/instituut voor de
Nederlandse taal/

jaarverslag
2018

verslag aangaande de werkzaamheden, gebeurtenissen enz. in het afgelopen jaar; jaarlijks verslag

(Woordenboek der Nederlandsche Taal)

Inhoud

Inleiding

Missie en visie van het instituut
Missie

- Het bevorderen van kennis en gebruik van de Nederlandse taal door het uitvoeren van toepassingsgericht wetenschappelijk onderzoek
- Het stimuleren en coördineren van de wetenschappelijke beschrijving van de Nederlandse woordenschat en grammatica in al zijn facetten door de eeuwen heen.
- De productie, koppeling en ontsluiting van bronnenmateriaal voor het Nederlands in de vorm van historische en eigentijdse corpora, woordenboeken, lexicale digitale databases, grammatica's en de daarbij behorende technologische hulpmiddelen.

Visie
Het Instituut voor de Nederlandse Taal wil een breed toegankelijk wetenschappelijk instituut zijn op het gebied van het Nederlands. Daarnaast wil het instituut een centrale positie innemen in het hele Nederlandse taalgebied (Nederland, Vlaanderen, Suriname en de Antillen) op het vlak van het wetenschappelijk verantwoord ontwikkelen, bewaren en duurzaam beschikbaar stellen van taalmateriaal.

Wat hebben we gerealiseerd in 2018?
Er waren vier vergaderingen van de Raad van Toezicht en een vergadering van de Raad van Advies (zie bijlage 1voor de samenstelling).

De Raad van Toezicht kwam in 2018 bijeen op 21 februari, 30 mei, 12 september en op 10 december. Belangrijke thema's die aan bod kwamen waren personeel, herschikking functieboek, nieuwe accountant, financiële zaken, begroting en beleidsplan.

De Raad van Advies kwam in 2018 bijeen op 14 februari. Thema's die werden besproken waren: de afwikkeling van het Nederlab-project en de daarna volgende disseminatie en hosting, het CLARIN-rapport voor EWI (Vlaanderen), de mogelijke samenwerking met het IVN en wat het INT voor deze vereniging kan betekenen. Verder werd ook de planning van het project "hoger onderwijsterminologie" besproken en de werkzaamheden rond het Surinaams-Nederlands verhelderd. Ten slotte werd de raad gevraagd om advies over welke accenten we leggen in de wetenschapscommunicatie.

Het jaar 2018 begon met een heel belangrijke wending: het Instituut verhuisde op 1 februari van het Matthias de Vrieshof naar het Rapenburg. De volledige campus van de faculteit Geesteswetenschappen van de Universiteit Leiden wordt hertekend in de komende jaren, alle gebouwen worden gerenoveerd en/of vervangen door nieuwbouw. Dat betekent dat de kantoren in het Matthias De Vrieshof in de komende jaren door andere departementen en groepen van de universiteit zullen worden gebruikt. De dienst huisvesting van de Universiteit zocht actief mee naar een geschikte locatie en vond deze in het prachtige historische "Snouck Hurgronjehuis". De verhuizing verliep heel vlot en de medewerkers genieten ten volle van het nieuwe pand. Bedoeling is dat het INT gedurende de volledige looptijd van de verbouwingen op het Rapenburg blijft; pas na de volledige afwerking van alle bouwprojecten keren we terug naar het Matthias de Vrieshof om dan samen met het LUCL (Leids Universitair Centrum voor Linguïstiek) onderdak te vinden.

De ingebruikname van het nieuwe pand zorgde voor meer bekendheid, we hielden een opendeurdag en het pand werd ook opengesteld tijdens de open monumentendagen. We zetten verder in op een algemene bekendmaking van het instituut, zowel in Vlaanderen als in Nederland. Het INT vervult een duidelijke maatschappelijke functie en staat ten dienste van alle taalgebruikers. Er was in 2018 meer aandacht voor bekendmaking naar het bredere publiek en dienstverlening voor docenten en studenten. Getuige daarvan zijn de stagiaires die sinds 2018 bij het INT aan de slag gingen. Verder leverden medewerkers van het INT ook bijdragen aan onderwijs bij de universiteiten van Leiden en Leuven.

Daarnaast is er de blijvende ondersteuning voor onderzoekers en de deelname aan onderzoeksprojecten. De CLARIN-workshops in Vlaanderen werden druk bijgewoond. Door het vervangen van de oude naam "TST-Centrale" door de nieuwe rubriek "Taalmaterialen", die vlotter toegankelijk en doorzoekbaar is, kunnen we heel wat meer gebruikers met alle aspecten van taal- en spraaktechnologie in contact brengen.

In 2018 werd van start gegaan met intensiever overleg met de Taalunie. Er zijn twee vormen van overleg gestart: er is enerzijds "werkvloeroverleg" en anderzijds "directieoverleg".

Bij het werkvloeroverleg, dat twee keer per jaar plaatsvindt, wordt een aantal projecten besproken die door het INT worden uitgevoerd. De projectleiders van het INT en de beleidsmedewerkers van de Taalunie die bij deze projecten betrokken zijn krijgen zo meer voeling met elkaar en kunnen de plannen, prioriteiten en eventuele pijnpunten meteen bespreken. In 2018 was er twee keer werkvloeroverleg, met name op 29 maart en op 16 oktober. Projecten die werden besproken waren: spelling, terminologie, onderwijs Nederlands in het buitenland (samenwerking met IVN), ANS en Taalportaal, Taaladvies.net, vertaalwoordenboeken en de corpora rond eenvoudige taal.

Het directieoverleg richt zich op de bespreking van het beleidsplan, het prioriteren van bepaalde projecten en de financiële zaken: subsidie en geoormerkt projectgeld. Ook de begroting komt hier aan bod. Aan het directieoverleg nemen naast de algemeen secretaris van de NTU en de directeur van het INT ook de financiële verantwoordelijken deel. In 2018 waren er overlegmomenten op 2 oktober en 27 november.

Wat betreft de grote lijnen van de werkzaamheden wordt het meerjarenbeleidsplan gevolgd. De algemene woordenboekprojecten, met name het *Algemeen Nederlands Woordenboek* (ANW) en de historische woordenboeken, worden voortgezet. De vernieuwde versie van de Geïntegreerde Taalbank werd in 2018 gelanceerd. Ook de neologismen krijgen veel aandacht en worden apart verwerkt. Het project "Woordcombinaties" wordt verder uitgewerkt; het is een proeftuin voor nieuwe computationele lexicografische methodes. Het tweede vertaalwoordenboek, deel van een hele collectie vertaalwoordenboeken uitgebracht door de Commissie Lexicografische Vertaalvoorzieningen (CLVV), is online gezet: het betreft de collectie Nederlands/Portugees, Portugees/Nederlands. Wat betreft de terminologie werd er in 2018 ingezet op het HOTNeV-project rond hoger onderwijsterminologie en het beter beschikbaar stellen van terminologische informatie via de nieuwsbrief en de website.

Op vlak van deelnamen aan wetenschappelijke projecten scoorden we uitstekend: niet alleen werd CLARIAH-PLUS goedgekeurd door NWO en betekent dit voor het INT deelname aan verschillende werkpakketten voor de komende jaren, ook werd onze aanvraag voor een Lorentz-workshop goedgekeurd; deze Lorentz-workshop draagt als titel "the Future of Academic Lexicography" en zal ons toelaten met tal van experts uit verschillende disciplines in te gaan op de uitdagingen van de moderne computationele lexicografie met Artifical Intelligence en Big Data. De workshop zelf vindt plaats in de week van 4 tot en met 8 november 2019. Op Europees vlak werd er hard gewerkt aan het ELEXIS-project waarbij het INT verantwoordelijk is voor het werkpakket "Lexicographic data and workflow". In het kader van het Het COST-netwerk enetCollect over computer assisted language learning and crowd sourcing techniques organiseerde het INT een speciale workshop voor twee werkgroepen van dit COST-netwerk in Leiden. Via deze Europese projecten houdt het INT de vinger aan de pols van de allernieuwste methodes op het vlak van computationele lexicografie.

Nederlab werd afgerond medio 2018 en er werd een publieksevenement door INT en het Meertens Instituut samen georganiseerd voor de presentatie van de eindresultaten. Er werd in dit verband ook een beheersovereenkomst afgesloten om het project in de komende jaren draaiend te houden en de gebruikers toegang te geven tot het portaal. Ook hier heeft het INT een belangrijke taak in het verspreiden van de kennis over Nederlab en het stimuleren van onderzoekers om van het portaal gebruik te maken.

Naast de lopende ondersteunende taken zoals financiën, personeelsadministratie etc. is er de nodige aandacht voor wetenschapscommunicatie. Het instituut werkte in 2018 ook weer actief mee aan de Week van het Nederlands, het DRONGO talenfestival, het Jaar in Taal en andere activiteiten.

1. Algemeen Nederlands Woordenboek

Taak en werkwijze
Het *Algemeen Nederlands Woordenboek* (ANW) is een corpusgebaseerd, digitaal woordenboek dat het eigentijdse Nederlands in Nederland, Vlaanderen, Suriname en het Caribisch gebied zo uitgebreid mogelijk beschrijft. Deze beschrijving strekt zich uit over de hele algemene woordenschat en wordt aangevuld met actuele neologismen. In 2018 werd vooral gewerkt aan de beschrijving van substantieven die behoren tot de kern van de Nederlandse woordenschat. Daarnaast zijn gangbare woorden van andere woordsoorten, woorden uit Suriname en het Caribisch gebied en tal van woorden met een beperkt profiel (d.w.z. met alleen informatie over grammaticale gegevens, de woordvorming, de uitspraak e.d.) bewerkt en deels ook online gezet.

Dagelijkse update
Voorheen werd het ANW ieder jaar geactualiseerd door middel van een aantal grote updates, bestaande uit volledige bewerkte artikelen en artikelen met een minimaal betekenisprofiel (de zogeheten "kopjesartikelen"). Met ingang van 22 mei 2018 is daar verandering in gekomen en vindt er dagelijks een update van het ANW plaats, met uitzondering van de weekenden. Daarmee is een langgekoesterde wens in vervulling gegaan, die de redactie in staat stelt om snel gebleken onvolkomenheden te herstellen, direct actuele informatie toe te voegen en onmiddellijk te reageren op suggesties van gebruikers van het woordenboek.

Databewerking
In totaal bewerkte de redactie in het verslagjaar 550 woordenboekartikelen en werden er 470 volledig bewerkte artikelen aan het ANW toegevoegd.
Een belangrijk project dat gerealiseerd werd, was de vervollediging en uniformering van de categorie diernamen. In samenwerking met stagiair Thomas Haga werden alle 2120 diernamen in het woordenboek geanalyseerd, wat leidde – vooral ook met het oog op de zoekmogelijkheden – tot inhoudelijke en stilistische aanpassing van de semagrammen en zo nodig ook van de definities. Een aantal van 126 ontbrekende diernamen werd door Thomas Haga bewerkt en alsnog aan het ANW toegevoegd.

Door Tanneke Schoonheim, Wil de Ruyter en enige tijd ook Thomas Haga is verder gewerkt aan de ontbrekende fonetische schrijfwijzen van het veld "Uitspraak" in het ANW. De transcripties worden nu nog niet direct getoond in het ANW online, maar worden achter de schermen bewerkt in de bewerkingsomgeving lex'it.

Op 23 november 2018 bereikte het ANW de grens van 250.000 opgenomen woorden. Aan het eind van het jaar telde het ANW 252.026 woorden, behandeld in 74.861 trefwoorden. Die trefwoorden bevatten in totaal 42.286 betekenissen, 29.580 synoniemen, 236.656 voorbeeldzinnen, 119.214 combinaties, 8768 vaste verbindingen, 316 spreekwoorden, 4726 afbeeldingen, 730 videofilmpjes en 179 geluiden.

Medewerkers
In 2018 werd dataproductiemedewerker Wil de Ruyter definitief aan het ANW-team toegevoegd.
Op 5 maart 2018 eindigde het dienstverband van redacteur Josefien Sweep, die afscheid nam in verband met het aanvaarden van een nieuwe betrekking bij de lerarenopleiding Duits bij de Hogeschool van Amsterdam.
Met ingang van 1 september 2018 werd de redactie van het ANW versterkt met Boukje Verheij.

Technische ondersteuning
De redactionele werkzaamheden worden ondersteund door een computerlinguïst en een ontwikkelaar/programmeur. Op technisch gebied werden zowel voor als achter de schermen de nodige verbeteringen gerealiseerd. Afgezien van de dagelijkse update vallen o.a. te noemen een nieuw experimenteel werkstation met een nieuwe afdrukweergave, het zichtbaar maken van semagrammen bij vaste verbindingen in het ANW online, het toevoegen van de mogelijkheid om te zoeken op een

woord dat in de definitie gebruikt wordt en de mogelijkheid om bij het zoeken in het ANW direct officiële varianten te vinden, zodat bijvoorbeeld *kieviet* bij *kievit*, *giraffe* bij *giraf*, *sardien* bij *sardine* en *libelle* bij *libel* gevonden worden.

Vooruitblik
Nu de vervollediging van de categorie diernamen gereed is, zal er in 2019 vooral veel aandacht besteed worden aan de medische termen en aan woorden uit de categorie verkeer en vervoer, een samenlevingsdomein dat sterk aan verandering onderhevig is en tal van nieuwe woorden en gefixeerde woordcombinaties oplevert.

In het komende jaar wordt opnieuw gewerkt aan het maken van de nog ontbrekende fonetische schrijfwijzen. Het streven is om in 2020 alle transcripties aan het ANW toe te voegen.
Het blijft de bedoeling dat het ANW voor het INT ook een experimenteel platform wordt, waarvan de resultaten ook andere projecten ten goede komen. Zo zullen de ideeën die ontwikkeld zijn in het kader van het ANW met betrekking tot het zoeken op kenmerken nader uitgewerkt en geïmplementeerd worden, niet alleen voor het ANW, maar ook voor andere woordenboekapplicaties.

2. Neologismen

In 2018 is er gewerkt aan het ontwerpen van een neologismeneditor die speciaal gebruikt wordt voor het *Neologismenwoordenboek*. Dit woordenboek wordt voortdurend aangevuld en geüpdatet. Een apart woordenboek voor neologismen is nodig, omdat veel neologismen de weg naar gewone woordenboeken, bijvoorbeeld het ANW, niet vinden omdat ze maar kort in gebruik zijn. Toch is het interessant om die woorden vast te leggen en te beschrijven, bijvoorbeeld omdat veel neologismen morfologisch gezien in een bepaalde reeks passen en er patronen zichtbaar zijn: zo worden er zeer regelmatig nieuwe samenstellingen gevormd met het eerste lid *lok-*, *sjoemel-* en *troetel-* of met het tweede lid op *-gate* (in de betekenis ‘schandaal’) en nieuwe afleidingen op *-ing*, vooral in benamingen voor nieuwe trends of rages.

Een eerste testversie van het *Neologismenwoordenboek* is in 2018 in gebruik genomen en er is een proef gedaan met het bewerken van neologismen in deze nieuwe woordenboekomgeving. Ook zijn de neologismen die al in het ANW zijn opgenomen toegevoegd aan het *Neologismenwoordenboek*. Bovendien is weer verder gegaan met het automatisch selecteren van neologismen door middel van het programma *Neoloog*. Aan *Neoloog* zijn ook nieuwe functies toegevoegd die het mogelijk maken vanuit dit programma woordenboeklemma´s toe te voegen aan het ANW en het *Neologismenwoordenboek* en alvast een voorbewerking uit te voeren. Op 30 november heeft Vivien Waszink op het INT een lezing gegeven over dit nieuwe project en het verloop van de werkzaamheden daaraan. In 2019 komt een eerste versie van het *Neologismenwoordenboek* online. Vivien Waszink zal het woordenboek dan ook presenteren op de conferentie van Globalex van de Dictionary Society of North America (DSNA) in Bloomington.

3. Historische woordenboeken

In het voorbije jaar werd in april, conform de planning, de geheel vernieuwde interface voor de onlineapplicatie van de historische woordenboeken geïmplementeerd (nu zonder Adobe Flash). Daarnaast werd, mede in het kader van *ELEXIS* – het Europees project gericht op de ontwikkeling van een duurzame e-lexicografische infrastructuur – in 2018 intensief gewerkt aan het omzetten van de XML-bestanden van de historische woordenboeken naar een codering conform TEI-P5.
In het verslagjaar werd ook een digitaal bestand verworven van het *Supplement op het Middelnederlandsch Handwoordenboek* van J. J. van der Voort van der Kleij, een nog niet gedigitaliseerd INL-product uit 1983. Dit bestand zal t.z.t. aan de onlineapplicatie van de historische woordenboeken toegevoegd worden. Op het bestand is al een aantal bewerkingen uitgevoerd ter optimalisering van de data.

Tot slot: in de webrubriek *Terug in de Taal*, waarmee we onze historische woordenboeken ruimer onder de aandacht willen brengen, werd ook dit jaar een twintigtal bijdragen rond historische woorden en begrippen gepubliceerd.

4. Terminologie

In 2018 is in het kader van de verbrede taakstelling van het Instituut voor de Nederlandse Taal begonnen met de ontwikkeling van het Expertisecentrum Nederlandstalige Terminologie (ENT). Daarvoor is een beleidsnota geschreven en gerealiseerde werkzaamheden kaderen binnen de twee geformuleerde beleidslijnen: ondersteuning van het veld door voorzieningen op het gebied van de Nederlandstalige terminologie en de positionering van het ENT in het terminologische netwerk. Onder de eerste doelstelling valt ondersteuning door voorzieningen die worden verzameld en beschikbaar gesteld op het INT via de terminologiesite van het ENT, alsook ondersteuning waarbij het ENT geïnteresseerden in de terminologie zelf benadert via een nieuwsbrief.
In 2018 zijn vier nieuwsbrieven verstuurd met onder andere terugblikken op waardevolle evenementen, verslaglegging over nieuwe ontwikkelingen en publicaties, nieuwsberichten over terminologische bronnen en tools, en een agenda. Deze nieuwsbrieven werden ook gearchiveerd op de terminologiepagina van het ENT.

Deze terminologiesite is in ontwikkeling binnen de INT-website. Een reeds beschikbare rubriek in 2018 betrof het overzicht van terminologie-evenementen, zowel voor Nederland en Vlaanderen als in Europees perspectief.

Andere rubrieken zullen mede gebruik maken van de data van de NedTerm-website van de Taalunie. Voor de overdracht hiervan vond coördinatieoverleg plaats met medewerkers van de Taalunie en het INT, evenals verdere analyse en selectie van de gegevens. Voor de rubriek ‘Bronnen’ werd, vooralsnog in een digitale proefsetting, de basisstructuur voor termenlijsten opgezet overeenkomstig de onderwerpencatalogus van de Library of Congress. Rubrieken en subrubrieken dienden daarbij vertaald te worden naar het Nederlands. Ook nieuwe links naar termcollecties op het internet werden verzameld en toegevoegd. Daarnaast biedt het INT zelf eveneens termcollecties aan of stelt deze beschikbaar op basis van projecten of door hosting. Op het vlak van de medische vaktaal is overleg gevoerd over *Pinkhof Geneeskundig Woordenboek*. Het digitale bestand daarvan is aan het INT overgedragen en zal vanaf 2019 worden bewerkt. Diverse malen is er in 2018 ook overleg geweest over het *Juridisch Woordenboek Nederlands-Spaans* van M.C. Oosterveld-Egaz en J.B. Vuyk-Bosdriesz dat in september 2019 zal worden overdragen.

Voor de rubriek ‘Hulpmiddelen’ van de terminologiepagina vonden voorbereidingen plaats om er de extractietool TermTreffer en de beheerstool TermBeheerder aan te bieden. Naar aanleiding van de overdracht door de producenten aan het INT hielden deze werkzaamheden onder meer in het testen van de opgeleverde codes en patches en van ato-versies (acceptatie- en testomgeving), alsook de administratieve afhandeling. Voor het onderhouden en verder ontwikkelen van beide tools zijn voorstellen geformuleerd voor functionele verbeteringen. In 2019 volgt het onderzoek van de IT-afdeling naar de mogelijkheden hiervoor. Voor de huidige, publiek beschikbare applicatie op de INT-website TermTreffer.org werden op verzoek van externe gebruikers accounts uitgegeven.
De terminologiesite van het ENT gaat in de loop van 2019 online en wordt dan verder uitgebreid.
Werkzaamheden hadden eveneens betrekking op de tweede algemene doelstelling: de positionering van het ENT in het terminologische netwerk. Daarvoor kwam op verschillende wijzen contact en/of samenwerking tot stand binnen het Nederlandse taalgebied en in internationaal verband. Voor de veldvereniging NL-Term droeg het ENT bij aan de tiende editie van de TiNT-dag door redactionele en digitale ondersteuning via de evenementenwebsite van het INT. Voor het pilootproject *Hoger Onderwijs Terminologie in Nederland en Vlaanderen* (*HOTNeV*) vond overleg plaats op het niveau van de werkgroep en stuurgroep en werd de tool *qTerm* voor meertalig termbeheer geconfigureerd naar de structuur van de terminologische fiches van de Europese IATE-termbank. Deze tool werd ook ingezet voor de stagebegeleiding in twee terminologiestages voor de ULeiden. Een dertigtal studenten van de KULeuven, Campus Sint-Andries Antwerpen kreeg in het kader van colleges werkopdrachten

waarvan de resultaten in 2019 bijdragen aan het *HOTNeV*-project. Het ENT beoogde de terminologische positionering ook door ondersteuning van projecten als het *Algemeen Nederlands Woordenboek* door bewerking van *HOTNeV*-termen en proefartikelen voor plantennamen die in de algemene taal doordrongen. Contact en samenwerking in internationaal verband werden verder gerealiseerd door bijdragen met lezingen aan de tweedaagse workshop terminologie tijdens de twintigste editie (2-6 juli) van het International Congress of Linguists in Kaapstad. Vanuit het ENT is ook deelgenomen aan de door ECQA gecertificeerde opleiding terminologiebeheer van de internationale terminologievereniging TermNet en aan de cursus *Vertaaltechnologie. Bronnen voor terminologie* aan de KU Leuven Campus KulaK Kortrijk.

5. Woordcombinaties

Woordcombinaties wordt een online taaltool die leerders van het Nederlands als vreemde taal en taalgebruikers in het algemeen ondersteunt bij het gebruiken van woorden in context. In een pilot met werkwoorden wordt bekeken in hoeverre een collocatie- en idioomwoordenboek en een patroonwoordenboek geïntegreerd kunnen worden.
Het project is corpusgebaseerd en gebruikt Sketch Engine als corpusquerytool. Voor de beschrijving van patronen en hun betekenis wordt de Corpus Pattern Analysis (CPA) gebruikt van Patrick Hanks. *Woordcombinaties* toont hoe woorden gebruikt worden in goede voorbeeldzinnen, welke woorden met elkaar gecombineerd worden en hoe (valentie)patronen samen met collocaten gebruikt worden voor het bouwen van zinnen. Taalleerders en taalgebruikers leren met de tool niet alleen woorden kennen, maar ook woorden gebruiken.

Corpuswerkzaamheden
Er is een pilotcorpus samengesteld dat voornamelijk bestaat uit recent krantenmateriaal uit Nederland en Vlaanderen (NRC en De Standaard) uit het *Corpus Hedendaags Nederlands*. Het corpusmateriaal is omgezet naar 1 woord per lijn formaat en in het Corpus Query Systeem, de Sketch Engine, geladen. Voor de extractie van goede voorbeeldzinnen met Sketch Engine zijn een aantal GDEX-configuraties ontwikkeld en getest, waarmee voorbeeldzinnen automatisch gesorteerd worden volgens bepaalde criteria. Hierdoor wordt de redactionele bewerking van voorbeeldzinnen versneld. Daarnaast is een sketch grammar geschreven waarmee grammaticale relaties uit het corpus kunnen worden geëxtraheerd ten behoeve van de bewerking van de combinatiemogelijkheden.
Naast het getagde pilotcorpus is ook begonnen met het parseren van corpusmateriaal. De teksten van NRC en De Standaard sinds 2000, die beschikbaar zijn in het *Corpus Hedendaags Nederlands*, werden automatisch syntactisch geannoteerd met Alpino (van Noord, 2005)[1]. De tools die uit de resulterende syntactische bomen (xml-files) de dependents van de 124 voor het pilootproject geselecteerde werkwoorden trekken, zijn gebouwd en getest op enkele jaargangen. Daarnaast zijn de syntactische bomen geconverteerd naar universal dependencies (UDs) in CoNLL-formaat (Conference on Natural Language Learning, i.e. 1 woord per lijn, informatie in kolommen), volgens het algoritme van Bouma & van Noord (2017)[2]. Dit formaat staat toe om de data in te laden in Sketch Engine, waardoor het makkelijk toegankelijk wordt voor lexicografen.

Ontwikkeling lexicografische tools
Begin oktober 2018 is het eerste onderdeel van de editor gereedgemaakt voor de bewerking van de voorbeeldzinnen. Begin 2019 wordt deze editor verder uitgebreid met bewerkingsmogelijkheden voor de combinatiemogelijkheden.

[1] Gertjan van Noord (2005). At Last Parsing is Now Operational. TALN. Leuven.

[2] Gosse Bouma and Gertjan van Noord (2017). Increasing Return on annotation investment: the automatic construction of a Universal Dependency treebank for Dutch in: Proceedings of the Universal Dependencies Workshop, Gothenburg, 22 May 2017.

Binnen het Elexisproject is ook gestart met de ontwikkeling van een patrooneditor die gebruikt kan worden voor de bewerking van de (valentie)patronen in *Woordcombinaties*. Daarmee kan vanaf de tweede helft van 2019 begonnen worden met de beschrijving van de werkwoordbetekenissen en (valentie)patronen, waarin collocaten gekoppeld worden aan betekenissen en patronen.

Redactionele werkzaamheden
In de loop van 2018 zijn de verschillende woordenboekfeatures en de daarvoor benodigde editoronderdelen verder uitgewerkt en vastgelegd ten behoeve van het datamodel en de te ontwikkelen bewerkingstools. Na vergelijkende studies van een aantal woordenboeken en lexicale databases zijn inventarissen gemaakt, o.a. van argument- en complementtypes, gebruikslabels en kennisdomeinen. Er zijn een aantal modelartikelen gemaakt met patroonbeschrijving ten behoeve van de patrooneditor. Er is een vergelijkende studie gemaakt van de behandeling van reflexieve werkwoorden in de lexicale databases *DuELME*, *RBN* en *Cornetto*, de woordenboeken van *Van Dale* en de treebank *LASSY*. Dat heeft geresulteerd in een beleidsnotitie over de behandeling van reflexieve werkwoorden in *Woordcombinaties*.

In het laatste kwartaal van 2018 is begonnen met de selectie en naredactie van semiautomatisch gegenereerde voorbeeldzinnen bij de werkwoorden. De verschillende GDEX-configuraties zijn ook door de redactie uitgetest en vergeleken, waarna gekozen is voor de meest bevredigende configuratie. De bewerking van de voorbeeldzinnen is in de eerste week van 2019 voltooid. Begin 2019 kan dan begonnen worden met de redactionele bewerking van de combinatiemogelijkheden. De bewerking van de patronen is te verwachten vanaf de tweede helft van 2019.
Naast de redactionele werkzaamheden is ook aandacht besteed aan communicatie over het project. Er is een paper geschreven voor het EURALEX-congres, en abstracts en een poster voor presentaties en workshops. Er zijn een aantal lezingen en presentaties gegeven over het project.

6. Taalbank Nederlands

De Taalbank Nederlands omvat corpora en computationele lexica van zowel modern als historisch Nederlands. Het lexicon is bedoeld om ingezet te worden voor onderzoek. Hieronder een stand van zaken (zie bijlage 4 voor een uitgebreid verslag).

GiGaNT
GiGaNT (Groot Geïntegreerd Lexicon van de Nederlandse Taal) is het computationele woordvormenlexicon dat de door het INT beschreven woordenschat moet gaan bevatten van het Nederlands vanaf de 6e eeuw tot nu en de centrale database vormt van het Nederlands van het INT. In GiGaNT zit de formele beschrijving van de woorden. De semantische informatie zit in DiaMaNT, en in de diverse woordenboeken van het INT.

Werkzaamheden aan deze centrale data-infrastructuur zijn modulair opgezet. De twee grote componenten zijn GiGaNT-Molex, waarin de hedendaagse woordenschat wordt beschreven', en GiGaNT-Hilex, de historische lexiconcomponent. De historische lexiconcomponent heeft als kern de vier historische woordenboeken van het Nederlands (ONW, VMNW, MNW en WNT) en wordt aangevuld met nieuw materiaal. De moderne lexiconcomponent bevat het hedendaags taalmateriaal. Afgelopen jaar is gewerkt aan de verdere integratie van de historische en de moderne lexiconmodule, en gewerkt aan de uitbreiding van de moderne lexiconcomponent, een continue taak. Er is ook werk gedaan aan de koppeling van het ANW en GiGaNT-Molex. Hiermee wordt op termijn niet alleen de data van het ANW toegevoegd aan de moderne lexiconcomponent, het is ook een stap die nodig is om te komen tot een geïntegreerde lexiconworkflow voor het beschrijven van de hedendaags Nederlandse woordenschat.

DiaMaNT
DiaMaNT (Diachroon seMantisch lexicon van de Nederlandse Taal) is een project dat uitgevoerd wordt binnen CLARIAH (zie hoofdstuk 17). Het is gestart op 15 oktober 2015 en loopt tot en met 31

december 2018. In het project zal gewerkt worden aan de ontwikkeling van een diachroon semantisch lexicon van het Nederlands. Dit project moet het ontwerp, de bouwwijze en een eerste versie opleveren van een diachroon semantisch lexicon. Het diachroon semantisch lexicon heeft als doel een hulpmiddel te bieden bij tekstontsluiting en bij het onderzoek naar begrippen door de eeuwen heen. Het lexicon zal relaties leggen tussen woordvormen en betekeniseenheden (concepten), en deze in de tijd plaatsen. Er is gewerkt aan het datamodel, aan de controle en de uitbreiding van de data, en een eerste versie is opgeleverd aan CLARIAH. Er is een eerste prototype zoekapplicatie gebouwd met querybuilder en visualisaties voor lexicale Linked Open Data. Tot slot is er verder geëxperimenteerd met distributionele semantiek en met word sense disambiguation.

Modern corpusmateriaal: Corpus Hedendaags Nederlands (CHN)
Het Corpus Hedendaags Nederlands (CHN) bevat hedendaags taalmateriaal met teksten voornamelijk uit kranten, tijdschriften, journaaluitzendingen en juridisch materiaal. Hoewel er geen release is geweest, is ook dit jaar binnenkomend materiaal verder verwerkt. Winst van het in 2017 door de Taalunie georganiseerde werkbezoek is dat dataleverantie van materialen uit Suriname die ook al ten behoeve van het Groene Boekje 2015 waren toegevoegd, opnieuw tot stand is gekomen. Daar is een dataleverancier bijgekomen en een lijst contacten waar hopelijk in de toekomst nog meer taalmateriaal uit zal komen. Momenteel bevat het interne corpus 3,8 miljoen teksten. Een update van het corpus, intern en extern (dat kleiner is vanwege IPR-beperkingen) is gepland voor het voorjaar van 2018. Het zwaartepunt van de werkzaamheden voor het CHN lag bij de werkzaamheden aan de corpusworkflow. Het uitbreiden van het Corpus Hedendaags Nederlands is zoveel mogelijk geautomatiseerd. Ontvangen en opslaan van materiaal, conversies, verrijking en indexering zijn stadia in een corpusworkflow, uitgevoerd binnen een daartoe ontwikkelde tool DUCT (Data Update Creation Tool). DUCT is een tool voor het converteren van bestanden in verschillende stappen. Aan het corpuszoeksysteem BlackLab zijn, onder andere in de context van CLARIAH, diverse verbeteringen doorgevoerd waaronder een verbeterde user interface en verbeteringen van de performance.

Historisch corpusmateriaal
Werkzaamheden aan het historisch corpusmateriaal zijn uitgevoerd in het kader van het project Nederlab. Sedert 1 januari 2017 is het INT verantwoordelijk voor de corpusprocessing van Nederlab. Dat betekent dat digitaal corpusmateriaal van derden wordt geconverteerd naar Nederlabformaat (XML-FoLiA) en voorzien van taalherkenning getokeniseerd aan de projectpartners wordt opgeleverd. De teksten worden voorzien van correcte metadata, inclusief thesaurering van de auteurs. In het afgelopen jaar zijn negen collecties aan het Nederlabcorpus toegevoegd, met in totaal ca. 31 miljoen woorden. Daarnaast is er met de toolstrack samengewerkt om de taalkundige verrijking op een zo goed mogelijk peil te krijgen. Er is geëvalueerd en er zijn evaluatiesets gemaakt.
De werkzaamheden aan de zoekapplicatie van het INT, die uitgevoerd zijn in de context van het CLARIAH-project, en met name het verbeteren van de user interface, komt ook de historische corpora van het INT ten goede, en zullen zichtbaar zijn in nieuwe releases van bijvoorbeeld het *Corpus Gysseling* en *Brieven als Buit* in 2018.

Externe projecten
Nederlab
Nederlab wil een gebruiksvriendelijke webinterface inrichten vanwaaruit onderzoekers losse, historische corpora als eenheid kunnen doorzoeken en analyseren. Het INT werkt eraan mee om historisch lexicaal materiaal inzetbaar te maken voor zoeken in het historisch corpusmateriaal van Nederlab. Een substantieel deel van het werk aan de historische lexica (zie GiGaNT-HILEX) wordt in het kader van Nederlab uitgevoerd. Daarnaast heeft het INT de taak op zich genomen een substantieel deel van het totale Nederlabcorpus af te ronden.

Nederlab klein deelproject: corpus 15^{e} en 16^{e}-eeuws Nederlands
In samenwerking met de RU Nijmegen is gewerkt aan het bijeenbrengen van een corpus van 15^{e} en 16^{e}-eeuws ambtelijk materiaal voor Nederlab.

7. Spelling

Woordenlijst
Sinds 1995 is de Woordenlijst Nederlandse Taal bij ons instituut ondergebracht. Het gedrukte boekje werd in 2005 en 2015 geactualiseerd, terwijl de onlineversie sinds 2015 driemaandelijks geüpdatet wordt. Dit vervolgtraject werd voortgezet in 2018: er werden ruim 6000 trefwoorden opgeleverd voor de updates van het onlinespellingbestand. Naast nieuwe trefwoorden werden ook extra woordvormen en andere informatie aangevuld en vonden ook de nodige correcties aan het reeds aanwezige taalmateriaal plaats. Alle correcties werden verzameld in een dynamische erratalijst, die nauwgezet wordt bijgehouden. Bij elk van deze updates werden ook de nodige datawerkzaamheden en inhoudelijke en technische controles uitgevoerd door en bij het INT.

In 2018 werd ook een lijst met Surinaams-Nederlandse trefwoorden samengesteld door het INT (uit het beschikbare bronnenmateriaal), die werd voorgelegd aan enkele deskundigen in Suriname, teneinde de onlinewoordenlijst in 2019 aan te vullen met bijkomend Surinaams-Nederlands materiaal. Een soortgelijk traject is in 2019 ook voor het Antilliaans-Nederlands voorzien.
In 2019 bekijken we, in overleg met de Nederlandse Taalunie, de mogelijkheid om de hosting van woordenlijst.org bij het INT onder te brengen.

Commissie Spelling
Alle toevoegingen en wijzigingen aan het onlinespellingbestand werden ook in 2018 voorgelegd aan en goedgekeurd door de Commissie Spelling. In 2018 werden 3 vergaderingen van deze Commissie georganiseerd. Door het INT werden voor deze vergaderingen o.a. volgende taken uitgevoerd: samenstellen en aanleveren van woordenlijsten; voorbereiden van inhoudelijke kwesties; verwerken van opmerkingen, aanvullingen en correcties in het spellingbestand.

Keurmerk Spelling
In 2018 werden onderstaande controles uitgevoerd in het kader van het toekennen van het spellingkeurmerk:

- Uitgeverij Gateway heeft twee controles van taalmateriaal van het *Juridisch-Economisch Lexicon* laten uitvoeren, goed voor 2454 trefwoorden en 1435 voorbeeldzinnen in april en 1548 trefwoorden en 1001 voorbeeldzinnen in oktober.
- Voor de vertaalwoordenboeken Nederlands-Portugees en Nederlands-Estisch (Vertaalwoordenschat INT) werd in resp. januari en december de spelling van resp. 8.827 en 37.471 trefwoorden gecontroleerd en zo nodig gecorrigeerd.

Vanaf 2019 houden de reguliere keurmerkcontroles op te bestaan. Voor de zeer omvangrijke doch voor het INT erg interessante woordverzamelingen van OpenTaal (ruim 230.000 trefwoorden) en TaalTik (ruim 915.000 trefwoorden) lopen aangepaste controletrajecten, gezien de grootte en complexiteit van deze materialen. De spellingcontrole van (substantiële delen van) deze trefwoordenverzamelingen betekent voor beide partijen een win-winsituatie.

Spelspiek & spellingvragen
In het kader van Spelspiek is in 2018 een zeventigtal vragen beantwoord. Daarnaast kwamen ook via e-mail spellingvragen (en bij uitbreiding taalvragen) binnen, die zo snel mogelijk werden beantwoord of doorgestuurd aan de juiste persoon of dienst.
Zoals ook vermeld in het meerjarenbeleidsplan, willen we bekijken of het potentieel van de interactieve toepassing Spelspiek niet beter kan worden benut door het ‘inhouse’ ontwikkelen van een versie 2.0, gevuld met eigen en dynamische data. Hiervoor zal in 2019 worden gestart met het nodige research- en voorbereidingswerk.

8. TST-materialen

Inleiding
In februari 2018 zijn er een aantal grote veranderingen geweest. Ten eerste is de naam 'TST-Centrale' vervangen door de naam 'Taalmaterialen'. Daarbij is een nieuwe website gelanceerd waarbij het eenvoudiger werd om producten te vinden. Die website bood ook de mogelijkheid om de meeste producten rechtstreeks te downloaden nadat de gebruiker zich akkoord verklaard heeft met de gebruiksvoorwaarden. En, last but not least, de producten werden ook voor commercieel gebruik officieel gratis te beschikking gesteld.

Nieuwe downloadpartner: PCLOUD.COM
Zoals hierboven reeds gemeld waren de meeste producten rechtstreeks te downloaden. Dat geldt echter alleen voor niet-commercieel gebruik van de producten. Voor commercieel gebruik moet een licentie worden ondertekend. De gebruikers die daarin geïnteresseerd zijn downloaden dan een zogenaamd bestelpakket, waarin een licentie en een instructie zit. Nadat de ingevulde en ondertekende licentie door ons is ontvangen wordt een link opgestuurd waarmee het product gedownload kan worden van de website pCloud.com. Die biedt een snelle en betrouwbare service aan onze gebruikers.
Overigens geldt voor sommige producten dat voor zowel commercieel als niet-commercieel gebruik een licentie dient te worden ondertekend.

Nieuwe producten
In het afgelopen jaar hebben we de volgende producten toegevoegd aan de catalogus. Een aantal daarvan werden daarvoor op een andere wijze door het INT beschikbaar gesteld. In verband met de introductie van de Taalmaterialenwebsite zijn die toegevoegd aan de Taalmaterialencatalogus.

- Wablieft Corpus. Deze verzameling bevat het digitaal archief van de Wablieft-krant (periode 2011-2017).
- BasiScript Corpus. Dit is een geannoteerde verzameling van teksten geschreven door kinderen in de basisschoolleeftijd.
- BasiScript Lexicon. Een lexicon afgeleid van het BasiScript Corpus.
- Philosophical Integrator of Computational and Corpus Libraries (PICCL). Deze onlineservice is ontwikkeld in het kader van CLARIAH en biedt een workflow aan voor het samenstellen van corpora waarbij een aantal bestaande tools zijn samengevoegd.
- INT IMPACT NE Lexicon. Een lexicon voor het Nederlands, met historische namen en varianten uit de periode 1750-1945.
- INT Historische Woordenlijst. De INT Historische Woordenlijst bestaat uit 2 lijsten met ieder ca. 500.000 historische woordvormen ten behoeve van OCR en OCR-postcorrectie, voor de periode ca. 1550 - ca. 1970.
- Brieven als Buit - Gouden Standaard. De circa 1000 met hoofdwoordsoort en modern lemma verrijkte bronbestanden van het Brieven als Buit-programma.
- OpenConvert. Een online tool om tekst te converteren naar XML-formaat (TEI) en te voorzien van taalkundige annotaties. Alleen toegankelijk met een CLARIN-account.
- Nerd. Online tool met een 'Named Entity Recognizer'. Alleen toegankelijk met een CLARIN-account.
- Memory Based Morphological Parser (MBMP). Een geheugengebaseerde morfologische parser voor de programmeertaal Python.
- INL Labs. Online tool voor het taggen/lemmatiseren van (historische) teksten met o.a. een tagger voor eigennamen (named entities) en een tagger speciaal getraind voor historisch materiaal.
- Cobalt. Applicatie om een verzameling tekstbestanden in te laden en taalkundig te annoteren.
- Blacklab. Corpuszoeksysteem op basis van Apache Lucene.
- Autosearch. Een online tool om geannoteerde teksten te uploaden (voorzien van lemma's en woordsoortinformatie in TEI- of FoLiA-formaat), één of meerdere corpora te definiëren en deze te doorzoeken. Alleen toegankelijk met een CLARIN-account.

- Attestation Tool. Multifunctionele gebruikersinterface voor de productie van computationele lexica, inclusief gouden standaard voor named entity tagging.
- OpenSoNaR. Online zoeksysteem voor het SoNaR-corpus, een tekstverzameling van hedendaags geschreven Nederlands dat uit meer dan 500 miljoen woorden bestaat.
- Brieven als buit. Online applicatie voor de ontsluiting van taalkundig verrijkte 17e- en 18e-eeuwse brieven tussen Nederlanders in verre oorden en hun families en geliefden aan het thuisfront.
- WebCelex. Online applicatie waarmee de CELEX-lexicale databases van het Duits, Engels, Nederlands kunnen worden geraadpleegd.
- Vertaalwoordenschat. Online applicatie voor tweetalige woordenboeken met Nederlands als bron- of doeltaal.
- Corpus Oudnederlands. Een verzameling van al het overgebleven Nederlandse woordmateriaal uit de periode 475-1200.

Geografische verspreiding van gebruikers
Gebruikers moeten zich te registreren om producten te kunnen downloaden. Daarbij wordt ook hun e-mailadres vastgelegd. Aan de hand van de extensie van die adressen kan een beeld worden verkregen vanuit welke landen belangstelling bestaat voor onze taalmaterialen. De meesten komen uit Nederland (105), daarna België (49) en verder Duitsland (17), UK (7), Zwitserland (5), Hongarije (3), British Indian Ocean Territory (3), Tsjechië (2), Zuid-Afrika (2), Ierland (2), en vervolgens Frankrijk, Canada, IJsland, Denemarken, Rusland, India, Zweden ieder één.

Overzicht downloads commercieel
In totaal is er negen keer een product afgenomen voor commerciële toepassingen.

Corpus Gesproken Nederlands (CGN)	3
Basilex Corpus	1
CGN-annotaties	1
BasiScript Corpus	1
BasiScript Lexicon	1
SoNaR Groot	1
SoNaR Klein	1

Overzicht downloads niet-commercieel
Er zijn in totaal 681 producten gedownload voor niet-commercieel gebruik.

Corpus Gesproken Nederlands (CGN)	146
SoNaR-corpus	63
CGN-annotaties	57
Cd-rom Middelnederlands	57
Frequentielijsten Corpora	52
Lassy Klein-corpus	33
Dutch Parallel Corpus (DPC)	32
e-Lex	29
DuOMAn-subjectivitylexicon	21
Referentiebestand Nederlands (RBN)	21

Corpus Middelnederlands	20
Corpus Gysseling	18
CombiLex	12
BasiLex Corpus	12
Wablieft-corpus	11
Corpus Pathologische en Normale Spraak (COPAS)	10
DuELME	9
PAROLE-lexicon	8
COREA-coreferentiecorpus	8
RBN-klein	7
Brieven als Buit - Gouden Standaard	6
BasiLex Lexicon	6
Afrikaans Custom Dictionary for Government Domain	5
DAESO-corpus: Parallelle Nederlandstalige monolinguale treebank	4
Jasmin	4
Afrikaans Genre Classification Corpus	3
D-TUNA-corpus	3
AUTONOMATA-namencorpus	3
Meertalige Ondertiteldata 2BDutch	3
BasiScript Corpus	3
AUTONOMATA-POI-corpus	2
OMBI Arabisch-Nederlands	2
Paco-MT Parallelle Corpora	2
Referentiebestand Belgisch-Nederlands (RBBN)	2
Dupira	2
SoNaR-Nieuwe Media	2
IFA-dialoog-videocorpus	1
OMBI Nederlands-Arabisch	1
OMBI Nederlands-Indonesisch	1

9. Taalportaal

Voor het *Taalportaal* zijn in 2018 twee reguliere updates uitgevoerd, waarmee er opnieuw enkele honderden artikelen aan het grammaticaportaal zijn toegevoegd. Daarnaast is er onderhoud verricht aan de auteursomgeving en de achterkant van de webapplicatie. In het najaar is, op initiatief van vIvA (Potchefstroom, Zuid-Afrika), begonnen met het updaten van de auteursomgeving naar de nieuwste versie van Oxygen. De Zuid-Afrikaanse auteurs en de Nederlandse werkten tot nog toe in twee verschillende databases voor literatuurverwijzingen, en deze zijn het afgelopen jaar samengevoegd. Beide klussen zullen naar verwachting in het voorjaar 2019 zijn afgerond.

Het afgelopen jaar is ook de discussie begonnen over de toekomst van *Taalportaal* en de mogelijke overgang naar een groot, geïntegreerd grammaticaportaal van *Taalportaal*, *eANS* en *Taaladvies.net*. In 2019 zal de verkenning hiernaar worden doorgezet. Ook zal in 2019 worden begonnen met het integreren van zoekresultaten van het project *Taalportaal-Zuid-Afrika* in de Taalportaalapplicatie, zal er verder onderhoud plaatvinden aan de bestaande webapplicatie, en een nieuw hoofdstuk van de *Syntax of Dutch* van Hans Broekhuis worden toegevoegd, met behulp van conversiesoftware van het INT.

10. Algemene Nederlandse Spraakkunst

Voor WP4 van het project *Herziening eANS* is in 2018 een demoversie van de nieuwe *Algemene Nederlandse Spraakkunst* (*ANS*) ontwikkeld voor intern gebruik en voor presentatiedoeleinden. Daarnaast is er verder gewerkt aan de ontwikkeling van de webapplicatie, de structuur van de inhoud en de mogelijke integratie van externe kennisbronnen en Taalportaalartikelen, en heeft er ondersteuning plaatsgevonden van de auteurs en hun werkomgeving. Dit werk zal voortgezet worden in 2019, waarbij ook een bestaand hoofdstuk uit LaTex naar XML moet worden getransformeerd. Er is een stageplaats opengesteld, maar hierop is vooralsnog geen reactie gekomen.

11. Nederlands in Cijfers

Dit project ligt feitelijk stil, maar de voor dit project eerder ontwikkelde Pythonscripts zijn wel gebruikt om data te genereren voor de theateravond van het *Jaar in Taal* in Brussel in december 2018. Deze data bestaat uit lijsten van woorden uit *De Standaard* en *NRC Handelsblad* die in 2018 sterk in gebruik zijn toegenomen in vergelijking met voorgaande jaren.

12. Vertaalwoordenschat

De *Vertaalwoordenschat* is een applicatie voor tweetalige woordenboeken met Nederlands als bron- of doeltaal, ontwikkeld door het Instituut voor de Nederlandse Taal. Rond de eeuwwisseling zijn er verschillende tweetalige bestanden ontwikkeld voor talen die voor de Nederlandstalige wel relevant zijn, maar op de commerciële markt niet spontaan aan bod kwamen, zoals Nederlands - Deens, Nederlands - Nieuwgrieks en Nederlands - Arabisch. Deze bestanden zijn veelal gemaakt in opdracht van een speciale *Commissie Lexicografische Vertaalvoorzieningen*, ingesteld door de toenmalige ministers van Onderwijs van Nederland en Vlaanderen. In de meeste gevallen beschikt de Taalunie over het volledige auteursrecht op deze bestanden en zijn met diverse uitgevers afspraken gemaakt over de papieren publicatie ervan. In enkele gevallen deelt de Taalunie het auteursrecht met andere partijen of beschikt ze enkel over een uitgavelicentie. Enkele talenparen zijn nu niet meer in druk ,omdat uitgevers er geen commerciële mogelijkheden meer in zien.

Het gevolg is dat deze bestanden niet meer beschikbaar zijn voor gebruikers en vertalingen tussen diverse talenparen niet meer worden ondersteund.

Daarom is in september 2017, het onlineplatform de *Vertaalwoordenschat* gelanceerd, waarmee de tweetalige bestanden ter beschikking worden gesteld aan gebruikers. Nederlands – Nieuwgrieks / Nieuwgrieks – Nederlands was het eerste taalpaar dat via het platform werd ontsloten. In mei 2018 is Nederlands - Portugees / Portugees - Nederlands toegevoegd. In de toekomst zal de applicatie nog verder worden uitgebreid met andere talen. Zo zal Nederlands-Estisch begin 2019 verschijnen. Naast de webversie, is in 2018 ook een app ontwikkeld. De app is gratis te downloaden en werkt zowel op Android als iOs. De mobiele app is momenteel nog een bètaversie, maar komt qua inhoud en functionaliteit overeen met de website.

13. Externe communicatie & wetenschapscommunicatie

In 2018 werd op het gebied van externe communicatie de huisstijl verder doorgevoerd, de bestaande communicatiemiddelen werden voortgezet en waar nodig verbeterd en er werden nieuwe populairwetenschappelijke activiteiten ontwikkeld. Net als ieder jaar trad het instituut ook weer op als sponsor voor CLIN en de TABU-dag (zie bijlage X voor een uitgebreid communicatieverslag).
Ook op het gebied van wetenschapscommunicatie gingen we op dezelfde voet voort: in 2018 publiceerden we in populairwetenschappelijke tijdschriften en werkten we mee aan onderzoeken of artikelen van onlineplatforms of tijdschriften. Ook het begeleiden van scholieren en studenten bij het maken van profielwerkstukken, scripties en andere onderzoeken valt onder wetenschapscommunicatie, net als het verzorgen van optredens op de radio (bijvoorbeeld in De Taalstaat) en tv en het doen van interviews voor kranten, tijdschriften en blogs. Medewerkers van het INT houden regelmatig populairwetenschappelijke lezingen, vaak over neologismen, spelling, taalverandering, jongerentaal etc., bijvoorbeeld op taalcongressen, maar ook in het (middelbaar) onderwijs.

Huisstijl
De nieuwe huisstijl is verder doorgevoerd in twee nieuwe folders en ‘gadgets’ voor bijeenkomsten en congressen. Met name het geelgekleurde linnen tasje, speciaal ontworpen als cadeautje voor onderzoeksdeelnemers tijdens het DRONGO talenfestival, was een groot succes.

Website
De website is het gezicht van het instituut naar buiten en daarmee het belangrijkste communicatiemiddel. Bijna alle communicatie-uitingen zijn erop gericht om bezoekers naar de website te trekken. In 2018 zijn er weer meer bezoekers op de website geweest dan het jaar ervoor, ondanks dat er dit jaar geen eindejaarsverkiezing georganiseerd werd die altijd voor veel extra aandacht zorgde. Het onderbrengen van de website van de TST-Centrale op de INT-website onder de noemer ‘Taalmaterialen’ heeft in ieder geval wel bijgedragen aan het groeiende bezoekersaantal.
De grote aandachttrekkers op de website zijn de populairwetenschappelijke webrubrieken *Woordbaak* over onder andere etymologie en *Neologisme van de week* over nieuwe woorden. Het is daarom van belang om deze rubrieken in stand te houden. Informatie over de historische woordenboeken wordt ook goed bezocht, waarschijnlijk mede dankzij de lancering van een nieuwe versie van gtb.ivdnt.org, de historische woordenboeken online.

Informatie over het *Algemeen Nederlands Woordenboek* is opvallend veel meer bekeken dan in het voorgaande jaar, en ook de Gelegenheidswoordenboekjes waren een stuk populairder. Daarvan zijn er in 2018 drie nieuwe verschenen.

Tijdens de Week van het Nederlands is eenmalig de nieuwe rubriek *Term van de dag* in het leven geroepen, die redelijk wat bezoekers trok. Het is goed om te bekijken of we deze rubriek op regelmatige basis kunnen voortzetten, bijvoorbeeld als onderdeel van de terminologienieuwsbrief, om op die manier ook wat meer aandacht voor terminologie te genereren.

Nieuwsbrieven, mailings & persberichten
De nieuwsbrieven (algemeen & terminologie) worden verstuurd om geïnteresseerden te blijven binden aan en te informeren over het instituut. In 2018 zijn er 5 algemene nieuwsbrieven verschenen en 4

terminologienieuwsbrieven. Daarnaast zijn er nog diverse aparte mailings verstuurd (o.a. aankondigingen symposia, uitnodigingen en persberichten). Voor de verzending en beheer van het adressenbestand wordt gebruikgemaakt van de nieuwsbriefsoftware MailChimp.

Het totale abonneebestand van de nieuwsbrieven bestaat uit ongeveer 6.500 e-mailadressen. De nieuwe aanmeldingen in 2018 zijn gering, en het abonneebestand voor de terminologienieuwsbrief is zelfs iets gekrompen. Ook liggen de cijfers voor het aantal opens (31%) onder het gemiddelde van de Nationale E-mail Benchmark van 2018 (38% opens). De algemene nieuwsbrief zit daar ruim boven met 46% opens. Het is goed om het komende jaar te bekijken hoe de nieuwsbrieven, en dan met name de terminologienieuwsbrief, verbeterd kunnen worden.

Sociale media

Sinds 2010 heeft het instituut een Twitteraccount. Via Twitter worden links naar de website en taalweetjes/taalnieuws gedeeld, en contacten met volgers en partnerorganisaties onderhouden. Het aantal volgers groeit gestaag. In 2018 zijn er 270 nieuwe volgers bij gekomen.
De Facebookpagina voor *Weg met dat woord!* is niet meer heel actief en er moet bekeken worden wat we met deze pagina gaan doen. De pagina kan bijvoorbeeld omgevormd worden tot een algemene Facebookpagina voor het INT, of worden gebruikt voor een nieuwe eindejaarsactie.

Populairwetenschappelijke activiteiten

Tijdens de Week van het Nederlands in september werd de eenmalige rubriek *Term van de dag* gelanceerd, om aandacht te vragen voor terminologie. Eind november stond het instituut op het DRONGO talenfestival met een lab over Taalradar en dialectloket.be.
In november verscheen bij uitgeverij AUP het populairwetenschappelijke boek *Kids, koffietjes & comfortzone. Waarom taal soms irritant is.* Het boek werd geschreven door Laura van Eerten en Vivien Waszink en is een afronding van vijf jaar de verkiezing *Weg met dat woord!*.
Eind december organiseerde het INT samen met het Vlaams-Nederlands Huis deBuren voor de tweede keer ‘Het jaar in taal’: een avond over woorden die het publieke debat in 2018 domineerden.
Tot slot beantwoorden we door het jaar heen veel vragen over het Nederlands aan onder andere studenten, journalisten en programmamakers. In 2018 werd het instituut regelmatig door radio en tv benaderd.

14. EnetCollect

2018 was het tweede jaar van de COST-actie, *enetCollect, European Network for Combining Language Learning with Crowdsourcing Techniques*. Deze COST-actie heeft vijf werkgroepen die alle aspecten van het combineren van crowdsourcing met het leren van (een) taal dekken. Frieda Steurs is vicevoorzitter van werkgroep 3 ‘User-oriented design strategies for a competitive solution’. Carole Tiberius en Tanneke Schoonheim zijn lid van het Management Comittee.
Frieda Steurs en Tanneke Schoonheim woonden de bijeenkomst bij van Management Committee en werkgroepen op 14, 15 en 16 maart in Iasi (Roemenië). Op 24 en 25 oktober werd vervolgens in Leiden een gezamenlijke bijeenkomst georganiseerd van de werkgroepen 3 en 5 **‘**Application-oriented specifications for an ethical, legal and profitable solution**’**. De organisatie daarvan lag in handen van de (vice)voorzitters van de beide werkgroepen, Tanneke Schoonheim en het secretariaat. Peter Dekker en Tanneke Schoonheim gaven een presentatie over hun ervaringen met het crowdsourcingsplatform Pybossa, getiteld ‘Recognizing blends: first experiments with Pybossa’. De bijbehorende publicatie, aangevuld met gegevens van nieuwe experimenten, verschijnt in het voorjaar van 2019. Namens het INT waren ook Kris Heylen en Carole Tiberius bij deze bijeenkomst aanwezig.

Op 5, 6 en 7 december vond er in Gotenburg (Zweden) de workshop *Learning materials through crowdsourcing: teachers, perspectives & scenarios* plaats over het gebruik van Pybossa. Peter Dekker en Tanneke Schoonheim waren hierbij aanwezig met een presentatie over hun ervaringen met Pybossa, ditmaal met nieuwe experimenten op het gebied van het herkennen van neologismen en het

verzamelen van dialectwoorden. Deze presentatie droeg de titel ‘When to use PYBOSSA? Case studies on crowdsourcing for Dutch’.

Voor 2019 staan er verschillende bijeenkomsten en workshops gepland. In januari is er een zogeheten crowdfest in Brussel, waarop taalkundigen, computerlinguïsten en softwaredevelopers gezamenlijk aan het ontwikkelen van crowdsourcingstaken zullen werken. De jaarlijkse plenaire bijeenkomst van het project is in maart in Lissabon. Frieda Steurs, Tanneke Schoonheim en Peter Dekker zullen hier namens het INT aanwezig zijn. In het kader van dit project worden ook de verschillende experimenten met Pybossa voortgezet om te onderzoeken in hoeverre dit crowdsourcingplatform kan bijdragen aan het verzamelen en beoordelen van taaldata voor taalleerders.

15. European Infrastructure for Lexicography (ELEXIS)

ELEXIS is een Europees Horizon 2020 project met als doel een duurzame infrastructuur voor e-lexicografie te creëren. Deze infrastructuur moet het mogelijk maken de kwalitatief hoogwaardige semantische informatie die momenteel nog veelal opgesloten zit in individuele lexicografische bronnen verspreid over Europa op grote schaal te koppelen, delen, verspreiden en op te slaan. Tevens zal een infrastructuur die speciaal gericht is op e-lexicografie bijdragen aan het verkleinen van de kloof tussen gemeenschappen met veel en weinig lexicografische expertise.

ELEXIS komt voort uit de COST-actie IS1305 European Network of eLexicography, die in oktober 2017 afliep. Binnen dit netwerk kwam duidelijk de behoefte naar voren voor een bredere en meer systematische uitwisseling van expertise, voor het vaststellen van gemeenschappelijke standaarden en oplossingen voor de ontwikkeling en integratie van lexicografische materialen en voor het uitbreiden van de toepassingsmogelijkheden van deze kwalitatief hoogwaardige materialen, o.a. binnen het semantische web, de kunstmatige intelligentie, NLP en de Digital Humanities.

ELEXIS is een samenwerkingsverband tussen 17 partners uit Europa en Israël. Het project is in februari 2018 begonnen en heeft een looptijd van 4 jaar.

Het INT leidt het werkpakket “Lexicographic data and workflow”. Speerpunten van dit werkpakket zijn a) het maken van een inventarisatie van de behoeften van de lexicografische gemeenschap om zo het maken van woordenboeken optimaal te kunnen ondersteunen; b) het vastleggen en ondersteunen van gezamenlijke standaarden en werkwijzen voor het lexicografische proces en c) het ontwikkelen van methoden en tools voor de conversie, automatische segmentatie en identificatie van lexicografische inhoud.

Het zwaartepunt van de activiteiten in 2018 lag op het maken van een inventarisatie van de behoeften van de lexicografische gemeenschap. Hiervoor zijn in samenwerking met drie partnerinstituten, twee enquêtes gemaakt: een voor lexicografen en een voor lexicografische instituten. De resultaten bieden een waardevol inzicht in de lexicografische praktijk van dit moment en zullen de komende jaren binnen *ELEXIS* gebruikt worden bij de ontwikkeling van richtlijnen en tools. Daarnaast is in 2018 eerste aanzet gemaakt voor een gezamenlijk datamodel voor lexicografische materialen en is er gewerkt aan interoperabele formaten (o.a. Ontolex-Lemon en TEI Lex-0).

projectbijeenkomst: *ELEXIS-Kick off* (15-17 februari, Ljubljana) aanwezig namens het INT: Bob Boelhouwer en Carole Tiberius
projectbijeenkomst: *ELEXIS TMB* (6-7 november, Leiden) aanwezig namens het INT: Carole Tiberius, Bob Boelhouwer, Katrien Depuydt en Jesse de Does
workshop: *TEI Lex-0* (16 juli, Ljubljana) aanwezig namens het INT: Katrien Depuydt, Jesse de Does en Carole Tiberius
workshop: *Ontolex-Lemon* (5 november, Leiden) aanwezig namens het INT: Katrien Depuydt, Jesse de Does en Carole Tiberius

16. European Language Resources Coordination Initiative (ELRC)

Het INT is betrokken bij het *ELRC*-initiatief. *ELRC* heeft als doel tekstdata te verzamelen, in alle EU-lidstaten, IJsland en Noorwegen, die gebruikt kunnen worden om CEF eTranslation verder te ontwikkelen. CEF eTranslation is een automatische vertaaldienst die door de Europese Commissie ter beschikking wordt gesteld om meertalige communicatie tussen openbare diensten, ministeries en burgers mogelijk te maken. De kwaliteit van een automatische vertaling hangt onvermijdelijk samen met de kwaliteit en kwantiteit van de taalbronnen die worden gebruikt om het systeem te "trainen". Grote hoeveelheden taaldata zijn dan ook nodig om de kwaliteit van de Nederlandse vertalingen te verbeteren. Dit alles is van groot belang om taalbarrières in Europa te slechten en de nationale talen, in dit geval de Nederlandse taal, te behouden in de digitale informatiemaatschappij.
Carole Tiberius is samen met Jan Odijk (Universiteit Utrecht) Technical National Anchor Point voor Nederland voor het *ELRC*-project.

Maandelijks wordt er door *ELRC* een online Q&A-sessie gehouden om de voortgang van het project te bespreken. Daarnaast is er twee keer per jaar een Language Resource Board-bijeenkomst. De eerste vond plaats op 19 april in Nice en werd bijgewoond door Jan Odijk. Op 19-20 september vond de tweede LRB-bijeenkomst plaats in Parijs. Namens het INT was Carole Tiberius hierbij aanwezig

Op 5 oktober 2018 is een tweede nationale workshop (de eerste was in 2016) voor potentiële dataleveranciers gehouden. De focus lag deze keer op de Digitial Service Infrastructures (DSIs). Deze workshop werd medegeorganiseerd door het INT. Namens het INT waren Carole Tiberius en Bob Boelhouwer aanwezig. In de laatste maanden van 2018, bestonden de *ELRC*-werkzaamheden voornamelijk bestaan uit het verzamelen van de taalbronnen die tijdens of na de workshop waren geïdentificeerd. Deze werkzaamheden lopen in 2019 door.

17. CLARIAH en CLARIAH-PLUS

In 2015 ging het NWO-project CLARIAH van start: Common Lab Research Infrastructure for the Arts and Humanities. In 2018 is het INT-aandeel in dit project grotendeels afgerond.

Binnen Werkpakket 2 (infrastructuur) is het project Diamant (diachroon semantisch lexicon) afgerond met een nieuwe dataversie. Aan de *PICCL*-webapplicatie, die workflows voor OCR, conversie en taalkundige verrijking implementeert, wordt nog de laatste hand gelegd om de connectie met Autosearch te maken. Binnen *CLEVER* zijn best practices gedocumenteerd voor CLARIN-centra om samen met software- en dataontwikkelaars tot duurzaam beheerbare applicaties te komen.

Binnen Werkpakket 3 (taalkunde) is vooral gewerkt aan tools die de doorzoekbaarheid van taalkundige data bevorderen. De onderdelen *OpenSonar*$^{+}$ en *Autosearch* hebben geleid tot een solide en gebruikersvriendelijk corpuszoeksysteem, waarin niet alleen de belangrijke corpora Sonar en CGN beschikbaar gemaakt zijn, maar waarin gebruikers (Autosearch) ook hun eigen corpora kunnen maken. Ook andere door het INT gepubliceerde corpora zullen profiteren van deze ontwikkeling.
De onderdelen *federated search* en *chaining search* worden in het eerste kwartaal van 2019 afgerond. Het resultaat is dat corpora, treebanks en lexica op een uniforme manier en in combinatie doorzoekbaar zijn in de CLARIAH infrastructuur.

De research pilots *NAMES*, waarin een referentiedatabase voor de normalisering van eigennaamvarianten is ontwikkeld, *SERPENS*, waarin de historische ontwikkeling van de publieke perceptie onderzocht is van met name als schadelijk ervaren diersoorten aan de hand van het KB-krantencorpus (http://www.delpher.nl) en *DB:CCC* (text en concept mining om de goederen en mensenstromen samenhangende met de diamanthandel in Borneo beter in beeld te krijgen) zijn afgerond. Het INT is bij deze projecten verantwoordelijk geweest voor de database-infrastructuur en de koppeling van lexicale data. De research pilot *TICCLAT* (extractie van diachrone lexicale informatie, met name met behulp van TICCL en het Nederlab-corpus) loopt nog.

In 2018 is voorts het vervolgproject CLARIAH-PLUS gehonoreerd, waarin infrastructuur wordt toegevoegd voor de disciplines die teksten niet op de gebruikte taal, maar op de inhoud bestuderen, zoals letterkunde, geschiedenis, maar ook filosofie en theologie. Het INT zal hierin verder ontwikkelen aan de doorzoekbaarheid van taalmateriaal (met name treebanks en parallelle corpora, alsmede de mogelijkheid voor onderzoekers om zoekresultaten te annoteren) en op twee manieren bijdragen aan de ontwikkeling van (verrijkt) tekstmateriaal voor onderzoek: door de ontwikkeling van een infrastructuur voor het taalkundig verrijken van historisch tekstmateriaal en het opzetten van een digitalisatieworkflow.

18. CLARIN ERIC

Medewerkers van het INT vertegenwoordigen Vlaanderen in verscheidene overlegorganen van het Europese infrastructuurproject CLARIN (Common Language Resources and Technology Infrastructure) en zijn op die manier actief betrokken bij de beleidsvorming.

Als vertegenwoordiger van Vlaanderen binnen de CLARIN ERIC en als CLARIN B-centrum[3] voor Vlaanderen heeft het INT o.a. als taak om de zichtbaarheid van de CLARIN-infrastructuur bij Vlaamse onderzoekers en studenten te vergroten en om het gebruik van de CLARIN-diensten en -materialen actief te promoten. Het afgelopen jaar werd hier een prioriteit van gemaakt.
In de lente van 2018 heeft CLARIN DLU/Flanders meegewerkt aan de Tour de CLARIN: het consortium werd voorgesteld, er werden tools en materialen belicht (Text2Picto, Picto2Text, Corpus Hedendaags Nederlands) en er werd een interview gepubliceerd met de onderzoekster Cora Pots (KU Leuven) over haar onderzoek en haar ervaringen i.v.m. CLARIN.
Het INT verzorgde CLARIN-infosessies op de Universiteit Gent, de KU Leuven en de Vrije Universiteit Brussel en er waren INT-medewerkers als CLARIN-vertegenwoordigers aanwezig op congressen (al dan niet met een CLARIN-stand). Tijdens deze contactmomenten met het Vlaamse onderzoeksveld bleek o.a. dat er behoefte was aan informatie over OCR-technieken, wat resulteerde in een door het INT georganiseerde OCR-workshop op de KU Leuven eind 2018.

En ten slotte werd er het afgelopen jaar ook een nieuwe INT-folder ontwikkeld waarin verduidelijkt werd wat het INT als CLARIN-centrum te bieden heeft en werd het voor EWI bedoelde rapport "CLARIN ERIC Toegevoegde waarde voor onderzoek en ontwikkeling in Vlaanderen", geschreven samen met de Taalunie, de consortiumpartners van CLARIN DLU/Flanders en CLARIN ERIC.

Ook in 2019 zal er actief gewerkt worden aan het promoten van alle aspecten van de CLARIN-infrastructuur in Vlaanderen en zullen er workshops georganiseerd worden, bv. een Nederlab-workshop, met de bedoeling om onderzoekers te informeren over de inhoud en de mogelijkheden van het Nederlab-onderzoeksportaal.

Het INT CLARIN B-centrum heeft in 2018 het 'Core Trust Seal' gekregen, een kwaliteitskeurmerk voor duurzame en betrouwbare data-infrastructuren. Sinds 2014 is het INT gecertificeerd als CLARIN B-centrum met het 'Data Seal of Approval', dat sindsdien opgegaan is in het 'Core Trust Seal'. In 2017 was de certificering verlopen. Daarop is een nieuwe beoordelingsronde gestart die in 2018 met succes beëindigd werd.

Het afgelopen jaar werd er gewerkt aan de vindbaarheid van de INT-taalmaterialen binnen de CLARIN-infrastructuur door het uitbreiden en het optimaliseren van de relevante metadata. In 2018 intensifieerde de samenwerking met DARIAH Vlaanderen. Zo werd er een gezamenlijke CLARIN/DARIAH-vragenlijst voor Vlaamse onderzoekers uit de humane wetenschappen opgesteld. Het doel was een aantal noden en wensen betreffende digitale bronnen te inventariseren. Sommige respondenten werden persoonlijk gecontacteerd om hun antwoorden te bespreken en hen te wijzen op bestaande of toekomstige oplossingen voor de door hen aangebrachte kwesties. DARIAH Vlaanderen

[3] Een CLARIN B-centrum is een leverancier van data en (web)services, voornamelijk aan onderzoekers en studenten die deel uitmaken van de CLARIN-gemeenschap.

was ook medeorganisator van de CLARIN-infosessies op de universiteiten. En ten slotte werd er ook ingezet op een formalisering van de samenwerking. CLARIN DLU/FLanders en DARIAH Vlaanderen schreven namelijk samen een CLARIAH-VL-projectvoorstel, waarbij het opzetten van een openservice-infrastructuur met gebruiksvriendelijke state-of-the-art tools en data voor de humane wetenschappen centraal stond.

19. Systeembeheer

Het belangrijkste wat er op het gebied van backend en systemen is gebeurd heeft betrekking op het faciliteren van de nieuwe CLARIAH-projecten die vooral in het begin van het jaar veel resources vereisten. Het betrof hierbij projecten als TICCL, Chaining- en Federated search en CLEVER. Voor de nieuwe terminologie-gerelateerde activiteiten zijn verschillende applicaties opgezet. Ook zijn de *Termenlijst* en *Termenbank* in overdracht van de NTU naar het INT. Verder is de storagecapaciteit uitgebreid om ook projecten met visuele data te kunnen faciliteren en is een deel van de kantoor-hardware zowel op locatie Antwerpen als locatie Leiden voor een deel vernieuwd. Personeelstechnisch is er afscheid genomen van 1 systeembeheerder en is er geworven voor een nieuwe. Deze is in januari 2019 begonnen.

20. IT-afdeling: projecten en organisatie

Projecten
Dit jaar is hard gewerkt om de CLARIAH-projecten af te ronden. Dat is grotendeels gelukt: de onderdelen Federated Content Search (corpora), PICCL, AutoSearch, NAMES, DIAMANT en CLEVER zijn gereed. De onderdelen OpenSonar+ (met extra zoekmogelijkheden en verbeterde performance) en Federated Content Search (lexica en treebanks) zijn in een vergevorderd stadium en worden begin 2019 afgerond. Verder wordt in 2019 het (in overleg wat afgeslankte) onderdeel Chaining Search opgepakt.

De integratie van de taalmaterialen van de TST-Centrale in de INT-website is dit jaar afgerond. Het overnemen van de terminologiewebsite van de Taalunie is in gang gezet. In 2019 worden ook de producten Termtreffer en Termbeheerder in beheer genomen en online gebracht.

Aan Vertaalwoordenschat is de taal Portugees toegevoegd; Estisch volgt in 2019. Een stagiair ontwikkelde daarnaast een mobiele app om deze vertaalwoordenboeken te raadplegen. Deze app is inmiddels beschikbaar voor zowel iOS als Android.
De verbeterde versie van de Geïntegreerde Taalbank (GTB) en ons aandeel in het Nederlab-project werden afgerond. Er werd veel gewerkt aan het geïntegreerde lexicon GIGANT en de verwerking van divers corpusmateriaal. Voor het Neologismenproject werd een workflow voor materiaalverzameling en een bewerkingsomgeving opgezet. Verder zijn databewerkingstools ontwikkeld voor MentalLex, een psycholinguïstische associatiedatabank.

Er is een project gestart om middels machine learning de taalkundige verrijking van historisch materiaal te verbeteren. In het kader van COST is het crowdsourcingplatform PyBossa uitgeprobeerd. Ook is voorbereidend werk verricht voor het koppelen van gegevens van verschillende projecten, waaronder Spelling, het *Algemeen Nederlands Woordenboek* (ANW), het Neologismenproject en het Woordcombinaties-project. In 2019 zal de Spelling-database de spil worden voor het koppelen van de andere projecten.

Organisatie
Begin 2018 is een nieuwe ontwikkelaar aangetrokken om de voortgang van onze projecten te waarborgen. Om met name onze corpuswerkzaamheden te versnellen, zal in 2019 een dataconversiespecialist worden gezocht.

Technologie
In 2018 is verder gewerkt aan het efficiënter maken van het ontwikkelproces, zoals het automatisch testen en bijwerken van servers. Eind 2018 is er op het INT een workshop Vue.js georganiseerd, om meer te leren over deze tijdbesparende technologie die snel aan populariteit wint. Ook werd een workshop gevolgd over de eXist XML database en werden een aantal systemen voor het ontwikkelen van applicaties vergeleken. Ook in 2019 zullen we onze kennis bijhouden middels research, workshops en (online) cursussen.

21. DSDD (Database of the Southern Dutch Dialects)

Het DSDD-project behelst de samenvoeging en standaardisering van drie grote dialectlexicografische databases, waarvan de samenstelling tientallen jaren heeft gekost: het *Woordenboek van de Vlaamse Dialecten* (WVD, 1972 -), het *Woordenboek van de Brabantse Dialecten* (WBD, 1961-2005) en het *Woordenboek van de Limburgse Dialecten* (WLD, 1961-2008). De woordenboeken werden 40 à 50 jaar geleden parallel opgezet, om een uiteindelijke samenvoeging mogelijk te maken. Ondanks deze parallelle opzet lopen de woordenboeken zowel methodologisch als in de keuze van de behandelde concepten uiteen. Ook in technisch opzicht (bestandsformaten en logische structuur) is sprake van heterogeniteit.

De geïntegreerde database heeft tot doel vernieuwend onderzoek mogelijk te maken, vooral op het gebied van kwantitatieve lexicologie en dialectgeografische analyse. Het project, dat loopt van 1/1/2017 tot 30/4/2020, zal resulteren in geharmoniseerde dataset, een API ten behoeve van onderzoekers en een portaalapplicatie waarin het resultaat van het project voor een breed publiek beschikbaar zal zijn.

Het INT heeft een relationele databasestructuur ontwikkeld, alsmede een daarop werkende, met behulp van het lex'it platform geïmplementeerde bewerkingsomgeving waarin de integratie wordt doorgevoerd. De datastructuren zijn geharmoniseerd; de woordenboeken worden onderling gekoppeld door het toevoegen van een overkoepelende laag concepten.

In 2018 is vooral verder gewerkt aan de database-inhoud. Concepten zijn gekoppeld aan lemmata in de drie woordenboeken; definities uit WLD en WBD zijn toegevoegd aan de database; de inhoud van de lemma- en trefwoordvelden is gedeeltelijk opgeschoond, en de onderdelen vaktaal en landbouw zijn bijgeladen in de database. Daarnaast is gewerkt aan het ontwerp voor de uiteindelijke portaalapplicatie.

Bijlage 1: Raad van Toezicht en Raad van Advies

Raad van Toezicht

Drs. Paul Rüpp, voorzitter
Drs. Gertine van der Vliet
Prof. mr. Jan Cerfontaine

Raad van Advies
Prof. dr. Antal van den Bosch (voorzitter)
Prof. dr. Willy Vandeweghe (vicevoorzitter)
Prof. dr. Dirk Geeraerts
Prof. dr. Veronique Hoste
Prof. dr. Reinhild Vandekerckhove
Prof. dr. Jack Hoeksema
Prof. dr. Paula Fikkert
Drs. Jan Jaap Knol
Lic. Wim Vanseveren
Prof. dr. Gert Oostindie

Bijlage 2: Medewerkers

Prof. dr. Frieda Steurs - directeur/bestuurder

Dr. Jan Theo Bakker – systeemarchitect
Petra Belt – administrateur
Marjolijn van Bennekom – taalkundig assistent
Dr. Bob Boelhouwer – computerlinguïst
Lic. Lut Colman – onderzoeker/taalkundige
Iulianna Ciudin MA – onderzoeker/taalkundige
Peter Dekker MSc– systeemontwikkelaar
Lic. Griet Depoorter – CLARIN-coördinator
Lic. Katrien Depuydt – onderzoeker/taalkundige
Dr. Jesse de Does – computerlinguïst
Laura van Eerten MA – communicatieadviseur
Drs. Mathieu Fannee – programmeur/ontwikkelaar
Dr. Karlien Franco (tot 31 december 2018) – onderzoeker/taalkundige
Lic. Dirk Geirnaert – onderzoeker/taalkundige
Dr. Kris Heylen – onderzoeker/taalkundige
Dr. Dirk Kinable – onderzoeker/taalkundige
Drs. Marco van der Laan – systeemontwikkelaar/programmeur
Dr. Frank Landsbergen – computerlinguïst
Jorrit Linnert (tot 1 oktober 2018) – programmeur/ontwikkelaar
Koen Mertens – programmeur/ontwikkelaar
Drs. Jan Niestadt – systeemontwikkelaar/programmeur
Wil de Ruyter – taalkundig assistent
Dr. Tanneke Schoonheim – onderzoeker/taalkundige
Rob van Strien – programmeur/ontwikkelaar
Dr. Josefien Sweep (tot 5 maart 2018) – onderzoeker/taalkundige
Paulette Tacx – managementassistent
Drs. Rob Tempelaars – onderzoeker/taalkundige
Dr. Carole Tiberius – computerlinguïst
Lic. Katrien Van pellicom – projectleider Spelling
Dr. Vincent Vandeghinste – onderzoeker/taalkundige
Drs. Boukje Verheij – taalkundig assistent, vanaf 1 september 2018 onderzoeker/taalkundige
Drs. Vivien Waszink – onderzoeker/taalkundige
Karin van Weerlee – managementassistent
Dr. Hans Westgeest – onderzoeker/taalkundige

Bijlage 3: Publicaties, lezingen, media etc.

Publicaties

Colman, Lut, Carole Tiberius (2018). A Good Match: a Dutch Collocation, Idiom and Pattern Dictionary Combined. In: *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, Ljubljana, Slovenia, 17-21 July 2018. 233-246.

Depuydt, Katrien and Jesse de Does (2018), "The Diachronic Semantic Lexicon of Dutch as Linked Open Data." In: I. Kernerman and S. Krek, *Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets".* [Miyazaki], 2018, pp. 23-28.

Eerten, Laura van (2018). Van astronaut tot vlinder. De namen van vuurwerk. In: *Onze Taal* 87, 12, p. 18-19.

Eerten, Laura van, Rob Tempelaars, Vivien Waszink (2018), Lubberiaans Lexicon. Op: www.ivdnt.org.

Eerten, Laura van, Vivien Waszink (2018). Bijdragen aan: Taalkalender 2019.

Eerten, Laura van, Vivien Waszink (2018). Coachwoordenboekje. Op: www.ivdnt.org.

Eerten, Laura van, Vivien Waszink (2018). Kids, koffietjes & comfortzone. Op: www.taalvoutjes.nl.

Erp, Marieke van, Jesse de Does, Katrien Depuydt , Rob Lenders and Thomas van Goethem (2018) . Slicing and Dicing a Newspaper Corpus for Historical Ecology Research. *Proceedings of European Knowledge Acquisition Workshop – EKAW* (2018).

Gantar, Polona, Lut Colman, Carla Parra Escartín, Hector Martínez Alonso (pre-publicatie). Multiword Expressions: Between Lexicography and NLP. *International Journal of Lexicography*.

Id-Youss H., Steurs F., Alswlaiman AA. (2018). Legal-based Ontologies Between serious needs and challenging realities. In: *TOTh 2017 Proceedings* (249-264). Presented at the Toth, Chambery, France, 08 Jun 2017-09 Jun 2017. Chambery. ISBN: 978-2-919732-80-7.

Kosem, Iztok, Kristina Koppel, Tanara Zingano Kuhn, Jan Michelfeit, Carole Tiberius (2018) Identification and automatic extraction of good dictionary examples: the case(s) of GDEX, *International Journal of Lexicography*, ecy014, Op: https://doi.org/10.1093/ijl/ecy014

Krek, Simon, Iztok Kosem, John McCrae, Roberto Navigli, Bolette S. Pedersen, Carole Tiberius, Tanja Wissik (2018). European Lexicographic Infrastructure (ELEXIS). In: *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts (EURALEX 2018)*, Ljubljana, Slovenia, 17-21 July 2018. 881-891.

Pancirov Cornelisse Z., Žagar AM., Steurs F. (2018). Vertalen voor toerisme: een oefening in persuasieve teksten. Casus: de Nederlandstalige brochure over Zagreb. In: *Regionaal Colloquium Neerlandicum 2017 Op reis!* Presented at the Regionaal Colloquium Neerlandicum 2017, Wroclaw, Poland, 24 May 2017-27 May 2017. (professional oriented).

Pedersen, Bolette S., John McCrae, Carole Tiberius, Simon Krek (2018). ELEXIS - a European infrastructure fostering cooperation and information exchange among lexicographical research communities. In: Francis Bond, Takayuki Kuribayashi, Christiane Fellbaum and Piek Vossen (eds) *Proceedings of the 9th Global WordNet Conference* (GWC 2018), Global Wordnet Association, Singapore. 339-344.

Schoonheim, Tanneke. *Panama, Paddenpoel, Popswoude and Poederooijen. Geographical names in the historical dictionaries of Dutch*. In: Rolf Bergmann und Stefanie Stricker (ed.). *Namen und Wörter. Theoretische Grenzen – Űbergänge im Sprachwandel*. Heidelberg: Universitätsverlag Winter, 2018. Pp. 145-169.

Steurs F. (2018). Taal en variatie. Variatie en taal. Het INT als schatkamer van alle taalmaterialen. In: Colleman T., De Caluwe J., De Tier V., Ghyselen A., Triest L., Vandenberghe R., Vogl U. (Eds.), *Woorden om te bewaren*(135-149). Gent: Skribis Gent. ISBN: 9789492944221.

Steurs F. (2018). Taal is business: (vak)taal als centrale schakel voor de economie en samenleving (accepted). In: *Dagen van het Nederlands "Het (on)vertaalbare vertaald"* Presented at the Twee dagen van het Nederlands, Lublin,Polen, 09 Nov 2017-10 Nov 2017. doi: 10.18290/rh

Tempelaars, Rob, Vivien Waszink, Tanneke Schoonheim (2018). Internetwoordenboekje. Op: www.ivdnt.org.

Tempelaars, Rob, Vivien Waszink, Rubriek 'Neologisme van de week'. Op www.ivdnt.org.
Van der Lek-Ciudin I., Rigouts Terryn A., Heyman G., Lefever E., Steurs F. (2018). Translator's methods of acquiring domain-specific terminology. Information retrieval in terminology using lexical Knowledge Patterns. In: *Proceedings of the 21st European Symposium on Languages for Special Purposes* (LSP) Presented at the European Symposium on Languages for Special Purposes (LSP), Bergen, Norway, 28 Jun 2017-30 Jun 2017. Bergen.
Vandeghinste V., Steurs F. (2018). Nederlandse Taaltechnologie voor Blinden. *Dixit*.
Vandeghinste V., Vanallemeersch T., Bulté B., Augustinus L., Van Eynde F., Pelemans J., Verwimp L., Wambacq P., Heyman G., Moens M-F., Van der Lek-Ciudin I., Steurs F., Rigouts Terryn A., Lefever E., Tezcan A., Macken L., Coppers S., Brulmans J., Van den Bergh J., Luyten K., Coninx K. (2018). Smart Computer-Aided Translation Environment (SCATE): Highlights. (367-367). Presented at the 21st annual conference of the European Association for Machine Translation - EAMT 2018, Alacant, Spain, 28 May 2018-30 May 2018.
Waszink, Vivien (2018), Niemendalletje, onesie en slimme jas. Klerenwoorden en modetaal. In: *Onze Taal*, februari/maart 2018.
Waszink, Vivien (2018). Kleedjes, onesies en rokjesdag. Klerenwoorden & modetaal. Op: www.vandale.nl.
Waszink, Vivien (2018). *Klerenwoorden en modetaal*. Van Dale Uitgevers, Utrecht.
Waszink, Vivien (2018). Klerenwoorden en modetaal. Op: www.taalvoutjes.nl.
Waszink, Vivien, Alex Reuneker, Ton van der Wouden (2018). Als ik praat, dan praat ik money. De hiphopste woorden. Op: www.neerlandistiek.nl.
Waszink, Vivien, Laura van Eerten (2018). *Kids, koffietjes & comfortzone. Waarom taal soms irritant is*. Uitgeverij AUP, Amsterdam.
Waszink, Vivien (2018): Taalbaas. Op: KRO/NCRV Data.

Lezingen, presentaties en onderwijs

Bloothooft, Gerrit, David Onland, Martin Reynaert, Katrien Depuydt, Tanneke Schoonheim. *NAMES*. Posterpresentatie op de CLARIAH toogdag. Den Haag, 9 maart.
Bloothooft, Gerrit, David Onland, Martin Reynaert, Katrien Depuydt, Tanneke Schoonheim. *Towards gold standards for personal names*. Posterpresentatie op DH Benelux. Amsterdam, 8 juni.
Bloothooft, Gerrit, David Onland, Martin Reynaert, Katrien Depuydt, Tanneke Schoonheim. *NAMES* Posterpresentatie op de CLARIAH toogdag. Amsterdam, 19 oktober.
Colman, Lut. *Van woorden kennen naar woorden gebruiken*. Presentatie op de LOWAN-studiedag. Amersfoort, 10 april 2018.
Colman, Lut. *Woordcombinaties*. Presentatie op de workshop Formulaic Language Processing and acquisition Research (FLIPR), georganiseerd door de projectgroep Idiomatic Second Language Acquisition (ISLA). Nijmegen, 18 juni 2018.
Colman, Lut. *Woordcombinaties*. Presentatie Mathias De Vriesgenootschap. Leiden, 27 juni 2018.
Tiberius, Carole. *A good match: a Dutch collocation, idiom and pattern dictionary combined*. Presentatie op het Euralex-congres. Ljubljana, 19 juli 2018.
Colman, Lut, Vandeghinste, Vincent. *Woordcombinaties: van woorden kennen naar woorden gebruiken*. Posterpresentatie IVN-colloqium. Leuven, 29 augustus 2018.
Dekker, Peter, Tanneke Schoonheim. *Recognizing blends: first experiments with Pybossa*. Presentatie op de enetCollectbijeenkomst van de werkgroepen 3 en 5. Leiden, 24 oktober.
Dekker, Peter, Tanneke Schoonheim. When to use PYBOSSA? Case studies on crowdsourcing for Dutch. Presentatie op de enetCollectworkshop *Learning materials through crowdsourcing: teachers, perspectives & scenarios.* Gotenburg (Zweden), 6 december.
Depuydt, Katrien, Maarten van Gompel, Jesse de Does, Hennie Brugman, Gosse Bouma. *Nederlab progress and challenges in linguistic enrichment of historical Dutch texts*. CLIN in Nijmegen op 26 januari 2018
Depuydt, Katrien en Jesse de Does hebben meegewerkt aan de presentatie:Marieke van Erp, Jesse de Does, Thomas van Goethem and Katrien Depuydt (2018) *Good lynx, bad lynx: document enrichment for historical ecologist.* Gepresenteerd door Marieke van Erp op de CLIN in Nijmegen op 26 januari.

Depuydt, Katrien. Presentatie CLARIAH Namesproject op de CLARIAH Toogdag van 9 maart 2018 in Den Haag.

Depuydt, Katrien, Jesse de Does. *The Diachronic Semantic Lexicon of Dutch as Linked Open Data.* LREC in Miyazaki, 8 mei op de Globalex 2018 pre-conference workshop "Lexicography and Wordnets"

Depuydt, Katrien, Margit Rem, Nicoline van der Sijs. *CLVN Corpus Laatmiddel-en Vroegnieuwnederlands.* Workshop "*Using Nederlab for Humanities Research*" op 6 juni 2018 op de Digital Humanities Benelux conferentie in Amsterdam

Depuydt, Katrien. *INT - corpus creation/preparation and provenance.* CLARIAH provenance workshop, 3 september 2018

Depuydt, Katrien. *From Dictionaries to a Lexicographical Infrastructure for Historical Dutch.* Leuven, 28 september 2018

Depuydt, Katrien. *Niks, bietekwiet!* Slotevent van Nederlab, 26 oktober 2018

Eerten, Laura van, Vivien Waszink. *Kids, koffietjes & comfortzone*. Boekpresentatie. Galerie Café Leidse Lente, Leiden, 20 november 2018.

Hofmeester, Karin Ashkan Ashkpour, Martin Reynaert, Katrien Depuydt, Jesse de Does, Marieke van Erp. *Diamonds in Borneo: Commodities as Concepts in Context.* CLARIAH Toogdag, vrijdag 19 oktober 2018

Schoonheim, Tanneke. MA-college *Word Senses and Definitions.* Universiteit Leiden, 25 april.

Schoonheim, Tanneke. *Former specialist terms and the general Dutch vocabulary.* Lezing op de workshop *The revival of the lexicon: on the crossroads of lexicology, lexicography and terminology* op het 20ste International Congress of Linguists. Kaapstad (Zuid-Afrika), 5 juli.

Schoonheim, Tanneke. *Language Variety at the Dutch Language Institute.* Presentatie op de 16de EFNIL-conferentie *Language variation: a factor of increasing language complexity and a challenge for language policy within Europe.* Amsterdam, 12 oktober.

Steurs, Frieda. *Clarin, Clariah and Dariah: Towards a complete Infrastructure for Digital Humanities in Europe.* Viering 30 jaar ESAT, KU Leuven, 30 januari 2018

Steurs, Frieda. *Taal is Business! Over de impact van taalkundige toepassingen in de samenleving*. Talencongres Groningen, 16 februari 2018

Steurs, Frieda, Branislav Bedi. *User oriented design strategies for a competitive solution.* Openingslezing Annual conference van het COST Netwerk "enetCollect: Computer Assisted Language Learning and Crowd Sourcing Techniques", 15 maart 2018

Steurs, Frieda. *The Dutch language institute: Developer, keeper and distributor of sustainable language resources*. CLARIN-workshop aan de KU Leuven, faculteit Letteren, 20 maart 2018

Steurs, Frieda. *Language is Business: the Single Digital Market and Language and Cultural Diversity*. CIPSH-congres Xiamen, 14-19 april 2018

Steurs, Frieda. *Het INT: de schatkamer van alle taalmaterialen*. De Orde van den Prince in Leiden, 29 mei 2018

Steurs, Frieda. *The narrow line between lexicology and terminology: the case of the legal dictionaries*. ICL-congres in Kaapstad, 5 juli 2018

Steurs, Frieda. *On the future of the translation profession*. Openingslezing derde Translation Technology Summer School, 3 t/m 7 september 2018

Steurs, Frieda. Lezing op het afscheid van Prof Dr Jacques van Keymeulen en overhandiging Matthias de Vriespenning voor bewezen verdiensten aan de lexicografie, 14 september 2018

Steurs, Frieda. *The changing profile of the translator: a challenge for universities*. "Translation: past-present-future", congres dat werd georganiseerd door het Europees comité van de regio's en het Europees economisch en sociaal comité, 28 september 2018

Steurs, Frieda. *Terminology Management in Technical Documents*; masterclass voor studenten van de KU Leuven, 9 oktober 2018

Steurs, Frieda. Openingslezing jaarlijkse EFNIL-congres, Brakke Grond Amsterdam, 11 oktober 2018

Steurs, Frieda. *Algemene en vakspecifieke woordenschat: valkuilen voor vertalers*. Expertisecentrum Nederlandse Taalkunde in Wroclaw, 4 t/m 6 december 2018

Tempelaars, Rob. 'Krijg een loopoor (van hier tot Hollands Spoor)! Over vloeken, scheldwoorden en andere krachttermen'. Webinar van twee uur voor Kennisnet voor Taal en Vakopleidingen (KTV), 15 januari 2018

Tempelaars, Rob. 'Algemeen Nederlands Woordenboek (ANW)'. Presentatie ten behoeve van de leden van de Raad van Toezicht van het INT, 30 mei 2018

Tiberius, Carole. *ELEXIS-WP1: Lexicographic Data and Workflow.* Presentatie op ELEXIS-kickoff. 16 februari 2018, Ljubljana.

Tiberius, Carole. *ELEXIS – a European Infrastructure for Lexicography*. Presentatie op LT Summit. 29 mei 2018, Brussel.

Tiberius, Carole. *A good match: a Dutch collocation, idiom and pattern dictionary combined*. Presentatie op het EURALEX-congres. Ljubljana, 19 juli 2018.

Tiberius, Carole. MA-collegereeks *Corpus Lexicography*. Universiteit Leiden, februari-mei 2018.

Van Keymeulen Jacques, Sally Chambers, Veronique De Tier, Jesse de Does, Katrien Depuydt, Schoonheim, Tanneke , Roxane Vandenberghe, Lien Hellebaut. *Sustaining the Dictionary of the Southern Dutch Dialects (DSDD): a case study for CLARIN and DARIAH*. Posterpresentatie op de CLARIN Annual Conference. Pisa, 8-10 oktober

Waszink, Vivien. Straattaal. Lezing op Leidse Scholierenconferentie ´'Migration first, Netherlands second?!' van Leiden Global. Leiden, 2 februari 2018.

Waszink, Vivien. Boekpresentatie Klerenwoorden en modetaal. Boekhandel Kooyker, Leiden, 13 februari 2018.

Waszink, Vivien. Kennisclip voor de opleiding Nederlands van de Universiteit Utrecht. Utrecht, mei 2018.

Waszink, Vivien. Lezing over hiphoptaal op de Carrièredag Nederlandse Taal en Cultuur Universiteit Groningen. Groningen,, 16 mei 2018.

Waszink, Vivien. Lezing Rymklets en zingzeggen op het Zuid-Afrika Huis. Amsterdam, 9 november 2018.

Waszink, Vivien. Het Algemeen Nederlands Woordenboek. Lezing op de Onderzoeksmarkt Association des Néerlandistes de Belgique francophone et de France. Lille, 7 december 2018.

Waszink, Vivien, Alex Reuneker, Ton van der Wouden. Rapshit is mijn taal. Lezing op de Onderzoeksmarkt Association des Néerlandistes de Belgique francophone et de France. Lille, 7 december 2018.

Waszink, Vivien. Het jaar in taal. de Buren, Brussel, 18 december 2018.

Wouden, Ton van der, Walter Haeseryn, Maaike Beliën, Adrienne Bruyn, Mario van de Visser, Timothy Colleman, Matthias Hüning, Frank Landsbergen. De Nieuwe ANS. Panelsessie op Nederlands in Beweging: het 20e Colloquim Neerlandicum. Leuven, 29 augustus 2018.

Deelname en organisatie congressen en workshops

Colman, Lut. Deelname aan de workshop *Formulaic Language Processing and acquisition Research (FLIPR)*, georganiseerd door de projectgroep Idiomatic Second Language Acquisition (ISLA) (18-19 juni, Nijmegen).

Colman, Lut. Deelname aan de *Dag van de Nederlandse zinsbouw* (21 december, Gent).

Colman, Lut. Deelname aan: *Rondetafelconferentie NTU taalvariatiebeleid* (14 maart, Rotterdam).

Depoorter, Griet, Frieda Steurs. Organisatie CLARIN-workshop KU Leuven, 20 maart 2018

Depuydt, Katrien en Jesse de Does hebben bijgedragen aan de *Serpens* workshop op 25 januari in Nijmegen.

Depuydt, Katrien en Jesse de Does hebben deelgenomen aan de workshop ter ontwikkeling van TEI Lex-0 op 2 en 3 mei in Berlijn.

Depuydt, Katrien en Jesse de Does hebben een workshop over OCR gegegeven bij de KU Leuven voor CLARIN Vlaanderen op 22 november.

Depuydt, Katrien heeft deelgenomen aan de Mentallex workshop op 31 oktober in Amsterdam

Depuydt, Katrien en Jesse de Does hebben deelgenomen aan de workshop van Ontolex over de ontwikkeling van Ontolex over lexicografische data op het INT in Leiden op 5 november 2018.

Eerten, Laura van, Peter Dekker, Marjolijn van Bennekom. Laborant voor Taalradar, op het DRONGO talenfestival in Nijmegen, 10 november 2018.

Schoonheim, Tanneke. Medeorganisatie van en deelname aan de Streektaalconferentie *De (on)zin van taal- en dialectpromotie*. (8 juni, Leeuwarden).

Schoonheim, Tanneke. Deelname aan het 20ste International Congress of Linguists (2 - 6 juli, Kaapstad, Zuid-Afrika).

Schoonheim, Tanneke. Medeorganisatie van en deelname aan het Euralexcongress *Lexicography in global contexts* (17 - 21 juli, Ljubljana, Slovenië).

Schoonheim, Tanneke. Medeorganisatie van en deelname aan de enetCollectbijeenkomst van de werkgroepen 3 en 5 (24 - 25 oktober, Leiden).

Schoonheim, Tanneke. Medeorganisatie van en deelname aan de 16de EFNIL-conferentie *Language variation: a factor of increasing language complexity and a challenge for language policy within Europe* (10 - 12 oktober, Amsterdam).

Schoonheim, Tanneke. Deelname aan de enetCollectworkshop *Learning materials through crowdsourcing: teachers, perspectives & scenarios.* (5 - 7 december, Gotenburg, Zweden).

Steurs, Frieda. Mede-organisator van het Tolk- en vertaalcongres "The Language Industry 4.0: Embracing the future?" in Breda. Zij organiseerde er twee workshops samen met Iulianna Van der Lek-Ciudin. Workshop 1 : Are we ready to embrace Language Industry 4.0? Workshop 2: How will your CAT tool look like in 5-10 years from now?, 9-10 maart 2018

Steurs, Frieda, Iulianna van der Lek-Ciudin. Organisatie derde Translation Technology Summer School. Frieda Steurs gaf de openingslezing en verzorgde daarna een onderdeel van de workshop: "Futurebound - A matrix for translation professionals", 3 t/m 7 september 2018

Steurs, Frieda. Medeorganisatie van en deelname aan de 16de EFNIL-conferentie *Language variation: a factor of increasing language complexity and a challenge for language policy within Europe* (10 - 12 oktober, Amsterdam).

Steurs, Frieda. Deelname aan CIPSH-congres in Xiamen als vertegenwoordiger van CIPL, 14-19 april 2018

Tiberius, Carole. Deelname aan projectbijeenkomst: *ELEXIS-Kick off* (15-17 februari, Ljubljana)

Tiberius, Carole. Deelname aan workshop: *TEI Lex-0* (16 juli, Ljubljana)

Tiberius, Carole. Deelname aan het congres: *LT Summit* (28-29 mei, Brussel).

Tiberius, Carole. Deelname aan het congres: *EURALEX* (17-21 juli, Ljubljana).

Tiberius, Carole. Deelname aan projectbijeenkomst: *ELRC LRB meeting* (19-20 september, Parijs)

Tiberius, Carole. Deelname aan het congres: *TiNT-dag* (26 oktober, Antwerpen).

Tiberius, Carole. Deelname aan workshop: *Ontolex-Lemon* (5 november, Leiden)

Tiberius, Carole. Deelname aan projectbijeenkomst: *ELEXIS TMB* (6-7 november)

Tiberius, Carole. Deelname aan webinar: *Introduction to web analytics for on-line cultural heritage collections* georganiseerd door het IMPACT Centre of Competence (18 april 2018)

Tiberius, Carole. Medeorganisatie workshop: *ELRC workshop* samen met Jan Odijk in Den Haag, 5 oktober 2018

In de media

Depuydt, Katrien. Interview over het project Nederlab in het programma NPO Focus op radio 1, 24 oktober 2018

Eerten, Laura van, Tanneke Schoonheim. Interview over verhuizing naar het Rapenburg. *Taalinstituut terug aan Rapenburg.* In Leidsch Dagblad, 6 maart 2018.

Eerten, Laura van. Interview over Kids, koffietjes & comfortzone. *Kijk in de Vegte* Radio 2, 14 november 2018.

Eerten, Laura van, Vivien Waszink. Interview over Kids, koffietjes & comfortzone. *Boek van Instituut Nederlandse Taal over onuitroeibare taalergernissen*. In o.a. Leidsch Dagblad en Noord-Hollands Dagblad, 21 november 2018.

Eerten, Laura van, Vivien Waszink. Boekbespreking Kids, koffietjes & comfortzone. *Straattaal en dooddoeners in de schoen*. In NRC, 21 november 2018.

Eerten, Laura van, Vivien Waszink. Boekbespreking Kids, koffietjes & comfortzone. *5 uur live* RTL 4, 23 november 2018.

Eerten, Laura van, Vivien Waszink. Recensie Kids, koffietjes & comfortzone. *Tenenkrommende taalgrillen vormden ook in het afgelopen jaar weer goed materiaal voor Nederlandse schrijfsters.* In Trouw, 1 december 2018.

Eerten, Laura van, Vivien Waszink. Interview over Kids, koffietjes & comfortzone. *Lunchroom* Radio Noord-Holland, 11 december 2018.

Eerten, Laura van, Vivien Waszink. Interview over kantoorjargon. *Vitaliteit, je rol pakken en nog 9 clichés: stem op ergste kantoortaal.* In: AD Werkt, 13 december 2018.

Instituut voor de Nederlandse Taal. Item over Weg met dat woord en Kids, koffietjes & comfortzone. *De Wereld Draait Door*, 14 november 2018.

Tempelaars, Rob. Interview over vloeken, verwensingen en scheldwoorden voor het programma Nu Al Wakker, rubriek 'Hoe zit dat eigenlijk?', NPO radio 1, 9 januari 2018

Tempelaars, Rob. Interview over vloeken, verwensingen en scheldwoorden, Radiozender FunX, 24 januari 2018

Tempelaars, Rob. Interview met Jacqueline Postma over het WNT in verband met een te publiceren boek over de Nederlandse kunst en cultuur, 5 februari 2018

Tempelaars, Rob. Interview met journaliste Kaat Schaubroeck over groeten ten behoeve van een artikel in De Standaard. Gepubliceerd in De Standaard van 31 maart 2018 onder de titel '[Hoogachtende groetjes. De Basisbeginselen van de e-mail-etiquette]'. In gewijzigde vorm, onder de titel 'Hooggeachte groetjes. Tobben met de e-mail-etiquette', ook opgenomen in *Onze Taal* 87 (2018), 9 (september), pp. 22-24.

Tempelaars, Rob. Interview met redacteur RTL Nieuws, ter voorbereiding van een item over het vermijden en vervangen van (vermeend) discriminatoire woorden, zoals *blank* voor de huidskleur, dat door de NOS vervangen wordt door *wit*, 27 maart 2018

Tempelaars, Rob. Interview met Marieke Buijs (Quest Psychologie) over de vraag "Door welke woorden voelden mensen zich 100 jaar geleden beledigd?", 29 augustus 2018

Tempelaars, Rob. Interview EditieNL (RTL-tv) over de uitspraak van een rechter dat het woord kankerhomo inmiddels geen homofoob scheldwoord meer is, maar behoort tot het dagelijks taalgebruik (uitgezonden op 2 november 2018 om 18.15 uur).

Waszink, Vivien. Interview in de Volkskrant over teksten van rapper Boef. In: de Volkskrant, 3 januari 2018 en diverse radio-interviews over teksten van Boef (oa op Radio 1 en FunX)

Waszink, Vivien. Interview over boek Klerenwoorden en modetaal, NH Radio, 9 februari 2018.

Waszink, Vivien. Interview over boek Klerenwoorden en modetaal, Ekdom in de ochtend, Radio 2, 13 februari 2018.

Waszink, Vivien. Interview over boek Klerenwoorden en modetaal, Nachtkijkers, Radio 1, 19 februari 2018.

Waszink, Vivien. Interview over Klerenwoorden en modetaal. BNR Nieuwsradio, 20 februari 2018.

Waszink, Vivien (2018). Boekbespreking Klerenwoorden en modetaal. In: Trouw, 21 februari 2018.

Waszink, Vivien. Interview/debat met Sylvana Simons, Anne-Fleur Dekker en Marlies Dekkers over vrouwonvriendelijkheid in hiphopteksten. Op: onlineplatform Voor de Ommekeer, 7 maart 2018.

Waszink, Vivien. Interview over oa Klerenwoorden en modetaal. Wat een Week Show!, Radio 2, 16 maart 2018.

Waszink, Vivien. Interview over het Internetwoordenboekje. De Taalstaat, Radio 1, 5 mei 2018.

Waszink, Vivien. Interview over het woord upskirting. BNR Nieuwsradio, 10 juli 2018.

Waszink, Vivien. Interview over Taalbaas en over onderzoek naar de hiphopste woorden. De Taalstaat, Radio 1, 28 juli 2018.

Waszink, Vivien. Interview over Kids, koffietjes & comfortzone. Slam FM, 13 november 2018.

Waszink, Vivien. Interview over Kids, koffietjes & comfortzone. Thuis op 5 NPO Radio 5, 21 november 2018.

Waszink, Vivien. Interview over Kids, koffietjes & comfortzone. Haandrikman! Radio 5, 19 december 2018.

Waszink, Vivien. Taalpanel met Japke-d. Bouma, René Appel en Ewoud Sanders. De Taalstaat, Radio 1, 29 december 2018.

Diversen

Schoonheim, Tanneke:

- Voorzitter van de Commissie Spelling van de Nederlandse Taalunie
- Penningmeester van Euralex, European Association for Lexicography
- Penningmeester van de Stichting Nederlandse Dialecten
- Bestuurslid van de Stichting Rechtstaal (Juridisch Woordenboek Spaans – Nederlands)
- Lid van de Raad van Advies van Hogeschooltaal (Noordhoff Uitgevers)
- Redacteur van *Trefwoord*, digitaal tijdschrift voor lexicografie

Steurs, Frieda:

- Lid Wetenschappelijk comité van het Regionaal colloquium Neerlandicum, Bratislava.
- Voorzitter van de Stichting Bibliographie Linguistique
- Secretaris-generaal van CIPL (Comité International Permanent des Linguistes)
- Vice-chair van WG3 enetCollect
- Lid van de doctoraatsjury van Leen Sevens (14 december 2018). Onderwerp: Words Divide, Pictographs Unite: Pictograph Communication Technologies for People with an Intellectual Disability. KU Leuven, Centrum voor Computerlinguïstiek

Tempelaars, Rob:

- Dagvoorzitter leden Matthias de Vriesgenootschap, 27 juni 2018
- Redacteur en redactiesecretaris van *Trefwoord¸digitaal tijdschrift voor lexicografie*

Bijlage 4: Verslag Taalbank Nederlands

De Taalbank Nederlands omvat corpora en computationele lexica van zowel modern als historisch Nederlands. Het lexicon is bedoeld om ingezet te worden voor onderzoek in de vorm van diverse producten die niet alleen in principes van verrijking en structuur op elkaar aansluiten, maar bovendien in toenemende mate inhoudelijk aan elkaar gekoppeld zijn. Hieronder een stand van zaken.

I Lexica

I.1 GiGaNT
GiGaNT is het computationele woordvormenlexicon dat de door het INT beschreven woordenschat moet gaan bevatten van het Nederlands vanaf de 6e eeuw tot nu en de centrale database vormt van het Nederlands van het INT. In GiGaNT zit de formele beschrijving van de woorden. De semantische informatie zit in DiaMaNT, en in de diverse woordenboeken van het INT.
Werkzaamheden aan deze centrale data-infrastructuur zijn modulair opgezet. De twee grote componenten zijn GiGaNT-Molex, waarin de hedendaagse woordenschat wordt beschreven, en GiGaNT-Hilex, de historische lexiconcomponent. Voor de historische lexiconcomponent moeten de vier historische woordenboeken van het Nederlands (ONW, VMNW, MNW en WNT) de kern vormen, aangevuld met nieuw materiaal. De moderne lexiconcomponent bevat het hedendaags taalmateriaal. Afgelopen jaar is met name gewerkt aan de verdere integratie van de historische en de moderne lexiconmodule, en gewerkt aan de uitbreiding van de moderne lexiconcomponent. De koppeling tussen GiGaNT-Molex en het ANW is afgerond.

a. GiGaNT- HILEX: Historische lexiconcomponent

De werkzaamheden aan de historische lexiconcomponent hebben zich in 2018 geconcentreerd op enerzijds het verwerken van de update van de WNT-module, waar nieuw citatenmateriaal uit voortkwam dat als attestaties in de database moest worden toegevoegd. Daarnaast is verder gewerkt aan de toevoeging van de VMNW-module aan GiGaNT-hilex. Toevoeging van ONW-module en VMNW-module hopen we in 2019 te kunnen realiseren. Tot slot zijn er optimalisaties doorgevoerd aan lex'it, de bewerkingsomgeving voor de database.

b. GiGaNT- MOLEX: Moderne lexiconcomponent

De moderne lexiconcomponent wordt continu geactualiseerd. De werkzaamheden daarvoor gebeuren in de context van het project Spelling (zie aldaar). Daarnaast is een additionele subset van woorden die niet voor *woordenlijst.org* zijn geselecteerd, verwerkt. In 2019 zal een eerste release van het lexicon als downloadbare dataset worden gedaan.
Integratie GiGaNT-Molex - ANW
De in 2017 gestarte werkzaamheden om de moderne lexiconcomponent te koppelen aan het ANW is in 2018 afgerond. De werkzaamheden aan de ca. 56.000 koppelingen hebben ook tot substantiële verbeteringen geleid van zowel GiGaNT-Molex als het ANW. In 2019 wordt het resultaat van die koppelingen niet alleen toegevoegd aan het ANW. Er zal een workflow geïmplementeerd worden waarbij bij bewerking van nieuwe woorden in het Neologismenproject en het ANW de koppeling met Molex meteen tot stand wordt gebracht. Paradigmainformatie en controle spelling van nieuwe woorden in deze producten wordt daardoor binnen de context van GiGaNT-Molex gerealiseerd en via de koppeling ANW/Neologismen aan GiGaNT-Molex aan deze woordenboekproducten toegevoegd.

c. Integratie GiGaNT-Hilex - GiGaNT-Molex

Er is in het najaar van 2016 gestart met fase 1 van de integratie van de lexiconmodule gebaseerd op het WNT met de moderne lexiconcomponent GiGaNT-Molex (zie hierna). Net zoals voor de koppeling ANW-Molex is de koppeling automatisch tot stand gebracht, en wordt de koppeling

geverifieerd in IKEA (zie hierboven, bij integratie GiGaNT-Molex - ANW). In 2017 zijn de ruim 60.000 koppelingen een eerste keer nagelopen door de taalkundig assistenten van het INT. Omdat een koppeling historisch Nederlands – hedendaags Nederlands niet altijd evident is, is er eind 2017 gestart met een redactionele controle. Deze redactionele controle daarvan loopt nog steeds. Ook deze koppelwerkzaamheden leiden tot substantiële verbeteringen in zowel GiGaNT-Molex als ook in GiGaNT-Hilex.

I.2 DiaMaNT
In dit project, dat gedeeltelijk uitgevoerd is binnen CLARIAH, zal gewerkt worden aan de ontwikkeling van een Diachroon semantisch lexicon van het Nederlands (DiaMaNT). Dit project moet het ontwerp en bouwwijze en eerste versie opleveren van een diachroon semantisch lexicon. Het diachroon semantisch lexicon heeft als doel een hulpmiddel te bieden bij tekstontsluiting en bij het onderzoek naar begrippen door de eeuwen heen. Het lexicon legt relaties tussen woordvormen en betekeniseenheden (concepten), en plaatst deze in de tijd. De bedoeling van het diachrone semantische lexicon is om diachrone onomasiologie, d.i. de veranderende uitdrukking/verbalisatie van een concept, en semasiologie, d.i. de verschuiving van betekenis(nuance) van woorden in de tijd, systematisch vast te leggen op een zodanige wijze dat de informatie voor mens en computer bruikbaar is.
Enerzijds dient de onomasiologische component de zoekmogelijkheden, omdat gerelateerde historische concepten kunnen worden toegevoegd aan een zoekvraag (slager → beenhouwer, beenhakker, vleeshouwer; boer → landman). Anderzijds draagt de semasiologische component (het in kaart brengen van betekenisverandering), bij aan de toegankelijkheid van historische tekst door de gebruiker erop te attenderen dat woorden in een bepaalde historische context een geheel andere betekenis kunnen hebben; zo is bv. de oudste betekenis van appel 'vrucht in het algemeen' (dus ook peren, pruimen etc.). Het lexicon vormt een laag op GiGaNT-HILEX, en heeft dus als belangrijkste bron de historische woordenboeken van het INL. Er is gewerkt aan het datamodel, aan de controle en de uitbreiding van de data. Een verbeterde dataversie is opgeleverd aan CLARIAH. Er is gewerkt aan een applicatie om het lexicon in de huidige vorm doorzoekbaar te maken.

Conceptvorming
In het kader van een opdracht bij het college Computationele Lexicografie is door Amos van Baalen een inventarisatie en analyse uitgevoerd van de diverse typen metonymische relaties in het WNT. Deze analyses zijn een basis voor het toevoegen van betekenisbetrekkingen in het lexicon.

Data en datamodel
Er is gewerkt aan een verbetering van het datamodel in RDF. Een aantal verbeteringen is mede tot stand gekomen n.a.v. de verdere ontwikkeling van Ontolex voor lexicografische producten. In het kader van dit project en Elexis is meegewerkt aan de verdere ontwikkeling van deze module, die volgend jaar gereleased wordt.

Data
Na het afronden van de correctie van de synoniemendetectie in de betekenisomschrijvingen van het MNW, dat ruim 120.000 relaties tussen een hedendaags en historisch trefwoord heeft opgeleverd, is in 2018 gestart met het nakijken van de ruim 180.000 automatisch gedetecteerde synoniemen in het WNT. Daarvan is reeds ruim een kwart nagekeken. Omdat het datamodel provenance-informatie bevat, kan de dataset opgeleverd worden voordat alle correctie is afgerond. Op basis van het aangepaste datamodel is een nieuwe versie van de dataset gemaakt.

Koppelen met externe datasets
Onder andere om retrieval in het KB-krantencorpus te ondersteunen in de CLARIAH research pilot DB-CCC (zie onderdeel CLARIAH) is een koppeling gemaakt tussen een door de onderzoeker geselecteerde termen uit de binnen dat project gedigitaliseerde bron *Geïllustreerde encyclopaedie der diamantnijverheid* van Felix Leviticus en DiaMaNT. De met DiaMaNT verrijkte termen zijn gebruikt in een onderzoek naar de diamantindustrie in Borneo.

Zoekapplicatie
Er is verder gewerkt aan het userinterface van de zoekapplicatie op de DiaMaNT dataset. De gebruiker zal kunnen zoeken naar historische synoniemen van moderne concepten en daarvan een visualisatie zien door de tijd heen. Deze applicatie maakt gebruik van een sparql endpoint. De applicatie zal een eerste release hebben medio 2019.

II Corpora

II.1 Modern corpusmateriaal: Corpus Hedendaags Nederlands

Corpusdata
Het Corpus Hedendaags Nederlands (CHN) bevat hedendaags taalmateriaal met teksten voornamelijk uit kranten, tijdschriften, journaaluitzendingen en juridisch materiaal. Hoewel er geen release is geweest, is ook dit jaar regulier binnenkomend materiaal verwerkt. Nieuw is dat het krantenmateriaal van Standaard en NRC van 2000 tot heden syntactisch is geannoteerd. Daarbij is gezorgd dat er een koppeling blijft bestaan tussen de met PoS en lemma verrijkte bestanden met de door Alpino geparseerde bestanden.
Wat betreft het toevoegen van extra materiaal is er een aantal stappen genomen. Voor sociale media is er gewerkt aan het automatisch binnenhalen en converteren van een Nederlands (https://forum.fok.nl/) en Vlaams online forum (https://www.9lives.be/). Deze zullen in 2019 in de corpusworkflow in Duct worden opgenomen en systematisch geharvest. Het materiaal zal worden aangevuld met een forum uit Suriname (http://forum.waterkant.net/).
Verder is geregeld dat de data van de Tweede Kamerverslagen systematisch binnenkomen.
Tot slot zijn er verkennende gesprekken gevoerd met de organisatie achter *oefenen.nl* om te kijken of een corpus eenvoudig Nederlands gerealiseerd zou kunnen worden, op basis van o.a. hun data.
In 2019 willen we het corpus substantieel uitbreiden met nieuwe collecties. Daarom is besloten tot werving van een nieuwe dataconversiespecialist.
Tot slot is een aantal verbeteringen gedaan aan de conversies van NRC en Twente Nieuwscorpus t.a.v. de metadata.

Corpusworkflow (DUCT)
Het uitbreiden van het Corpus Hedendaags Nederlands is zoveel mogelijk geautomatiseerd. Ontvangen en opslaan van materiaal, conversies, verrijking en indexering zijn stadia in een corpusworkflow, uitgevoerd binnen een daartoe ontwikkelde tool DUCT (Data Update Creation Tool). Dit is een tool voor het converteren van bestanden in verschillende stappen. In 2018 is het uploaden van alle corpusdata afgerond. In de corpusworkflow is het automatisch verwerken van het materiaal van de Tweede Kamer (NL) geregeld. DUCT is uitgebreid met een nieuw stadium in de taalkundige verrijking, de syntactische annotatie. Voor 2019 is gepland dat het parseren met Alpino binnen DUCT in productie genomen kan worden.

Corpus Front End en BlackLab/BlackLab server
In het kader van CLARIAH is substantieel werk verricht aan het corpus-frontend, en zijn waar nodig aanpassingen gedaan aan het backend. Omdat deze ontwikkelingen in 2018 nog volop bezig waren en de verbeteringen het CHN ten goede komen, is besloten om de update van het CHN, inclusief nieuw front-end met nieuwe zoekmogelijkheden, voor 2019 in te plannen.

II.2 Historisch corpusmateriaal

Corpusdata
Werkzaamheden aan het historisch corpusmateriaal zijn uitgevoerd in het kader van het project Nederlab. Sedert 1 januari 2017 is het INT verantwoordelijk voor de corpusprocessing van Nederlab. Dat betekent dat digitaal corpusmateriaal van derden wordt geconverteerd naar Nederlab-formaat (XML-FoLiA) en voorzien van taalherkenning getokeniseerd aan de projectpartners wordt opgeleverd.

De teksten worden voorzien van correcte metadata, inclusief thesaurering van de auteurs. In het afgelopen jaar zijn de volgende collecties aan het Nederlabcorpus toegevoegd:

- (Huygens-ING) Dagboeken van P.J.M. Aalberse
- (Huygens-ING) Van Gogh Letters
- (Huygens-ING) Correspondenties 1900.
- CLVN (Corpus Laatmiddel- en VroegnieuwNederlands) (15e en 16e eeuw)
- INT: CD-Rom Middelnederlands, uitgebreid met de teksten van de uitgever Brill waarvoor de rechten voor distributie verkregen zijn
- TU Eindhoven Corpus: een nieuwe versie die een integratie met verbeteringen is van eerdere versies van het Corpus. Het corpus zal bij de taalmaterialen ter beschikking worden gesteld
- MI Corpus Van Reenen Mulder (14e eeuw)
- INT: Corpus Gysseling
- INT: Corpus Oudnederlands, dat voor het eerst in de vorm gebracht is van een verrijkt tekstcorpus i.p.v. een database met verrijkte zinnen uit diverse documenten.

Uit het overleg met de toolstrack over de taalkundige verrijking ten slotte is de noodzaak gebleken om de reeds opgeleverde metadata van de DBNL qua datering verder uit te breiden zodat selectie van teksten van een bepaalde taalperiode gemaakt kan worden. De hele dataset is van de DBNL is in dit opzicht gereviseerd en verrijkt.

Taalkundige verrijking
In het kader van Nederlab is samengewerkt met de toolstrack van het project om de taalkundige verrijking conform de huidige state of the art op een zo goed mogelijk peil te krijgen. Omdat de toolstrack ervoor gezorgd heeft dat Frog hertrainbaar was, is er besloten om voor de collecties waarvoor trainingsmateriaal beschikbaar was, Frog opnieuw te trainen. *Corpus Gysseling, Corpus van Reenen Mulder, Brieven als Buit* en de *Gentse spelen* hebben een conversie gehad naar de CGN-tagging, waarbij de tagset van het CGN uitgebreid is met specifieke features voor de twee eerste corpora.
Er zijn verder evaluatiesets gemaakt van ca. 10.000 woorden voor de 15e -18e eeuw. Samen met de twee evaluatiesets van 18e en 19e eeuw vormen deze sets de basis voor verder werk aan de optimalisatie van de taalkundige verrijking van historisch Nederlands die vanaf 2019 uitgevoerd zal worden in het kader van CLARIAH plus. De ervaringen met de tekortkomingen van de CGN-tagset voor historisch Nederlands zal ook in dat verband gebruikt worden om binnen de community van historisch taalkundigen te komen tot een uniforme en breed gedragen tagset voor corpusannotatie van historisch Nederlands.

Corpusapplicatie
Net zoals voor het corpus hedendaags Nederlands, wordt voor de online historische corpora, momenteel het *corpus Gysseling* en de *Brieven als buit*, gebruikgemaakt van Blacklab server. De update van de historische corpora, die voorzien was in 2018, is vanwege uitvoerige werkzaamheden aan het userinterface uitgesteld en voorzien voor 2019.

III Externe projecten

Nederlab
Datatrack lexica
Het doel was om diachroon lexicaal materiaal inzetbaar te maken voor zoeken in het diachroon corpusmateriaal van Nederlab. De taken van dit jaar concentreerden zich op het produceren van evaluatiemateriaal voor taalkundige verrijking en het omzetten van de taalkundige verrijking van Corpus Gysseling, Corpus Van Reenen Mulder en het Brieven-als-Buitcorpus naar een uniform formaat dat compatibel is met de CGN-tagging. Deze werkzaamheden hebben geleid tot een vervolgproject in de context van CLARIAH plus over verrijking van historisch taalmateriaal (zie II.2).
Datatrack corpora

Het INT heeft corpustrack overgenomen van het Meertens Instituut. De taak was aan het eind van het project minimaal 20 deelcorpora in Nederlab verwerkt te hebben. Bij aanvang van de werkzaamheden waren dat er 7. Er zijn uiteindelijk 25 collecties opgeleverd. Voor een verdere beschrijving, zie II.2. Management
Katrien Depuydt heeft de datatracks lexica en corpora geleid, was lid van de corpuscommissie die verantwoordelijk is voor de selectie van corpusmateriaal voor Nederlab en maakte deel uit van de stuurgroep.

CLARIAH
Het CLARIAH-project heeft de ontwikkeling van de lexiconcomponent DiaMaNT (zie hierboven) ondersteund, als onderdeel van WP2, waarin naast een technische infrastructuur ook gewerkt wordt aan een data-infrastructuur van personen, plaatsen en concepten. DiaMaNT valt onder het laatste. In de context van kleine deelprojecten *Serpens* en *Diamonds in Borneo* (gestart in 2018) is het lexicon toegepast en verder uitgebreid. (Zie verder I.2 en de projectinformatie van CLARIAH.)
In de context van het CLARIAH Namesproject is gewerkt aan een voornamen- en familienamenlexicon dat net zoals GiGaNT ook via de lexicon service ter beschikking zal worden gesteld. Het INT brengt met name naamkundige expertise in voor zowel het bouwen van gouden standaardmateriaal als het evalueren van automatisch verwerkt namenmateriaal. Het lexicon komt in 2019 in een lexiconservice ter beschikking.

IV Advies en support

Katrien Depuydt heeft een adviserende rol gehad in de ontwikkeling van de data voor de pilot van het Mental lexiconproject.
De CoBaLT-tool, voor de Taalbank ontwikkeld voor handmatige correctie van corpusmateriaal wordt gebruikt door Rita van der Poel voor het taggen van Oudfries materiaal in het kader van haar promotieonderzoek aan de Universiteit Leiden (https://www.universiteitleiden.nl/en/staffmembers/rita-van-de-poel#tab-1) en door Karina van Dalen-Oskam voor het handmatig corrigeren van door het INT automatisch uitgevoerde tagging en lemmatisering van boekreviews in het kader van het project *the Riddle of Literary Quality* (http://literaryquality.huygens.knaw.nl/).

V Lezingen, presentaties, media
.
Katrien Depuydt heeft een presentatie gehouden op de CLIN in Nijmegen op 26 januari met als titel: *Nederlab progress and challenges in linguistic enrichment of historical Dutch texts*. Katrien Depuydt, Maarten van Gompel, Jesse de Does, Hennie Brugman, Gosse Bouma.

Katrien Depuydt en Jesse de Does hebben meegewerkt aan de presentatie: Marieke van Erp, Jesse de Does, Thomas van Goethem and Katrien Depuydt (2018) *Good lynx, bad lynx: document enrichment for historical ecologist.* Gepresenteerd door Marieke van Erp op de CLIN in Nijmegen op 26 januari.

Katrien Depuydt heeft het CLARIAH Namesproject gepresenteerd op de CLARIAH Toogdag van 9 maart 2018 in Den Haag.

Katrien Depuydt heeft een presentatie gehouden op LREC in Miyazaki op 8 mei op de Globalex 2018 pre-conference workshop “Lexicography and Wordnets”, met als titel: *The Diachronic Semantic Lexicon of Dutch as Linked Open Data.* Katrien Depuydt, Jesse de Does.

Katrien Depuydt heeft een presentatie gegeven op de workshop “*Using Nederlab for Humanities Research*” op 6 juni op de Digital Humanities Benelux conferentie in Amsterdam met als titel: *CLVN Corpus Laatmiddel-en Vroegnieuwnederlands.* Katrien Depuydt, Margit Rem, Nicoline van der Sijs

Katrien Depuydt en Tanneke Schoonheim hebben meegewerkt aan de poster die op 8 juni gepresenteerd werd op de Digital Humanities Benelux conferentie in Amsterdam, met als titel: *Names, a Clariah research project*, DH-Benelux conference, Utrecht. Gerrit Bloothooft, David Onland, Richard Oosterlaken, Martin Reynaert, Katrien Depuydt, and Tanneke Schoonheim.

Katrien Depuydt heeft op 3 september een lezing gehouden op de CLARIAH provenance workshop met als titel: *INT - corpus creation/preparation and provenance.*

Katrien Depuydt heeft op 28 september 2018 in het kader van haar promotieonderzoek een minipresentatie gehouden in Leuven met als titel: *From Dictionaries to a Lexicographical Infrastructure for Historical Dutch.*

Katrien Depuydt heeft op 24 oktober een interview gehad over het project Nederlab in het programma NPO Focus op radio 1.

Katrien Depuydt en Jesse de Does hebben bijgedragen aan de presentatie door Karin Hofmeester gegeven op de CLARIAH Toogdag van vrijdag 19 oktober met als titel: *Diamonds in Borneo: Commodities as Concepts in Context.* Karin Hofmeester, Ashkan Ashkpour, Martin Reynaert, Katrien Depuydt, Jesse de Does, Marieke van Erp.

Katrien Depuydt heeft een presentatie gehouden op het slotevent van Nederlab op 26 oktober met als titel: *Niks, bietekwiet!*

VI Workshops

Katrien Depuydt en Jesse de Does hebben bijgedragen aan de *Serpens* workshop op 25 januari in Nijmegen.
Katrien Depuydt en Jesse de Does hebben deelgenomen aan de workshop ter ontwikkeling van TEI Lex-0 op 2 en 3 mei in Berlijn.
Katrien Depuydt en Jesse de Does hebben een workshop over OCR gegegeven bij de KU Leuven voor CLARIN Vlaanderen op 22 november.
Katrien Depuydt heeft deelgenomen aan de Mentallex workshop op 31 oktober in Amsterdam.
Katrien Depuydt en Jesse de Does hebben deelgenomen aan de workshop van Ontolex over de ontwikkeling van Ontolex over lexicografische data op het INT in Leiden op 5 november 2018.

VII Publicaties

Marieke van Erp, Jesse de Does, Katrien Depuydt , Rob Lenders and Thomas van Goethem (2018) . Slicing and Dicing a Newspaper Corpus for Historical Ecology Research. *Proceedings of European Knowledge Acquisition Workshop – EKAW* (2018).
Katrien Depuydt and Jesse de Does (2018), "The Diachronic Semantic Lexicon of Dutch as Linked Open Data." In: I. Kernerman and S. Krek, *Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets".* [Miyazaki], 2018, pp. 23-28.

VIII Overig

Katrien Depuydt en Jesse de Does nemen deel aan de verdere ontwikkeling van de RDF Ontolex standaard voor lexicografische bronnen en aan de verdere ontwikkeling van TEI lex-0, een XML-standaard voor woordenboeken.

Katrien Depuydt is samen met Vincent Vanderghinste betrokken bij de organisatie van *Datech 2019* (http://datech.digitisation.eu/) van 8-10 mei in Brussel.

Bijlage 5: Rapport Externe Communicatie

In dit rapport worden de in 2018 ingezette communicatiemiddelen en de belangrijkste cijfers gepresenteerd.

1. Website

De website is het gezicht van het instituut naar buiten en daarmee het belangrijkste communicatiemiddel. Bijna alle communicatie-uitingen zijn erop gericht om bezoekers naar de website te trekken. Voor onderstaande analyse is gebruikgemaakt van Google Analytics.

Cijfers

Algemeen	**2017**	**2018**
Totaal aantal paginaweergaven	480.129	492.446
Bezoeken uit Nederland	176.388	184.590
Bezoeken uit België	52.374	42.269
Gemiddelde tijd op pagina	00.01.45	00.01.50

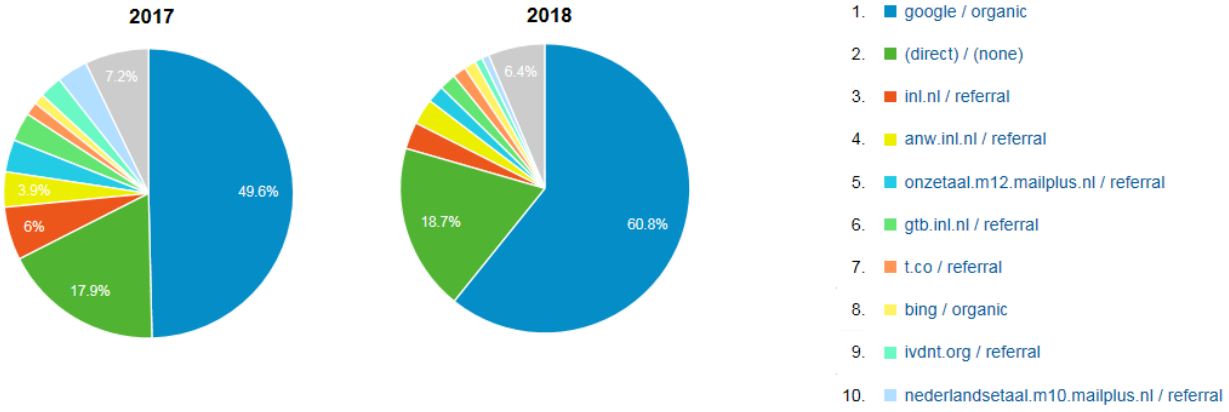
Paginaweergaven rubrieken	**2017**	**2018**
Woordbaak*	108.542	110.296
Neologismen (neo v/d week + neo v/d week archief + neologismen algemeen)**	85.504	77.323
Index/home	47.327	38.997
Historische woordenboeken (algemeen + ONW, VMNW, MNW, WNT, WFT)	27.664	36.355
Nieuws	16.790	13.342
Gelegenheidswoordenboekjes	4.157	9.966
Taalmaterialen	6.755	9.243
Terug in de taal	13.215	7.909
Algemeen Nederlands Woordenboek	2.426	7.524
Jaaroverzicht in neologismen	4.510	3.181
Etymologie	3.247	3.133
Spelling	1.795	2.087
CLARIN-ERIC	860	1.364
Contact/medewerkers	1.411	1.299
Terminologie (Term van dag (374) + evenementen (362) + SNT (357))	-	1.093

** Woordbaak is geen vaste wekelijkse rubriek meer; in 2018 zijn er 3 Woordbaakartikelen verschenen. Het blijft wel de meest bekeken webrubriek, en dan vooral het artikel 'Zijn pappa en mamma wel correct?".*

***Populairste rubriek: Neologisme van de week*

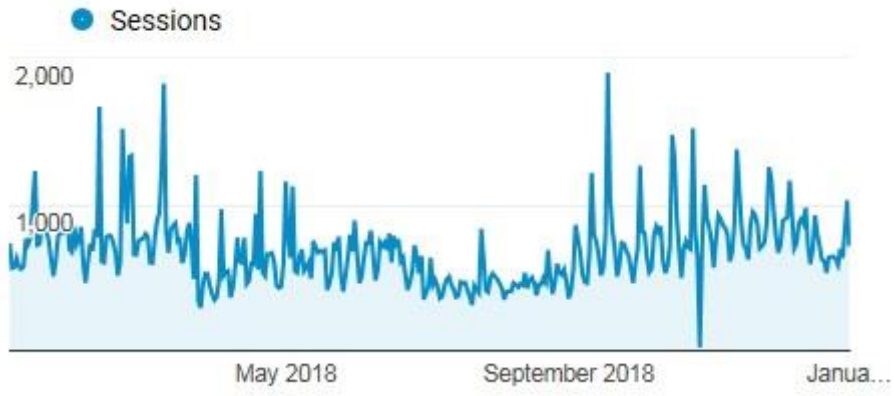
Zoekterm	hits
lassy	163
sonar	79
cgn	72
nieuwe woorden	71
basilex	60
ombi	41
autonomata	39
dialect	39
neologisme	35
corea	33
basiscript	30
cornetto	29
anw	28
neologismen	28
diamant	26

Bronnen/verwijzingen



- Er is een duidelijke toename (ongeveer 10%) in verwijzingen vanaf Google. Er zijn iets meer bezoekers direct naar de website gegaan. Ook staan de verwijzingen vanuit de nieuwsbrief van *Onze Taal* (Taalpost) en de Taalunie in de top tien.

Verloop bezoekersaantallen



In vergelijking met voorgaande jaren is er minder duidelijk een piek in bezoekersaantallen te zien aan het einde van het jaar, omdat er geen verkiezing georganiseerd werd. De opvallende pieken die er zijn hebben te maken met veelgelezen artikelen in nieuwsbrieven (zoals in februari *plogging* in Taalpost en in november *lok-bh* in Taalpost). In het najaar zijn de bezoekersaantallen het grootst, in de zomerperiode wordt de website het minst goed bezocht.

Conclusies & aanbevelingen
De website heeft in 2018 iets meer bezoekers getrokken dan het voorgaande jaar. Dat is opvallend omdat er in 2018 geen '*Weg met dat woord!*'-verkiezing georganiseerd werd: een eindejaarsactie die de afgelopen vijf jaar steeds zorgde voor een enorme piek in bezoekersaantallen op de website. Iets dat in 2018 wel van invloed is geweest op de bezoekersaantallen is het samenvoegen van de website van de TST-Centrale met die van het INT. De oude TST-website is opgeheven en onder de noemer 'Taalmaterialen' ondergebracht op de INT-website. Aan de top 15 zoektermen is ook te zien dat er het vaakst gezocht wordt op taalmaterialen zoals Lassy, Sonar en CGN.
Daarnaast kunnen we uitgaan van een groeiende naamsbekendheid van het INT en meer activiteit binnen het instituut, dat zijn weerslag heeft op de website.

Er is ten opzichte van 2017 een stijging in het aantal bezoeken vanuit Nederland, van 69,8% naar 73,2%, en een daling in het aantal bezoeken afkomstig uit België, van 20,7% naar 16,8%. De gemiddelde tijd die een bezoeker doorbrengt op een pagina is iets gestegen.

De grootste aandachttrekkers zijn de populairwetenschappelijke webrubrieken *Woordbaak* (etymologie) en *Neologisme van de week* (nieuwe woorden). *Nieuwe woorden*, *neologisme* en *neologismen* zijn populaire zoektermen op de website die allemaal in de top 15 staan. Ook informatie over de historische woordenboeken wordt opvallend vaak bezocht: in april 2018 is er een nieuwe versie van de online historische woordenboeken gelanceerd (gtb.ivdnt.org). De rubriek *Terug in de taal* (historische woorden) wordt goed bekeken, alhoewel minder vaak dan in het voorgaande jaar. *Woordbaak* is geen wekelijkse rubriek meer, maar gezien de populariteit is het nuttig om de rubriek te behouden en met enige regelmaat een artikel te blijven plaatsen.
Tijdens de Week van het Nederlands is eenmalig de nieuwe rubriek *Term van de dag* in het leven geroepen, die redelijk wat aandacht trok. Het is goed om te bekijken of we deze rubriek op regelmatige basis kunnen voortzetten, bijvoorbeeld als onderdeel van de terminologienieuwsbrief, om op die manier ook wat meer aandacht voor het onderwerp terminologie te genereren.
In 2018 zijn er (na de laatste in 2015) drie nieuwe gelegenheidswoordenboekjes verschenen, wat de meer dan verdubbelde bezoekersaantallen ten opzichte van het voorgaande jaar verklaart. Het *Internetwoordenboekje* is veruit het meest bekeken (hier is een apart persbericht over verstuurd), gevolgd door het *Lubberiaans Lexicon* en het *Coachwoordenboekje*.

Wat ook opvalt is dat de informatie over het *Algemeen Nederlands Woordenboek* veel meer is bezocht dan in het voorgaande jaar, en *ANW* komt ook voor in de top 15 zoektermen. De aanleiding hiervan is niet helemaal duidelijk. Het zou bijvoorbeeld kunnen dat er aandacht aan het woordenboek besteed is in een college(reeks).

Meer dan de helft van de bezoekers vindt de INT-website via Google en 18,7% gaat direct naar ivdnt.org. Ook staan de nieuwsbrieven van *Onze Taal* (Taalpost) en de Taalunie weer in de top tien verwijzingen.

Website 'Weg met dat woord!'
In 2017 werd de verkiezing *Weg met dat woord!* voor het laatst georganiseerd. De website bestaat nog wel en is in 2018 gebruikt ter promotie van het boek *Kids, koffietjes & comfortzone* (2018), over de resultaten van 5 jaar *Weg met dat woord!*. Daarnaast geeft de website een overzicht van de verkiezingsuitslagen van 2013 t/m 2017. In 2018 is wegmetdatwoord.org 6.302 keer bekeken.

2. Sociale Media

Sinds 2010 heeft het instituut een Twitteraccount. Via Twitter worden links naar de website en taalweetjes/taalnieuws gedeeld, en contacten met volgers en partnerorganisaties onderhouden.

Cijfers

	December 2017	**December 2018**
Aantal volgers	3.913	4.183

Facebook
Het instituut heeft geen algemene Facebookpagina, maar er is wel een specifieke pagina voor de verkiezing *Weg met dat woord!*: https://www.facebook.com/wegmetdatwoord/. De pagina heeft op dit moment 1.213 likes. In 2018 is de pagina gebruikt ter promotie van het boek *Kids, koffietjes & comfortzone*.

Conclusies & aanbevelingen
Twitter blijft gestaag groeien met dit jaar een toename van 270 volgers. De Facebookpagina voor *Weg met dat woord!* is niet meer heel actief en er moet bekeken worden wat we met deze pagina gaan doen. De pagina kan bijvoorbeeld omgevormd worden tot een algemene Facebookpagina voor het INT, of worden gebruikt voor een nieuwe eindejaarsactie.

3. Nieuwsbrieven, mailings & persberichten

De algemene nieuwsbrief is om geïnteresseerden te blijven binden aan en te informeren over het instituut. De nieuwsbrief terminologie is een aparte thematische nieuwsbrief, waarvoor abonnees zich los van de algemene nieuwsbrief kunnen inschrijven. In 2018 zijn er 5 algemene nieuwsbrieven verschenen en 4 nieuwsbrieven over terminologie. Daarnaast zijn er nog diverse aparte mailings verzonden. Voor het versturen van nieuwsbrieven, mailings (ook intern) en persberichten maken we gebruik van de software MailChimp.

Cijfers

Nieuwsbrief algemeen	2017	**2018**
Aantal inschrijvingen	3.872	3.894
Gemiddeld aantal opens	48%	46%
Gemiddeld aantal kliks (CTO)*	32%	29%
Nieuwsbrief terminologie		
Aantal inschrijvingen	2.671	2.660
Gemiddeld aantal opens	33%	31%

Gemiddeld aantal kliks (CTO)	17%	16%

** CTO = click-to-open ratio, een percentage dat wordt bepaald door het aantal ontvangers dat de mail opent en doorklikt (clicks e-mail gedeeld door geopende e-mails)*

- De augustuseditie van de algemene nieuwsbrief is het vaakst geopend (48,3%).
- Best gelezen artikelen: Woordbaakartikel over *rinsig* (320 kliks), gelegenheidswoordenboekje *Lubberiaans Lexicon* (266 kliks), Terug in de taal over *oktember* (198 kliks) en Neologisme van de week *tepelgate* (198 kliks).
- Meest geopende nieuwsbrief terminologie is de septemberedítie (36,5% opens, 15,3% kliks)
- Links waar het vaakst op geklikt is: boek van Onze Taal *Grammatica, 150 begrippen verklaard en toegelicht* (67 kliks), financiële begrippen uit 2017 (52 kliks) en *Algemeen letterkundig lexicon* in DBNL (43 kliks).

Mailings & persberichten	**opens**	**kliks (CTO)**
Verhuisbericht	46,2%	8,7%
Persbericht: Nieuw jasje voor historische woordenboeken	50%	34,4%
Persbericht: Internetwoordenboekje	31,9%	9,6%
Aankondiging TiNT-dag	36,4%	13,1%
Taalradar	47,2%	29,7%
Persbericht: Kids, koffietjes & comfortzone	35,3%	1,5%
Aankondiging Kids, koffietjes & comfortzone (WMDW-deelnemers)	54,3%	14,1%
Nieuwjaarswens relaties	43,8%	2,1%

Conclusies & aanbevelingen
Het abonneebestand van de nieuwsbrieven is in 2018 minder hard gegroeid dan andere jaren. De inschrijvingen voor de terminologienieuwsbrief zijn zelfs licht gedaald. De cijfers voor het aantal opens en kliks zijn voor de algemene nieuwsbrief relatief hoog in vergelijking met nieuwsbrieven van andere organisaties (gemiddeld 38% opens en 16,3% CTO volgens de Nationale E-mail Benchmark van 2018). Dat laat zien dat lezers geïnteresseerd zijn, zich betrokken voelen en optimaal kennis nemen van de inhoud. De nieuwsbrief Terminologie zit iets onder het gemiddelde percentage opens van de Benchmark. Daarnaast wordt er in de Terminologienieuwsbrief minder vaak op links geklikt dan in de algemene nieuwsbrief. Dat verschil zit waarschijnlijk in de opzet: in de algemene nieuwsbrief staan vaak korte berichten gevolgd door een ‘lees meer’ die naar de website verwijst. In de Terminologienieuwsbrief staan meestal volledige berichten, met links naar extra informatie. Het is nuttig om te onderzoeken hoe deze nieuwsbrief verbeterd kan worden, bijvoorbeeld door middel van een korte enquête onder de lezers.

De cijfers van de mailings en persberichten zijn lastig met elkaar te vergelijken omdat ze naar verschillende doelgroepen binnen het adressenbestand worden verstuurd. Aan de percentages is wel te zien dat de losse berichten beter gelezen worden dan nieuwsbrieven waarin meerdere berichten staan.

4. Huisstijl, sponsoring en populairwetenschappelijke activiteiten

In 2018 zijn er twee nieuwe folders uitgebracht, vormgegeven in de INT-huisstijl: een specifieke Engelstalige folder over het INT als Clarin centre en een algemene Nederlandstalige folder over de werkzaamheden van het INT. Daarnaast zijn er nog diverse flyers over verschillende projecten (o.a. Taalradar, stages, woordcombinaties) in huis gemaakt. Ook is er een serie populairwetenschappelijke posters ontworpen, zijn er linnen tasjes gemaakt voor het DRONGO talenfestival en zijn er koffiebekers gemaakt in de huisstijl. Met name de geelgekleurde linnen tasjes met daarop het woord

tas (inclusief betekenissen) afgedrukt waren een succes; bezoekers van DRONGO kwamen bij de stand vragen om de gele tas.

Net als voorgaande jaren trad het instituut op als sponsor voor onder andere CLIN en de TABU-dag. In februari 2018 verhuisde het INT naar een monumentaal pand aan het Rapenburg, en in september werd een deel van het pand opengesteld voor publiek tijdens Open Monumentendag. Daar kwamen bijna 3.000 bezoekers op af.
Tijdens de Week van het Nederlands in september werd de eenmalige rubriek *Term van de dag* gelanceerd, om aandacht te vragen voor terminologie. Eind november stond het instituut op het DRONGO talenfestival met een lab over Taalradar en dialectloket.be.
In november verscheen bij uitgeverij AUP het populairwetenschappelijke boek *Kids, koffietjes & comfortzone. Waarom taal soms irritant is.* Het boek werd geschreven door Laura van Eerten en Vivien Waszink en is een afsluiting van vijf jaar de verkiezing *Weg met dat woord!*.

Eind december organiseerde het INT samen met het Vlaams-Nederlands Huis deBuren voor de tweede keer 'Het jaar in taal': een avond over woorden die het publieke debat in 2018 domineerden.

Tot slot beantwoorden we door het jaar heen veel vragen over het Nederlands aan studenten, journalisten en programmamakers. In 2018 werd het instituut regelmatig door radio en tv benaderd. Een uitgebreid overzicht is te vinden in het jaarverslag 2018.

5. Resultaten

De gecombineerde communicatieactiviteiten in 2018 hebben geleid tot meer bezoekers op de website, nieuwe nieuwsbriefabonnees, meer volgers op Twitter, publicaties in vaktijdschriften en kranten, artikelen op goedbezochte nieuwswebsites en aandacht op radio en tv. Aandachtspunten voor 2019 zijn de populairwetenschappelijke webrubrieken, verbetering van de nieuwsbrieven en een eventuele nieuwe eindejaarsactie.