Introducing Catch 2020:
Computer Assisted Transcription of Complex Handwriting

Wout Dillen and Dirk Van Hulle

For a long time, the idiosyncrasy of handwriting with its marked differences in character shapes between writers (or even within a corpus of a single writer over time), lead many to believe that automating their transcription through OCR would be impossible. More recently, however, computer science has caused a breakthrough in this area, allowing us to use artificial neural networks to apply deep learning techniques in the field of computer vision ('Continuous Handwritten Script Recognition', Frinken and Bunke 2014). The most promising tool in this field is Transkribus (https://transkribus.eu), which has developed a robust HTR engine, an elaborate platform with a wide range of services for humanities scholars, archivists, librarians, public users, and computer scientists alike, and an elaborate consortium (READ, or Recognition and Enrichment of Archival Documents) that includes 50 partner institutions.

This poster presents CATCH 2020, a project at the University of Antwerp that recently received funding by the Flemish Research Foundation (FWO) to develop an API that builds on the existing Transkribus platform in order to facilitate the computer-assisted transcription of more complex handwritten documents by the end of the year 2020. Rather than producing flat transcripts of digital facsimile images (the default output of OCR and HTR engines), CATCH 2020 aims to produce structured texts by providing tools to add (1) textual and (2) linguistic dimensions to the transcription – thus combining the state of the art of the research field of textual scholarship with the state of the art of the research field of computational linguistics. (1) For the former, the proposed infrastructure will provide tools for (semi-) automatically identifying textual features on the document (i.e. layout analysis, such as additions, deletions, or structural elements such as paragraphs or stanzas). (2) For the latter, the proposed infrastructure will use linguistic and stylistic information to improve Transkribus's transcription algorithm. This combination will allow for the automatic generation of qualitative, structured digitized text that may serve as a sound basis for further literary, linguistic and historical research.

Lead by ACDC (the Antwerp Centre for Digital humanities and literary Criticism), this is a multidisciplinary project that is developed in collaboration with CLiPS, GaP, ISLN, the CSG, and with external partners Transkribus and the Antwerp based Flemish literary archive Letterenhuis.