Intra-author variation in negation: the letters of P.C. Hooft

Feike Dietz, Marjo van Koppen, Irene Kramer, Marijn Schraagen – Universiteit Utrecht

Eindrapportage pilot Nederlab – Augustus 2017

Inhoud rapportage:

1. Achtergrond en doelstellingen pilotstudie

In de Gouden Eeuw was het Nederlands volop in beweging. Als eenheidstaal van de nieuwe Republiek werd het Nederlands in steeds meer domeinen van de samenleving gebruikt (zoals het religieuze en wetenschappelijke domein) en er werden vele pogingen ondernomen om de positie van de moedertaal te versterken en de taal te standaardiseren. Ook natuurlijke taalontwikkelingen hadden een impact op het Nederlands: steeds meer eigenschappen uit het Middelnederlands (bijvoorbeeld naamval) verdwenen om plaats te maken voor nieuwe eigenschappen (zoals het gebruik van voorzetselgroepen) (Van der Sijs & Willemijns 2009; Van der Sijs 2004).

Het onderzoek naar het Nederlands in de Gouden Eeuw richt zich tot nu toe veelal op (i) de invloed van standaardisatie op taalverandering (cf. Nobels & Rutten 2014), (ii) diachrone ontwikkelingen (cf. Zeijlstra 2004), en (iii) de variatie tussen (groepen) sprekers (hoog-/laagopgeleid, mannen/vrouwen) (cf. Nobels 2013; Vosters & Vandenbussche 2012). Het Utrechtse onderzoeksproject 'Language Dynamics in the Dutch Golden Age' (NWO, Vrije Competitie, 2016-2020, https://languagedynamics.wp.hum.uu.nl/), daarentegen, richt zich op een type variatie die tot nu toe nog weinig is bestudeerd: de variatie binnen individuele taalgebruikers, ofwel intra-auteurvariatie. Via een interdisciplinaire aanpak – we combineren theoretisch-taalkundige, letterkundige, sociolinguïstische en computationele methodiek – proberen wij te begrijpen welke factoren variatie in de geschreven taal van een taalgebruiker tot stand brachten. Wat zijn de patronen van intra-auteurvariatie in de Nederlandse 17e eeuw, en hoe kunnen we die verklaren? Onze hypothese is dat de variatie het resultaat was van een dynamische interactie tussen het *interne taalsysteem* van taalgebruikers enerzijds en hun *sociaal/literair-culturele context* anderzijds. Het taalsysteem van een taalgebruiker maakte variatiemogelijkheden beschikbaar, die vervolgens door een taalgebruiker systematisch en vaak strategisch werden ingezet, afhankelijk van bijvoorbeeld het publiek of de doelstellingen en literaire vormgeving van zijn tekst.

Eén van de terreinen waarop het zeventiende-eeuws Nederlands variatie binnen taalgebruikers vertoont, is negatie. In het Middelnederlands werden zinnen ontkennend gemaakt door tweeledige negaties van het type *en…niet* (vergelijk het Franse *ne…pas*). In de zeventiende eeuw maakte deze vorm van negatie langzaam plaats voor eenledige negaties: *ic **en** sal **niet** moghen gaen* werd steeds vaker *ik zal **niet** mogen gaan* (Jespersen 1917; Van der Wouden 2007). Auteurs gebruikten de oude en de nieuwe vorm van negatie vaak door elkaar. Zo weten we dat P.C. Hooft in 1638 plotseling stopte met het gebruik van tweeledige negatie (Paardekooper 2006, p. 126), maar in

de periode daarvóór een- en tweeledige negatie door elkaar heen gebruikte. Was dat toeval of zit er systematiek achter zijn keuzes? Omdat Hooft een groot corpus aan teksten heeft nagelaten - schreef tussen 1600 en 1638 bijna 800 Nederlandstalige brieven - kunnen we zo'n vraag systematisch en op kwantitatieve wijze onderzoeken. Daarvoor is het echter noodzakelijk dat die brieven morfo-syntactisch en sociolinguïstisch verrijkt zijn, zodat ze digitaal doorzocht kunnen worden. Zo'n verrijkt corpus stelt ons in staat om vast te stellen of Hooft tweeledige negatie vaker gebruikte in combinatie met bepaalde type werkwoorden of bepaalde type adressanten.

Nederlab biedt de mogelijkheid om teksten van de DBNL – waaronder ook Hoofts brieven – digitaal te doorzoeken met Corpus Query Language. Het Nederlab-corpus is getagd met Frog, geschikt voor de tagging van het moderne Nederlands. Het resultaat van de Frog-tagging is niet toereikend om onze onderzoeksvragen te kunnen beantwoorden. Niet alleen maakt de tagger in het geval van zeventiende-eeuwse teksten veel fouten, maar ook worden niet-moderne taalverschijnselen zoals tweeledige negatie niet herkend. Sociolinguïstische annotatie ontbreekt daarnaast geheel in Nederlab.

Onze pilotstudie is erop gericht om met behulp van een nieuwe annotatietool Hoofts brieven uit de periode 1600-1638 te voorzien van zowel morfosyntactische als sociolinguïstische tagging, en de verrijkte brieven terug te plaatsen in Nederlab, waardoor er verbeterde CQL-zoekopdrachten uitgevoerd kunnen worden binnen de context van Nederlab. Binnen de context van ons eigen onderzoeksproject kunnen de verrijkte data niet alleen gebruikt worden voor verdere analyse, maar ook fungeren als trainingsdata zijn waarmee digitale tagging van vroegmoderne teksten verbeterd kan worden.

2. Aanpak pilotstudie

Stap 1: PoS-Tagging van brieven van Hooft door Adelheid (Rem & Van Halteren 2011), beschikbaar gesteld door Hans van Halteren, digitale editie brieven Hooft (Tricht 1977) beschikbaar gemaakt door de DBNL

1.1 Data

De teksten van het project zijn afkomstig van DBNL (http://dbnl.nl/tekst/hoof001hwva00_01/). De eerste stap van het tagging-proces bestond uit het verkrijgen van het oorspronkelijke document in platte tekst-formaat, zodat de tagger en de annotatietool de teksten zouden kunnen gebruiken.

Op de website van DBNL worden de teksten in html- en pdf-formaat aangeboden, waarbij het pdf-bestand automatisch wordt gegenereerd uit de gedigitaliseerde data. De pdf-bestanden zijn als eerste bekeken, omdat deze de platte tekst van een document bevatten zonder de html-code eromheen. Echter, de pdf bevat nog wel alle para- of metatekst, zoals voetnoten, annotaties van de editie, paginanummers, etc. Zonder formele structuur is het lastig om deze elementen automatisch uit de tekst te verwijderen. Daarom zijn uiteindelijk de html-bronbestanden gebruikt, waarbij de tekst kon worden geïsoleerd met behulp van de opmaak-tags. Een belangrijke uitzondering hierop zijn de korte samenvattingen van de editie die bij veel brieven aanwezig zijn, die in de html op dezelfde manier worden weergegeven als de originele tekst. Om dit op te lossen hebben we de annotatoren gevraagd om de samenvattingen te markeren (zie sectie 4 en 5).

Na overleg met DBNL hebben we ook de beschikking gekregen over de XML-bestanden die als basis dienen voor het genereren van de html op de website van DBNL. Echter, deze bestanden bieden geen duidelijk voordeel t.o.v. de html-bestanden, waardoor besloten is de xml-bestanden niet te gebruiken.

1.2 Tagging

Handmatige tokenisatie, lemmatisatie en Part-of-Speech tagging is een tijdsintensieve en foutgevoelige taak. Een manier om de taak te vereenvoudigen is het gebruiken van een bestaande tagger en de fouten daarvan te corrigeren, in plaats van het volledig zelf annoteren van de teksten. Voor het huidige project is de tagger *Adelheid* gebruikt, ontwikkeld door Hans van Halteren en Margit Rem aan de Radboud Universiteit Nijmegen voor het taggen van Middelnederlands (Rem en Van Halteren 2011). Adelheid is getraind op 14e-eeuwse documenten uit het Corpus van Reenen-Mulder. Deze teksten verschillen taalkundig sterk van de 17e-eeuwse brieven van Hooft, maar bepaalde elementen uit het Middelnederlands zijn ook in het Hooft-corpus nog aanwezig (zoals naamvalgebruik en negatieclitics), waardoor de Adelheid-annotaties een goed uitgangspunt vormen voor de huidige annotatietaak.

Adelheid is beschikbaar met een Clarin-federated login via een webinterface (http://adelheid.ruhosting.nl/, zie ook http://dev.clarin.nl/node/1918). Voor het taggen van grote hoeveelheden bestanden en het automatiseren van het proces is deze webinterface echter minder geschikt. Daarom heeft Hans van Halteren na overleg een stand-aloneversie van Adelheid beschikbaar gesteld. Deze versie werkt onder Linux, en het bleek dat het programma zonder enige problemen overgezet kon worden op onze ontwikkel-pc.

Adelheid is vervolgens toegepast op alle brieven van Hooft. Dit kost een paar minuten per brief, in totaal ongeveer een dag en een nacht rekentijd. Adelheid levert twee uitvoerformaten: TEI-XML en een tekstgebaseerd kolomformaat. Het XML-formaat is iets uitgebreider, zo worden er alternatieve tags en waarschijnlijkheden opgeslagen die ontbreken in het tekstformaat. Echter, voor onze doeleinden was deze extra informatie niet nodig, terwijl het XML-formaat nadelen heeft wat betreft leesbaarheid, manier van verwerken en bestandsgrootte. Daarom is besloten het tekstformaat aan te houden, en dit ook als datamodel te gebruiken in de annotatietool.

Stap 2: Ontwikkelen annotatietool, GuSTAVE (**G**eautomatiseerde **S**ociolinguistische en **T**aalkundige **A**nnotatie van **V**roegmoderne **E**dities)

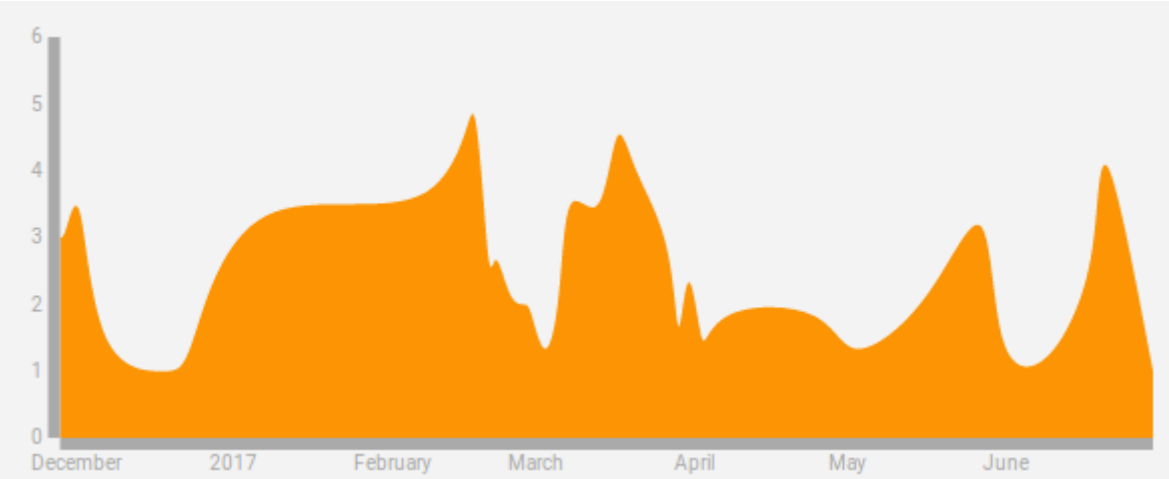
2.1 Features

De opzet van de annotatietaak zou met bestaande annotatietools niet eenvoudig kunnen worden geïmplementeerd, wat aanleiding gaf tot de ontwikkeling van een nieuwe annotatietool. De tool bevat de volgende features die een rol hebben gespeeld bij deze overweging:
– integratie van authenticatie, bestandsorganisatie en metadata voor gebruiksvriendelijke taakverdeling
– toegankelijkheid via webinterface
– ondersteuning van Adelheid-uitvoer, wat betreft het bestandsformaat, tagset, tagset-conversie, en specifieke features zoals definities van speciale tekens en tokens met dubbele tags
– ondersteuning voor aanpassen tokenisatie (combineren en splitsen van tokens)
– propagatie van handmatige correcties in een document
– domeinspecifieke sociolinguïstische tagging met verwijzingen naar externe bronnen
– integratie van uitvoer met huidige FoLiA-annotaties van Nederlab
– volledige controle over in- en uitvoer, mogelijkheid tot het maken van aanpassingen in de functionaliteit op elk niveau

2.2 Ontwikkeling

De tool is ontwikkeld vanaf eind november 2016 tot eind april 2017. De laatste paar weken van de ontwikkeling overlappen met het begin van het annoteren door de student-assistenten. In deze tijd zijn nog veel bugs, speciale gevallen, gewenste extra functionaliteit, en details van het toepassen

door meerdere gebruikers zichtbaar geworden en aangepast in de code. In de daaropvolgende periode is aandacht besteed aan het vergelijken van annotaties tussen gebruikers, en de integratie met de FoLiA-annotaties van Nederlab. De onderstaande afbeelding is een illustratie van het ontwikkelproces, getoond als het aantal wijzigingen in de code per dag.



De annotatie-interface maakt gebruik van een HTML/Javascript/jQuery frontend met een php/SQLite backend. De tools voor pre- en postprocessing zijn geprogrammeerd in Perl, Python en bash-scripting, met libxml voor de integratie met FoLiA. Adelheid maakt gebruik van (gecompileerde) C++ en Perl.

2.3 Resultaten

De resultaten van het taggen zijn beschikbaar in tekstgebaseerd kolomformaat, in afzonderlijke bestanden per brief en annotator. Daarnaast is de conversie naar FoLiA beschikbaar, waarbij automatische annotaties van Frog en handmatige annotaties naast elkaar in elk bestand zijn opgenomen. De sociolinguistische annotaties zijn beschikbaar in csv-bestanden. Op het moment van schrijven staat het precieze formaat van zowel de verrijkte FoLiA als de csv-uitvoer nog niet helemaal vast, hiervoor is aparte documentatie beschikbaar die nog zal worden aangepast waar nodig.

Stap 3: Uitbreiden van de PoS-tagging met aanvullende grammaticale kenmerken en toevoegen van sociolinguistische/letterkundige informatie.

Taalkundige tagging:

De annotatie bestaat uit een Lemma, een Part-of Speech-tag en een of meer Features. De PoS-tags en de features die door Adelheid worden aangeleverd zijn specifiek ontwikkeld om het Middelnederlands te annoteren. Het vroegmoderne Nederlands heeft veel kenmerken die ook in het Middelnederlands voorkomen, waardoor we het grootste deel van de Adelheid tagset konden handhaven. Voorbeelden van tags die specifiek zijn voor oudere taalfases zijn tags voor de specifieke morfologische uitgangen op lidwoorden, bijvoeglijke naamwoorden en andere modificeerders (zoals –en/-er/-es) waarmee de naamval kan worden herleid en een tag voor het negatieclitic. Er zijn echter ook nog tags en kenmerken die wij hebben toegevoegd om onze onderzoeksvragen beter te kunnen beantwoorden, zoals bijvoorbeeld of een element deel uitmaakt van een partikelwerkwoord en of een element nominatiefnaamval heeft of niet. Een volledig overzicht van de gebruikte tags en features kan worden gevonden in de annotatiehandleiding.

Sociolinguïstische/letterkundige tagging:

We weten dat taalgebruikers hun taal aanpassen aan de context waarin zij functioneren: zij spreken anders op hun werk dan tegen een vriendin, en anders over zaken dan over huiselijke aangelegenheden. Ook het genre dat taalgebruikers beoefenen bepaalt mede hoe zij hun taal inzetten. Om te kunnen onderzoeken of Hoofts negatiegebruik zich aanpaste aan zijn sociale en retorische context, annoteren we:

- het doel van een brief, denk aan overtuigen of bedanken
- het onderwerp van een brief, denk aan een brief over zakelijke aangelegenheden of literatuur
- de geadresseerde van een brief, en informatie over die persoon: naam, geboortedatum, gender, beroep

Omdat uit eerder onderzoek naar zeventiende-eeuwse brieven bleek dat formulaire delen van een brief, vaker 'ouderwetse' taal bevatten (cf. Nobels & Rutten 2014), annoteren we ook de structuur van een brief: wat is de formulaire opening, waar is de kern van de brief, waar is de afsluiting? Een voorbeeld met achtereenvolgens de begroeting, opening, narratio, afsluiting en eindgroet:

811 Aen mê Joffrouwe, Mê Joffre Tesselscha Visschers, weduwe van Sr Crombalgh z.g. in de Langestraet, tot Alkmaer.

loont.

Mê Joffre

De prujmen beginnen al teffens op een bodt te rijpen, en te roepen Tesseltje, Tesseltjes mondtje. Etlijke deuntjes van Belusar en andre roepen daer tegen aen, Tesseltje, Tesseltjes keeltje; daer zij gejrne van gezongen waeren, ende wenschten wel dat U.E. Joffre Francisca te hulpe meêbraght. Wat ik haer zeg, Tesseltje suft; Tesseltjen heeft pen nocht inkt om een briefken te beantwoorden; zij neemen 't niet aen, ende willen dat ik U.E. ujt den droom wekke. Op, op dan.

10 Rozemondt, hoor je speelen nocht zingen?

Wij verwachten U.E. op 't spoedighste, met U.E. dochter, ende Joffre Duart met haer' E. man: maer een briefken voor ujt, om wat gissings te moghen maeken. Ondertussen zullen wij in den windt zien, ende happen nae den geenen die van Alkmaer komt en snuffen oft hij nae U.E. adem riekt. Godt beboede U.E. op de *zejze*, en eeuwlijk in genaede, met alle die haer lief zijn; gelijk van heelen, heeten harte wenscht,

Mê Joffre

U.E.

verplichte, dienstwste

P C Hóóft.

Van den Hujze te Mujden.

j Aug. 1636.

Uitnodiging.

NB: voor details over definiëring en afbakening van categorieën verwijzen we naar de handleiding, die verder wordt toegelicht onder stap 4.

Stap 4: Ontwikkelen annotatiehandleiding

Voordat we begonnen met de controle van de annotatie door de student-assistenten hebben we een eerste versie van de annotatiehandleiding gemaakt. De eerste versie van de handleiding (van maart 2017) bevatte de uitleg over de verschillende PoS-categorieën en features, inclusief voorbeeldzinnen. We konden grotendeels de bestaande handleiding van Adelheid gebruiken. Aan het eind van de handleiding voegden we twee tabellen toe, met daarin twee volledig uitgewerkte zinnen. Aan de hand van deze eerste handleiding gingen de student-assistenten aan de slag met een brief van ca 1000 woorden. Dit deden we om te zien op welke vlakken nog moeilijkheden optraden en waar de annotaties van de student-assistenten van elkaar verschilden. Al tijdens het annoteren van deze brief, kwamen de student-assistenten met vragen. Voor sommige gevallen bood de handleiding geen uitkomst. Deze vragen stuurden zij naar Irene, die ze in overleg met Feike, Marjo en Marijn beantwoordde. Alle vragen plus antwoorden werden verzameld en via een groepsmail naar alle student-assistenten gestuurd. Hierdoor was iedereen snel op de hoogte van de nieuwste ontwikkelingen.

Gedurende de maanden waarin de studenten hun selectie brieven toegewezen hadden gekregen, hielden we deze manier van communiceren aan. Daarnaast publiceerden we vier GoogleDocs. De eerste diende als 'notitieblok' voor de studenten, waarin zij de gemaakte keuzes in het lemma verantwoorden. Het tweede diende als forum waarop de gemaakte keuzes omtrent de annotaties werden verantwoord en zo nodig bediscussieerd. Het derde was een document waarin technische aspecten werden aangekaart (zoals de locaties van de samenvattingen van Van Tricht e.a., of conversie-artefacten die per ongeluk meegekomen waren met de brieftekst). Het vierde was een document waarin de aantekeningen van vergaderingen en andere updates werden gepubliceerd.

Naast deze vormen van overleg, kwamen we ongeveer een keer per maand samen met het hele team, om moeilijkheden en de voortgang te bespreken. De aantekeningen van deze vergaderingen verschenen meteen op de GoogleDoc, zodat iedereen ze later nog kon teruglezen. In de GoogleDoc noteerden we specifieke beslissingen en antwoorden op vragen die tijdens de vergadering werden gesteld, zowel van taalkundige als sociolinguïstische aard.

In de meest recente versie van de handleiding (juli 2017) hebben we de vragen van de studenten verwerkt, plus de updates en opmerkingen uit de GoogleDocs. Dit heeft geresulteerd in een handleiding die bestaat uit twee hoofdstukken: het eerste is van toepassing op het POS-taggen, het tweede op de sociolinguïstische annotatie.

Stap 5: Handmatig corrigeren van deze PoS-tagging door student-assistenten

Aanpak:

- Een team van negen studentassistenten heeft de Adelheid-output gecontroleerd en sociolinguïstische annotaties gemaakt.
- Iedereen een eigen set brieven toegewezen, geheel random (adressanten, periode etc. door elkaar).
- Er waren werkplekken beschikbaar om samen te werken.
- Afstemming en uitwisseling via tweewekelijkse bijeenkomst en Google Docs-bestand
- Eén van de studentassistenten – Irene Kramer – was het vaste aanspreekpunt voor vragen.

Waarborging betrouwbaarheid tussen de annotatoren:

- We zijn begonnen om allemaal dezelfde testbrief te doorlopen en te zien waar de problemen/tegenstrijdigheden optraden.
- Daarna kreeg iedereen een eigen set brieven toegewezen. Per tien brieven werd er één door twee assistenten gedaan.
- Halverwege hebben we een analyse gemaakt van fouten en inconsequenties, op basis van een steekproef van drie brieven van verschillende annotatoren. Hieruit bleek dat er vrij veel verschillen tussen de annotatoren optreden (340 verschillen op 1422 woorden), maar dat de inconsequenties vaak bepaalde patronen vertonen.
    - Bijvoorbeeld: In ongeveer 20%: verschil op het terrein van werkwoorden:
        - Ongeveer de helft van de fouten komt doordat de ene annotator méér features invult dan de andere.
        - In ongeveer 20% van de gevallen: verschil van mening over of het een conjunctivus is of niet.
        - In ongeveer 20% van de gevallen: verschil van mening over of het verleden of tegenwoordige tijd is.
        - In ongeveer 10% van de gevallen: verschil van mening over of het werkwoord lexicaal is of niet.

- In ongeveer 20%: verschil op het terrein van naamvallen: nominativus of niet-nominativus (‘nonnom’)?

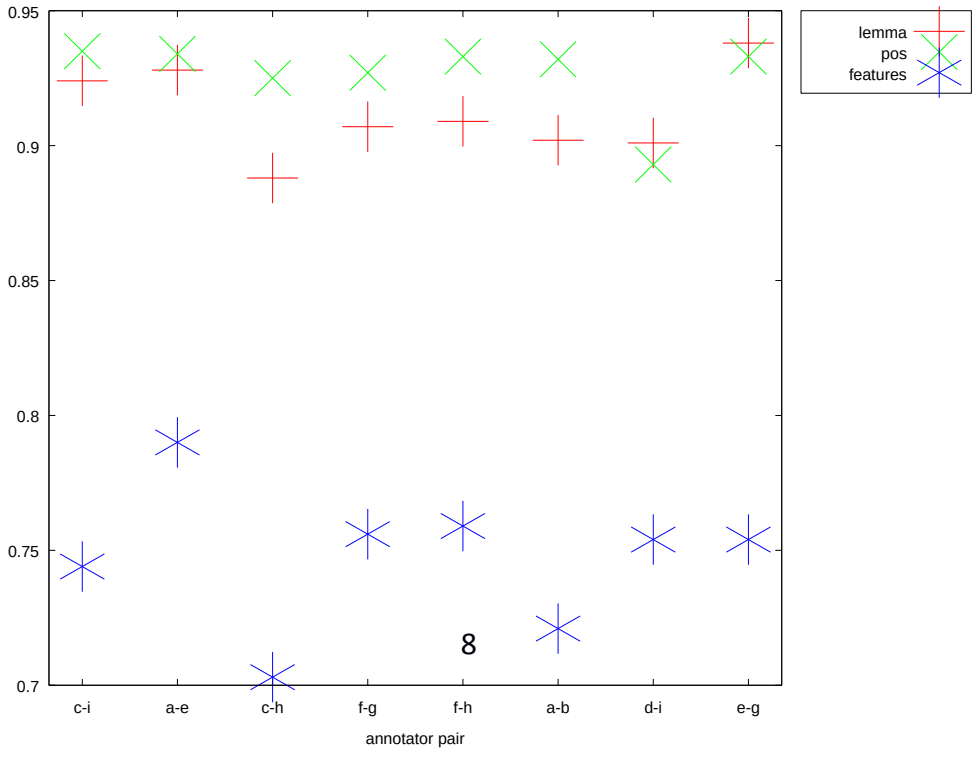
De opbrengsten van de foutenanalyse zijn gedeeld met de studentassistenten en ze hebben bijgedragen aan het verbeteren van de handleiding.

Inter-annotator agreement

Na overleg met een statisticus van de UU is besloten om van elke annotator een bepaald aantal brieven met een totale lengte van minimaal 1000 woorden ook door een andere annotator te laten taggen. Hiermee kan de mate van overeenstemming tussen annotatoren (de *inter-annotator agreement*) worden gemeten, door middel van Cohen’s Kappa (κ). Dit is op drie niveaus afzonderlijk berekend: de overeenstemming op lemma’s, POS en features. Voor de features zijn er verschillende analyses mogelijk, zoals agreement per feature-categorie of een speciale berekening voor features die niet zijn ingevuld. Hier zal in het vervolg van het onderzoek nog aandacht aan worden besteed. Voor het huidige rapport is de eenvoudige (strikte) definitie gebruikt dat alle features volledig moeten overeenkomen om als agreement meegeteld te worden. Hiermee zijn de volgende waardes berekend voor verschillende paren annotatoren:

Annotatoren	κ lemma’s	κ POS	κ features
c-i	0.924	0.935	0.744
a-e	0.928	0.934	0.790
c-h	0.888	0.925	0.703
f-g	0.907	0.927	0.756
f-h	0.909	0.933	0.759
a-b	0.902	0.932	0.721
d-i	0.901	0.893	0.754
e-g	0.938	0.933	0.754

Als grafiek:



Het gemiddelde van de agreement-waardes is 0.91 (lemma's), 0.93 (POS) en 0.75 (features). Dit wordt over het algemeen beschouwd als een hoge mate van overeenstemming. Ook is te zien dat de spreiding van de waardes tussen de verschillende annotatorparen klein is, wat een aanwijzing is dat de agreement consistent is tussen alle annotatoren.

Stap 6: Integreren van de verrijkte brieven in de Nederlab-interface

Na afloop van de annotatietaak zijn de uitvoerbestanden geconverteerd naar FoLiA. Het doel hierbij was om een paralelle annotatie op te slaan, waarbij voor elk token zowel de oorspronkelijke Nederlab-annotatie als de handmatige annotatie uit het huidige project beschikbaar zijn. Als Nederlab-annotatie is de uitvoer van Frog met de standaardinstellingen gebruikt, dit is echter eenvoudig aan te passen met een andere basis-annotatie in FoLiA-formaat indien nodig.

Het samenvoegen van de twee annotaties is niet triviaal, omdat er verschillen optreden in de tokenisatie. Daarom is er een alignment-algoritme geïmplementeerd dat de correspondenties tussen de tokens bepaalt en vervolgens de annotaties toevoegt. Dit betekent dat er enige speciale notatie is geïntroduceerd voor gesplitste of samengevoegde tokens, wat technische implicaties heeft voor de integratie van de data.

Op het moment van schrijven is Nederlab aan het bekijken of dit samengevoegde formaat geschikt is voor gebruik in de interface, en zo ja, op welke manier de nieuwe annotaties het beste kunnen worden getoond. Op korte termijn zal daarover verder overleg worden gepland.

Naast de annotaties zelf is ook de annotatie-interface en de daarbij behorende pre- en postprocessing mogelijk interessant voor integratie met Nederlab. Een mogelijk scenario is het annoteren en/of verrijken door gebruikers van een Nederlab-subcorpus, met eventueel een eigen tagset, via de interface van Nederlab. Deze integratie is echter technisch en conceptueel ingewikkeld. In overleg met Nederlab zal op een later moment besloten worden of dit haalbaar en wenselijk is.

3. Opbrengst pilotstudie

- Morfosyntactisch en sociolinguïstisch verrijkte brieven van Hooft (in totaal 108,000 woorden). Dit corpus geleverd aan Nederlab in de vorm van verrijkte Folia-bestanden en een database met meta-annotaties.

- Tool *GuSTAVE* met handleiding, beschikbaar via Git. Met deze tool kan Adelheid-output verbeterd en aangevuld worden, en sociolinguïstische informatie aan teksten worden toegevoegd.

- Enkele presentaties:
    - 2 februari 2017: Language Sciences Day 2017 (UU intern): poster met opzet annotatietaak
    - 22 mei 2017: Artikel en presentatie op NoDaLiDa 2017 (Nordic Conference on Computational Linguistics, Göteborg, Zweden). Referentie: Marijn Schraagen, Marjo van Koppen, & Feike Dietz. *Data-driven Morphology and Sociolinguistics for Early Modern Dutch*, Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, Pages 47-53. Linköping University Electronic Press, 2017.
    - 4 juli 2017: Extended abstract en presentatie op Digital Humanities Benelux Conference 2017. Titel: "Corpus enrichment for 17th century Dutch: a pilot study".

- Analyses met de geannoteerde data. Zoals in het onderzoeksplan aangegeven, werken wij toe naar een publicatie. Echter: pas als het verrijkte corpus in Nederlab is teruggeplaatst en daar via CQL doorzocht kan worden, kunnen die analyses verder vormgeven. In dit stadium hebben wij ons beperkt tot enkele grove proefboringen die inzicht geven in de functionaliteit van de data.

Proefboring 1:
Voor haar BA-eindwerkstuk onderzocht Irene handmatig 108 brieven van Hooft (Kramer 2016). Een deel was geadresseerd aan zijn goede vriend en zwager Joost Baak, en het andere deel aan zijn literaire vriendin Tesselschade Visscher. Irene onderzocht in welke taalkundige en retorische omgevingen Hooft gebruik maakte van de enkele of tweeledige negatie. Dit handmatig verrichte onderzoek turfde de verschillende soorten negaties in zes soorten zinsstructuren:

	Enkele negatie		Tweeledige negatie		Totaal
	N	%	N	%	
Bijzin	69	83%	14	17%	83
Inversie	22	76%	7	24%	29
Hoofdzin	50	74%	17	26%	67
Niet = niets	4	67%	2	33%	6
Local	80	93%	6	7%	86
V1	4	100%	0	0%	4
	229	**83%**	**46**	**17%**	**274**

Met behulp van de door Adelheid en GuSTAVE geannoteerde resultaten, konden we onderzoeken of deze handmatige uitkomsten stand zouden houden op grotere schaal. In een kort digitaal vervolgonderzoek richtten we ons op de taalkundige omgevingen. We kozen ervoor om te zoeken naar bijzinnen en naar V1-zinnen (zinnen met de persoonsvorm op de eerste positie). Dit deden we,

omdat deze twee categorieën relatief makkelijk op te sporen waren op korte termijn. In de tabel hieronder is te zien wat onze digitale zoektocht opleverde (in zwart), naast de handmatige resultaten (in grijs):

	single	%	bipartite	%	total
subclause	237	81	75	19	312
subclause	69	83	14	17	83
V1	34	100	0	0	34
V1	4	100	0	0	4

De digitaal verkregen resultaten ondersteunen de uitkomsten van het handmatig verrichte onderzoek. Wanneer we werkbare zoekcriteria hebben ontwikkeld voor de verschillende zinstypen, kunnen we verder onderzoek doen naar de omgevingen waarin enkele en tweeledige negatie voorkomen.

Proefboring 2:

Eerste proefboringen zetten ons bijvoorbeeld op de volgende sporen voor vervolganalyse:

We tagden in totaal:

Zinnen zonder negatie	5147
Zinnen met eenledige negatie	507
Zinnen met tweeledige negatie	143
Zinnen met alleen een negatieclitic	83

Dat betekent dat nog steeds een substantieel deel van de ontkennende zinnen een tweeledige negatie bevatten. Er zijn zelfs zinnen (ongeveer 1,5 %: meer dan gedacht) die enkel een negatief clitic hebben: *Ik en ken Piet*. Die clitic-zinnen vragen om verdere analyse: in welke gevallen wordt alleen een clitic gebruikt?

Sociolinguïstische omstandigheden stimuleerden mogelijk de keuze voor een bepaald type negatie. In de 108 brieven die als 'informerend' werden getagd, zitten 60 tweeledige negaties. In de 20 brieven die als 'overtuigend' werd aangemerkt, vonden we 26 tweeledige negaties. Deze resultaten geven aanleiding om het effect van de sociolinguïstische context op de aan- of afwezigheid van tweeledige negatie verder te onderzoeken.

Eerste proefboringen wijzen ook uit dat de rol van de adressant verdere analyse behoeft. In de 4 brieven aan raadsheren vonden we 3 gevallen van tweeledige negatie, maar in de 8 brieven aan koopmannen maar liefst 25.

4. Vervolgstappen

Vervolgstappen Language Dynamics & Nederlab:

- Nederlab gaat de verrijkte Folia integreren in hun interface, waarna we binnen de Nederlabomgeving zoekopdrachten kunnen uitvoeren. We moeten nog afwachten hoeveel mogelijkheden dit voor ons gaat opleveren: kunnen wij ook goed zoeken op sociolinguïstische tags in Nederlab?
  (Wij houden voor onszelf de optie open dat we een eigen interface moeten ontwikkelen met opties die zijn toegespitst op onze behoeftes. Hierover loopt overleg met de opleiding Informatica, die wellicht stagairs kan bieden om ons hiermee te helpen.)
- We gaan onderzoeken of GuSTAVE ook in Nederlab opgenomen kan worden. Op korte termijn zal een proef worden uitgevoerd, waarna een inschatting kan worden gemaakt: is integratie mogelijk, wat komt erbij kijken aan tijd en werk? Voor Nederlab is het belangrijk dat de tool zo breed mogelijk inzetbaar is (denk aan: niet alleen te gebruiken voor door Adelheid getagde teksten, en mogelijkheid tot aanpassing tagset). Mogelijk kan deze ambitie resulteren in een vervolgproject waarin Nederlab en het Language Dynamics-project samenwerken.

Vervolgstappen Language Dynamics:

- Verder werken aan de inhoudelijke analyses en het publiceren daarvan.
- Geannoteerde data inzetten als trainingsdata voor vroegmoderne tagging.

5. Leerpunten & aanbevelingen voor Nederlab

Moeilijkheden waarmee we geconfronteerd werden:

- De startperiode was tijdsintensiever dan gehoopt: student-assistenten hadden tijd nodig het werk in de vingers te krijgen en er traden onvoorziene technische problemen op met de tool-in-ontwikkeling. Gaandeweg ging het werk steeds sneller. Uiteindelijk hebben we grofweg de helft van de woorden die Hooft in de periode 1600-1638 schreef getagd (108.000 van de ongeveer 150.000 woorden). Ook hebben we ongeveer de helft van de brieven uit die periode van sociolinguïstische tagging voorzien.
- Het project vond plaats op basis van matching: Nederlab betaalde onze studentassistenten, maar niet de uren die besteed moesten worden aan organisatie en coördinatie. Hiermee moet in het vervolg beter rekening gehouden worden.
- Het bleek moeilijk om alle annotatoren op één lijn te krijgen. Op hoofdlijnen gaat het meestal goed (alle studenten zien of iets een WW is), maar op detailniveau blijven er veel consequenties optreden (niet alle studenten vullen bij zo'n WW dezelfde features in).

Aanbevelingen voor Nederlab:

- We willen Nederlab graag vragen om de mogelijkheid te ontwikkelen om sociolinguïstische tags in zoekopdrachten te betrekken.
- Voor het vervolg van onze analyses is het ook belangrijk dat we het subcorpus van de verrijkte brieven binnen Nederlab als afgebakend deelcorpus kunnen doorzoeken.
- Mogelijk kan (een doorontwikkelde versie van) GuSTAVE binnen Nederlab gaan functioneren.

6. Literatuur

- Jespersen, O., *Negation in English and other languages*. Copenhagen, 1917.
- Kramer, I. Variatie in Negatie, een syntactisch en retorische analyse van het gebruik van enkele en tweeledige negatie in de brieven van P.C. Hooft van 1633 tot 1638 aan Joost Baek en Tesselschade Roemersdochter Visser. BA-eindwerkstuk, Utrecht, 2016.
- Nobels J. and G. Rutten, 'Language norms and language use in seventeenth-century Dutch: Negation and the genitive'. In: G. Rutten, R. Vosters and W. Vandenbussche (eds), *Norms and usage in language history, 1600-1900. A sociolinguistic and comparative perspective* Advances in Historical Sociolinguistics no. 3. Amsterdam/Philadelphia: John Benjamins Publishing Company 2014, p. 21-48.
- Nobels, J. *(Extra)Ordinary Letters. A View from below on 17th-century Dutch*. PhD dissertation, Leiden, 2013.
- Paardekooper, P., 'Bloei en ondergang van onbeperkt ne/en, vooral dat bij niet-woorden', *Neerlandistiek.nl* 2006.
- Rem, M. and H. van Halteren, *Tagging and Lemmatization Manual for the corpus van Reenen-Mulder and the Adelheid 1.0 Tagger-Lemmatizer*. Radboud University Nijmegen (2011).
- Sijs, N. van der & R. Willemijns, *Het verhaal van het Nederlands*. Amsterdam, 2009.
- Sijs, N. van der, *Taal als mensenwerk. Het ontstaan van het ABN.* Den Haag, 2004.
- Tricht, H.W. van, ed., *De briefwisseling van P.C. Hooft*. Culemborg 1977.
- Vosters, R. and W. Vandenbussche, 'Bipartite negation in 18th and early 19th century Southern Dutch. Sociolinguistic aspects of norms and variation'. In: *Neuphilologische Mitteilungen* 3 (2012), p. 343-364.
- Wouden, T. van der, 'Meer over dubbele ontkenningen. Reactie op Piet Paardekooper'. *Neerlandistiek.nl*, 2007.
- Zeijlstra, H., *Sentential negation and negative concord.* PhD dissertation, Utrecht, 2004.