An Approach for Agile Interdisciplinary Digital Humanities Research — a Case Study in Journalism

Eetu Mäkelä¹, Anu Koivunen², Antti Kanner¹, Maciej Janicki¹, Auli Harju², Julius Hokkanen², and Olli Seuri³

¹ Department of Digital Humanities
University of Helsinki
² Faculty of Social Sciences
Tampere University
³ Faculty of Information Technology and Communication Studies
Tampere University

Abstract In this paper, we present insights into the collaboration between 1) computer scientists, 2) linguists and 3) media scholars gained from work related to a pilot study in the project Flows of Power: media as site and agent of politics (Academy of Finland 2019–2022). In lieu of describing the actual results from our analyses, we instead describe our approach, and how it led into a versatile general template for an iterative and discursive approach to digital humanities research, which moves toward questions of interest both fast, as well as with high capability to truly capture the phenomena from all viewpoints of interest.

1 Introduction

In this paper, we present insights into the collaboration between 1) computer scientists, 2) linguists and 3) media scholars gained from work related to a pilot study in the project Flows of Power: media as site and agent of politics (Academy of Finland 2019–2022). As a whole, the project investigates the agency of journalistic media in the flows of information, public opinion and power. Through a large-scale empirical analysis of Finnish news journalism between 1998 and 2018, the project explores the strategies of journalistic news media in staging and managing political processes, as well as seeks to find out if they have changed with the advent of social media.

While the full time scale of the project is vast, for our first pilot we decided to focus on a more tightly-defined case study: how was affectivity mediated, modulated and managed by the media in the reporting of, and commentary on, a particular yearlong political conflict between the right-wing conservative government and the trade unions in Finland (2015–2016). Focusing on the government’s goal to increase the competitiveness of the Finnish economy by lowering labour costs, this yearlong, multi-peak process was interesting, because while it was centred around a single issue, it also had enough duration and variability to be
on the upper boundary for manual research, and thus amenable to benefit from scalable computational means of analysis.

On the other hand, in choosing the management of affectivity (or emotiveness) as our object of interest, we were setting the bar quite high. Journalism in general has an innate culture of seeking to project an appearance of balance and objectivity, of dealing with issues fairly and without bias [13, 3]. According to Tuchman [14], this objectivity norm manifests so strongly that it can be described as a ‘strategic ritual’, whereby journalists utilise various strategies to appear dispassionate, disembodied and impartial. Given such an environment, it was clear that we could not use for example off-the-self approaches to sentiment analysis to obtain a credible and nuanced analysis of the phenomena in question. Instead, we needed to work closely together between the domain experts, linguists and computer scientists to derive new indicators for capturing the unique ways by which affect was expressed in this domain. Further, we also needed to develop tools to identify the ways by which this affect was then mediated, modulated and displaced in the service of the ritual of objectivity.

In the following, we will not describe the actual results from our analysis, but instead describe our approach in this pilot study, and how it led into a versatile general template for an iterative and discursive approach to digital humanities research, which moves toward questions of interest both fast, as well as with high capability to truly capture the phenomena from all viewpoints of interest.

## 2 Facilitating Discussion in a Shared Space

At the core of our approach is to, as early as possible, build a common environment where all partners of the project can not only view, but highlight to each other and discuss what is interesting in the data from their perspective. By doing this, for example, humanities scholars are able to highlight to the computer scientists new phenomena of interest in the data derived from their close reading, while the computer scientists can show what they are currently automatically able to bring forth from the data. Through this, everyone is kept on the same page, misunderstandings are avoided, and the most fruitful avenues for development can be negotiated in a shared space where everyone contributes equally.

In the FLOPO project, this began with making use of existing tools. First, a Slack chat space was established for discussion amongst project participants. Second, the whole 20 years of FLOPO data from three different news sources was loaded into Octavo [8], a custom-built Lucene index for rich queries, which also includes a web user interface for easy end-user access to query results. Crucially, the state of the query and the interface are serialised in the URL, facilitating easy passing of a particular view between the project participants.

In deciding what to focus on for the pilot case study, this interface was used for qualitatively verifying the interestingness of the material for analysis. Further, through passing query state URLs back and forth between the media scholars and the computer scientists, the Octavo interface was used to iteratively
develop and verify the query constraints used to extract the relevant subset of competitiveness pact-related articles for further study.

Soon after, a repeatable pipeline was designed which downloaded the results of a given Octavo query and converted them into plain text and CSV files, which could then be further processed computationally, but also close read as is by the media scholars to get an overview of the data subset. Further, the metadata CSVs for each data source were loaded into a shared Google Drive folder, which the media scholars then used for recording their notes in a shared location for discussion on Slack.

After multiple iterations using a combination of the Octavo interface, Google Drive sheets and the local plain text files, we arrived at a much more refined definition of what to include and exclude from the pilot study data set. In this, the ability to flexibly rerun the pipeline to produce new versions of the pilot data set proved crucial. Later, it also proved crucial in enabling us to flexibly add a fourth data source to our analysis, which became available to us only later in the research process.

3 Building a Pipeline for the Agile Development of Computational Indicators

While the Octavo interface and local text and metadata files allowed us to zero in on an interesting subset of the data, as well as explore its basic structure, it didn’t facilitate the computer scientists highlighting to the media scholars the results of their developments of computational indicators. As the adequate identification of the handling of affectivity desperately required this, we sought to add new pieces to our toolbox.

First, we needed to decide on a main format for storing our computational analyses in. Our primary toolbox for statistical analysis is R. This motivated using a ‘tidy data’ CSV-based format [19] as our main data format. As most linguistic analysis is currently based on variations of the CONLL-U format, we sought to design our core data model in similar terms. However, we could not use CONLL-U directly due to two reasons. First, while being based on TSV, the CONLL formats are an extension to it, with non-tabular means to denote paragraph and sentence boundaries. Second, even within the tabular part of CONLL-U, the semantics of all its columns are predetermined, with just a single "miscellaneous" column for recording additional information. As a result, we ended up with a core tidy table format that reproduces the CONLL-U output from the TurkuNLP pipeline [6], but with added columns for documentID, paragraphId and sentenceId. Together, these facilitate combining the per-document CONLL files into a single unified table, as well as remove the need to separately handle the way CONLL codes paragraph and sentence boundaries.

Given this base format, all additional annotation layers, beginning with named entity recognition using the rule-based FINER tool [12], could then

4 https://universaldependencies.org/format.html
be relegated to separate CSV files, where tuples like \((\text{documentId}, \text{sentenceId}, \text{spanStartId}, \text{spanEndId}, \text{value})\) were stored. Such tabular data are easy to produce, join together and manipulate dynamically within R.

For visualising to our media scholars what our computational approaches could extract from the texts, we decided to employ WebAnno [2]. While this tool was originally intended for the creation of datasets for language technology tasks, its functionality is designed to be very general, which enabled its use in a wide variety of projects involving text annotation\(^5\). In WebAnno, annotations to the texts are shown as highlighted text spans, with feature values shown as colourful bubbles over the text (see Figure 1). Crucially, in addition to the usual linguistic layers of annotation, like lemma or head, it allows the creation of custom layers and feature sets. Further, the various annotation layers can be shown or hidden on demand. For us, this provided the means to show what our computational tools could dig up from the data in an easy way. As an added benefit, the media scholars could use the manual annotation functionalities of WebAnno to both explicitly highlight to us new features of interest, as well as to provide corrections and further training data for our algorithms.

What remained to be done was figure out ways in which WebAnno could be hooked into the rest of our components as part of a reproducible and iterative pipeline. A first requirement was that we needed to be able to link to the view of a particular document from outside. Here, while WebAnno did provide URLs taking the user to a particular document, these were based on internal document ids instead of any external identifies we used. However, we were able to derive a mapping between the two, which was then loaded into the shared Google Sheets that tracked our pilot study data.

\(^5\) see: https://webanno.github.io/webanno/use-case-gallery/
As a second requirement, we needed to be able to iteratively synchronise the data between WebAnno and our environment for running the computational enrichments. Here, WebAnno supported several data formats for import and export, all of them assuming one document per file. Among others, different variants of the CONLL format were supported. However, while in the CONLL formats, the semantics of all fields are predefined, WebAnno also provided WebAnno-TSV as its own tab-separated text format, which included support for custom annotation layers. Because it is a text format and is well documented, we were able to implement a fully automatic bidirectional conversion between our corpus-wide, per-annotation CSV files and per-document WebAnno-TSV files.

To reiterate, now that we had WebAnno working, we had a better interface for close reading, but crucially also one which was able to highlight to the user all results from automated computational annotation. Through this interface, the users were thus also able to judge the quality of our computational analysis. Further, as the interface supported manual annotation of the material, the media scholars could use it both to 1) highlight new aspects of interest for the computer scientists to try to capture and 2) correct and further teach our machine-learned approaches.

4 Indicator Development and Analysis

Making use of these tools, we were finally able to develop indicators for the study of the management of affectivity in the news texts, as well as to develop the axes by which we’d compare them.

Previous studies, especially corpus-based studies based on appraisal theory (e.g. [5]) and those applying computational methods such as sentiment analysis, have approached affectivity and emotions through both overt emotive lexical choices as well as lexemes signalling clear negative or positive evaluations. While we knew this would not suffice for our purposes, we also started from this established baseline in our approach.

We first started with the word class of adjectives, from which we extracted those that had evaluative meaning. To ensure coverage, we automatically extracted from the corpora all 4,732 adjective lexeme (word) types. Evaluative lexemes (N = 2,857) were then picked manually. The excluded, non-evaluative lexemes consisted mostly of technical and temporal qualifications and other classifications such as nationalities.

However, in a given setting, almost any linguistic structure can express affectivity. Thus, we also close-read the news articles in our corpus to obtain a data-informed handle on their emotive and evaluative vocabulary beyond adjectives. We were struck by the large amount of metaphorical language – an evidently affective practice [17], but one that was fundamentally different from the evaluative expressions and also seemed to appear particularly in response to inherently affectively charged events and segments of text. Using metaphors that transposed events from the source domain (the labour market) into e.g. one
of sports or armed conflict enabled writers to utilise affective intensity in a safe manner without appearing confrontational.

While such metaphors are an interesting indicator of affective intensity, they are also content specific and cannot be identified using general-purpose word lists; nor can they be described using syntactic rules. In order to operationalise the identification of metaphors for this particular case, we manually analysed a sample of the articles and marked all passages with metaphorical language. Because we started this phase early, most of this work was done purely manually through the media scholars copy-pasting sentences into Google Sheets, but as soon as we had WebAnno set up, we migrated to using that.

We then examined the central vocabulary of these annotations and manually extracted a seed word list of around 500 words that marked strongly metaphorical discourse. Using pre-trained word embeddings to discover other words that were used in similar ways, we then automatically expanded this list to 2,100 words. For analysis, we further divided this set into two subsets, one for verbs and the other for nouns.

To go beyond emotive and evaluative lexemes, we also wanted to study how affect was modulated or ‘hedged’ in the material. Here, we turned to grammatical and lexical structures used to express ‘evidentiality’ and epistemic modality. In communication, these structures are typically used to express reservations or to communicate degrees of uncertainty or conviction and have been shown to be heavily involved, especially in academic genres, in politeness strategies (Brown and Levinson, 1987; Myers, 1989; White, 2003). In our corpora, we identified these structures using a list of grammatical structures that expressed evidentiality and epistemic modality in Finnish, based on a review of established research (Laitinen, 1992; Kangasniemi, 1991; Hakulinen et al., 2004).

Given the assumption that affect in journalistic contexts would be heavily managed and hedged, our linguist looked into research on grammatical and lexical structures used to express ‘evidentiality’ and epistemic modality. In communication, these structures are typically used to express reservations or to communicate degrees of uncertainty or conviction and have been shown to be heavily involved, especially in academic genres, in politeness strategies [1, 11, 18]. Semantically, these grammatical structures are often used in language to express the degrees of likelihood of some event or to express some attitude towards the source of the information. The logic of their face-protecting pragmatic function is that the affective content of the message appears to be less strong if it is expressed as a mere possibility instead of a certainty. Our hypothesis was that, as with academic text, these structures would also appear in journalistic texts in managing and containing affect in the service of the ritual of objectivity. In our corpora, we identified these structures using a list of grammatical structures that expressed evidentiality and epistemic modality in Finnish, based on a review of established research [10, 7, 4].

After listing the linguistic markers used in the study, we delineated the axes by which to compare them. First, we wished to study the hypothesis of the ritual of emotionality [15, 16]. This hypothesis starts from the understanding that
appealing to emotions is a good way to rouse and maintain audience interest, and thus should be available also in the repertoire of journalists. However, due to the needs of the ritual of objectivity, this affectivity cannot appear in the journalists’ own discourse, but will be outsourced to quotes. We thus developed methods to identify both direct and indirect quotations in the articles. Due to the nature of the material, where the accurate attribution of information and claims is deemed extremely important, this was a relatively easy matter. First, each source had their own clear and consistent markers used to identify direct quotations. Identifying indirect quotations was based on automatic syntactic dependency parsing to extract syntactic structures used for reported speech in Finnish.

We next sought to examine possible external factors that could have influenced the quantity of evaluative and emotive words. We tested the ‘click-bait’ hypothesis, which posits that affectively intensive content will be located at the beginning of the article as a means to solicit reader interest. We also investigated whether the affectivity and intensity of the actual events affected the journalism by (1) comparing the reporting on the competitiveness pact with baseline political journalism of a randomly chosen other period and (2) comparing articles written during periods of peak activity with those written at other times.

Next, we analysed the differences among different news media, ranging from the tabloid Iltalehti to the Finnish news agency STT, whose direct clients are not the public but other news outlets, which may choose to republish or extend their stories on their own sites.

Finally, we focused on affectivity in various types of news articles: (1) news reporting focused on facts; (2) news analyses with an emphasis on contextualising, explaining and interpreting events; (3) news commentaries or columns where the writer’s voice and opinion are foregrounded; and (4) editorials and leading articles, either signed or non-signed, that express the viewpoints of the paper. To distinguish between these, we used the iterative capabilities of our pipeline to draft classification rules for each news source, making use of both structured metadata such as section information, as well as explicit article type markers in for example the article headlines.

After all of the above, in the end we had more than twelve different indicator signals available for analysis, combined with six axes or dimensions of analysis. Naturally, this leads to an explosion of possible combinations to analyse, only some of which are interesting. Identifying which aspects to focus on again required closely working together between the computer scientists, the linguists and the media scholars. To facilitate this, we added one final piece of software to our ecosystem: a web-based tool for explorative statistical analysis, which allowed anyone in the project to 1) get statistical overviews of a desired indicator variable across a custom combination of dimensions and 2) to also get a statistical overview of how each data point was placed with respect to others from the viewpoint of an indicator (see Figure 2). Naturally, this interface was also programmed to facilitate the sharing of queries and UI state through copying and pasting the URL visible in the browser. In addition, the individual articles
Figure 2. View in the explorative analysis tool highlighting news articles from a user-selected timespan with an abnormally large amount of metaphors.

visible in the second view were all linked to WebAnno, so that outliers could be investigated, and reasons for particular distributions qualitatively explored.

5 Discussion

As already said, the actual results of our analyses will not be discussed here. For them, the reader is referred to an article currently submitted for evaluation in the journal Journalism [9]. Instead, we will summarise the aspects pertinent to our process here.

First, in this paper, we’ve presented an example of a general approach to digital humanities that allows combining computational analysis with the knowledge of domain experts in all steps of the process, from the development of computational indicators to final analysis.

Once fully functioning, our solution rests on three pillars. The first of these is an interface for close reading, but crucially one which is able to highlight to the user all results from automated computational annotation. Beyond pure close reading, through this interface, the user is thus also able to evaluate the quality of computational analysis. Further, the interface supports manual annotation of the material, facilitating correction and teaching of machine-learned approaches.

The second of our pillars is an interface for statistical analysis, where the phenomena of interest can be analysed en masse. However crucially, this interface is also linked to the close-reading one to further let the users delve into interesting outliers. Through this, they are not only able to derive hypotheses
and explanations of the phenomena, but can also identify cases where outliers are more due to errors and omissions in our computational pipeline.

Finally, our third pillar is an agile pipeline to move data between these interfaces and our computational environment. In application, this third pillar is crucial, as it allows us to iteratively experiment with different computational indicators to capture the objects of our interest, with the results quickly making their way to experts for evaluation and explorative analysis. Through this analysis and evaluation, we then equally quickly get back information on not just the technical accuracy of our approach, but also if it captures the question of interest. Further, beside direct training data, we also get suggestions on new phenomena of interest to try to capture.

While in any given project, some of these pillars will not be available from the very start or will be replaced with better ones during the course of the project, the idea and principles remain the same. By maintaining from the start environments and interfaces that allow both computer scientists and humanities scholars to not only view, but highlight to each other all aspects of the data, we further a shared understanding between the participants. For example, humanities scholars are easily able to highlight to the computer scientists new phenomena of interest in the data derived from their close reading, while the computer scientists can easily show what they are currently automatically able to bring forth from the data. Through this, everyone is kept on the same page, misunderstandings are avoided, and the most fruitful avenues for development can be negotiated in a shared space where everyone contributes equally.

In fact, we argue that it is precisely getting such an environment going as early as possible (even if by ready tools that are not completely ideal) that facilitates a fruitfully directed cycle of discussion, agile development and experimentation. Together, we feel that these insights provide a versatile template for an iterative and discursive approach to digital humanities research, which moves toward questions of interest both fast, as well as with high capability to truly capture the phenomena from all viewpoints of interest.

References


