READ
Recognition and Enrichment of Archival Documents

Public Models in Transkribus

Last update of this guide: 29.08.2019

This paper should give you an overview of the publicly available models in Transkribus we offer so far. You will find a short description of the training material, which languages the model can be useful for and who about the project/institution/person who/which has created and trained it. We are working on making more models available for Transkribus users in future, so they can benefit from the network effect and save work and time.

Download the Transkribus Expert Client, or make sure you are using the latest version:

- https://transkribus.eu/

Consult the Transkribus Wiki for further information and other How to Guides:

- https://transkribus.eu/wiki/

Transkribus and the technology behind it are made available via the following projects and sites:

- https://read.transkribus.eu/
- https://transcriptorium.eu/
- https://github.com/transkribus/

Contact

- The Transkribus Team: email@transkribus.eu

Contents

The READ project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943.

Dutch handwriting– National Archives Netherlands

The digitisation team around Liesbeth Keyser from the National Archives in the Netherlands is working hard on creating training data for their collections in order to prepare HTR processing on a large scale. As a first result a model based on 475.769 words is now made available for Transkribus users. The model shows a Character Error Rate of 7.48% on the training set and 6.15% on the validation set. It is based on the careful transcription of dozens of different handwritings and comprises scans from the Incoming Documents from the Dutch East India Company (Overgekomen Brieven en Papieren van de VOC) of the National Archives of the Netherlands and of 19th century Notarial deeds from the Noord-Hollands archief. The model is named: NAN/NHA_GT_M3+ Enjoy!

Russian Church Slavonic – Achim Rabus (University of Freiburg)

Prof. Achim Rabus from the University of Freiburg has released two specialized models which are able to read Russian Curch Slavonic. The first model is called VMC_Test_4+: Training data consist of parts of the Russian Church Slavonic Great Reading Menology (16th century). The model is tailored towards transcribing Cyrillic semi-uncial script from the 16th century. Character Error Rates for the training data are 3.72% and for the validation set 3.92% and for the validation set 3.82%.

The second model is called: Combined_Full_VKS_2: Training data consist of parts of the Russian Church Slavonic Great Reading Menology (16th century), Old Church Slavonic Codex Suprasliensis (11th century), and the 11th century manuscript of the Catecheses of Cyril of Jerusalem. This is a generic model suitable for transcribing a variety of Old Cyrillic script styles including uncial and semi-uncial. Character Error Rates for the training data are 4.42% and for the validation set 3.92%.

Achim has written a detailed report about his usage of Transkribus. Though it deals with Church Slavonic it is definitely interesting for other users as well. Thanks a lot!

Fraktur – Austrian National Library and NewsEye project

Thanks to the Library Labs of the Austrian National Library and the NewsEye project we are happy to announce the release of a free model which is capable to read German Fraktur documents especially from the 19th and 20th century in a convincing quality outperforming most standard OCR engines. The model is based on training data coming from the ANNO collection of the Austrian National Library and comprises 442.141 words. It shows a CER of 1,55% on the training set and 1,65% on the test set without any dictionary support. Note: the model is trained on German language documents. It will provide less convincing results for other languages, such as Swedish or Finnish Fraktur. However models for these languages are also in preparation and may be released in the coming months. The Fraktur model is available for every registered user in Transkribus and called: ONB _Newseye_GT_M1+. Have fun!

Credits

We would like to thank the many users who have contributed their feedback to help improve the Transkribus software.

Transkribus is made available to the public as part of H2020 e-Infrastructure Project READ (Recognition and Enrichment of Archival Documents) which received funding from the European Commission under grant agreement No 674943.