STSM REPORT: Sinai Rusinek

Sinai Rusinek
Reach Hadas 256
Givat Yearim
Israel 90970
Tel. 972-528397737
Sinai.rusinek@mail.huji.ac.il

Purpose of the STSM (from original letter)

Research of 'Republics of Letters' entails the study of various kinds of connections and relations between nodes of a network. The most commonly studied relation in the analysis and visualization of correspondence is that between the correspondents. The study of the texts themselves, however, opens up additional nodes and dimensions of such networks. Unlike plagiarism, which normally refers to the use of more or less exact phrases or text sections, the humanist phenomenon of text re-use includes a broad range of methods, from citations and direct quotes to paraphrases and hidden references. It therefore also presents a bigger computational challenge. In recent years various text reuse algorithms and tools have been applied to various textual genres - from ancient literature to the Encyclopedie, to modern fiction and nonfiction and to newspaper collections, but not, to our knowledge, to correspondence collections. The aim of this mission is therefore to explore the application of text reuse detection and visualization methods to correspondence corpora, first and foremost the application of the Tracer tool, developed by Marco Büchler, on the corpus of the project "Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic". An adaptation of the Tracer tool (which is still being developed and will be available as open source) to correspondence will enable, in addition to enriching the reading of the text by hyperlinking to the quoted/referenced source, also to analyze and visualize the flow of ideas, tropes, motives, quotes and commonplaces in the corpus. This STSM would naturally contribute to the work of WG3, text and topics, which aims to engage the text of correspondence with text mining approaches.

Description of the work carried out during the STSM

As is often the case, requirements for various stages of the analytical process often take us back to preliminary stages of data preparation. Thus, for example, even after securing the results and consulting with the developer of the visualization tool regarding the best visualization scenarios, the need came up to revisit the initial pre-processing and extract the text in different ways. Such back and forth is in fact conducive to both the implementation and the interpretation of the process and results. In what follows, however, I describe the work process in logical, rather than chronological order.

The Corpus, and Data Preparation:

The Epistolarium corpus used for the mission is made available on the Dutch DANS (Data Archiving and Networked Services) platform. It consists of several collections of letters, marked up with three slightly different XML schemas. I used a python script (adapted to each schema) to extract the metadata (author, recipient and full date) for each unit of text.
The unit of text that we chose (after first experimenting with segmentation to sentences) for initial analysis was a paragraph (see discussion on segmentation***). The paragraphs were extracted only from the Body element of the marked-up letters, in order to avoid formulae typical of openers and closers. As it became evident in the results, this was not enough to keep formulae out, and see discussion ***.

Configuration and Analysis

The Tracer tool enables a large choice of parameters in its implementation. In order to configure the tool to yield the most productive analysis a preliminary examination of several possible methods and various parameters has to take place, in order to compare the number of results, the time each implementation takes, and to have a first impression of the nature of results detected.
For reasons of time one specific scenario was selected (in consultation with Marco Büchler, the developer of the Tracer Tool), as follows:

```
<property name="PREPROCESSING_IMPL" value="de.gcdh.tracer.preprocessing.WordLevelPreprocessingImpl" />
<property name="TRAINING_IMPL" value="de.gcdh.tracer.featuring.semantic.WordBasedTrainingImpl"/>
 <property name="SELECTION_IMPL"
value="de.gcdh.tracer.selection.localglobal.LocalMaxFeatureFrequencySelectorImpl"/>
<property name="LINKING_IMPL" value="de.gcdh.tracer.linking.InterCorpusLinkingImpl"/>
<property name="SCORING_IMPL"
value="de.gcdh.tracer.scoring.feature.selected.symmetric.SelectedFeatureResemblanceSimilarityImpl"/>
```

Preprocessing according to this scenario yielded: 30,099 potential reuse units (paragraphs), 129,732 word types and 1, 265, 417 word tokens in the corpus. The case was ignored and diacritics erased by the tracer tool.

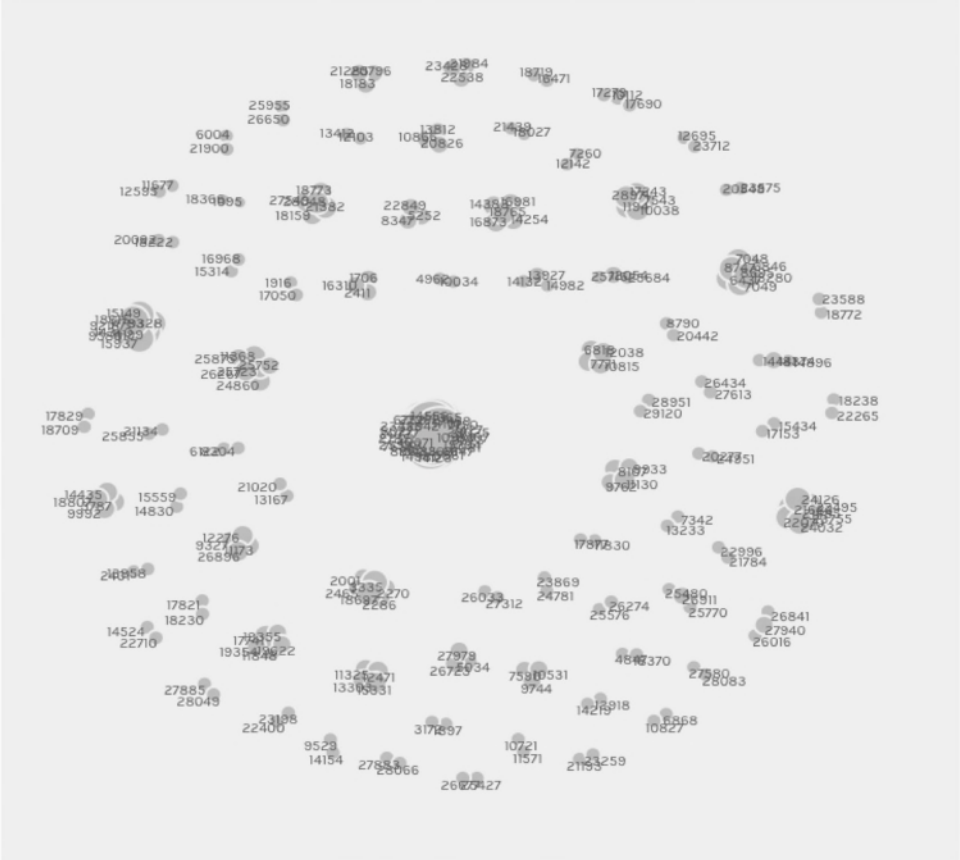
Two additional parameters were not set in advance but explored: feature density (which determines the percentage of most frequent features removed in pre-processing) and scoring threshold. Low feature density speeds the process significantly and reduces the number of results, which can be raised by lowering the scoring threshold.
The summary of results is given in the table below:

Scoring Threshold	Feature Density	1	2	3	4	5	6	7
3		34	606	4326	9070	16520	56580	28014

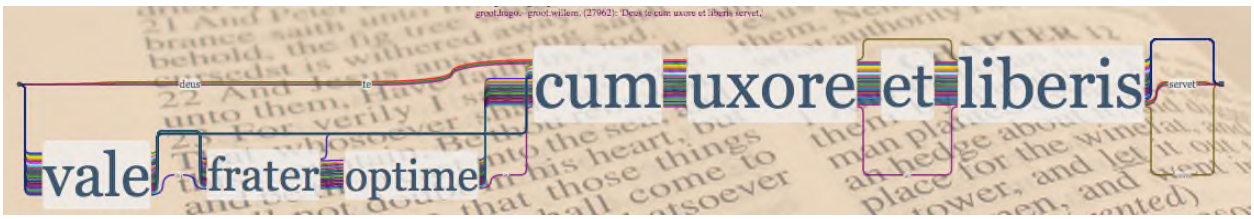
4	32	500	3480	6262	9372	19346	12668
5	26	360	2704	4252	5656	9954	6792
6	20	252	2026	2218	2990	6000	2912
7	12	122	1216	842	1516	1444	1492
TOTAL_NUMBER_OF_LINKS	138	6126	69078	397206	1821058	6251546	19082392

For the preliminary stage of interpretation, a sample of 1216 scored links (out of 69078; the results of running Tracer with feature density 3 and scoring threshold 7) was chosen as being a large enough number, which is still convenient enough for the following stages, which, at present, have to be done manually. For continuation of the work, other combination of these parameters as well as other methods will have to be explored.

The tracer tool does not, at the moment, provide an interface for interpretation of the results. The scored links are given as CSV of the number ID's of the units linked. Using Excel lookup functions, the linked pairs of units were converted from number ID back to the text. From here on, then, the rest of the work was done using external tools:
Looking at the intermediate result, in a table format, I could see that many of them are formulae, and these repeat in several pairs. The need came up to cluster these results; this can be done in stages, by a combination of sorting, counting and lookup functions. I found, however, that a network analysis tool would give a convenient, interesting angle; for this, I used Palladio network analysis tool. See the figure for a view of the resulting graph.
It is interesting to see different clustering patterns in different sets of results, and I intend to explore this further with a more sophisticated tool, which will enable representing weighted directionality and coloring according to changing categories.



Since Palladio does not enable exporting the clustered metadata, I resorted, for this exploratory trial, to a very cumbersome manual solution of screen capturing each cluster and typing in the ID numbers. The results can then be further visualized: each cluster could be inputted into a word-cloud or a word-tree tool, which could give a general impression of its contents. The best tool for this purpose, however, would be an alignment tool. I used TRAVIZ, which was created by Stefan Jänicke and communicates well with the format of Tracer results. Running the tool on the largest cluster from the center of the graph above, for example, we find the formula 'vale (...(frater optime)) cum uxore (et/ac) liberis(...)' visualized as in the figure below, in an interactive interface which enables exploring the variants:



Description of the main results obtained;

The goals of this study can be divided to three types: data collection and interpretation; theoretical advancement, and recommendations for tool development. The two first goals require much more time for application of the tools to the data, and would therefore benefit from further development of tools, the third goal was already advanced significantly, and a general pipeline of required features can already be conceived.

1. Data collection and interpretation:
   In order to interpret the scored and linked results of the text reuse detection applied to the Epistolarium corpus, it would be best to collaborate with a person who has intimate acquaintance of the corpus; a scholar of Dutch late humanism, or specifically of Grotius. It would also be easier to do with improved tools. The number of results is still daunting and hard to handle 'manually'. A preliminary impression could be gained by looking at the top results, by simply sorting them according to the number of common features or the scoring of the suggested links.
   The main challenge in understanding the data, however, would be to categorize the results, and separate the uses of formulae (which are the main bulk of results) from those results that point to quotations, direct or indirect, and ideas running through the network of corresponding humanists. Though some advance was made in the days following the STSM, it is still too early to report on it.
2. Theoretical advancement:
   The phenomenon of text reuse is broad and varied. In the results interpreted so far, some very different types were found: one 'epitextual' note, which was attached to several letters, one duplicity which must have stemmed from errors in coding, a few paragraphs that seems to come from a template of letter writing, and a few - too few, at this point - quotes and paraphrases. The main bulk of the results, at least with the parameters that were used so far, is repeating formulae in the opening and closing of letters. Identifying them calls for several ideas: either add their keywords to an edited stop-word list, or, what I would suggest is much better practice - to mark them up with a refined module for letters, if not simply include them in openers and closers – a practice that would enable improving Topic Modeling and Text reuse detection as well as other forms of text analysis. In either case, though it is tempting to discard formulae as the 'noise', rather than a research results, but in fact the use of formulae in letters is in itself a fascinating subject, which, with distant reading methods can open new ways to understand social phenomena. In looking for a way to identify the formulaic text reuses, I had a hypothesis that was not confirmed: that the larger clusters would point to formulae, while the other sorts of reuse would appear in smaller clusters. It is true that in this corpus, the larger reuses appear only in pairs or triplets, but so do many formulae. And indeed, one could also imagine that a letter template would be used many times, and that a quotation would circle in many letters. The number of common features or the score did not prove as good indicators either. At this point, then, one would have to

characterize types of text reuse without machine help, until a deeper understanding of the parameters is gained, possibly by running more scenarios.

3. Recommendations for tool development:
   In working towards a future digital platform for letters, one can already build on existing tools and check how they should be adapted for correspondence study. In the process of applying the Tracer tool and TRAVIZ reuse visualization tools I found several gaps which I had to "patch" using external tools and manual work. These gaps can be translated to recommendations for the further development of the tools:
   a. As is already known, TRACER requires a convenient user interface, documentation, and in particular a convenient "interpreter" that would enable the user to examine the units in text in VARIOUS stages of analysis, even before the linking and scoring. At the moment they can only be viewed as text from the visualization tools, and this creates a 'black box' feeling for the analyzing process. (In order to fill this gap I used excel functions, and participants in previous TRACER workshops used their own scripts to devise it. This should be internal to the Tracer package).
   b. Once a front end interface is built, it is highly recommended that the user will be able, if interested, to only linking inter-corpora units, and have flexibility in designing the corpora according to the metadata, without the need to return to the data preparation stage.
   c. The third gap consists of several stages: cases of text reuse are very often not singular but repetitive. A clustering mechanism is very beneficial in helping to manage the data and understand it. I used a combination of simple tools, where a more flowing pipeline could bring much better understanding:
      i. Once identified, each cluster should be legible as text units (without having to use Excel) and as text-alignment (see the TRAVIZ figure above)
      ii. An interactive graph will enable organizing manually the clusters, moving closer together those that are judged by user to be of a similar type. Alternatively, this could be done automatically by further clustering stages (grouping clusters that share vocabulary).
      iii. Adding a temporal dimension to the graph and to each cluster is a bigger challenge, but one that would be very beneficial to the study of the phenomenon, whether it is formulae, quotes, or a flow of ideas that each cluster is showing.
   d. TRAVIZ was devised to visualize specific scenarios, and could be much more flexible in adapting to the questions of the user:
      i. This point was raised in a meeting with Marco Büchler and Stefan Jänicke. At the moment, the format of the CSV files, which TRAVIZ accepts is limited to 4 columns, including the ID and the textual unit. This leaves room for only two metadata columns: author and work, or date. When running the corpus/grid visualization tool, only one comparative dimension can be

explored. If another dimension interests the user, the table has to be revised. In correspondence, as well as in other types of corpora, the metadata is often very rich, and each column of metadata may enable an interesting question. In the data I used, for example, authors' or recipients' years of birth/age, and addresses, could be relevant. The same is true for periodical studies, where titles, dates and genre (op-eds, news, advertisements) are relevant facets to compare.

ii. The TRAVIZ alignment tool was found very useful here, but here too I used it for a rather different scenario than originally conceived. Instead of applying it to given versions (as in the bible translations visualization) or to each single link given by the Tracer (as in the alignment embedded in the grid view), it would be helpful to apply TRAVIZ for a chosen cluster, once a clustering mechanism is devised. This might be a serious challenge, but perhaps a clustering mechanism could be introduced already into the alignment code?

e. During the work I often found myself looking for the letters in the Epistolarium interface or in the full XML corpus, in order to interpret the reuse in context. For both the Tracer and TraViz it would be helpful to add the option of linking the unit to a viewer that would enable to see it with its co-text.

f. Finally, a caveat: Tracer is not the only existing platform for text reuse exploration, nor is TraViz the only visualization package. In working towards a future platform for the re-assembled republic of letters it would be necessary to also have a look at alternatives that could be adapted for our use, evaluate and compare them.

Future collaboration with the host institution

Naturally, I continue to consult with the Göttingen based E-Trap team and with Stefan Jaenicke from Leipzig via email. In addition, Marco Büchler and I are working together towards opening an ADHO special interest group for the historical text reuse, based on the community assembled at the mailing list, and in workshops and conferences of the last years.

Foreseen publications/articles resulting from the STSM

Though it may be too early to think of full-blown publication, I definitely intend to continue the work and hopefully present it in some of the coming DH events. I will present the results in the COST conference in Warsaw, and also intend to submit an application to the 3rd Workshop on Computational History (HistoInformatics) in DH2016 in Krakow.

In view to publication, I would prefer to collaborate - with the Tracer team, with tool designers and with humanist scholars, on two or three articles addressing different aspects of the work.

I especially look forward to collaborate with visualization experts following the COST visualization workshop in Como in April.