Application form

Investment Subsidy NWO Large 2011-2012

The completed application form including attachments should be submitted electronically via the Iris system (personal account of the applicant). Please ensure that the form and attachments are saved in PDF format. Iris is accessible via the NWO website: www.iris.nwo.nl. Please use the Verdana font, size 8.5, line spacing 13 and keep to a **maximum of 30 pages A4 (incl. attachments)**.

Please read the brochure Investment Subsidy NWO Large 2011-2012 before completing this form.
The closing date for the submission of applications is 1 November 2011, 17.00 hrs. The date and time on which you upload via Iris is the valid submission date and time.

Algemene informatie

Aanvragend instituut

University/institute	KNAW
Faculty	Meertens Instituut

Projectleider/coördinator

Title(s)	prof.dr.
First name	Hans
Initials	H.J.
Surname	Bennis ☒male ☐female
Address for correspondence	Joan Muyskenweg 25, 1096 CJ Amsterdam
Telephone number	020-4628523
Fax	020-4628555
Email	Hans.Bennis@Meertens.knaw.nl
Website (optional)	http://www.meertens.knaw.nl/cms/nl/medewerkers/142447-hansb
Preference for correspondence	☐English ☒Dutch

Titel van het project

Nederlab, een laboratorium voor onderzoek naar de veranderingspatronen in de Nederlandse taal en cultuur

Samenvatting

Samenvatting van het project

Het doel van dit project is alle gedigitaliseerde teksten die relevant zijn voor ons nationaal erfgoed, de geschiedenis van de Nederlandse taal en cultuur (ca. 800 – heden), bijeen te brengen in een gebruiksvriendelijke, algemeen toegankelijke en met tools verrijkte gebruikersomgeving. Onderzoekers kunnen vanuit deze gebruikersomgeving uit teksten die de volledige geschreven geschiedenis van de Lage Landen omvatten, gegevens zoeken en analyseren.

Taal en cultuur zijn dynamische verschijnselen. Geesteswetenschappers - taalkundigen, letterkundigen, historici – trachten de veranderingsprocessen en variatie binnen taal en cultuur te begrijpen, en de interne en externe factoren die hieraan ten grondslag liggen te achterhalen. Om de relevante onderzoeksvragen te kunnen beantwoorden, zijn grote hoeveelheden gegevens nodig. Tot nu toe vond het onderzoek noodzakelijkerwijs plaats in de vorm van detailstudies over relatief korte tijdsperiodes en beperkt tot een enkele discipline. Naarmate er meer Nederlandstalige historische teksten digitaal beschikbaar komen, komen er nieuwe, longitudinale onderzoeksvragen in het vizier, en doemt de mogelijkheid op voor systematisch onderzoek naar de interactie tussen veranderingen in de cultuur, maatschappij, letteren en taal. De hypothese is dat veranderingen in de taal en cultuur - beide uitingen van de menselijke cognitie - aan elkaar zijn gerelateerd en dat er identieke of vergelijkbare wetmatigheden aan ten grondslag liggen. Onderzoek naar deze wetmatigheden leveren nieuwe interdisciplinaire inzichten in fundamentele theoretische problemen zoals het nature-nurturedebat, het ontstaan van canons en een nationale culturele identiteit, culturele integratie, en de verbreiding van kennis, cultuur en taal.

Vooralsnog echter is dergelijk langetermijnonderzoek om praktische redenen niet mogelijk. De beschikbare diachrone corpora worden op verschillende plaatsen en met verschillende metadata aangeboden en zijn niet gezamenlijk te doorzoeken. De ontwikkelde tools zijn vaak alleen bruikbaar voor een specifiek gebruiksdoel en niet generiek. De mogelijkheden van de beschikbare technologische middelen, die berusten op het toepassen van statistische analysemethodes en wetenschappelijke waarden als verifieerbaarheid en herhaalbaarheid, worden lang niet ten volle benut.

Om aan deze bezwaren een einde te maken dient een groep experts op het gebied van diachrone onderzoeksvragen, corpus- en infrastructuurbouw, zoektechnologieën en toolsontwikkeling gezamenlijk een aanvraag in voor de oprichting van Nederlab. Nederlab biedt een gebruikersvriendelijke webinterface van waaruit het mogelijk wordt de bestaande diachrone corpora gedistribueerd te doorzoeken, zowel op tekstniveau als op metadata-niveau. Dit zal niet alleen leiden tot een betere toegankelijkheid van de bestaande corpora, maar ook tot kwaliteitsverbetering en standaardisering van de data en metadata.

Het project bouwt voort op verschillende initiatieven: voor corpora werkt Nederlab samen met de wetenschappelijke bibliotheken en instellingen, voor infrastructuur met CLARIN, DARIAH (en mogelijk CLARIAH), voor tools met eHumanities-programma's als Catch en IMPACT. Nederlab voegt hieraan belangrijke meerwaarde toe: een gebruikersvriendelijke infrastructuur voor onderzoekers die automatisch leidt tot samenwerking en synergie binnen de geesteswetenschappen en tot het stellen van nieuwe, veelal interdisciplinaire, onderzoeksvragen. Veel aandacht zal tijdens de projectfase worden besteed aan de disseminatie van informatie; de infrastructuur zal in nauw overleg met de onderzoeksgemeenschap worden ingericht, en getest worden in concreet pilotonderzoek.

Samenvatting van het voorstel voor het algemene publiek

Het doel van Nederlab is onderzoekers de mogelijkheid te bieden antwoorden te vinden op nieuwe, longitudinale onderzoeksvragen. Nederlab wil hiervoor een gebruikersvriendelijke webinterface inrichten van waaruit geesteswetenschappers de digitale historische teksten, die beschikbaar gesteld worden door de wetenschappelijke bibliotheken en instellingen, tegelijkertijd kunnen doorzoeken en met tools kunnen analyseren, zowel op tekstniveau als op metadata-niveau.

Trefwoorden

eHumanities, diachronic, corpora, tools , workspace

Top 10 van relevante publicaties

Project summary

- Barbiers, S., H.J. Bennis *et al*. (2005-2008). *Syntactic Atlas of the Dutch Dialects, Volume I, II*. Amsterdam: Amsterdam University Press.
- Coupé, Griet & Ans van Kemenade (2009). 'Grammaticalization of modals in English and Dutch: uncontingent change'. In: P. Crisma and G. Longobardi (eds.) *Historical Syntax and Linguistic Theory*. Oxford: Oxford University Press, 250-270.
- Daelemans, W. & A. van den Bosch (2005). *Memory-based language processing*. Cambridge, UK: Cambridge University Press.
- Hunt, L., M.C. Jacob & W. Mijnhardt (2010). *The Book That Changed Europe*. Cambridge Mass., Harvard University Press/Belknap Press.
- Kennedy, James (2008). 'Religion, Nation and European Representations of the Past'. In: Stefan Berger and Chris Lorenz (eds.), *The Contested Nation: Ethnicity, Class, Religion and Gender in National Histories*. Basingstoke: Palgrave Macmillan, 104-134.
- Leerssen, J.T. (2010). 'Viral Nationalism: Romantic intellectuals on the move in 19th-century Europe'. In: *Nations and Nationalism, 17*(2), 257-271.
- Nerbonne, John (2010). 'Measuring the Diffusion of Linguistic Change'. In: *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365, 3821-3828. DOI: 10.1098/rstb.2010.0048.
- Sijs, Nicoline van der (2001). *Etymologie in het digitale tijdperk. Een chronologisch woordenboek als praktijkvoorbeeld*. Doctoral dissertation Leiden.
- Steen, Gerard J., Aletta G. Dorst & J. Berenike Herrmann (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Amsterdam/Philadelphia: John Benjamins.
- Stipriaan, R. van (2007). 'Words at War: the Early Years of William of Orange's Propaganda'. In: *Journal of Early Modern History*, 11, 331-349.

Investeringsvoorstel

Onderzoeksveld en onderzoeksplannen

1. De doelstellingen[1]

Er zijn Nederlandstalige teksten bewaard gebleven van de achtste eeuw tot heden, en vele daarvan – zowel fictie als nonfictie - zijn inmiddels gedigitaliseerd. In die teksten vinden we de neerslag van de Nederlandse taal en cultuur. Zoals bekend zijn zowel de taal als de cultuur voortdurend in verandering. Daarbij wisselen perioden van vertraging en versnelling elkaar af. Geesteswetenschappers van verschillende disciplines - historici, letterkundigen, taalkundigen - trachten de historische veranderingen binnen hun werkterrein te beschrijven, het tempo waarin die plaatsvinden en de factoren die eraan ten grondslag liggen. Veel is echter nog onbekend, en er heeft nog nauwelijks systematisch onderzoek plaatsgevonden naar de interactie tussen veranderingen in de cultuur, maatschappij, letteren en taal. Toch bestaat er evidentie dat veranderingen in de verschillende domeinen op elkaar inwerken: zo gaat men er binnen de taalkunde van uit dat veel taalveranderingen het gevolg zijn van taalcontact als gevolg van migratie en immigratie, dus veranderingen in de maatschappij. Historici hebben aangetoond dat taal een sturende en vormende factor in politieke en sociale veranderingen kan zijn. En literaire werken vormen een weerspiegeling van de contemporaine cultuur.

Nederlab biedt een gebruikersvriendelijke webinterface van waaruit het mogelijk wordt de digitale historische teksten, die beschikbaar gesteld worden door de wetenschappelijke bibliotheken en instellingen, tegelijkertijd te doorzoeken. Met behulp van Nederlab kunnen geesteswetenschappers voor het eerst systematisch veranderingen in de taal en cultuur en hun onderlinge relatie onderzoeken op basis van een corpus van Nederlandse teksten van de oudste tijd tot heden. De hypothese is dat veranderingen in de taal en cultuur - beide uitingen van de menselijke cognitie - aan elkaar zijn gerelateerd en dat er identieke of vergelijkbare wetmatigheden aan ten grondslag liggen. Op basis van Nederlab willen we deze wetmatigheden blootleggen. Uit onderzoek naar de wetmatigheden zal blijken welke onderdelen van de Nederlandse taal en cultuur onderhevig zijn aan veranderingen en welke constant blijven. De veranderingen zullen deels verklaard moeten worden als autonome ontwikkelingen en deels als gevolg van de invloed van een ander (binnenlands of buitenlands) verschijnsel. In dat geval is er sprake van culturele overdracht en culturele integratie. Constanten kunnen behoren tot de inherente eigenschappen van taal en cultuur; nadere kennis hierover zal een bijdrage kunnen leveren aan het nature-nurturedebat. Constanten kunnen echter ook wijzen op een traditie, en leiden tot het ontstaan van canons, ofwel een nationale culturele identiteit.

Onderzoek naar de patronen van veranderingen is niet nieuw: vooraanstaande onderzoekers zoals, om slechts enkelen te noemen, R. Aerts, R. Bod, A. van Kemenade, J. Kennedy, J. Leerssen, W. Mijnhardt, T. Vaessens, hebben hun expertise op allerlei deelterreinen getoond in baanbrekende publicaties. Alle erkende experts hebben een sleutelpositie gekregen binnen Nederlab als supervisor of in een van de vier adviesraden (zie de organisatiestructuur in 'Relatie tot andere onderzoeksgroepen'), en zij zullen binnen Nederlab een voortrekkersrol gaan vervullen, in technische zin gesteund door prominente onderzoekers als A. van den Bosch, F. de Jong, J. Nerbonne, N. Oostdijk, L. Schomaker, door specialisten in zoektechnologieën als J. Kamps en A.P. de Vries, en door ervaren corpusbouwers als J. Hoeksema, G.J. Postma, P. van Reenen, M. Rem, N. van der Sijs, R. van Stipriaan en INL-medewerkers.

Al het bestaande onderzoek is, door het ontbreken van een omvangrijk diachroon corpus, beperkt tot detailstudies over relatief korte tijdsperiodes. Nederlab zal het mogelijk maken wetmatigheden van veranderingen binnen de Nederlandse taal en cultuur te onderkennen, waardoor het onderzoek een grote sprong voorwaarts zal maken. Nederlab biedt een onderzoeksobject en gebruikersvriendelijke tools om dit object mee te doorzoeken. Het

[1] Het ontwikkelen en schrijven van deze aanvraag is mede mogelijk gemaakt door financiële ondersteuning van het Meertens Instituut, INL en de universiteiten van Nijmegen en Amsterdam.

onderzoeksobject bestaat uit een diachroon corpus teksten van de achtste eeuw tot heden; die teksten zijn her en der al digitaal beschikbaar, maar worden binnen Nederlab voor het eerst gezamenlijk gedistribueerd doorzoekbaar gemaakt. Een belangrijk kenmerk van de Nederlandse taal en cultuur is de veelvormigheid ervan: zowel binnen de taal als de cultuur bestond en bestaat veel variatie. Juist in deze variatie zal de verklaring voor een deel van veranderingen moeten worden gezocht. Om deze variatie bij onderzoeksvragen te kunnen betrekken, wordt aan iedere tekst binnen het corpus informatie toegevoegd over vier variabelen: tijd, plaats, auteur(sgegevens) en tekstsoort. De variabelen 'tijd' en 'plaats' geven het corpus een diachrone en geografische dimensie. Met de vier variabelen kan zowel de input als de output van een onderzoeksvraag op diverse manieren worden gesorteerd. Dit maakt het bijvoorbeeld mogelijk de verbreiding van een verandering in de loop van de tijd te volgen binnen de dialecten of binnen bepaalde categorieën literaire teksten.

Doordat het diachrone corpus van Nederlab meer dan tien eeuwen omspant, kunnen er longitudinale onderzoeksvragen worden gesteld: zo kan de visie op vreemdelingen door de eeuwen heen worden onderzocht, het ontstaan, de verbreiding en het verdwijnen van literaire genres, de veranderende oordelen over literaire schrijvers, geleerden of historische gebeurtenissen, en de ontwikkeling van het gebruik van voorzetsels of voegwoorden (zie verder de gebruiksgevallen).

Voor de beantwoording van dergelijke longitudinale onderzoeksvragen levert Nederlab een gebruikersvriendelijke webinterface met een scala aan computertechnieken en aanwijzingen voor niet-technische onderzoekers over hoe die toe te passen. Onderzoekers kunnen met behulp van de tools het corpus of delen daarvan gericht bevragen. De computertechnieken kunnen ook gebruikt worden voor het automatisch analyseren van grote hoeveelheden tekstbestanden. Deze laatste methode zal zonder enige twijfel leiden tot onverwachte vondsten en tot het ontdekken van patronen in de gegevens die niet aan het licht komen bij handmatig onderzoek en die nadere verklaring door de onderzoekers behoeven. Ook zal nader licht geworpen worden op de Constant Rate Hypothesis van taalveranderingen, die stelt dat taalveranderingen zich in verwante contexten in hetzelfde tempo verbreiden, en op de relatie tussen taalveranderingen en sociale veranderingen.

Het vinden van wetmatigheden dient noodzakelijkerwijs te gebeuren door grote fundamentele onderzoeksvragen te vertalen in kleinere, concrete onderzoeksvragen waarvan de antwoorden dienen als bouwstenen voor inzicht in het grotere geheel. Juist hiervoor is Nederlab uiterst geschikt: enerzijds doordat hier alle benodigde tools bijeengebracht worden, en anderzijds doordat onderzoekers binnen Nederlab successievelijk maar ook tegelijkertijd kunnen werken aan deelvragen die bijdragen aan een grotere onderzoeksvraag, en zo voort kunnen bouwen op elkaars onderzoek. Zowel de door een onderzoeker verrijkte data als de door hem gebruikte methode, in de vorm van aangepaste tools, blijven namelijk binnen Nederlab bewaard, zodat volgende onderzoekers er gebruik van kunnen maken. Geesteswetenschappers kunnen hierdoor voortbouwen op de resultaten die geboekt zijn door collega's; door taalkundigen ontwikkelde automatische namenherkenning kan bijvoorbeeld door historici verder uitgebouwd worden door het toevoegen van concrete biografische gegevens aan de namen, terwijl letterkundigen auteursnamen kunnen linken aan bibliografische data. Zo leidt Nederlab automatisch tot een versnelde integratie van uiteenlopend geesteswetenschappelijk onderzoek.

Het binnen Nederlab geboden instrumentarium maakt verifieerbaar, betrouwbaar, geobjectiveerd, herhaalbaar en statistisch onderbouwd onderzoek binnen de geesteswetenschappen mogelijk. Voor veel disciplines in de geesteswetenschappen is dit een methodologische vernieuwing die een sterke empirische basis kan verschaffen aan het onderzoek: het casuïstische, subjectieve en kwalitatieve onderzoek wordt kwantitatief toetsbaar. De verwachting is dat de toepassing van deze methodes kan leiden tot veel nieuwe inzichten in de dynamiek van veranderingen in de Nederlandse taal en cultuur. De onderzoeksaanpak wordt hieronder toegelicht aan de hand van gebruiksgevallen.

2. Gebruiksgevallen

Nederlab biedt het instrumentarium om feitelijk onderbouwde antwoorden te vinden op vragen die geesteswetenschappers aan een diachroon corpus stellen, om bestaande hypothesen te toetsen aan een omvangrijk corpus en om door verkennende data-analyse nieuwe hypotheses op te stellen. Het instrumentarium kan een in principe eindeloze reeks specifieke onderzoeksvragen beantwoorden: ter voorbereiding van deze subsidieaanvraag

hebben wij circa 150 onderzoekers uit de geesteswetenschappen, die zich bezighouden met diachroon onderzoek en/of eHumanities (zie Appendix 1), mondeling of schriftelijk geconsulteerd over de vraag welke onderzoeksvragen ze op basis van Nederlab beantwoord willen zien. De veelheid van door hen genoemde onderzoeksvragen kan op basis van vraagstelling en methodiek worden verdeeld in vragen rond het opsporen van vernieuwingen ofwel het begin van een verandering (2.1), het vaststellen van de verbreiding van veranderingen (2.2), het leggen van (statistische) verbanden en netwerken (2.3), en het detecteren van overeenkomsten en verschillen tussen teksten (2.4). De antwoorden op deze vragen zullen leiden tot grote, samenhangende onderzoeksthema's of 'scholarly narratives'.

Om recht te doen aan de variatie binnen het corpus worden aan iedere afzonderlijke digitale tekst in ieder geval de volgende vier metadata toegevoegd:

- **tijd**: het jaar van schrijven of drukken van een tekst, en een link naar tijdlijn/periodisering;
- **plaats**: plaats van herkomst van een tekst, en een link naar geografische namenlijst en GPS;
- **auteursgegevens**: naam, en een link naar biografische gegevens;
- **tekstsoort**: fictie, nonfictie, kranten, egodocumenten, etc, en een link naar bibliografische gegevens en naar scans van het gedrukte of geschreven origineel.

Daarnaast worden metadata toegevoegd over de kwaliteit van de teksten, zodat onderzoekers de betrouwbaarheid van de gevonden resultaten kunnen wegen:

- **kwaliteit**: a. gecorrigeerde tekst (= diplomatische transcriptie of gecorrigeerde ocr), b. kritische transcriptie of omgespelde tekst, c. herkend door ocr (dus met leesfouten).

Tot slot zal een deel van het corpus op woordniveau inhoudelijk verrijkt worden, waardoor gestructureerd zoeken mogelijk wordt. Ook dit wordt met metadata aangegeven, waarbij meerdere opties tegelijk mogelijk zijn:

- **inhoudelijk verrijkt**: door onderzoeker geannoteerd, (semi-)automatisch gelemmatiseerd/POS, morfologisch geanalyseerd, syntactisch geparst, tijdsaanduidingen getagd, plaatsnamen getagd en voorzien van georeferentie, persoonsnamen getagd, informatie over taalgebruik (dialect, sociolect, Amerikaans-Nederlands, Belgisch- Nederlands e.d.). De inhoudelijke tagging van tijdsaanduidingen, plaatsnamen, persoonsnamen en informatie over taalgebruik corresponderen met de vier corpus-metadata tijd, plaats, auteur en tekstsoort. De inhoudelijke tagging wordt automatisch opgeslagen via bijvoorbeeld *FoLiA* (Format for Linguistic Annotation), waarbij de oorspronkelijke weergave van de tekst ongewijzigd blijft, maar iedere onderzoeker een eigen laag van verrijking of annotatie kan toevoegen.

2.1. Onderzoeksvragen m.b.t. het opsporen van het begin van vernieuwing in de taal en cultuur

Voor veel onderzoek naar de geschiedenis van de Nederlandse taal en cultuur is het belangrijk het verloop van veranderingsprocessen vast te leggen. Dit begint met de vaststelling wanneer en waar een bepaalde vernieuwing voor het eerst is opgetreden (2.1); daarna wordt de verbreiding van de vernieuwing onderzocht in tekst, tijd en ruimte (2.2).

2.1.1 Het opsporen van nieuwe begrippen

Een historicus wil weten hoe snel technische vernieuwingen (*windmolens*, *elektriciteit, gaslicht, telefoon, fotografie*) vanuit het buitenland de Lage Landen hebben bereikt. Die gegevens kunnen vergeleken worden met informatie over de snelheid waarmee woorden voor gedachtegoed zoals *democratie, marxisme, socialisme* of buitenlandse literaire stromingen zoals *romantiek* of *detectives* geïntroduceerd werden in de Lage Landen. Uit de gegevens kan naar voren komen dat een bepaald type buitenlandse innovaties (bijvoorbeeld concrete uitvindingen) sneller of juist langzamer worden geïntroduceerd dan een ander type (bijvoorbeeld gedachtegoed), en ook kan blijken of in een bepaalde periode sneller en meer buitenlandse innovaties in de Nederlandse maatschappij zijn doorgedrongen dan in een andere periode. Dit zegt iets over de openheid of geslotenheid van de Nederlandse maatschappij en cultuur door de tijd heen. Taalkundigen zijn daarnaast geïnteresseerd in de naamgeving van de nieuwe concepten, terwijl letterkundigen naar beeldvormingsaspecten van nieuwe concepten kijken.

Voor het opsporen van nieuwe begrippen levert Nederlab materiaal en gereedschap dat tot nu toe niet beschikbaar is. Momenteel moet een onderzoeker verschillende corpora raadplegen die zich op verschillende websites bevinden en verschillende interfaces hebben. Binnen Nederlab krijgt de onderzoeker een compleet

diachroon corpus gedistribueerd onder handbereik. Daarbij verschaft Nederlab *automatische indexeringsprogramma's* van alle woordvormen uit het complete corpus, die men op verschillende manieren kan bekijken: als woordenlijst of in context (KWIC, Key Word In Context) via *concordantieprogramma's*. Omdat iedere tekst voorzien is van metadata, kan men alle in het corpus voorkomende woordvormen chronologisch sorteren en zo voor iedere woordvorm eenvoudig de oudste vindplaats vaststellen. Via de metadata plaats, auteur en tekstsoort kan men vervolgens vaststellen in welke plaats, bij welke auteur en in wat voor soort tekst een woordvorm voor het eerst is opgetreden.

Deze automatische service betreft álle woordvormen in het complete corpus. Onderzoekers kunnen op basis van de kwaliteitsmetadata de gevonden woordvormen uit gecorrigeerde teksten vergelijken met die uit teksten die zijn herkend met ocr. Wanneer de gegevens opvallend afwijken (bijvoorbeeld wanneer het begrip *televisie* volgens ocr-teksten dateert uit 1886) dan is dit een reden de originele bron te raadplegen (waaruit blijkt dat *televisie* in 1886 een ocr-leesfout is voor *ter visie*).

De automatische indexering van woordvormen legt geen link tussen spelling- en vormvarianten van één en hetzelfde woord. Voor veel onderzoeksvragen wil men echter van dergelijke variatie kunnen abstraheren. Ook daarvoor biedt Nederlab het instrumentarium:

1. *Tools voor tokenization*: de inputtekst (een sequentie van karakters) wordt omgezet in een sequentie van tokens (= voorkomens van woordvormen). Tokenization gebeurt op basis van spaties, leestekens etc.; een gedeelte van het corpus is al getokenized.
2. *Tools voor spellingnormalisatie*: deze mappen spellingvarianten, spelfouten, ocr-fouten etc. op een canonieke spellingvorm zoals een woordenboekingang; voorbeelden van dergelijke tools zijn het door de Universiteit Tilburg ontwikkelde TICCLOPS, tools die binnen het Europese project IMPACT beschikbaar komen, en tools die varianten van persoonsnamen aan elkaar koppelen (ontwikkeld binnen het CATCH-project LINK).
3. *PoS-tagging*: toekenning van een Part of Speech-code aan een voorkomen van een woordvorm (= token) in context; een gedeelte van het corpus heeft dat al.
4. *Lemmatisering*: toekenning van een lemma (= canonieke vorm voor een buigingsvorm) aan een voorkomen van een woordvorm; een gedeelte van het corpus heeft dat al.
5. Bij al deze bewerkingen speelt een *computationeel historisch lexicon*, zoals ontwikkeld bij INL voor teksten uit de 16de tot de 20ste eeuw, een grote rol. Dit lexicon herleidt automatisch bij elkaar behorende woorden tot woordenboekingangen.
6. Vervolgens kunnen *indexen* gemaakt worden op basis van:

- de set van woordvorm-types
- de set van lemma-types
- de set van types van de woordvorm in canonieke spelling.

Met behulp van deze indexen kan men de precision en recall verbeteren.

2.1.2 Het opsporen van nieuwe woordvormen

Taalkundigen willen niet alleen nieuwe woorden en begrippen opsporen maar ook nieuwe woordvormen; zij stellen vragen als: sinds wanneer en in welk dialect is het deelwoord met *ge-* voor het eerst gesignaleerd en van welk type werkwoorden zijn deelwoorden op *ge-* het eerst gevormd?

Voor de historische morfologie zal Nederlab een grote vernieuwing betekenen: er bestaan momenteel slechts weinig historische studies naar de Nederlandse woordvorming, en dat komt voornamelijk doordat het materiaal – data en tools – hiervoor ontbreekt. Binnen Nederlab kan voor het eerst systematisch onderzoek plaatsvinden naar de manier waarop samenstellingen en afleidingen in de loop van de tijd binnen het Nederlands zijn gevormd.

Als voorbeeld kan het onderzoek naar afleidingen dienen. Afleidingen zijn woorden die een affix bevatten, een element dat niet zelfstandig voorkomt, zoals *-te* in *diepte* en *-ig* in *groenig*. Het aantal woorden dat met een bepaald affix kan worden afgeleid is in principe onbeperkt (hoewel aan bepaalde restricties gebonden), maar het aantal affixen is eindig en kan voor de verschillende periodes van het Nederlands worden vastgesteld.

Onderzoekers zoeken antwoorden op vragen als: Sinds wanneer en waar komen de verschillende typen afleidingen binnen het Nederlands voor? Waarom zijn sommige typen afleidingen (*diepe* 'diepte') verdwenen en

vervangen door andere (*diepte*)? Wat waren de combinatorische restricties van de verschillende affixen in de loop van de tijd? Wat is de verklaring voor het ontstaan van nieuwe typen afleidingen of nieuwe combinatorische restricties? Hoe verhouden synonieme affixen, zoals vrouwelijke persoonsaanduidingen op *-in* (*boerin*) en *–esse* (*secretaresse*), zich tot elkaar?

Nederlab biedt de hulpmiddelen die een onderzoeker nodig heeft om alle afleidingen die in de loop van de tijd met een bepaald affix zijn gevormd, te inventariseren en analyseren:
1. Er komt een uitputtende *affixenlijst* van het Nederlands beschikbaar uit alle tijden. Momenteel bestaat een dergelijke lijst slechts van het moderne Nederlands.
2. Er worden *geanalyseerde datasets* beschikbaar gesteld, woorden met een morfologische analyse (die gebruikt zijn of worden als trainingsmateriaal voor de ontwikkeling van morfologische parsers).
3. Een deel van het corpus is (semi-)automatisch *morfologisch getagt*.
4. Er komen verschillende *morfologische parsers* beschikbaar voor het moderne Nederlands en de 18e-19e eeuw. Voor het oudere materiaal moeten parsers worden ontwikkeld: daarvoor kan een taalonafhankelijke, lerende parser worden gebruikt, of bestaande tools voor middeleeuws materiaal kunnen worden uitgebreid met een morfologische module, zoals Adelheid (ontwikkeld bij RU/MI voor flectie en spellingvariatie), en INPOLDER (voor syntactische parsing).

2.1.3 Het opsporen van nieuwe patronen

Veel geesteswetenschappelijke onderzoeksvragen betreffen niet zozeer het opsporen van nieuwe begrippen of woordvormen alswel van nieuwe woordcombinaties, nieuwe patronen. Letterkundigen zijn bijvoorbeeld geïnteresseerd in de vraag welke nieuwe woordcombinaties zijn geïntroduceerd door de Tachtigers, of welke nieuwe rijmwoorden door bepaalde dichters zijn ingevoerd. Voor taalkundigen is het opduiken van nieuwe patronen een manier om betekenisveranderingen van woorden te herkennen. Zo werd in het oudste Nederlands *varen* vooral gecombineerd met woorden als *wagen* of *paard*; pas in de loop van de middeleeuwen komt het voor met namen voor vaartuigen. Hieruit blijkt dat de betekenis van het woord *varen* is veranderd van ‘gaan, reizen’ in het algemeen naar ‘reizen te water’. Ook kunnen taalkundigen aan de hand van woordpatronen grammaticale verschuivingen vinden zoals de verandering van *wijzigen* van overgankelijke werkwoord (*ik wijzig een nota*) in onovergankelijk werkwoord (*een nota wijzigt*).

Nederlab biedt tekstanalyserende tools aan (zie 2.4) die vaststellen dat bepaalde woorden vanaf een bepaald moment regelmatig bij elkaar in de buurt worden genoemd terwijl zij eerder nooit samen voorkwamen. Op deze manier kunnen ook nieuwe metaforen en voorbeelden van ‘framing’ aan het licht komen (zie 2.4.3).

2.2 Onderzoeksvragen m.b.t. de verbreiding van veranderingen in de taal en cultuur

Veel geesteswetenschappelijke onderzoeksvragen gaan niet (alleen) over het begin van een verandering maar over de verbreiding ervan door de tijd heen. Onderzoekers trachten te achterhalen welke verschijnselen en welk type verschijnselen zich in de loop van de tijd hebben uitgebreid - van de ene tekstsoort naar een andere, van de ene auteur naar andere auteurs, van het ene dialect naar een of meer andere of naar de standaardtaal - en welke verschijnselen zijn verdwenen. Op basis van een inventarisatie hiervan zoeken zij verklaringen voor de bestaande variatie, die gebaseerd zijn op invloed en overname.

Nederlab maakt het mogelijk de verbreiding van veranderingen in teksten op te sporen. Dat gebeurt door het meten van de frequentie van verschijnselen, zoals de volgende casussen laten zien.

2.2.1 Begripsgeschiedenis

Onderzoek naar de verandering van begripshistorische termen zoals *burgerschap, nationaliteit, Nederlanderschap, verzuiling, moderniteit* is een aparte tak van de geschiedwetenschap. Binnen de letterkunde onderzoekt men de verschuivende opvattingen van bijvoorbeeld *romantiek* of *sentimentalisme*. Dergelijk receptie-onderzoek is tot nu toe verricht op basis van handmatig literatuuronderzoek en intuïtie. Nederlab maakt het mogelijk dit onderzoek te systematiseren en statistisch te onderbouwen, zodat de empirie kwantitatief getoetst kan worden aan de feiten, wat zeker tot allerlei nieuwe inzichten zal leiden.

De verbreiding van een verschijnsel blijkt uit de frequentie ervan door de tijd heen. Nederlab biedt een groot aantal *zoekprogramma's* waarmee onderzoekers in een corpus of een subset daarvan, het voorkomen van een bepaald begrip kunnen opzoeken. Om in één zoekactie ook de spelling- en vormvarianten van een begrip te vinden, kunnen onderzoekers zoeken in alleen gelemmatiseerde tekst of via een computationeel historisch lexicon (zie 2.1.1.).

Nederlab geeft diverse programma's die de gevonden resultaten *visualiseren* door ze te presenteren als lijndiagram, staafdiagram, taartdiagram, puntenwolk. Zo wordt de frequentiecurve van een bepaald begrip door de tijd heen in één oogopslag duidelijk. Ook is het mogelijk de frequentie van meerdere verwante begrippen in één grafiek op te nemen, zodat men deze met elkaar kan vergelijken. Dit is vergelijkbaar met Googles Ngram Viewer (die momenteel geen Nederlandse teksten bevat), zie J.-B. Michel, E.L. Aiden *et al.* 'Quantitative Analysis of Culture Using Millions of Digitized Books, in: *Science* December 16, 2010. Nederlab is echter veel geavanceerder: binnen Nederlab kan men de frequentiegegevens namelijk sorteren op basis van de metadata. Zo kan men bijvoorbeeld achterhalen welke tekstsoorten (politieke teksten, juridische of literaire) in welke periode verantwoordelijk zijn voor de pieken in het gebruik van een bepaalde term. Vervolgens kan men een relatie leggen met maatschappelijke ontwikkelingen. Als *slavernij* bijvoorbeeld piekt in de tweede helft van de 19e eeuw, kan met een verband leggen met (inter)nationale discussies over de afschaffing van de slavernij. Een ander methodologisch verschil met het werk van Aiden *et al.* is dat Nederlab uitgaat van open access: iedereen kan de resultaten van een zoekopdracht controleren. Aiden *et al.*'s werk daarentegen is gebaseerd op de Google Books-database, waarvan de achterliggende gegevens niet toegankelijk zijn voor buitenstaanders – sommige van de auteurs zijn in dienst van Google.

2.2.2 Het systematisch in kaart brengen van taalveranderingen

Taalkundigen zijn al decennia bezig met onderzoek naar de fundamentele vraag welk systeem er zit achter de deflexie in het Nederlands. Deflexie is het verschijnsel dat de uitgangen van woorden afslijten, verdwijnen of samenvallen. Als gevolg van deflexie worden de functies van naamvallen en verbogen vormen (dus synthetische constructies) overgenomen door analytische omschrijvingen met lidwoorden, voorzetsels en hulpwerkwoorden. Zo veranderde bijvoorbeeld *sconinx boec* in *het boek van de koning*. Deflexie leidt tevens tot allerlei verschuivingen in de morfologie en de woordvolgorde in een zin. Al in de oudste fase van het Nederlands zijn de eerste tekenen van deflexie te vinden en het verschijnsel loopt tot op heden door. Vrijwel alle taalveranderingen die in het Nederlands zijn opgetreden, zijn op enigerlei wijze gerelateerd aan het verschijnsel deflexie.

Hoewel er in de loop van de jaren steeds meer details bekend zijn gekomen over de deflexie in het Nederlands, zijn er nog veel onbeantwoorde vragen, die alleen op basis van een diachroon corpus kunnen worden beantwoord. Vragen zoals: Wanneer en in welk dialect zijn bepaalde veranderingen begonnen, hoe hebben ze zich verbreid van het ene dialect naar een ander, wat zijn de verschillen en overeenkomsten van deflexie in de verschillende Nederlandse dialecten? Zijn er in de geschiedenis van het Nederlands versnellingen, vertragingen of zelfs omkeringen in het proces van deflexie opgetreden? Hoe hebben de verschillende veranderingen in de loop van de tijd op elkaar ingegrepen en welke (talige, maatschappelijke, sociale) factoren hebben op het verschijnsel deflexie gewerkt?

Om inzicht te verkrijgen in de deflexie van het Nederlands moet een zeer groot aantal deelvragen worden beantwoord. Bijvoorbeeld hoe, waar en wanneer ieder afzonderlijk lidwoord, voornaamwoord, voorzetsel en voegwoord is opgekomen en wat de verdere ontwikkeling ervan is geweest, hoe naamvallen in de verschillende Nederlandse dialecten in de loop van de tijd zijn verdwenen, wanneer en waar hulpwerkwoorden als *hebben, zijn, worden, zullen* zijn ontstaan en hoe deze zich in de loop van de tijd hebben ontwikkeld, hoe onpersoonlijke constructies (zoals *mi lanct na di*) in de loop van de tijd zijn veranderd in persoonlijke (*ik verlang naar jou*).

Als voorbeeld van hoe taalveranderingen in het corpus gemeten kunnen worden, kan de verandering van de werkwoordsuitgang van de tweede persoon enkelvoud verleden tijd dienen. Het is bekend dat die tweede persoon in de loop van de tijd onder andere is uitgedrukt door vormen als: *du hoordes, ghi hoordet, gij hoorde, jij hoorde, u hoorde*. Het is echter niet bekend wanneer de veranderingen in de persoonsvormen precies hebben plaatsgevonden, in welke dialecten de veranderingen zijn begonnen, en of er bij de veranderingen een sociale component een rol speelde: golden bepaalde vormen als beschaafd en andere als onbeschaafd?

Om dit te kunnen onderzoeken, moet van al deze werkwoordsvormen vastgesteld worden wat hun frequentie door de tijd heen is geweest, en dit moet worden gerelateerd aan de plaats (het dialect) waar ze voorkwamen. Dergelijke syntactische onderzoeksvragen kunnen niet worden beantwoord door op concrete woordvormen te zoeken; zowel de werkwoordsvorm als de vorm van het voornaamwoord zijn immers variabel en die variatie is onvoorspelbaar. Dit soort syntactisch onderzoek kan alleen plaatsvinden op een verrijkt corpus. Hiervoor levert levert Nederlab diverse hulpmiddelen:
1. Een deel van het corpus is (semi-)automatisch *syntactisch geparst*, dat wil zeggen voorzien van grammaticale informatie.
2. Er worden verschillende *syntactische parsers* beschikbaar gesteld zodat onderzoekers relevante delen van het corpus kunnen verrijken. Voor het moderne Nederlands bestaan diverse syntactische parsers, deels taalonafhankelijk en zelflerende; voor middeleeuws materiaal komt binnenkort INPOLDER beschikbaar.

Op basis van het geparste materiaal kunnen onderzoekers de relevante werkwoordsvormen bijeenzoeken. Met behulp van de metadata kunnen ze de verspreiding over de dialecten door de tijd heen bepalen; om dat te visualiseren biedt Nederlab verschillende *kaarttekenprogramma's* aan, zoals het CLARIN-project gabmap.

2.3 Onderzoeksvragen m.b.t. het leggen van (statistische) verbanden en netwerken

Voor allerlei geesteswetenschappelijke onderzoeksvragen is het van belang herhalende, min of meer vaste woordpatronen te herkennen – zo vindt men bijvoorbeeld literaire motieven of woordvelden van aan elkaar gerelateerde woorden. Door het in kaart brengen van de relaties tussen verwante verschijnselen, zoals persoons- en plaatsnamen, kunnen netwerken worden blootgelegd. Dergelijke netwerken geven inzicht in de verbreiding van kennis, cultuur en taal. Binnen een diachroon corpus kunnen verschuivingen en ontwikkelingen in patronen worden afgeleid, wat weer kan wijzen op ontwikkelingen in taal en cultuur. Nederlab biedt allerlei tools die statistische verbanden tussen woorden leggen zodat automatisch patronen worden herkend.

2.3.1 Het vaststellen van patronen en motieven

Historici willen bijvoorbeeld weten hoe opvattingen over astronomie wijzigden door de uitvinding van betere instrumenten. Of hoe er in de loop van de tijd tegen verschillende religies is aangekeken of hoe de visie op het buitenland in de loop van de tijd is gewijzigd. Letterkundigen willen onderzoeken hoe allochtone inwoners van de kolonie in (post)koloniale romans worden omschreven, met welke motieven ze worden geassocieerd en wat daarin de variaties per genre, gebied, auteur, periode zijn. Voor de beantwoording van dergelijke onderzoeksvragen biedt Nederlab allerlei *textmining- of dataming-tools*. Deze stellen voor grote hoeveelheden teksten, op basis van statistische verbanden, semantische woordvelden vast: woorden die veel in elkaars buurt voorkomen. Zo vormen ze een woordveld rond het begrip 'astronomie' met woorden als *ster* en *planeet*, maar ook het vroeger gebruikte *dwaelder*. Een woordveld rond het begrip 'religie' bevat woorden als *mohammedaans, islam, heidens*. Nederlab biedt de mogelijkheid om dergelijke woordvelden te visualiseren in *woordwolken*, maar ook om die woordvelden statistisch te onderscheiden en veranderingen te monitoren. Zo kunnen significante verschuivingen in concepten worden gemeten en afgeleid. Voor het moderne Nederlands is aan de VU een semantisch verrijkt corpus ontwikkeld, DutchSemCor, dat als hulpmiddel bij textmining kan worden gebruikt. Dit is een meer-precieze semantische duiding van de tekst die het mogelijk maakt om betekenis te koppelen aan de collocationele context, maar ook om te zoeken naar concepten ongeacht de vorm (bijv. zowel *planeet* als *dwaelster*, *dwaelder* en *hemellichaam*). Een dergelijke verrijking kan ook worden toegepast op het historische corpus. Dit biedt verdere mogelijkheden om betekenisontwikkeling en conceptgebruik door de tijd te volgen.

Letterkundigen gebruiken textmining-tools en topical modelling om automatisch motieven en onderwerpen in verhalen, teksten en liederen te herkennen. Aan de ontwikkeling van tools hiervoor wordt binnen het project Tunes & Tales verkennend onderzoek uitgevoerd met subsidie van de KNAW, mede voortbouwend op het NWO-project Dutch Songs Online.

2.3.2 Het vaststellen van subjectieve oordelen

Het voordeel van onderzoek op grond van een diachroon corpus is dat hiermee grote hoeveelheden data kwantitatief kunnen worden doorzocht. Maar veel geesteswetenschappelijke onderzoeksvragen hebben een

kwalitatieve vraagstelling: zij onderzoeken bijvoorbeeld subjectieve oordelen zoals de receptie van een bepaalde auteur, een bepaald werk of een bepaalde gebeurtenis. Of zij willen achterhalen hoe emoties werden opgeroepen in politieke teksten, preken en op het toneel in de 18e eeuw. Of met welke taalmiddelen politieke achterban wordt gecreëerd in de 20e eeuw. Of in welk teksttype, in welke tijdperiode en door welke auteurs de meeste verbale agressie wordt getoond.

Om dit te kunnen onderzoeken biedt Nederlab tools aan voor *sentiment mining*, zoals WEKA. Dit is een pakket machinelerende software waarmee aan de hand van gebruikte bijvoeglijke naamwoorden en werkwoorden wordt bepaald of het sentiment van een tekst positief, negatief of neutraal is. Dit gebeurt meestal door de woorden die in een tekst gebruikt worden, te vergelijken met een tabel waarin de sentimentwaarden van die woorden zijn opgenomen - die tabel moet door onderzoekers handmatig worden aangelegd, zeker voor de oudere tijden. Met behulp van deze software kan men bijvoorbeeld positieve en negatieve recensies van literaire werken analyseren. Onderzoekers kunnen de data op allerlei manieren filteren en de resultaten met elkaar vergelijken.

2.3.3 Het bepalen van de relaties tussen personen en plaatsen

Voor letterkundig receptie-onderzoek en onderzoek naar processen van reputatie- en canonvorming is het cruciaal te achterhalen hoe vaak en door wie een bepaalde auteur in de eigen tijd en in latere tijden wordt vermeld. Zo kunnen netwerken tussen elkaar beïnvloedende auteurs worden vastgesteld. Die netwerken kunnen worden gevisualiseerd, waarbij bijvoorbeeld de dikte van de lijnen afhangt van de frequentie van de vermeldingen. Ook het bepalen van de relaties tussen romanfiguren of figuren in een toneelstuk leidt tot interessant nieuw onderzoek. Historici willen achterhalen hoe er in de loop van de tijd is aangekeken tegen historische figuren zoals bepaalde heersers, politici, geleerden.

Voor dit type onderzoek moeten persoonsnamen worden geïdentificeerd. Namen kunnen niet betrouwbaar worden opgezocht in een ongestructureerd corpus. Ze zijn namelijk in principe niet uniek: meerdere personen dragen al dan niet gelijktijdig dezelfde naam. Voor onderzoek is het belangrijk de verschillende personen van elkaar te onderscheiden. Ook moeten namen worden onderscheiden van woorden: meneer Smid is niet dezelfde als de smid in een bepaald dorp.

Nederlab levert verschillende tools voor het automatisch herkennen en taggen van persoonsnamen en plaatsnamen, de zogenoemde *Named Entity Recognition*, die echter alleen voor modern Nederlands zijn getraind. De getagde persoonsnamen worden zoveel mogelijk automatisch gelinkt met bibliografische en biografische bestanden, zoals het Biografisch Portaal, parlement.com, de auteursgegevens van de DBNL, de bibliografische bestanden van de KB, genealogische databestanden of internationale namencollecties als Wikipedia of VIAF. Oudere plaatsnamen worden zoveel mogelijk gelinkt aan moderne gemeentenamen, en deze worden gelinkt aan *GPS-coödinaten*, zodat kaarttekenprogramma's ermee overweg kunnen. Men kan dan bijvoorbeeld een kaart laten tekenen van de geboorteplaatsen van de belangrijkste schrijvers in een bepaalde periode.

Ook biedt Nederlab zogenoemde *Coreference Resolution tools*, die automatisch verwijzingen naar personen in de vorm van voornaamwoorden (‘hij’, ‘haar’) of omschrijvingen (‘de voorzitter’, ‘de stagiaire’) relateren aan de ermee bedoelde persoonsnaam. De tool is ontwikkeld bij Informatica aan de UvA voor Staten-Generaal Digitaal, maar kan ook toegepast worden op literaire werken en toneelstukken.

Uiteindelijk wordt zo binnen het diachrone corpus een enorm netwerk van aan elkaar gerelateerde gegevens vastgelegd, die allerlei verbanden en dwarsverbanden zichtbaar maken, waarbij wordt voortgebouwd op de resultaten van de NWO-projecten CKCC Geleerdenbrieven, Historical Timeline Mining and Extraction (HITIME, een tool die historische gebeurtenissen, personen, organisaties, beroepen, tijdsaanduidingen en locaties herkent en linkt), en LINKS (linking system for historical family reconstruction), dat de reconstructie van alle families in Nederland uit de negentiende en begin twintigste eeuw beoogt op basis van de gegevens uit de gedigitaliseerde registers van de burgerlijke stand in Genlias. Door zoveel mogelijk gegevens aan elkaar te linken, kunnen belangrijke onderzoeksvragen beantwoord worden zoals: Welke invloed hebben binnenlandse en buitenlandse literaire en culturele stromingen in de loop van de tijd in de Lage Landen gehad en hoe heeft de verbreiding en doorwerking van hun invloed plaatsgevonden? Welke contacten hebben er bestaan tussen Noord- en Zuid-Nederlandse auteurs van 1100 tot heden, in welke periodes waren deze contacten het meest frequent en in welke

richting verliepen ze? Dergelijke vragen zijn niet nieuw, maar voor het eerst wordt het mogelijk ze systematisch te onderzoeken en oude hypothesen te toetsen.

2.4 Onderzoeksvragen gebaseerd op systematische tekstvergelijking

Het feit dat de computer in staat is teksten automatisch met elkaar te vergelijken, heeft geleid tot een groot aantal nieuwe onderzoeksvragen. Onderzoekers passen tekstvergelijking toe om overeenkomsten in teksten te detecteren en zo plagiaat, citaten of parafrasen op te sporen (2.4.1), om verschillen tussen bijvoorbeeld dialecten of het taalgebruik van auteurs te vinden (2.4.2), om metaforen te herkennen (2.4.3) of om de herkomst van onbekende teksten te determineren (2.4.4 en 2.4.5).

2.4.1 Het detecteren van overeenkomsten in teksten: plagiaat, parafrasen, citaten

Letterkundigen zijn geïnteresseerd in de vraag welke auteurs aan elkaar schatplichtig zijn en in hoeverre. Om dit te achterhalen biedt Nederlab allerlei tools. *Plagiaatherkenners* herkennen letterlijk overgenomen fragmenten en citaten uit andere werken. Omdat auteurs overgenomen teksten vaak niet letterlijk overnemen maar erop variëren, worden er ook *parafraseherkenners* aangeboden: tools die variaties van teksten herkennen, zoals het door de universiteit van Tilburg ontwikkelde DAESO (Detecting And Exploiting Semantic Overlap). Dit werkt echter alleen op moderne teksten. Wanneer men grote hoeveelheid literaire werken door plagiaat- en parafraseherkenners laat doorploegen, zullen er ongetwijfeld allerlei nieuwe vondsten worden gedaan waarbij auteurs onverwacht schatplichtig blijken te zijn aan andere.

Parafraseherkenners of intelligente zoekmachines kunnen worden ingezet om te onderzoeken waar en wanneer min of meer vaste combinaties zoals pregnante zegswijzen, catchphrases, spreekwoorden, spreuken zijn opgekomen en hoe ze zijn overgenomen door auteurs en door verschillende tekstgenres. Van citaten kan men onderzoeken of de (vermeende) oorspronkelijke auteur erbij wordt vermeld en in hoeverre er op een citaat wordt gevarieerd. Zo krijgt men een beeld van de bekendheid en populariteit van citaten en schrijvers door de eeuwen heen. Hiernaar is nog nooit systematisch onderzoek verricht, omdat de bronnen ervoor ontbraken.

Citaten uit veel geciteerde werken kunnen worden getagt en gelinkt aan een bepaalde vaste editie. Met name voor de bijbel is dit interessant: dit werk is het meest geciteerd, in zowel literaire werken als nonfictie-teksten, terwijl er op verschillende manieren naar wordt verwezen (Gen. 10:12, Gene. X, 12, etc.) en ook de indeling van de bijbel in hoofdstukken en verzen in de loop van de tijd is gewijzigd. Binnen het corpus van de DBNL is een test gedaan met het automatisch linken van alle bijbelverwijzingen aan de Nieuwe Bijbelvertaling. Hieruit kan men allerlei gegevens afleiden zoals in welke tijden welke bijbelfragmenten het meest geciteerd zijn en welke bijbelfragmenten onder literaire auteurs en journalisten het populairst zijn.

2.4.2 Het detecteren van verschillen tussen tekstsoorten, auteurs, dialecten

Nederlab biedt een aantal taalonafhankelijke tools die een statistische analyse maken van teksten. Deze *corpusanalyse-tools* leggen van iedere tekst de teksteigenschappen (features) vast op basis van het aantal types en tokens en de verhouding daartussen, de frequentie, het aantal lettergrepen, woordlengte, zinslengte, meest voorkomende woordcombinaties, etc. Deze tools stellen onderzoekers in staat om de verschillen te vinden tussen specifieke delen van het corpus. Zo kan men een corpus met zakelijke teksten zoals advertenties uit een bepaalde periode vergelijken met literaire teksten uit diezelfde periode en daaruit conclusies trekken over het verschil in taalgebruik. De resultaten van het analyseprogramma moeten uiteraard door de onderzoekers worden gewogen, want niet alle verschillen zijn even belangrijk. De resultaten, inclusief het oordeel van de onderzoekers, kunnen binnen Nederlab bewaard worden voor volgende onderzoekers.

Er zijn veel soorten onderzoeksvragen die door tekstanalyse beantwoord kunnen worden. Taalbeheersers kunnen onderzoeken welke en hoeveel verschillen er bestaan tussen het Belgisch-Nederlands en het Nederlands-Nederlands in de 21e eeuw en of die verschillen groter of kleiner zijn geworden ten opzichte van bijvoorbeeld 1950. Taalkundigen kunnen de bewering dat de taal van de Statenvertaling representatief is voor het Nederlands zoals dat rond 1637 werd gesproken of geschreven, toetsen aan de feiten door de tekst van de Statenvertaling te vergelijken met bijvoorbeeld een corpus literaire teksten uit dezelfde periode, een corpus geleerdenteksten of egodocumenten. Historici willen het politieke taalgebruik in de 21e eeuw vergelijken met dat uit de tweede helft van de 20e eeuw om

zo de vraag te kunnen beantwoorden of het taalgebruik inderdaad verruwt, zoals vaak wordt beweerd. Voor boekhistorici is het interessant om de relaties tussen verschillende afschriften of drukken van een hetzelfde werk in de loop van de tijd te bekijken, of de relaties tussen klassieke werken en bewerkingen daarvan voor de jeugd. Letterkundigen willen de overgang van een stijlfiguur als de vrije indirecte rede ('Ze vroeg zich af: moest ze gaan?') van literaire naar journalistieke teksten onderzoeken.

Nederlab biedt voorts tools aan die de afstand berekenen tussen verschillende dialecten, ofwel *dialectometrische analyse*. Deze tools zijn ontwikkeld voor moderne dialecten, en zij berekenen de afstand tussen de verschillende dialecten en het Standaardnederlands. Onderzoekers kunnen deze tools binnen Nederlab uitbreiden zodat zij ook voor oudere taalperioden van het Nederlands werken. Hierdoor kunnen taalkundigen de verschillen laten berekenen tussen Oudnederlandse, Middelnederlandse en moderne dialecten, en de afstand van de verschillende dialecten tot het Standaardnederlands. Deze vergelijkingen zullen helpen bij de beantwoording van onderzoeksvragen als: Hoe is het Standaardnederlands ontstaan uit de verschillende Middelnederlandse dialecten? Hoe zijn de moderne Nederlandse dialecten ontstaan uit de Middelnederlandse en Oudnederlandse dialecten? Heeft de westelijke standaard de dialecten beinvloed, of hebben de locale regionale standaarden als substraat gediend bij het ontstaan van de standaard in de Hollandse steden? Hoe zijn dialecten in de loop van de tijd uit elkaar of naar elkaar toegegroeid, en valt dat samen met politieke veranderingen?

2.4.3 Het vinden van metaforen

Onderzoekers uit alle geesteswetenschappelijke disciplines zijn geïnteresseerd in metaforen: letterkundigen en taalbeheersers stellen onderzoeksvragen als: door welke auteur, in welke literaire stroming, in welke tekstsoort is een bepaalde metafoor geïntroduceerd? Hoe worden metaforen van kleine locale uitdrukkingen 'opgeblazen' tot complete tekstorganiserende principes die een gedicht of zelfs een roman tot een spannende exercitie in metaforisch en vernieuwend denken maken? Taalkundigen zijn geïnteresseerd in de vraag welke woorden zich lenen voor metaforisch gebruik, en zoeken naar bewijzen voor de stelling van embodied cognition van Lakoff en Johnson, die ervan uit gaan dat de menselijke cognitie wordt bepaald door de vorm van het lichaam. Bovendien wordt geponeerd dat metaforiek een belangrijk mechanisme is in de afvlakking van betekenissen in concrete taalvormen en daarmee een essentiële component is in het grammaticalisatieproces dat alle talen doormaken, maar wat voor het Nederlands op deze manier nog niets is onderzocht. Historici onderzoeken de introductie van een nieuwe metafoor omdat deze vaak een aanwijzing is voor een omslag in het denken. Eerst wordt een idee in bijvoorbeeld de krant geconceptualiseerd (drankzucht als 'kanker van de maatschappij'), voordat daadwerkelijke veranderingen in de maatschappij of de politiek plaatsvinden, zoals belastingverhoging op sterkedrank. Conceptualisaties van organisaties en management, politiek en overheid zijn ook gestoeld op onderliggende basismetaforen (de staat is een persoon, of een gezin; een organisatie is een plant of een boom die kan groeien en gesnoeid kan worden), waarvan de ontwikkeling nu voor het Nederlands, en voor allerlei Nederlandse communicatiedomeinen, op lange termijn getraceerd kan worden.

Metaforen kunnen tegenwoordig – mede dankzij NWO-programma's die aan de VU zijn uitgevoerd - met een hoge betrouwbaarheid in corpora worden opgespoord en geannoteerd. Onderzoek heeft aangetoond dat de distributie van metaforen sterk samenhangt met de taalvariëteit – conversatie, geschreven tekst etc. - waarin ze voorkomen. Bovendien is aangetoond dat de functie van metaforen in specifieke registers door de tijd heen sterk zijn veranderd, door de differentiatie van de media en hun evoluerende en uiteenlopende culturele functies. Omdat metaforiek een wezenlijke rol speelt bij de modellering van allerlei abstracte concepten is het van belang te onderzoeken hoe metaforiek daarin een rol heeft gespeeld, als oorzaak én als gevolg.

Daarbij kan ook voor het eerst worden gekeken naar de historische ontwikkeling van hele metafoorcomplexen als cognitief en cultureel gedeelde uitdrukkingsmiddelen voor de conceptualisatie van allerlei emotionele, mentale, sociale en culturele verschijnselen en ervaringen, die op hun beurt ook een sterke evolutie hebben doorgemaakt over de afgelopen duizend jaar. Onze taal staat bol van de watermetaforen—waar komen die vandaan en hoe hebben die het denken en het publieke debat over allerlei kwesties vormgegeven, tot aan de stroom/golf/vloedgolf/tsunami aan immigranten toe? Wanneer kwam het 'poldermodel' op en hoe hangt dat samen met de historische ontwikkeling van onze opvattingen omtrent democratie, participatie en cultuur? Hoe worden zij gebruikt en misbruikt in de taal van de media? Dergelijke metaforen en metaforische modellen worden soms tot

voorwerp van intens publiek debat, waarna ze meer of juist minder geaccepteerde manieren van spreken zijn geworden voor specifieke inhoudelijke domeinen waarmee ze uiteindelijk effect uitoefenen op ons individuele wereldbeeld, de (politieke en andere) keuzes die we maken, en ons gedrag. Soortgelijke processen spelen zich af in de financiële sector (banken die omvallen), het milieu (het broeikaseffect) en de gezondheidszorg (aids als een plaag). Nederlab geeft geesteswetenschappers de mogelijkheid al deze kwesties op ambitieuze schaal voor de Nederlandse taal en cultuur aan te pakken, een project dat uniek is in de wereld.

2.4.4 Lokalisering en datering van een onbekende tekst

In principe heeft iedere tekst binnen het corpus een (globale) indicatie over tijd, plaats, auteursgegevens en tekstgenre. Er zijn echter teksten waarvan één of meerdere van deze gegevens vooralsnog onbekend zijn. Deze teksten vormen een interessante onderzoekscasus. Taalkundigen willen op basis van dialectkenmerken een bepaalde tekst herleiden tot een bepaalde plaats of regio. Zowel voor taalkundigen als voor letterkundigen is het interessant om op basis van tekstkenmerken een bepaalde tekst zo exact mogelijk te dateren. Voor letterkundigen is het interessant om teksten op grond van formele kenmerken onder te brengen bij één van de verschillende tekstgenres of bij een bepaalde auteur of vertaler.

Nederlab maakt het determineren van onbekende teksten mogelijk. Stel, een onderzoeker heeft een tekst uit de 15e eeuw waarvan hij niet weet uit welke regio deze afkomstig is. Om de onbekende tekst te lokaliseren, selecteert de onderzoeker, op grond van de metadata ‘plaats’ en ‘tijd’, alle teksten uit de 15e eeuw waarvan de lokalisering precies bekend is. Deze verdeelt hij handmatig in dialectgroepen: een corpus 15e-eeuws Limburgs, 15e-eeuws Brabants, 15e-eeuws Nedersaksisch, 15e-eeuws Hollands et cetera. Hiervoor zal hij allerlei beslissingen moeten nemen: welke dialectgroepen wil hij onderscheiden voor de 15e eeuw: Gronings, Drents, Twents, of liever ruimer: Nedersaksisch, of kleiner: Oldenzaals? Welke concrete plaatsen horen bij een bepaalde dialectgroep? Zijn beslissingen legt de onderzoeker vast in metadata, waardoor het corpus wordt verrijkt.

Vervolgens laat de taalkundige ieder dialectcorpus afzonderlijk analyseren door een corpusanalyse-tool. De tool stelt op basis van een statistische analyse de teksteigenschappen van ieder corpus vast. De onderzoeker vergelijkt de features van de verschillende dialectcorpora met elkaar en weegt deze: hij bepaalt welke woorden, woordvormen, klanken en woordcombinaties kenmerkend zijn voor een bepaalde regio. Het resultaat is dus een beschrijving van kenmerkende taalkundige eigenschappen van de dialecten in een bepaalde tijd, zoals de 15e eeuw. Deze dialecteigenschappen kunnen op een kaart worden getekend via een kaarttekenprogramma. Hierna laat de onderzoeker de onbekende tekst analyseren en bekijkt hij met welk dialectcorpus de onbekende tekst de meeste overeenkomsten heeft. Zo kan hij met meer of minder waarschijnlijkheid bepalen uit welk dialect een tekst afkomstig is.

Het resultaat van het onderzoek is niet alleen dat een onbekende tekst is gelokaliseerd, maar ook dat een deel van het corpus voor verder onderzoek is verrijkt met metadata en dat een bestaande tool is toegepast voor een nieuwe onderzoeksvraag, wat heeft geleid tot nieuwe, reproduceerbare data/criteria. Dankzij Nederlab krijgen andere onderzoekers exact inzicht in de onderzoeksmethoden van collega's, en kunnen zij bovendien gebruikmaken van de verrijkte gegevens. Zij kunnen bijvoorbeeld het onderzoek herhalen voor de 14e en de 16e eeuw, of voor nieuwe teksten uit de 15e eeuw. Op min of meer dezelfde manier kan een onderzoeker een ongedateerde tekst binnen Nederlab dateren.

2.4.5 Tekstgenre-herkenning

Genres zijn interpretatiekaders waarbinnen een oordeel over een tekst wordt gevormd; genres en genre-indelingen hebben dus een ideologische kracht. Genres en genre-indelingen veranderen in de tijd. Kwantitatief onderzoek biedt de mogelijkheid de deels tot ‘intuitie’ geworden kennis van genres en genresystemen te onderbouwen én te corrigeren, zoals in de VS door Moretti e.a. wordt onderzocht.

Stromingen en periodes in de literatuurgeschiedenis (inclusief de literatuurgeschiedschrijving) onderscheiden zich door een eigen jargon, een eigen stijl, een eigen thematiek, die in de verschillende genres (roman, poëzie, toneel, boekbesprekingen, essays, geschiedschrijving) tot uiting komt. Ook het genre-systeem (inclusief genre-aanduidingen) en/of de hiërachie van genres kan verschillen. Diachroon onderzoek naar de

ontwikkeling van bepaalde tekstsoorten maakt het mogelijk veranderingen op het spoor te komen, zonder de historische diversiteit en de gelijktijdigheid van het ongelijktijdige uit het oog te verliezen.

De verklaring van de veranderingen in genres en genresystemen kan langs verschillende wegen worden gezocht. Invloed van buitenlandse ontwikkelingen, bijvoorbeeld, kan worden getraceerd door comparatieve kwantitatieve analyse van woordvelden. Genres en genresystemen veranderen bovendien onder invloed van maatschappelijke ontwikkelingen, zodat de studie van genrevorming bij uitstek een manier is om de wisselwerking tussen literatuur en maatschappij in kaart te brengen. Een voorbeeld is de meisjesroman waarvan de plotstructuur lijkt te zijn veranderd onder invloed van de veranderde positie van de vrouw in de samenleving vanaf de eerste feministische golf. Of om een recent voorbeeld te noemen: de zogeheten 'literaire thriller' is een nieuwkomer in het literaire genresysteem. Wanneer duikt dit genre op? Is dit genre een uitvinding van de uitgevers? Betekent dit dat het domein van de literatuur meer dan voorheen door commerciële partijen wordt gedefinieerd, ten koste van critici en letterkundigen? Dergelijk onderzoek sluit aan bij het eHumanities project 'The Riddle of Literary Quality' van de KNAW, waar gezocht wordt naar formele kenmerken van het literaire genre.

Beschrijving en motivatie van de investering

Motivatie

Er is en wordt veel overheidsgeld gestopt in het digitaliseren van corpora en in de ontwikkeling van tools om deze te doorzoeken, analyseren en bewerken. Het toepassen van statistische analysemethodes en de introductie van beta-wetenschappelijke waarden als verifieerbaarheid en herhaalbaarheid beloven al enige tijd een ingrijpende vernieuwing van het onderzoek in de alfawetenschappen in gang te zetten. Er bestaat echter een kloof tussen de geboden middelen en de onderzoekspraktijk van de gemiddelde geesteswetenschapper, waardoor de beloftes nog niet worden waargemaakt. Het is de ambitie van Nederlab om die problemen in één klap op te lossen.

Het probleem van de diachrone corpora[2] is dat deze op verschillende plaatsen en door verschillende instellingen worden aangeboden. Grote corpora zijn aanwezig bij DBNL, Huygens ING, INL, KB, Meertens Instituut. Daarnaast worden vele kleinere corpora aangeboden op websites van universiteiten of individuele onderzoekers. Al deze corpora kunnen nu slechts naast elkaar - en niet tegelijkertijd en samen - worden doorzocht en geanalyseerd. Bovendien verschillen de zoekinterfaces en zoekmogelijkheden per corpus, bestaan er aanzienlijke kwaliteitsverschillen tussen de verschillende corpora, en voegt iedere instelling zijn eigen metadata toe.

Voor de beantwoording van longitudinale onderzoeksvragen is het absoluut noodzakelijk dat alle losse corpora als eenheid doorzoekbaar gemaakt worden. De grootste uitdaging van Nederlab zal eruit bestaan te inventariseren welke corpora er zijn en er vervolgens voor te zorgen dat alle (bestaande en nog te vervaardigen) diachrone corpora gedistribueerd doorzoekbaar worden, zowel op tekstniveau als op metadata-niveau (met behulp van een metadata-harvester). De toevoeging van uniforme metadata zal al snel zorgdragen voor een betere toegankelijkheid van de bestaande corpora, wat de behoefte aan kwaliteitsverbetering en standaardisering van de data zal versterken. Kwaliteitsverbetering is des te belangrijker omdat momenteel het gevaar bestaat dat Google de toon gaat aangeven als het gaat om de kwaliteit van de corpora: de door Google geleverde ocr is echter voor wetenschappelijk onderzoek onvoldoende, waardoor statistische resultaten onbetrouwbaar zijn. Bovendien is de manier waarop Google werkt niet geheel duidelijk: de gescande boeken kunnen op allerlei, door Google vastgestelde manieren doorzocht worden, maar onderzoekers krijgen geen directe toegang tot het onderliggende corpus. Het is belangrijk dat studenten en onderzoekers binnen Nederlab een betrouwbaar corpus aangeboden krijgen voor hun onderzoek en de mogelijkheid om onbetrouwbare teksten te (laten) corrigeren en annoteren.

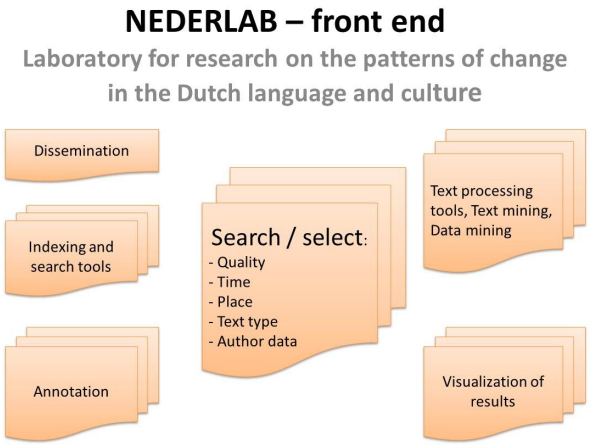
Het tweede probleem vormen de tools. Momenteel worden er allerlei tools ontwikkeld, die uiteindelijk in principe beschikbaar komen via CLARIN-NL. Deze tools zijn echter vaak niet bruikbaar voor een andere situatie dan waarvoor ze zijn ontwikkeld (bijvoorbeeld voor morfologische verrijking van teksten uit de 14e eeuw die op een bepaalde manier zijn voorbewerkt). Nederlab wil, in nauw overleg met CLARIN, de bestaande tools op een centrale plaats bijeenbrengen en de gebruikersinterfaces zo aanpassen dat ze binnen de geboden infrastructuur direct

[2] Zie ook 'Het digitale drama' in *NRC Handelsblad*, Wetenschap, 10 september 2011.

bruikbaar zijn voor niet technisch onderlegde onderzoekers, en toepasbaar zijn op diachrone teksten. Daarbij sluit Nederlab aan bij de standaarden in formaten en metadata van CLARIN.

Nederlab bouwt dus voort op de vele initiatieven met betrekking tot corpusopbouw en toolsontwikkeling die door de Nederlandse overheid en onderzoeksinstellingen als KNAW en NWO zijn ontwikkeld, maar voegt hieraan belangrijke meerwaarde toe: een comfortabele infrastructuur voor onderzoekers en studenten die automatisch leidt tot samenwerking en synergie binnen de geesteswetenschappen en tot het stellen van nieuwe, veelal interdisciplinaire, onderzoeksvragen. Met een betrekkelijk geringe investering wordt zo een enorme meerwaarde verkregen. Dit is conform de toekomstvisie van eurocommissaris Neelie Kroes in de inleiding van het rapport *Riding the wave*[3]: "My vision is a scientific community that does not waste resources on recreating data that have already been produced, in particular if public money has helped to collect those data in the first place. Scientists should be able to concentrate on the best ways to make use of data. Data become an infrastructure that scientists can use on their way to new frontiers." Nederlab zal automatisch leiden tot standaardisering in metadata en formaten van corpora, zoals hieronder blijkt.

Technische beschrijving



Front-end en back-end

Nederlab functioneert als een virtuele onderzoeksomgeving voor diachroon taalonderzoek. De front-end werkt via de webbrowser van de gebruiker. De gebruikersinterface is helder en efficiënt en wordt zo ingericht dat eindgebruikers (onderzoekers, studenten) er gemakkelijk mee kunnen werken. Hierin verschilt Nederlab van bestaande gebruikersinterfaces zoals die van CLARIN, die vooral gericht zijn op het aanbieden van tools en data. De front-end biedt tabs of keuzemenu's voor verschillende functionaliteiten: een tab voor informatie (met forum en helpdesk), een tab voor navigeren, zoeken en selecteren van subsets van het corpus, een tab voor het visualiseren van de resultaten, een tab voor annotateren en ten slotte een tab voor het uitvoeren van tools voor tekstbewerking. De back-end draait op meerdere, gedistribueerde servers.

Basisdiensten

Iedere gebruiker van Nederlab heeft toegang tot een aantal basisdiensten – afhankelijk van het soort gebruiker - en iedereen kan door het diachrone corpus bladeren en zoeken. De teksten van het diachrone corpus worden beschikbaar gesteld door een aantal wetenschappelijke bibliotheken en instellingen (zie Appendix 2). Binnen Nederlab worden de teksten en metadata van de verschillende instellingen onderling gelinkt en er wordt een laag toegevoegd met informatie over de kwaliteit. Dit maakt het makkelijker om geschikte teksten te vinden, en deze

[3] http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

informatie biedt grote toegevoegde waarde voor zowel Nederlab-gebruikers als corpusleveranciers. De kwaliteit van de collecties is heel wisselend. Tijdens de samenstelling van het diachrone corpus informeren we de gebruikers over welke subcorpora er allemaal zijn en welke kwaliteitsverschillen ertussen bestaan, zodat ze in staat zijn de juiste keuzes te maken voor het verkrijgen van betrouwbare onderzoeksresultaten. In overleg met de dataleveranciers en de gebruikers trachten we te komen tot zo homogeen mogelijke data. Dit kan worden bereikt door iedere bibliografische eenheid (zelfstandige titel, tijdschriftnummer, fragment, artikel) tijdens de samenstelling van het diachrone corpus te voorzien van aparte administratieve informatie, in de vorm van een xml-document dat tegelijkertijd dient als 'container' van het gedigitaliseerde document. In de containers worden ook de metadata opgenomen en tagging die in een later stadium door onderzoekers wordt toegevoegd.

Onderzoekers krijgen een eigen virtuele werkruimte binnen Nederlab waar ze data kunnen verzamelen, data met andere gebruikers kunnen delen, en data met tools kunnen bewerken. Onderzoekers kunnen hun eigen datasets uploaden en hieraan metadata toevoegen, of ze kunnen data uit het diachrone corpus selecteren. Data in afzonderlijke werkruimten kunnen ook worden gedeeld met andere onderzoekers om er gezamenlijk aan te werken. Andere onderzoekers kunnen wijzigingen voorstellen in de metadata, of ze kunnen de data gebruiken voor verdere bewerking. Wijzigingen in de metadata moeten dan goedgekeurd worden door de eigenaar van het metadata-document, d.w.z. de onderzoeker die de metadata oorspronkelijk creëerde. Als data verder worden bewerkt, worden de resultaten, dus de afgeleide data, opgeslagen in de werkruimte van de onderzoeker. Het oorspronkelijke document blijft zo in zijn originele vorm behouden, en via de herkomstinformatie kunnen de secundaire data gelinkt worden aan het oorspronkelijke document. Toepassing van verdere operaties (b.v. lemmatisering > PoS tagging > Named Enitity Recognition) levert een reeks secundairedata-documenten op, waarbij het resultaat van elke afzonderlijke stap door de onderzoeker gecontroleerd kan worden. De onderzoeker kan de verschillende processen naar behoefte aanpassen. Nederlab levert een aantal standaarddiensten of workflows, waaruit de onderzoeker de tools kan kiezen die voor zijn type onderzoek nodig zijn. Die keuzes worden opgeslagen in de virtuele werkruimte.

Het diachrone corpus zal worden uitgebreid met bijdragen van deelnemende onderzoekers. Het toevoegen van materiaal aan het corpus wordt beoordeeld door een redactieteam, waarvan de leden worden aangewezen door de supervisoren van de datacuratie. De beoordeling en goedkeuring zullen formeel geregeld worden binnen Nederlab. Daarbij zullen afspraken gemaakt worden over de manier waarop nieuwe data aan het corpus worden toegevoegd en er zal worden geregeld dat de nieuwe data opgeslagen en duurzaam bewaard worden in een geschikt archiveringscentrum.

Authenticatie en autorisatie

Authenticatie en autorisatie zijn sleutelbegrippen voor iedere infrastructuur: virtuele werkruimten, services en data moeten beveiligd worden tegen ongeoorloofde toegang, en onderzoekers moeten andere onderzoekers toegang kunnen verlenen tot hun data en metadata om samen te kunnen werken. Nederlab moet ervoor zorgen dat de juiste autorisatieprocedures worden gevolgd bij bijvoorbeeld het zoeken door documenten. Als onderzoekers hun onderzoeksresultaten vrij toegankelijk maken, kunnen ze resultaten publiceren die (nog) niet bij archiverende instellingen zijn opgeslagen. Autorisatieprocedures zijn ook nodig wanneer nieuwe onderzoeksgegevens formeel voorgelegd worden om in het corpus opgenomen te worden: bij de voorbereiding van het materiaal voor opname in het corpus wordt de toegang tot metadata en onderzoeksgegevens overgeheveld van de onderzoeker naar het redactieteam. Het project zal voortbouwen op de ervaring die is opgedaan bij projecten als CLARIN-NL, CatchPlus en SURF, waar momenteel projecten lopen die verschillende aspecten van autorisatie en authenticatie onderzoeken. Authenticatie zal via de SURFFederation plaatsvinden, met Single Sign-on zoals verschaft door Shibboleth, hoewel ondersteuning van andere systemen niet wordt uitgesloten.

Beheer

De Beheersmodule levert standaarddiensten als het toevoegen en verwijderen van gebruikers of groepen binnen de Nederlab-omgeving. Als onderdeel van de algemene gebruikersinformatie biedt Nederlab de mogelijkheid om binnen de Nederlab-omgeving verschillende gebruikersrollen te onderscheiden, met verschillende autorisatieniveaus, zoals onderzoeker, redacteur, administrator. Deze autorisatieniveaus zorgen ervoor dat de eindgebruiker in de gebruiksomgeving verschillende functionaliteiten tot zijn beschikking krijgt.

Beheer van de virtuele werkruimten

Virtuele werkruimten bevatten informatie van een bepaalde gebruiker: persoonlijke onderzoeksgegevens, verwijzingen naar onderzoeksgegevens in het diachrone corpus, gebruikersvoorkeuren (zoals op het gebied van tools) en administratieve gegevens zoals indexen. De toegewezen schrijfrechten ondersteunen het beheer van de virtuele werkruimten. Gebruikers kunnen tools kiezen die zij in hun gepersonifieerde Nederlab-omgeving willen opnemen, en ze kunnen data uploaden, data organiseren en de toegang beperken tot data die ze hebben opgeslagen in hun werkruimte. De Nederlab-omgeving gebruikt de werkruimte ook voor de toevoeging van gebruikersinformatie aan de werkruimte, zoals indexen op data binnen de werkruimte, zodat het beheer van de werkruimte soepel loopt. Voor de opslag van de data in de werkruimten zal met SARA/BigGrid worden samengewerkt.

Metadata

Alle onderzoeksgegevens in Nederlab, zowel die in de werkruimten als die van het diachrone corpus, worden voorzien van metadata-beschrijvingen. Om een groot aantal verschillende metadata-beschrijvingen mogelijk te maken, bestaande uit zowel beschrijvende velden als uit structuurbeschrijvingen, is er een flexibel en uitbreidbaar metadata-schema nodig, zodat de metadata-beschrijvingen gebruikt kunnen worden voor verschillende onderzoeksgebieden en projecten. De vervaardigde metadata-beschrijvingen zijn ook van belang buiten de Nederlab-omgeving, dus de technologie moet erin voorzien dat de metadata zo openbaar worden gemaakt dat andere gebruikers uit verschillende gebieden het materiaal kunnen vinden dat relevant is voor hun onderzoek. Het CLARIN-project werkt met een flexibel metadata-schema (CMDI) dat de flexibiliteit verschaft om willekeurige metadata-beschrijvingen te produceren die semantisch interoperabel zijn, uitgaande van de ISOcat Data Category Registry. De CLARIN-infrastructuur verschaft ook het noodzakelijke technologische kader waarmee grootschalige metadata kunnen worden gepubliceerd, met gebruikmaking van OAI-PMH. Nederlab gaat gebruikmaken van dit CMDI-schema voor het creëren van metadata-beschrijvingen, en zal redactietools leveren die aangepast zijn voor het gebruik als Nederlab-metadataprofilen, zodat het materiaal van het diachrone corpus met OAI-PMH kan worden 'geharvest'. Dit garandeert dat de voor het corpus gecreëerde data en metadata in gestandaardiseerde vorm beschikbaar komen, terwijl dankzij de samenwerking met CLARIN de informatie ook op Europees niveau beschikbaar komt.

Procedure voor het opnemen van onderzoeksdata in het corpus

Nederlab zal een procedure opzetten voor het opnemen van onderzoeksdata in het diachrone corpus. Iedere onderzoeker kan nieuwe data voorstellen om in het corpus opgenomen te worden, of hij kan wijzigingen voorstellen voor het bestaande corpus of de metadata. Een dergelijk verzoek dient goedgekeurd en geaccepteerd te worden door het redactieteam en de Raad van Bestuur. Elk verzoek moet vergezeld gaan van een beschrijving van de relevantie voor het corpus, en het moet het subcorpus bevatten waaraan de nieuwe data moeten worden toegevoegd.

Nederlab zal voldoende feedback verstrekken over de status van de voorgestelde onderzoeksdata tijdens de hele procedure van voorstellen, evalueren en goedkeuren, en zal de nodige autoriseringsprocedures faciliteren. Het redactieteam krijgt een werkomgeving die de evaluatie- en validatieprocedures adequaat ondersteunt. Als de validatie is doorlopen en goedkeuring is verkregen, worden de voorgestelde onderzoeksdata toegevoegd aan het corpus en gearchiveerd bij de deelnemende archiveringscentra.

Zoeken

Nederlab zal onderzoekers de mogelijkheid bieden te zoeken in het hele diachrone corpus en in de vrituele werkruimten van onderzoekers. Er worden twee zoekdomeinen onderscheiden: binnen metadata en binnen data, maar een eindgebruiker kan beide ook gezamenlijk doorzoeken. Het metadata-zoekdomein omvat alle metadata-documenten uit het corpus. Deze worden opgeslagen in een of meer CLARIN-centra die zorgen voor de publicatie van de metadata in de grotere CLARIN- infrastructuur. Dit garandeert dat de metadata kunnen worden 'geharvest'

en dat ze beschikbaar komen via het centrale Virtual Language Observatory van CLARIN[4] en vergelijkbare zoekmachines. Bovendien zullen de metadatadocumenten in werkruimten van onderzoekers die door onderzoekers openbaar zijn gemaakt, centraal verzameld worden en binnen Nederlab beschikbaar gemaakt met de technologie van CLARIN, zodat er een metadata-zoekdomein ontstaat voor alle binnen Nederlab gepubliceerde metadatadocumenten. Iedere gebruiker zal bovendien kunnen zoeken binnen de metadatadocumenten in de eigen werkruimte en in werkruimten van andere onderzoekers waartoe zij toegang hebben gekregen. De hierbij gecreëerde zoekindexen kunnen in de werkruimte van de onderzoeker worden opgeslagen. Op die manier beslaat het metadata-zoekdomein drie gebieden: corpus-metadata, openbare metadata en particuliere/met anderen gedeelde metadata.

Voor het doorzoeken van data in de werkruimten wordt de situatie bemoeilijkt doordat deze data nog geen deel uitmaken van een instellingsarchief en daarom niet toegankelijk zijn voor een zoekmachine. Het is belangrijk om hiervoor een oplossing voor Nederlab-gebruikers te vinden. Een mogelijkheid die nog verder uitgewerkt moet worden, is om verschillende soorten bronnen te linken aan de verschillende kwaliteitslagen die aan het diachrone corpus worden toegekend. Nederlab kan per formaatsoort speciale indexeringsstrategieën volgen die toegespitste toegang bieden tot verschillende verrijkingsniveaus, zoals morfosyntactische of syntactische informatie. Dit zal bevorderen dat onderzoekers standaardformaten voor hun bronnen gaan hanteren in de verschillende werkfasen, wat bijdraagt aan homogenisering van het corpus. De zoekindexen voor iedere werkruimte kunnen worden opgeslagen in de werkruimte van de onderzoeker. Ook het data-zoekdomein beslaat drie gebieden: corpusdata, openbare Nederlab-data en particuliere/met anderen gedeelde data.

Ter ondersteuning van de eindgebruiker moet bijzondere aandacht geschonken worden aan het ontwerp van de gebruikersinterface, zodat de eindgebruiker de weg wordt gewezen bij het kiezen uit de verschillende beschikbare zoekopties.

Tools

Nederlab zal een serie standaardtools leveren die door onderzoekers gebruikt kunnen worden voor het bewerken van delen van het corpus of data in de werkruimte. Er wordt een eenvoudige, gebruikersvriendelijke XML-editor geboden om nieuw materiaal te creëren en bestaand materiaal te annoteren. Nederlab zal standaardtools opnemen die beschikbaar worden gesteld door projecten zoals CLARIN-NL en CATCHPlus. Waar nodig en waar mogelijk zullen aanpassingen gemaakt worden om deze tools geschikt te maken voor relevante tijdsperiodes of specifieke praktijkgevallen. Omdat de meeste tools tegenwoordig als webservices worden geleverd, kunnen ze waarschijnlijk gecombineerd worden tot complexere workflows. Hierdoor komen volledig vastgelegde productieprocessen ter beschikking van de onderzoeker, zoals text-to-named-entities, zonder dat hij veel af hoeft te weten van de werking van de afzonderlijke processen. Tools en processen worden centraal ter beschikking van onderzoekers gesteld; onderzoekers kunnen bladeren en tools/processen kiezen om binnen hun virtuele werkruimte mee te werken. Alle tools en workflows zullen worden beschreven met behulp van CMDI-metadata. Er wordt samengewerkt met de cloud-infrastructure van SARA/BigGrid.

Archivering

Nederlab zal complexe informatiestructuren leveren in de vorm van datastructuren, formaten en dataversies. Versiebeheer en archivering verdienen speciale aandacht, zodat de data van het diachrone corpus in de toekomst volledig beschikbaar blijven. Nederlab zal een NWO-DANS-datacontract met DANS afsluiten om duurzame kwaliteit en toegankelijkheid van de data te garanderen in een opslagplaats die het Data Seal of Approval heeft gekregen. Het archiveren van verrijkte data en het databeheer voor lange termijn worden verzorgd door The Language Archive (TLA) van het MPI, als onderdeel van de samenwerkingsovereenkomst tussen TLA en KNAW. Adviezen over archiveren zullen worden ingewonnen bij DANS en SURF. Er zullen persistent identifiers worden gecreëerd voor verschillende data- en metadata-versies, waardoor deze uniek identificeerbaar en raadpleegbaar worden. Metadata

[4] http://catalog.clarin.eu/ds/vlo/

worden los van de oorspronkelijke bron opgeslagen als afgeleide, verrijkte bron. De broncodes (convertoren, redactie-omgeving, gebruikersinterface enz.) komen beschikbaar als opensource-software.

Harmonisering en standaardisering

Er worden op dit ogenblik zeer veel verschillende formaten gebruikt, vooral in de manier waarop taalkundige annotaties worden toegevoegd. Weliswaar lopen er diverse initiatieven voor de standaardisering van taalkundige annotaties van taaldata, zoals het LAF (Linguistic Annotation Framework; ISO/FDIS 24612) en MAF (Morpho-syntactic Annotation Framework (ISO CD 24615), maar geen daarvan is algemeen geaccepteerd door de onderzoekers of de toolbouwers. Recentelijk hebben Linked Open Data en Open Annotation Collaboration, uitgaande van Linked Open Data, ruime aanhang gekregen. Nederlab is van plan deze benaderingen als uitgangspunt te nemen bij het vastleggen van zowel handmatige als automatische annotaties in het ondersteunde Nederlab-formaat.

Maatschappelijke relevantie

Het diachrone corpus van het Nederlands van de achtste eeuw tot heden vormt gezamenlijk het talige deel van het Nederlandse nationale erfgoed; het is dan ook in het algemene belang dat dit corpus – met een veelheid aan plaats- en tijdgebonden informatie - op één plaats beschikbaar wordt gesteld, niet alleen voor wetenschappers en studenten maar voor iedereen die zich bezighoudt met het Nederlandstalige erfgoed, zowel professionals als het algemene publiek. Zowel de Nederlandse regering als de Europese Unie hechten veel waarde aan digitale duurzaamheid en toegankelijkheid van digitale informatie. Nederlab zal ons digitale geheugen gemakkelijker toegankelijk maken, in overeenstemming met de doelstellingen en actieplannen van de *Digital Agenda* voor Europa (http://ec.europa.eu/information_ society/digital-agenda/index_en.htm). Het corpus kan met enkele kleine ingrepen ook geschikt worden gemaakt voor onderwijstoepassingen, en het kan vanwege de veelheid aan verfijnde metadatering en tagging worden verbonden met eindproducten als woordenboeken, encyclopedieën, navigatiesystemen; dit zal ongetwijfeld ook een impuls geven aan productvernieuwing in de uitgeverij.

Nationale context van de faciliteit

De onderzoeksvragen op het gebied van de geschiedenis van de Nederlandse taal en cultuur die met behulp van Nederlab beantwoord kunnen worden (zie 1. Doelstellingen), sluiten aan op recente ontwikkelingen binnen de geesteswetenschappen en leveren materiaal voor fundamentele theoretische vraagstukken binnen de cognitieprogramma's en eHumanities-programma's van KNAW en NWO. Onderzoekers krijgen met behulp van Nederlab nieuwe inzichten in belangrijke vragen zoals het nature-nurturedebat, de nationale culturele identiteit, culturele integratie, het ontstaan van canons en de verbreiding van kennis, cultuur en taal. Deze onderwerpen staan alle op de Nederlandse wetenschapsagenda die de KNAW in mei 2011 heeft gepresenteerd.

Nederlab is bedoeld als specifiek onderzoeksinstrument. Er wordt voortgebouwd op bestaande technologieën, die voor Nederlab worden aangepast. Als basis van de technische infrastructuur zal Nederlab de technologieën benutten die beschikbaar worden gesteld door de grotere (inter)nationale infrastructuurprogramma's CLARIN en DARIAH (inclusief het nieuwe roadmap-voorstel CLARIAH) en die uitgaan van internationale standaards voor datamanagement en infrastructuur (zie Appendix 3 voor een overzicht van de Nederlandse eHumanities-infrastructuur). De virtuele gebruikersomgeving van Nederlab, die speciaal is gericht op de onderzoeksgemeenschap, wordt gebouwd op deze generieke infrastructuur. Voor het onderzoek zal Nederlab tools aanbieden die zo zijn ontwikkeld dat ze gebruikt kunnen worden binnen de CLARIN-DARIAH infrastructuur. Tot slot zal Nederlab een diachroon corpus voor onderzoek aanbieden. De data die tezamen dit diachrone corpus vormen, worden geleverd door grote dataleveranciers zoals de Koninklijke Bibliotheek, de universiteitsbibliotheken, DBNL, INL, die zich alle bereid hebben verklaard het Nederlab te steunen. Binnen Nederlab zullen de data en metadata van de dataleveranciers aan elkaar worden gelinkt en gedistribueerd doorzoekbaar gemaakt. Op die manier kunnen onderzoekers corpora tegelijkertijd analyseren die zich op verschillende websites bevinden, waardoor de kwaliteit van de data bij de verschillende dataleveranciers wordt verhoogd.

Nederlab neemt zo een speciale en unieke positie in tussen de grote (inter)nationale infrastructuurprogramma's en grote dataleveranciers. Deze unieke positie maakt ons mogelijk specifieke applicaties te ontwikkelen voor het zoeken en organiseren van Nederlandse teksten, in samenwerking met zowel de internationale infrastructuur die momenteel wordt gebouwd als de dataleveranciers, die steeds meer teksten digitaal beschikbaar maken. De vijf toekomstige Nederlandse CLARIN-centra (DANS, Huygens ING, INL, MI, MPI) hebben een cruciale rol in Nederlab.

Nederlab vormt het eerste gemeenschappelijke platform voor geesteswetenschappers, corpusaanbieders en technici. Onderzoekers, studenten en geïnteresseerde leken kunnen via Nederlab vanuit één centraal punt alle bestaande diachrone corpora tegelijkertijd doorzoeken. Een belangrijk effect van de oprichting van Nederlab is dat alle betrokken partijen komen tot afspraken over standaardisering en harmonisering. Daartoe werkt Nederlab samen met Nederlandse toolontwikkelaars, zoals de door NWO-gesubsidieerde projecten Catch en CatchPlus om het culturele erfgoed digitaal te ontsluiten, het Europese IMPACT-project dat zich bezighoudt met de verbetering van de optische tekenherkenning (ocr), en ICT'ers en taaltechnologen van alle Nederlandse universiteiten.

Nederlab zal een dynamische vorm krijgen, en uiteraard uitgaan van open access en open source. Voor de teksten waarop copyright of IPR berust, zullen algemene afspraken worden gemaakt met de rechthebbenden en hun vertegenwoordigers zodat de data voor onderzoeksanalyses beschikbaar komen, eventueel via een (door instellingen) betaald abonnement/licenties.

Nederlab wil een belangrijke educatieve functie voor onderzoekers en studenten gaan vervullen. Daarom zal er tijdens de subsidieperiode veel aandacht worden besteed aan de disseminatie van informatie. Een eerste versie van Nederlab zal zo snel mogelijk worden opgeleverd, binnen een jaar na het begin van de subsidieperiode. Daarbij wordt een educatief programma geschreven. Er worden eenvoudige tools aangeboden, 'tools light' , met vaste basisfuncties die vervolgens door de onderzoekers gelaagd uitgebreid en gefinetuned kunnen worden. Ook worden er concrete voorbeelden gegeven van de werking van de tools in de vorm van workflows, waarvan de verschillende tussenstappen bewaard kunnen worden - die kunnen dan weer door andere onderzoekers worden gebruikt.

Tijdens de duur van het project wordt een helpdesk ingericht. Voorts komt er een digitaal forum waar onderzoekers met elkaar kunnen overleggen. Er zullen op alle instellingen masterclasses worden georganiseerd, en jaarlijkse summer schools. Op deze manier wordt gegarandeerd dat onderzoekers, studenten, promovendi en postdocs optimaal gebruik gaan maken van Nederlab. Door de veelvuldige contacten met onderzoekers wordt ook duidelijk wat de exacte wensen zijn en waar knelpunten of moeilijkheden voor de onderzoekers zitten; die kunnen in interactie met de onderzoekers tijdens de subsidieperiode worden opgelost.

Relatie tot andere onderzoeksgroepen

Uit de organisatiestructuur blijkt dat Nederlab wordt gesteund door een groot aantal Nederlandse universiteiten en KNAW-instituten (zie ook Appendix 1).

Organizatiestructuur

Nederlab wordt geleid door een **Raad van Bestuur (RvB)**, gesteund door een nationale coördinator. De RvB bestaat uit vertegenwoordigers van de drie instituten die garant staan voor duurzame hosting, onderhoud en beheer: het Meertens Institute (prof. dr. H.J. Bennis), Huygens ING (dr. H. Wals), DBNL (C.A. Klapwijk), en de projectleider (dr. N. van der Sijs). De RvB is verantwoordelijk voor het dagelijks bestuur van het project. De leden van de RvB zullen toezicht houden op de lopende activiteiten door regelmatig overleg met de supervisoren van de deelprojecten en de voorzitters van de adviesraden, die, samen met de RvB, het **Algemeen Bestuur (AB)** vormen. De AB bepaalt het jaarlijkse budget en werkplan, en vergadert daarover eenmaal per jaar.

De activiteiten betreffen het onderzoek, de techniek en de data (voor details zie het managementplan): **Deelproject 1**: wetenschappelijke inbedding: supervisie Meertens Instituut/Utrecht University (prof. dr. L.C.J. Barbiers) en Uva (prof. dr. J.C. Kennedy);

Deelproject 2a: infrastructuur: supervisie Meertens Instituut (gebruikersinterface, workflow, algemene technische coördinatie van Nederlab; ir. M. Kemps-Snijders) en Huygens ING (virtuele werkruimte; ing. R. Haentjes Dekker);
Deelproject 2b: toolsaanpassing: supervisie Radboud University (prof. dr. A.P.J. van den Bosch) en Groningen University (prof. dr. ir. J. Nerbonne);
Deelproject 3: datacuratie: supervisie DBNL (teksten; dr. R. van Stipriaan) en INL (lexicale data; lic. K. Depuydt).

Voorts zijn er vier adviesraden:
1. **Wetenschappelijke adviesraad** (WAR): prof. dr. A.M.C. van Kemenade (RUN, voorzitter); prof. dr. G.J. Dorleijn, prof. dr. J. Hoeksema, prof. dr. B.A.M. Ramakers (RUG); prof. dr. R.A.M. Aerts, prof. dr. J.B. Oosterman, dr. M. Rem (RUN); prof. dr. R. Bod, prof. dr. J.T. Leerssen, prof. dr. T.L. Vaessens, prof. dr. F.P. Weerman (UvA); prof. dr. J.L. Goedegebuure, prof. dr. T. van Haaften, prof. dr. H. te Velde, prof. dr. A. Verhagen, prof. dr. M.J. van der Wal (UL); prof. dr. G. Buelens, prof. dr. W.W. Mijnhardt, prof. dr. E. Stronks (UU); prof. dr. I. Leemans, prof. dr. B.J. Peperkamp, prof. dr. G.J. Steen (VU); dr. K.H. van Dalen-Oskam (Huygens ING); dr. G.J. Postma (Meertens Institute); prof. dr. A.P. Versloot (Fryske Akademy).
2. **Technische adviesraad** (TAR): prof. dr. F.M.G. de Jong (UT, voorzitter); prof. dr. G.J.M. van Noord, prof. dr. L.R.B. Schomaker (RUG); dr. N.H.J. Oostdijk (RUN); dr. M. Bouwhuis (SARA); dr. M.W.C. Reynaert (TU); prof. dr. A.P. de Vries (TUD); dr. ir. J. Kamps, dr. M. Marx (UvA); dr. J. de Kruif, prof. dr. J.E.J.M. Odijk (UU); prof. dr. P.T.J.M. Vossen (VU); dr. P. Doorn (DANS); drs. J.J. van Zundert (Huygens ING); dr. J. de Does (INL); ir. D. Broeder (MPI).
3. **Corpusadviesraad** (CAR): dr. J. Beeken (INL, voorzitter), dr. W. van Bergen (UBL), dr. C. Cucchiarini (Nederlandse Taalunie), drs. P. Doorenbosch (KB), drs. M. de Niet (DEN), drs. J.F. Oomen (Beeld en Geluid), drs. M. Slabbertje (UBU), dr. B. Zeeman (UvA).
4. **Internationale adviesraad** (IAR): prof. dr. F. Willaert (Antwerp, voorzitter), drs. T. Roselaar (IVN); prof. dr. J. van Keymeulen (Ghent), prof. dr. E. Leijnse (Namen), prof. dr. J. Tollebeek (Louvain), prof. dr. W. Vandenbussche (Brussels), lic. E. Vanhoutte (Ghent); prof. dr. R. Grüttemeier (Oldenburg), prof. dr. M. Hüning (Berlin); prof. dr. J. Pekelder (Paris); dr. R. Vismans (Sheffield); dr. R. de Bies (Paramaribo), dr. K. Groeneboer (Jakarta), prof. dr. R.B. Howell (Madison WI, VS), prof. dr. D. Prinsloo (Pretoria, South Africa), prof. dr. R. Severing (Curaçao).

De wetenschappelijke, technische en corpusadviesraden komen eenmaal per jaar bijeen (of vaker indien nodig) en geven gevraagd en ongevraagd advies. De TAR adviseert over activiteiten in Deelproject 2a (infrastructuur) en Deelproject 2b (toolsaanpassing). Leden van de TAR testen de infrastructuur en de tools. De WAR adviseert over de keuzes die in Deelproject 1 (wetenschappelijke inbedding) moeten worden gemaakt. The pilotprojecten die worden ingediend om de infrastructuur, tools en data te testen en aan te vullen, worden beoordeeld door leden van de WAR en/of TAR, in overleg met de supervisoren van Deelproject 1 en de voorzitter van de WAR. De CAR zal worden geconsulteerd over zaken die te maken hebben met Deelproject 3 (datacuratie). Tot slot zal de Internationale Adviesraad erop toezien dat het diachrone corpus ook Nederlandstalige teksten bevat die zijn geschreven buiten Nederland en Vlaanderen, en dat het Nederlandstalige diachrone corpus gelinkt kan worden aan diachrone corpora van andere, vooral naburige, landen, zodat het mogelijk wordt onderlinge contacten en invloeden binnen West-Europa te onderzoeken.

Internationale context

Het diachrone corpus van Nederlab is ook belangrijk voor buitenlandse onderzoekers, waarmee nauw wordt samengewerkt. De onderhavige aanvraag wordt gesteund door de Nederlandse Taalunie, de Internationale Vereniging voor Neerlandistiek (IVN) – beide in adviesraden vertegenwoordigd - en door CLARIN-ERIC. De Vlaamse geesteswetenschappers zullen de aanvraag voor Nederlab concreet steunen door het indienen van een

(middelzware) aanvraag bij de Herculesstichting voor de samenstelling van een corpus Nederlandstalige egodocumenten uit de 15e en 16e eeuw. Dankzij deze Vlaamse aanvraag zal de lacune die het digitale diachrone corpus voor deze periode vertoont, worden gedicht.

Inhoudelijk is het diachrone corpus interessant voor buitenlandse onderzoekers omdat het niet alleen Nederlandstalige teksten uit Nederland en Vlaanderen bevat, maar ook Surinaams-, Antilliaans-, Amerikaans- en Indisch-Nederlandse teksten. Ook wordt er samengewerkt met Zuid-Afrika, waar men bezig is met de samenstelling van een diachroon corpus van het Afrikaans.

De inrichting van Nederlab kan exemplarisch worden voor buitenlands onderzoek. In geen enkel land bestaat een met Nederlab vergelijkbare infrastructuur. Wel loopt het Engelse taalgebied duidelijk voorop in de aanbieding van goed doorzoekbare en taalkundig verrijkte diachrone corpora (zie bv. http://www.helsinki.fi/varieng/CoRD/index.html), en in de oprichting van The Stanford Literary Lab door Franco Moretti, waar voor de letterkunde baanbrekend werk wordt verricht. Initiatieven in andere landen zijn veel kleinschaliger: in Frankrijk is een diachroon corpus voor abonnees beschikbaar (http://www.frantext.fr/). In Duitsland loopt een initiatief om een diachroon corpus voor de complete historische periode samen te stellen. Overige Europese landen kennen ook allerlei initiatieven: een overzicht van synchrone en diachrone corpora die voor de verschillende talen beschikbaar zijn, is te vinden op http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/links-en/korpora_links.

Het opvallendste verschil met Nederlab is dat alle diachrone corpora in het buitenland bestaan uit een selectie van de beschikbare teksten uit de verschillende perioden. Binnen Nederlab worden echter *alle* historische teksten aangeboden die digitaal beschikbaar zijn of komen. Dit is op uitdrukkelijk verzoek van de Nederlandse onderzoeksgemeenschap, die unaniem heeft laten weten dat er geen restricties aan het corpus moeten worden opgelegd, wil het tegemoetkomen aan alle gemeenschappelijk onderzoeksvragen.

Wanneer Nederlab in 2013 van start gaat, kan het dus het model worden voor buitenlandse onderzoeksinstituten, met als bijkomend voordeel dat buitenlandse diachrone corpora automatisch aansluiten op het Nederlandse, zodat het diachrone onderzoek een internationale dimensie kan krijgen: taalgrenzen en politieke grenzen vallen lang niet altijd samen en het Nederlands moet zeker diachroon worden gezien als een continuüm van dialecten dat over de landsgrenzen heen gaat. Veel historische en letterkundige verschijnselen zijn bovendien in het buitenland begonnen en hebben zich van daaruit verbreid naar de Lage Landen, waarna sommige weer zijn 'geëxporteerd'.

Locale context

In de strategische agenda van de KNAW voor 2010-2015 (*Voor de wetenschap. De Akademie in de kennissamenleving*, Amsterdam 2010, p. 23) wordt als belangrijk doel genoemd de inzet van nieuwe informatietechnologieën voor de modernisering van de geesteswetenschappelijke instituten en wordt de ambitie van de Academie uitgesproken 'om met haar instituten een initiërende en sturende rol te spelen in het creëren en onderhouden van infrastructurele ict-voorzieningen en de toepassing van deze voorzieningen in het wetenschappelijk onderzoek'. Om aan deze ambitie vorm te geven zijn de twee KNAW-instellingen Meertens Instituut (penvoerder) en Huygens ING de trekkende krachten achter de oprichting van Nederlab. Genoemde twee instituten garanderen, samen met DBNL/Nederlandse Taalunie, gezamenlijk het onderhoud van Nederlab ook na de subsidieperiode, en ze verschaffen serverruimte. De helpdesk die in de projectperiode wordt opgezet, zal ook daarna nog worden onderhouden. Binnen de webinterface van Nederlab zullen de beschikbare diachrone corpora aan elkaar worden gelinkt op tekst- en metadata-niveau, waardoor niet alleen de bestaande corpora beter toegankelijk worden, maar ook de kwaliteit en standaardisering van de data in de metadata worden verbeterd. Het aan elkaar linken en de kwaliteitsverbetering zijn blijvende resultaten, die met behulp van persistent identifiers worden bewaard. Archivering van verrijkte data en het datamanagement voor de lange termijn worden geregeld via The Language Archive (TLA) van het MPI, als onderdeel van de samenwerkingsovereenkomst tussen TLA en KNAW.

Budget

Gevraagde NWO-subsidie Beschrijving	2013	2014	2015	2016	2017	**Totaal**
Deelproject 0						
Projectleider (0.8 fte)	47,2	48,1	49,1	50,0	72,2	266,6
Helpdesk/assistentie (0.8 fte)	42,3	43,1	43,9	44,8	64,7	238,8
Servers & data-archivering	15,9	16,4	23,7	23,7	31,0	110,7
Deelproject 1						
3 pilotstudies besloten oproep		90,0				90,0
6 pilotstudies open oproep			90,0	90,0		180,0
Internationale workshop			15,0			15,0
Geëditeerde boekpublicatie					5,0	5,0
Cursussen en kennisverspreiding					5,0	5,0
Deelproject 2a						
Software-ontwikkelaar (4.6/2.6/2.5/1.9/1.9 fte)	301,4	167,6	174,1	127,2	183,7	954,0
Deelproject 2b						
Software-ontwikkelaar (0.5 fte)	31,6	32,2	32,8	33,5	48,3	178,4
Postdocs (0.5/0.5/0.5/0.5/0.5 en 0.5/0.5/0.5/0.5/0.0 fte)	59,0	60,2	61,4	73,0	45,1	298,7
Internationale workshop					15,0	15,0
Publicatie (boek & software)					5,0	5,0
Deelproject 3						
Editor (0.2/0.2/0.1/0.1/0.1 fte)	12,6	12,9	6,6	6,7	9,7	48,5
Seniorredacteur (0.5/0.5/0.0/0.0/0.0 fte)	26,4	31,4				57,8
Redacteurs (1.8/1.8/0.8/0.8/0.8 fte)	79,2	88,1	36,5	37,2	53,8	294,8
Postdoc (0.8 fte)	47,2	48,1	49,1	50,0	72,2	266,6
Software-ontwikkelaar (0.4 fte)	25,3	25,8	26,3	26,8	38,7	142,9
Opwerken databestanden	40,0	50,0	50,0	50,0	30,0	220,0
Total	728,1	713,9	658,5	612,9	679,4	3392,8

Matching Beschrijving	2013	2014	2015	2016	2017	**Totaal**	**Door**
Raad van Bestuur (0.3 fte)	36,3	36,3	36,3	36,3	36,3	181,5*	MI, Huygens, DBNL
Supervisoren deelprojecten (0.8 fte)	80,1	80,1	80,1	80,1	80,1	400,5*	Universiteiten
Totaal	116,4	116,4	116,4	116,4	116,4	582,0*	

*Geschatte reëele kosten (volgens NWO-standaarden, salarisschaal 11, zou het totaalbedrag komen op 366,7).

Uitwerking van het aangevraagde budget

Nederlab: Werkplan 2013-2017

Deelproject	Beschrijving
Deelproject 0	**Algemene organisatie**
Supervisie: Raad van Bestuur: prof. dr. H.J. Bennis, C.A. Klapwijk, dr. N. van der Sijs, dr. H. Wals	
Projectleider	Projectleiding, coördineren van de werkzaamheden van de vier deelprojecten, disseminatie van informatie onder onderzoekers, schrijven van educatief programma (0.8)
Assistent	Assistentie t.b.v. projectleiding en praktische organisatorische zaken als helpdesk, pr, masterclasses/workshops/zomerscholen Ondersteuning van projectleider en supervisoren van deelprojecten (0.8)
Deelproject 1	**Wetenschappelijke inbedding**
Supervisie: prof. dr. L.C.J. Barbiers, prof. dr. J. Kennedy	
	Uitzenden besloten oproep voor drie pilotstudies: geschiedenis, letterkunde, taalkunde (2013) Evaluatie van de besloten oproep Supervisie van drie pilotstudies; coördinatie van de feedback naar de software-ontwikkelaars; evaluatie van pilotstudies; advies over publicaties op basis van de projecten (2014) Uitzenden open oproep 1 Evaluatie en selectie van drie projectvoorstellen van open oproep 1 (drie disciplines) Supervisie van drie pilotstudies; coördinatie van de feedback naar de software-ontwikkelaars; evaluatie van pilotstudies; advies over publicaties op basis van de projecten Workshop: wetenschappers tonen de resultaten geboekt m.b.v. de infrastructuur (2015) Uitzenden open oproep 2 Evaluatie en selectie van drie projectvoorstellen van open oproep 2 (drie disciplines) Supervisie van drie pilotstudies; coördinatie van de feedback naar de software-ontwikkelaars; formuleren van de eisen voor de infrastructuur; advies over publicaties n.a.v. oproep 2 (2016) Coördinatie van de feedback van de wetenschappelijke projecten Boekpublicatie over het project, samen met de software-ontwikkelaars Organisatie van cursussen, colleges en een afsluitende workshop (2017)
Deelproject 2a	**Infrastructuur**
Supervisie: ir. M. Kemps-Snijders, ing. R. Haentjes Dekker	
Software-ontwikkelaars	Harmonisatie van dataformaten en standaarden om bestaande dataformaten voor verschillende annotatielagen te synchroniseren en te linken aan bestaande standaarden (1.1 fte) Indexeringsmodule om het indexeren van metadata en data. Metadata en data moeten worden geïndexeerd op corpus-, document- en werkruimteniveau (1.1 fte) Werkruimtemodule voor het beheren van toegang tot de individuele virtuele werkruimten en het individuele databeheer (2.3 fte) Module voor gebruikersbeheer en autorisatie voor gemeenschappelijk gebruik (0.5 fte) Autorisatiemodule om autorisatie van informatie over metadata en data-bronnen te beheren (0.5 fte) Zoek/Blader-module voor het zoeken en bladeren door eerder gecreëerde indexen (2.7 fte) Metadata-editors voor verschillende profielen in het diachrone corpus (0.4 fte) Metadata-publicatie voor publicatie van openbare metadata en data-documenten (0.2 fte) Toolsintegratie: het integreren van verschillende tools (2.2 fte) Implementeren van een algemene gebruikersinterface voor Nederlab (2.5 fte)

Deelproject 2b	**Toolsaanpassing**
Supervisie: prof. dr. A.P.J. van den Bosch, prof. dr. ir. J. Nerbonne	
Software-ontwikkelaar + postdocs	Integratie van algemene tools voor tekstnormalisatie (OCR, spellingvariatie), met focus op integratie en interoperabiliteit* (0.5 + 1.5 fte)
	Integratie van tools voor lemmatisering, morfologie, PoS tagging (0.5 + 1.0 fte)
	Integratie van tools voor syntactische analyse (0.5 + 1.0 fte)
	Overige NLP-integratie (0.5 + 1.0 fte)
	Onderhoud, verbetering, bijstellingen (0.5 fte)
	Organisatie van twee 'shared task'-sessies op CLIN of een internationaal congres (taak postdocs)

Deelproject 3	**Datacuratie**
Supervisie: dr. R. van Stipriaan, lic. K. Depuydt	
Editor	Ontwerp van layout van 'xml-containers' & beheersomgeving° (0,2 fte)
	Kwaliteitsbeoordeling (0.5 fte)
Senior-redacteur	Implementatie van beheersomgeving in DBNL-workflow (0.2 fte)
	Planning & controle van het proces van opwerken van (meta)data (0.4 fte)
	Kwaliteitsbeoordeling (0.4 fte)
Redacteurs	Opwerken van corpora (zie Appendix 2 voor een overzicht van te bewerken materiaal) (2.5 fte)
	Opwerken van metadata (zie Appendix 2 voor een overzicht) (2.5 fte)
	Linken aan biografische data (zie Appendix 2 voor een overzicht) (1 fte)
postdoc + software-ontwikkelaar	Curatie van Oudnederlands Corpus, Corpus van cd-rom Middelnederlands (1.0 + 0.5 fte)
	Opwerken van lexicons: integratie van historische lexicografische bronnen in één diachroon lexicon en integratie met de inhoud van een modern lexicon (3.0 + 1.5 fte)

*De integratie van bestaande tools in de infastructuur wordt zoveel mogelijk naar voren geschoven, teneinde te voorkomen dat tools eerst ontwikkeld worden (of dat bestaande tools aangepast worden voor historische corpora) en dat pas later, tijdens de integratie, blijkt dat ze niet compatibel zijn. Voorts is het belangrijk om vanaf het begin het wetenschappelijk gebruik (Deelproject 1) te bevorderen, en de infrastructuur te testen met concrete casussen, data en gebruikers. Uit ervaring is gebleken dat bij daadwerkelijk gebruik vaak de vraag naar speciale aanpassingen of de toevoeging van uitgebreidere (algemene) analysemogelijkheden opkomt, zoals voorzieningen voor speciale statistische analyses, visualisaties of kaarten. In het algemeen zijn we niet van plan deze binnen Nederlab te ontwikkelen, maar waar nodig zullen we ze invoeren uit bestaande open bronnen (b.v. uit het R-statistiekpakket). Bovendien hebben wij het werk dat nodig is ter ondersteuning van letterkundige en historische analyses (tekstopschonen, normaliseren, woordenschatanalyses, concordanties, annotaties van diverse aard en toegankelijkheid, direct beschikbare lemmatisering en part-of-speech tagging) in de eerste twee jaar geconcentreerd, en we hebben werk dat geavanceerdere tools vereist (ingewikkeldere taggers, lemmatiseerders, named entity recognition, sentimentanalyse, georeferencing, syntactische parsing) na het eerste jaar gepland. Op die manier hopen we van begin af aan meer historici en letterkundigen bij het project te betrekken, en zo ook meer belangstelling voor de geavanceerdere tools te wekken.

° Om een uniformerende laag van noodzakelijke metadata voor het diachrone corpus te creëren, zal er binnen de bestaande werkomgeving van DBNL een beheersomgeving worden gecreëerd waarin de aangeleverde metadata worden verwerkt en de toe te voegen gedigitaliseerde data op kwaliteit worden beoordeeld op basis van documentatie of door persoonlijke observatie. Met behulp van nieuwgevormde 'xml-containers' zullen de metadata gelinkt worden met bronnen die gedistribueerd zijn opgeslagen; digitaal beheer van de basismetadata geschiedt in de DNBL-omgeving. De data worden op vast te stellen tijden gesynchroniseerd met Nederlab.

Tijdens de opbouw van het corpus zullen zoveel mogelijke bibliografische en biografische metadata worden geïntegreerd, afkomstig uit o.a STCN, CBK, de krantenbank van de Koninklijke Bibliotheek, DBNL en INL. Bestanden kunnen worden gevonden, doorzocht en verrijkt met behulp van de xml-containers; voorts zullen voor het onderzoek onmisbare databestanden worden opgewerkt, zodat ze voldoen aan de minimumeisen voor betrouwbaar wetenschappelijk onderzoek.

Exploitatie en overige kosten

De kosten die nodig zijn om de infrastructuur na de projectperiode in de lucht te houden, worden gedragen door de drie constituerende instellingen: het Meertens Instituut, Huygens ING en DBNL/Nederlandse Taalunie, zoals vermeld onder de kop 'Locale context' hierboven. Deze instellingen garanderen serverruimte voor Nederlab ook nadat de subsidieperiode is verstreken.

Duur van het project

Geplande startdatum	01-01-2013
Verwachte einddatum	31-12-2017

De faciliteit zal enkele decennia in stand blijven, omdat ze is gebouwd op corpora en (technische) kennis die afkomstig zijn van sterke instellingen als universiteitsbibliotheken en (KNAW-)onderzoeksinstituten. De virtuele gebruikersomgeving van Nederlab zal, zowel in de uitvoeringsfase als in de exploitatiefase, profijt trekken van toevoegingen die de verschillende dataleveranciers aan de corpora doen, doordat het diachrone corpus dynamisch is opgezet. Bovendien zal Nederlab profiteren van technische vernieuwingen die in de uitvoeringsfase worden gedaan. De KNAW-instituten zullen, met hun geavanceerde technologische afdelingen, garanderen dat de tools binnen Nederlab up-to-date blijven.

Managementplan

De tabel hieronder geeft een overzicht van de belangrijkste mijlpalen en deliverables in de uitvoeringsfase. Hierna zal de exploitatie worden gegarandeerd door de constituerende instellingen, zie boven.

Nederlab: Mijlpalen en deliverables

Deelproject		Beschrijving van deliverables
Deelpr. 0		
	a	Eduactief programma en kennisverspreiding
	b	Masterclasses, workshops, zomerscholen
Deelpr. 1		
	a	Projectoproepen en beoordelingen
	b	Geëditeerd boek
	c	Internationale workshop
	d	Colleges en verdere verbreiding van projectresultaten
Deelpr. 2a		
	a	Algemene gebruikersinterface (eerste jaar: Demonstrator)
	b	Indexeringsmodule
	c	Module voor individuele virtuele werkruimten
	d	Module voor gebruikersbeheer en autorisatie voor gemeenschappelijk gebruik
	e	Autorisatiemodule
	f	Zoek/Blader-module, o.a. metadataharvester

	g	Metadata-editors
	h	Module voor de publicatie van metadata
Deelpr. 2b		
	a	(eerste jaar) Nederlab Tools 0.1, implementeren normalisatietools (b.v. TICCLops): opgezet als webservice (b.v. via CLAM); kan toegepast worden op alle corpusdata die converteerbaar zijn naar data
	b	Nederlab Tools 0.2, met diachrone normalisatie
	c	Technische rapporten met beschrijving van de normalisatietools
	d	Nederlab Tools 0.3, met diachrone lemmatisering, generieke en periode-specificieke lemmatisering, morfologische analyse (b.v. MBMA), en PoS-tagging (b.v. Adelheid)
	e	Technische rapporten met beschrijving van de lemmatisering, morfologische analyse en PoS-tagging
	f	Nederlab Tools 0.4, met diachrone syntactische analyse (b.v. INPOLDER of ALPINO) en integratie van andere NLP-tools: coreferentie, NER, georeferencing, sentiment, shared task outcome
	g	Technische rapporten met beschrijving van de syntactische analyse
	h	Technische rapporten met beschrijving van de integratie met andere tools
	i	Software-release: Nederlab Tools 1.0, en alle modules die de kwaliteitstest doorstaan
	j	Release van algemene documentatie over Nederlab Tools: naslaggids, algehele schattingen van de empirische prestaties
	k	Twee 'shared task'-sessies op CLIN of een internationaal congres
Deelpr. 3		
	a	(eerste jaar) Integratie van teksten en metadata met hoge kwaliteit in de Nederlab-infrastructuur
	b	Opwerken van corpora van mindere kwaliteit (van wetenschappelijke bibliotheken en instellingen) naar de vereiste standaarden en onderling linken
	c	Opwerken van metadata van mindere kwaliteit naar de vereiste standaarden en onderling linken
	d	Linken van corpora aan biografische gegevens
	e	Gecureerd Oudnederlands Corpus, Corpus van cd-rom Middelnederlands
	f	Geïntegreerd computationeel historisch lexicon van de 6e-20e eeuw

Risico's en onvoorziene gebeurtenissen

De uitvoering van Nederlab met een zeer groot aantal partijen is een ambitieuze onderneming. Drie dragende partijen zullen er gezamenlijk voor instaan dat de werkzaamheden in goede banen worden geleid, ook als één van de andere partijen onverhoopt verstek laat gaan.

Er is sprake van heterogeniteit binnen de data: een fors deel van het diachrone corpus is matig van kwaliteit doordat het bestaat uit slechte ocr. Dit wordt gecompenseerd doordat er enkele representatieve corpora van hoge kwaliteit zijn die direct als kerncorpus binnen de demonstrator worden opgenomen. Het is de ambitie om via datacuratie de matige delen van het corpus geleidelijk op te waarderen. Dat is ook nodig omdat de zwakkere data met tools moeilijk zijn te benaderen. Dit probleem zal onder andere worden opgelost door de tools zo te ontwerpen dat ze de problemen van de zwakkere data omzeilen, en door onderzoekers in staat te stellen om subsets van het corpus samen te stellen waaruit de zwakkere data worden geweerd.

Het tempo van de technische uitvoering van Nederlab wordt bepaald door een combinatie van het beschikbaar komen van de interface en het corpus. Om te zorgen voor een goede voortgang is gekozen voor een modulaire opzet: na een jaar zal een demonstrator met gecorrigeerde teksten beschikbaar komen. De demonstrator zal direct getest worden door onderzoekers (waarvoor pilots worden uitgeschreven). Geleidelijk worden nieuwe componenten aan de demonstrator toegevoegd, waarmee de onderzoekers kunnen experimenteren.

Het is gebleken dat een deel van de onderzoekers moeite heeft de technische mogelijkheden ten volle te benutten. Om de techniek dichter bij de onderzoekers te brengen wordt binnen Nederlab een speciaal educatief programma opgericht.

Aanvullende informatie

Bij deze aanvraag zijn drie appendices gevoegd en twee ondersteuningsbrieven (als pdf):
- Appendix 1. Lijst van onderzoekers die zijn geconsulteerd over de opzet van Nederlab
- Appendix 2. Overzicht van te verwerken data en metadata
- Appendix 3. Overzicht van de infrastructuur- en toolprogramma's die in deze aanvraag zijn vermeld-
Ondersteuningsbrief van dr. Theo Mulder, directeur van de KNAW
- Ondersteuningsbrief van de constituerende partijen: prof. dr. H. Bennis, directeur van het Meertens Instituut; drs. L. van den Bosch, secretaris van de Nederlandse Taalunie; C.A. Klapwijk, directeur van de DBNL; dr. H. Wals, directeur van het Huygens ING

Verklaring en handtekening

Have you requested funding for this research elsewhere?

☒ No

☐ Yes,

Please include details of any additional grants you have requested for this research project

Declaration

☒ By submitting this form through Iris, I declare that I have completed this form truthfully and completely.

prof.dr. H.J. Bennis, Meertens Instituut, Amsterdam