PeregrXML:
The challenges posed by a 17th century corpus of Medieval Christian Hebrew

Tamás Biró
Jewish Theological Seminary – University of Jewish Studies (OR-ZSE)
Eötvös Loránd University, Budapest (ELTE)

Our ongoing project "Hebrew Carmina Gratulatoria of the Hungarian Peregrines in the Seventeenth Century" (K 125486, National Research, Development and Innovation Fund of Hungary [NKFIH], 2017–2021) focuses on gratulatory poems written in Hebrew by Hungarian protestant students studying theology at Dutch and German universities ("peregrines") in the seventeenth century. These poems – together with similar poems written in Latin, Greek, Syriac and other languages – were recited at public events, such as theses defences, and published in print subsequently.

The project consists of collecting, analysing and publishing these poems. First, we bring together photocopies from 350-year-old rare books dispersed around the globe. Second, we transcribe them, normalize them and translate them. Third, we pose and attempt to answer research questions on the linguistic, literary, cultural and religious aspects of these compositions. The project aims at better understanding the linguistic skills and cultural backgrounds of the peregrines, their social network, the influences of various periods of the Hebrew language, the use of biblical and rabbinic sources, the contemporaneous educational systems and theological debates. Examples relevant to general theories in literary, linguistic and cultural studies (intertextuality, language transfer in foreign language production etc) are also of interest.

Building a serviceable corpus lies at the centre of the project, which could also serve in the future as a sample for an unusual variety of the Hebrew language, viz. "Medieval Christian Hebrew". This short paper shall focus on the challenges posed both by the input-side and by the output-side.

On the input-side, the transcription of the photocopies raises difficulties. The quality of the picture and the quality of the original edition might both be suboptimal, typos occur not infrequently, and the imperfect Hebrew skills of the authors have also introduced textual problems. Since these poems were published only once – and with low circulation, at that – we have no access to second, corrected editions. Most texts appear in Hebrew script, with or without vocalization, but a few ones in transliteration with Latin characters. Hence, normalization is indispensable. Yet, normalization introduces arbitrary decisions, and it is unclear to what extent we should compensate for the authors' lack of linguistic competence. It all depends on the research questions to be posed. At the same time, non-linguistic aspects of the texts – the typography of the poetic structures – are very clear. Therefore, we introduce PeregrXML, a markup language that allows for linguistic uncertainty on several levels (the original text, the normalized text and the interpreted text), while maintaining the poetic structures.

Finally, on the output-side, our corpus should help answer an extremely broad and open range of research questions, within the current project, but also by ourselves and others in the future. The corpus will be annotated for several features, while other features will be introduced only on-demand, should a specific research require them. Finding the balance between excessive and insufficient investments into corpus building is a major issue, as always in DH.