From Cyclopaedia to Encyclopédie: Using Machine Translation and Sequence Alignment to Identify Encyclopedia Articles across Languages

Glenn Roe, Mark Olsen, Robert Morrissey

Google Colaboratory (Colab): https://colab.research.google.com/

LINCS project: https://lincsproject.ca/

Stanford Named Entity Recognizer: https://nlp.stanford.edu/software/CRF-NER.html

Voyant Tools: https://voyant-tools.org and Spyral: https://voyant-tools.org/spyral

Bibliography

Finkel, J. R., Grenager, T., and Manning C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf (accessed 21 May 2022).

Zafar, H. (2021). Linked Data Conversion using Microservices [video file]. Zenodo. https://doi.org/10.5281/zenodo.6551465 (accessed 21 May 2022).

Land, K., MacDonald, A. and Rockwell, G. (2021). Spyral Notebooks as a Supplement to Voyant Tools. CSDH-SCHN 2021 conference online. http://dx.doi.org/10.17613/2bsr-xp53 (accessed 21 May 2022).

Rockwell, G. and Sinclair, S. (2016). Hermeneutica: Computer-Assisted Interpretation in the Humanities. Cambridge, Massachusetts, MIT Press.

Rockwell, G., Land, K., and MacDonald, A. (2021). Social Analytics Through Spyral. Pop! Public. Open. Participatory. no. 3 (2021-10-31). https://popjournal.ca/issue03/rockwell (accessed 21 May 2022).

Vrandečić, D. (2012). Wikidata: A new platform for collaborative data collection. In Proceedings of the 21st international conference on world wide web, pp. 1063-1064.

From *Cyclopaedia* to *Encyclopédie*: Using Machine Translation and Sequence Alignment to Identify Encyclopedia Articles across Languages

Roe, Glenn
glenn.roe@sorbonne-universite.fr
Sorbonne University, France

Olsen, Mark
MarkyMaypo57@gmail.com
ARTFL Project, University of Chicago

Morrissey, Robert
rmorriss@uchicago.edu
ARTFL Project, University of Chicago

It is well known that the great 18th-century French *Encyclopédie* began first as a modest translation project of Ephraim Chambers' *Cyclopaedia* in 1745. And, although their project grew into something much more significant, the *Enyclopédie* editors (Diderot and d'Alembert) were not shy in incorporating translations of the *Cyclopaedia* as filler for their expanded work. Indeed, as Paolo Quintili remarks, 'the they left a good part of these articles almost unchanged, or with only minor changes' (Quintili, 1996: 75). Given the scale of the two works under consideration, however, systematic evaluation of the extent of the *philosophes'* use of Chambers has remained, even today, a daunting task. John Lough, in 1980, framed the problem thusly: 'So far no one has had the patience to make a detailed study of the exact relationship between the text of Diderot's *Encyclopédie* and the work of Ephraim Chambers. This would no doubt require several years of arduous toil devoted to comparing the two works article by article' (Lough, 1980: 221).

Recent developments in machine translation and sequence alignment now offer new possibilities for the systematic comparison of digital texts across languages. This paper outlines some recent experimental work in leveraging these new techniques in an effort to reduce the 'arduous toil' of textual comparison through automatic translation. In essence, we aimed to generate French translations of *Cyclopaedia* articles and then use sequence alignment to identify similar passages also found in the *Encyclopédie* [1].

We examined two of the most widely-used resources in this domain, Googe Translate and DeepL. Both systems provide useful APIs as part of their respective subscription services, and both provide translations based on cutting-edge neural network language models. While DeepL provided somewhat more satisfying translations from a reader's perspective, we ultimately opted to use Google Translate for the ease of its API and its ability to parse TEI-XML. The latter is of critical importance as we wanted to keep the overall document structure of our dictionaries to allow for easy navigation between the versions.

Our objective here was *not* to produce a good translation of the text, or even one that might serve as the basis for a readable edition. Rather, this machine-generated edition serves as a 'pivot-text' between the two corpora, allowing for an automatic comparison of the two (or three) versions using ARTFL's highly fault-tolerant sequence alignment package, Text-PAIR [2]. In order to determine the parameters for this task, we ran a series of tests with different matching parameters on a representative selection

of 100 articles where Chambers was identified as the possible source. It is important to note that even with the best parameters, which we adjusted to get favourable recall and precision results, we were only able to identify 81 of these 100 articles.

Once settled on the optimal parameters, we then used Text-PAIR to generate both an alignment database, for interactive examination, and a set of static results tables. The alignment database contains some 7,304 aligned passage pairs. The system allows queries on metadata, such as author and article title as well as words or phrases found in the aligned passages. Each aligned passage is presented as a facing page representation and the user can toggle a display of the variations between the two aligned passages. As seen below, the variations between the texts can be extensive (*fig. 1*).

Figure 1. Text-PAIR interface showing differences in the article “Air”.

Text-PAIR also contextualises results back to the original document(s). For example, the following is the article “Almanach” by d'Alembert, showing the aligned passage from Chambers in blue (*fig. 2*).

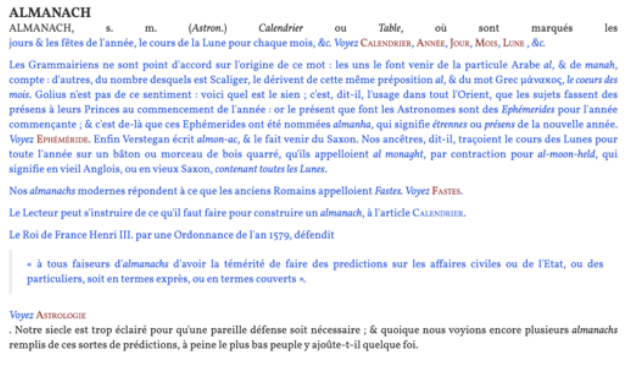

Figure 2. Article “Almanach” with shared Chambers passages in blue.

In this instance, d'Alembert reused almost all of Chambers’ original article “Almanac”, with some minor variations, but does not to appear to have indicated the source of the first part of his article.

To accumulate results and to refine their assessment, we developed an evaluation algorithm for each alignment, with parameters based on the length of the matching passages and the degree to which the headwords were close matches. This simple evaluation model eliminated a significant number of false positives, which we found were typically short text matches between articles with different headwords. The output of this algorithm resulted in two tables, one for matches that were likely to be valid and one that was less likely to be valid, based on these simple heuristics.

In all, we found some 3,778 articles in the *Encyclopédie* that upon evaluation seem highly similar in both content and structure to articles in the 1741 edition of Chambers’ *Cyclopaedia*. Whether or not these articles constitute real acts of historical translation is the subject for another, or several other, articles. There are simply too many outside factors at play, even in this rather straightforward comparison, to make blanket conclusions about the editorial practices of the *encyclopédistes* based on this limited experiment [3]. What we can say, however, is that of the 1,081 articles that include a ‘Chambers’ reference in the *Encyclopédie*, we only found 689 with at least one matching passage, although even here, the recall may in fact be higher than the numbers suggest, given that some citations function more like cross-references. Nonetheless, beyond testing this ground truth, we are also left with the rather astounding fact of 3,089 articles with no reference to Chambers whatsoever, all of which seem to be at least somewhat related to their English predecessor.

Notes:

[1] Our two comparison datasets are the ARTFL *Encyclopédie* and the recently digitised ARTFL edition of the 1741 Chambers *Cyclopaedia*. See https://artflsrv03.uchicago.edu/philologic4/encyclopedie1117/ and https://artflsrv03.uchicago.edu/philologic4/chambers_new/. The 1741 edition was selected as it was one of the likely sources for the translation original project and we were able to work from high quality pages images provided by the University of Chicago Library. On the possible editions of the *Cyclopaedia* used by the *encyclopédistes*, see (Passeron, 2006). On Text-PAIR, see https://github.com/ARTFL-Project/text-pair.

[2] See Clovis Gladstone, Russ Horton, and Mark Olsen, " TextPAIR (Pairwise Alignment for Intertextual Relations)", ARTFL Project, University of Chicago,

2008-2021, and, more specifically, (Olsen, Horton and Roe, 2011).

[3] The question of the *Dictionnaire de Trévoux* is one such factor, as it is known that both Chambers and the *encyclopédistes* used it as a source for their own articles—so matches we find between the Chambers and *Encyclopédie* may indeed represent shared borrowings from the Trévoux and not a translation at all. Or, more interestingly, perhaps Chambers translated a Trévoux article from French to English, which a dutiful *encyclopédiste* then translated back to French for the *Encyclopédie*—in this case, which article is the 'source' and which the 'translation'? For more on these particular aspects of dictionary-making, see our previous article (Allen et al., 2010) and a response (Leca-Tsiomis, 2013).

Bibliography

Allen, T. et al. (2010). Plundering philosophers: identifying sources of the Encyclopédie", J *ournal of the Association for History and Computing* **13.1**.

Leca-Tsiomis, M. (2013). The use and abuse of the digital humanities in the history of ideas: How to study the *Encyclopédie*. *History of European Ideas* **39.4**: 467-76.

Lough, J. (1980). The *Encyclopédie* and the Chambers' *Cyclopaedia*. *SVEC* **185**: 221-24

Passeron, I. (2006). Quelle(s) édition(s) de la Cyclopœdia les encyclopédistes ont-ils utilisée(s) ? *Recherches sur Diderot et sur l'Encyclopédie* **40-41**: 287-92.

Olsen, M., Horton, R. and Roe, G. (2011). Something borrowed: Sequence alignment and the identification of similar passages in large text collections. *Digital Studies / Le Champ numérique* **2.1**.

Quintili, P. (1996). D'Alembert 'traduit' Chambers. Les articles de mécanique de la *Cyclopædia* à l' *Encyclopédie*. *Recherches sur Diderot et sur l'Encyclopédie*, **21**:75-90.

Establishing parameters for stylometric authorship attribution of 19th-century Arabic books and periodicals

Romanov, Maxim
maxim.romanov@uni-hamburg.de
Universität Hamburg, Germany

Grallert, Till
t.grallert@fu-berlin.de
Humboldt-Universität zu Berlin, Germany

The vast majority of articles in Arabic periodicals from the late Ottoman Eastern Mediterranean (c.1850–1918) carried no explicit authorship information (Grallert 2021, Khayat 2019). Yet, the question of authorship has not received much attention in existing scholarship and is strikingly absent from Ayalon (1995), the standard work in the field. The common implicit hypothesis considers editors-cum-owners listed in mastheads and imprints as the sole authors of all the anonymous texts. This results in the conflation of periodicals with the intellectual output of a single person. Such a synonymous use of, for example, "Muḥammad Kurd ʿAlī" (1876–1953) and the monthly "*al-Muqtabas*" (published in Cairo and Damascus, 1906–1918) can be observed across the board (e.g. Seikaly 1981, Ezzerelli 2017). However, the hypothesis a) remains empirically untested, b) negates the known realities of periodical production and individual biographies, and c) ignores specific contexts of individual periodicals.

Computational stylistics or stylometry is a well-established approach in linguistics and literary studies for authorship attribution and genre detection for major languages of the Global North and has been successfully applied in English and German periodical studies (Benatti and King 2017; Kestemont, Martens, and Ries 2019). "Style", in this context, refers to patterns in the distribution of most frequent linguistic features (most commonly, token or character n-grams). This pattern can be captured statistically and then used to identify authorship of specific texts with high accuracy. This identification works through clustering texts by their similarity according to a variety of distance measures (for example, delta, cosine, euclidean, manhattan, etc.; see, Burrows 2002; Eder 2015; Koppel, Schler, and Argamon 2009). The precision of the approach tends to improve with the length of analyzed samples and Eder (2015) recommends at least 5,000 tokens as a safe threshold for meaningful attribution of prose in English, German, Hungarian, and Polish.

Arabic is a prime example for severely under-resourced languages and scripts of the Global South in the digital realm. Infrastructures of methods, tools, and funding often treat Arabic as an afterthought. Consequently, the rich textual heritage of Arabic-speaking and Islamicate societies is largely absent from debates in Digital Humanities (Miller, Savant, and Romanov 2018). Yet, it is one of the major languages of human cultural production. Arabic script is the second most common after the Latin alphabet and is used for 14 modern languages. Among them, Arabic is the fifth most common language globally with more than 420 million speakers in 26 countries.

Our paper presents the first systematic test of stylometry as implemented in the "stylo" package for R (Eder, Rybicki,