Golden Agents: Detectie van entiteiten in boedelinventarissen

Maarten van Gompel, Bram Buitendijk, Leon van Wissen, Harm Nijboer, Menzo Windhouwer

23 juni 2022

Introductie

- Boedelinventarissen beschrijven de huisraad, activa en passiva.
  - geen onroerend goed
  - nav huwelijk, faillisement, nalatenschap etc. . .
- Deze willen we makkelijker ontsluiten voor onderzoekers
  - lange onderzoekstraditie (archeologisch, sociaal-historisch, etc..)

Onze doelstelling: Hoe kunnen we *automatisch* namen van personen, locaties, en vooral objecten herkennen in boedelinventarissen?

Dataverwerkingspipeline (1)

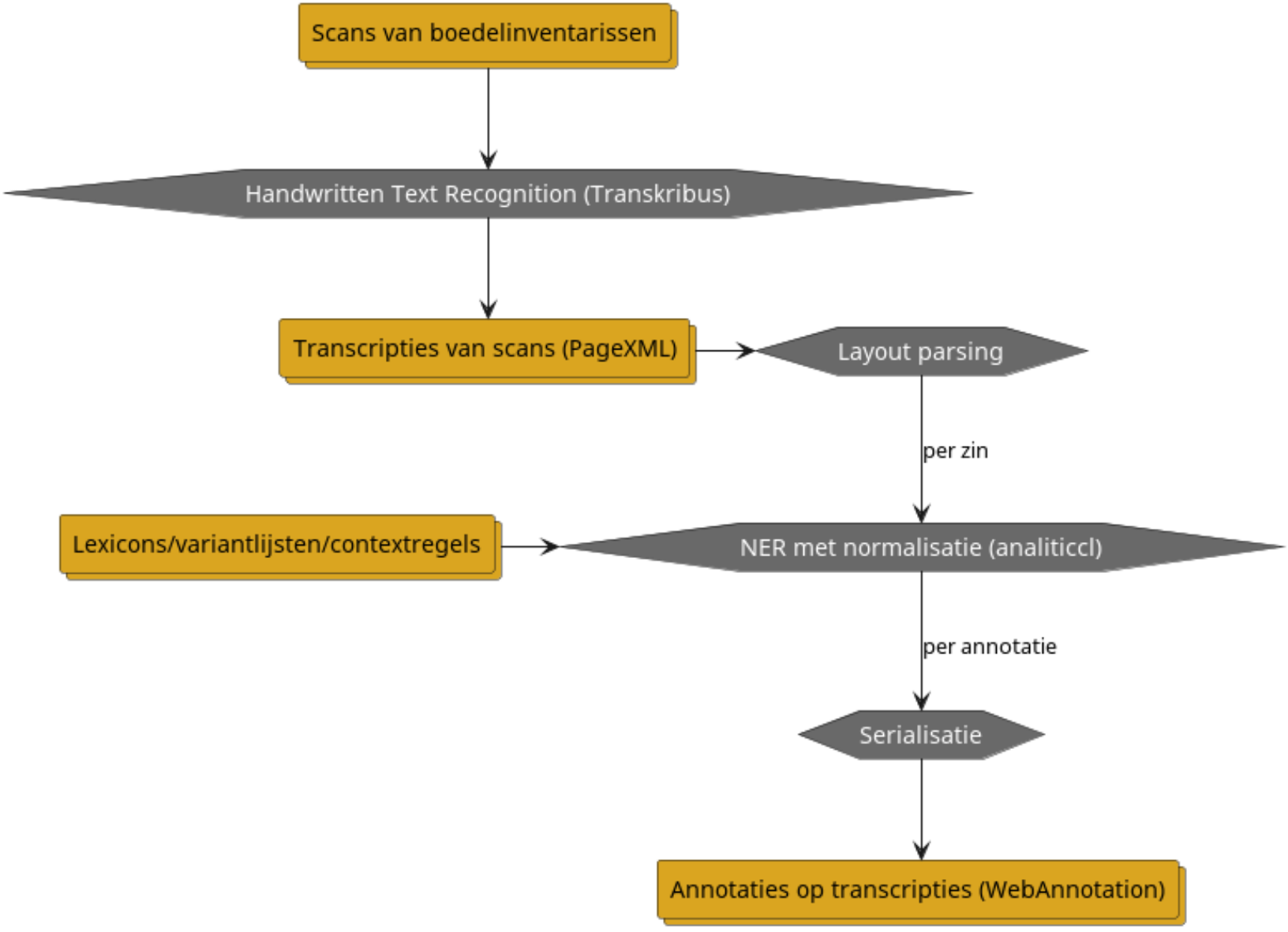

Figure 1: Pipeline

Uitdagingen

- De HTR output bevat allerlei spellingsvariatie:
  - door HTR-fouten
  - door diachronische spellingsvariatie en gebrek aan standaardisatie indertijd door spatiëringsfouten (splits/run-ons)
- Normale entiteitherkenning werkt hierdoor niet goed
  - Standaard NER-modellen zijn niet getraind op objecten

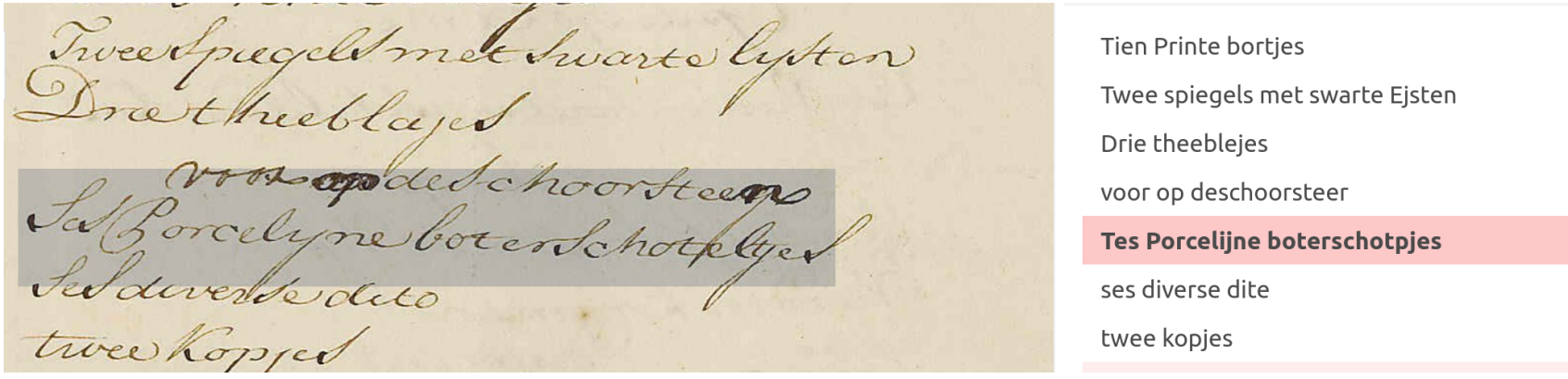

Figure 2: HTR Voorbeeld

Onze strategie

Wat hebben we?

- Een flink aantal lexicons/thesauri met namen van personen, locaties en objecten (diverse bronnen)
- Met name de objecten hebben we handmatig verrijkt:
  - bron
- INT Historisch Lexicon; gecureerde lijst die historische varianten koppelt, tevens geschikt als *achtergrondlexicon*
  - helaas niet openbaar beschikbaar ivm beperkte rechten

Aanpak

- We doorzoeken boedelinventarissen op termen uit deze lexicons, rekening houdend met:
  - spellingsvariatie; zoek de begrippen in de lexicons die het meeste op de aangetroffen vorm lijken
- combinatie van named entity recognition en tekstnormalisatie in één

Dataverwerkingspipeline (2)

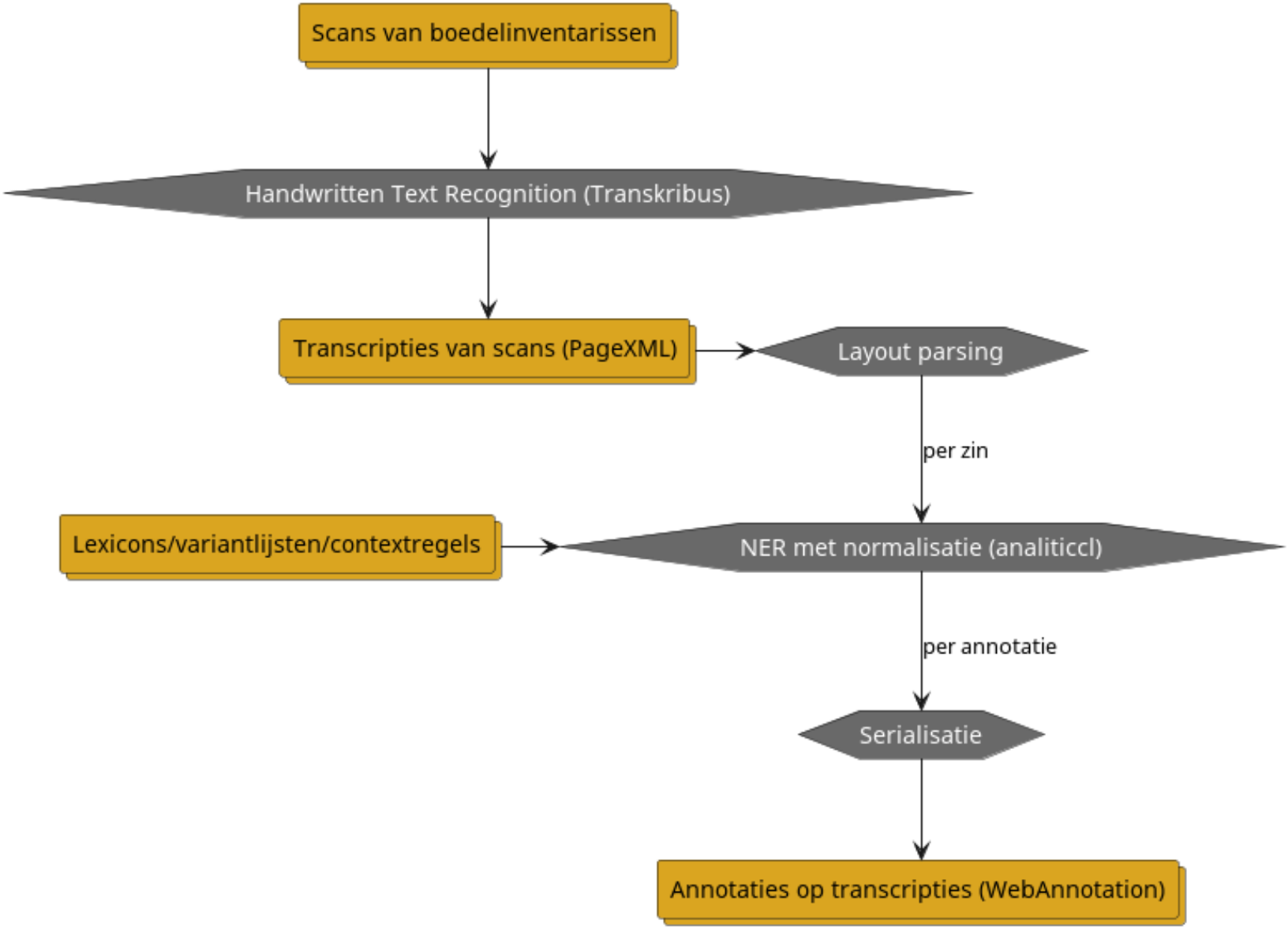

Figure 3: Pipeline

Analiticcl

- Software voor spellingscorrectie en tekstnormalisatie
  - koppelt woorden en frasen aan varianten
  - leest en genereert *variantenlijsten*
  - doet zowel *correctie* als *detectie*
  - verschillende afstandsmetrieken (waaronder Damerau-Levenshtein)
  - *schaalbaarheid:* gaat efficiënt om met grote zoekruimten en zoekafstanden
  - lexicons kunnen *frequentieinformatie* bevatten
  - *context*: weegt contextinformatie mee d.m.v. taalmodellen of opgestelde contextregels
  - deze contextregels maken een soort tagging mogelijk
  - *command-line tool* en *library*: geïmplementeerd in Rust, met Python bindings
- Herimplementeert en bouwt voort op kernideeën van Martin Reynaert (TICCL)
- Technisch-inhoudelijke presentatie: https://diode.zone/w/kkrqA4MocGwxyC3s68Zsq7

Voorbeeld Analiticcl

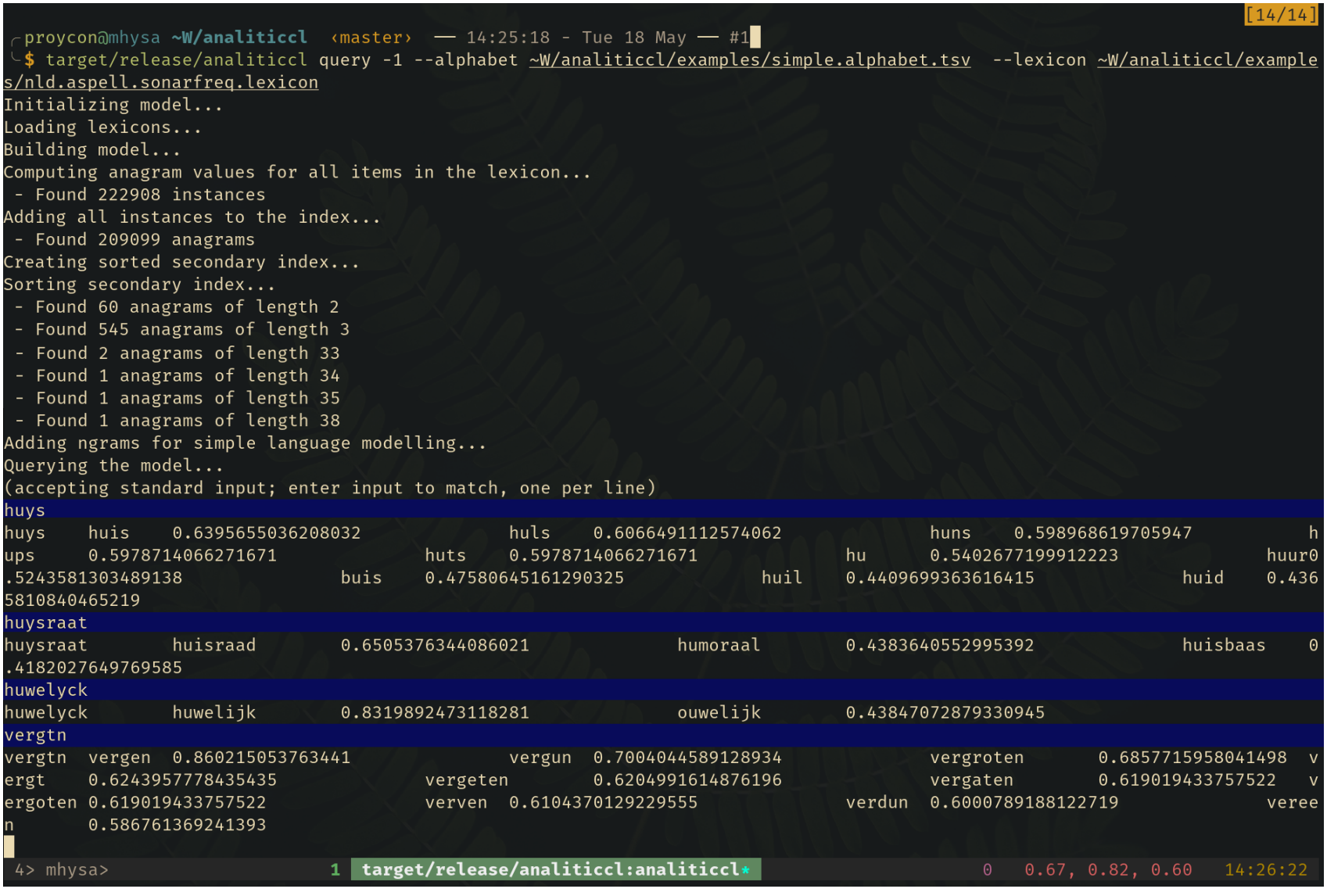

Figure 4: Voorbeeld, blauwe regels zijn input

Lexiconcuratie

- Kwaliteit is erg afhankelijk van de kwaliteit van de input (lexicons, variantlijsten)
- Handmatige **lexiconcuratie** om tot een lijst 'boedeltermen' te komen. Focus op objecten, maar ook categorieën die helpen deze vindbaar te maken (denk aan materialen, telwoorden en andere eigenschappen).
- Belangrijke rol voor het achtergrondlexicon (INT Historisch Lexicon)

Referentiedata en annotatie

Om te kunnen evalueren hebben we referentiedata nodig.

- Een klein aantal boedelinventarissen zijn handmatig geannoteerd, zowel op categorie (persoon, locatie, object) als op tekstnormalisatie
- Een annotatieomgeving (gebaseerd op Recogito-JS) is speciaal hiervoor ontwikkeld:
  - Source: `https://github.com/knaw-huc/golden-agents-htr/tree/maste`
- Annotatoren hebben een eerste output van analiticcl gecorrigeerd en aangevuld om tot een ground-truth the komen
  - dank ook aan Jirsi Reinders & Judith Brouwer
- Op deze development-set hebben we verdere parameters getest en input lexicons verbeterd
- Een evaluatietool vergelijkt systeemoutput (analiticcl) met de referentiedata en berekent Precisie, Recall en F1.

Voorbeeld Annotatietool

Golden Agents: Annotation Evaluation (v2022.03.23) (?)

Text: ☐ ☑ ☐ A25555000113 | Checked: Harm ☐ Jirsi ☑ | Judith ☐ | Save annotations

Tag Legend: | (ambiguous) | firstname | familyname | person | occupation | material | property | object | picture | animal | category | quantifier |

Page in Stadsarchief

t
18
24:
ƒ
Inventaris en taxatie van de gereetslappen
door den versuerder overgelebe t en door de
guerdens vande voorsz. brouv
en bekennen bijde onderteker
haer ontfangen zijn, onder c
huer sedul vermelt Staet, Na
de nieuwe waterschuijt cost
een entachtigh gulde sacht s
-
Segge aengenomen voor
-
ƒ 1481:8
de oude Waterschuijt aengenomen
voor hondert en vijftigh guld Segge ƒ 150:

gereedschap
gereedschappen
Add a reply...
object
Add tag...
Cancel
Ok

Voorbeeld resultaten

Muylen
f2 28
in een klijn Vast kasje
Een Doosje meediverse sakjes Knoope
Eenige Rommeling
Twee wolle dam dte borstrokken
Twee wolle en eenig Catoene deeken
Een kessen en een klijnder dito
Drie Mattrassen
Eenonde Velthafel
Een Vengel kooijje
Eenkleere bakje
Een Tuweel Koffertje nevens eenige Rommeling
op 't Comptoire
Een Bed Peulerenses Kussens
Eenhhoene deken
Twee Materessen
Een klijn wit deekenje
Een Vrouwe Jak en eendito borstrek
Twee Gryne Rokken
In het voorhuijs
Eerstoombank ende winkel plakke
Eenspeecel met eenswartelijst
Een gescheldert gemak Koffertje

Evaluatie

Classificatie & Normalisatie

Metriek	#out	#ref	Precisie	Recall	F1
Objecten	448	584	0.498	0.382	0.432
Personen	219	130	0.612	0.638	0.625
Locaties	14	26	0.143	0.077	0.100
Vertrekken	33	33	0.455	0.455	0.455
Totaal	-	-	0.525	0.435	0.475

Alleen Classificatie

Metriek	Precisie	Recall	F1
Objecten	0.681	0.522	0.432
Personen	0.685	0.714	0.699
Locaties	0.214	0.115	0.150
Vertrekken	0.455	0.455	0.455
Totaal	0.662	0.550	0.601

Referenties

- **Analiticcl**: https://github.com/proycon/analiticcl
- **Golden Agents HTR Repo:**
  https://github.com/knaw-huc/golden-agents-htr
  - Bevat: annotatietool, evaluatietool, NER wrapper tool, gecureerde lexicons, alle experimenten (uitvoer, evaluatie logs, sparql queries)

Publicaties:

- Reynaert, Martin. (2004) Text induced spelling correction. In: Proceedings COLING 2004, Geneva (2004). https://doi.org/10.3115/1220355.1220475
- Reynaert, Martin. (2011) Character confusion versus focus word-based correction of spelling and OCR variants in corpora. IJDAR 14, 173–187 (2011). https://doi.org/10.1007/s10032-010-0133-5