exhibitions, artworks series, literary works and agents in their main life events, roles, places, and dates.

WRITE data model documentation and some representative case studies (one artwork for each collection) are available to foster the understanding and reusability of the model. The contribution of the dataset is twofold: first, it supports scholars in investigating the WRITE domain by recording and classifying its data to allow a systematic analysis of its resources to help the domain experts answer their main research questions (e.g. the definition of a spectrum of forms and practices ranging from traditional calligraphy to modern/contemporary artistic expression: can those forms/practices still be defined as calligraphy?) and, hopefully, discover new knowledge; second, it is a challenge in the modelling of different sources from GLAM by categorizing and connecting the visual dimension of iconographic elements (calligraphy as visual artwork) with the respective textual elements (lexical features, text transcription and translation, literary sources of text) to attribute an actual meaning to the new concept of “calli-writing units” through the analysis of the hybridization of multimedia resources.

Bibliography


Notes

1. See https://writecalligraphyproject.eu/
2. The project WRITE - New Forms of Calligraphy in China: A Contemporary Culture Mirror is an European Research Council (ERC) Starting Grant funded project based in the Department of Interpreting and Translation of the Alma Mater Studiorum – University of Bologna (GA n. 949645).
3. See https://www.britishmuseum.org/search?search_api_fulltext=chinese+calligraphy
4. See https://www.mediawiki.org/wiki/Wikibase/DataModel
5. OmekaS is a web publishing platform to collaboratively create collections with a shared pool of online resources.
6. See https://write-dataset.github.io/documentation/

Extracting and providing online access to annotated and semantically enriched historical data. The AGODA project

Puren, Marie Anna
marie.puren@epitech.eu
Epitech, MNSHS, France; Centre Jean-Mabillon, Ecole nationale des chartes, France

Vernus, Pierre
pierre.vernus@msh-lse.fr
LARHRA, France; Université Lyon 2, France
The AGODA project \(^1\) (Puren and Vernus, 2021) is one of five pilot projects supported by the DataLab of the Bibliothèque nationale de France. It aims to create an online platform facilitating the exploration and use of the parliamentary debates of the Chamber of Deputies published in the *Journal officiel* from 1881 to 1940. In the framework of the DataLab, we are working on a test sub-corpus, namely the parliamentary cycle from 1889 to 1893, to test our hypotheses on a smaller dataset.

Over the past sixty years, a great deal of work has been done on parliamentary debates (Chester and Bowring, 1962; Franklin and Norton, 1993). It is indeed a valuable source for historians (Marnot, 2000; Ouellet and Roussel-Beaulieu, 2003; Ihalainen, 2016; Lemercier, 2021), political scientists (Van Dijk, 2010), sociologists (Cheng, 2015) or linguists (de Galembert et al., 2013; Hirst et al., 2014; Rheault et al., 2016). Access to digitised and OCRised debates thus seems to have a positive effect on the number of historical works using these documents (Mela et al., 2022). The same effect can be observed for other disciplines using contemporary debates (Fišer et al., 2018; Fišer et al., 2020). AGODA is thus part of a wider movement to facilitate the use and analysis of parliamentary data, following the example of ParlaClarin (Fišer and Lenardič, 2018) and ParlaMint (Erjavec et al., 2022a; Erjavec et al., 2022b), which propose to produce comparable and multilingual Parliamentary Proceedings Corpora according to the XML-TEI standard. Naomi Truan has also produced a corpus of parliamentary debates encoded in XML-TEI (Truan, 2016; Truan and Romary, 2021). The production of this type of resource facilitates the publication of works exploiting this data to better understand French political discourse (Diwersy et al., 2018; Blaette et al., 2020; Diwersy and Luxardo, 2020).

Between 1881 and 1899, 2596 issues of the *Journal Officiel* were published (50791 JPG images). The debates are also in TXT format but put online without extensive post-correction: the quality of the OCR is not sufficient to provide a satisfactory online browsing experience, and it could have a negative impact on the analyses performed on these texts (van Strien, 2020). Therefore, we chose to OCRise the text, to obtain a better-quality result. We use the PERO OCR (Kodym and Hradiš, 2021; Kohút and Hradiš, 2021; Kišš et al., 2021) based solution developed by the SODUCO project \(^2\). This tool, still in private alpha version, has been used to prepare the data in (Abadie et al., 2022) that will be accessible via Zenodo.

OCRised texts are obtained in JSON format; we are developing Python scripts to convert this output into an XML file corresponding to the chosen TEI model. This model is formalised with an adapted XML schema, created using an ODD (Rahtz and Burnard, 2013). We chose to use the ODD created by ParlaClarin (Erjavec and Pančur, 2021) which can be easily adapted to annotate historical parliamentary debates. In the case of France, the rules for transcribing debates were set in the 19th century; thus, the recordings of today's debates are very similar to those produced during the Third Republic. The TEI-encoded corpus will be stored in an eXist-db database, and it will be visualised using the TEI Publisher application, which can transform the source data into HTML web pages. The parliamentary debates will thus be made available to online users as a digital edition and integrated into an application context.

We will also present the first analyses we have carried out on this corpus with "bag-of-words" techniques - these being not too sensitive to the quality of the OCR. We first used topic modelling, an unsupervised learning method that allows us to discover the latent semantic structures of a corpus of texts, without using semantic and lexical resources (Blei et al., 2003). This method is well suited to study parliamentary debates (Bourgeois et al., 2022).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{distribution.png}
\caption{Distribution of four different topics over time.} \label{fig:topic_distribution}
\end{figure}

Alternatively, we can use word embeddings to reduce the dimension of the original space from several tens of thousands of forms to a hundred axes, and then apply classical data science tools such as clustering or correlation analysis on the reduced space (Mikolov et al., 2013). Word embedding has thus shown its interest in the study of parliamentary debates (Rheault and Cochrane, 2020). We used a continuous bag-of-words model for dimension reduction and an unsupervised classification algorithm - in this case DBSCAN - to group words into clusters.
t-SNE projection of the centroids of the clusters

Bibliography


Traven between the impostors.

Preliminary considerations on an authorship verification case

Rebora, Simone
simone.rebora@univr.it
University of Verona, Italy

Salgaro, Massimo
massimo.salgaro@univr.it
University of Verona, Italy

This paper sets up the groundwork for an authorship verification project dedicated to the German novelist B. Traven, author of novels such as *The Death Ship* (1926) and *The Treasure of the Sierra Madre* (1927), whose real identity is still a mystery. Among the different theories, the most established is the one that sees B. Traven as the pseudonym of Otto Feige, author of a series of political pamphlets for the metal workers' union in Gelsenkirchen, who then changed his name into Ret Marut (publisher of the anarchist periodical *Der Ziegelbrenner*), before moving to Mexico and acquiring his final, world-famous pseudonym (Goldwasser, 1993; Hauschild, 2012).

From the point of view of stylometry, that of Feige/Matut/Traven is a typical authorship verification problem, where the goal is not that of attributing an anonymous text to a candidate author, but that of verifying if two (or more) texts were written by the same author. Extensive research...