How Corpus Analysis Helps Operationalize Research Questions and Entices Literary Scholars to Learn Programming

Cinková, Silvie
cinkova@ufal.mff.cuni.cz
Charles University, Faculty of Mathematics and Physics, Czech Republic

Cvrček, Václav
vaclav.cvrcek@ff.cuni.cz
Charles University, Faculty of Arts, Czech Republic

Janssen, Maarten
janssen@ufal.mff.cuni.cz
Charles University, Faculty of Mathematics and Physics, Czech Republic

Křen, Michal
michal.kren@ff.cuni.cz
Charles University, Faculty of Arts, Czech Republic

The *Skills Gap Analysis*, a recent survey among scholars of all academic stages (Rossum / Šeļa 2022), describes the current distribution of (self-attested) practical text-processing skills and the corresponding education supplies in the Computational Literary Studies (CLS), stating that "[...] involvement in a computational field and corresponding methodological knowledge often are paired in current practice" (p. 7).

The most desired skills included corpus building and text modeling, contrasted to remarkably low scores for corpus analysis and annotation. The respondents apparently do not recognize corpus queries as an information extraction method, which implies that they would not reach beyond the bag-of-words approach with the advanced analytical methods they fancy to use. At the same time, the survey revealed despair from overwhelming skill requirements and the perceived absence of "an entry point or structured path in(to) CLS education" (p. 48).

We argue that corpus analysis is the most natural entry point into CLS education and spotlight it as the "missing link" between an interested scholar and a reasonably proficient text-miner, for the following reasons:

1. It integrates methodological as well as implementation consideration straight in the initial research design: *which concepts am I going to extract for analysis*? *How do I operationalize them*?
2. It generates concrete automation desires and hence naturally specifies the relevant computational skills – typically boiling down to deploying a few data-analytical Python or R libraries – a curriculum manageable even through MOOCs.
3. It encourages further learning: a corpus query language it is in fact programming, yet its operational command turns out to be a low-hanging fruit for most scholars.

This is what we had observed in previous setups and intentionally promoted to the leading principle of our 3-day CLS Infra Training School Data and Annotation (Cinková et al. 2022) in the following building blocks:

1. conceptualization and operationalization of a research question,
2. building and editing a TEI-XML annotated corpus,
3. automatic parsing with Universal Dependencies (Marneffe et al. 2014),
4. understanding the UD annotation scheme,
5. implementation of the initially operationalized concepts into corpus query languages: CQL (Evert / CWB Development Team 2022; Machálek 2020) for linear search and Grew (Guillaume 2021) for tree search,
6. Filtering and aggregation to frequency tables with CQL,
7. Using elementary statistics to estimate differences in frequencies.

Most of the hands-on work by the students was done using TEITOK (Janssen 2018). TEITOK is an online corpus environment for (potentially) heavily annotated TEI/XML documents, sporting a GUI to create, annotate, correct, visualize, and search these TEI documents in a number of different ways and run NLP chains over the TEI files, without having to look at the underlying XML code (although the source code is always accessible). The interface has been fine-tuned over time in response to feedback from the many projects using TEITOK (Janssen 2021) and has proven user-friendly for people with limited computational background.

In the corpus-building and annotation part, each student used source documents according to their own choice that they had been asked to produce for the class. The files could be in any of the 80+ languages for which a UDPipe (Straka / Straková 2022) language model exists. The students learned the basics of TEI-XML using the built-in XML editor in TEITOK (ACE).

To explain the conceptualization and operationalization of research questions, we simulated research of Shakespeare's dramas in DraCor, drawing on the actual research on Shakespeare's style presented in *Think On My Words*, a popular book by David Crystal (Crystal 2008), and on the seminal work on Language registers by Douglas Biber (Biber / Conrad 2009). The students were discussing concepts such as *narrativity* vs. *descriptivity* and how to operationalize them through their linguistic manifestation in texts (e.g., frequency of verbs vs. nouns).

Once the students found out which linguistic information they would need in order to operationalize various rhetorical and stylistic concepts, we explained the fundamental principles of UD for morphology and syntax and let them express their operationalizations more formally with UD tags.

Only then we presented the query languages, focusing on CQL but introducing tree search with Grew as a way to abstract away from the word order. The students learned to use wildcards as well as to formulate constraints on order and number of elements within the queried sequence. In self-paced sessions, they had the opportunity to fiddle with their own corpora in their languages.

Finally, they learned to use the simple web-based statistical cal-culators QuitaUp (Cvrček / Čech / Kubát 2020) and Calc (Cvrček 2021) to compare word frequencies e.g., to extract key words or calculate the precision and reliability of a random sample.

The building blocks in this teaching setup are for the most part not innovative on their own, but together they create synergy.

They spotlight the research questions, presenting corpus building and proper markup as mere means to an end rather than research in its own right. The intimate experience with automatically tagged corpora also teaches the students to deal with noise, and the cumbersome operationalization of abstract concepts teaches them that it is more appropriate to document the way a concept is operationalized, with all its evident limitations (e. g. narrativity based on the ratio of verbs and nouns), than pretending to have a universal grasp of it. The fact that all this NLP is done while keeping the TEI annotation that is of clear relevance within any DH project, makes it an integrated experience rather than additional work.

Although it might seem that promoting information extraction with corpus linguistic methods within the DH community is like bringing owls to Athens and the bespoke Skills Gap survey an anomaly, we have had many interactions with scholars whose main concern with a corpus was the TEI encoding of rather vaguely conceived “NLP features” (implicitly boiling down to named entity recognition), but never its searchability. Hence, the CLS Infra survey confirmed our impression gathered throughout the years rather than becoming a sudden inspiration, and we have drawn pedagogical conclusions from it.

Supported by H2020 101004984 — CLS INFRA.

Bibliography

Biber, Douglas / **Conrad, Susan** (2009): *Register, Genre, and Style*. (= Cambridge Textbooks in Linguistics) Cambridge: Cambridge University Press.

Cinková, Silvie / **Cvrček, Václav** / **Janssen, Maarten** / **Křen, Michal** (2022): *CLS-Infra Trai-ning School on Data and Annotation*. Prague, Czech Republic. https://campus.dariah.eu/resource/events/cls-infra-trai-ning-school-on-data-and-annotation. [18.11.2022]

Crystal, David (2008): *Think On My Words: Exploring Shakespeare’s Language*. Cambridge: Cambridge University Press.

Cvrček, Václav (2021): *Calc: Corpus Calculator.* https://korpus.cz/calc/ [18.11.2022]

Cvrček, Václav / **Čech, Radek** / **Kubát, Miroslav** (2020): *Qui-taUp – a tool for quantitative stylometric analysis.* Czech Natio-nal Corpus and University of Ostrava. https://korpus.cz/quitaup/[18.11.2022].

Evert, Stephanie / **CWB Development Team** (2022): *The IMS Open Corpus Workbench (CWB) CQP Interface and Query Language Manual*. (= CWB Version 3.5).

Guillaume, Bruno (2021): *Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion*. in: *EACL 2021 - 16th conference of the European Chapter of the Association for Computational Linguistics*. Kiev, Ukraine (online).

Janssen, Maarten (2018): *Adding Words to Manuscripts: From PagesXML to TEITOK*. in: *TPDL 2018: Digital Libraries for Open Knowledge* (= Lecture Notes in Computer Science). Universidade do Porto: Springer International Publishing. 152–157. (= Lecture Notes in Computer Science).

Janssen, Maarten (2021): *Integrating TEITOK and KonText/PMLTQ at LINDAT*. in: *Selected Papers from the CLARIN Annual Conference 2020* (= Linköping Electronic Conference Proceedings). Linköping, Sweden: Linköping University Electronic Press, Linköpings universitet. 104–110. (= Linköping Electronic Conference Proceedings).

Machálek, Tomáš (2020): *KonText: Advanced and Flexible Corpus Query Interface*. in: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association. 7003–7008. https://www.aclweb.org/anthology/2020.lrec-1.865.

Marneffe, M.-C / **Dozat, T.** / **Silveira, N.** / **Haverinen, K.** / **Ginter, F.** / **Nivre, Joakim** / **Manning, C.D.** (2014): "Universal Stanford Dependencies: A cross-linguistic typology", in: *Proceedings of the 9Th International Conference on Language Resources and Evaluation (LREC)*: 4585–4592.

Rossum, Lisanne M. van / **Šeļa, Artjoms** (2022): "CLS INFRA D4.1 Skills Gap Analysis", 10.5281/zenodo.6421513.

Straka, Milan / **Straková, Jana** (2022): *UDPipe 2*. http://hdl.handle.net/11234/1-4816.