Using GaLAHaD and LAnCeLoT for historical corpus annotation

CLARIAH-PLUS WP3 Thomas Haga (CP-WP3-22-004)

In het kader van CLARIAH-PLUS WP3 zijn op het INT GaLAHaD (**G**enerating **L**inguistic **A**nnotations for **H**istorical **D**utch) en LAnCeLoT (**L**inguistic **A**nnotation **C**orpus **L**aundry **T**ool) ontwikkeld.

GaLAHaD is een platform voor taalkundige verrijking van historisch Nederlands. GaLAHaD biedt:

- taggers waarmee historische corpusdata verrijkt kunnen worden met woordsoort en lemma;
- benchmarkcorpora waarmee taggers geëvalueerd kunnen worden;
- de mogelijkheid om je eigen corpus te uploaden om taalkundig te verrijken;
- de mogelijkheid om de performance van de aangeboden taggers te vergelijken (evalueren) op de benchmarkcorpora;
- de mogelijkheid om te evalueren hoe goed de aangeboden taggers het doen op je eigen corpusmateriaal, hetzij aan de hand van een eigen gouden standaard van een gedeelte van het eigen materiaal, hetzij door het vergelijken van de tagging door de verschillende aangeboden taggers;
- de mogelijkheid om taggers te laten toevoegen en vergelijken met de andere beschikbare taggers;
- modellen (taggers) en datasets zelf; hierbij word je verwezen naar de GitHub (Open Source).

Na tagging van een corpus kan het corpus met annotaties geëxporteerd worden. LAnCeLoT maakt het vervolgens mogelijk om de verrijking handmatig te corrigeren. Kenmerkend voor LAnCeLoT is:

- het corrigeren via types met alle daarbij behorende tokens en bijbehorende analyses;
- het gemakkelijk tegelijkertijd (en dus consistenter) kunnen beoordelen van alle tokens per type;
- de beschikbaarheid van suggesties uit het GiGaNT-Hilex-lexicon[1] die horen bij een type;
- de mogelijkheid om met meerdere mensen tegelijkertijd in een project te kunnen werken;
- de mogelijkheid om de context van een match (token) te zien en (de context en) het gehele document in het geüploade corpus.

Ik heb gewerkt met beide omgevingen, omdat ik een corpus van zevenentwintig onverrijkte fragmenten van 19e-eeuwse kranten[2] taalkundig wil verrijken. Ik doe verslag van mijn ervaringen en geef per applicatie nog een aantal wensen en verbeterpunten door.

GaLAHaD biedt verschillende taggers en de mogelijkheid om te evalueren. Omdat ik geen gouden standaard heb voor mijn dataset, was het kiezen van een goede (en nog moeilijker de beste) tagger een uitdaging. Die gouden standaard kan je maken met behulp van LAnCeLoT. Met LAnCeLoT kan je de automatisch aangebrachte analyses eenvoudig (en per type) verbeteren.

1. GaLAHaD

1.1 Selectie van tagger en verrijking corpus

Allereerst maak ik een nieuw corpus aan dat ik “19e-eeuwse kranten” noem, waarin ik de onverrijkte XML-bestanden upload. Hierbij kies ik voor de TDN-tagset, die ontwikkeld is voor diachroon tekstmateriaal van historisch Nederlands.[3] Bij elk geüpload document krijg je een kort voorbeeld van

[1] Zie https://ivdnt.org/corpora-lexica/gigant/.
[2] De data zijn afkomsting uit de IMPACT dataset, https://www.digitisation.eu/impact-dataset/.
[3] Zie https://ivdnt.org/wp-content/uploads/2021/05/TDN_INT_WP_1.pdf voor versie 1. De herziene versie verschijnt binnenkort.

het begin van je documenten te zien. In Figuur 1 laat de kolom *source annotations* ook zien dat er geen bronlaag in de documenten zit. Dat is logisch, aangezien mijn corpus nog niet taalkundig verrijkt is.

NAME	FORMAT	PREVIEW	SOURCE ANNOTATIONS (TOKEN / POS / LEMMA)	LAST MODIFIED	ACTIONS
00530995.xml	tei-p5	pc-00530995 Anno 1853. N°. 959. Donderdag 11 Augustus. DE NEDER-LANDER. NIEUWE UTRECHTSCHE COURA...	0 / 0 / 0	2-07-24 16:20	
00531003.xml	tei-p5	pc-00531003 Verpachten: Het Grasgewas van: 1°. Een Perceel Hooiland, het Slag, in het Haerster...	0 / 0 / 0	2-07-24 16:20	

Figuur 1: documenten in het corpus

Daarna kunnen een of meerdere taggers gekozen worden om de documenten taalkundig te verrijken. Momenteel hangen er verscheidene Huggingface- en PIE-taggers in die getraind zijn op reeds verrijkte datasets met materiaal van de 13e tot en met de 19e eeuw. Er is afzonderlijk getraind op een kleine hoeveelheid data, op materiaal uit een bepaalde tijdsperiode en op al het materiaal.

Omdat het corpus onverrijkt is, is het lastig om te kiezen welke taggers goede resultaten opleveren. Ik verwacht dat de taggers die getraind zijn op al het materiaal, materiaal uit die periode en op het type materiaal (kranten) goed zullen scoren, maar kan dat niet zien bij de jobs zelf. Daarom heb ik ook naar de benchmarks gekeken, waar wel een verrijkte set op alles (tdn-core-all), op materiaal uit die periode (tdn-1600-1900) en op kranten (couranten) beschikbaar is. Hierbij is hug-tdn-all-enhanced een versie met verbeteringen voor scheidbare werkwoorden. Figuur 2 laat zien hoe goed de taggers scoren op de PoS-annotatie en Figuur 3 hoe goed de taggers scoren op lemmatisering. Voor de helderheid zijn de taggers waarvan ik op voorhand dacht dat ze goed zouden kunnen scoren geelgemarkeerd.

Dataset: tdn-core-all | Annotation: PoS | Group by: PoS | Single/multiple analysis: Both

TAGGER	MACRO PRECISION	MACRO RECALL	MACRO F1	MICRO ACCURACY	DETAILED EVALUATION
hug-tdn-all-enhanced	0.52	0.49	0.49	0.95	Details
hug-tdn-all	0.54	0.48	0.49	0.95	Details
hug-tdn-dbnldg	0.48	0.47	0.46	0.94	Details
hug-tdn-1600-1900	0.43	0.40	0.40	0.93	Details
hug-tdn-1400-1600	0.43	0.40	0.40	0.93	Details
hug-tdn-bab	0.31	0.27	0.28	0.91	Details
pie-tdn-all	0.42	0.36	0.38	0.91	Details
hug-tdn-clvn	0.35	0.25	0.26	0.90	Details
hug-tdn-cour	0.29	0.27	0.26	0.90	Details
pie-tdn-1600-1900	0.35	0.28	0.30	0.88	Details
pie-tdn-dbnldg	0.38	0.31	0.33	0.87	Details
pie-tdn-1400-1600	0.31	0.28	0.29	0.85	Details
pie-tdn-bab	0.25	0.18	0.19	0.80	Details
pie-tdn-clvn	0.21	0.17	0.17	0.73	Details
pie-tdn-cour	0.19	0.16	0.16	0.70	Details

Dataset: tdn-1600-1900 | Annotation: PoS | Group by: PoS | Single/multiple analysis: Both

TAGGER	MACRO PRECISION	MACRO RECALL	MACRO F1	MICRO ACCURACY	DETAILED EVALUATION
hug-tdn-all-enhanced	0.56	0.57	0.55	0.96	Details
hug-tdn-all	0.56	0.55	0.53	0.96	Details
hug-tdn-1600-1900	0.50	0.53	0.50	0.96	Details
hug-tdn-dbnldg	0.50	0.53	0.49	0.95	Details
hug-tdn-bab	0.35	0.33	0.32	0.93	Details
hug-tdn-1400-1600	0.36	0.39	0.35	0.93	Details
pie-tdn-all	0.47	0.42	0.43	0.92	Details
hug-tdn-cour	0.32	0.35	0.31	0.91	Details
pie-tdn-1600-1900	0.47	0.39	0.41	0.91	Details
hug-tdn-clvn	0.31	0.28	0.27	0.90	Details
pie-tdn-dbnldg	0.46	0.40	0.41	0.88	Details
pie-tdn-bab	0.29	0.22	0.23	0.83	Details
pie-tdn-1400-1600	0.29	0.30	0.28	0.82	Details
pie-tdn-clvn	0.21	0.20	0.18	0.72	Details
pie-tdn-cour	0.23	0.20	0.20	0.71	Details

Dataset: couranten | Annotation: PoS | Group by: PoS | Single/multiple analysis: Both

TAGGER	MACRO PRECISION	MACRO RECALL	MACRO F1	MICRO ACCURACY	DETAILED EVALUATION
hug-tdn-all-enhanced	0.73	0.69	0.70	0.96	Details
hug-tdn-1600-1900	0.72	0.69	0.70	0.95	Details
hug-tdn-all	0.71	0.69	0.70	0.95	Details
hug-tdn-cour	0.73	0.73	0.73	0.95	Details
hug-tdn-dbnldg	0.64	0.66	0.65	0.94	Details
hug-tdn-bab	0.50	0.48	0.48	0.92	Details
pie-tdn-all	0.72	0.70	0.70	0.92	Details
hug-tdn-1400-1600	0.52	0.51	0.51	0.92	Details
pie-tdn-1600-1900	0.64	0.66	0.65	0.92	Details
hug-tdn-clvn	0.51	0.50	0.49	0.91	Details
pie-tdn-cour	0.56	0.49	0.50	0.87	Details
pie-tdn-dbnldg	0.50	0.50	0.50	0.86	Details
pie-tdn-1400-1600	0.50	0.50	0.50	0.85	Details
pie-tdn-bab	0.37	0.37	0.37	0.82	Details
pie-tdn-clvn	0.49	0.41	0.43	0.77	Details

Figuur 2: benchmarks PoS-toekenning

Dataset: tdn-core-all | Annotation: Lemma | Group by: Lemma | Single/multiple analysis: Both

TAGGER	MACRO PRECISION	MACRO RECALL	MACRO F1	MICRO ACCURACY	DETAILED EVALUATION
hug-tdn-all-enhanced	0.72	0.71	0.71	0.93	Details
hug-tdn-all	0.69	0.68	0.68	0.90	Details
hug-tdn-dbnldg	0.68	0.68	0.67	0.90	Details
hug-tdn-1600-1900	0.67	0.66	0.66	0.89	Details
hug-tdn-1400-1600	0.68	0.67	0.67	0.89	Details
hug-tdn-bab	0.63	0.62	0.62	0.88	Details
hug-tdn-clvn	0.64	0.63	0.63	0.88	Details
hug-tdn-cour	0.63	0.62	0.62	0.88	Details
pie-tdn-all	0.48	0.47	0.47	0.88	Details
pie-tdn-dbnldg	0.44	0.43	0.43	0.86	Details
pie-tdn-1600-1900	0.36	0.36	0.35	0.84	Details
pie-tdn-1400-1600	0.29	0.29	0.28	0.81	Details
pie-tdn-bab	0.13	0.12	0.12	0.72	Details
pie-tdn-cour	0.11	0.10	0.10	0.67	Details
pie-tdn-clvn	0.07	0.07	0.07	0.66	Details

Dataset: tdn-1600-1900 | Annotation: Lemma | Group by: Lemma | Single/multiple analysis: Both

TAGGER	MACRO PRECISION	MACRO RECALL	MACRO F1	MICRO ACCURACY	DETAILED EVALUATION
hug-tdn-all-enhanced	0.76	0.76	0.75	0.94	Details
hug-tdn-all	0.72	0.71	0.71	0.90	Details
hug-tdn-1600-1900	0.71	0.71	0.70	0.90	Details
hug-tdn-dbnldg	0.71	0.71	0.70	0.90	Details
hug-tdn-1400-1600	0.70	0.69	0.69	0.90	Details
hug-tdn-bab	0.68	0.67	0.67	0.89	Details
hug-tdn-cour	0.68	0.67	0.67	0.89	Details
hug-tdn-clvn	0.68	0.67	0.67	0.89	Details
pie-tdn-all	0.52	0.52	0.51	0.88	Details
pie-tdn-1600-1900	0.47	0.47	0.46	0.87	Details
pie-tdn-dbnldg	0.49	0.48	0.48	0.86	Details
pie-tdn-1400-1600	0.28	0.28	0.27	0.78	Details
pie-tdn-bab	0.16	0.16	0.16	0.75	Details
pie-tdn-cour	0.14	0.13	0.13	0.69	Details
pie-tdn-clvn	0.07	0.08	0.07	0.64	Details

Dataset: couranten | Annotation: Lemma | Group by: Lemma | Single/multiple analysis: Both

TAGGER	MACRO PRECISION	MACRO RECALL	MACRO F1	MICRO ACCURACY	DETAILED EVALUATION
hug-tdn-all-enhanced	0.72	0.72	0.72	0.92	Details
pie-tdn-all	0.60	0.61	0.60	0.89	Details
hug-tdn-all	0.66	0.67	0.66	0.88	Details
hug-tdn-1600-1900	0.66	0.67	0.66	0.88	Details
pie-tdn-1600-1900	0.58	0.58	0.58	0.88	Details
hug-tdn-dbnldg	0.65	0.66	0.65	0.88	Details
hug-tdn-cour	0.65	0.66	0.65	0.88	Details
hug-tdn-1400-1600	0.65	0.65	0.65	0.87	Details
hug-tdn-bab	0.64	0.64	0.64	0.87	Details
hug-tdn-clvn	0.63	0.64	0.63	0.87	Details
pie-tdn-dbnldg	0.49	0.49	0.49	0.83	Details
pie-tdn-cour	0.44	0.44	0.44	0.83	Details
pie-tdn-1400-1600	0.39	0.39	0.39	0.79	Details
pie-tdn-bab	0.26	0.25	0.25	0.72	Details
pie-tdn-clvn	0.17	0.18	0.17	0.66	Details

Figuur 3: benchmarks lemmatisering

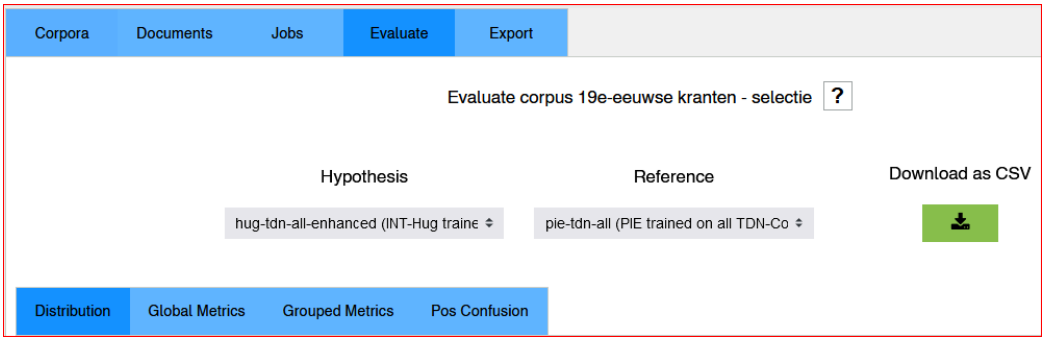
GaLAHaD bevat de nodige metrics, die in de helpsectie nader toegelicht worden. Een belangrijk verschil hierbij is macro naast micro, waarbij de macroscore de performance per categorie laat zien en de microscore evenveel gewicht geeft aan frequente als aan infrequente categorieën. De macroscore is dus belangrijk om te kijken of infrequente categorieën ook van een goede analyse worden voorzien. De benchmarks geven een globaal overzicht, maar er zijn nog veel andere waarden die in meer detail kunnen worden bekeken.

Wat hierbij meteen opvalt, is dat de Huggingface-taggers over het algemeen bij alle metrics beter scoren. Zelfs de taggers die getraind zijn op weinig data scoren hoog. Ook zie ik dat de taggers die getraind zijn op al het materiaal in al deze gevallen beter scoren dan de taggers die enkel op een deel van het materiaal getraind zijn. Ik kies er nu voor om zeven taggers te draaien, hoewel ik verwacht dat de drie taggers die op alles getraind zijn het beste resultaat opleveren. Het resultaat van de zeven getagde jobs is te zien in Figuur 4. Daar is ook te zien dat taggers anders omgaan met tokeniseren. Bij de Huggingface-taggers zijn er 63389 tokens, terwijl er bij de PIE-taggers 64169 tokens zijn. Op tokenisering kom ik aan het einde van deze sectie terug.

TAGGER	TAGSET	TYPE	TOKENS	PERIOD	LAST MODIFIED	PROGRESS	ACTIONS
pie-tdn-1600-1900	TDN-Core	TOK, POS, LEM	64169	**1600 – 1900**	2-07-24 17:31	100%	View & Tag
pie-tdn-cour	TDN-Core	TOK, POS, LEM	64169	**1600 – 1700**	2-07-24 17:33	100%	View & Tag
hug-tdn-all-enhanced	TDN-Core	TOK, POS, LEM	63389	**1400 – 1900**	2-07-24 19:55	100%	View & Tag
pie-tdn-all	TDN-Core	TOK, POS, LEM	64169	**1400 – 1900**	2-07-24 17:22	100%	View & Tag
hug-tdn-all	TDN-Core	TOK, POS, LEM	63389	**1400 – 1900**	2-07-24 19:54	100%	View & Tag
hug-tdn-cour	TDN-Core	TOK, POS, LEM	63389	**1600 – 1700**	2-07-24 19:46	100%	View & Tag
hug-tdn-1600-1900	TDN-Core	TOK, POS, LEM	63389	**1600 – 1900**	2-07-24 19:51	100%	View & Tag

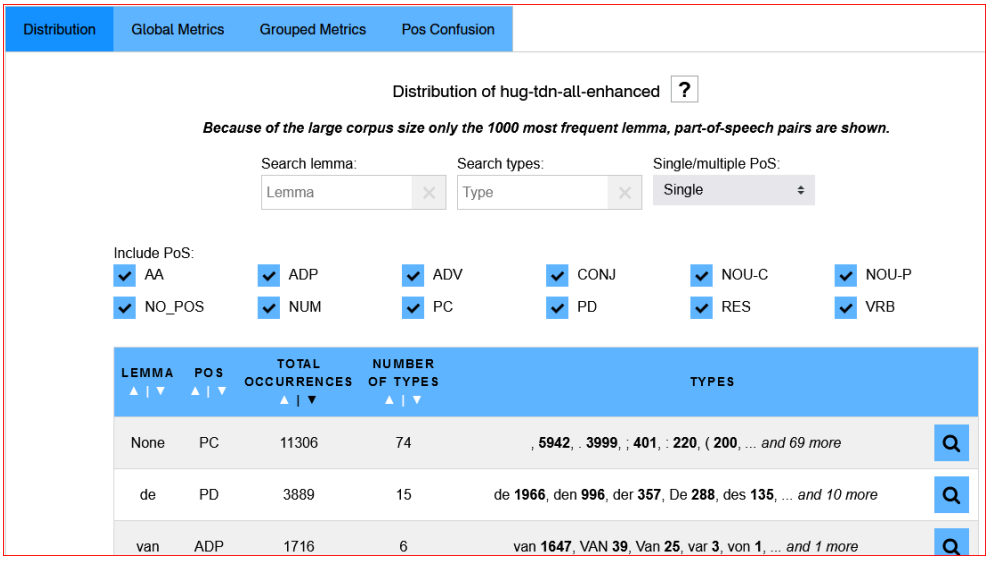
Figuur 4: de gebruikte taggers

Hierna kunnen de data geëvalueerd worden. De *confusion, distribution* en de metrics (in elf verschillende bestanden) kunnen ook gedownload worden als CSV en in Excel bekeken worden. Omdat ik geen gouden standaard heb, bieden de metrics voor mij weinig meer uitkomst dan wat de benchmarks al aangaven. Daarom is de evaluatiesectie voor mij nuttig om in meer detail te kijken naar *distribution* (één tagger) en naar *PoS confusion* (twee taggers). Ik kan twee taggers vergelijken door de ene als hypotheselaag te kiezen en de andere als referentielaag, zoals ook in Figuur 5 te zien is.



Figuur 5: een overzicht van de evaluatiepagina

De distributiesectie (Figuur 6) geeft een overzicht van de meest voorkomende (in kleine corpora alle) lemmata met alle types die daarbij horen. Ook is het mogelijk om een of meerdere woordsoorten te bekijken door gebruik te maken van de vinkvakjes, hoewel het wel wat veel klikken is als je een specifieke woordsoort wil bekijken. Deze pagina is vooral bedoeld voor een algemeen beeld, maar kan ook nuttig zijn om taggerspecifieke resultaten weer te geven. De Huggingface-taggers geven bij de lemmaloze tags met PoS PC bijvoorbeeld ook enkele tientallen woorden waar leestekens aan vastzitten. De tokenisering zou op dit front verbeterd moeten worden (zie verder sectie 1.2).



Figuur 6: de distributiepagina

Bij de PoS Confusion is het mogelijk om de door mij gekozen taggers te vergelijken en te zien hoe de resultaten per woordsoort verschillen. Figuur 7 laat bijvoorbeeld de vergelijking tussen Hug-tdn-all-enhanced en pie-tdn-all zien. Alle genummerde velden kunnen hier in detail bekeken en indien gewenst ook gedownload worden. Wat hier opvalt, is dat er veel woordsoorttoekenningen overlappen (in groen). Dit is gezien de hoge (boven 90%) *micro accuracy* van beide taggers ook niet onverwacht.

Part-of-speech confusion ?

PART-OF-SPEECH (HUG-TDN-ALL-ENHANCED→) (PIE-TDN-ALL↓)	AA	ADP	ADV	CONJ	INT	NOU-C	NOU-P	NUM	PD	RES	VRB	PC	NO_POS	MISSING MATCH	MULTIPLE
AA	3409	16	56	2	0	159	44	11	20	3	95	3	3	7	16
ADP	5	7309	25	10	0	10	83	18	1	1	20	8	6	7	22
ADV	28	64	2077	42	1	70	8	6	26	3	33	7	0	1	12
CONJ	3	2	16	2600	0	47	0	3	25	3	2	5	0	2	0
INT	0	0	0	0	4	2	0	2	0	0	0	0	0	0	2
NOU-C	228	30	61	4	2	10664	512	80	20	96	107	5	66	65	186
NOU-P	125	40	31	0	3	841	2758	19	62	53	16	5	16	50	182
NUM	6	5	12	6	0	40	38	2528	16	14	0	9	36	21	68
PD	10	3	32	21	0	30	40	39	8872	26	8	9	5	13	4

Figuur 7: een voorbeeld van de PoS confusion

Als ik in detail kijk naar de verschillen, dan blijkt met name de PIE-tagger meer onjuiste resultaten oplevert, zoals de toekenning NOU-P voor tokens als *Adres, Javaas, Rouwbeklag* en *VRAAGT* van pie-tdn-all. Dit, gecombineerd met de lagere scores van de benchmarks, maakt het voor mij duidelijk dat de Huggingface-taggers waarschijnlijk een beter resultaat geven. Ik bekijk hierna Hug-tdn-all en Hug-tdn-all-enhanced. De resultaten liggen hierbij heel dicht bij elkaar, waardoor het voor mij aannemelijk wordt dat beide taggers een goed resultaat zullen geven bij de download. Omdat hug-tdn-all-enhanced getraind is op scheidbare werkwoorden, bekijk ik bij de distributie nog de werkwoorden van de andere Huggingface-taggers. Die taggers hebben het lemma *aanwerken* met drieëntwintig keer het type *aan* en *afvangen* met negen keer het type *af*. Hoewel dit in theorie mogelijk is, verwacht ik dat deze lemmatisering niet klopt (en dit bleek later ook het geval). Omdat de scheidbare werkwoorden vaker een correct lemma krijgen, kies ik Hug-tdn-all-enhanced als tagger.

Na de evaluatie kunnen de verrijkte documenten geëxporteerd worden. Als het exportformaat hetzelfde is als het inputformaat kan de verrijkte laag erbij komen in de oude bestanden door voor de *merge*-optie te kiezen. LAnCeLoT ondersteunt momenteel het TEI P5-formaat, dus ik kies dit formaat om de bestanden te downloaden.

1.2 Tokeniseerproblemen

Bij het vergelijken in de PoS confusion en bij de XML in de download blijkt dat tokeniseren niet altijd goed gaat. Het is voor de taggers niet altijd duidelijk wat bij een woord hoort en wat, met name als er leestekens voor of achter (of midden in) het token staan. Die parsering heeft invloed op de tagging en zou idealiter verbeterd moeten worden. Wat hoort bij een token en wat niet? In de meeste gevallen zouden de leestekens (waaronder ook haakjes vallen) geen deel uit moeten maken van de tokens. Anders zijn *aangekondigd* en *aangekondigd]* twee aparte types, terwijl het eigenlijk één type zou moeten zijn. Dit leidt bovendien soms tot andere problemen, zoals het toekennen van een PC-tag aan *[het*, terwijl dat eigenlijk als woord getagd moet worden. De afbrekingspunt (zoals *gebr.* voor *gebroeders*) zou deel uit moeten maken van een token, maar de zinseindepunt niet. De Huggingface-taggers tokeniseren (soms) beide soorten punten aan de lemmata vast, terwijl de PIE-taggers nooit een punt erbij tokeniseren. In de toekomst moet er (meer) aandacht aan tokeniseren worden gegeven, zodat dergelijke problemen (nagenoeg) niet meer voor zouden moeten komen.

1.3 Wensen en verbeterpunten

- **Meer taggers.** Alle taggers zijn nu getraind op de tagset TDN-Core, maar bij het overzicht van de tagsets wordt ook Gysseling/CRM en CGN/D-Coi gegeven. Op termijn zou het mooi zijn als er ook taggers met een andere tagset (zoals deze) in zouden hangen, zodat daar ook mee vergeleken kan worden.
- **Ongekoppelde tokens.** Bij de export en met name later in LAnCeLoT bleek dat tokens als *aan* en *drong* weliswaar beide dezelfde analyse hadden, maar dat ze niet aan elkaar verbonden waren (geen groep). Die informatie zou eigenlijk ook gedownload moeten worden, zodat deze vervolgens meegenomen kan worden in LAnCeLoT.
- **Veel soortgelijke taggers.** Er hangen nu vrij veel taggers in het platform met vergelijkbare resultaten. Daarbij valt op dat de taggers beter scoren als ze getraind zijn op meer materiaal, op (vrijwel) alle metrics. Dit geldt met name voor de PIE-taggers. De vraag is of al die taggers hier aangeboden zouden moeten worden. Aan de andere kant is het verlies aan informatie misschien ook niet wenselijk. Een optie zou kunnen zijn om ze wel in de benchmarks te tonen (puur informatief) en niet als tagger aan te bieden in het platform. Anderzijds zou er ook een pagina met een uitgebreide taggerselectie kunnen komen, die niet bij de benchmarks staat, en de benchmarks alleen weer te geven voor taggers die ook in het platform aangeboden worden.
- **Vinkvakjes distribution.** Als je de types bij één specifieke woordsoort wil bekijken, moet je ze nu allemaal een voor een uitklikken tot er een overblijft. Alle vinkjes staan weer aan als je naar een andere tab gaat, dus als je wisselt tussen de evaluatietabbladen, is dit onwenselijk. Het zou een (kleine) verbetering zijn om met minder klikken de relevante woordsoorten te selecteren.

2. LAnCeLoT

2.1 Handmatige correctie van lemmatisering en part-of-speechtagging

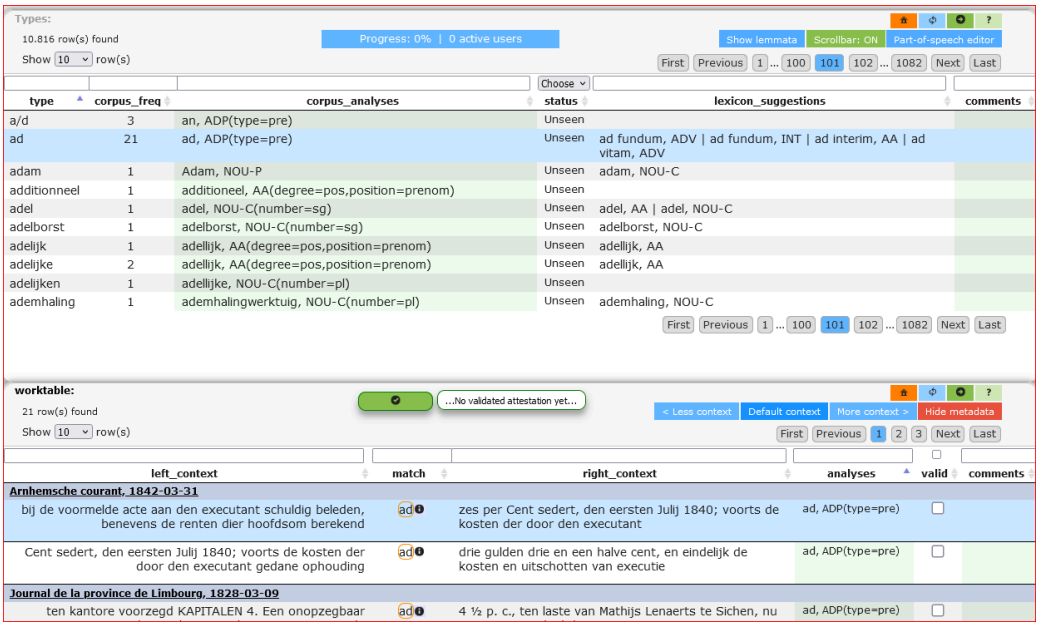
LAnCeLoT is een omgeving waarin automatisch taalkundig verrijkte bestanden gecorrigeerd kunnen worden. Ik maak hierin een nieuw project aan en doorloop de stappen om mijn corpus te uploaden:

1. Ik geef het project een naam, in dit geval *19e-eeuwse kranten*;
2. Ik upload het corpus in LAnCeLoT Search. Door op de knop te drukken, open ik een nieuw tabblad. Daar maak ik een corpus aan, dat ik de naam *19e-eeuwse kranten* geef. In dat corpus upload ik vervolgens het zipje met de gedownloade documenten. Hierna keer ik terug naar het oude tabblad en druk ik op Ok;[4]
3. Ik kies de default tagset TDN-core en klik op Ok. Daarna volgt een kort overzicht van 'verboden parts-of-speech'. Deze zijn vooral in het leven geroepen om analyses te voorkomen waar een waarde *unclear* (alle waardes zijn onduidelijk) wordt verwacht. Zo is NOU-C(number=sg|pl) verboden, zodat alle onduidelijke gevallen aangegeven kunnen worden met uncl, zoals in NOU-C(number=uncl). Ik klik vervolgens nog een keer op Ok;
4. Ik kies het default lexicon GiGaNT-HiLex en klik een laatste keer op Ok. Voor dit project zou het evengoed mogelijk zijn om alleen de suggesties uit de WNT-tijd (vanaf 1500) te kiezen.

In Figuur 8 is te zien hoe het project eruit ziet (zoekresultaat van een willekeurige pagina): de bovenstaande tabel bevat alle types, de onderstaande alle tokens die horen bij een specifiek type (in onderstaande voorbeeld het type *ad,* met 21 tokens die daarbij horen). Naast de toegekende analyses worden er ook lexiconsuggesties gegeven die corresponderen met het type. *Ad* komt in het lexicon dus als token voor bij de lemmata *ad fundum, ad interim* en *ad vitam*. Deze lexiconsuggesties kunnen helpen om het lemma en/of de woordsoort te bepalen van de match in kwestie. Dit voorbeeld laat

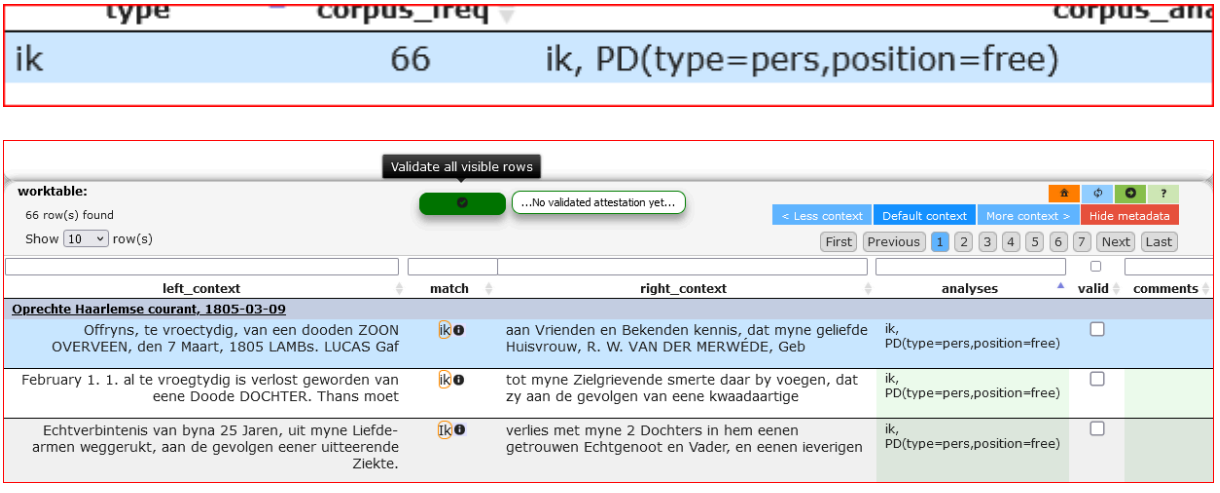
[4] Vanuit de landingspagina of een individueel project is het altijd mogelijk om het corpus in zijn geheel te bekijken door naar LAnCeLoT Search te gaan.

echter ook zien dat niet alles in het woordenboek is opgenomen. Er is namelijk geen voorzetsel *ad* in de woordenboeken te vinden, terwijl *ad* hier wel 21 keer als voorzetsel fungeert. Misschien zou dat op termijn wel kunnen worden toegevoegd, zie verder sectie 2.2.



Figuur 8: een voorbeeld van de typetabel (boven) en de worktable (onder)

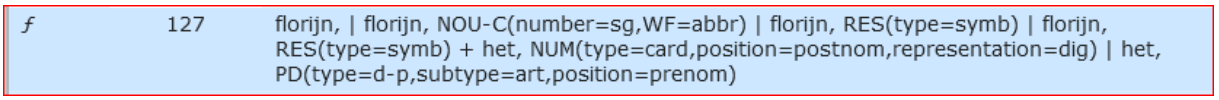
Vanuit de typetabel kunnen per type alle bijbehorende tokens in de *worktable* worden opgevraagd, zoals in Figuur 9 te zien is. Je kunt hier (net als in GaLaHaD) ook sorteren of filteren door iets in te typen in de informatievelden (reguliere expressies worden ook ondersteund). Een van de grote voordelen van werken met deze applicatie is dat analyses dus ook per type kunnen worden beoordeeld. Dit betekent bijvoorbeeld dat niet alle zesenzestig tokens van *ik* afzonderlijk in de loop van de tekst hoeven te worden gevalideerd, maar dat ze tegelijkertijd (maar bijvoorbeeld ook per tien) kunnen worden weergegeven in de *worktable* en met een druk op de knop (zie hieronder) allemaal kunnen worden gevalideerd.



Figuur 9: (boven) het type ik en (onder) de tokens die behoren tot dat type

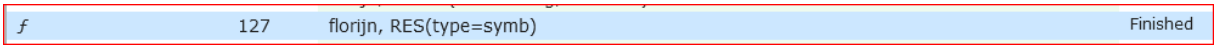
Een groot voordeel hiervan is de consistentie van de data, zeker wanneer meerdere mensen aan hetzelfde project werken. Zo kan het in een project voorkomen dat het florijnteken soms geen analyse heeft, soms als zelfstandig naamwoord is getagd (de ene keer met WF voor afkorting, de andere keer

geen WF voor afkorting) of soms als symbool is getagd, zoals hieronder te zien is bij de automatische PoS-toekenning in Figuur 10.



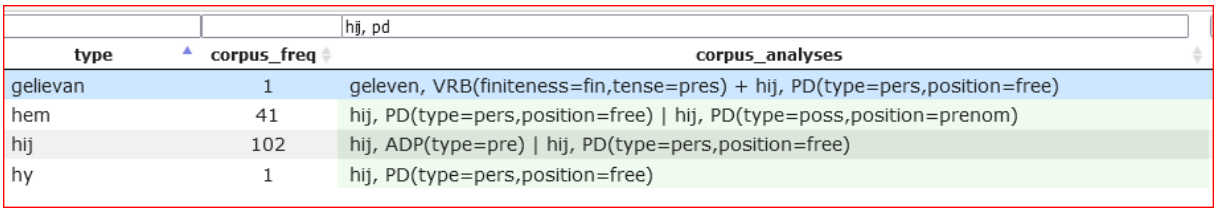
Figuur 10: automatisch toegekende analyses aan het type f

Door alle gevallen tegelijkertijd te behandelen, kunnen ze allemaal gemakkelijk dezelfde analyse krijgen. In Figuur 11 is te zien dat alle 127 tokens nu dezelfde analyse hebben.



Figuur 11: handmatig gecorrigeerde analyses bij het type f

Het is ook eenvoudig om vreemde of overduidelijk foute analyses op te sporen (als je weet hoe je moet zoeken). In Figuur 12 heb ik bijvoorbeeld gezocht op *hij*, PD om alle types op te vragen waarin *hij* als voornaamwoord is getagd.



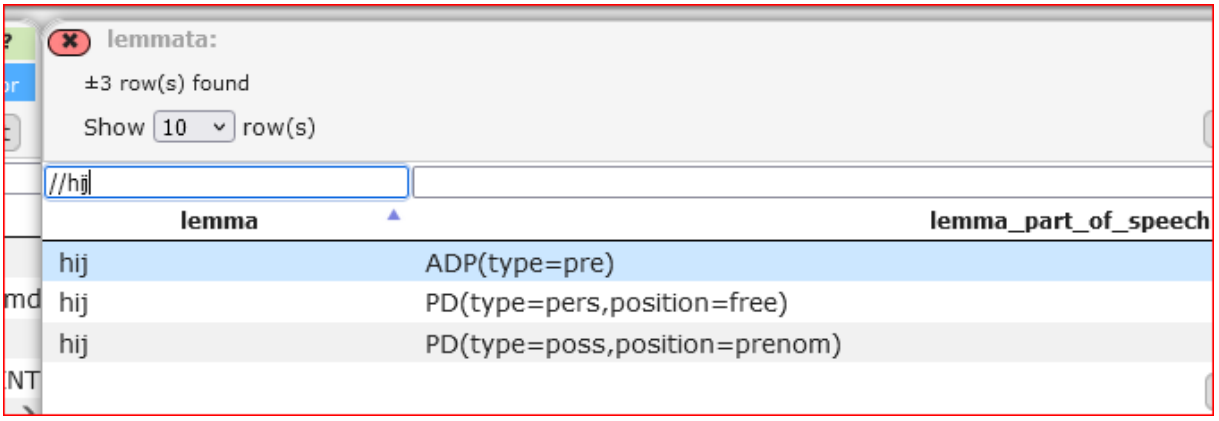
Figuur 12: alle voorkomens van hij als persoonlijk voornaamwoord

Hierbij kunnen gelijk al een aantal fouten worden opgespoord, namelijk dat:

- *hij* bij het type *hem* een of meer keer is getagd als bezittelijk voornaamwoord
- *hij* bij het type *hij* een of meer keer is getagd als voorzetsel (hier een toevallige bijvangst, omdat ik niet heb gezocht op **hij, ADP** of op **hij,**)
- *hij* bij het type *gelievan* deel uitmaakt van een multiple lemma

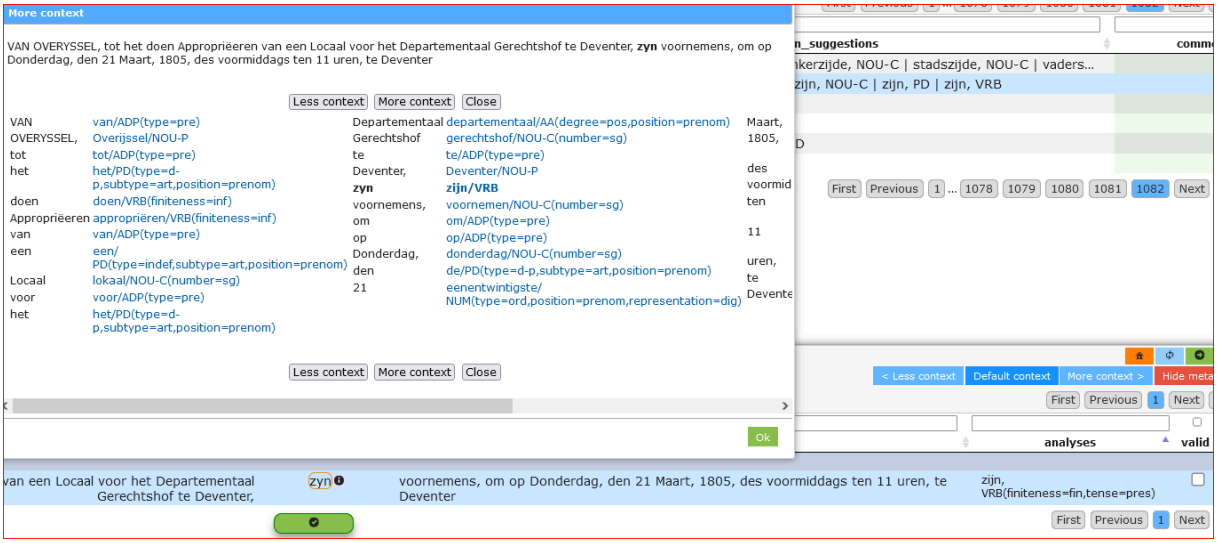
De analyses kunnen door de types in de worktable te laden eenvoudig worden verbeterd, waarna ze ook verdwijnen in de typetabel wanneer ik opnieuw op *hij*, PD zou zoeken.

De lemmatatabel kan gebruikt worden als referentiepunt, maar ook om vreemde analyses na te lopen of op te sporen, als alternatieve mogelijkheid. In Figuur 13 (hetzelfde als hierboven) zocht ik bijvoorbeeld naar alle analyses die horen bij een bepaald lemma, in dit geval *hij*. Net als hierboven wordt duidelijk dat aan *hij* de foutieve analyses *type=poss* (bezittelijk voornaamwoord) en ADP (voorzetsel) zijn toegekend. Het is natuurlijk ook mogelijk dat de PoS-analyse wel klopt, maar de lemmatisering niet (in dit geval klopt de PoS wel).



Figuur 13: alle analyses behorend tot het lemma hij

Soms kan *more context* helpen om te kijken hoe omliggende woorden zijn getagd en/of om te kijken hoe de match kan worden geïnterpreteerd. In het onderstaande voorbeeld, *zyn voornemens*, zou het immers kunnen gaan om een bezittelijk voornaamwoord *zijn* bij een meervoud *voornemens* of om het werkwoord *zijn* met het bijwoord *voornemens*. Een blik op de context (en de automatisch toegekende analyse) helpt dan om te bepalen dat *zyn* in dit geval als werkwoord getagd moet worden, zoals in Figuur 14 te zien is.

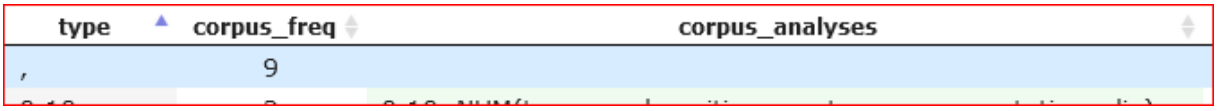


Figuur 14: more context in de worktable

Nadat ik klaar ben met het verbeteren van de annotaties kan ik het project exporteren. Bij de export zitten alle bestanden in een zip-folder. De bestanden hebben dezelfde naam als bij de upload, zodat ze makkelijk teruggevonden kunnen worden.

<u>Aandachtspunt</u>

- Bij het verrijken van een bestand in GaLAHaD kan het voorkomen dat een tagger interpunctie als token meeneemt, zoals een komma als type. Idealiter worden die niet getagd en meegenomen, maar dat kan wel voorkomen. De analyse kan in LAnCeLoT makkelijk worden weggehaald, zodat er geen vreemde analyse aan een leesteken wordt meegegeven, zoals bij de komma in Figuur 15.

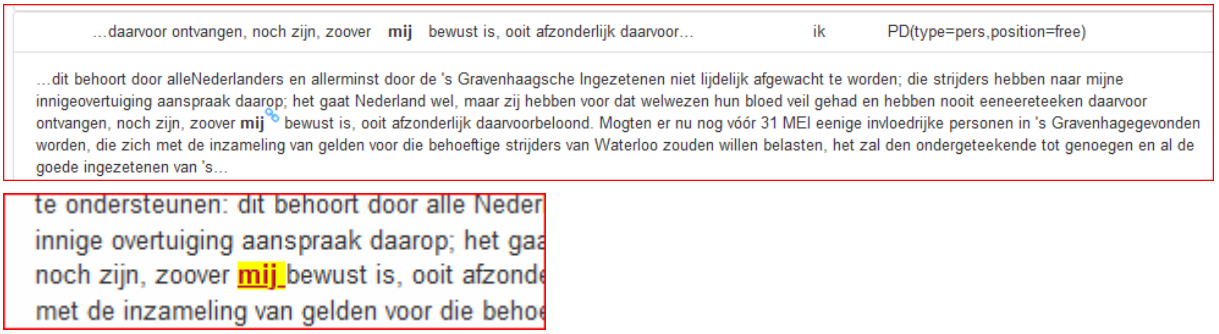


Figuur 15: geen toegekende analyse(s) bij het type ,

2.2 Wensen en verbeterpunten

- **Toevoegen/suggereren nieuwe woordvormen voor het lexicon (contribute).** In het corpus zaten de nodige woordvormen (soms zelfs woorden) zonder lexiconsuggestie. Dit is logisch voor de getallen, waar veel variatie zit in de woordvorm (vanwege punten, komma's en dergelijke), of voor de types waar leestekens in zitten, maar ook met die types buiten beschouwing gelaten, zijn er nog de nodige woordvormen zonder lexiconsuggestie. Het zou misschien een idee zijn om in een contributesectie (deze zit nog niet in de huidige versie van LAnCeLoT) data ter beschikking te stellen, zodat op termijn het centrale lexicon aangevuld kan worden ten behoeve van andere gebruikers. Ook zou het handig zijn om suggesties uit het hedendaagse lexicon, GiGaNT-Molex,[5] aan te bieden, omdat daar in principe alle standaardtalige paradigmavormen van elk lemma in zitten. Dit is misschien minder nodig, omdat die vormen vaak makkelijk te doorgronden zijn (hoewel dit wel weer wat kan opleveren met betrekking tot bijvoorbeeld (met name geografische) eigennamen).

- **Verwijzing naar plek in de tekst.** Vanuit het corpus (in LAnCeLoT Search) kan je altijd direct naar de relevante passage, omdat er pagina-informatie wordt meegegeven. Elke pagina heeft duizend tokens en met de link uit het corpus ga je direct naar het relevante stukje in de tekst, zoals te zien is bij Figuur 16.

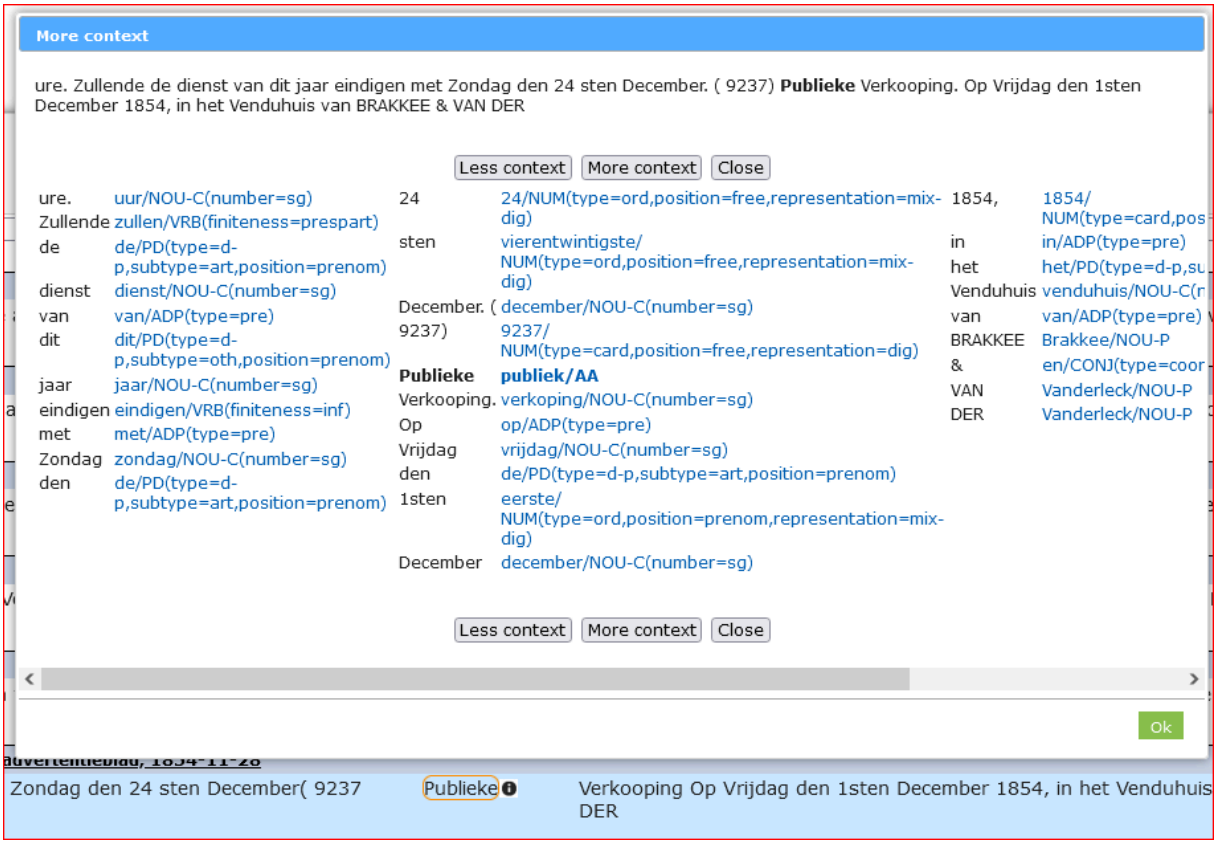


Figuur 16: (boven) een voorbeeld van mij in het corpus en (onder) de plek van mij in de tekst

In LAnCeLot is dit nog niet mogelijk. Idealiter zou je bij *more context* een link willen van de match naar de relevante plek in de tekst. Dat werkt dan alleen als de pagina-informatie ook wordt meegegeven, zoals bij de zoekvraag in het corpus. Die informatie moet worden meegegeven, zodat je ook bij grotere bestanden altijd met één klik op bij de relevante passage in het document terechtkomt.

- **Weergave haakjes.** In Figuur 17 zijn geen sluitingshaakjes te zien (alleen in *more context*) en het openingshaakje hangt verkeerdelijk vast aan het voorgaande token.

[5] Zie https://ivdnt.org/corpora-lexica/gigant/.



Figuur 17: de weergave van haakjes in more context

Als de tokenisering wordt verbeterd, wordt daarmee hopelijk de weergave van de haakjes ook verbeterd.

3. Slotopmerkingen

Met de bovengenoemde applicaties is het nu voor alle gebruikers eenvoudig en mogelijk om zelf een (al dan niet verrijkt) corpus of bestanden van taalkundige informatie te voorzien en deze indien gewenst ook zelf te corrigeren. De applicaties zijn toegankelijk voor iedereen met een CLARIN-account.

GaLAHaD: https://portal.clarin.ivdnt.org/galahad/home

LAnCeLoT: https://portal.clarin.ivdnt.org/lancelot/lexit2/?db=lancelot

Tot slot zou ik Katrien Depuydt graag bedanken voor het nalezen en becommentariëren van dit verslag.