OCR4all: Modern tool for old texts

April 24 2019

Comme deux folz car ceſt tout vng
Si lung eſt bien noir lautre eſt brun
Se lung eſt fol lautre leſt plus
Si lung boiteux lautre perclus
¶ Jadis eſtoit grande louange — iij. Regū. x.
Qui maintenant ſeroit eſtrange
Aux anciens peres porter
Grande barbe. et deuez noter
Que a leyemple de ſocrates — Socrates
Tous les philozophes apres
Et auant quil fuſt mort portoient
La barbe grant leſquelz eſtoient
Rempliz de grande ſapience
Mais euly decedez leur ſcience
Sop voyant par nous contempnee
Sen eſt laſſus au ciel volee
Nous laiſſant ca bas tous inhertes
Dont ſont innarrables les pertes
¶ Libidinite corrumpue
A commence faire repue
Et tenir ſon cours par le monde
Toutes vertuz et choſe monde
Dont les haultz cieulx ſont decorez
Ny ſont plus mais ſont demorez
Tous mauly tous vices et pechez
Dont les humains ſont entachez
¶ Tout le monde ſe contrefait
Et veullent ce que dieu a fait
Par preſumpcion contrefaire
En cuidant mieulx que dieu les faire
Qui eſt vng peche par trop grant
Honteux ſont et honte les prant
De porter grant barbe au viſaige
De peur de monſtrer leur vieulx aage
Mais leurs corps et viz ſi bien gardent
Si bien les acoutrent et fardent
Que iamais ne deuiennent vieulx
Se ſemble. et auſſi leurs cheueux
Les vngz comme ſicambriens — Sycambri.
Et comme les ethiopiens — Ethiopes.
Les portent tous craiſpes et tors
Faiſant a nature grans tors

Ce qui fut iadis vicieux
Criminel ignominieux
Aboli laiſſe reprouue
Mat. xviij. Eſt par les aultres approuue
Maintenant en ce preſent aage
Par les aultres nouuel vſaige
Nouueau rit couſtume nouuelle
Eſt gardee la choſe eſt telle
Mais ie ne puis pas bien penſer
En mon cueur ne contrepenſer
Lequel eſt le plus fol des deux
Vſant du nouueau rit et vieux
Du cil qui prent les grans coudetz
Comme font vng tas de lourdetz
Du cil qui porte manches larges
Comme font maintenant les paiges
Si non de dire quil me ſemble
Que mectre ſe peuent enſemble

Page from a french version of the "Narrenschiff" (Ship of Fools). Such old fonts

can be reliably converted into computer-readable text with OCR4all. Credit: Dresden State and University Library, CC BY-SA 4.0

Historians and other humanities' scholars often have to deal with difficult research objects: centuries-old printed works that are difficult to decipher and often in an unsatisfactory state of conservation. Many of these documents have now been digitized—usually photographed or scanned—and are available online worldwide. For research purposes, this is already a step forward.

However, there is still a challenge to overcome: bringing the digitized old fonts into a modern form with text recognition software that is readable for non-specialists as well as for computers. Scientists at the Center for Philology and Digitality at Julius-Maximilians-Universität Würzburg (JMU) in Bavaria, Germany, have made a significant contribution to further development in this field.

With OCR4all, the JMU research team is making a new tool available to the scientific community. It converts digitized historical prints with an error rate of less than one percent into computer-readable texts. And it offers a graphical user interface that requires no IT expertise. With previous tools of this kind, user-friendliness was not always a given, as the users mostly had to work with programming commands.

Developed in cooperation with the humanities

The new OCR4all tool was developed under the direction of Christian Reul together with his computer science colleagues Professor Frank Puppe (Chair of Artificial Intelligence and Applied computer science) and Christoph Wick as well as Uwe Springmann (Digital Humanities expert) and numerous students and assistants.

OCR4all originates from the JMU Kallimachos project, which is funded by the German Federal Ministry of Education and Research. This cooperation between the humanities and computer science will be continued and institutionalized in the newly founded JMU Center for Philology and Digitality.

In developing OCR4all, computer scientists have collaborated with the humanities at JMU—including German and Romance studies and literature studies in the project "Narragonien digital." The aim was to digitize the "Narrenschiff," a moral satire by Sebastian Brant, a bestseller of the 15th century that was translated into many languages. Furthermore, OCR4all has been frequently used in the JMU's Kolleg "Medieval and Early Modern Times."

OCR4all is freely available to the public on the GitHub platform (with instructions and examples): https://github.com/OCR4all

Each print shop had its own font

Christian Reul explains the challenges involved in the development of OCR4all: Automatic text recognition (OCR = Optical Character Recognition) has been working very well for modern fonts for some time now. However, this has not yet been the case for historical fonts.

"One of the biggest problems was typography," says Reul. 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 carved by themselves, each printing house practically had its own letters."

Error rates below one percent

Whether "e" or "c," whether "v" or "r"—it is often not easy to

distinguish in old prints, but software can learn to recognize such subtleties. To do so, it has to be trained on sample material. In his work, Reul has developed methods to make training more efficient. In a case study with six historical prints from the years 1476 to 1572, the average error rate in automatic text recognition was reduced from 3.9 to 1.7 percent.

Not only was the methodology was improved, JMU computer scientist Christoph Wick has also decisively further refined the technical component by developing the Calamari OCR tool, which is also freely available and has since been fully integrated into OCR4all, promising even better results. Now, even for the oldest printed works, error rates of less than one percent can be achieved in general.

Lexical projects

Reul has also convinced external partners of the quality of Würzburg's OCR research. In cooperation with the "Zentrum für digitale Lexikographie der deutschen Sprache" (Berlin), Daniel Sanders' "Wörterbuch der deutschen Sprache" (Dictionary of the German Language) has been digitally indexed, and a scientific publication on this work is currently being prepared. The various lines of this text often contain different fonts, representing different semantic information. Here, the existing approach to character recognition was extended in such a way that not only the text but also the typography and thus the complex content structure of the lexicon may be reproduced very precisely.

The computer scientist from Würzburg will soon complete his doctoral thesis, but he is also willing to continue working with OCR in the future: "The computer science behind OCR is extremely exciting," he says. A possible project in the near future: the creators of the "Idiotikon," a dictionary of the Swiss-German language, have indicated their interest in

collaboration since they might well need the Würzburg's specialist knowledge.

More information: github.com/OCR4all
github.com/Calamari-OCR

jlcl.org/content/2-allissues/1 … 18/jlcl_2018-1_1.pdf
jlcl.org/content/2-allissues/1 … 18/jlcl_2018-1_4.pdf

Provided by University of Würzburg