Riddles in the Dark.
Corpus building

Katrien Depuydt
Senior researcher/linguist
katrien.depuydt@ivdnt.org
Gelde...<br>eind tekst onleesbaar
Digitisation, conversion
Born digital

Data suppliers
UNPO Statement of Solidarity

The Hague, 17 March 2003 - The Presidency of the Unrepresented Nations and Peoples Organization (UNPO), led by Reverend Mi Lhakpa Tsering, expresses its solidarity on behalf of all UNPO members with the people of Tibet in this period of extreme tension and reiterates its support for their decolonizing nonviolent campaign against Chinese oppression.

UNPO condemns the demonic Chinese response that has led to substantial loss of life and countless detentions and beatings, and calls upon the Chinese authorities at all levels to enter into a constructive dialogue designed to end the violence and promote a return to peace within Tibet as soon as possible.

UNPO urges the international community to demonstrate its commitment to human rights by responding to the events in Tibet with an unmistakable and immediate response to the events that have led to such violence within Tibet.

Standing on recent developments, UNPO General Secretary Mr. Marzio Radić added that “The
Steps

1. Corpus design
2. Corpus building strategy + quality criteria
3. Corpus acquisition [IPR]
4. Processing
   - analysis
   - conversion + metadata
 [linguist – SE]
To Lukas Prujs, 13 januari 1661 by Elisabeth Philipse Amelingh (1661)
... heeft - - - - - - - 4 gl - 0 Aen boeken en pampier en penen en ...

To Lukas Prujs, 6 juni 1664 by Elisabeth Philipse Amelingh (1664)
... lijue man stelt alles te boeck waet ghij uijt geft dat ...

To Simon Been, 9 december 1664 by Jacob Been (1664)
... alzo hij niet op de bouck en stondt en hij heeft ...

To J.H. Martens, 11 november 1780 by J.D. Piest (1780)
... Schripampier + + a 6 Sr t boek. 50 ditto ongsneeden ditto a ...

To Gabriel Lixraaven, 10 januari 1781 by Simon Quirijn Lixraaven (1781)
... voorne missive, dat UEDs mijn boeken bij berntrop laat verkoopen, als ...
... verbeeld hebbe, dat wat geleerde boeken aan gaat, schouten altoos geprefereert ...

To David Wendorp, 16 december 1780 by M. Reuvenhagen (1780)
... beijde, accuratese goede order der boecken, Voorsigtigheid, die prijse ik ...

To Magdalena Wendorp-Bagge, 16 december 1780 by M. Reuvenhagen (1780)
... rollen matten MR Eenige boeken Maar also die twee Scheepen ...
... te Senden, wat de boeken aan betreft Zo Gelieft maar ...

To Christiaan Frederik Coller, december 1780 by JVC Jur Luijt (1780)
... laaten wagt tot ik mijn boeken heeft, want weet anders niet ...
... houw nu alle de boeken bonte die sijn ...

To Christina Bakker-Smits, 26 september 1780 by Carsten Bakker (1780)
... gantsche regel die in myn Boek niet te vinden is, en ...

To Maartje Nannings, 12 februari 1665 by Paulus Pietersz. Schuitenmaker (1665)
... al frij hoch bij hente bock soo het scijnt en men ...

To Michiel Heusch jr., 19 oktober 1664 by Michiel Heusch sr. (1664)
... reecck. van hem wt v boeken Is heden versuijmt wij sijn ...

To Eldert Jansz., 6 november 1672 by Pieter Jansz. (1) (1672)
... maer ick behoevede wel een boeck pampier te hebben daerom sal ...

To Willem Lukaszw., november 1672 by Meintje Meinerts (1672)
... dan ick behoevede wel een boeck Pampier daerom sal ick opbreecken ...

To Johan Godfried Meeler, 5 januari 1781 by Geertruyda Westman-Erasmus (1781)
... van concept en wilde syn boeken absolut ter secretarij versorgen Eyndelyk ...
... 't Eijschen, van onvueruen van boeken ligten van copien Extracten of ...
General principles corpus processing

- What is a text?
- What is the format?
- Each token in the text has the correct metadata
- Metadatascheme for all corpora
  - core metadataset
  - date text vs. date witness
  - collection specific metadata
Corpusprocessing for Nederlab

- 2013 – 30 June 2018
- Create a research environment for historians, linguists, literary scholars...
- Diachronic corpus of existing digital collections (ca. 500 – present )
- User-friendly and tool-enriched web interface for scholars
- Uniform text format + metadata
- Enrichment of data by team and by scholarly users
- Focus on data quality by including an editorial staf
- http://www.nederlab.nl/onderzoeksportaal
  10 billion words currently; aimed at incorporating min. 20 different collections
Corpus processing

- Acquisition
- Analysis
- Conversion of text format > Folia
- Conversion of available metadata > relational database
- Curation of metadata (persons)
Huygens ING

- OCR (RGP 450 volumes)

- Text editions in TEI XML (GT)
Heinsius letters

- Daniel Heinsius (or Heins) (9 June 1580 – 25 February 1655) was one of the most famous scholars of the Dutch Renaissance.
- From Flanders to the Netherlands (Spanish)
- Prof of Latin and Greek in Leiden and librarian
- Publications: scientific and literary (poems, emblemata)
Letters

Over 20,000 letters

Published in 19 volumes

Digitised and made available by Huygens ING

http://resources.huygens.knaw.nl/retroboeken/heinsius/#page=0&accessor=toc_1&view=homePane
Challenges

- OCR quality
- Parsing structure (format!)
- Editorial matter vs original text
- Metadata
Hreda:
PHTLIPUB VnoBtxoa / Breda
I). D. PKTKHS PI,AS< HAART predicant tot ) Etten
Grn..iKi \n II.u:\irnr- ( Jilse.
Nevens d' heer PAIJLUS SNELEEN, schepen de r stad Hreda en ouder-
lingh aldaer.

U y t welke aengekome broederen tot directores van de aenstaeiele
verhandelinge verkoren zijn, t o t :
Praeses: D . DAVID A M I A .
Assessor: D. DIONYSIÜS KEL< I .
Scriba: D. DANIEL VAN PEKNE.

De tijt van ordinaire vergaderinge is volgens gewoonte gestelt des
morgens van acht t o t elf u y r e n , en des naemiddaghs van drie tot
ses u y r e n ; behalven des Saturdaeghs van seven t o t tien uyren, en des
Maendaeghs van vier t o t seven u y r e n ; en yder absent sal met twaalf-
t u w e r s , en te laet komende met ses stuyvers geboedt worden.
Sijn in dese Synodus verschenen de ordinaris Gedeputeerde des
Bynodi: I). D. JOIIANNS VOLLENHOVE, S. S. Theol. doctor, THEODORUS
VÁN e n LEK, Grunjium ANSLAEK en PETRUS DU BOIS, respective predi-
< uiten in 's-Gravenhage, L e y d e n , Delft en Gorichem.
Sijn door D. I). directores Synodi in dese vergaderinge geintroduceert
de Edele Hooghachtbare heeren Commissarisen van de Edele Groot-
Moogende heeren Staeten van Hollandt en West-Frieslandt, de heer
en meester HENDRIC DRUYF, raadordinary in den Hoogen R a d e , ende
d' a c h t b a r e heer en meester Puuroo M' B o » , borgenneester der stad
522
cependant le leur suis obligé de la permission qu'elles m'ont accordée pour la faire sortir en payant les droits. Je suis - - - N. Lillieroot.

Le 3 brc 1702.

1073. van VAN REEDE VAN DE LIER, 3 novembre (1702).
Eigenh. orig. H.A. 788.

Weledest Gesterne Heer, Desen sal allez dienen tot recipiscie van Uw E. t d van den 27 octob.1 en sullen wij ons nae derselve teneur reguleren. Ik kan ik niet wel begrijpen wat Uw E. belijf te verstaen met het recommanderen aen H. r grave Malbrough, ten waere Uw E. daarmede meent het compliment aen mij voor desen geden; so ik het heb geraden ben ik daerover aan Uw E. vervolget, maer versoete dat daervan niet gelieve te tonen de minste kennis te hebben. Ik desperere daer nogh niet van, overmits mylord thresorer uyt zijn eynig mouvement niet lange geleden, alhoewel in duystere termen, daervan heeft gesproken; het moet en with goede gratie geschieden of ik snode daer van allens en terwijlen ik togh geresolueft ben mij uyt het geweld te begeven en op het lant te gaan uytrusten, schriet het mij weynigh, maar ik kan niet afsyn mijn beklagh te doen over de missive van Haer E. Gr. Mog. geschreven aen de heeren Staten van Utrecht dat men met reght een Ural's brief magh noemen en waerdoor men mij genoeghsaem vastmackt aen de bank van justitie aldahr, regelreght tegen mijne intentie ende alle billicheyt, synde een vasal van Hollant en het goet waer-over de questie is leenroerigh aen deselve provintie.2 Maer als Haer Ed.G.Mog. de saek te degen belijfden in te zien, als de justitie 't uijs, hetsey haer daermede bemoeit is het tot deklyn en veragting van haer souvereyne maght uyt welkhen hoofte sij die resoluto hebben genomen. Nu ik wil hopen dat alles sal werden geredesereert en ver-soeke Uw E. gunste daerontrent, waarmede blijve F.b. van Reede.

Londen, den 5 nov.

1074. van SAUNIERE DE L'HERMITAGE, 3 novembre 1702.
Eigenh. orig. H.A. 792.

Nouvelles uit Londen over de toespraak van de koningin in het Parlement, het mislukken van de expeditie naar Cadiz en het aanwijzen van de schuldigen daaraan, de unie met Schotland en de houding van Nottingham tegenover de Anglicaanse kerk.

1075. van VAN VRIJBERGEN, 3 novembre 1702.
Eigenh. orig. H.A. 800.

Hooghevede Gesterne Heer, Desen namiddagh ontfongh men dry Hollantse maelen en daarmede de gewenschte tijdinge van het bemagtighen der citadelle van Luyck, dat hier veel vreughe geheet.
Het Lagerhuyys heeft heeden nemine contradicente geresolueft Haer Maj.1 in allens bij te staaen tot het volvoeren haerer alliancien en to intueugelinge der ongelimiteerde ambitie van Vrankrijk. Het addres sal morgen door commissarissen worden opgestelt. Mij dunkt dat men alle de reeden van de weerl heeft om te vertrouwen dat dese senh met gewencht sach en spoedigh sullen afloopen.
D'heeren Van de Lier en Van Haeren hebben mij gecomuniceert Haer Ho. Mog. resoluto en ordens van de 27 octob. behelsende de deschephes voor de heer Schoenenberg ende de mesures bij dit hoff en den Staat omtrent de croon Portugal te

1073. 1. Deze brief is niet gevonden.
2. Zie over deze zaak hiervóór p. 474.

1073. van VAN REEDE VAN DE LIER, 3 november (1702).
Eigenh. orig. H.A. 788.

Weledest Gesterne Heer, Desen sal alleen dienen tot recipiscie van Uw E. t d van den 27 octob.1 en sullen wij ons nae derselve teneur reguleren. Ik kan ik niet wel begrijpen wat Uw E. belijf te verstaen met het recommanderen aen H. r grave Malbrough, ten waere Uw E. daarmede meent het compliment aen mij voor desen geden; so ik het heb geraden ben ik daerover aan Uw E. vervolget, maer versoete dat daervan niet gelieve te tonen de minste kennis te hebben. Ik desperere daer nogh niet van, overmits mylord thresorer uyt zijn eynig mouvement niet lange geleden, alhoewel in duystere termen, daervan heeft gesproken; het moet en with goede gratie geschieden of ik snode daer van allens en terwijlen ik togh geresolueft ben mij uyt het geweld te begeven en op het lant te gaan uytrusten, schriet het mij weynigh, maar ik kan niet afsyn mijn beklagh te doen over de missive van Haer E. Gr. Mog. geschreven aen de heeren Staten van Utrecht dat men met reght een Ural's brief magh noemen en waerdoor men mij genoeghsaem vastmackt aen de bank van justitie aldahr, regelreght tegen mijne intentie ende alle billicheyt, synde een vasal van Hollant en het goet waer-over de questie is leenroerigh aen deselve provintie.2 Maer als Haer Ed.G.Mog. de saek te degen belijfden in te zien, als de justitie 't uijs, hetsey haer daermede bemoeit is het tot deklyn en veragting van haer souvereyne maght uyt welkhen hoofte sij die resoluto hebben genomen. Nu ik wil hopen dat alles sal werden geredesereert en ver-soeke Uw E. gunste daerontrent, waarmede blijve F.b. van Reede.

Londen, den 3 nov.

1 0 7 4 . van SAUNIERE DE I/HERMITAGE, 3 novembre 1 7 0 2 .
Eigenh. orig. H.A. 792.

Nouvelles uit Londen over de toespraak van de koningin in het Parlement, het mislukken van de expeditie naar Cadiz en het aanwijzen van de schuldigen daaraan, de unie met Schotland en de houding van Nottingham tegenover de Anglicaanse kerk.

1 0 7 5 . van VAN VRIJBERGEN, 3 novembre 1 7 0 2 .
Eigenh. orig. H.A. 800.

Hooghevede Gesterne Heer, Desen namiddagh ontfongh men dry Hollantse maelen en daarmede de gewenschte tijdinge van het bemagtighen der citadelle van Luyck, dat hier veel vreughe geheet.
Het Lagerhuyys heeft heeden nemine contradicente geresolueft Haer Maj.1 in allens bij te staaen tot het volvoeren haerer alliancien en to intueugelinge der ongelimiteerde ambitie van Vrankrijk. Het addres sal morgen door commissarissen worden opgestelt. Mij dunkt dat men alle de reeden van de weerl heeft om te vertrouwen dat dese senh met gewencht sach en spoedigh sullen afloopen.
D'heeren Van de Lier en Van Haeren hebben mij gecomuniceert Haer Ho. Mog. resoluto en ordens van de 27 octob. behelsende de deschephes voor de heer Schoenenberg ende de mesures bij dit hoff en den Staat omtrent de croon Portugal te

1 0 7 3 . 1. Deze brief is niet gevonden.
2. Zie over deze zaak hiervóór p. 474.
Re-OCR

- Abbyy FineReader Engine CLI for Linux [OCR4Linux.com] (Finereader 11)
- Parameters // Engine and ABBYY XML and ALTO available
- 120,000 pages for 1000 euro
- 10 seconds per page; OCR with 6 threads
Formats

Plain text
Docx/RTF/HTML/….
PDF
XML
  • Alto
  • → Abbyy xml
    • Why?
      • Blocks and paragraphs both explicitly coded
      • Detailed layout information on glyph level
      • Other details (separators, etc)
Gerritsz,
Artic.

van

leger

lasten.

Ad

art.
Amst.,
den
25
\textsuperscript{e}
Structure recognition: notes

1. Deze brief is niet gevonden.
2. Zie over deze zaak hiervóór p. 474.

1073. Mog. resolutie en ordres van de 27e ochtend, behelsende de deschrook nieuwelijke leden en Staet omtrent de croon Portugaule te
Structure recognition: letter segmentation

Task:
- find start and end of letter
- remove editorial matter

Method: approximate matching of (corrected) table of contents + typographical structure

Would not have been possible with less structured formatting information
Weledele Gestrenge Heer, Desen sal alleen dienen tot recipiscie van Uw E.

vl. rends="superscript">ts</vl>

van den 27 octob.

vl. rends="superscript">1</vl>

en sullen wij ons nae derselver teneur reguleren. Ick kan niet wel

vl. rends="superscript">1</vl>

begrijpen wat Uw E. belieft te verstaen met het recommerden aan H.

vl. rends="superscript">1</vl>

grafe Malbo-

vl. rends="superscript">1</vl>

rough, ten waere Uw E. daarmede meet het compliment aan mij voor desen gedaen;

vl. rends="superscript">1</vl>

so ick het heb geraden ben ick daerover aan Uw E. verobligirte, maer versoeke dat

darvan niet gelieve te tonen de minste kennisse te hebben. Ick desperere daer nogh

vl. rends="superscript">1</vl>

niet van, overmits mylord thersoer uyt sijn egen movemunt niet lange geleden, al-

vl. rends="superscript">1</vl>

hoewel in duystere termen, daervan heeft gesproken; het moet met goede gratie geschie-

vl. rends="superscript">1</vl>

den of ick soude daer van afsien en terwijlen ick togh geresolveert ben mij uyt het

vl. rends="superscript">1</vl>

gewoel te begeven en op het lant te gaan uytrusten, scheelt het mij weynigh, maer ick

vl. rends="superscript">1</vl>

can niet afsyn mijn beklagh te doen over de missive van Haer E. Gr. Mog. geschreven

vl. rends="superscript">1</vl>
aen de heeren Staten van Utrecht dat men met reght een Uria’s brief magh noemen en

vl. rends="superscript">1</vl>
waerdoor men mij genoeghaem vastmaakt aen de bank van justitie aldaer, regelrechte

vl. rends="superscript">1</vl>
tegen mijne intentie ende alle billicheyt, sijnde een vasal van Hollant en het goet waer-

vl. rends="superscript">1</vl>
over de questie is leenroerigh aan deselve provintie.

vl. rends="superscript">2</vl>
Maer als Haer Ed.G. Mog. de saek

vl. rends="superscript">1</vl>
to degen beliefden in te sien, als de justitie ’tsij waer, hetsy hea daermede benoemoet is

vl. rends="superscript">1</vl>
het tot deklyn en verwagting van haere souvereine maagt uyt welkers hoofdie sij de

vl. rends="superscript">1</vl>
resolutie hebben genomen. Nu ick wil hopen dat alles sal worden geredresseert en ver-

vl. rends="superscript">1</vl>
soeke Uw E. gunste daerontrent, waarmede blijve - F.b. van Reede.

vl. rends="superscript">1</vl>

Londen, den 3 nov.
Metadata

- The collection of Heinsius’ letters is written by over 200 different authors
- Identification of authors + biographical information
Corpus Hedendaags Nederlands
chn.inl.nl

- Integration: 5, 27, 38 miljoen woorden-corpora, Parolecorpus, monitorkrantencorpus (internal vs. external)
  - format
  - metadata scheme
- Surinam, Dutch Caribbean
- Corpus workflow
- Corpus design future
Corpus workflow

Workflow tool:

Steps:

1. FTP (NRC/Standaard)
2. Conversion to TEI (utf8-pid; fixup-tei)
3. Remove duplicates (content/source-id)
4. Linguistic annotation
5. Indexing with BlackLab (incremental)
6. Errors
7. Maandelijkse export naar Neoloog
8. *Automatic update*
Build your own corpus

- Born digital
- Digitisation
  - IMPACT CoC (digitisation.eu)/ advice on digitisation
  - PICCL pipeline (CLARIAH)
- Conversion:
  conversion tools
  Open Convert
  (Openconvert.clarin.inl.nl)