RESEARCH

04/24/2019

Computer reads old texts

The software from Würzburg scientists can decipher digitized historical prints with almost no errors.

Photo: Klaus-Dietmar Gabbert, dpa

Würzburg scientists have developed a tool that deciphers historical prints with almost no errors. where the problems lay

Scientists at the Julius-Maximilians-Universität Würzburg (JMU) have developed a computer tool for ancient texts. It converts digitized historical prints into a computer-readable text with an error rate of less than one percent, as the university announced. In addition, OCR4all offers a graphical user interface that requires no IT expertise. The new electronic tool was designed under the direction of Christian Reul with computer science colleagues and many students.

According to the announcement, OCR4all has its roots in the university's Kallimachus joint project, which is funded by the Federal Ministry of Education . This cooperation between the humanities and computer science will be continued and institutionalized in the newly founded Center for Philology and Digitality. During the development, the computer scientists worked closely with the humanities disciplines at the JMU, including German and Romance studies. The aim there was to digitally process the "Ship of Fools", a moral satire by Sebastian Brant from the 15th century.

According to Reul, one of the biggest problems was the typography. One of the reasons for this is that the first printers of the 15th century did not use uniform fonts. "Their printing stamps were all self-carved, each printer practically had its own letters and symbols." In old prints, it is often not easy to distinguish between e or c, v or r. However, software can learn to recognize such subtleties.

It is said that Reul has also convinced external partners of the quality of the research. With the "Center for Digital Lexicography of the German Language" in Berlin, Daniel Sanders' "Dictionary of the German Language" has been made

digitally accessible. This work often contains different fonts per line of text, each of which stands for different semantic information. Here, the existing approach to character recognition has been expanded in such a way that, in addition to the text, the typography and thus the complex content structure of the lexicon can also be mapped exactly. (kna)