Workshop on Tool Criticism in the Digital Humanities

Myriam C. Traub[1]*, Jacco van Ossenbruggen[1,2]

Abstract
This document reports on the discussions and results of the Workshop on Tool Criticism in the Digital Humanities, that took place on May 22, 2015 in Pand 020, Amsterdam. The workshop was co-organized by Centrum Wiskunde & Informatica, the eHumanities group of KNAW and the Amsterdam Data Science Center.

Keywords
Digital Humanities, Tool Criticism, #toolcrit

[1] *Centrum Wiskunde & Informatica, Amsterdam, The Netherlands*
[2] *Vrije Universiteit Amsterdam, The Netherlands*
***Corresponding author**: traub@cwi.nl

Contents

1. Motivation and background

In digital humanties (DH) research there is a trend to the use of larger datasets and mixing hermeneutic/interpretative with computational approaches. As the role of digital tools in these type of studies grows, it is important that scholars are aware of the limitations of these tools, especially when these limitations might bias the outcome of the answers to their specific research questions. While this potential bias is sometimes acknowledged as an issue, it is rarely discussed in detail, quantified or otherwise made explicit.

On the other hand, computer scientists (CS) and most tool developers tend to aim for generic methods that are highly generalisable, with a preference for tools that are applicable to a wide range of research questions. As such, they are typically not able to predict the performance of their tools and methods in a very specific context. This is often the point where the discussion stops.

The aim of the workshop was to break this *impasse*, by taking that point as the start, not the end, of a conversation between DH and CS researchers. The goal was to better understand the impact of technology-induced bias on specific research contexts in the humanties. More specifically, we aimed to identify:

- typical research tasks affected by by technology-induced bias or other tool limitations
- the specific information, knowledge and skills required for scholars to be able to perform tool criticism as part of their daily research
- guidelines or best practices for systematic tool and digital source criticism

1.1 Tool Criticism

With *tool criticism* we mean the evaluation of the suitability of a given digital tool for a specific task. Our goal is to better understand the impact of any bias of the tool on the specific task, not to improve the tools performance.

While source criticism is common practice in many academic fields, the awareness for biases of digital tools and their influence on research tasks needs to be increased. This requires scholars, data custodians and tool providers to understand issues from different perspectives. Scholars need to be trained to anticipate and recognize tool bias and its impact on their research results. Data custodians, tool providers and computer scientists, on the other hand, have to make information about the potential biases of the underlying processes more transparent. This includes processes such as collection policies, digitization procedures, optical character recognition (OCR), data enrichment and linking, quality assessment, error correction and search technologies.

1.2 Organisation and format

The scope and format of the workshop was developed during an earlier meeting of the workshop organisers (see Appendix A) on March 6 at CWI in Amsterdam. Participants were asked to use the workshop website to submit use cases in advance, and we received seven use cases in total.

The program of the workshop was split in several parts. The morning was dedicated to introducing the concept of *tool criticism*, pointing out the goals and non-goals of the workshop and a short presentation of the use cases (see 2. During an informal lunch, participants could express interest in a specific use case. The participants choose 4 out of all 7 use cases as a target for the afternoon sessions, and formed teams around these 4 cases. After lunch, each of the four breakout groups were asked to work out their use cases further. The organizers provided a list of questions to guide and inspire the breakout sessions (see Appendix B). Afterwards, the results were presented and discussed in the plenary. All use case leaders were kind enough to send us their notes by email. These notes were used as input for section 2.

1.3 Workshop opening

Day-chair Sally Wyatt opened the workshop and welcomed all the participants. Before the use cases were presented, Jacco van Ossenbruggen briefly explained the goals (see Section 1) and non-goals of the workshop. The non-goals included: discussions on how to *reduce* tool-induced bias (i.e. by improving the tool), to down-play the role of the tools ("the tool is only used in exploratory phase of research") or discussions about the pros and cons of digital versus non-digital approaches ("we would just hire 20 interns to do this by hand").

In the following discussion, a participant pointed out that scholars tend to overestimate the certainty and trustworthiness of their own data and rely too much on intuition.

How can scholars differentiate signal from noise? Understanding how this can be done is not only relevant for tool-induced errors, human errors may have a similar effect (example: relevance assessments). The intuition that more meta data is better is not always true (see IR).

2. Use cases

The following use cases were submitted to the workshop:

- Co-occurrence of Named Entities in newspaper articles
- SHEBANQ
- Word frequency patterns over time
- Polimedia
- Location extraction and visualisation
- contaWords
- Quantifying Historical Perspectives

From this list, the participants chose to discuss the first 4 use cases in the breakout sessions. The participants were asked to form groups with at least one researcher from (Digital) Humanities as well as Computer Science.

2.1 Constructing social networks with co-occurrence

This use case was submitted by Jacqueline Hicks (KITLV) under the original title "Co-occurrence of Named Entities in Newspaper Articles".

Use case description

The computational strategy is to use the co-occurrence of named entities in newspaper articles to represent a real-world relationship between those entities.

Main discussion points

The discussion started with explaining the purpose of the tool: As well as locating names of people appearing together in one sentence in a newspaper article, it was also used in the project to help disambiguate entities.

The tool makes use of the widely known and used Stanford NER, its performance is documented on CoNLL 2002 and 2003 NER data[1]. This data is not similar to the data used in the example use case. To be able to evaluate the performance of the Stanford NER in the new domain, the researcher would need a corresponding "ground truth" data set, that is, manually constructed reference data that can be used to check the results of the automatic NER process. Developing a ground truth for a new domain is a very time consuming operation.

The research task is to find out whether the tool can help detect changes in communities of elite that changed over regime transitions when the Indonesian authoritarian government fell after 30 years in power. However, the task turned out to be difficult to solve as insufficient data was available for the time before 1998. More time is needed to add linguistic context to the co-occurrences to find what sort of relationships ties the entities together in a sentence. A co-occurrence of two entities can mean that they participated in the same event, that one person commented on the other or that they were in competition with each other. With such diverse relations, it is difficult to draw conclusions from the automatically generated graph.

Biases of the source selection The data was collected from several listserves of news articles on Indonesian politics. The articles on these listserves were handpicked by those running them and so could not be considered free from bias. They include, for example, the articles in English language, chosen for the interest of foreign and Indonesian readers generally interested in political reform, as it was originally started to share information among activists under the authoritarian government. Since these biases are known, they are easily dealt with as limitations of the study in the same way that research limitations are usually explained when writing in the social sciences. This is in contrast to the computational filtering which introduces biases which are not known to the social scientist.

Provenance of the data All articles had date and newspaper source on them.

[1] http://nlp.stanford.edu/ner/

Utility of the tool Utility is limited and only good for some initial explorations. The idea of the session was to find ways to integrate qualitative information from interviews with Indonesian elites about their network with the computational techniques. The group discussed the idea to investigate the changes in the political system by creating two networks for the time before and another two for the time after the transition. The networks could then be compared and in case the networks of the same period coincide, but the networks across periods do not, they may be used to reveal interesting differences as basis for further research. Jacky would have to explain the differences by investigating through political sciences literature how a military group fragmented in this way around person X and/or came together (again). Jacky already wrote a paper on how social scientists have identified populations of elites in the past and how this can be done differently with computational tools [1].

Summary

In general, to methodically evaluate tools is extremely time intensive. It requires intensive exchange between the user and developer. As publishing papers is an incentive to work in academia, the lack of forums to publish about tool criticism is a problem.

2.2 SHEBANQ

This use case was submitted by Dirk Roorda (DANS).

Use case description

SHEBANQ[2] allows users to query the Hebrew text database created over the years by the ETCBC group at the VU University Amsterdam. There is an associated, offline tool, LAF-Fabric for more refined and intense processing of the data. The data is encoded in Linguistic Annotation Framework, an ISO standard. LAF-Fabric is a python tool to deal with big LAF resources efficiently. There are several modules on top of it that exploit the structure of this particular research.

Main discussion points

The tool At the beginning of the breakout session, Dirk Roorda introduced the participants to some of the functionalities. SHEBANQ should actually be seen as a collection of tools to annotate and query the Hebrew Bible. It is developed at the Eep Talstra Centre for Bible and Computing which is located at VU University. Not only is the tool freely available through the *ETCBC's organizational github account*[3], a user can also download the documentation of the tool as well as executable *IPython Notebooks* that demonstrate some uses of the tool.

The data SHEBANQ is designed to support analysis of a specific version of the Hebrew Bible and is therefore tailored to cater to the specific requirements of the data set. Using the tool on a different data set therefore does not seem reasonable. The data is encoded in the Linguistic Annotation Framework (LAF), an ISO standard [2].

The user SHEBANQ encourages a community of people to come forward with their attempts to answer research questions by means of formalizing questions into tasks that can be run on the data. A unique feature of SHEBANQ is the possibility to share and publish queries:

> "If you want to cite your shared query in a publication, you can also publish it. Your query and its results on a particular version om the database will be frozen, so that others will see exactly the same results later on. When newer versions of the database arrive in SHEBANQ, you can run the same query on the new data. You can modify that version of your query and publish it separately from earlier versions."[4]

This is seen as an important and novel feature that can facilitate the discussion among users on the fitness of a query for a given research task.

Summary

It is vitally important to make explicit how the data in the ETCBC database has been encoded. Who has done it by what methods? Especially when the same researcher draws conclusions from the database as the one who has contributed relevant parts of the encoding. That is not necessarily bad, as long as his/her method of encoding is well described and can be subject to criticism. Another matter is whether other researchers are willing to contribute data to SHEBANQ. That will only happen if others can identify with the way of encoding and trust that SHEBANQ is impartial. Maybe SHEBANQ should allow multiple encoding styles and give other researchers partial ownership.

2.3 Word frequency patterns over time

This workshop was submitted by Marijn Koolen (UvA).

Use case description

The use case aims at looking into tools that chart word frequencies using timebased counts of n-grams found in digital sources. Examples of such tools are the Google Ngram Viewer[5] and the Ngram Viewer bases on historic newspapers which was developed by KB[6].

Main discussion points

Criticism is not only a playing field for Computer Sciences and Humanities but also for libraries and social sciences. It is, however, sometimes difficult to distinguish tool criticism from data criticism since tools have been used to create the data. These tools may not be available for criticism, which needs to be explicitly accounted for.

[2] http://shebanq.ancient-data.org/
[3] https://github.com/ETCBC
[4] http://shebanq.ancient-data.org/
[5] https://books.google.com/ngrams
[6] http://lab.kbresearch.nl/find/Ngrams#

The tool The chosen tools are designed to visualize word counts on a time line. In the experience of the researchers, this task is not as simple as it may appear: three different programs give three different counts for the total word count. In linguistic annotation, when different people annotate the same text, different schemes are used. The resulting conflicts need to be resolved by writing down the choices and agreements. This could be done similarly for coding / tools. The different results show that a tool does interpretation, too (in the sense of defining what a "word" is). Humanists are often put off if such counts are off by 1, because they tend to have precise ideas about text length. Without statistics, it is hard to say how much difference/variance is 'allowable' for a humanities researcher. This also applies for search engines. One participant recalls different answers from different search engines on the same query. She concluded that tools are not neutral, and that accuracy/concreteness are an illusion.

Interestingly, though, textual scholars seem to cope with this lack of precision very well *until* they start using a technical tool? We should remind people that tools and code are human engineered creative contraptions that have all queer human decisions and ambiguities built in. History depends on who writes it, code depends on who writes it.

The user The group further discussed the skills required from a humanities scholar to perform the tool criticism. They agreed that to understand a program, to some extent programming skills are required and that it should be part of the education in DH. A better documentation of the program code could make it easier to understand it and/or code in a way that is easier understandable. The readability of the code, however, may affect the efficiency. Understanding the implications of program code requires deep inspection and knowledge of the code. It can be very complicated for good reasons. If scholars cannot invest a considerable amount of time in understanding the code, they will have to trust the experts. In that case, the developers need to explain the tool, for example how it counts words and why it may come up with different totals for word counts than another tool.

The scholars may also need the programming skills to understand the methodology that a certain tool may force on them. This is foremost a task for humanities researchers to experiment and judge the methodology. If the source code is not (freely) available, it requires the scholar to experiment in order to to find major shortcomings or bias [3]. The use of tools should be embedded in a research process that iterates between distant reading and close reading to foster understanding. Tool support for this could be provided with "Sub Corpus Modeling" [4]. Ideally, there are multiple tools available that a scholar can choose from. In order to make an informed decision for choosing one above the other, the criteria need to be clear.

An important aspect of criticism is seen in its publication. Results of tool criticism should be reported to other users. This, in turn, raises the question of trust. Criticism cannot be considered as neutral, as it depends on the persona, background, status, etc. of the critic.

Tool builders Tool builders and computer scientists could learn from the humanities that there are more perspectives / more possible choices what the 'data' are. However, computer scientists are *not* interested in what DH does. CS studies process and abstraction. We should NOT suggest that DH is the field where Humanities and CS meet. It is maybe where AI and Humanities meet.

Summary

CS/AI need to evaluate a tool in a way that is tailored to a humanities researcher. The commonly used computational metrics usually do not answer that question. The DH, however, are in a 'it's all up in the air' period (as opposed to times in science where things seem to be clear and generalized); and scholars are not even sure about the standards against which they should be evaluated. Therefore, in order to define the requirements of humanities scholars, more discussions between the two disciplines is needed.

2.4 Polimedia

This use case was submitted by Laura Hollink (CWI).

Use case description

PoliMedia[7] is designed for specific humanities research tasks that require the possibility to do a cross-media analysis [5]. An example use case might be: studying several events related to "the resignation of Aantjes" by comparing information from different media. With PoliMedia, researchers can search among the debates in the Dutch Parliament (Dutch Hansard), Dutch historical newspapers archive and ANP radio bulletins, in a uniform search interface. The functionality is proven useful and the system design is highly valued. However, there are still obvious limitations.

Main discussion points

During the discussion, the PoliMedia group particularly wrote down a list of deficiencies of resource bias, then brainstormed about the solutions from the "tool side".

Biasses of the Source Selection Some bias issues of the dataset are known and might be quantified or circumvented. One problem is the coverage and selection of the resources: PoliMedia does not make use of data from television programs and news (but it does have ANP as a data source); the dataset covers only one radio station, so opinions might be limited; the selection of KB newspaper items for PoliMedia are significantly different in amount related to different newspaper brands. Additionally, there are technical issues such as OCR errors in the database, hindering users in retrieving the complete results. There are also biasses that the creators of the system cannot circumvent or quantify. On one hand, some of the links/search results are definitely lost due to the bias in the phase collecting the resource and system's technical issues. We do not know what we are missing in the database and how

[7] http://www.polimedia.nl/

those missing files would influence researchers' conclusions. On the other hand, bias can be caused by a chain of uncertainty: we don't know what the bias is of the off-the-shelf topic extraction tool.

Data Provenance Provenance of the data is clear and all search results link to original sources where a user could check if the digital versions are correct. However, the provenance of the algorithm is unclear: e.g. the system limits links to articles written within 7 days of the debate. This would be a limitation if the user needs more information. Such issues could be solved by a collaboration/dialog between tool makers and users, to explain and point out the impact of the algorithms on specific research questions. It is also possible to change the tool so the user can define a time period.

Solution Brainstorm In regard to to the limitations discussed, the group wrote down some questions and brainstormed solutions for them:

How to convince a reviewer that dataset and tool are good enough to draw quantitative conclusions from it?

Solution: *Sandbox:* on the spot evaluation of that particular query: The general goal is to provide the user the means on the spot get a feeling or even a measure of the bias. For dealing with the bias of data selection, practically user can always manually go to KB archive for more complete files that should be in there. The system can also compare the results with present links to on the spot, evaluating for that particular query. Till here users might still miss some links, but at least they cannot systematically miss out on things. For dealing with OCR issues, the system can provide relevance feedback, and does a query expansion to help users finding miss-OCRed versions of their query.

Quality can vary per query (e.g. simple/complex query, OCR errors, etc.), how to deal with the specific quality issues?

Solution: *Queries Sharing:* If a user took time making queries for a particular topic and find meaningful results, other users may also need the "accurate queries" when searching for similar topics. Solution: Triangulation: could be possible if we had multiple versions of the linking algorithm.

Solution: *Sharing the research process:* validated subsets that could be reused and criticized.

Future questions and research direction Given what we know about the quality of the tool/data, what can we do in our research:

- What can prove that something is there: e.g. media said x, this debate is discussed by x
- We can never prove that something is not there: e.g. nobody said x, this debate is never discussed.
- We can find preliminary results for quantitative questions: e.g. this debates is discussed more in the media than another one. Further research would be needed for definite conclusions about these kinds of questions

Summary
The group discussed different biases that influence research tasks. Some biases were found easy to assess and circumvent (limited number of sources included), others were more difficult (missing links and cascading of biases from tools used for preprocessing). The question was raised, how a reviewer could be convinced that tool and data are suitable to perform the task. Solutions could be a "sandbox" approach (on the spot evaluation of a query, ask user to give some manual results and check), a community approach (share queries, quality of queries, validate queries and answers) or cross validation with other tools.

3. Results

3.1 Discussion

At the end of the workshop, the participants agreed that the idea of *Tool Criticism* as part of the Digital Humanities' research practices should be fostered. This could be achieved in different ways. A traditional way that would reach a large audience could be a journal article (Digital Scholarship in the Humanities[8], Digital Humanities Quarterly[9] or a conference contribution (DH 2016, DH BeNeLux 2016).

Complementary to this, a more "interactive" approach in the form of a website[10] or a forum was suggested. This could be used to obtain feedback from users on a selected set of powerful tools. It would be interesting to be able to collect use cases and to compare evaluations of different tools that were designed to support similar tasks (such as named entity extraction). The insights gained from these examples could be used to create checklists and guidelines for both, tool builders and users. The checklists should, however, not only focus on general tasks, but also on very specific ones.

In order to encourage the direct exchange of ideas between tool builders and humanities scholars and to complement creation and evaluation of tools, hackathons could be organized. This could be done in one-day events, such as a follow-up workshop or at larger scale as part of a Dagstuhl or Lorentz Center seminar. Ideally, those activities should result in the establishment of a European network for tool criticism.

Acknowledgments

We would like to thank the eHumanities group of KNAW, CWI, COMMIT/ and the Amsterdam Data Science for their support, and all the participants for the valuable contributions.

[8] http://dsh.oxfordjournals.org/
[9] http://www.digitalhumanities.org/dhq/
[10] see for example http://programminghistorian.org/

References

[1] Jacqueline Hicks, Ridho Reinanda, and Vincent Traag. Old questions, new techniques: A research note on the computational identification of political elites. *Comparative Sociology*, 2015.

[2] Nancy Ide and Keith Suderman. The linguistic annotation framework: a standard for annotation interchange and merging. *Language Resources and Evaluation*, 48(3):395–418, 2014.

[3] Alberto Acerbi, Vasileios Lampos, Philip Garnett, and R. Alexander Bentley. The expression of emotions in 20th century books. *PLoS ONE*, 8(3):e59030, 03 2013.

[4] Timothy R. Tangherlini and Peter Leonard. Trawling in the sea of the great unread: Sub-corpus topic modeling and humanities research. *Poetics*, 41(6):725 – 749, 2013. Topic Models and the Cultural Sciences.

[5] Martijn Kleppe, Laura Hollink, Max Kemman, Damir Juric, Henri Beunders, Jaap Blom, Johan Oomen, and Geert-Jan Houben. Polimedia-analysing media coverage of political debates by automatically generated links to radio newspaper items. In *LinkedUp Veni Competition on Linked and Open Data for Education*. CEUR-WS, 2014.

1. Organizers

Chair

Sally Wyatt, eHumanities Group, KNAW

Organization

Jacco van Ossenbruggen, CWI
Myriam C. Traub, CWI
Victor de Boer, VU
Serge ter Braake, VU
Jackie Hicks, eHumanities
Laura Hollink, CWI
Wolfgang Kaltenbrunner, UL
Marijn Koolen, UvA
Daan Odijk, UvA

2. Breakout session

These questions are intended to provide starting points for the breakout sessions and to stimulate the discussion, in case is comes to a standstill. You do not need to answer all questions and it may well be that the splitting into two separate categories does not work well for your use case. If so, please feel free to add, remove, merge, move, or reword the questions in a way that they fit your needs.

Data set criticism

1. What type of data does the tool make use of?
   (a) Is the tool able to cope with multiple data sets (of different types)?
   (b) What is the relation between data set and tool?
   (c) How does the tool deal with anomalies and outliers?
2. Is documentation on the curation, representativity, biases and pitfalls of the data set available?
3. Is provenance data on the data set available?
4. Who created the data set?
   (a) Who was involved? What is the reputation / scientific impact / qualification of the people involved?
   (b) What institutions were involved? What is the reputation / scientific impact of the institutions involved?
5. When and how was the data set published?
6. Was the data collected for a specific task / research question?
   (a) How does this differ from your intentions?
   (b) Is the data set credible and objective?
7. Do other versions of the data set exist?
   (a) Are there older / more recent versions of the data set?
   (b) How do the versions differ?
8. Does the data show a particular political or cultural bias?
   (a) Is this bias of importance for your research question?
9. Do similar data sets from other sources exist?
   (a) Can you use the other data set(s) to answer the same research question?
   (b) Can you use the other data set(s) to detect / quantify biases in your data set (triangulation)?

Tool criticism

1. Was the tool developed to perform a specific task?
   (a) How does this task differ from yours?
   (b) For which part of your research cycle do you think the tool is suited (exploration, hypothesis generation, . . .)?
2. Is documentation on the precision, recall, biases and pitfalls of the tool available?
3. Is provenance data available on the way the tool manipulates the data set? (i.e. algorithms, choices when selecting, NLP pipeline)
   (a) What would it take to make the tool suitable for drawing quantitative conclusions?
4. Which versions of the tool are available?
   (a) What are the differences between the versions?
   (b) Which version caters best to the requirements of your research task?
5. Who are the developers behind the tool?
   (a) Who was involved? What is the reputation / scientific impact / qualification of the people involved?

   (b) What institutions were involved? What is the reputation / scientific impact of the institutions involved?
   (c) Do you know which scientific discipline the tool was built for? Does this matter for your research task?

6. Do you know similar tools?
   (a) Can you use other tools to answer the same research question?
   (b) Can you use the other tools to detect / quantify biases in your data set (triangulation)?