Centre for
Manuscript Genetics
University of Antwerp

Transkribus in Practice

...

The University of Antwerp's Centre for Manuscript Genetics

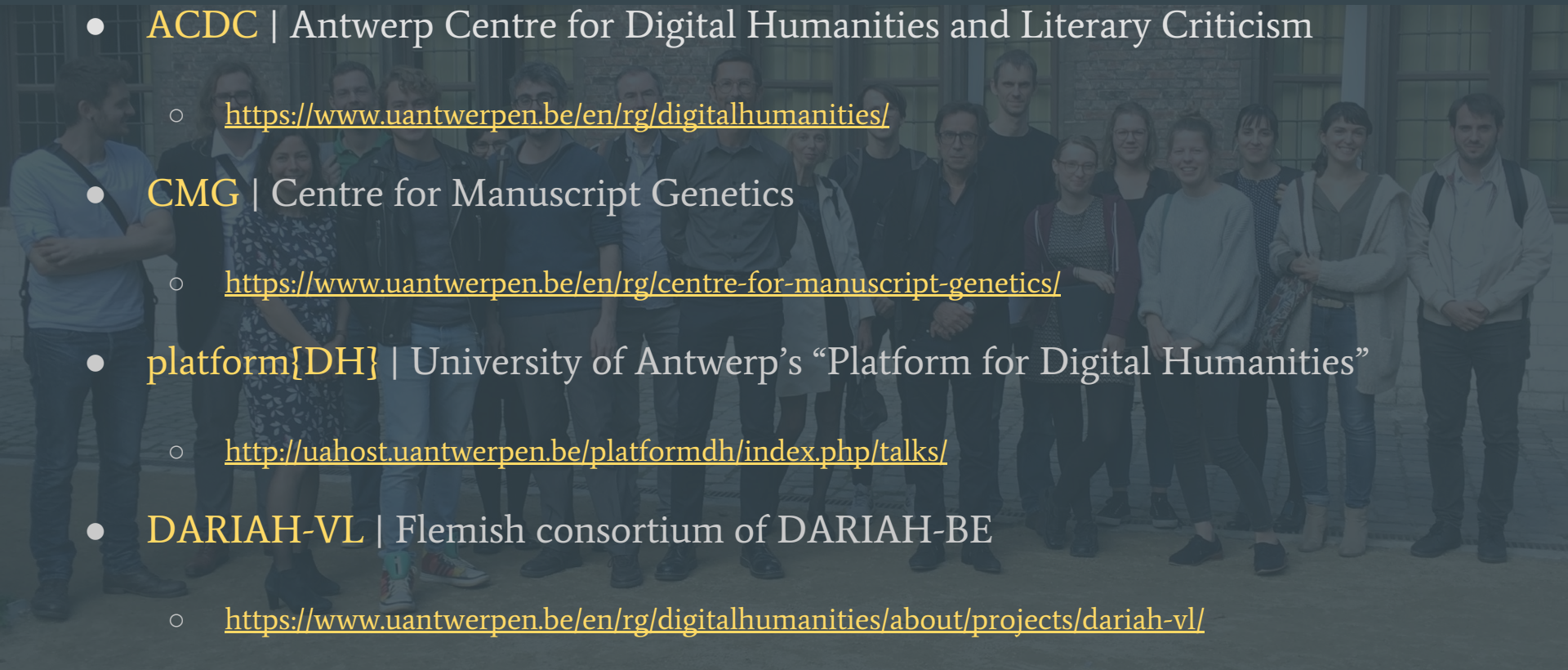
Wout Dillen

@WoutDLN | @CMG_UA | @platform_dh

Who are we?

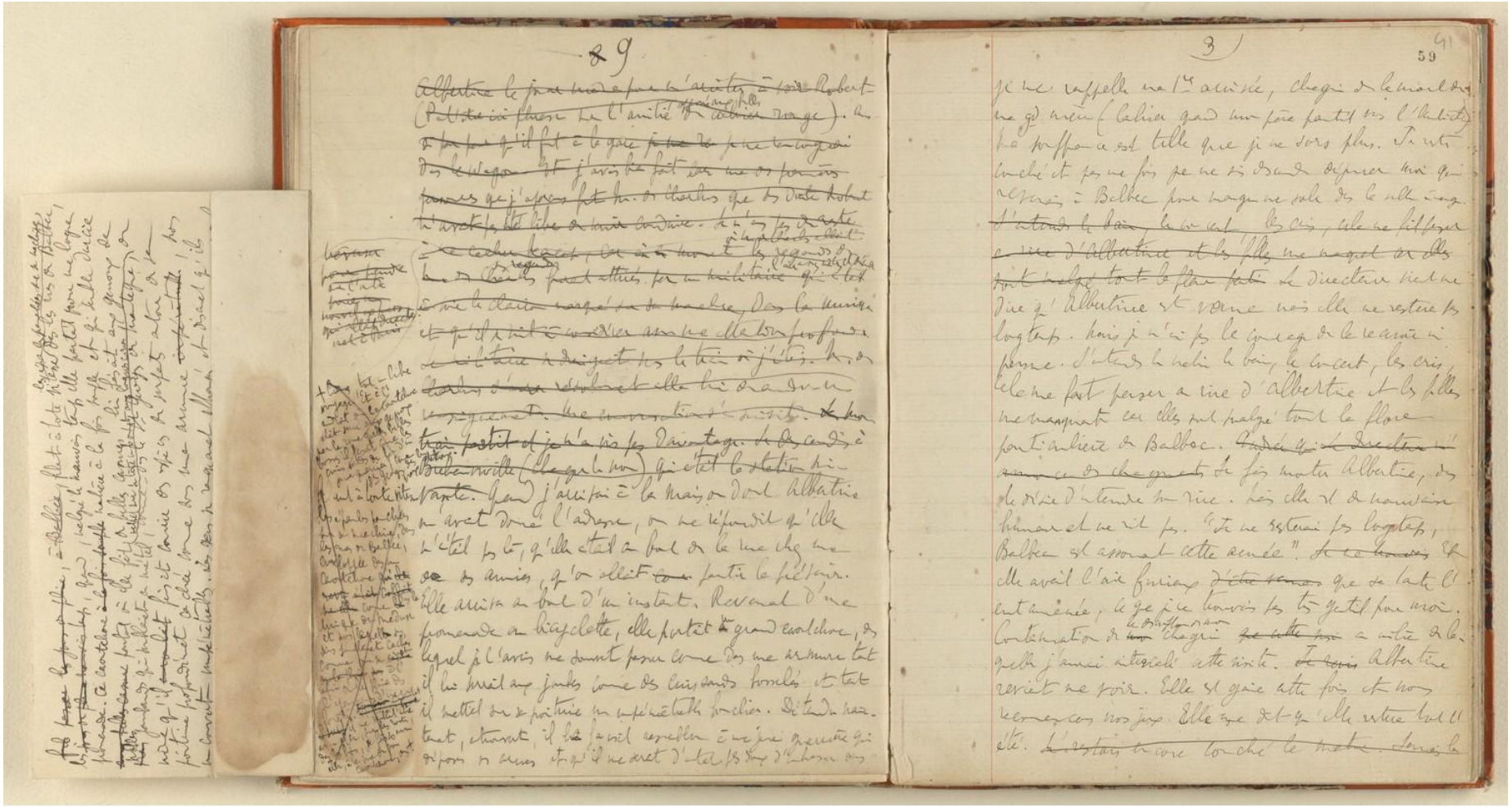
Who are we?

Who are we?



- ACDC | Antwerp Centre for Digital Humanities and Literary Criticism
  - https://www.uantwerpen.be/en/rg/digitalhumanities/
- CMG | Centre for Manuscript Genetics
  - https://www.uantwerpen.be/en/rg/centre-for-manuscript-genetics/
- platform{DH} | University of Antwerp's "Platform for Digital Humanities"
  - http://uahost.uantwerpen.be/platformdh/index.php/talks/
- DARIAH-VL | Flemish consortium of DARIAH-BE
  - https://www.uantwerpen.be/en/rg/digitalhumanities/about/projects/dariah-vl/

What do we do?



Marcel Proust. II — À LA RECHERCHE DU TEMPS PERDU. Notebook 46, pages 58v-58Av-59r

Samuel Beckett

Digital Manuscript Project

www.beckettarchive.org

How do we use Transkribus?

How do we use Transkribus?

Antwerp Academy in DH 2016: Demystifying Digitisation: A Hands-On Master Class in Text Digitisation with Transkribus Workshop.

Since then:

- +/- 400 pages of ground truths
- 350 of which using the 'image2text' tool

Bilingual corpus, so two algorithms:

- English: 11% character error rate on test documents
- French: 18% character error rate on test documents

How would we like to use Transkribus?

How would we like to use Transkribus?

Modern manuscripts: constructed for personal use (not dissemination / publication)

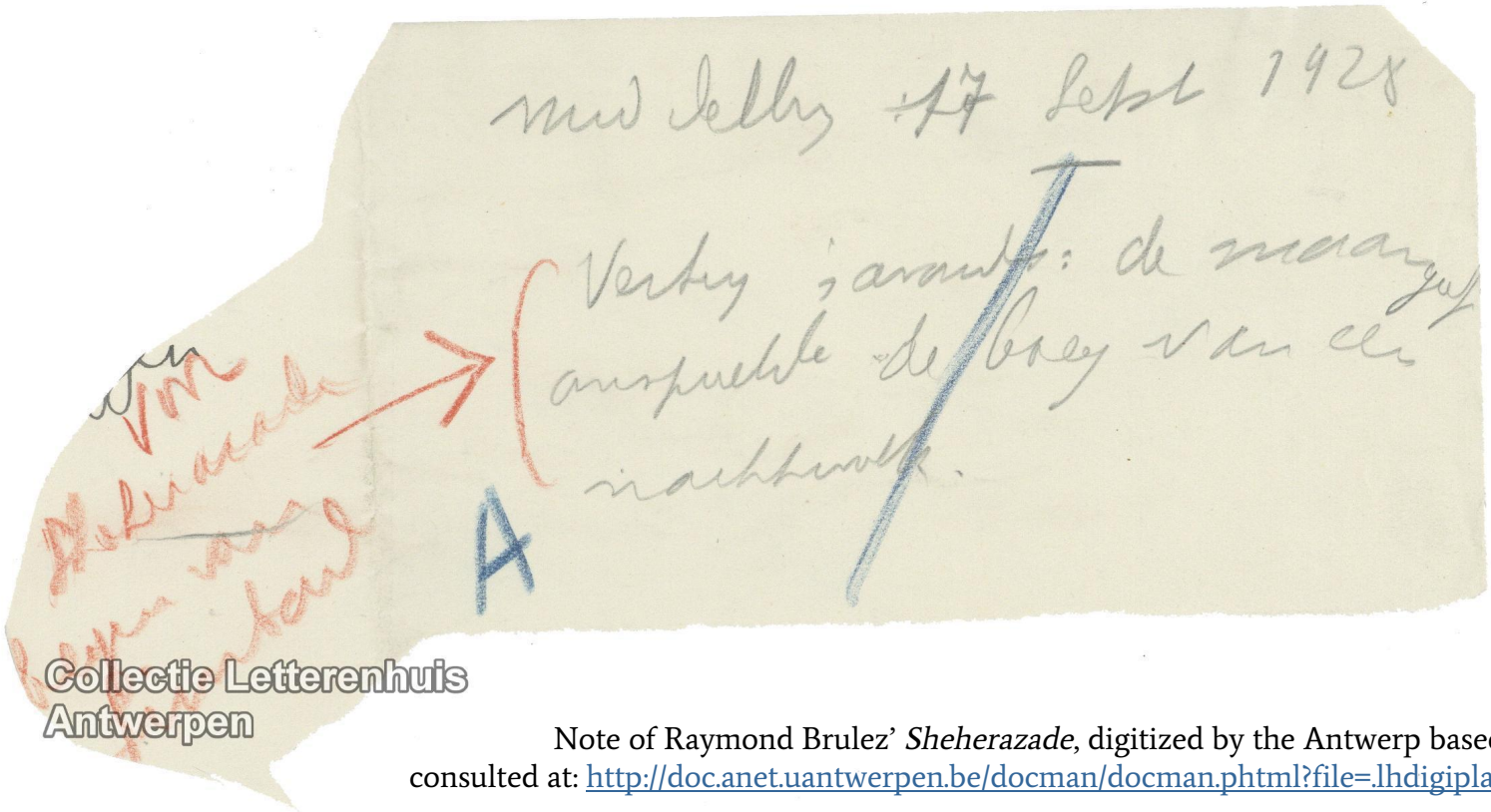
→ bad handwriting (and prone to change)

→ multi-layered (e.g. writing stages)

→ internal logic (e.g. metamarks)

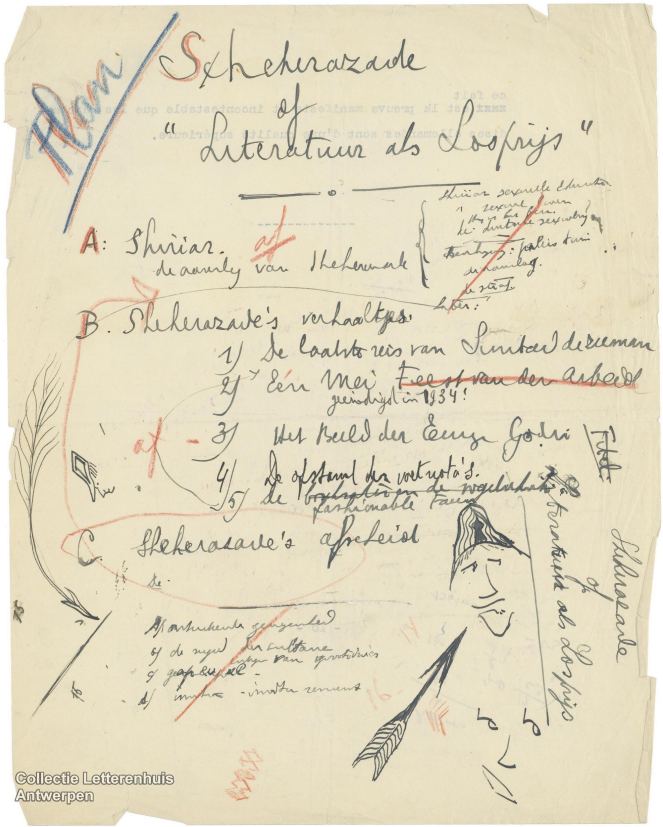
→ high concentration of noise (e.g. heavy deletions)

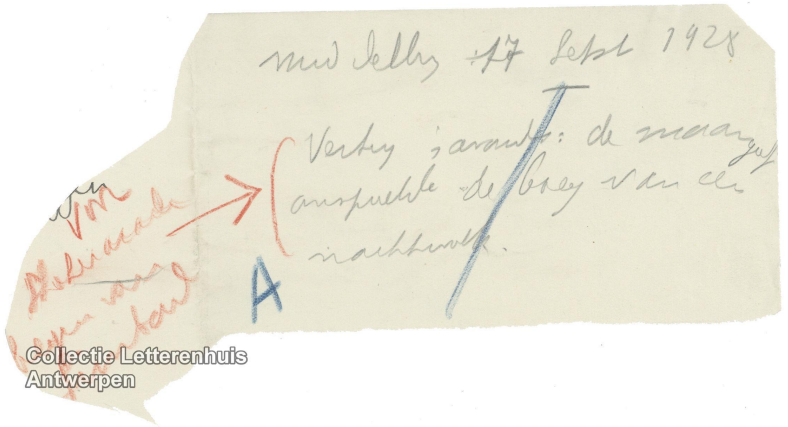
How would we like to use Transkribus?



Note of Raymond Brulez' *Sheherazade*, digitized by the Antwerp based Letterenhuis, consulted at: http://doc.anet.uantwerpen.be/docman/docman.phtml?file=.lhdigiplat.e03770.13.jpg.

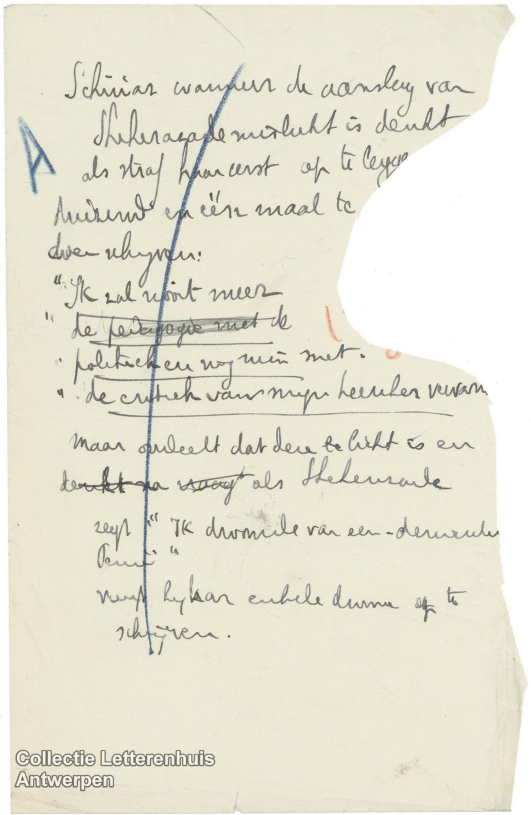
How would we like to use Transkribus?

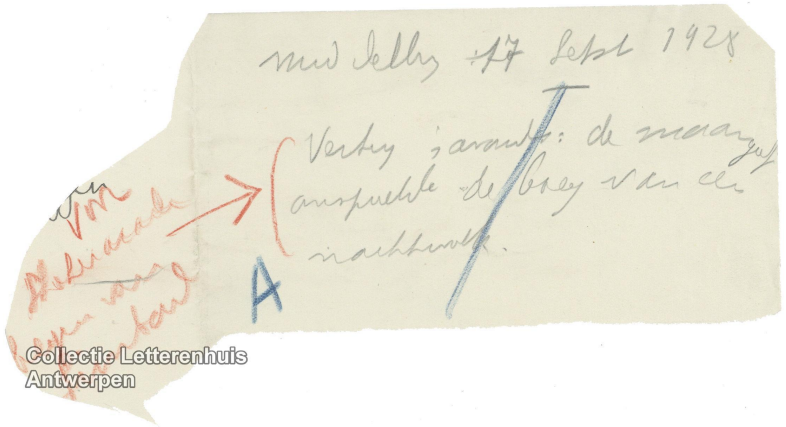




Note of Raymond Brulez' *Sheherazade*,
digitized by the Antwerp based Letterenhuis, consulted at:
http://doc.anet.uantwerpen.be/docman/docman.phtml?file=.lhdigiplat.bd658a.04.jpg.

How would we like to use Transkribus?





Note of Raymond Brulez' *Sheherazade*,
digitized by the Antwerp based Letterenhuis, consulted at:
http://doc.anet.uantwerpen.be/docman/docman.phtml?file=.lhdigiplat.7002f9.28.jpg.

How would we like to use Transkribus?

Funding proposal Research Foundation Flanders FWO (medium-sized infrastructure):

CATCH 2020 | Computer-Assisted Transcription of Complex Handwriting by 2020

- Textual Dimension | Genre, Text versus Document, Deleted Text
    - e.g. www.beckettarchive.org
- Linguistic Dimension | Language-Aware HTR (Style, Syntax)

Credits

This work was presented at the Transkribus User Conference 2017 hosted by READ at the Technical University Vienna (Austria) on 2-3 November 2017. Most of the (especially BDMP related) research leading up to these results was part of a project called CUTS (Creative Undoing and Textual Scholarship) supervised by Dirk Van Hulle, which received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) under ERC grant agreement n° 313609. The author contributed to the development of this collaboration with Transkribus (and the design of the CATCH 2020 project) as part of his work as the coordinator of the Antwerp division of DARIAH-VL, the Flemish consortium of DARIAH-BE. This function received funding from the Research Foundation Flanders (FWO), which also allowed the author to participate in the conference. The author would also like to thank the Letterenhuis for their kind permission to use the reproduction of Brulez' manuscripts in these slides (see copyright license on next slide).

License