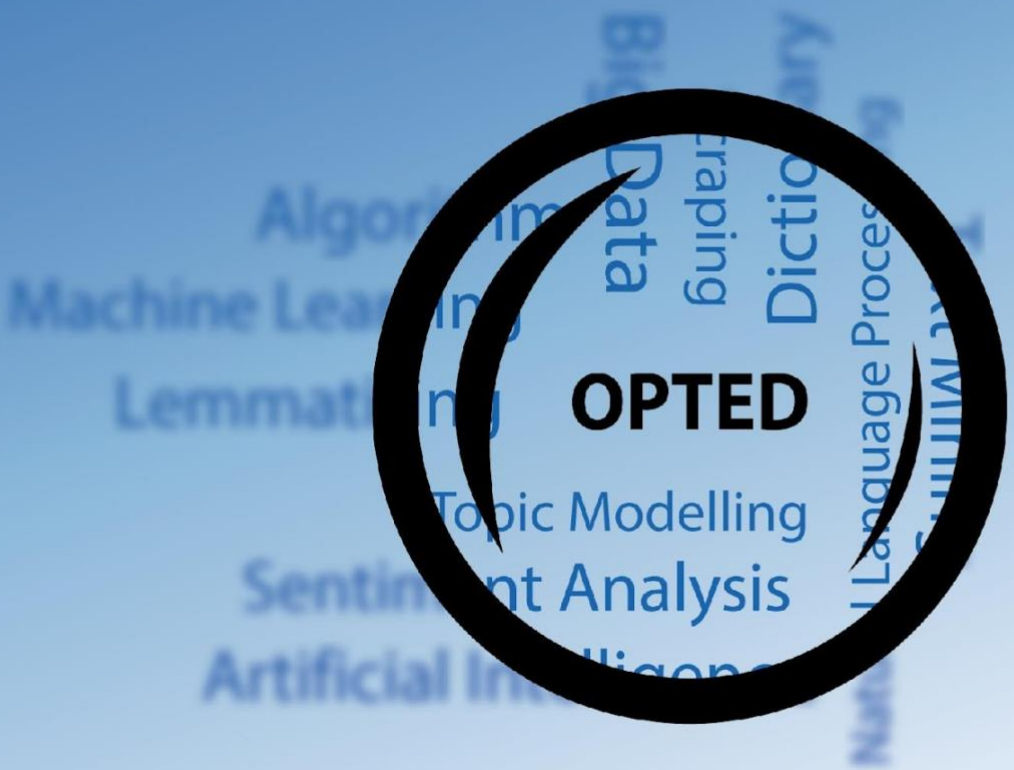
OPTED

Living Hub for Textual Research in a Multilingual World

Christian Baden, Alona Dolinsky, Farzam Fanitabasi, Fabienne Lind, Christian Pipal, Martijn Schoonvelde, Guy Shababo, & Mariken A.C.G. van der Velden



Disclaimer

This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement No 951832. The document reflects only the authors' views. The European Union is not liable for any use that may be made of the information contained herein.

Dissemination level

Public

Type

Demonstration

OPTED
Observatory for Political Texts in European Democracies:
A European research infrastructure

Living Hub for Textual Research in a Multilingual World

D6.1

Authors: Christian Baden[1], Alona Dolinsky[2], Farzam Fanitabasi[3], Fabienne Lind[4], Christian Pipal[5], Martijn Schoonvelde[6], Guy Shababo[1], & Mariken A.C.G. van der Velden[3]

[1] Hebrew University of Jerusalem
[2] University College Dublin
[3] Vrije Universiteit Amsterdam
[4] University of Vienna
[5] University of Amsterdam
[6] University of Groningen

Due date: March 2022

Executive Summary

Social science text research, and the study of political texts in particular, is undergoing rapid internationalisation. Internationally comparative research is growing in importance. Both developments have brought to the fore a range of challenges that arise from different languages used in the studied texts. On the one hand, much existing technology for computational textual analysis assumes English-language text; their application to textual corpora in different languages often reveals mismatches that may give rise to consequential biases. On the other hand, equivalent meanings are expressed in different ways in different languages, raising questions concerning comparative validity. It is the purpose of WP6 to map the challenges that arise from the computational analysis of textual data in a multilingual world, to identify best practices and guidelines, and chart a research agenda toward cross-lingually valid computational text analysis.

In this Deliverable, we provide the conceptual work for a research infrastructure for computational text analysis in a multilingual world. That is, we provide information of the languages studied in the academic literature, as well as indicate the strategies authors have taken to either (implicitly) compare their results to the English language norm, or explicitly used more than one language in the research design. In the following, we introduce the "living hub" as part of a research infrastructure informing and facilitating researchers' use of computational text analysis in a multilingual world, and describe how it, in this first step, enables users to access existing scholarly experience.[1] Specifically, we have taken stock of the state-of-the-art by providing an information hub for computational text analysis. This hub collates (a) the existing cutting-edge knowledge in the field; (b) the available tools; as well as c) an overview of validation and methodological evaluation used in the field. Our hub aims to enable researchers in the field to (i) draw upon the fast-growing stock of experience and capacities; (ii) apply transparent, context-sensitive standards for strategy selection and validation; and (iii) feed their insights back into a cohesive methodological debate advancing the field. WP6 thus builds towards a systematic and knowledge-based integration of multilingualism into social science research methodology. Hence, it lays the foundations for the future development of new, solution-oriented methodological capacity to analyse multilingual political discourse in a validated, transparent fashion.

1 Building a Living Hub for Textual Research in a Multilingual World

As part of the overall research infrastructure developed by OPTED, a living hub dedicated to the challenges of computational textual analysis in a multilingual world grants users access to a variety of critical knowledge resources. Eventually, the living hub introduced hereunder will serve as a one-stop-shop for a wide variety of resources for multilingual textual research, including validity benchmarks and benchmarking data sets, tools, as well as an inventory of key issues that arise from the application of computational tools to different languages. In its present, first stage, it enables users to rapidly identify and access existing research that has applied different textual analysis methods to textual corpora in different languages. In particular, users can browse existing research by language, methodological approach, validation strategy, and the kind of variable under investigation. The living hub addresses both novice and experienced users of computational text analysis: To the novice, it offers initial orientation by enabling a rapid identification of strategies that have been used by others in the analysis of similar textual variables in any given language. To the expert, it offers a quick overview of tested tools and (open access) resources, as well as validation strategies applied to a given method, measurement problem and language.

2 How to Use the "Living Hub"

Inspired by the biotools website, our Living Hub has a search bar where a user can put in keywords to retrieve relevant papers (linked via their unique identifiers (DOI)) from the database. Additionally, our website has a filter with the following options:

- Language. This indicator shows the amount of languages analysed in the paper. A user can also use this option using the filter to only search for papers on a particular (set of) language(s).

[1] In the other deliverables of this work package, we will build further on the living hub as a tool for multilingual text analysis.

- Method. This option helps researchers to search for papers employing specific methods: unsupervised, supervised, rule-based (e.g., dictionaries, actors), or manual. It can be both used in the search bar.
- Validation. This shows the type of validation the researchers have employed.

For any entry, a range of tags enable the user to quickly survey additional features of those research works retrieved by the query. For instance, a user may search for supervised methods applied to French or Spanish text, which will retrieve a total of 53 studies. Each entry then additionally informs the user which studies offer access to open data or open materials, used additional methods or languages, applied specific validation strategies, and measured what kinds of textual variables. In a future development, we aim to additionally identify studies that did not use the exact language queried for, but focused on other, but similar languages (e.g., a query for Estonian language might also identify work on Finnish text, which is linguistically close to Estonian). This feature will be useful especially for languages that are weakly represented in social scientific text research.

3 Data & Methods for the "Living Hub"

The data available on the website derives from a manual coding of 854 articles. These were obtained via a content analysis of all quantitative text-based research published over the past five years in the top 20 highest ranked journals in the Web of Science categories of communication, political science, sociology (including mathematical models) and psychology (multidisciplinary and mathematical), selected according to their 2019 SSCI 1-year impact factors. The initial content analysis began with an inventory of all articles 45,437 published in the selected journals (for more information, see this article) between January 2016 and September 2020. Using a keyword search on the Web of Science, we then identified a total of 7,296 potentially relevant articles whose abstracts referred to some kind of textual contents or text analytic procedures. We then accessed the full text of these articles to determine whether the presented research included any form of quantitative textual analysis.

For the initial screening of all published articles, quantitative textual analysis was defined broadly to include any form of processing natural language that identified specific kinds of textual contents with the purpose of classification and quantitative analysis. Analyses that relied solely on metadata or pre-existing classifications were excluded, as were investigations accessing only formal properties of the sampled texts (e.g., length). We included analyses of multi-modal media (e.g., posters, television) as long as textual contents were informative toward classification. Purely methodological contributions discussing specific potentials or limitations of available methods were excluded, unless they included applied demonstrations wherein actual textual data was processed. Articles were considered relevant as soon as they used any form of quantitative textual analysis, even if it was used merely in an auxiliary capacity (e.g., a content analysis to identify frames to be used in an experiment; sentiment analysis of open-ended survey responses). This screening yielded a total of N = 854 articles, after which a second, manual screening, filtered out a further 25 articles that did not analyse textual materials. The final set of articles coded was N = 829.

3.1 Coded Variables

To enable the targeted identification and retrieval of scholarly work related to specific queries, we classified all relevant studies based on a list of variables that represent different entry points for a user accessing the living hub. Specifically, we anticipate that users will have specific languages or language combinations in mind when accessing the hub for additional information; users may wish to preselect studies by their use of specific methods or focus on specific variables or broad analytical perspective, their use of validation strategies, or their affordance of open access data and materials. In addition, we coded several variables that aid users to quickly assess the similarity of studies' key properties to their intended usage (e.g., the total N of texts processed).

For all relevant articles included in the analysis, we manually coded 26 variables. First, we determined the type of textual data examined in the study, followed by the open-sourced availability of the data and code/instruments used in the study (including hyperlinks), and recorded the number of texts analysed. Additionally, we coded the analytical approach of the study as inductive, deductive or other the methodological strategy used, manual or computational (and if the latter, what specific methods were used). Next, we coded

the number of languages of the textual materials, and in case of studies using computational strategies, whether a validation attempt was mentioned and demonstrated. If a validation attempt was demonstrated, we coded the number of such validation attempts and the specific approach taken: rationale for pre-processing steps; validation of classification rules/criteria; comparing model output to gold standard; and validation of output by comparison to related concepts. For each of these we also coded additional details on the validation attempt. For multilingual studies using computational strategies, we also coded for whether the validation was reported for all languages, and whether (and if so which) bridging method was used. For studies using a manual strategy, we coded for whether a reliability test of manual coding was detailed. Finally, we coded for where the validation strategy used in the study was reported: all in the main text, partly in the main text, partly in online appendices or other; and only in online appendices or other supplementary materials.

3.2 Coded Variables

Each article selected for manual screening and content analysis was read and assessed by one coder. To ensure the reliability of our coding, a second coder repeated both the screening procedure and the coding process, using a random sample of articles classified at each stage. We then assessed the agreement between coders using Cohen's kappa. For the initial relevance screening of articles based on their use of quantitative text analysis, this sample consisted of 200 randomly selected articles from each discipline, sampled from the results of the keyword search. For the subsequent classification based on the four groups of coded variables, 10% from all relevant articles (n = 85) were coded by a second coder. Inter-coder reliability exceeded $\kappa = 0.9$ for both stages, and for all variables.

The same data required to enable user access to relevant research through the living hub also provides a valuable snapshot of the present state of multilingual text analysis, and quantitative text analysis more broadly, in the social sciences. Besides a descriptive overview of findings (working paper under preparation), this has also informed an initial publication by the WP6 team that identifies three major gaps in presently available technology and research practice for computational text analysis. Positioning multilingualism as one key domain in need for methodological attention and development, this article helps focus the attention of the research community on the issues addressed by this WP. Building upon this attention, the living hub introduced in this Deliverable, and to be further expanded toward D6.2 and D6.3, will serve as a central venue for the community to accumulate and exchange knowledge.