Theoretical Basis and Creation of a Parallel Corpus: Delineating New Perspectives in the Processing of Library Material

Dragana Stolić
stolic@unilib.rs
University Library "Svetozar Marković" Belgrade, Serbia

PAPER SUBMITTED: 24 November 2022
PAPER ACCEPTED: 31 March 2023

1 Introduction

Jelena Andonovski's publication *A Parallel Serbian-German Corpus of Literary Texts: Production, Information Retrieval, and the Semantic Web*[1] joins the valuable publishing collection of the University Library "Svetozar Marković," launched in 2016 under the name "Homilies." Within the collection are published scientific monographs, most of which are edited doctoral dissertations defended in the social sciences and humanities, i. e. digital humanities. In the text we are presenting the adapted text of the dissertation *Linked Open Data and Language Resources in Creating Serbian-German Literary Corpus*, which the author defended at the Faculty of Philology in Belgrade in 2020.

The publication presents a highly precise systematization of knowledge and skills related to the processing, description and exploration of linguistic corpora in order to create basis for the latter lexical, terminological and semantic analyses. Gradually and logically, the work leads us to an understanding of this specific segment of librarianship and information sciences, outlines a model to be applied when approaching this task, not forgetting to indicate the paths of further research. In general, these kind of publications help employees in the library and information industry to see more clearly the entire domain of research, which, conditioned by technological progress and the development of new tools, has enabled analytical penetration into the publications themselves, joining them to available superior units and

1. Jelena Andonovski: *Paralelni srpsko-nemački korpus književnih tekstova: izrada, pronalaženje informacija i semantički veb*, Beograd: Univerzitetska biblioteka „Svetozar Marković", 2021.

providing previously impossible approach to literary works and writers. Just as classical librarianship implies a significant number of various activities, collection, processing and presentation of publications in order to provide the user with the desired source of information, so new information technologies allow the analysis and processing of language materials in order to understand the nature of language and its use and to pave paths for easier translation and understanding of other languages.

2 A Road to Corpus Research

Although the publication has several chapters, two major units can be observed: one, which we can call the theoretical or *conceptual starting point* and the other, which included the application of all the above to a particular corpus (SrpNemKor), which implied the use of certain tools, conducting searches and analysis of the obtained results.

The mentioned first part of the publication first of all defines the concept of *corpus*, as the basis of *corpus linguistics*, and within that – *parallel corpora*. This part provides a clear typology of corpora, which can be monolingual and multilingual, which can further be divided into *parallel* and *comparative*. Parallel corpora are composed of texts in the original language and their translations into one or more target languages and as such are important for learning foreign languages, various translations, especially machine translations, then terminological and linguistic research, etc. It was pointed out that the basic element of a bilingual corpus is a *bitext* or a *parallelized text*, where the connection of the text and its translation is carried out at different levels – from the entire document, through chapters, paragraphs to sentences and words, with the aim of *aligning texts* (*parallelization*) or creating links between variant units of translation and building a set of translation units ("Translation Unit").

After a review of the first known attempts to form parallel corpora, an entire chapter is devoted to achievements in Serbia in this field, primarily to the work of the Association for Language Resources and Technologies – JeRTeh[2]. In this way, a short and concise, but very informative overview of the results of multi-year projects was provided.

The procedure for creating parallel corpora is presented step by step, which significantly facilitates the understanding of the scope of this approach, where annotations have special significance, mostly those that include the use of the EXtensible Markup Language (XML) and increasingly

2. JeRTeh

present, but it seems still insufficient accepted, Text Encoding Initiative (TEI). Marks (tags) are added to the selected segments, which assign additional descriptors for Part-Of-Speech, canonical form, or morphological categories to certain words. Description of the material preparation is then followed by a presentation of the prerequisites for retrieving information, where properly assigned metadata plays a key role. The term "metadata", which as a matter of fact takes precedence in modern librarianship over the classical bibliographic data, is carefully explained in all forms of its manifestation (descriptive, structural, administrative) and by presenting basic schemes (Dublin Core, METS, MODS).

At the very end of this fundamental segment, in which definitions are formulated and the entire system of means necessary to create a corpus is outlined, the most important and complex concept of *Semantic Web* is explained. Professionally, but understandably enough, this part of the work grapples with the "Linked Open Data" initiative, trying to clarify the functioning of the web as a global database whose structure is no longer based on static pages, but through a mechanism of linking data at the level of their meaning, that is, it is based on content connected according to semantic characteristics.

3 Creation of the Serbian-German Parallel Corpus

After an exhaustive presentation and analysis of the tools and resources that can be used in the creation of a parallel corpus, the second part of the study (Chapter 6) deals with the application of all of the above, considering the possibilities, scopes or limitations of some resources on a chosen material. To form the corpus, seven novels by different Serbian writers and the same number of works by different writers from the German-speaking areas were used, guided by criteria such as: award-winning authors, availability of works in both languages in Serbian libraries, popularity of works and their volume.

The steps in the creation of the corpus are shown, from the digitization of the material, the application of optical character recognition, structural annotation and, finally, to the placement of the corpus in the *Bibliša* digital library,[3] the provision of conditions for sophisticated information retrieval, as well as the inclusion of the processed corpus in the network of open linguistic data. The whole presentation is an authentic illustration of all the necessary actions, but also challenges, problems and limitations, which does not diminish the relevance of this work.

3. Biblisha

4 The Role of Libraries

This study practically, through the integration of several elements, created itself a new "corpus" of concepts and meanings that look at the text in a different way. In this sense, it seems that classic library and information activities do not necessarily have points of contact with such an approach. However, the special quality of this work is precisely the consistent emphasis on the indispensable role of libraries, which, despite their dominantly "non-semantic" approach, must be part of this process, not only because of digitization, where they most actively participate, or the fact that they possess materials that are the starting point for creation of the corpus. Namely, the libraries, in which various metadata and printouts are already created, have the prerequisites to make an appropriate step towards the meaningful linking of the obtained data and, therefore, the provision of more complete information for users.