TNA ACTIVITY REPORT

BUILDING A DIGITAL CORPUS OF VAUDEVILLE TO STUDY HUMOUR: ISSUES AND CHALLENGES

Author: Lara Nugues

Current position: Ph.D. candidate within the SNSF-PRIMA project “Le Rire des vers / Mining the Comic Verse”

Affiliation: University of Basel

Host institution: Trier Center for Digital Humanities (TCDH)

Mentor(s): Christof Schöch, Evgeniia Fileva

Period of stay: February 27th to April 6th

1. Presentation of the project

The project proposal I submitted to the CLS INFRA fellowship is related to my ongoing PhD thesis. I am a PhD candidate in French literature at the University of Basel within the SNSF PRIMA project “Le Rire des vers / Mining the Comic Verse” led by Prof. Dr. Anne-Sophie Bories.[1] This digital humanities project aims to study the links between humour and versification in distant and close reading using XML-TEI tagged and annotated corpora. The statistical study of verse is made possible by a partnership with Richard Renault of CRISCO at the University of Caen, who has developed a software capable of automatically measuring the length of verse[2]. The manual annotation of the corpus is based on a famous humour theory developed by the American linguists Salvatore Attardo and Victor Raskin: the General Verbal Theory of Humour.[3] For my part, I am working on a corpus of 300 vaudevilles (short plays interspersed with songs) between 1830 and 1835 and more particularly on the versified and sung parts of these vaudevilles, also called “couplets”. With the support of the CLS INFRA, I wanted to carry out a project that would allow me to benefit from the expertise and advice of researchers in digital humanities in order to continue the digital constitution of my corpus and to produce a reflection on this constitution by examing the issues at stake. Indeed, creating one’s research material, in this case a digital corpus, conditions the research itself and its results, and deserves a reflexive stance towards it.

One of the first and crucial steps in building a digital corpus is OCR. When texts are not available in plain text, the use of OCR technology is inevitable and often a challenge. It is a

[1] https://slw-comicverse.dslw.unibas.ch/
[2] Delente, Éliane ; Renault, Richard (2005), Projet Anamètre : le calcul du mètre des vers complexes. *Langages* (N°199),125-146.
[3] Attardo, Salvatore (2020), *The Linguistics of Humor: An Introduction*. Oxford: Oxford University Press.

CLS INFRA has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101004984

question of knowing the existing programs and choosing the most suitable according to specific criteria. The second fundamental step is the issue of text preparation. What position should be adopted in relation to the original text? Should one, for example, be conservative and not change any old spellings or error introduced during the publishing process at the risk of not being able to achieve a good part-of-speech tagging? Or should one be more interventionist and modify the text with the risk of losing information? The construction of a digital corpus also requires XML tagging in order to be able to take targeted data and analyse it. This raises the question of the level of granularity of the tagging as well as possible adaptations of the TEI standards. When the researcher doesn't build a digital edition but mines a corpus guided by a research objective, the issues of XML are not quite the same. Finally, when it is planned to annotate a corpus, what is the best way to proceed? With which criteria and which tools? All these questions depend, of course, on the type of documents the researcher is working on, as well as on his or her research objectives, in my case 19th century plays and the analysis of humour. But despite the specificities of each corpus, certain questions are common to those who wish to build a digital corpus. As these questions are often seen as a preliminary stage of a research project, they are rarely examined in detail. As for me, I wished to place them at the heart of my reflection and to take advantage of a stay at the Trier Center for Digital Humanities to develop this reflection and to find concrete solutions for some of the challenges I faced.

2. Research visit and its outcomes

2. 1. How to get the plain text?

My aim during my stay in Trier was to work on 50 plays that I wanted to include in my digital corpus. It should be noted that the vaudeville plays are relatively short, and that I mainly only tag the couplets, which are 10-15 on average per play. These vaudeville plays had been selected beforehand according to specific criteria (authors, place of performance, success of the plays), and their PDF versions have been extracted mostly from Google Books, but also from Gallica, and to a lesser extent from the Internet Archive. I had first tried to OCRize these PDF with Abby Fine Reader, but my attempts were not positive, this software being rather adapted to contemporary texts. When I arrived at TCDH, I was at that point trying to get the plain text provided by Google Books, Gallica or Internet Archive, but with some difficulties. Indeed, all the texts on Gallica are not OCRized and, when the documents are divided into columns, Google's OCR is very often of poor quality – the segmentation not taking this division into account.

My meeting with Johanna Konstanciak, member of the digital humanities project "Mining and Modeling Text" then allowed me to discover OCR4all, a software mainly designed for OCR of printed documents and originally developed by Christian Reul's team at the University of Würzburg.[4] This software can run on mac OS (which I use), its installation remains complicated since it requires a Docker, but once installed it is a very handy software. It works, as desired, with LAREX, Dummy or Kraken for the segmentation phase, and Calamari for the character recognition phase.[5] For the documents in several columns, it is possible to adjust the segmentation, either manually, page by page, or to establish an identical setting for all the pages. But the tests that I carried out on my corpus did not prove fruitful due to the linguistic state of my texts. Indeed, OCR4all offers a single model for French, called "historical French",[6]

[4] https://github.com/OCR4all/OCR4all
[5] For further information, see Reul, Christian; Christ, Dennis; Hartelt, Alexander; Balbach, Nico; Wehner, Maximilian; Springmann, Uwe; Wick, Christoph; Grundig, Christine; Büttner, Andreas; Puppe, Frank (2019). OCR4all - an open-source tool providing a (semi-)automatic OCR workflow for historical printings. *Applied Sciences*, 9(22):1-30.
[6] https://github.com/OCR4all/ocr4all_models

and it turns out that this model built on 17th, 18th and 19th century corpora was not precise enough for my corpus.[7] It is certainly possible to train a more adequate model, but it is an extremely time-consuming operation. Furthermore, the MiMoText team has developed a model for texts from the second part of the 18th century, a period quite close to my period of study. I was able to download their model called "*18th_century_french"* from their Github repository,[8] but unfortunately it is no longer working with the new version of OCR4all. However, it remains a very valuable resource and I intend to find a way to use it in the near future.

Afterwards, I discovered the Tesseract tool thanks to Radoslav Petkov and Frank Queens, both computer scientists at the TCDH.[9] Tesseract was initially designed by engineers from Hewlett Packard (HP), then its development was taken over by Google. It is a less manageable program than OCR4all, since it works with the Terminal command lines. But Tesseract has the advantage of offering two models for French.[10] The first model called "frm" was not suited to my corpus (it concerns the broad period 1400-1600), but the second model called "fra" worked well on my texts. Furthermore, Tesseract also offers the possibility to choose between 13 different automatic page segmentation methods.[11] In my case, the method 1 called "Automatic page segmentation with OSD" gave quite good results, much better in any case than the OCR performed by Google with double column documents.

Along with my research on OCR tools, I also learned at the TCDH what the Double key technique is. Dr. Thomas Burch responsible for several digital edition projects took the time to explain it to me in detail. He taught me that for the Wörterbuchnetz project[12] the manual double entry was carried out by a team in China, while the automatic comparison and quality assurance was performed by a team in Trier. To process the Grimm's Deutsches Wörterbuch, a work of around 300,000,000 characters, it took 15 months, and cost around € 170.000. Although the method is expensive, it can be used, for example, to build up a corpus of only a few tens of thousands of characters in order to train a reusable OCR model on a much larger corpus. This is, for instance, what the MiMoText team did to train its "18th_century_french" model.

For my part, in order to process the 50 plays during my stay in Trier, I opted for the following method. I continued to use OCR results from Google Books, Gallica or Internet Archive, wherever possible. Where this was not possible, either because the OCR was of poor quality in the case of double column documents, or because there was simply no OCR layer, I used Tesseract with either the default page segmentation method or the method 1 ("Automatic page segmentation with OSD"). I was thus able to obtain a corpus in .txt of relatively good quality.

2.2 Cleaning and tagging

Once the corpus was obtained in .txt format, it had to be corrected and tagged. The correction of the text concerns both OCR errors (OCR never has a 100% success rate) and the possible modifications that one wants to make to the text. But one has to decide whether the text should remain the same as the original print or whether it can be modified. I discussed this issue with Dr. Claudia Bamberg when she presented various digital editions projects to me (the Arthur Schnitzler Edition project[13] and the Johann Caspar Lavater Edition project).[14] In the case of digital editions based on a paper edition, the text is kept as it is, as close as possible to the

[7] https://github.com/Calamari-OCR/calamari_models
[8] https://github.com/MiMoText/roman18
[9] https://github.com/tesseract-ocr/tesseract
[10] https://tesseract-ocr.github.io/tessdoc/Data-Files-in-different-versions.html
[11] https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html#page-segmentation-method
[12] https://woerterbuchnetz.de/#0
[13] https://www.schnitzler-edition.net/genetisch
[14] https://jclavater-briefwechsel.ch/home

original, since the aim is to preserve a certain state of the text. However, when it comes to data mining projects, the objective is quite different. It is often a question of extracting data or applying certain analysis software to texts in order to obtain data. Thus, leaving typos, or not modifying certain states of text can affect the quality of the data collected. But there is also a danger in correcting or making changes. Information is lost by changing the original text, and it is also possible to make mistakes by correcting what one thinks is wrong, when in fact it is intentional. The MiMoText team rarely corrects these XML files directly, instead the team uses a modernization pipeline on a .txt version of their corpus.[15] This pipeline makes it possible, for example, to replace long “s” (“ſ”) with modern “s”, or to modify old spellings by means of a word list built thanks to Wikisource[16] and Wiktionary among others.[17] This provides the MiMoText team with a dataset that can be used for subsequent Topic Modeling, Name Entity Recognition, Sentiment analysis. For the moment, I am keeping the original text and do not modernise it. Later on, I plan to build a word list inspired by the MiMoText pipeline to obtain a modernised version of my corpus and thus get better results in POS Tagging. Until now, I make only one exception in my protocol. When certain edition errors are likely to interfere with the measure of the verse length or the recognition of the rhyme scheme, then I make corrections. Indeed, I do not want these edition errors to distort my statistics on vaudeville verse.

Regarding the body markup, the MiMoText team has chosen a very minimalist markup based on the Eltec Guidelines (level 1).[18] This choice is justified by the fact that the MiMoText team does not use XML tags to extract information from the body or to run analysis software. I find myself in a different situation, since the software developed by Richard Renault requires certain standards, mostly TEI compatible, to work. My markup is thus a compromise between the standards required by this software, and other needs that concern other aspects of my research. For instance, I am interested in the tunes to which vaudeville couplets are sung. The names of these tunes are indicated in vaudeville editions. I have chosen, in agreement with my colleague Dr. Nils Couturier from the project “Le Rire des vers / Mining the Comic Verse” (who is also interested in collecting tune names) to mark them with the tag <stage type="tune"> so that they can be extracted and studied.

Concerning the header, the talk “GND und Normdaten für europäische Literatur? Personen und Werke in den multilingualen Korpora von ELTeC” given by Nanette Rißler-Pipka and José Calvo Tello during the Dhd 2023 in Trier led me to review the way I standardized my information on the authors. I most often used the standards of the French National Library (BnF) and those of Wikidata, but I was unaware that there were other standards, such as GND (Gemeinsame Normdatei), and that I could compare different standards thanks to the VIAF (Virtual International Authority File), which “combines multiple name authority files into a single OCLC-hosted name authority service”.[19] Furthermore, Tinghui Duan, member of MiMoText and Phd Candidate, has kindly shared with me a script he has developed that allows him to automatically collect data on authors from Wikibase, such as their date of birth, date of death, place of birth, place of death, or gender.[20] This could be an option to fill in the headers of my XML documents more efficiently.

[15] https://github.com/MiMoText/roman18/tree/master/Python-Scripts/modernization_and_transformation_to_plaintext
[16] https://fr.wikisource.org/wiki/Wikisource:Dictionnaire
[17] https://fr.wiktionary.org/wiki/Cat%C3%A9gorie:fran%C3%A7ais_moderne_d%E2%80%99avant_1835
[18] https://distantreading.github.io/Schema/eltec-1.html
[19] https://viaf.org/
[20] https://github.com/t-duan/dissertation/tree/main/script

2.3 Annotation of humour

On the question of humour and its annotation, Prof. Dr. Claudine Moulin was unfortunately not available to discuss her research on “the cultural-historical negotiation of the comic in the context of German-French language contact”.[21] However, I had exciting exchanges with two historians, Prof. Dr. Damien Tricoire (project leader of Patron and Pamphlets)[22] and Dr. Simon Dagenais (member of the project) on subjects related to the issue of humour, such as the censorship. As they work on French polemical political writings under the Enlightenment, it was thus possible to compare censorial practices at the end of the 18th century and censorial practices under the July Monarchy. I also spoke with Prof. Dr. Damien Tricoire about the Jesuits in the 19th century, who are one of the main targets of humour in my corpus. I learned that their bad reputation was due, for example, to the fact that there were several quarrels between the Jesuits and other orders (Jansenists, Dominicans), among other things on their conception of divine grace, on their policy of inculturation, on their ultramontane doctrines or on their penchant for casuistry.

2. 4. Additional thoughts

Regarding my reflection on digital corpus construction, three new points appeared important to me that I did not have in mind before my stay in Trier. The first point concerns the OCR accuracy rate. No OCR software gives a 100% success rate. Only human intervention, which is very time-consuming, makes it possible to get close to a 100% rate (although humans can also make mistakes). But should one aim for 100%? A margin of error may perhaps be accepted if it does not affect the validity of the scientific results. The challenge would then be to define this acceptable margin of error. The second point concerns the availability of models for OCR software. OCR is a fundamental step that is problematic in many projects. Unfortunately, as far as I know, there are no platforms that centralize OCR models. Each project develops its own model, and one is not necessarily aware of what has been developed by others. Creating such a platform would probably be very fruitful for the advancement of Digital Humanities research. The third point concerns the negative consequence of this lack of availability of resources. By default, researchers therefore tend to use existing data in .txt format to build their digital corpus. This practice raises scientific issues: what values do the chosen texts have in relation to the research objective? How is the question of representativeness handled? More generally, these different points, as well as the other aspects mentioned above, have all been used to feed a chapter of my Ph.D. dissertation dealing precisely with the question of corpus constitution.

3. Considerations over future projects

This stay also had the transversal objectives of increasing my scientific network, broadening my culture in the field of digital humanities, and giving me ideas for future research. This was made possible thanks to Prof. Dr. Christof Schöch who took the time to explain to me the cartography of the digital humanities research in Germany, and invited me to participate in various events. Thus, the workshop "Using and Developing Software for Keyness Analysis" organized by the project "Zeta and company"[23] gave me a better understanding of the notion of “keyness”, also called “distinctiveness”, often used in the field of Authorship Attribution or in software for stylometry, and to meet researchers at the forefront of this field such as Serge

[21] https://tcdh.uni-trier.de/en/event/annual-lecture-german-historical-institute-paris-dhip
[22] https://papa.uni-trier.de/
[23] https://zeta-project.eu/en/

Heiden.[24] In addition, my participation in the workshop "SPARQL für (digitale) Geisteswissenschaftler:innen - Querying Wikidata und die MiMoTextBase" allowed me to become familiar with the SPARQL language and to really get to know the Wikidata database and how it works, which I only knew superficially. I was also lucky enough, thanks to Dr. Joëlle Weis who accepted a late registration, to be able to attend the DHD 2023 which took place in Trier and Luxembourg from March 13th to 17th.[25] I have thus acquired knowledge of recent research in the field of digital humanities in different areas (narratology, musicology, sentiment analysis, intertextuality, culturnomics). I was particularly interested in sentiment analysis and was excited to learn about the use of BERT to build a sentiment classifier. After my thesis, I would like to deepen this field and maybe to compare a double corpus (one in French and one in German) starting from a scientific question that would justify the use of sentiment analysis tools as the MiMoText project does. On the institutional side, my meeting with Dr. Daria Sambuk, early career scientist support officer, also gave me a better insight into the fellowships available in Germany, which will be of great help for my future projects.

To conclude, this stay has fully contributed to my scientific progress and I am very grateful to CLS INFRA for the opportunity that was offered to me and which was enriching both on a scientific and human level. It should be noted that the reception by the TCDH and the University of Trier was of great value, and that the professional exchanges were carried out in an atmosphere particularly favourable to collaboration.

[24] Heiden, Serge (2010), The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. *24th Pacific Asia Conference on Language, Information and Computation*. Sendai, Japon. Retrieved from http://halshs.archives-ouvertes.fr/docs/00/54/97/64/PDF/paclic24_sheiden.pdf
[25] https://dhd2023.dig-hum.de/