MEERJARENBELEIDSPLAN
2018 - 2022

/instituut voor
de Nederlandse
taal/

Inhoudstafel

E.2. INT Impact centrum en digitization.eu
E.3. ELRC

F. Digitale taalmaterialen (geïntegreerde webwinkel)
INT, TST en CLARIN-materialen

G. CLARIN
G.1. Het INT als CLARIN-centrum
G.2. DARIAH (Digital Research Infrastructure for the Arts and Humanities)

H. Samenwerking en netwerken

I. Onderzoek: Europese aanvragen en competitieve projecten

I.1. ELEXIS (Horizon 2020)
I.2. TermNeXT (Horizon 2020)
I.3. Inzicht in het mentale lexicon (KIEM NWO)
I.4. ENETCOLLECT (COST-actie)

J. Doelgroepenbeleid (inclusief onderwijs)
J.1. Onderzoekers, wetenschappers
J.2. Studenten en docenten
J.3. Algemeen publiek

K. Wetenschapscommunicatie
K.1. Twitter; Facebook
K.2. Woordbaak; Terug in de Taal; Neologisme van de Week
K.3. Populairwetenschappelijke uitgaven

Meerjarenbegroting

Inleiding

Algemene doelstellingen INT: instituut met een nieuwe missie

Het jaar 2016 was een scharnierjaar voor het instituut. Door de beslissingen van het Comité van Ministers om vanaf medio 2015 op jaarbasis 400.000 euro te bezuinigen op het instituut, liep de structurele basisfinanciering terug van 2.345.000 euro naar 1.945.000 euro. De reorganisatie werd in goede banen geleid door overleg met de Taalunie en de Nederlandse en Vlaamse overheden en leidde tot een ingrijpende statutenwijziging van de Stichting Instituut voor Nederlandse Lexicologie (INL) per 1 januari 2016.
In het jaar 2016 werd de reorganisatie afgerond en het INL werd hervormd tot het Instituut voor de Nederlandse Taal (INT), een breed toegankelijk wetenschappelijk instituut op het gebied van het Nederlands.

Het INT wil een centrale positie innemen in het hele Nederlandse taalgebied (o.a. Vlaanderen, Suriname en de voormalige Antillen) op het vlak van het wetenschappelijk verantwoord ontwikkelen, bewaren en duurzaam beschikbaar stellen van taalmateriaal.
Het INT streeft ernaar om het best gesorteerde en daarmee zeer goed, toegankelijk wetenschappelijk instituut te zijn op het gebied van de Nederlandse taal en de woordenschat. Het instituut ontwikkelt en levert data voor woordenboeken, (computationele) lexica, corpora en tools. De woordenboeken zijn online te raadplegen. Software en computerlinguïstische tools zijn open source beschikbaar.
Het instituut speelt in op de nieuwe ontwikkelingen in de Geesteswetenschappen, met name op het terrein van de Digital Humanities. Om deze rol te kunnen vervullen beheert en onderhoudt het INT een digitale infrastructuur voor het Nederlands, met aandacht voor taalvariatie (terminologie, dialecten etc.). Zowel academische als niet-academische partijen kunnen gebruikmaken van deze infrastructuur.

Het INT zal zich met name toeleggen op het al dan niet in samenwerking met derden ontwikkelen en ontsluiten van concrete producten: corpora, woordenboeken van standaardtaal en dialecten, terminologische databanken, vertaalwoordenboeken, grammatica's, taaladvieshandboeken, cursusmateriaal etc.

Op basis van hun verdrag en statuten streven de Taalunie en het INT gemeenschappelijke doelstellingen na. Hiervoor kennen ze een bijzondere samenwerkingsrelatie die is vastgelegd in een samenwerkingsovereenkomst. Het toezicht op het bestuur van het instituut wordt verzekerd door een uit drie personen bestaande Raad van Toezicht (RvT) op een wijze zoals vastgesteld in de statuten. Er is tevens een Raad van Advies (RvA), die uit 10 leden bestaat: zeven hoogleraren taalkunde (waarvan de voorzitter de KNAW vertegenwoordigt, en de vicevoorzitter de KANTL), twee experts uit het culturele domein en één expert voor Suriname en de voormalige Antillen. De leden van de RvT en de RvA worden benoemd door het Comité van Ministers. Op deze wijze functioneert het Instituut voor de Nederlandse Taal als een zelfstandig instituut, maar toch in goed overleg met de Taalunie, die de subsidies beheert en beschikbaar maakt. De RvT die in 2016 de reorganisatie heeft afgerond, heeft afscheid genomen en op 5 december 2016 werd een nieuwe RvT benoemd. Deze bestaat uit de volgende leden: de heer P. Rüpp (voorzitter), de heer J. Cerfontaine en mevrouw G. van der Vliet. Tevens werd in 2016 een vacature uitgeschreven voor de functie van wetenschappelijk directeur/bestuurder. De vacature resulteerde in de aanstelling van Prof. dr. F. Steurs als nieuwe directeur. Zij startte haar werkzaamheden bij het instituut op 1 september 2016.

Het Instituut voor de Nederlandse Taal is dé plek voor iedereen die iets wil weten over het Nederlands door de eeuwen heen. Het is een breed toegankelijk wetenschappelijk instituut dat alle aspecten van de Nederlandse taal bestudeert, waaronder de woordenschat, grammatica en taalvariatie. Het instituut verzamelt de nieuwste Nederlandse woorden, actualiseert belangrijke naslagwerken zoals de *Algemene Nederlandse Spraakkunst* en maakt vaktaal toegankelijk via terminologielijsten.
Daarnaast neemt het instituut een centrale positie in het hele Nederlandse taalgebied in (o.a. Vlaanderen, Suriname en de voormalige Antillen) op het vlak van het wetenschappelijk verantwoord ontwikkelen, bewaren en duurzaam beschikbaar stellen van corpora, lexica, woordenboeken en grammatica's, ook wel taalmateriaal genoemd. Daarmee levert het Instituut voor de Nederlandse Taal noodzakelijke bouwstenen voor alle taaltoepassingen die bedrijven en publieke organisaties willen uitbouwen.

Deze rol willen we in de komende jaren veel sterker maken; we zetten dan ook in op het ontsluiten en duurzaam beschikbaar stellen van alle mogelijk taalmaterialen:

- de Nederlandse woordenschat, zowel historisch als hedendaags, zowel de standaardtaal als de dialecten, zowel de algemene taal als de vaktaal;
- nieuwe technologieën en technieken om het internet toegankelijk te maken voor taalkundig onderzoek en om, continu aangevulde, uitgebreide corpora van hedendaags Nederlands te blijven onderhouden;
- bijdragen aan het toegankelijker maken van historisch tekstmateriaal (komend van binnen en buiten het INT), waarbij de enorme variatie aan spelling niet langer een hinderende factor is bij het zoeken, en waarbij handvatten worden geboden voor het detecteren van en omgaan met variatie in woordgebruik;
- het gebruikmaken van, en het bijdragen aan nieuwe computerlinguïstische of taaltechnologische technieken voor informatie-extractie uit taalmateriaal;
- het formeel structureren van taalkundige informatie, zodat deze gebruikt kan worden voor computerlinguïstische toepassingen;
- de verdere uitbouw van de spellinginformatie;
- het realiseren van voorzieningen waarmee derden interactief kunnen bijdragen aan de beschrijving van de Nederlandse taal en het optimaliseren van de centrale digitale data-infrastructuur om dit alles mogelijk te maken.
- een aanspreekpunt worden voor alle taaldocenten en een infrastructuur uitbouwen van taalmaterialen die nuttig en nodig zijn voor de ondersteuning van het doceren van Nederlands aan alle geledingen van taalleerders.

Centrale data-infrastructuur van de Nederlandse taal

Het INT wil in de komende jaren zeer zichtbaar worden en als hét centrum voor de data-infrastructuur voor het Nederlands algemeen erkend worden. Om dat doel te bereiken moet worden ingezet op volgende aspecten: duurzaam beheer, permanent onderhoud, kennismanagement, beschikbaarstelling en dienstverlening. Daartoe werden in 2017 belangrijke stappen gezet in beheer van de data. De outsourcing die in 2016 was opgestart door de voormalige interim-directeur bleek niet adequaat voor de uitdagingen en de centrale missie van het INT: het beschikbaar stellen, upgraden, onderhouden van zeer veel digitale taalmaterialen en tevens kunnen inspelen op de wensen en noden van individuele gebruikers. Een commerciële samenwerking met een extern bedrijf maakt dat soort werk niet makkelijk en bovendien veel te duur. In de eerste helft van 2017 werd daarom besloten om het contract met de commerciële partij niet verder te zetten en de eigen servers opnieuw te benutten. Deze staan op het ISSC van de U Leiden. Met het ISSC zijn daartoe de nodige afspraken gemaakt. Er is een extra rekenserver aangekocht, en een extra systeembeheerder is aangeworven. Dit maakt het INT sterk en flexibel voor databeheer en -ontsluiting in de toekomst. Verder werd in 2017 een samenwerking afgesloten met DANS voor het archiveren van onze data. Op deze wijze wordt een veilig beheer gegarandeerd.

Het INT ziet een duidelijke overlapping tussen de eigen activiteiten - de centrale data-infrastructuur - en recente ontwikkelingen binnen de eHumanities. Vanuit zijn eigen expertise draagt het INT bij aan de digitale toekomst van de geesteswetenschappen in Nederland en Vlaanderen. Enerzijds worden kennis en producten geleverd ten bate van andere wetenschappelijke organisaties, anderzijds verhoogt de samenwerking binnen de eHumanities de kwaliteit van de centrale data-infrastructuur van het Nederlands.
Er zal dan ook in de komende jaren nauw worden samengewerkt met centra voor digital humanities aan de verschillende universiteiten en met netwerken zoals het KNAW digital humanities cluster (Nederland), Digital Humanities Benelux en de WOG Digital Humanities (Vlaanderen).

Beoogde resultaten 2017-2022

Specifieke projecten

A. Woordenboeken

Beschrijving van de woordenschat blijft één van de kerntaken van het instituut, zowel historisch als hedendaags, zowel de standaardtaal als de dialecten, zowel de algemene taal als de vaktaal, zowel monolinguaal als bilinguaal. Daarnaast zet het INT in de komende beleidsperiode ook in op een meer systematische inventarisatie en beschrijving van woordcombinaties.

Het maken van woordenboeken is de afgelopen decennia radicaal veranderd door de opkomst van allerlei technologieën die het productieproces ondersteunen. Daarnaast wordt er steeds meer gebruikgemaakt van NLP-technieken (Natural Language Processing) om lexicografische gegevens (collocaties, definities, voorbeeldzinnen en vertalingen) uit corpora en andere digitale bronnen te halen. Dit geldt ook voor het lexicografische werk van het INT.
Zo zal het INT, o.a. in samenwerking met Europese partners binnen het ELEXIS-project, meer onderzoek gaan doen naar automatische informatie-extractie. In eerste instantie moet hierbij gedacht worden aan het automatisch extraheren van voorbeeldzinnen, collocaties en multimedia. Later zullen soortgelijke technieken ook voor andere informatiecategorieën worden ontwikkeld. Het uiteindelijke doel is om het lexicografische werk en de intellectuele input zoveel mogelijk naar de post-editingfase te verplaatsen in plaats van het handmatig analyseren van inputdata. Dit betekent dat bestaande systemen voor zowel de analyse van het corpusmateriaal als voor het (post)editen en online presenteren van het materiaal moeten worden geoptimaliseerd voor deze taken (o.a. door ontwikkelen van verbeterde DWS).

Een andere belangrijke ontwikkeling binnen het semantisch web is (Linguistic) Linked Open Data ((L)LOD). Het Semantische Web gaat over het maken van koppelingen tussen datasets, niet alleen voor mensen, maar ook voor machines, en Linked Data biedt de methoden om die

links te maken. In de mate van het mogelijke zal het INT zijn lexica en woordenboeken ter beschikking stellen als LLOD.

Tot slot zal het INT, in het kader van de ENETCollect COST actie, de mogelijkheden om crowdsourcing te gebruiken ten behoeve van de lexicografie onderzoeken.

A.1. Synchroon: ANW

Projectleider: Rob Tempelaars

In de komende jaren zal worden verder gewerkt aan het Algemeen Nederlands Woordenboek (ANW). Dit is een corpusgebaseerd, digitaal woordenboek van het van het eigentijdse Nederlands in Nederland en Vlaanderen, in Suriname en in het Caraïbisch gebied. De taalperiode die het ANW bestrijkt, loopt van 1970 tot heden en valt min of meer samen met de naoorlogse generaties volwassen taalgebruikers.
Het INT beschouwt het als een kerntaak om als toegepast wetenschappelijk instituut een dergelijk woordenboek van het moderne Nederlands te bouwen. Commerciële uitgeverijen doen geen onderzoek naar een nieuwe vorm van een digitaal woordenboek dat bijzonder rijk is in structuur en inhoud.
Het ANW richt zich op de standaardtaal. Het 'Algemeen' in de titel moet worden opgevat als: niet gebonden aan een bepaalde regio, een bepaalde groep personen of een bepaald vakgebied. Het hoofdaccent ligt op het geschreven Nederlands. Naast de woorden uit de kernwoordenschat worden in het ANW ook neologismen (nieuwe woorden, nieuwe verbindingen, nieuwe uitdrukkingen, nieuwe betekenissen van al bestaande woorden) beschreven. Deze worden gedurende de looptijd van het project verzameld en toegevoegd.

Het ANW is een multimediaal onlinewoordenboek waaraan geen papieren versie ten grondslag ligt. De woordenboekartikelen zijn hierop afgestemd en er is van meet af aan rekening gehouden met de mogelijkheden, eisen en problemen die aan het maken van een nieuw digitaal woordenboek verbonden zijn.
Het ANW wordt gemaakt op basis van een corpus (een verzameling teksten) van ruim 100 miljoen woorden dat speciaal voor dit project is opgebouwd. Dit corpus bevat materiaal uit alle domeinen van de samenleving. Daarnaast wordt uitgebreid gebruikgemaakt van

relevant materiaal uit andere bronnen, zoals het Corpus Hedendaags Nederlands (CHN) en internet. Het woordenboek wil diensten bewijzen aan gebruikersgroepen die uiteenlopen van de geïnteresseerde leek tot de professionele taalwetenschapper. Het ANW wordt door het INT ook gebruikt als platform om nieuwe, innovatieve computerlinguïstische technieken uit te testen.

Investeringen 2018 :ANW € 211.469 2,4 ft ; Neologismen € 33.639 0,6 fte allemaal uit de vaste subsidies

A.2. Synchroon: Combinatiewoordenboek (nieuw project)

Projectleider: Lut Colman

Woorden krijgen vaak pas echt betekenis als ze gebruikt worden in context, dus in combinatie met andere woorden. Zo wordt pas duidelijk in welke betekenis het werkwoord *blazen* gebruikt is als we het zien in combinatie met *wind*, *rook*, *bestuurder*, *aftocht*, *lachen,* enz. *De wind blaast* is een ander *blazen* dan *hij blies de rook in mijn gezicht*, *hij blies de aftocht*, *dat is lachen geblazen* of *de bestuurder moest blazen*. Woordenboeken illustreren betekenissen dan ook meestal met voorbeeldzinnen zodat men woorden in context kan zien. Maar vaak zijn voorbeeldzinnen alleen niet genoeg. Wie een vreemde taal bijna even vloeiend wil leren spreken en schrijven als een moedertaalgebruiker, moet ook een behoorlijk aantal vaste en minder vaste woordcombinaties leren om goed te kunnen communiceren.

Het INT wil daarom werk maken van een meer systematische inventarisatie en beschrijving van combinaties in een nieuw project *Combinatiewoordenboek* dat zal bestaan uit een database en een onlineapplicatie voor gebruikers. Een pilot wordt eerst ontwikkeld voor een selectie werkwoorden, omdat een systematische beschrijving van zinspatronen met werkwoorden tot nog toe onderbelicht is gebleven in woordenboeken. ‘Combinaties’ gebruiken wij in het project als overkoepelende term voor:

- **collocaties**: frequente en/of typische semivaste combinaties als een *aanbod accepteren* of *afslaan*, *spelers fanatiek* of *enthousiast aanmoedigen*, *supporteren voor, rekenen op*, *huiswerk maken*, *boodschappen doen*.
- **idiomen**: vastere combinaties, vaak met een figuurlijke betekenis, bv. *de boot afhouden*, *de kat de bel aanbinden*, *Spreken is zilver, zwijgen is goud*.
- **patronen**: syntactische constructies die corresponderen met bepaalde betekenissen. Patronen met werkwoorden zijn de zogenaamde valentiepatronen waarin zinsdeelplaatsen bezet worden door sets van woorden (lexicale sets, lexical sets) uit een bepaalde semantische categorie (semantisch type, semantic type).

Met de systematische beschrijving van bovengenoemde combinatietypes ontstaat op termijn als het ware geen lexicon (een inventaris van woorden), maar een 'constructicon' van de Nederlandse taal (een inventaris van constructies).

Gevorderde NT2-leerders en NT2-docenten zijn belangrijke doelgroepen van het project, maar ook NVT-leerders en -docenten. We kunnen hiermee tegemoet komen aan de vragen van de Internationale Vereniging voor Neerlandistiek (ivn) die graag een gratis online woordenboek voor gevorderde leerders ter beschikking willen hebben. Het INT zal dan ook een belangrijke leverancier worden van taaldata op combinatorisch gebied voor taalleerders en docenten, maar daarnaast hebben meer gebruikers baat bij het *Combinatiewoordenboek*:

- tekstschrijvers in de ruimste zin van het woord: professionele schrijvers, amateurschrijvers, leerlingen en studenten die werkstukken en ander proza schrijven. Zij kunnen het woordenboek als schrijfhulp gebruiken.
- lexicografen van algemene woordenboeken en vertaalwoordenboeken. Zij kunnen het materiaal in hun woordenboeken opnemen. Ook het INT-project *Algemeen Nederlands Woordenboek (ANW)* kan het materiaal verwerken.
- taalkundigen in het algemeen en computerlinguïsten. Zij kunnen het materiaal gebruiken voor taalkundig onderzoek of voor toepassingen in natural language processing (NLP), bijvoorbeeld automatisch vertalen (machine translation). Bestaande computationele lexica kunnen uitgebreid worden met nieuwe meerwoordexpressies en het materiaal kan gebruikt

worden als trainingsmateriaal voor machine learning t.b.v. semantisch parseren (semantic parsing), d.i. automatische zinsontleding met de semantische types als toegevoegde betekenisinformatie, hetgeen automatische vertaalprogramma's aanzienlijk kan verbeteren.

- ontwikkelaars van taalprogramma's voor specifieke doelgroepen. Te denken valt aan *Woordcombinaties on demand* in de toekomst, waarbij een opdrachtgever een lemmalijst uit zijn domein kan aanleveren als input voor uitdrukkingen uit dat specifieke domein. Bijvoorbeeld: uitdrukkingen die voor nieuwkomers relevant zijn voor een goede communicatie in de spreekkamer van een zorgverlener, bv. de huisarts; uitdrukkingen die gebruikelijk zijn in bepaalde situaties of in iemands werkomgeving; terminologische combinaties uit een bepaald vakgebied, enz.

Het *Combinatiewoordenboek* zal in fasen opgebouwd worden. In de eerste fase zullen we goede voorbeeldzinnen aanbieden en woordschetsen met het globale gebruiksprofiel van een selectie werkwoorden in lijsten met woorden of woordgroepen die vaak of typisch voorkomen bij het werkwoord. In een volgende fase komen de patronen aan bod. Na de werkwoorden kan het woordenboek uitgebreid worden met combinaties van de naamwoorden (adjectieven en substantieven).
Fase 1 loopt van 2017 tot eind 2018; daarna wordt verder gewerkt en de selectie wordt uitgebreid met eventueel andere doelgroepen. De vlotte voortgang en uitbouw van het project is ook afhankelijk van externe financiering.

Investering 2018: € 82.787 1.1 fte + aanvraag externe financiering

Er komt vanaf 1 januari 2018 een begeleidingscommissie voor het ANW, het corpusonderzoek en de lexica van het hedendaags Nederlands. Deze commissie is samengesteld uit taalkundigen en lexicografen met bijkomende expertise in de meest experimentele e-lexicography.
De voorgestelde namen zijn:

- Iztok Kosem (Researcher at Faculty of Arts, University of Ljubljana; Director of Trojina, Institute for Applied Lexicography, lid van het E-lexicology netwerk)

- Lars Trap Jensen (Society for Danish Language and Literature, Copenhagen, Denmark), lid van het E-lexicology netwerk)
- Alexander Geyken (Digitales Wörterbuch der deutschen Sprache Berlin-Brandenburgische Akademie der Wissenschaften) Of: Lothar Lemnitzer (idem als Alexander Geyken)
- Albert Oosterhof (KU Leuven)
- Timothy Colleman (UGent)
- Dirk Geeraerts (KU Leuven en lid van de adviesraad van het INT)

De beslissing over de samenstelling van de begeleidingscommissie wordt gemaakt door de raad van advies tegen eind 2017.

A.3. Diachroon: historische woordenboeken en GTB

Projectleider: Katrien Depuydt

De beschrijving van de historische woordenschat is te vinden in de historische woordenboeken van het INT. Deze woordenboeken zijn online beschikbaar in de historische woordenboekenportal van het INT. Daarin zijn de belangrijkste historische woordenboeken van het Nederlands opgenomen, het *Oudnederlands Woordenboek* (ONW), het *Vroegmiddelnederlands woordenboek* (WNT), het *Middelnederlandsch Woordenboek* (MNW) en het *Woordenboek der Nederlandsche Taal* (WNT). De data van deze woordenboeken zijn een bron voor GiGaNT en DiaMaNT (zie aldaar).

Werkzaamheden voor de komende jaren zijn:

- Structuurverbeteringen van de opnoemerssecties van met name het WNT om het aantal geattesteerde lemmata in GiGaNT substantieel te verhogen.
- Updates van het portaal met deze structuurverbeteringen, en met correcties van gevonden digitaliseringsfouten, verbeterde lemmatiseringen, metadata etc.
- Koppeling van de woordenboekenapplicatie aan de historische corpora die het INT beheert (zie aldaar).

- Uitbreiding van de inhoud:
    - Mogelijke opname van het *Rhetoricael Glossarium* van Mak en van Van der Voort Van der Kleijs supplement op het Middelnederlandsch Handwoordenboek.
    - Onderzoek (en gedeeltelijk realisatie) naar de opname van eenvoudige lexicografische beschrijvingen van niet in de woordenboeken opgenomen vocabulair uit historische corpora en GiGaNT.

Investering:	€ 39.964	0.4 fte + uitbreiding via stageplaatsen en/of doctorandus

Er wordt een begeleidingscommissie ingesteld voor de historische woordenboeken en het historisch taalmateriaal. De voorgestelde namen zijn:

- Gijsbert Rutten (ULeiden)
- Mike Kestemont (UAntwerpen)
- Matthias Hüning (Freie Universität Berlin, Institut für Deutsche und Niederländische Philologie)
- Margit Rem (Radboud U)
- Rik Vosters (VUB)
- Freek van de Velde (KU Leuven)
- Folgert Karsdorp (Meertens)
- Gunther de Vogelaar (Münster, Institut für Niederländische Philologie)

De beslissing over de samenstelling van de begeleidingscommissie wordt door de Raad van Advies gemaakt tegen eind 2017.

A.4. Dialectwoordenboeken

Projectleider: Tanneke Schoonheim

In 2017 is de expertise van het INT ingeroepen om mee te werken aan het realiseren van een geïntegreerde databank van de drie grote dialectlexicografische databanken, 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). Daarin is de goeddeels verdwenen dialectwoordenschat van het Zuidelijk Nederlands beschreven. Dit werk wordt uitgevoerd in de context van het Herculesproject *Dictionary of the Southern Dutch Dialects (DSDD). An integrated lexicological infrastructure for the Southern Dutch Dialects*, dat onder de leiding staat van Prof. Jacques Van Keymeulen van de Universiteit Gent. Het project is aangevraagd door een consortium van 11 taalkundigen, informatici en geografen. Het project is gestart op 1 januari 2017 en eindigt op 31-12-2019. Het is de bedoeling dat het INT de database en de publieksapplicatie ontwikkelt, waarbij de universiteit Gent het cartografische deel van de applicatie op zicht neemt. Het INT zal applicatie en database in beheer nemen en hosten, ook na afloop van het project. Het is zo opgezet dat op termijn ander dialectwoordenboekenmateriaal kan worden toegevoegd, en dat er gelinkt zal kunnen worden naar andere databases, zoals de historische woordenboeken van het INT. Ook het Woordenboek van de Vlaamse Dialecten (WVD) zal na 2018 door het INT worden gehost en verder ontsloten.

Investering:	€ 49.509	0.5 fte (gefinancierd door externe middelen Herculesfonds)

A.5. Vertaalwoordenboeken

Projectleider: Carole Tiberius

In de afgelopen decennia zijn, onder meer in opdracht van de Commissie Lexicologische Vertaalvoorzieningen (CLVV, 1993-2003), verschillende tweetalige bestanden ontwikkeld waaruit vertaalwoordenboeken van en naar het Nederlands kunnen worden afgeleid. Het betrof talenparen die op de commerciële markt niet spontaan aan bod kwamen. 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 beschikbaar, omdat uitgevers er geen commerciële mogelijkheden meer in zien. Daardoor zijn deze bestanden niet meer beschikbaar voor gebruikers en worden vertalingen tussen diverse talenparen niet meer ondersteund.

Begin 2017 zijn deze bestanden daarom aan het INT overgedragen om ze te beheren, binnen zijn eigen mogelijkheden bij te werken en binnen de grenzen van de vigerende auteursrechten en uitgavelicenties te gebruiken en publiek ter beschikking te stellen via een eigen online(vertaal)woordenschatplatform. De Taalunie behoudt echter haar eigen (gedeelde) auteursrechten en uitgavelicenties.
In september 2017 is het onlineplatform, de vertaalwoordenschat, gelanceerd. Nederlands-Nieuwgrieks en Nieuwgrieks-Nederlands is het eerste taalpaar dat via de applicatie is ontsloten. In de komende beleidsperiode zullen andere taalparen volgen.

Investering 2018:	€ 16.925	0,2 fte (eenmalig extra budget NTU 85,000 euro)

A.6. Vaktaal en terminologische databanken

Projectleiders: Frieda Steurs en Dirk Kinable

Een nieuwe taak voor het Instituut voor de Nederlandse Taal is de studie van vaktaal. De daarmee verbonden terminologische databanken zullen ook op het INT worden gehost en verder ontsloten. Daartoe heeft een van de medewerkers, Dirk Kinable, zich in 2016-2017 bijgeschoold. Het INT zal zich in de komende jaren verder ontwikkelen als een Expertise Centrum Terminologie, dat nationaal en internationaal op de kaart staat. Daartoe zal het INT ook een waardevolle partner worden voor de TermRaad Academy, die studenten uit vertaalopleidingen, overheidsdiensten en Europese instellingen samenbrengt om meer terminologische expertise voor het Nederlands uit te bouwen bij toekomstige vertalers.
Het INT zet zich ook op de kaart als stageplaats voor Nederlandse studenten van vertaalopleidingen die voor de TermRaad onderzoek willen doen.
Tegelijk zal het INT zich inzetten om de werking van de veldvereniging NL-Term te ondersteunen en een nieuw terminologieplatform uit te bouwen waarop actuele kennis, informatie en terminologische hulpmiddelen worden aangeboden. Het INT neemt aldus de oude website “Nedterm” volledig over van de NTU over en zal de informatie via de eigen website aanbieden.
Tenslotte wensen we ook samen te werken met CRITI om zo de banden met Suriname aan te trekken.

Onderwijsterminologie

Een eerste onderzoeksproject terminologie is gestart in 2017 als een pilot: het betreft het onderzoek naar de terminologie van het hoger onderwijs in Nederland en Vlaanderen.
Dit project wordt door het INT samen met U Groningen geleid en gestuurd door het Tuning project. Er is een werkgroep en een stuurgroep van experts uit het hoger onderwijs en van de Europese Nederlandstalige terminologieafdeling. Een eerste database met een beperkte set terminologie wordt begin 2018 opgeleverd en gepresenteerd aan de Taalunie. Dit leidt hopelijk tot een vervolgproject op grotere schaal mits financiering wordt gevonden.

Terminologiecollecties

De eerste belangrijke digitale terminologiecollectie die in 2018-2019 bij het INT zal worden gehost, is het Juridisch Woordenboek Nederlands-Spaans.[1] Dit belangrijke werk is volledig volgens de regels van de terminologieleer opgesteld en vormt een rijke collectie. Aangezien het hier over het Nederlandse rechtssysteem gaat, is het wenselijk dat er in de komende jaren een vervolgproject voor het Belgische rechtssysteem kan worden opgestart. Dat kan alleen mits projectfinanciering.
Voor de komende jaren hopen wij ook vooruitgang te maken met het hosten en ontsluiten van medische thesauri en vaktaalcollecties, zoals de thesaurus Zorg en Welzijn en het Pinkhof Woordenboek. Beide producten bevatten een schat aan Nederlandstalige terminologie en komen onder druk te staan wegens te weinig ondersteuning voor doorontwikkeling.
Het streven is om in de komende jaren meer en meer dergelijke vaktalige collecties bij het INT beschikbaar te stellen voor de gebruiker.

Investering: € 100.774 1.28 fte (gedeeltelijk gefinancierd met extra budget NTU van 70,000 euro)

A.7. Gebarentaalwoordenboeken

Eerste contacten: Frieda Steurs,
Opvolging projectbeschrijving: Vincent Vandeghinste

[1] Juridisch woordenboek Nederlands-Spaans. M.C. Oosterveld-Egas Reparaz & J.B. Vuyck-Bosdriesz.

Er zijn reeds in 2017 voorafgaande gesprekken geweest met onderzoekers gebarentaal in Nederland en Vlaanderen. Nederlandse Gebarentaal (NGT) en Vlaamse Gebarentaal (VGT) zijn heel verschillende talen die morfologisch en linguïstisch totaal andere patronen hebben. Toch gebruiken de Doven en de gebarentaaltolken deze talen steeds in combinatie met het Nederlands.
De onderzoekers gebarentaal uit Nederland en Vlaanderen zouden graag samen met het INT een project uitvoeren in de komende jaren waarbij het lexicon van de beide gebarentalen en de relatie van dat lexicon in verhouding tot het Nederlandse lexicon centraal staat. Een belangrijk thema is het verschil in omvang tussen het Nederlandse lexicon en de "gebarenschat" van respectievelijk NGT en VGT. Ook de invloed vanuit het Nederlands op de gebarentaallexica is een interessant gegeven. Hiervoor zal een onderzoeksproject worden aangevraagd.

Investering 2018:	€ 46.984 0.5 fte + aanvraag externe financiering

B. Corpora

Projectleiders: Carole Tiberius en Katrien Depuydt
Technische leiding Blacklab: Jan Niestadt

Voor het hedendaags taalmateriaal is er ook een corpus samengesteld, en wordt er dagelijks nieuw materiaal aangevoerd bij het INL vanuit kranten (het zogenaamde monitorcorpus). Dat nieuwe materiaal wordt doorzocht om o.a. nieuwe woorden op te sporen (neologismen). Zoeken in een corpus gebeurt met behulp van zoeksoftware. In de komende beleidsperiode zal onderhoud gedaan worden aan het oudste corpusmateriaal, zodat het online beschikbaar kan worden gesteld. Daarnaast zullen ook op regelmatige tijdstippen nieuwe releases van het monitorcorpus beschikbaar worden gesteld. Dat beschikbaar stellen van het INL-corpusmateriaal gebeurt via internet met behulp van een corpuszoekmachine die door het INL is ontwikkeld, BlackLab, die in een eerste release gepresenteerd wordt in 2012 en in de onderhavige beleidsperiode verder uitgewerkt zal worden. Daarnaast zal onderzocht worden of de bestaande lemmatiseerders en PoS-taggers voldoende geschikt zijn om ook middeleeuws materiaal uit de cd-rom Middelnederlands aan het INL-corpus te kunnen toevoegen. Mocht dat het geval zijn, dan wordt ook dit materiaal voor onderzoek en aan het

publiek ter beschikking gesteld. Om de verrijkingswerkzaamheden zo efficiënt mogelijk te laten verlopen zal er een workflow voor PoS-tagging en lemmatisering van taalmateriaal worden opgezet.

Het materiaal dat in de woordenboeken en lexica van het INL terechtkomt, wordt afgeleid uit tekstverzamelingen (corpora). Voor zowel het hedendaags Nederlands als het historisch Nederlands wordt er bij het INT aan corpora gewerkt. Daarvoor bestaan de werkzaamheden uit acquisitie van materiaal, analyse, conversie, toevoegen van metadata, taalkundige verrijking indexeren en beschikbaar stellen, Dat laatste gebeurt met een corpuszoekmachine, BlackLab, die door het INT is ontwikkeld.
Het aanleggen van corpora gebeurt niet alleen op het INT, maar ook door andere partijen. Daar waar mogelijk streeft het INT naar samenwerking. Het kan corpusbouwers ook een plek bieden om hun materiaal permanent te beheren en ter beschikking te stellen. Daarvoor moeten we wel pro-actief de vinger aan de pols houden bij de externe corpuscompileerders: proberen al betrokken te worden bij de corpuscompilatie van in het begin en te adviseren over bv. metadata- en tekstencoderingsformaten.

Voor de komende beleidsperiode zijn de werkzaamheden voorzien zoals beschreven in de volgende drie hoofdstukken.

B.1. Hedendaags Nederlands

Projectleider: Carole Tiberius

De focus van de komende beleidsperiode wordt de verdere uitbouw van een monitorcorpus Hedendaags Nederlands. Dit is een corpus waar continu nieuw materiaal aan wordt toegevoegd. Het materiaal zal in de eerste plaats bestaan uit een verdere uitbreiding van het krantenmateriaal. Een eerste stap daarbij wordt de merger van het materiaal van het INT en QLVL.
Naast het krantenmateriaal zullen we binnen de mogelijkheden ook andere typen taalmateriaal verzamelen. Hierbij moet gedacht worden aan gesproken Nederlands, meertalig materiaal, vaktaal en gebarentaal.

Het verzamelen van materiaal met gesproken Nederlands is geen sinecure. Er zal daarom in eerste instantie gekeken worden naar mogelijkheden om geschreven materiaal dat gezien kan worden als een benadering van gesproken materiaal te verzamelen. Daarbij kan worden gedacht aan ondertitels (captions in eigen taal bij tv-programma's) en parlementaire verslagen. Daarnaast zal er naar financiering gezocht worden om samen met andere partijen een nieuwe versie van het Corpus Gesproken Nederlands te realiseren.
Voor het opsporen van neologismen (nieuwe woorden) zal ook gekeken worden naar materiaal van het web. Daarbij wordt het belangrijk om een goede methode te ontwikkelen om tot een selectie van de daartoe meest relevante sites komen.
Tot slot zal om voeling met het jonge Nederlands [term komt uit CLARIN feedback Vlaanderen] te houden, ook verkend worden welke bronnen van de sociale media gebruikt kunnen worden.
Voor wat betreft meertalig materiaal, vaktaalmateriaal en gebarentaal, zal samenwerking gezocht worden met externe partijen met expertise op die gebieden. Daarbij zien we het INT in de eerste plaats als plaats waar het materiaal uit deze projecten in beheer wordt genomen en ter beschikking gesteld. Naast het uitwisselen van kennis hopen we synergie te kunnen realiseren t.a.v. de metadata die bij het materiaal wordt opgenomen. Voor meertalig materiaal in het bijzonder continueren we de aansluiting bij ELRC (European Language Resource Coordination).
Bij de uitbouw van het corpusmateriaal zal steeds gestreefd worden om het hele taalgebied (Nederland, Vlaanderen, Suriname en de voormalige Antillen) zo goed mogelijk te representeren.

Het verzamelde materiaal zal door het INT taalkundig verrijkt worden. In de komende beleidsperiode zullen wij naast taggen en lemmatiseren het materiaal ook van syntactische annotatie voorzien. Hiervoor zal het INT een bestaande parser gebruiken. Welke dat wordt, zal bepaald worden door de requirements van het INT en de andere beoogde gebruikers van het materiaal.

Er zal gewerkt moeten worden aan de schaalbaarheid van BlackLab om de steeds groter wordende datasets aan te kunnen. Daarnaast zal BlackLab geschikt gemaakt moet worden voor het doorzoeken van syntactisch geannoteerd materiaal.

Vanwege IPR-restricties is het distribueren en toegankelijk maken voor een ruimer publiek van hedendaags materiaal bijzonder lastig. De verwachting is dat het INT niet van elke dataleverancier de toestemming zal krijgen om het materiaal aan derden ter beschikking te stellen in een corpusapplicatie zoals voor het huidige Corpus Hedendaags Nederlands, dan wel als distribueerbare dataset. Daarom zal onderzocht worden hoe aan externe gebruikers toch toegang tot het materiaal verschaft kan worden, zonder daarbij de afspraken met betrekking tot IPR te schenden. Hierbij kan gedacht worden aan andere exploitatiemogelijkheden van het tekstmateriaal in de vorm van frequentielijsten, ngrammen en statistische informatie met behulp van BlackLab server. Maar er zal ook onderzoek gedaan worden naar mogelijkheden om toegang te verschaffen voor toolontwikkelaars en onderzoekers die direct toegang tot de dataset nodig hebben. Hierin is voorzien in de in het dit jaar ingediende project CLARIAH+.

Investering 2018: €153.886 1,9 FTE

B.2. Historisch Nederlands

Projectleider: Katrien Depuydt

Voor het INT is, zeker gezien de teruggang van de historische taalkunde aan de universiteiten, het onderhoud van historisch taalmateriaal een kerntaak.
Het INT heeft momenteel twee historische corpora online staan, het *corpus Gysseling* en de *Brieven als buit* (brievenalsbuit.inl.nl).

De opbouw van historisch diachroon corpusmateriaal wordt in deze beleidsperiode in eerste instantie voortgezet in de context van het NWO-project Nederlab (zie aldaar). In dat project is het INT verantwoordelijk voor zowel de lexicon- als de corpusbouw.
In de komende periode zal er op drie hoofdlijnen gewerkt worden:

1. Betere ontsluiting van het corpusmateriaal dat door het INT beheerd wordt. Dit geldt met name de corpora die gekoppeld zijn aan de historische woordenboeken: ONW-corpus, Corpus Gysseling, Corpus cd-rom Middelnederlands. Voor deze corpora zal een verbeterde applicatie worden ontwikkeld, nauw gekoppeld aan de relevante woordenboeken. Hiermee is dan tevens het probleem van de (technische) veroudering van de cd-rom Middelnederlands ondervangen.
2. Verbetering van de kwaliteit van de automatische taalkundige verrijking (met name part-of-speech tagging, lemmatisering en naamherkenning) voor historisch materiaal. Hiervoor zal het INT vooral werken aan de uitbreiding van beschikbaar evaluatie- en trainingmateriaal. Op basis hiervan
    - Kunnen bestaande systemen worden geëvalueerd
    - Kunnen bestaande systemen worden verbeterd door hertraining
    - Kunnen competities worden uitgeschreven die computationeel linguïsten stimuleren op dit gebied actief te zijn
3. Uitbreiding van het beschikbare corpusmateriaal, vooral voor teksttypen die in het op dit moment beschikbare materiaal ondervertegenwoordigd zijn. Dit geldt met name voor Vlaams materiaal; het zou ook wenselijk zijn een meer gebalanceerd historisch corpus, gebaseerd op een teksttypologie, te kunnen aanbieden aan onderzoekers.

Hiervoor zal samenwerking met andere partijen en externe financiering gezocht moeten worden.

Investering 2018:	€ 19.826	0.2 fte + zoeken naar externe financiering, stages etc.

B.3. Nederlab

Projectleider: Katrien Depuydt

Digitale historische teksten worden nu nog op allerlei verschillende plaatsen, door verschillende instellingen en op verschillende manieren beschikbaar gesteld. Eind 2011 werd een NWO-grootaanvraag ingediend onder de naam 'Nederlab: Laboratory for research on the patterns of change in the Dutch language and culture'. Het project heeft als doel alle digitale teksten die belangrijk zijn voor ons nationale erfgoed samen te brengen in een gebruikersvriendelijke en vrij toegankelijke webinterface. Het consortium bestaat uit onder meer het Meertens Instituut, het Huygens ING, de Digitale Bibliotheek voor de Nederlandse Letteren, de Radboud Universiteit en het INT. De financiering bedraagt 3.4 miljoen euro waarvan ca. 2 miljoen door NWO wordt gedragen. Het project is gestart op 1 januari 2013 en loopt af op 30 juni 2018. Het INL is vanaf aanvang track supervisor lexicale data geweest en met name verantwoordelijk voor lexicondata om het historische materiaal mee te verrijken. Sinds 1 januari heeft het INT de verantwoordelijkheid over de hele datatrack, en dus ook de corpusbouw op zich genomen; hieraan zal nog tot medio 2018 worden gewerkt.

Investering 2018:	€ 50.000	0,5 FTE	(volledig externe financiering van NWO)

C. Lexica

Projectleider: Katrien Depuydt

Om de inventarisatie en de beschrijving van de historische en hedendaagse woordenschat van het Nederlands nog beter te faciliteren, wordt een centrale lexicale data-infrastructuur aangelegd in de vorm van een computationeel lexicon, waarin aan ieder woord uit de Nederlandse taal taalkundige en semantische informatie wordt gekoppeld. Het lexicon is een computationeel lexicon, omdat het primair ingezet kan worden door een computer om corpusmateriaal automatisch te verrijken. Het lexicon fungeert ook als een thesaurus, waarin taalkundige informatie over de totale woordenschat van het Nederlands wordt beschreven, die in diverse producten van het INT terug kan komen. Het is de bedoeling om uit te komen op een database waarin ieder door het INT in diverse producten beschreven woord een persistent id heeft (een burgerservicenummer of rijksregisternummer) waaraan binnen en buiten het INT informatie gekoppeld kan worden.

Deze data-infrastructuur wordt naar buiten toe ontsloten door middel van webservices (de bestaande lexiconservice die operationeel is in bv. het KB-platform Delpher, is een voorbeeld) en publicatie in het in CLARIAH opgezette linked open data platform. Een afgeleid historisch woordvormenlexicon wordt als dataset beschikbaar gesteld t.b.v. bijvoorbeeld verbeterde OCR.

Om deze data-infrastructuur aan te leggen is voor een modulaire aanpak gekozen en voor de realisatie zijn twee projecten geformuleerd. In het GiGaNT-project wordt gewerkt aan de taalkundige beschrijving van de woordenschat, waardoor een computationeel lexicon gerealiseerd wordt waarin spelling- en paradigmatische variatie van de woordenschat door de eeuwen heen wordt beschreven. In het DiaMaNT-project wordt gewerkt aan een semantische laag, waarin woorden aan elkaar gekoppeld worden omdat ze conceptueel verwant zijn.

C.1. GiGaNT (Groot Geïntegreerd lexicon van de Nederlandse Taal)

Projectleider: Katrien Depuydt

GiGaNT beoogt een geïntegreerd lexicon te zijn, maar is opgebouwd uit modules (componenten voor het vocabulair voor de historische woordenboeken en het modern lexicon). De afgelopen jaren is al gewerkt aan enerzijds de onderlinge koppeling van de lemmata van ONW, VMNW, MNW en WNT, die in een vergevorderd stadium is, en de koppeling van het historische deel (HILEX) en het moderne deel (MOLEX) van het lexicon, en koppeling van MOLEX aan het ANW. Dit werk zal worden afgerond. Aan GiGaNT zijn verder de volgende werkzaamheden voorzien:

- Uitbreiding van de morfosyntactische informatie bij de woordvormen in het historisch deel van het lexicon.
- Het werk aan de morfologische component zal weer worden opgepakt, met het accent op diachrone morfologie.
- Uitbreiding op basis van corpusmateriaal. In eerste instantie is te denken aan het integreren van materiaal uit reeds taalkundig verrijkte bronnen.
- Voor de historische component (HILEX) zijn alle woorden gekoppeld met attestatie-informatie via de woordenboekcitaten. Attestatie-informatie zal ook worden toegevoegd aan de moderne component (MOLEX).

Investering 2018:	€ 127.966	1.4 fte	eigen vaste subsidie

C.2. DiaMaNT (Diachroon seMANtisch lexicon van de Nederlandse Taal)

Projectleider: Katrien Depuydt

In dit project dat mede uitgevoerd wordt binnen CLARIAH, wordt gewerkt aan de ontwikkeling van een Diachroon semantisch lexicon van het Nederlands (DiaMaNT). Dit project levert een bouwwijze en eerste versie van een diachroon semantisch lexicon op. Het diachroon semantisch lexicon heeft als doel een hulpmiddel te bieden bij tekstontsluiting en bij het onderzoek naar begrippen door de eeuwen heen. Het legt lexiconrelaties 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.
Het lexicon vormt een laag op GiGaNT-HILEX, en heeft dus als belangrijkste bron de historische woordenboeken van het INL. In de komende beleidsperiode zullen de volgende werkzaamheden worden verricht:

- Voltooiing van de structurering van de extractie van synoniemdefinities (aanbrengen van koppelingen op lemmaniveau).
- Integratie met beschikbare semantische datasets (Open Dutch Wordnet, Brouwers, …).
- Onderzoek naar de manier waarop betekenisbetrekkingen (metonymie, metafoor) kunnen worden opgenomen, en dataontwikkeling in dat kader.
- Werk aan automatische verwerving van relevante semantische informatie uit corpusmateriaal door middel van distributionele semantiek.

Investering 2018:	€ 151.664	1.9 fte

D. Bredere taalinfrastructuur: grammatica, spelling

D.1. e-ANS

Projectleider: Frank Landsbergen

Binnen het ANS-project verzorgt het INT WP4, 'ergonomie en digitale omgeving'. Dit betekent dat het INT verantwoordelijk wordt voor de technische kant: (1) de conversie van de huidige digitale ANS naar xml (in samenwerking met Peter-Arno Coppen van de RU). (2) het ontwikkelen en bijhouden van de auteursomgeving, waarin auteurs (a) het bijwerken van de bestaande ANS-teksten, en (b) nieuwe teksten zullen schrijven, (3) het verzorgen van opslag en backups van het materiaal en (4) ontwerp, bouw en beheer van de nieuwe ANS-website. Dit betekent dat de site een moderne look-and-feel krijgt, en dat we optimaal gebruik gaan maken van de huidige mogelijkheden om de site te koppelen aan allerlei mogelijk interessante digitale bronnen zoals woordenboeken, corpora, ontologieën en sites als taalportaal en taaladvies.net. Door de informatie offline beschikbaar te maken in een databank, kan die ook aan andere bronnen worden gekoppeld en op andere platforms worden gepresenteerd. Daarnaast zal de site zo worden opgebouwd dat er ruimte blijft voor de toevoeging van een zogenaamde 'didactische laag', een laag die de ANS beter inzetbaar moet maken voor het onderwijs, maar waarvan de precieze invulling op dit moment nog niet vastligt.

Investering 2018 :	€ 16.550	0.2 fte	(extra financiering van NTU)

D.2. Taalportaal

Projectleider: Frank Landsbergen

Het Taalportaalproject is officieel beëindigd aan het einde van 2015. Het INT vormt, samen met de Fryske Akademie en het Meertens Instituut, de *governance* die het project ook na deze einddatum blijft beheren. Omdat er nog steeds materiaal wordt vernieuwd en toegevoegd (met name in Zuid-Afrika maar ook in Nederland) voert het INT reguliere updates (tweemaal per jaar) uit van de site. Daarnaast is de binnen het INT ontwikkelde auteursomgeving in gebruik voor de ANS.

Investering 2018:	€ 5.781	0,1 fte

D.3. Spelling

Projectleider: Katrien Van pellicom

Sinds 1995 is de Woordenlijst Nederlandse Taal bij ons instituut ondergebracht. De gedrukte woordenlijst werd in 2005 en 2015 werd geactualiseerd, volgens de besluiten van de Nederlandse Taalunie, terwijl de onlineversie sinds 2015 driemaandelijks geüpdatet wordt. Op die manier spelen we beter en sneller in op wijzigingen en vernieuwingen in het taalgebruik. Tijdens de komende beleidsperiode 2018-2022 is geen actualisering van de gedrukte woordenlijst gepland, maar de online-updates worden vanzelfsprekend wel verder uitgevoerd. Voor deze uitbreiding van het spellingmateriaal zullen trefwoorden worden geselecteerd uit het modern materiaal van zowel België en Nederlands als van Suriname en de voormalige Antillen, waaraan de nodige verrijking zal worden toegevoegd. Concrete selectie kan o.a. gebeuren op basis van de aanwezigheid van potentiële spellingproblemen, denken we maar aan de tussenletters, spelling met c of k enz. Niet alleen op trefwoordniveau is er uitbreiding voorzien: ook ruimer zien we mogelijkheden. Zo kan de koppeling tussen het ANW en het spellingmateriaal / modern lexicon (GiGaNT MOLEX) (die momenteel wordt uitgevoerd) worden aangewend om betekenisinformatie mee te geven bij spellinggerelateerde opzoekingen. Ook zijn nieuwe functionaliteiten te verwachten voor de onlineversie. Daarnaast zou een nieuw te ontwikkelen spellingcontrolemodule een significante bijdrage kunnen leveren aan toepassingen als woordenlijst.org en Spelspiek. Ook blijven de HulK-controles ten behoeve van de toekenning van het officiële spellingkeurmerk aan allerlei uiteenlopende producten een taak van het instituut.

Ze werd hierboven al genoemd: de Spelspiektoepassing. De huidige versie dateert uit de opstartfase als STEVIN-demonstratieproject in 2007 en werd toentertijd opgezet als een tijdelijk project met verschillende partners, waardoor het jammer genoeg een zwarte doos is met een erg gesloten structuur. Omdat deze interactieve, automatische spellingtoepassing nog steeds gebruikt wordt en volgens ons veel potentieel bezit, willen we deze graag een tweede leven geven: Spelspiek 2.0. Uiteraard moet dan eerst worden gekeken welke (soorten) modules voor spellingcontrole er al bestaan. Op basis daarvan kunnen dan de nodige

voorstellen en aanpassingen worden gedaan aan Spelspiek. Het is ook expliciet de bedoeling dat we onze eigen steeds groeiende data erin kwijt kunnen.
Verder willen we – uiteraard in samenspraak met de NTU - bekijken of de mogelijkheid bestaat om de hosting van *woordenlijst.org* (spelling-API) bij het instituut onder te brengen. Zo kunnen we korter op de bal spelen als het gaat over updates, errata enz. Bovendien kunnen we vanuit een dergelijke structuur beter inspelen op bepaalde noden of wensen, zoals bv. het uitbreiden van *woordenlijst.org* met bepaalde nieuwe functionaliteiten.

Tenslotte willen we in de komende jaren in overleg met de NTU bekijken hoe de rol van het INT kan evolueren in een nieuw en breder samenwerkingsverband ten behoeve van Taaladvies.net en de onderlinge koppeling van diverse bronnen en materialen voor spelling, woordenschat, grammatica e.d. om de online taaladviesverlening tot een nog hoger niveau te tillen dan nu al het geval is.

Investering 2018:	€ 94.616	1 fte	(waarvan 40 000 euro extra financiering NTU)

D.4. Kennisbank Begrijpelijke Taal

Projectleider: Frank Landsbergen

De Kennisbank Begrijpelijke Taal is het eindresultaat van een NWO-project dat uitgevoerd is aan de Universiteit Utrecht. Deze brengt het begrijpelijkheidsonderzoek in kaart voor wetenschappers, communicatie- en taaladviseurs, studenten, beleidsmakers en andere belangstellenden. De kennisbank bevat honderden onderzoeken naar de begrijpelijkheid van teksten. Dit zijn uitsluitend publicaties over empirisch onderzoek naar het verband tussen begrip (of verwerking) en kenmerken van teksten. Het grootste deel van de publicaties is Engelstalig, aangezien het meeste relevante onderzoek in die taal gepubliceerd is. Het doel van de kennisbank is de wetenschappelijke expertise op dit terrein beschikbaar te maken voor iedereen die in onderzoek of in de praktijk bezig is met de begrijpelijkheid van teksten en documenten.

In april 2017 heeft het INT op verzoek van Leo Lentz (UU) besloten de kennisbank over te nemen. In het najaar van 2017 zal de site worden overgezet. Daarnaast zal er elke 5 jaar een

update plaats gaan vinden van de inhoud van de kennisbank. Het initiatief hiervoor ligt bij het INT; omdat de expertise vooral bij de UU ligt zal daar ook de uitvoering plaatsvinden.

Investering 2018:	€ 332	8 uur

E. Softwaretools (ontwikkeling, beheer en onderhoud)

E.1. CLARIAH Nederland en Vlaanderen

Projectleider: Jesse De Does,
in samenwerking met Kris Heylen en Vincent Vandeghinste

In het kader van CLARIAH en het mogelijke vervolgproject CLARIAH-plus draagt het INT bij aan de digitale ontwikkeling van de geesteswetenschappen in Nederland.
CLARIAH (Common Lab for Research in the Arts and Humanities) is erop gericht een gemeenschappelijke infrastructuur tot stand te brengen voor data-intensief wetenschappelijk onderzoek.

Het CLARIAH-project is gestructureerd in vijf werkpakketten, waarbij het INT deelneemt aan werkpakketten 2 (technische infrastructuur) en 3 (focusgebied taalkunde).
In het kader van de technische infrastructuur werkt het INT aan de deeltaken CLEVER (vaststellen van best practice richtlijnen voor inbeheername van onderzoeksresultaten en tools), PICCL (een workflowsysteem dat onderzoekers ondersteunt in het traject van scans naar doorzoekbaar corpus) en het diachroon semantisch lexicon DiaMaNT (zie aldaar), dat voor CLARIAH een centraal aanknopingspunt voor een gestructureerde collectie concepten biedt.
Binnen het aandachtsgebied taalkunde wordt door het INT gewerkt aan:

- Standaarden en dataformaten ten behoeve van interoperabiliteit, waarbij Linked Open Data (RDF) een belangrijke rol speelt.
- Verbeterde ontsluiting van belangrijke databronnen (inbeheername en aanpassing van webCELEX, optimalisatie en uitbreiding van OpenSoNaR).
- Ondersteuning van de geesteswetenschappelijke onderzoeker in het traject van tekstverzameling naar onderzoeksinstrument (Autosearch, een op BlackLab gebaseerde tool waarmee de onderzoeker zelf een doorzoekbaar corpus kan creëren).
- WP3 search, waarbij een op CLARIN voortbouwende infrastructuur wordt gebouwd waarin diverse typen databron (corpora, treebanks, lexica) gecombineerd doorzoekbaar zijn. Hier wordt eerst zorg gedragen dat de afzonderlijke bronnen op een

bij de centrale infrastructuur aansluitende manier beschikbaar zijn ("local search"). "Federated search", voortbouwend op CLARIN, zorgt voor de gecombineerde bevraagbaarheid van gelijkaardige materialen; door middel "Chaining search" is het mogelijk gecombineerde zoektrajecten over heterogene databronnen, zoals corpora en lexica, op te zetten.

Het werk aan het uitbreiden en nader invullen van deze onderzoeksinfrastructuur stopt niet bij het aflopen van het CLARIAH-project eind 2018. Een vervolgproject CLARIAH-plus is aangevraagd, waarin het accent nog meer gericht is op het concreet ondersteunen van de onderzoeker door middel van het totstandbrengen van (virtuele) onderzoeksomgevingen voor bepaalde taken.

Ook los van het doorgaan van dit project ziet het INT het als zijn verantwoordelijkheid de ontwikkelde infrastructuur te onderhouden en verder te ontwikkelen. Een belangrijke ontwikkeling is voorts dat in toenemende mate met Vlaamse onderzoeksinstituten samengewerkt wordt op het gebied van onderzoeksinfrastructuur. Het onder A4 genoemde Herculus-project is daarvan een goed voorbeeld. Het INT is nu samen met Vlaamse onderzoeksinstituten bezig te onderzoeken hoe een Vlaamse pendant van het CLARIAH-project tot stand gebracht kan worden.

Investering 2018 : € 178.930 2.4 fte (via externe financiering Clariah)

E.2. INT Impact centrum en digitization.eu

Projectleiders: Frieda Steurs en Katrien Depuydt

Het INT heeft mede aan de wieg gestaan van het IMPACT Centre of Comptetence (www.digitisation.eu) en is momenteel chair van de executive board. Het IMPACT Centre of Competence is een non-profitorganisatie bestaande uit publieke en commerciële organisaties met het doel de digitalisering van historisch materiaal "beter, sneller, en goedkoper" te maken. Het centre voorziet data, tools, services en expertise op het gebied van document imaging, taaltechnologie en het processen van historisch tekstmateriaal. Het INT heeft in de afgelopen jaren ruim expertise gedaan op het gebied van digitalisering van historisch

taalmateriaal, zowel door OCR (Optical Character Recognition) als HTR (handwritten tekst recognition), en heeft ruime ervaring met het ter beschikking stellen van historisch taalmateriaal. In de vorm van trainingen en advies deelt het INT deze kennis. Daar waar het centre in eerste instantie de bibliotheken als doelgroep had, bereikt het, mede door de betrokkenheid van het INT in de digital humanities community en CLARIN, een steeds grotere doelgroep. Naast voortzetting van de huidige activiteiten, zal daarom worden ingezet op het uitbreiden van de doelgroep naar de digitale geesteswetenschappen, mede door samenwerking met gebruikers en ontwikkelaars van geesteswetenschappelijke infrastructuurprojecten.

E.3. European Language Resources Coordination Initiative (ELRC)

Projectleider: Carole Tiberius

Het INT is betrokken bij het ELRC-initiatief dat als doel heeft publieke tekstuele data beschikbaar te krijgen ten behoeve van de Connecting Europe Facility Automatic Translation (CEF.AT).

Het doel van ELRC is om op grote schaal taaldata te verzamelen die gebruikt kunnen worden voor het ontwikkelen van automatische vertaalsystemen voor publieke diensten in alle EU lidstaten, Noorwegen en IJsland, zodat er beter tegemoetgekomen kan worden aan de dagelijks noden van publieke diensten in heel Europa. ELRC is een unieke inspanning voor het collecteren van data uit de publieke sector en beoogt de CEF.AT te voorzien van taal- en vertaaldata (monolinguale en bilinguale data) die relevant zijn voor de dagelijkse noden van de Europese nationale overheden. 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.

De eerste fase van het ELRC initiatief liep af in het eerste kwartaal van 2017. Vanaf april 2017 is de tweede fase van het initiatief begonnen, die doorloopt tot 2019. In deze fase wordt het verzamelen van taalmateriaal verder uitgebreid, met name naar Digital Service Infrastructuren. Hiervoor zal onder meer een nationale workshop voor potentiële dataleveranciers georganiseerd worden

F. Digitale taalmaterialen (geïntegreerde webwinkel)

INT, TST en CLARIN-materialen

Projectleider: Bob Boelhouwer

In 2016 is de TST-Centrale (TSTC) door de NTU overgedragen aan het Instituut voor de Nederlandse Taal. De TST-Centrale is opgericht in het kader van het STEVIN-project. Dit was een groots opgezet onderzoeksprogramma dat erop gericht was om de ontwikkeling van Taal- en Spraaktechnologie in Nederland te bevorderen. Het doel van de TST centrale is het aanbod van taaldata actueel houden en uitbreiden, ondersteuning van dataproducenten, ondersteuning van gebruikers. Daarbij willen we in de komende jaren ook bekijken hoe we het gebruik van de TST in het Nederlands in nieuwe toepassingen kunnen stimuleren en kunnen we een aantal materialen integreren en doorontwikkelen. Hiervoor zullen voorstellen bij de NTU worden neergelegd en zijn extra eenmalige middelen voorzien.

Om de TST-Centrale goed te laten functioneren willen we in de komende jaren inzetten op de herkenbaarheid en duidelijke profilering. Daarbij moet het aanbod van taaldata actueel worden houden en ook voortdurend uitbreiden.
Een belangrijk deel van het huidige aanbod van de TSTC bestaat uit producten die zijn ontwikkeld in het kader van het STEVIN-programma. Die data zijn inmiddels 6 tot 12 jaar oud en reflecteren het taalgebruik van dat moment. Het is daarom van belang dat in ieder geval een aantal producten wordt geactualiseerd.

Daarnaast moet er een actief beleid ontwikkeld worden om de catalogus van de TSTC uit te breiden met nieuwe datasets. Een belangrijke doelstelling van de TSTC is om hergebruik en conservatie van data mogelijk te maken. Bij onderzoeksprojecten waarbij taaldata worden ontwikkeld is meestal geen voorziening getroffen om die data voor langere tijd beschikbaar te stellen. Overdracht van die data aan de TSTC maakt dat wel mogelijk.

Een ander belangrijk aspect van het uitbreiden met nieuwe datasets is het beschrijven van diverse variëteiten van het Standaardnederlands en onderzoek ondersteunen naar specifieke taal-en spraakstoornissen zoals dyslexie en afasie en de uitwisseling van open data in de cultuur – en erfgoedsector.

Daarnaast is bij het INT veel kennis aanwezig betreffende digitalisering en het verrijken van taaldata. Ook heeft het INT daarvoor een groot aantal tools ontwikkeld. Een product als Blacklab wordt inmiddels veel door andere partijen gebruikt. Het is een goed idee om die producten in de catalogus op te nemen en actief te promoten. Daarnaast is het wellicht mogelijk een webservice in te richten met een overzicht van beschikbare niet-commerciële en commerciële tools voor bewerking van taaldata[2] alsmede een verzameling 'Best Practices'. Om het gebruik van de producten te bevorderen is het zinvol om de procedure voor het verwerven van de data te vereenvoudigen waar dat kan. De bestelprocedure is momenteel vrij omslachtig. Met name waar het gaat om niet-commercieel gebruik is die procedure aanzienlijk te vereenvoudigen. Wel is het belangrijk om de gebruikers zorgvuldig te informeren over wat is toegestaan met de data.

Bij de overdracht van de TSTC aan het INT in 2016 is afgesproken dat er een integratie mogelijk is van de producten van de TSTC en die van het INT. De TSTC heeft momenteel een aparte website waarop de catalogus gepresenteerd wordt en waar gebruikers producten kunnen bestellen. Het INT heeft op de website twee pagina's waarop producten worden aangeboden: 'Taalmaterialen' met de producten die door het instituut (soms tezamen met partners) zijn ontwikkeld en de 'CLARIN Portal' waar de producten zijn te vinden die binnen CLARIN zijn ontwikkeld en die via de CLARIN-login beschikbaar zijn gesteld aan de CLARIN-gemeenschap. Vanuit het oogpunt van de gebruikers is het van belang dat er dat alle producten op een centrale plaats beschikbaar worden gesteld. Een belangrijke doelstelling van het INT is namelijk een centraal loket voor Nederlandse taalmaterialen. Daarom dient er een afstemming te worden bereikt tussen het aanbod van de TSTC en het INT. Daarbij zijn er nog een aantal zaken die aandacht nodig hebben.

- Veel producten in de portefeuille van de TSTC zijn geproduceerd door andere partijen of zijn eigendom van anderen. Die andere partijen stellen prijs op herkenbaarheid. Dat wordt bemoeilijkt als de TSTC en de TSTC-website volledig worden geïntegreerd in het INT en de producten worden gepresenteerd als INT-producten;

[2] Zie bijvoorbeeld: http://www.digitisation.eu/tools-resources/tools-for-text-digitisation/

- Anderzijds wil het INT wellicht voorkomen dat ze geassocieerd wordt met producten die zij niet zelf heeft geproduceerd.

Deze overwegingen zijn een aanleiding om te besluiten om het apart domein voor de TSTC op te heffen en de distributie van de taalmaterialen als functie van het INT naar buiten te brengen. De website van het INT zal worden uitgebreid met een uitgebreide zoekinterface waarbij de volledige catalogus doorzocht kan worden. De bestelprocedure voor een groot aantal producten zal daarbij worden vereenvoudigd. Het is daarbij belangrijk de rol van de makers van de producten en van de NTU als financier te benadrukken.
Er is overigens nog een andere reden om de naam van de TSTC te laten verdwijnen. Producten voor taal- en spraaktechnologie maken weliswaar een belangrijk onderdeel uit van catalogus, maar veel producten zoals de historische corpora zijn daar niet voor bedoeld.

Investering 2018:	€ 186.265 2.4 fte	(200,000 euro beschikbaar via de Taalunie)

G. CLARIN

G.1. Het INT als CLARIN[3]-centrum.

Verantwoordelijke : Griet Depoorter
In samenwerking met Frieda Steurs en Vincent Vandeghinste

Als CLARIN-centrum is het INT een van de knooppunten in een Europees netwerk dat ten dienste staat van zowel de linguïstiek en de geesteswetenschappen als de maatschappij in het algemeen. Gebruikers van dat netwerk krijgen laagdrempelig toegang tot talige data, tools en andere diensten, waar ze ook zijn en waar de materialen die ze gebruiken zich ook in het netwerk bevinden.
Via de NTU zijn we ook het CLARIN-centrum voor Vlaanderen. Om de zichtbaarheid in Vlaanderen te vergroten is bij de aanvang van 2017 een bevraging georganiseerd bij de

[3] CLARIN staat voor Common Language Resources and Technology Infrastructure.

Vlaamse onderzoekers. De resultaten hiervan zullen door het INT gebruikt worden om in Vlaanderen meer diensten te verlenen.

Een CLARIN-centrum zijn betekent ook dat we kwaliteit leveren in het werk dat we doen: als we onderzoeksresultaten in beheer nemen, zorgen we ervoor dat ze beschikbaar worden gesteld als veilige, betrouwbare en goed gedocumenteerde producten. We brengen desgewenst advies uit aan beleidsmakers, opdrachtgevers en financiers. Zo dragen we actief bij aan een sterke positie van het Nederlands in de informatiemaatschappij.

Het INT zal in de komende jaren zich heel sterk profileren als ondersteunend voor alle onderzoekers, docenten en studenten in Vlaanderen die meer nood hebben aan digitale taalmaterialen en tools om onderzoek te doen.

Investering 2018: € 43.057 0,4 fte

G.2. DARIAH (Digital Research Infrastructure for the Arts and Humanities)

Verantwoordelijke: Frieda Steurs en Vincent Vandeghinste

Er wordt hard gewerkt in Vlaanderen aan een DARIAH-VL Virtual Research Environment Service Infrastructure (VRE-SI). Hier wordt het beleid bepaald voor de digital humanities in Vlaanderen. Op 25 oktober 2017 zal er een strategische vergadering zijn voor Dariah VL waar het INT ook aanwezig zal zijn. Het INT zal meewerken aan het Exploitatie- en Werkplan met Kritische Prestatie-Indicatoren voor 2017-2018.

Prof. dr. Christophe Verbruggen (UGent) is zowel coordinator van het DARIAH-VL consortium en is de DARIAH-BE National Coordinator. Het INT zal in de komende jaren actief deelnemen aan Dariah-VL en werkt daartoe ook samen met Sally Chambers (UGent) die lid is van het Dariah-EU senior management team. Verder zijn ook Dirk van Hulle (UA) en Mark De Pauw (KU Leuven) hierbij betrokken.

Op beleidsniveau zal er overleg zijn met de Vlaamse verantwoordelijken voor onderzoeksinfrastructuur bij EWI : Michele Oleo, Peter Spyns en Bart Dumoulin.

H. Samenwerking en netwerken

Algemene leiding: Frieda Steurs

Het INT zal zijn positie in de komende vijf jaren versterken door deel te nemen in verschillende samenwerkingsverbanden en netwerken. Voorbeelden van samenwerking zijn bijvoorbeeld de CLARIN-projecten (www.clarin.nl en www.clarin.eu), het TaalPortaal (www.taalportaal.org), CLARIAH (www.clariah.nl), Nederlab, het Meldpunt Taal (www.meldpunttaal.be of www.meldpunttaal.nl).

Met de Fryske Akademy wordt een intensieve samenwerking opgezet, met als doel ook samen projecten aan te vragen voor het Nederlands én het Fries. Verder wordt de samenwerking met twee onderzoeksgroepen, nl. LUCL (Leiden) en QLVL (Leuven) in de komende jaren geïntensifieerd om samen onderzoeksprojecten uit te voeren.

Nieuw is dat het INT lid wordt van EFNIL, the European Federation of National Institutes of Languages, waar het een actieve rol kan spelen als aanbieder van digitale taalmaterialen.

Verder zijn er contacten en samenwerking met:

- Zuid-Afrika (voor het Taalportaal en lexicografisch onderzoek) (specifiek met het WAT (woordenboek Afrikaanse Taal), de universiteit van Stellenbosch en North West University
- KNAW digital humanities
- WOG Digital Humanities Vlaanderen en Dariah (zie G2.)
- NL-Term
- Termraad Academy
- European Association for Terminology
- TermNet
- DANS
- CTRITI (Suriname)

Het INT is deelnemer bij de stichting Nederlandse Organisatie voor Taal en Spraaktechnologie (NOTaS). NOTaS behartigt de belangen van de Nederlandse bedrijven en kennisinstellingen uit de sector taal- en spraaktechnologie en is voor het INL een belangrijk communicatiekanaal naar het bedrijfsleven. Voor Notas zal het INT in de komende jaren ook de brug vormen naar de Vlaamse bedrijven. Er zal contact worden gelegd met de Vlaamse innovatiecentra en met iminds-imec.

I. Onderzoek: Europese aanvragen en competitieve projecten

Algemeen verantwoordelijke: Kris Heylen
in samenwerking met Vincent Vandeghinste

Belangrijke hoofdtaken van een kennis- en onderzoeksinstituut zoals het INT bestaan erin om fundamenteel en toepassingsgericht onderzoek te verrichten en daarover te publiceren, universitair onderwijs te verzorgen, resultaatgericht te werken naar producten toe, werk te blijven maken van de kennisbenutting dan wel de popularisering van de taalwetenschap en samenwerkingsverbanden aan te gaan om de opgedane kennis te delen en uit te wisselen.

Voor de komende meerjarenbeleidsperiode zet het INT in op breed taalkundig onderzoek in verband met taalvariatie. Dit gebeurt met de nieuwste inzichten uit de corpuslinguïstiek en de computerlinguïstiek. Het INT wil in samenwerking met andere onderzoekscentra in Nederland en Vlaanderen nieuwe producten ontwikkelen.

Om de positie van het INT-onderzoek in het veld te verstevigen en om de bestaande onderzoeks- en samenwerkingsbanden tussen het INT en gerenommeerde (inter)nationale onderzoekers en instituten aan te halen, werken we samen in grotere Europese projecten. Dit zorgt voor de nodige inzichten in de allernieuwste ontwikkelingen in de taaltechnologie. Verschillende INT-medewerkers geven gastcolleges en studievakken (op bachelor- en masterniveau) aan verschillende Vlaamse en Nederlandse universiteiten en leggen ze werkbezoeken af bij vergelijkbare instituten en andere partners in binnen- en buitenland om kennis uit te dragen en nieuwe kennis op te doen.

I.1. European Lexicographic Infrastructure (ELEXIS) (2018-2021)

Projectleider: Carole Tiberius

ELEXIS is een Europese Horizon 2020 project met als doel een duurzame infrastructuur voor e-lexicografie te creëren die het mogelijk maakt 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 afloopt. 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 Europese partners: Jožef Stefan" Instituut, Slovenië (hoofdaanvrager); Lexical Computing CZ s.r.o., Tsjechië; Instituut voor de Nederlandse Taal, Nederland; Universita degli Studi di Roma "La Sapienza", Italië; National University of Ireland, Galway, Ierland; Oesterreichische Akademie der Wissenschaften, Oostenrijk; Centar za digitalne humanisticke nauke, Servië; Magyar Tudományos Akadémia Nyelvtudományi Intézetének, Hungarije; Institute for Bulgarian Language Prof Lyubomir Andreychin, Bulgarije; Faculdade de Ciencias Sociais e Humanas da Universidade Nova de Lisboa, Portugal; K Dictionaries Ltd, Israël; Consiglio Nazionale delle Richeche, Italië; Det Danske Sprog- og Litteraturselskab, Denemarken; Universiteit Kopenhagen, Denemarken; Eesti Keele Instituut, Estland; Universiteit Trier, Duitsland en Real Academia Española, Spanje.

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.

Investering 2018: €77.916 1 fte (extern projectgeld via Horizon2020 R&I)

I.2. TermNeXT aanvraag (Horizon 2020 Marie Curie ITN network)

Projectleiders: Frieda Steurs en Kris Heylen

Dit project werd reeds op 10 januari 2017 ingediend, maar niet gehonoreerd. De algemene evaluatie was echter zeer positief, vandaar wordt het opnieuw ingediend op 10 januari 2018 met de nodige verbeteringen. Hoofdaanvrager is KU Leuven (QLVL); het INT is een partner die stageplaatsen zal aanbieden aan onderzoekers in dit project.

TermNeXT stelt een geïntegreerd, interdisciplinair en marktgericht opleidingsprogramma voor onderzoek in terminologie en kennisbeheer. Een aantal onderzoekers zullen werken aan innovatief, oplossingsgericht onderzoek dat de terminologische uitdagingen in ons aanpakt. Het consortium brengt samen partners uit academische, industriële en non-profitorganisaties die samenwerken om 4 specifieke onderzoeksdoelen te realiseren.
TermNeXT heeft zich 4 specifieke onderzoeksdoelstellingen opgezet:

1. Toegevoegde waarde voor de kenniseconomie: hoe kunnen we de economische meerwaarde van het terminologisch beheer kwantificeren en maximaliseren voor bedrijven en organisaties.
2. Kennisoverdracht mogelijk maken: Welke procedures en software-infrastructuren zijn nodig om kennis in te zetten? Uitwisseling tussen verschillende terminologische bronnen en systemen. TermNeXT ontwikkelt interoperabiliteitsprocedures voor dataformaten.
3. Kenniskwaliteit verzekeren: hoe kan terminologie management helpen om de kwaliteit van kennisdatabanken en kennisuitwisseling te verzekeren?

4. De kenniskloof overbruggen: hoe kan het beheer van terminologie helpen om de kenniskloof tussen hoogopgeleide en minder opgeleide burgers in onze globaliserende kenniseconomie te verkleinen?

I.3. Inzicht in het mentale lexicon: Hoe worden woorden in het hoofd opgeslagen? (KIEM project NWO)

Projectaanvragers: Nicoline Van der Sijs (Meertens & RU), medeaanvrager Frieda Steurs

Er wordt momenteel vanuit diverse disciplines onderzoek gedaan naar de manier waarop woorden in het hoofd worden opgeslagen. Hiervoor worden verschillende methodes gebruikt, zoals ratings, reactietijden, woordassociaties, corpusonderzoek, semantische vectoren, distributionele modellen. Veel van dat onderzoek levert nieuwe datasets op , bestaande uit woorden die zijn verrijkt met o.a. informatie over frequentie, associatie, gebruik en bekendheid.

Die datasets staan momenteel allemaal los van elkaar en ook los van de traditionele woordenboeken. Het aan elkaar koppelen van de verschillende datasets zou een enorme toegevoegde waarde opleveren voor zowel onderzoek als praktische toepassingen. Dat was de conclusie van een groep Nederlandse en Vlaamse taalkundigen, letterkundigen, historici, psycholinguïsten en taaltechnologen, die in juni een week lang bijeen zijn gekomen op het Leidse Lorentz Center. Deze Kiem-aanvraag is de eerste vervolgstap van die workshop.

Voor het linken en doorzoekbaar maken van de verschillende verrijkte datasets is een goed doordachte infrastructuur nodig. Het doel van deze projectaanvraag is drieledig:

1. een eerste infrastructuur te bouwen waarop bestaande databestanden via shallow linking aan elkaar gekoppeld worden als pilotstudie,
2. een rapport op te leveren dat kan dienen voor een grotere subsidieaanvraag bij de creatieve industrie of NWO (middel)groot, met een overzicht van de nieuwe wetenschappelijke inzichten en praktische toepassingen die mogelijk worden op basis van de uitgewerkte infrastructuur,
3. een blauwdruk voor een praktische toepassing van de infrastructuur als showcase, en wel hoe de infrastructuur kan worden ingezet voor de ontwikkeling van digitale leermiddelen voor taalverwerving.

Looptijd 1 jaar: van 1 januari 2018 tot 31 december 2018

Indien goedgekeurd: 7500 euro voor het INT (extern projectgeld)

1.4. ENETCOLLECT : European Network for Combining Language Learning with Crowdsourcing Techniques

Verantwoordelijken: Frieda Steurs en Tanneke Schoonheim

Dit nieuwe project binnen het COST-netwerk loopt van maart 2017 tot maart 2021 en wordt geleid door EURAC, Bolzano. Frieda Steurs is lid van het managementcomité voor Nederland en is vicevoorzitter van werkpakket 3.
De enetCollect Action richt zich op de grote Europese uitdaging om de taalvaardigheid van alle burgers te bevorderen, ongeacht hun sociale, educatieve en taalkundige achtergronden. Het project streeft naar het verbeteren van de productie van leermateriaal om de toenemende vraag naar taalonderwijs en de diversificatie van leerprofielen te kunnen aanpakken. Het gaat over het basiswerk uit te voeren om een onderzoeks- en innovatietrend in gang te brengen die het gevestigde domein van Language Learning combineert met recente en succesvolle crowdsourcingbenaderingen om een crowdsourcingpotentieel beschikbaar te maken voor alle talen en een innovatie doorbraak voor de Productie van taalleermateriaal. Voor het INT is dit project belangrijk omdat we ervaring hebben met corpusopbouw en voor bepaalde projecten ook aan crowdsourding doen, en anderzijds omdat we ons meer en meer willen richten op het beschikbaar stellen van digitale taalmaterialen voor computer ondersteund talenleren.

Via dit netwerk worden alle reis- en netwerkkosten betaald, en zijn stages mogelijk.

J. Doelgroepenbeleid (inclusief onderwijs)

Het INT wil veel meer diversifiëren in de doelgroepen die kunnen worden bereikt.
Door verschillende activiteiten kunnen we een ruim publiek meer informatie en ondersteuning geven wat betreft taalmaterialen.

J.1. Onderzoekers, wetenschappers

Onderzoekers en wetenschappers kunnen worden geholpen via workshops op maat, internationale uitwisseling, STSM's binnen het COST-netwerk, gastcolleges aan universiteiten etc. In het verleden heeft het INL veel colleges gegeven op het gebied van de (computationele) lexicologie en de semantiek. We onderzoeken of dit nu weer tot de mogelijkheden behoort. Belangrijk is om de veranderende rol van onlinewoordenboeken en lexica te benadrukken en de gebruikers (wetenschappers, studenten en docenten) hierin meer te betrekken. Niet alleen colleges op dit gebied zijn wenselijk, ook gebruikersvideo's en workshops toegespitst op praktisch gebruik van onze onderzoeksmaterialen (zoals in 2017 in Antwerpen en Leuven).

J.2. Studenten en docenten

We bieden stages aan voor studenten in de taalkunde en computertaalkunde en we kunnen gerichte workshops organiseren voor deze doelgroep. Een communicatiestage zou ook zeer wenselijk zijn, vooral op het gebied van social media. Er zou bijvoorbeeld een student aangetrokken kunnen worden die zich richt op het opzetten van een Facebookpagina voor het instituut (zie ook K.1.) en alle sociale media rondom de verkiezing Weg met dat Woord!.

J.3. Algemeen publiek

Via activiteiten in de Week van het Nederlands, het Drongo Talenfestival, het promoten van het ANW-spel etc. bereiken we een groter publiek. In het kader van Weg met dat woord! (WMDW) maar ook in het algemeen is het nuttig om te onderzoeken of we ons meer kunnen richten op communicatie in videovorm. In 2016 zijn voor de Week van het Nederlands een

aantal video's opgenomen over bekende woorden en hun betekenissen en die zijn heel goed bekeken. Het past ook in deze tijd om meer met video en beeld in het algemeen te werken, vooral om een jongere doelgroep aan te trekken. In dit kader is ook de MOOC (Massive Open Online Course), een onlinecursus waarvoor iedereen zich mag inschrijven, is een interessante manier om een groot publiek te bereiken via videocolleges.

Weg met dat woord! wordt in 2017 voor het vijfde jaar georganiseerd en er wordt tijdens de Week van het Nederlands een speciale lezingenmiddag aan deze editie gewijd. In 2018 staat een populairwetenschappelijke publicatie gepland over vijf jaar WMDW. In de komende jaren moet bekeken worden of we met de verkiezing blijven doorgaan, en in welke vorm.

K. Wetenschapscommunicatie

Projectleiders: Vivien Waszink en Laura van Eerten
In samenwerking met J.J. Knol en W. van Severen

K.1. Twitter; Facebook

Het instituut gebruikt met name Twitter als socialmediakanaal voor het delen van nieuws, evenementen en kennis over taal, en zal dat de komende jaren blijven doen. Daarnaast moet er een algemene Facebookpagina voor het INT komen; nu is er alleen een pagina over WMDW die na oprichting in 2016 succesvol bleek. Een INT-Facebookpagina was ook een van de dingen die naar voren kwam uit het onderzoek van communicatiestudente Miranda Slootweg van de Hogeschool Den Haag. Op Facebook zouden de rubrieken gedeeld kunnen worden die nu ook op Twitter gedeeld worden, en bovendien zou er op Facebook ook meer aandacht kunnen worden besteed aan video.

K.2. Woordbaak; Terug in de Taal; Neologisme van de Week

De aandachttrekkers op de website zijn de populairwetenschappelijke rubrieken Woordbaak, Terug in de Taal en Neologisme van de Week. Ze worden goed gelezen, vaak gedeeld op de sociale media en overgenomen in taalnieuwsbrieven van de Taalunie en Onze Taal. Deze rubrieken zullen ook voor de aankomende jaren verzorgd, vernieuwd en uitgebreid worden.

K.3. Populairwetenschappelijke uitgaven

In 2017 komt het derde boek over alledaagse etymologie in de reeks ‘Waar komt *pindakaas* vandaan?’ uit, met als titel ‘Waar komt *suikerspin* vandaan?’. Ook dit boek is gemaakt door een aantal INT-medewerkers in samenwerking met Onze Taal. De boeken worden keer op keer zeer goed verkocht en de verwachting is dat de boeken om de twee jaar zullen blijven uitkomen. In 2018 staat een populairwetenschappelijke publicatie gepland over vijf jaar WMDW.

Investering 2018 €104.581 1,4 fte