texts, contained a wealth of relevant information on medical care. This demonstrates how in a digital humanities project, error analysis can provide new information to prompt reformulation of the original research hypotheses.

### 4 Conclusion

This paper explores issues around selecting an appropriate strategy for using word embeddings to construct semantic lexicons for literary and historical texts. Established default strategies often emerge in response to requirements from different domains. However, this work shows the importance of evaluating the assumptions behind established strategies, and considering the specific requirements of individual digital humanities projects.

## References

- Jose Camacho-Collados and Mohammad Taher Pilehvar. 2017. On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. *arXiv preprint arXiv:1707.01780*.
- William August Guy. 1856. *Hooper's Physician's Vade Mecum*.
- William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proc. EMNLP 2016*. NIH Public Access, volume 2016, page 595.
- Gabriella Lapesa and Stefan Evert. 2014. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association of Computational Linguistics* 2(1):531–545.
- Susan Leavy, Mark T Keane, and Emilie Pine. 2017. Mining the cultural memory of irish industrial schools using word embedding and text classification. In *Proc. Digital Humanities 2017*.
- Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Proc. ICLR 2013* pages 1–12.
- George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.
- D Sculley and Bradley M Pasanek. 2008. Meaning and mining: the impact of implicit assumptions in data mining for the humanities. *Literary and Linguistic Computing* 23(4):409–424.
- Gerhard Wohlgenannt, Ekaterina Chernyak, and Dmitry Ilvovsky. 2016. Extracting social networks from literary text with word embedding tools. In *Proc. Workshop on Language Technology Resources and Tools for Digital Humanities*. pages 18–25.



# Map visualization and quantification of literary places in a Spanish Corpus

José Luis Losada Palenzuela  
University of Wrocław

## 1 Introduction

With this communication I would like to address the possibilities of digital technologies for quantification and visualization of the narrative space in fictional texts. The analysis is based on a corpus of early modern Spanish Byzantine novels (16th-17th centuries). I would like to reflect around the procedure of making *post-authorial* maps (Bushell, 2012) in order to use them as an analytical tool and as part of scholarly digital edition.

## 2 Theoretical framework and corpus

Literary geography is an interdisciplinary crossroad of literary theory, geography, digital cartography and spatial analysis (Piatti/Hurni, 2011), (Bodenhamer et al., 2010). I will focus on the fictional literary geography: that is, the digital visualization of the spaces of fiction; both the quantitative analysis and the visualization of places and itineraries in the novels are taken into account, having in mind Moretti's work (1998, 2007), but trying partially to reproduce some methodological approaches of Barbara Piatti in her work *Die Geographie der Literatur* (2008), where she plotted on analogical maps the space action of some Swiss literary works.

The corpus belongs to the Byzantine genre, in which the chronotope is defined by sea travels, pirates, exoticism, shipwrecks, transcultural encounters. Cervantes' posthumous novel *The Trials of Persiles and Sigismunda. A Northern History* (1617) represents an important shift in the spatial genre classification, because unlike Heliodorus' Mediterranean model, Cervantes chooses a couple of Nordic lovers and

describes their pilgrimage along a North-South axis, starting at the Nordic countries and ending in Rome. Early modern sequels seem to imitate Cervantes' settings and characters by placing its actions in Muscovy, England or Poland, but without abandoning the Mediterranean basin as a setting area for the diegesis. On the other hand some novels have been criticized as lacking spatial coherence, geographical accuracy or verisimilitude, but the apparently spatial gibberish is based, however, on geographic and cartographic sources which determine its narrative organization. This is defined as *kartographisches Schreiben* (Dünne 2011), that is, maps such as *Carta Marina* (in the case of the Northern geography in Cervantes' *Persiles*) or *Theatrum Orbis Terrarum* (for Zuñigas's *Semprilis*) are used to configure spatial and topographical references in the novels (Losada, 2016). This artistic procedure connects texts and maps, and gives to the novels a certain affinity to cartographic representation called by Stockhammer (2007) *literarische Kartizität*.

The main question is whether a quantitative analysis of places and digital map visualization could shed light on a spatial distribution that defines the genre.

## 3 Methodological framework

As a proof of concept the corpus has been limited to a few works, in which the places have been automatically extracted with Stanford<sup>1</sup> and Freeling<sup>2</sup> Named Entity Recognizers. In order to

<sup>1</sup> Stanford Named Entity Recognizer, version 3.8.0, language models 2017/09/06 (Spanish, English). See at <https://nlp.stanford.edu/software/CRF-NER.html>

<sup>2</sup> FreeLing, version 4.0. See at <http://nlp.lsi.upc.edu/freeling>



automate the queries for a vector of places and be able to work within the R environment I adapted, forked from the ggmap package, my own package in R (editio/georeference<sup>3</sup>), which allows to geolocate places from three different gazetteers: Pelagios<sup>4</sup>, GeoNames<sup>5</sup>, and Wikipedia (georeferenced articles stored in the GeoNames database), gazetteers more suitable for historical and literary texts. For the proper georeference and tiling of historical maps I used Qgis<sup>6</sup>, and for the visualizations I used Leaflet for R<sup>7</sup>, which is, in part, available as a package in R, so it has the advantage of using just one environment for data processing, visualization, overlay of historic maps and export capabilities.

#### 4 Map visualization

The several visualizations plotted on a map (places most frequently mentioned, coverage of cluster's bounds, proximity clustering, places shared by novels, weighted importance for each place) correspond roughly with the knowledge we have about those novels, so that the insights into the corpus are not totally incorrect despite of some methodological and technical caveats: Firstly, reducing fiction to invented events in real places (Stockhammer, 2013) and ignoring the enormous topographical variance in literature (Piatti 2008) as well as different zones of narrative action, digressions, diegesis, etc. Secondly the accuracy of NERs are deeply limited by language (Bornet and Kaplan, 2017), by toponym variance or spelling in a Spanish Early Modern corpus, together with the fact that the automated geolocation by gazetteers has, as well, its limits (gazetteers automated returns in the corpus are around 60% of all places found by NERs).

On the other hand, to prove the dependence of geographical and cartographic sources, historical maps (Ortelius' map of *Fessae, et Marocchi*, and the historical borders of the Kingdom of Poland) have been overlaid together with the itinerary of the lead characters in the novel *Semprilis*. The visualizations will raise the question of whether the author could create the narrative space relying on this particular sources.

<sup>3</sup> See at <https://github.com/editio/georeference>

<sup>4</sup> See at <http://commons.pelagios.org>

<sup>5</sup> See at <http://www.geonames.org>

<sup>6</sup> See at <https://www.qgis.org>

<sup>7</sup> See at <https://github.com/rstudio/leaflet>

#### 5 Conclusion

The quantitative approach may fall short in explaining how the Byzantine genre operates spatially, but a visualization enriched with historical maps can, in this particular case, add knowledge to the analysis of the literary space.

#### References

- Bodenhamer, D. J., Corrigan, J., Harris, T. M. (eds.). (2010). *The Spatial Humanities. GIS and the Future of Humanities Scholarship*, Indiana University Press, Bloomington.
- Bornet, C., & Kaplan, F. (2017). A Simple Set of Rules for Characters and Place Recognition in French Novels. *Frontiers in Digital Humanities*, 4. <https://doi.org/10.3389/fdigh.2017.00006>
- Bushell, S. (2012). The Slipperiness of Literary Maps: Critical Cartography and Literary Cartography. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 47(3), 149-160. <https://doi.org/10.3138/cart0.47.3.1202>
- Dünne, J. (2011). *Die kartographische Imagination: Erinnern, Erzählen und Fingieren in der Frühen Neuzeit*. München: Fink.
- Losada Palenzuela, J. L. (2017). Desplazamiento de la imagen septentrional: Polonia en La historia de las fortunas de Semprilis y Genorodano. In H. Ehrlicher & J. Dünne (Eds.), *Ficciones entre mundos. Nuevas lecturas de «Los Trabajos de Persiles y Sigismunda» de Miguel de Cervantes* (pp. 253-273). Kassel: Reichenberger.
- Moretti, F. (1998). *Atlas of the European Novel, 1800-1900*. Verso.
- Moretti, F. (2007). *Graphs, Maps and Trees. Abstract Models for Literary History* (2.<sup>a</sup> ed.). London, New York: Verso.
- Piatti, B. (2009). *Die Geographie der Literatur. Schauplätze, Handlungsräume, Raumphantasien* (2.<sup>a</sup> ed.). Göttingen: Wallstein-Verlag.
- Reuschel, A.-K., & Hurni, L. (2011). Mapping Literature: Visualisation of Spatial Uncertainty in Fiction. *The Cartographic Journal*, 48(4), 293-308. <https://doi.org/10.1179/1743277411Y.0000000023>
- Stockhammer, R. (2007). *Kartierung der Erde: Macht und Lust in Karten und Literatur*. München: Fink.
- Stockhammer, R. (2013). Exokeanismo: the (un)mappability of literature. *Primerjalna Knjizevnost*, 36(2), 123-138.

# Toward a Tool for Sentiment Analysis for German Historic Plays

**Thomas Schmidt**

Media Informatics Group  
Regensburg University  
93040 Regensburg, Germany  
thomas.schmidt@ur.de

**Manuel Burghardt**

Computational Humanities Department  
Leipzig University  
04109 Leipzig, Germany  
burghardt@informatik.uni-leipzig.de

With the availability of large amounts of opinionated data through the Internet (social networks, online forums, product reviews, etc.), computational sentiment analysis has become popular in the early 2000s, especially in the context of social media and online reviews (Liu, 2016). Recently sentiment analysis has also found applications in the digital humanities, most notably in the field of literary studies. Sentiment analysis is used for genre classification (Kim et al., 2017), to investigate shifts in the meaning of words (Buechel et al., 2016), to predict the success of novels (Ashok et al., 2013), or to analyse fairy tales (Alm et al., 2005), novels (Kakkonen & Kakkonen, 2011; Jockers, 2015; Jannidis et al., 2016) and drama (Mohammad, 2011; Nalisnick & Baird, 2013). Many of the current projects in this domain use sentiment lexicons. A sentiment lexicon is a list of words with sentiment annotations (positive/negative values). These words are typically referred to as *sentiment bearing words* (SBW). By adding up the number of positive words and subtracting the number of negative words (or polarity annotations on a metric scale), the overall polarity of a text unit can be calculated (Kennedy & Inkpen, 2006).

We present a project on the exploration of different lexicon-based sentiment analysis techniques for the domain of historic, German drama texts, more concretely on a corpus of Lessing’s plays. The corpus is composed of twelve plays and was obtained from the *TextGrid*<sup>1</sup> platform. As historic German texts that, at the same time, also use poetic language challenge standard sentiment analysis lexicons, we conducted a systematic evaluation study, to investigate which configuration of dictionaries and NLP tools yields the best results.

We evaluated several combinations of sentiment lexicons and optimization steps:

- Five existing sentiment dictionaries (Remus et al., 2010; Vo et al., 2009; Mohammad &

Turney, 2010; Clematide & Klenner, 2010; Waltinger, 2010) for present German, as well as an accumulated combination of all lexicons were evaluated;

- The extension of each of the above lexicons with historical linguistic variants (Jurish, 2012) was evaluated;
- Different types of stopword lists und lists of most frequent words of the corpus (cf. Saif et al., 2014) were evaluated;
- Lemmatization with the pattern lemmatizer (De Smedt & Daelemans, 2012) and the treetagger (Schmid, 1995) was evaluated;

We evaluated the different configurations against a gold standard corpus of 200 single speeches of our corpus. This method of evaluation can be considered rather unique in this branch of sentiment research, as results are typically evaluated by comparing them to well-known observations that are already available from other, oftentimes hermeneutic, scholarly work (cf. Mohammad, 2011; Nalisnick & Baird, 2013).

The gold standard was created in a preliminary annotation study. Five annotators (all fluent in German language) annotated the polarity (positive or negative) of the character speeches. The annotation of the majority of the annotators defines the final polarity of a speech. The measure of agreement between the annotators point to a mediocre agreement (Fleiss’ kappa = 0.47; overall agreement in percent = 77%). These results are in line with related studies in the context of narrative texts (Alm & Sproat, 2005). The final gold standard corpus consists of 139 negative und 61 positive speeches.

We compared the performance of all aforementioned combinations of sentiment and NLP techniques by calculating the overall polarity and by analyzing typical performance metrics such as the accuracy (Gonçalves et al., 2013). During the evaluation study, we found that

<sup>1</sup> <https://textgridrep.org/> (note: all URLs mentioned in this article were last visited April 13, 2018)

- the extension of lexicons with historical linguistic variants and lemmas yields the highest performance boost,
- lexicons with polarity scales (e.g. from -1 to 1) instead of nominal sentiment-annotations (neg/pos) yield consistently better results,
- lexicons that come with explicit lemma and flexion forms typically perform better than generic lemmatization tools.

Going through all the metrics, we identified the following combination of techniques as the setup with the best overall performance:

- SentiWS lexicon (Remus et al., 2010),
- no stopword lists,
- pattern lemmatizer,
- extension with historical linguistic variants;

With an overall accuracy of 67%, the performance is above the random baseline, but still considerably worse than in other domains of sentiment analysis (cf. Vinodhini & Chandrasekran, 2012). However, since we use very basic lexicon-based sentiment analysis techniques and the human annotators who produced the gold standard also had severe problems and disagreements concerning the sentiment annotations, we consider these results as promising. We also found that the lower the agreement between annotators for a speech the more likely the sentiment analysis predicts a wrong class. Furthermore, for the gold standard annotation, annotators could only choose between positive and negative; annotations like neutral or mixed were not possible, which aggravates the annotation as well as the automatic prediction. However, other results of our annotation study show that these classes are indeed relevant for our corpus.

To further investigate the possibilities of sentiment analysis in German drama texts, we developed a web application<sup>2</sup> that can be used to explore the results of our current project. Users are able to analyze sentiment progressions and sentiment distributions on several different levels. The structural levels of analysis are the whole drama, single acts, scenes and speeches. Furthermore, by

<sup>2</sup>[http://lauchblatt.github.io/QuantitativeDramenanalyseDH2015/FrontEnd/sa\\_selection.html](http://lauchblatt.github.io/QuantitativeDramenanalyseDH2015/FrontEnd/sa_selection.html)

accumulating the speeches of single speakers, users can explore sentiment processes and distributions of specific characters. By using a heuristic described in Nalisnick and Baird (2013), we also integrated sentiment relationships of speakers. Sentiments of speakers and speaker relationships can be analyzed on all structural levels. Besides polarities (positive/negative), we also integrated our results on eight basic emotions as implemented in the NRC Emotion Lexicon (Mohammad & Turney, 2010). To allow for comparisons (e.g. between scenes), users can choose to normalize the results by the number of all words or SBWs.

We are currently working together with literary scholars to further explore requirements for computer-based sentiment analysis in literary studies. We also started a project to acquire more manually annotated data in the context of German historic plays and are also integrating more polarity classes like neutral and mixed in the annotation process. We are planning to use this data for more exact evaluations of the lexicon approach, but also as training data for machine learning approaches to sentiment analysis. Furthermore, we want to extend our current corpus beyond the scope of Lessing's plays, to enable comparisons of authors, genres and periods.

## References

- Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 579-586). Association for Computational Linguistics
- Alm, C. O. & Sproat, R. (2005). Emotional sequencing and development in fairy tales. In *International Conference on Affective Computing and Intelligent Interaction* (pp. 668-674). Springer Berlin Heidelberg.
- Ashok, V. G., Feng, S., & Choi, Y. (2013). Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1753-1764).
- Buechel, S., Hellrich, J., & Hahn, U. (2016). Feelings from the Past—Adapting Affective Lexicons for Historical Emotion Analysis. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)* (pp. 54-61).
- Clematide, S. & Klenner, M. (2010). Evaluation and extension of a polarity lexicon for German. In *Proceedings of the First Workshop on Computational*

- Approaches to Subjectivity and Sentiment Analysis* (pp. 7-13).
- De Smedt, T. & Daelemans, W. (2012). *Pattern for Python*. *Journal of Machine Learning Research*, 13: 2031–2035.
- Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks* (pp. 27-38). ACM.
- Jannidis, F., Reger, I., Zehe, A., Becker, M., Hettlinger, L. & Hotho, A. (2016). *Analyzing Features for the Detection of Happy Endings in German Novels*. arXiv preprint arXiv:1611.09028.
- Jockers, M. L. (2015). Revealing sentiment and plot arcs with the syuzhet package. Retrieved from <http://www.matthewjockers.net/2015/02/02/syuzhet/>
- Jurish, B. (2012). *Finite-state Canonicalization Techniques for Historical German*. PhD thesis, Universität Potsdam (defended 2011). URN urn:nbn:de:kobv:517-opus-55789.
- Kakkonen, T. & Kakkonen, G. G. (2011). SentiProfiler: creating comparable visual profiles of sentimental content in texts. In *Proceedings of Language Technologies for Digital Humanities and Cultural Heritage* (pp. 62-69).
- Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2), 110-125.
- Kim, E., Padó, S., & Klinger, R. (2017). Prototypical Emotion Developments in Literary Genres. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 17–26).
- Liu, B. (2016). *Sentiment Analysis. Mining Opinions, Sentiments and Emotions*. New York: Cambridge University Press.
- Mohammad, S. (2011). From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 105-114). Association for Computational Linguistics.
- Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 26-34). Association for Computational Linguistics.
- Nalisnick, E. T. & Baird, H. S. (2013). Character-to-character sentiment analysis in shakespeare's plays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 479–483).
- Remus, R., Quasthoff, U. & Heyer, G. (2010). SentiWS-A Publicly Available German-language Resource for Sentiment Analysis. In *LREC* (pp. 1168-1171).
- Saif, H., Fernandez, M., He, Y., Alani, H. (2014). On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter. In: *Proc. 9th Language Resources and Evaluation Conference (LREC)* (pp. 810-817).
- Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*.
- Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(6), 282-292.
- Võ, M. L., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin affective word list reloaded (BAWL-R). *Behavior research methods*, 41(2), 534-538.
- Waltinger, U. (2010). Sentiment Analysis Reloaded: A Comparative Study On Sentiment Polarity Identification Combining Machine Learning And Subjectivity Features. In *Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST '10)*

# Modeling Thematic Structure in Holiday Postcards

Kyoko Sugisaki, Nicolas Wiedmer, Marcel Naef, Heiko Hausendorf

German department, University of Zurich  
Schönberggasse 9, 8001 Zurich  
Switzerland

{sugisaki,nicolas.wiedmer,marcel.naef,heiko.hausendorf}@ds.uzh.ch

In this work, we present unsupervised and supervised machine learning methods and corpus-linguistic methods to model thematic structure in holiday postcards. We consider thematic structure from a point of view of text linguistics. Accordingly, a text consists of themes that are introduced, continued and terminated in the course of discourses (cf. Brinker [p. 24ff](2014)). So far, the automatic recognition of theme has been carried out mainly based on information structure, such as the Prague Treebank (Hajič, 1998) and the Potsdam Commentary Corpus (Stede and Mampin, 2016). In our work, the term *theme* is distinguished from the notion of *topic* in information structure, that is, ‘aboutness’ and ‘old/given entity’. In information structure, the topic is determined mainly by its syntactic position in a sentence and by the salience of its discourse in relation to the entities mentioned in the previous sentence(s). In contrast, in our work *theme* is rather a semantic frame that constitutes the thematic coherence of a certain genre of text (here: holiday picture postcards). We use the term semantic frame in the sense of Busse (2012, p. 563) who defines the frame as a structure of knowledge, in which the core of a frame is connected to the constituents of knowledge. Depending on the context of a concrete situation, possible constituents vary. These constituents define the conditions of the realisation of textual phrases. In the case of holiday picture postcards, the frame is *to be on holiday*. The constituents of knowledge (i.e. slots) can be filled with actual text (i.e. fillers) according to the concrete situation of writing a holiday postcard. For instance, a slot can be weather in holiday postcards, whose possible filler is, for example, “*hier regnet es*” (‘here, it is raining’) or “*Am 1. Oktober tragen wir kurze Hosen!*” (‘We are wearing short trousers on the first of October!’).

The primary goal of our linguistic research is

to find the core thematic structures (or slots) of the holiday postcards and their development over time. To this end, we have defined the categories of the super themes that potentially remain consistent over time (cf. Hausendorf and Kesselheim (2008, p. 103)); Hausendorf (2008, p. 333); Hausendorf (2009, p. 13)). Then, we extracted over 1000 holiday postcards from the *Ansichtskartenkorpus* ([anko] ‘picture postcard corpus’ (Sugisaki et al., 2018)) and annotated them according to this schema. In this workshop, we show how we have identified and annotated our thematic categories, and demonstrate various data-driven supervised and unsupervised methods such as topic models (Blei, 2013) and word embedding (Mikolov et al., 2013) as well as corpus-linguistic methods to automatically categorise sentences of holiday postcards in the *Ansichtskartenkorpus* into the thematic categories. We also discuss whether these computational methods can be a viable model for modelling semantic-thematic structures in holiday postcards. By doing so, we compare the automatic analysis with the manual analysis, and discuss whether the computational methods are compatible to human interpretation. We also investigate into whether the automatic analysis is able to identify previously undiscovered thematic patterns.

**Acknowledgments** This work has been funded under SNSF grant 160238. We thank all the project members, Joachim Scharloth, Noah Bubenhofer, Maaïke Kellenberger, David Koch, Dewi Josephine Obert, Jan Langenhorst.

## References

- David M. Blei. 2013. Topic modeling and digital humanities. *Journal of Digital Humanities*.
- Klaus Brinker, Hermann Cölfen, and Steffen Pappert.

2014. *Linguistische Textanalyse: eine Einführung in Grundbegriffe und Methoden*. Erich Schmidt Verlag, Berlin, 8., neu bearb. und erw. aufl edition.
- Dietrich Busse. 2012. *Frame-Semantik. Ein Kompendium*. De Gruyter, Berlin, Boston.
- Jan Hajič. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In E. Hajičová, editor, *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, Karolinum, Charles University Press, Prague, Czech Republic, pages 106–132.
- Heiko Hausendorf. 2008. Zwischen Linguistik und Literaturwissenschaft: Textualität revisited. Mit Illustrationen aus der Welt der Urlaubsansichtskarte. *Zeitschrift für germanistische Linguistik (ZGL)* 36(3):319–342.
- Heiko Hausendorf. 2009. Kleine Texte. Über Randerscheinungen von Textualität. *Germanistik in der Schweiz. Online-Zeitschrift der Schweizer Akademischen Gesellschaft für Germanistik* 6.
- Heiko Hausendorf and Wolfgang Kesselheim. 2008. *Textlinguistik fürs Examen*. Vandenhoeck & Ruprecht, Göttingen, Germany.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Manfred Stede and Sara Mamprin. 2016. Information structure in the Potsdam Commentary Corpus: Topics. In *Proceeding of the 9th International Conference on Language Resources and Evaluation (LREC 2016)*.
- Kyoko Sugisaki, Nicolas Wiedmer, and Heiko Hausendorf. 2018. ANKO – a picture postcard corpus: Transcription, annotation and part-of-speech tagging. In *Proceeding of the 11th International Conference on Language Resources and Evaluation (LREC'18)*. pages 255–259.