↘nederlab↗

On OCR ground truths and OCR post-correction gold standards, tools and formats

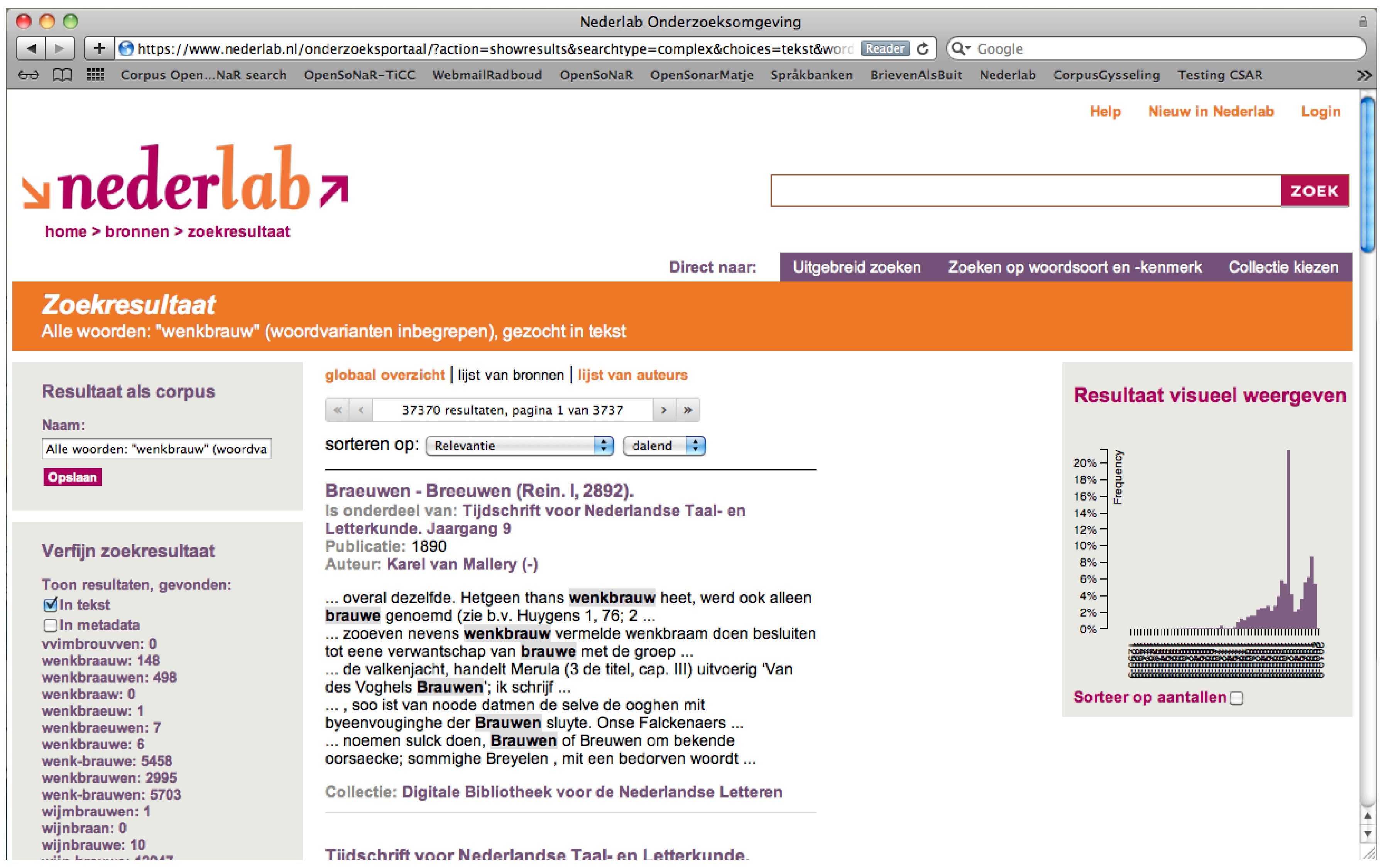
Martin Reynaert

TiCC - Tilburg University / CLST - Radboud University Nijmegen

DATECH 2014 Biblioteca Nacional de España, Madrid, Spain May 19-20, 2014

1. NWO project Nederlab

- The Nederlab project aims to bring together all digitized texts relevant to the Dutch national heritage (c. A.D. 800 – present) consisting of terabytes of data in one user-friendly and tool-enriched web interface, allowing scholars to simultaneously search and analyze textual data in a virtual research environment.
- The focus in Nederlab is currently on incorporating the vast digital text collections of the Koninklijke Bibliotheek (`http://www.kb.nl/en`) (KB or Dutch National Library) as well as the contents of the Digitale Bibliotheek voor de Nederlandse Letteren (`http://www.dbnl.org/`) (DBNL - The Digital Library of Dutch Literature).
- KB text collections comprise newspapers from 1618 to 1995 and the Early Dutch Books Online or EDBO (`http://www.delpher.nl/`).
- These were digitized by means of Optical Character Recognition. If there is one thing all results of large digitization programmes have, it is that they are riddled with OCR misrecognition errors.


Nederlab Onderzoeksomgeving

https://www.nederlab.nl/onderzoeksportaal/?action=showresults&searchtype=complex&choices=tekst&word

Zoekresultaat
Alle woorden: "wenkbrauw" (woordvarianten inbegrepen), gezocht in tekst

Resultaat als corpus
Naam: Alle woorden: "wenkbrauw" (woordva
Opslaan

Verfijn zoekresultaat
Toon resultaten, gevonden:
☑ In tekst
☐ In metadata
vvimbrouvven: 0
wenkbraauw: 148
wenkbraauwen: 498
wenkbraaw: 0
wenkbraeuw: 1
wenkbraeuwen: 7
wenkbrauwe: 6
wenk-brauwe: 5458
wenkbrauwen: 2995
wenk-brauwen: 5703
wijmbrauwen: 1
wijnbraan: 0
wijnbrauwe: 10

globaal overzicht | lijst van bronnen | lijst van auteurs
37370 resultaten, pagina 1 van 3737
sorteren op: Relevantie dalend

Braeuwen - Breeuwen (Rein. I, 2892).
Is onderdeel van: Tijdschrift voor Nederlandse Taal- en Letterkunde. Jaargang 9
Publicatie: 1890
Auteur: Karel van Mallery (-)

... overal dezelfde. Hetgeen thans **wenkbrauw** heet, werd ook alleen **brauwe** genoemd (zie b.v. Huygens 1, 76; 2 ...
... zooeven nevens **wenkbrauw** vermelde wenkbraam doen besluiten tot eene verwantschap van **brauwe** met de groep ...
... de valkenjacht, handelt Merula (3 de titel, cap. III) uitvoerig 'Van des Voghels **Brauwen**'; ik schrijf ...
... , soo ist van noode datmen de selve de ooghen mit byeenvoughinghe der **Brauwen** sluyte. Onse Falckenaers ...
... noemen sulck doen, **Brauwen** of Breuwen om bekende oorsaecke; sommighe Breyelen , mit een bedorven woordt ...

Collectie: Digitale Bibliotheek voor de Nederlandse Letteren

Tijdschrift voor Nederlandse Taal- en Letterkunde.

Resultaat visueel weergeven

Sorteer op aantallen ☐

Figure 1 : Querying Nederlab for diachronical spelling variants for 'wenkbrauw' (E.: 'eyebrow') .

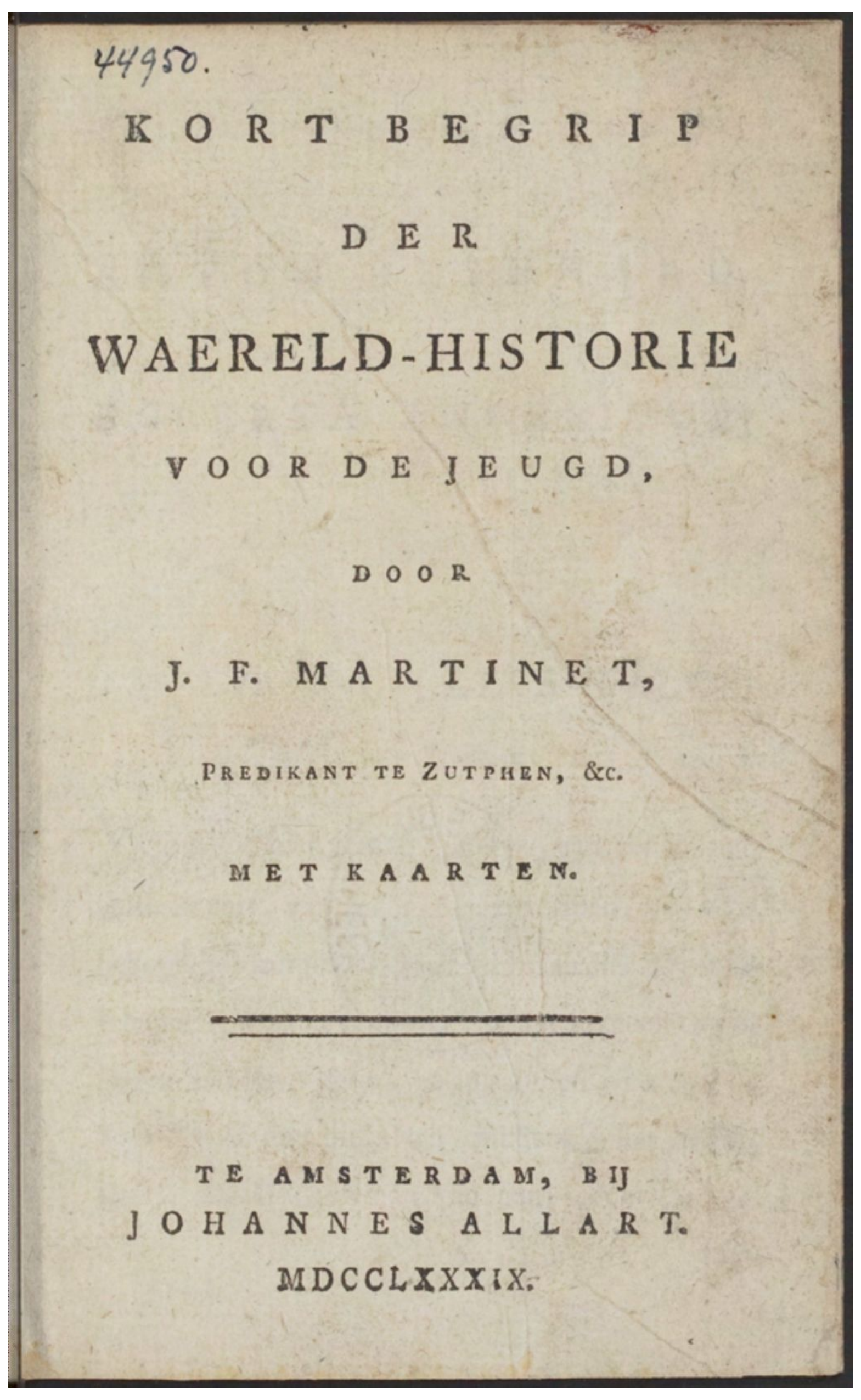
2. CLARIN-NL project @PhilosTEI

@PhilosTei

- Philosophers – as all other aspiring eHumanities researchers – today increasingly require high quality electronic versions of the works they study.
- In CLARIN-NL Call 4 project @PhilosTEI we are therefore building a work flow of web services which will allow individual researchers to upload digital images of the book's pages and receive back after processing a well formatted electronic text in TEI format fit for further building into e.g. a critical edition of the work.
- In the work flow, it is TICCL's task in its guise as the web service TICCLops, to enhance the text's quality, fully automatically.

3. Gold standards and tools for OCR post-correction evaluation

- We have aligned the KB OCRed version of the 1789 book by Martinet, known as DPO35, with the OCR ground truth built in the European project Impact.
  We have further extended this with both an aligned historical OCR post-correction gold standard and a contemporary one. These gold standards allow for accurately measuring the performance of OCR post-correction systems such as TICCL on this book, as well as to estimate their performance on other digitized books of the era.
- Gold standards are expensive. Nevertheless more are required for other time periods and text types. To facilitate their building, we have developed convertors to and from FoLiA XML from commercial as well as open source OCR output formats, i.e. Alto and Page XML and hOCR HTML.
  FoLiA XML has been extended towards untokenized OCRed text enriched with ground truth and gold standard versions on string level, in alignment with the noisy OCR strings.


44950.

KORT BEGRIP
DER
WAERELD-HISTORIE
VOOR DE JEUGD,
DOOR
J. F. MARTINET,
PREDIKANT TE ZUTPHEN, &c.
MET KAARTEN.

TE AMSTERDAM, BY
JOHANNES ALLART.
MDCCLXXXIX.

(a) URL: http://resolver.kb.nl/resolve?urn=dpo:35:mpeg21:0009

Figure 2 : Title page of 1789 book by Martinet

4. The case for OCR post-correction

Category	LD 1	LD 2	LD 3	LD 4	LD 5	LD 6 -12	Total	%
deletion	315	22	3	1			341	2.420
insertion	3378	190	11	8	4		3591	25.490
substitution	2342	341	92	42	2		2819	20.010
transposition		7					7	0.050
multisingle		3605	514	93	59	47	4318	30.650
multiple		853	767	391	134	168	2313	16.418
space deletion	52						52	0.369
space insertion	643						643	4.564
TOTAL	6734	5018	1387	535	199	215	14088	
%	47.80	35.62	9.85	3.80	1.41	1.51		100.0
SUMMED %	47.80	83.42	93.27	97.07	98.48	100.0		

- Spelling correction systems, of whatever kind, have a certain 'reach' in terms of Levenshtein or edit distance (LD) within which they operate.
- The sums of the totaled percentages in the table detailing the shifts of the OCR version compared to the modern Gold Standard show that 83.42% of the shifts lay within LD 2, 93.27% within LD 3.
- Given sufficiently powerful and comprehensive OCR post-correction systems, most errors should be resolvable.

5. FoLiA XML in Nederlab

- The route followed in Nederlab is to convert all the texts incorporated into a common format, FoLiA XML (CLIN Journal 3, van Gompel and Reynaert – 2013). All text processing and linguistic enrichment tools in the Nederlab work flow handle FoLiA XML. The format is further extended if and when required.
- In their turn all the research and analysis tools available in Nederlab (will) have been adapted to this format.
- If already available online, the texts remain as they are at their original location and the linguistically or otherwise enriched versions link to these.

6. Diachronical Text-Induced Corpus Clean-up

- The Text-Induced Corpus Clean-up system TICCL has now been largely ported from Perl to distributable (in both senses of being shareable and being parallelizable) C++ code.
  It has been rethought so as to be multilingual in the sense that anyone can now plug in an available open source spelling correction lexicon for the language of choice and set to work.
- We have incorporated into TICCL what must be the largest extant historical lexicon for Dutch as well as its accompanying historical name list. Both were developed at INL (`http://www.inl.nl/`), the Dutch Institute for Lexicology, partner in Nederlab. They were deliverables of the European project Impact and are available through the Impact Centre of Competence (`http://www.digitisation.eu/`).
  We will measure their effect on OCR post-correction of the Nederlab corpora.
- In Nederlab the main challenge for TICCL is to be able to distinguish between historical and OCR spelling variants.
- That is, apart from the enormous sizes of the text collections to be fully automatically post-corrected.
  EDBO has about 10K Dutch books, about 1.7M pages of digitised text, representing about 435M word tokens.

7. TICCL challenges

- A comparison of historical spelling variation with OCR derived lexical variation:

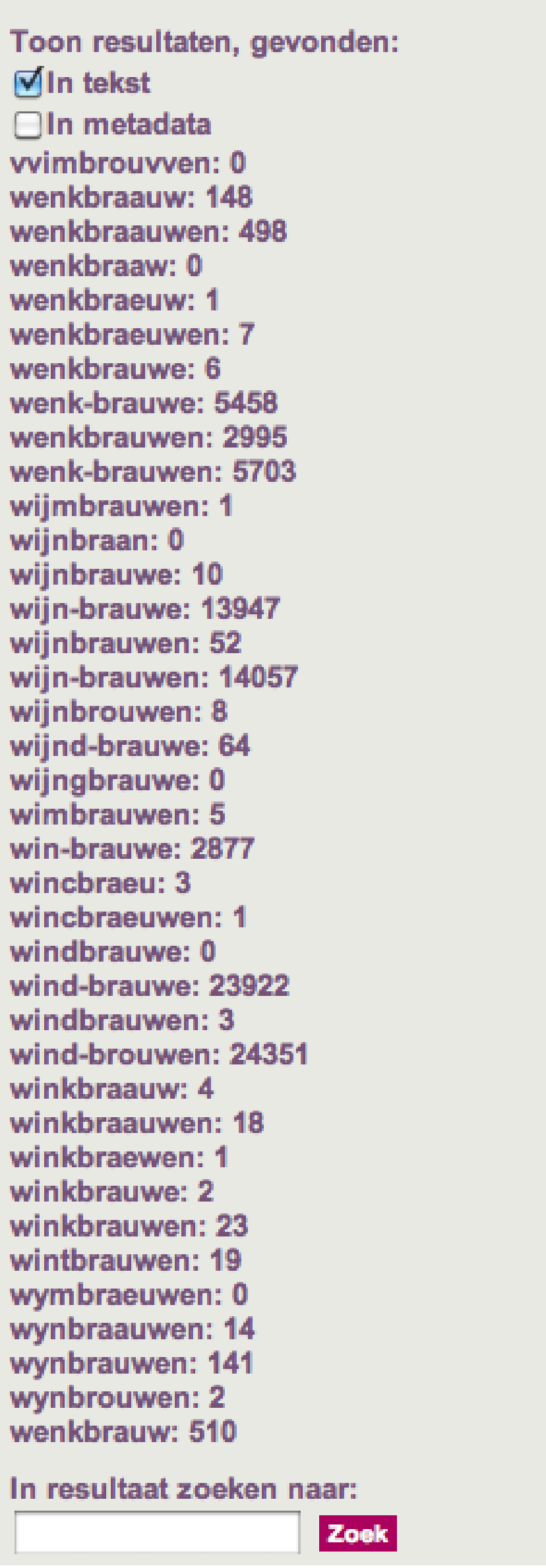

Toon resultaten, gevonden:
☑ In tekst
☐ In metadata
vvimbrouvven: 0
wenkbraauw: 148
wenkbraauwen: 498
wenkbraaw: 0
wenkbraeuw: 1
wenkbraeuwen: 7
wenkbrauwe: 6
wenk-brauwe: 5458
wenkbrauwen: 2995
wenk-brauwen: 5703
wijmbrauwen: 1
wijnbraan: 0
wijnbrauwe: 10
wijn-brauwe: 13947
wijnbrauwen: 52
wijn-brauwen: 14057
wijnbrouwen: 8
wijnd-brauwe: 64
wijngbrauwe: 0
wimbrauwen: 5
win-brauwe: 2877
wincbraeu: 3
wincbraeuwen: 1
windbrauwe: 0
wind-brauwe: 23922
windbrauwen: 3
wind-brouwen: 24351
winkbraauw: 4
winkbraauwen: 18
winkbraewen: 1
winkbrauwe: 2
winkbrauwen: 23
wintbrauwen: 19
wymbraeuwen: 0
wynbraauwen: 14
wynbrauwen: 141
wynbrouwen: 2
wenkbrauw: 510
In resultaat zoeken naar:

Figure 3 : 38 Diachronical spelling variants for 'wenkbrauw' from INL historical lexicon (E.: 'eyebrow')

- 11 Attested historical word forms for 'regering' (E.: 'government') with known attestation years: regeringe (1556), regheringhe (1562), regieringe (1581), regieringhe (1596), regieringh (1625), regeeringe (1631), regeeringhe (1658), regeeringh (n.a.), regeringhe (n.a.), regeering (n.a.), regering (1950)
- Random selection of just 50 OCR-variants for 'regering' from 1570 variants collected in one year's editions of newspaper 'Het Volk' (1918):
  egeering jregeerinff r2gecring rcgcerinc rcgeeiiog rcgeering- rcgeering3- rcgeering rcgeerint rcgeeriog rcgerricg re2rcring re2reoring re3ccring re3e2rins re3eerinf receeriiig regc2rint regccrin'g regceringt regcpring regctfring regecrbag regecriir regecrmg regecruig regeerin- regeerln regerrinj regerrins regetring regoerinfr regouring regut ring rejcering rekeerinft rereerinr rerocring reswring rfgccring rjegeering rogeerina rogeeriu rorcering rrjeering rvgccring tegeerings- tegeering tjtgeering ucgecring

8. TICCL companion paper

- An in-depth evaluation of TICCL on EDBO, as measured on Gold Standard DPO35, is to be presented in a companion paper titled 'Synergy of Nederlab and @PhilosTEI: diachronic and multilingual Text-Induced Corpus Clean-up' at LREC 2014.
- TICCL is now within Nederlab being set to correct the Dutch National Library digitized newspaper collection 1618 to 1899, working backwards in time. A crowd sourcing endeavour is being set up to have volunteers correct the 17th. century newspapers.
- Within Nederlab the move is currently towards modernising the spelling of the diachronical texts rather than adapting the tools to the diachronical text variations.

9. Acknowledgements

This work is supported by the Netherlands Organisation for Scientific Research (NWO), the Royal Netherlands Academy of Arts and Sciences (KNAW), and Common Language Resources and Technology Infrastructure (CLARIN-NL-12-006 project @PhilosTEI and CLARIN-NL-12-013 project OpenSoNaR).

10. Datech 2014

DATeCH 2014 Digital Access to Textual Cultural Heritage

For more information, see www.nederlab.nl. (c) Meertens Instituut (KNAW). Poster design by Junte Zhang, adapted by Martin Reynaert.