Chronotopos
A Journal of Translation History

Ondřej Vimr

Big Translation History and the Use of Data Mining and Big Data Approaches: Panel Report and Observations (EST Congress 2019 in Stellenbosch)

2/2019
DOI: 10.25365/cts-2019-1-2-11

Herausgegeben am / Éditée au / Edited at the: Zentrum für Translationswissenschaft der Universität Wien

ISSN: 2617-3441

Zum Zitieren des Artikels / Pour citer l'article / To cite the article:
Vimr, Ondřej (2019): Big Translation History and the Use of Data Mining and Big Data Approaches: Panel Report and Observations (EST Congress 2019 in Stellenbosch), *Chronotopos* 2 (1), 192-195. DOI: 10.25365/cts-2019-1-2-11

Ondřej Vimr

Conference Report

Big Translation History and the Use of Data Mining and Big Data Approaches: Panel Report and Observations[1] (EST Congress 2019 in Stellenbosch)

The 2019 Congress of the European Society for Translation Studies held in Stellenbosch, South Africa, featured a panel entitled *Big Translation History*. Convened by Diana Roig-Sanz (Open University of Catalonia), Ondřej Vimr (Czech Academy of Sciences & University of Bristol) and Laura Fólica (Open University of Catalonia), it brought together researchers applying Big Data inspired approaches to translation history.

The convenors opened the panel by outlining a Big Translation History approach as one that makes use of data mining, collection and computational processing and aims at visualizing, analyzing and interpreting the data to provide unique insights justifying the approach. Issues proposed for discussion included: a clear-cut definition of Big Translation History, various scales of analysis, the need for a flexible periodization of the corpus, the manageability and feasibility of data collection and visualization as well as the relationship between qualitative and quantitative analysis.

The subsequent presentations revealed a broad scope of interpretations of a Big Data approach to translation history and raised a number of methodological questions fundamental to this underexplored method. Diana Roig-Sanz and Laura Fólica presented their research focusing on the analysis of the literature translated from and into Spanish and published in the Spanish-speaking book market between 1900 and 1945. Their goals included a classification of texts, discovery of yet unidentified originals of translations and visualization of the market of translated books in the Hispanic realm and beyond. The methodological issue of metadata collection, cleaning and reliability when mining data on a global level was particularly pertaining to their research. The potential and challenges related to the mass harvesting and algorithmic

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 749871.

processing of bibliographic metadata was a recurrent issue discussed by both the panelists and the members of the audience.

It reappeared in Ceyda Elgül's contribution (Boğaziçi University) based on a large-scale analysis of bibliographies of biographies translated into Turkish from 1800 to 2020. Among other things, she discussed the challenges of dealing with a dataset covering a time span of over 200 years in Turkish history that included a fundamental change of script with severe implications for an automated processing of bibliographies. Furthermore, her presentation opened the question of what size of a dataset is big enough for Big Translation History. In most scientific fields, Big Data involves processing of terabytes or petabytes of data. In translation history, a well-defined dataset of one thousand entries may already be of relevance at least as basis for a pilot study. Moreover, given the fact that statistical methods have been used in translation history for decades, what is the potential advantage of a Big Data approach? The latter seems to involve processing of a dataset of such an extent that it cannot be analyzed manually and features a method reliant on using automated computational processing.

Josefína Zubaková (Palacký University of Olomouc) echoed the concern over the size, structure and quality of a big dataset in her analysis of the Czech theatrical system drawing on a well-curated bibliography of theatre productions of plays translated from English into Czech since 1989. While her dataset was comparatively small, slightly less than 1000 entries, the quality and depth of information multiplied the analytical options and interpretation avenues. Zubáková's contribution focused largely on individual and systemic traits in translators' biographies. This brought attention to a household topic of translation history for over two decades, which was otherwise suspiciously absent in other presentations. While Big Data researchers keep discussing the challenges of integrating close and distant reading, there is yet another dimension that they might be losing the sight of as they are busy coding, namely that of agency, or people of flesh and bone that make literature and translations happen.

Ninja Steinbach-Hüther (University of Leipzig) presented a team research project investigating geographical societies from 1821 to 1914, which is conducted within a broader framework of the Collaborative Research Centre (SFB) 1199. In cooperation with Thomas Efer and Dirk Hänsgen, she analyzes journals published in 13 languages by 34 geographical societies from Europe, North and South America as well as Asia, Northern Africa and Australia over almost 100 years. The aims of the analysis of over 60000 entries include providing a global and transnational comparison of spatial se-

mantics embedded in the journals' content. She attempted to shed light upon the synchronic and diachronic presence of spatial semantics, its transformation and translation as well as imagined meanings in various languages. Yet at the same time, her presentation tested the boundaries of translation history and showcased a research avenue that integrates translation history, global studies and digital humanities.

Observations

The panel proved very timely as the presenters and audience seemed to face similar methodological challenges while addressing rather different research questions dealing with geographically and historically diverse sets of data. In translation studies, statistical analysis has been the bread and butter of sociological and historical research for a long time with researchers regularly using various databases and bibliographies of translations to introduce topics to translation history and support their arguments. A Big Translation History approach is meant to provide a new Big Data perspective. Yet, analyzing a variety of sources, structured and unstructured data on a large scale with a vision to modify canonical narratives of translation history and comparative literature as well as reveal dynamics in less translated and explored settings, however tempting, is bound to require much effort and face at least three major methodological challenges.

First, mining data on translations from libraries and meta catalogues is largely dependent on the quality of the data in the catalogues. Big Data researchers may tend to control quality using algorithms that check the consistency and possibly improve the quality and enrich the metadata with the help of triangulation of several databases. This may prove insufficient and/or might result in a research on cataloguers and library practices in a given period of time rather than translations, translators, literature or any other actual subject of the research.

Second, the problem of the quality of the underlying data, very common in all large-scale research, is coupled with issues related to the multinational and multilingual nature of translation history. Attempts at combining catalogues from various geographical (and linguistic) settings will inevitably need to deal with the implications of the global technological imbalance. Hypothetically, wealthy countries have more comprehensive and detailed catalogues while poorer countries have patchy and unreliable catalogues. A translation history project aiming at an adjustment of a canonical West-centric narrative by including non-Western datasets from non-Western catalogues, for instance, may unintentionally replicate the canonical pattern due to an

overlap with the existing distribution of funds available in global territories that are necessary to maintain high-quality databases. Unless the quality and balance in the dataset is well controlled, research based on such databases will potentially show biased results mirroring the socio-economic status of the regions included.

Third, Big Translation History presupposes a hefty use of technology and computer analysis. Excellent skills in information technology, statistics and especially coding are a precondition *sine qua non*. Including a data scientist in a research team and offloading much of the technology-related issues sounds inevitable and like a reasonable solution. The drawback, however, is less apparent. The data scientist will become an important gatekeeper of the research project and the other researchers are running the risk of inserting black boxes into their research. Big Translation History researchers will need to get their hands dirty with coding in order to understand what is happening with their data below the surface of the algorithms and be able to correctly interpret the outputs. The acquisition of coding skills will also make them aware of the intrinsic limitations of algorithmic processing of their data (bibliographic or other) and teach them a lesson of research feasibility, for instance what kind of a dataset may or may not produces consistent results.

Overall, the panel showed great promise in terms of providing new questions, methodologies, producing and using new datasets as well as opening up to other bordering disciplines such as global studies and digital humanities. Yet, researches will need to make sure they also develop methods to bridge the gap between the digital and more traditional translation history.

The book of abstracts that includes the panel description and its contributions is available at https://www.est2019.com/.