/instituut voor de
Nederlandse taal/

# Beleidsplan 2022

november 2021

# Inhoudsopgave

## Inhoud

# 0. Algemene doelstellingen

Na de hervormingen van 2016 is het Instituut voor de Nederlandse Taal (INT) voluit aan de slag gegaan met het uitvoeren van al zijn (kern)taken. In 2017 werd een meerjarenbeleidsplan voor de periode 2018-2022 geschreven en goedgekeurd. In 2018 werd meteen gestart met het uitvoeren van deze doelstellingen, en in 2019 en 2020 werden de grote lijnen van het meerjarenbeleidsplan verder uitgewerkt, met een aantal nieuwe accenten en dit in nauw overleg met de Nederlandse Taalunie. Dit vertaalt zich ook in een heldere planning van projecten en in een begroting die deze taken duidelijk laat zien.

De missie van het INT staat daarbij centraal: het INT neemt een centrale positie in voor het hele Nederlandse taalgebied (Nederland, 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 een 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 digitale taalinfrastructuur. 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 gebruik maken van deze infrastructuur. Het instituut werkt dan ook hard aan het meer zichtbaar maken van al onze taalmaterialen. Zo worden er meer podcasts en webinars gemaakt. Tegelijk vinden meer en meer stagiairs de weg naar het INT en dit zorgt voor extra wisselwerking met de universitaire opleidingen in Nederland en Vlaanderen.

Het beleidsplan 2022 sluit aan op de hoofdlijnen van het meerjarenbeleidsplan van de NTU (2020 tot 2024). De aandachtsgebieden die de NTU benoemt, met name: 1. Standaardtaal, 2. Nederlands, taalvariëteiten en andere talen, 3. Onderwijs Nederlands binnen het taalgebied, 4. Onderwijs Nederlands buiten het taalgebied en 5. Taal en Cultuur komen terug in onze werkzaamheden. Dat vertaalt zich in diverse structurele werkzaamheden die hieronder geformuleerd staan, maar wordt ook zichtbaar in de diverse projecten met externe partijen.

Voor de uitvoeringen van de vele taken die het INT heeft, wordt gestreefd naar zoveel mogelijk samenhang en synergie in de planning van die taken en de uitvoering ervan. Dit wordt uiteengezet in

de sectie over de gedeelde infrastructuur van het INT en is ook terug te zien de daaropvolgende beschrijving van diverse taken van het INT die voor 2022 gepland staan.

Een belangrijke taak voor 2022 is het formuleren van een nieuw meerjarenbeleidsplan voor de periode 2023-2028. Dat beleidsplan zal rekening houden met het rapport van de internationale visitatiecommissie, die in 2021 op bezoek kwam om de werking van de afgelopen vijf jaar te evalueren en aanbevelingen te doen voor de toekomst. Voor de planning van de werkzaamheden van het komende beleidsjaar is in de mate van het mogelijke reeds rekening gehouden met de aanbevelingen van de visitatiecommissie. Met name vindt de visitatiecommissie onze benadering van het uitbouwen van de taalinfrastructuur zeer innovatief en moedigt ons aan om op deze weg verder te gaan. Het lexicografische werk dat op deze infrastructuur voortbouwt, heeft zich met succes gediversifieerd in kleinere projecten die op verschillende manieren aan elkaar gekoppeld zijn. Door deze nieuwe benadering en modulaire aanpak kunnen hedendaagse, historische, dialectologische, morfologische, syntactische en semantische informatie enz. voor een bepaald woord bijeengebracht worden in verschillende maar nauw samenwerkende projecten en kan deze informatie afhankelijk van de doelgroep op diverse manieren online aangeboden worden. Op deze manier kan het INT zijn opdracht vervullen en moderne wetenschappelijke lexicografische producten blijven aanbieden. Daarbij moet steeds een evenwicht worden gevonden tussen de beoogde en gewenste werkzaamheden en de beperkte capaciteit van het instituut.

De prioriteit voor 2022 ligt duidelijk op de verdere uitwerking van de lexicografische infrastructuur voor het hedendaags Nederlands, rekening houdend met de ideeën uit de white paper The Future of Academic Lexicography -- A White Paper (Steurs et. al. (eds.), 2020). In 2021 zijn reeds resultaten geboekt ten aanzien van het Corpus Hedendaags Nederlands (CHN), de workflow voor morfosyntactische informatie in het *Algemeen Nederlands Woordenboek* (ANW) en in Woordcombinaties en de koppeling van producten als het Referentiebestand Nederlands (RBN) en diverse vertaalwoordenboeken aan het centrale lexicon GiGaNT. Voor 2022 voorzien we verder onderzoek naar de optimalisatie van deze infrastructuur en naar de wijze waarop de lexicografische beschrijving van het Nederlands in het ANW in deze infrastructuur ingebed zal worden.

Er is in de afgelopen jaren meer en meer aandacht gekomen voor de disseminatie, en onze projectleider communicatie heeft zich samen met de webredactie erg ingezet op de verhoging van de zichtbaarheid van het instituut bij verschillende doelgroepen. Dit werk werd door de visitatiecommissie erg op prijs gesteld en we gaan in 2022 in deze richting verder.

Voor 2022 staat in Leiden het grote evenement “Leiden City of Science 2022” gepland. Het INT zal zeer actief meewerken aan dit grote programma en een aantal activiteiten organiseren. Deze moeten nog verder met de organisatoren worden afgesproken.

Het INT richt zich als toegepast wetenschappelijk instituut traditioneel op onderzoekers en taalkundigen. Bestaande contacten met onderzoekers uit binnen- en buitenland, verbonden aan wetenschappelijke instituten en universiteiten, worden al dan niet in samenwerkingsprojecten onderhouden en waar mogelijk geïntensiveerd en uitgebreid. Voor universitaire studenten verzorgt het INT twee verschillende collegereeksen over computationele lexicografie, waarbij ook masterproeven worden begeleid.

Daarnaast heeft het INT zijn werkterrein, gezien de brede taakomschrijving, nadrukkelijk naar docenten en leerlingen in het voortgezet/secundair onderwijs verschoven. In dat verband is het INT aanwezig op en profileert het zich op beurzen, conferenties (HSN-conferentie) en evenementen (Neerlandistiekdagen). Het INT wil verder de banden met het Onderwijsnetwerk Zuid-Holland en Alphalab Leiden aanhalen, en is een samenwerking aangegaan met de Taalkunde Olympiade. Daarnaast is er een posterpakket gemaakt voor in het klaslokaal dat docenten tegen productiekosten kunnen bestellen bij het INT.

De taalmaterialen van het INT zullen voor zover mogelijk beter toegankelijk worden gemaakt voor het secundair en het tertiair onderwijs. Op de website heeft onderwijs met een eigen menu-item een vaste plaats gekregen. De daar te vinden beschikbare informatie en materialen zoals lesbrieven worden bijgehouden en geregeld uitgebreid.

Ook het algemene publiek wordt niet uit het oog verloren. Op onze website verschijnt elke dag van de week een populairwetenschappelijke rubriek over woorden, zoals ‘Nieuw woord van de week’ (neologismen), ‘Terug in de taal’ (historische woorden), in 2021 uitgebreid met ‘WoordHoek’ (column Ewoud Sanders), ‘Woorden weten alles’ (column Ludo Permentier), ‘Uit de streek’ (dialectwoorden) en ‘Uitgelichte term’ (een keer per maand). Minimaal zes keer per jaar wordt een algemene nieuwsbrief verstuurd aan geïnteresseerden. Daarnaast is er een nieuwsbrief terminologie die vier keer per jaar verschijnt en die informatie geeft over vaktaal. Ook worden er regelmatig publieksevenementen georganiseerd, live en digitaal. Het INT levert jaarlijks een bijdrage aan de Week van het Nederlands in oktober, en vanaf 2022 komt daar het nieuwe festival Letterlijk Leiden bij. Naast de podcast ‘Waar komt pindakaas vandaan?’ over etymologie is het INT in 2021 een nieuwe podcast gaan produceren samen met Onze Taal: ‘Over taal gesproken’. Deze podcast verschijnt elke maand en snijdt taalkundige onderwerpen aan die van belang zijn voor het INT en Onze Taal, en toegespitst zijn op een breed publiek. Zoals het Nederlands in België en Nederland en grammatica.

Medewerkers houden regelmatig lezingen, zijn te horen in radioprogramma's en schrijven boeken en artikelen voor een algemeen publiek dat belangstelling heeft voor taal in het algemeen en Nederlands in het bijzonder.

Met webinars, livestreams van evenementen en berichten op de sociale media Instagram, Facebook, LinkedIn en Twitter brengen we ook online voortdurend (de werkzaamheden van) het instituut bij alle doelgroepen onder de aandacht.

De overgrote meerderheid van hierboven genoemde taken worden gefinancierd vanuit de lumpsum die ons via de Taalunie door het Comité van Ministers ter beschikking wordt gesteld. Het gaat hier met name over:

1) Hedendaags Nederlands: *Algemeen Nederlands Woordenboek* (ANW), Neologismen en *Woordenboek van Nieuwe Woorden* (WNW), Hedendaagse corpora, Woordcombinaties (senior: Katrien Depuydt; projectleiders Rob Tempelaars, Vivien Waszink en Lut Colman) € 463.389,-
2) Terminologie (senior: Vincent Vandeghinste; projectleider Dirk Kinable) € 123.570,-
3) Spelling (senior: Katrien Depuydt; projectleider Katrien Van pellicom) € 123.570,-
4) Historische taalmaterialen (senior Katrien Depuydt; projectleider Roland de Bonth) € 308.926,-
5) Dialectologie (senior Nicoline van der Sijs; projectleider Veronique De Tier) € 123.570,-
6) Taalmaterialen (senior Vincent Vandeghinste; projectleider Bob Boelhouwer) € 185.356,-
7) Grammatica (ANS, Taalportaal) (senior Frieda Steurs; projectleider Frank Landsbergen) € 123.570,-
8) Clarin B-centrum/Clarin K-centrum/Clarin.be (senior: Vincent Vandeghinste) € 61.785,-
9) IT en systeembeheer (senior Jesse de Does) € 308.926,-
10) Communicatie, website en evenementen (senior Frieda Steurs; projectleider Laura van Eerten) € 154.463,-
11) Algemene dienst, secretariaat, financiële dienst, directie (lonen van directeur, administratieve ondersteuning, financiële dienst (senior Frieda Steurs) € 185.356,-
12) Algemene kosten (huisvesting, verzekeringen, overhead, andere vaste kosten) € 371.675,-

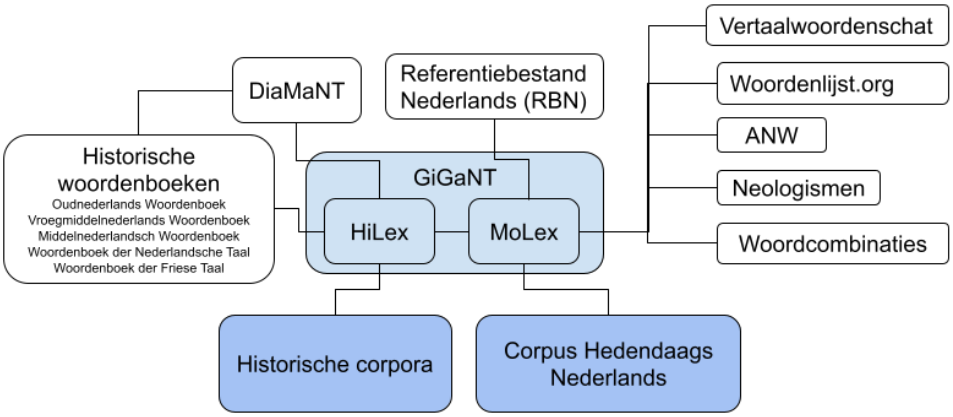
Het betreft 35 medewerkers (= equivalent van 28 fulltime medewerkers).

Meer gedetailleerde informatie kan worden gevonden in de begroting 2022.

# 1. Gedeelde infrastructuur

Figuur 1 schetst de globale gedeelde taalinfrastructuur bij het INT. Het Corpus Hedendaags Nederlands (CHN) is het centrale corpus dat voor de verschillende toepassingen m.b.t. hedendaags Nederlands als bronmateriaal dienst doet. Daarnaast heeft het INT diverse historische corpora, waarvan enkele ten grondslag hebben gelegen aan een aantal historische woordenboeken van het INT.

De kern van de lexicale infrastructuur is het computationeel lexicon GiGaNT,[1] bestaande uit Molex voor het hedendaags Nederlands en Hilex voor het historisch Nederlands, waaraan verschillende lexicale databases gekoppeld worden. De verbeteringen en uitbreidingen die de lexicografen uitvoeren in GiGaNT dringen zo door naar alle specifieke lexica en woordenboeken die door het INT gecreëerd worden. Door de koppelingen vloeien ook verbeteringen terug naar de centrale database.




*Figuur 1. Globale taalinfrastructuur bij INT*

Voor het ontwikkelen en exploiteren van de taaldata is ondersteuning door computationeel linguïstische en andere softwaretools essentieel. De computationele infrastructuur van het INT bestaat uit een aantal componenten:

- het corpuszoeksysteem BlackLab en het bijbehorende userinterface (corpus-frontend) voor ontsluiting van corpora
- de pijplijn DUCT voor bestandsconversie, taalkundige verrijking en indexering van corpusmateriaal

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

- het Rapid Database Application Development platform Lex'it voor bewerking van lexicale (en andere gestructureerde) data, in gebruik voor onder andere Hilex, Molex, DSDD en vele andere projecten
- de woordenboekeditor SwingLex (INL-DWS), met specialisaties voor ANW, Neologismen en Woordcombinaties
- de API van de lexiconservice waarmee het GiGaNT-lexicon toegankelijk gemaakt wordt
- generieke componenten voor het publiceren van woordenboeken, onder meer toegepast bij de publicatie van woordcombinaties en het Lexicon Frisicum

Daarnaast worden specialisaties ontwikkeld binnen specifieke projecten en voor specifieke doeleinden. Daarbij wordt er steeds naar gestreefd de ontwikkeling van de core-infrastructuur verder te zetten.

# 2. Hedendaags Nederlands

## 2.1. Infrastructureel

### 2.1.1. GiGaNT-Molex

De moderne lexiconcomponent is de centrale ruggengraat voor de lexicale beschrijving van het modern Nederlands, en wordt voortdurend uitgebreid met nieuwe data, niet alleen in de zin van toegevoegd vocabulaire, maar ook door uitbreiding van de features, zoals in 2020-2021 met toegevoegde werkwoordkenmerken die het project Woordcombinaties ten goede komen en het toevoegen van uitspraakinformatie (zie hiervoor het onderdeel spelling). Het lexicon is beschikbaar als dataset onder een niet-commerciële en commerciële licentie, en via de lexiconservice.
In 2022 zal met name gewerkt worden aan een nieuwe versie van de Spelling-API (zie bij het onderdeel spelling), aan uitbreiding van de structuur voor betere integratie van meerwoordscombinaties, met name vanuit het project Woordcombinaties en vanuit terminologie, en aan het aanpassen van de woordsoortinformatie aan de Tagset Diachroon Nederlands. De werkzaamheden aan de koppeling tussen Hilex en Molex worden afgerond, en de koppeling met de Vertaalwoordenschat wordt uitgebreid met nieuwe datasets (zie daarvoor onder Vertaalwoordenschat).

### 2.1.2. Corpus Hedendaags Nederlands

De moderne lexiconbouw en de beschrijving van het hedendaags Nederlands wordt gebaseerd op modern corpusmateriaal. De corpusdata voor het hedendaags Nederlands zitten in het Corpus Hedendaags Nederlands (CHN). De werkzaamheden bestaan enerzijds uit acquisitie, dataprocessing en taalkundige verrijking, en anderzijds uit het beschikbaar stellen van het materiaal in een corpusapplicatie voor intern en extern gebruik. Het materiaal is afkomstig uit Nederland, Vlaanderen, Suriname en de Antillen.

In 2021 is  een nieuwe versie van het corpus in de laatste versie van de corpusapplicatie gepubliceerd, waarbij ook het mechanisme voor wekelijkse updates van het interne corpus en maandelijks updates voor het externe corpus in productie is genomen, zodat gebruikers steeds in de gelegenheid zijn het meest recente Nederlands te onderzoeken.

**Data**

In 2020-2021 is het corpus flink uitgebreid met nieuwe bronnen, zoals bijvoorbeeld Vlaams krantenmateriaal en forumdata uit Nederland en Vlaanderen en de Antillen. Parlementair materiaal (Belgisch Federaal Parlement) is gestructureerd en verrijkt in het ParlaMint-project voor de ontwikkeling van vergelijkbare parlementaire corpora binnen Europa. In het kader van dit project is ook onderzoek gedaan naar syntactische verrijking volgens de universal dependencies guidelines. Dit materiaal zal in 2022 worden uitgebreid met materiaal van het Vlaams Parlement. Voor wat betreft de acquisitie zal de focus liggen op uitbreiding met krantenmateriaal uit Nederland en op de opvulling van gaten in de chronologie, zodat we werken richting een diachroon monitorcorpus.

**Workflow en verrijking**

Verder werken we aan de verbetering en uitbreiding van de taalkundige verrijking van het corpus. In 2022 integreren we syntactische verrijking in de corpusworkflow. Deze verrijking zal onder meer bestaan uit universal dependencies,, eventueel uitgebreid met meer specifieke informatie die nuttig is voor projecten binnen het instituut (zoals Woordcombinaties).

**Corpusapplicatie**

Om efficiënter te kunnen omgaan met de steeds groeiende hoeveelheid materiaal, zullen optimalisaties worden geïmplementeerd en zal worden gewerkt aan het mogelijk maken van een gedistribueerde opzet door de applicatie in Apache Solr te integreren. Dit is noodzakelijk om het CHN efficiënter te kunnen inzetten in de workflow voor de beschrijving van neologismen. Verder wordt binnen het CLARIAH-PLUS project (zie aldaar) gewerkt aan een uitbreiding van de zoekengine voor syntactische verrijking en voor parallelle corpora.

## 2.2. Lexicale beschrijving van het hedendaags Nederlands

### 2.2.1. Onderzoek naar nieuwe wegen

In 2022 wordt volop het onderzoek ingezet naar nieuwe wegen om de lexicografische beschrijving van het modern Nederlands tot stand te brengen. We analyseren voor *alle* componenten die deel uitmaken van ANW, Neologismen, Woordcombinaties en terminologie hoe deze het beste ingevuld kunnen worden. Hierbij moet ook bekeken worden hoeveel capaciteit nodig is om dit te realiseren. Bij ieder

aspect van de beschrijving analyseren we de huidige dekking en de dekking waarnaar we streven, waarbij we rekening houden met de diverse doelgroepen (eindgebruikers, onderwijs, computationele toepassingen, taalprofessionals, ….). Per component kijken we welke bewerkingsmethodes, NLP-technologieën en databronnen we het beste kunnen inzetten.
In de afgelopen jaren is deze aanpak al gerealiseerd voor de morfosyntactische component. In 2022 focussen we op 1. Verwerving van nieuw vocabulaire (neologismenworkflow); 2. de semantische component; 3. verbeterde ondersteuning van het bewerkingsproces voor Woordcombinaties en 4. Een update van de terminologie-extractietools zodat deze weer in lijn zijn met de huidige state-of-the-art.

### 2.2.2. Algemene woordenschat

De beschrijving van de algemene woordenschat gebeurt op drie verschillende terreinen: de spelling van woorden, de betekenis en de constructies waarin ze voorkomen. De informatie wordt beschikbaar gesteld via respectievelijk *woordenlijst.org, anw.ivdnt.org* en *woordcombinaties.ivdnt.org.* In 2021 is daaraan nog *neologismen.ivdnt.org* toegevoegd.

#### Spelling

In 2022 wordt de in 2020-21 ingezette lijn om bij de updates voor woordenlijst.org in te zetten op kwaliteit boven kwantiteit voortgezet. Hierbij is nu naast potentiële spellingproblemen ook corpusfrequentie in recent materiaal een belangrijk criterium. Het spellingbestand is in 2021 uitgebreid met informatie uit de bestanden van de werkgroep Buitenlandse Aardrijkskundige Namen (BAN) van de Commissie Anderstalige Namen. Daarnaast is gestart met het toevoegen van uitspraakinformatie aan de online woordenlijst. Hiervoor zijn in 2021 de basisprincipes vastgelegd. De data-werkzaamheden daarvoor worden voortgezet in 2022.

Daarnaast ronden we de nieuwe publieke zoekmachine voor de lexicale data (Spelling-API en woordenlijst.org) af. De nieuwe versie maakt het onder andere eenvoudiger nieuwe functionaliteit toe te voegen: zoeken op meerdere kenmerken, flexibeler omgaan met de presentatie van resultaten, relevantere suggesties bij het zoeken, en het gebruik van uitspraakinformatie.

#### Semantische beschrijving

Het *Algemeen Nederlands Woordenboek* (ANW) en het neologismenproject staan in voor de semantische beschrijving van het hedendaags Nederlands. De beide projecten maken gebruik van het CHN, en de beschrijving van de woorden wordt opgeslagen in een gemeenschappelijke databank. De semantische beschrijving wordt in 2022 voortgezet. Er vindt onderzoek plaats naar een betere workflow voor de detectie en beschrijving van neologismen, en naar hoe we de semantische beschrijving zo kunnen organiseren dat dit leidt tot een publiek beschikbare betekenisinventaris voor

het Nederlands, waarbij optimaal gebruik gemaakt wordt van reeds beschikbare vrije data. Tot slot zal er worden geïnvesteerd in kennisoverdracht in de vorm van opleiding van jonge lexicografen via cursussen en stages.

### Nederlands als vreemde taal

***Woordcombinaties***

Woordcombinaties is de online taaltool die leerders van het Nederlands als vreemde taal ondersteunt bij het gebruiken van woorden in context. Dit project inventariseert en beschrijft systematisch combinaties (collocaties, idiomen en patronen) in het Nederlands.

In 2022 zullen meer patronen beschreven worden en werkwoorden, geselecteerd op basis van hun frequentie, voorzien worden van voorbeeldzinnen en combinaties. In 2022 zal ook gewerkt worden aan een betere bewerkingsomgeving voor de idiomen en er zullen voorbereidende taken uitgevoerd worden voor de beschrijving van de substantieven.

***Oefenen.nl***

*Oefenen.nl* wil een app *EenvoudigNL* realiseren voor woordenschat en zinnen voor anderstaligen. De inzet is om te werken vanuit een corpus 'Eenvoudige taal' zodat mensen onbeperkt kunnen blijven oefenen met woorden en zinnen. De app voorziet in oefeningen op taalniveau A2 van het Raamwerk NT2 en in een plusversie waarin gewerkt kan worden aan woordenschat richting taalniveau B1. De rol van het INT is om de infrastructuur te bouwen die data aan de app kan leveren. Er is in 2021 gestart met een onderzoek naar hoe die infrastructuur eruit moet zien. Er zal een proof of concept worden gemaakt waarin twee thema's worden uitgewerkt. Deze werkzaamheden worden in 2022 voortgezet. De opschaling van e.e.a., het verder automatiseren en optimaliseren van de workflow en het in productie nemen is een substantieel project, waarvoor extra financiering gezocht zal worden.

## 2.3. Terminologie

De nieuwe website van het INT heeft een performante zoekmachine om snel en efficiënt alles over het Expertisecentrum Terminologie (ENT) te kunnen vinden. Het ENT wordt gestaag uitgebreid met nieuwe gegevens.

### Termenlijsten

Termenlijsten documenteren de mate waarin een taal zich voorbij de gangbare dagelijkse communicatiebehoeften in meer specialistische domeinen blijft ontwikkelen. Ze vormen ook een handige vraagbaak voor de vele taalgerichte beroepsbeoefenaars zoals vertalers en technische schrijvers. We blijven daarom permanent de hedendaagse termenlijsten op de ENT-webpagina's

updaten en uitbreiden. Analoog hiermee wordt in de lijn van het INT en zijn historische woordenboeken en corpora ook een luik voor lijsten met historische termen ingericht. Ook hiervoor is de internationale en grondige Library of Congress-classificatie goed bruikbaar.

### Tools

Wat de terminologietools betreft gaan we in 2022 onderzoek doen ten behoeve van de ontwikkeling van een nieuwe versie van Termtreffer. Deze versie zal door het INT zelf volledig worden ontwikkeld, rekening houdend met de vereisten van de gebruikers. Ook het ENT-advies wordt meegenomen.

### Veldondersteuning

Klassieke veldondersteuning in de vorm van verdere updates voor de bestaande websiterubrieken, vooral de opleidingen in Nederland en Vlaanderen, blijft aan de orde. Ook de evenementenrubriek voor terminologie dient continu te worden bijgehouden en er worden jaarlijks 4 uitgebreide nieuwsbrieven Terminologie gedistribueerd.

Nu het pilotproject over hogeronderwijstermen in de HOTNeV-termbank is afgerond, wordt dit voortgezet via stages en scripties. De groei van deze databank is uiteraard afhankelijk van de stages die bij het ENT worden aangevraagd en de begeleiding daarvan. Begeleiding van studenten blijft belangrijk en ons stageaanbod voor terminologiewerk wordt blijvend gepromoot, het is ook belangrijk voor de netwerkpositie van het ENT. Afstudeerscripties van studenten kunnen eveneens bijdragen aan de verdere invulling van de HOTNeV-termbank. Verder willen we dit project koppelen aan internationale samenwerking. We hebben kennis genomen van twee gelijkaardige projecten, met name bij EURAC (Bolzano) en bij de Universitat Autònoma de Barcelona. In 2022 willen we via deze contacten het project verder uitbreiden.

### *Speerpunt 1: de medische vaktaal*

Een eerste versie van het Pinkhof geneeskundig woordenboek werd in 2021 online gezet via een applicatie die bij het INT werd ontwikkeld. Dit woordenboek wordt bijgewerkt met als primaire gebruiksfunctie: verklarend hedendaags medisch woordenboek en een taalbaak voor medisch Nederlands te vormen. Om dit te realiseren wordt er samengewerkt met de Stichting Beheer Pinkhof-database. Er werd ook een raad van advies opgericht om afgeleide medische terminologieprojecten te begeleiden. Het verder optimaliseren van het Pinkhofbestand en het uitwerken van nieuwe deelprojecten wordt een taak voor 2022.

### *Speerpunt 2: de juridische vaktaal*

Wat betreft het juridisch woordenboek van M.C. Oosterveld-Egas Reparaz en Johanna Vuyk-Bosdriesz wordt verder gewerkt aan de updating en uitbreiding van het bestand. Het juridische woordenboek van Oosterveld Ragaz (Nederlands Recht) wordt in 2022 online beschikbaar via een nieuwe applicatie gebouwd bij het INT.

In samenspraak met dr. Karl Hendrikx (UA en KU Leuven) wordt in 2022 een studie gemaakt om de Belgische equivalenten aan het juridische woordenboek toe te voegen. Deze studie zal worden uitgevoerd op basis van bestaande lijsten, maar met een vergelijkend onderzoek naar equivalenten en semi-equivalenten.

### *Speerpunt 3: Nederlands als wetenschapstaal*

Er zijn heel wat initiatieven die de aandacht vestigen op de noodzaak aan talige hulpmiddelen om voor studenten de overstap van het middelbaar naar het hoger onderwijs makkelijker te maken. We werken in dit verband samen met het Proefproject Nederlands als wetenschapstaal - van corpora naar terminologielijsten. Dit project is een samenwerking tussen Stichting Nederlands / Vlaams Platform Taalbeleid Hoger Onderwijs, KU Leuven, UGent en het INT. De motivatie voor dit project is breed: het past bij "Nederlands in de aansluiting" tussen het middelbaar en hoger onderwijs, en het streven om de drempel voor het hoger onderwijs te verlagen. Tevens is het een ondersteuning van het gebruik van het Nederlands als wetenschapstaal. In 2021 werden door het INT ook aangepaste terminologielijsten ontwikkeld voor de vakken scheikunde en wiskunde (niveau BA1 NL en VL). Deze lijsten zijn via een speciaal ontwikkelde zoekapplicatie online gezet.

In 2022 gaan we op deze lijn verder met een termenlijst van natuurkunde (BA1) en volgen we verder de lijn van het Hoger Onderwijsplatform Vlaanderen-Nederland en het SaBeD project (Elke Peters) Daarbij wordt ook een databank gemaakt van "moeilijke woorden" op basis van examenblad Nederlands (eindexamen) en overig onderzoeksmateriaal.

## 2.4. Vertaalwoordenboeken

In september 2017 heeft het INT het online platform, de Vertaalwoordenschat, gelanceerd. Via dit platform worden de tweetalige bestanden, die in de afgelopen decennia, onder meer in opdracht van de Commissie Lexicologische Vertaalvoorzieningen (CLVV, 1993-2003), zijn ontwikkeld voor taalparen die op de commerciële markt niet spontaan aan bod kwamen, ontsloten.

Inmiddels staan het Nederlands-Nieuwgrieks / Nieuwgrieks-Nederlands, het Nederlands-Portugees / Portugees-Nederlands, het Nederlands-Estisch en het Nederlands-Fins / Fins-Nederlands online.

Om correcties en inhoudelijke updates in de toekomst op een gebruiksvriendelijke manier te kunnen realiseren is in 2021 gewerkt aan een nieuwe redactieomgeving voor de Vertaalwoordenschat. De werkzaamheden hieraan zullen begin 2022 worden afgerond, waarna correcties in de bestanden systematisch kunnen worden verwerkt. Hiervoor zal in eerste instantie gestart worden met het verwerken van de opmerkingen over het Nieuwgrieks-Nederlands/Nederlands-Nieuwgrieks die na het verschijnen van het papieren woordenboek in 2008 door de UvA zijn verzameld en die na het verschijnen van de onlineversie bij het INT zijn binnengekomen.

Daarnaast zal in 2022 de applicatie verder worden uitgebreid met nieuwe informatie en een nieuw taalpaar. In overleg met uitgeverijen zal gekeken worden naar de mogelijkheid om bepaalde woordenboeken online betaald aan te bieden. Ook zal verder gewerkt worden aan integratie van de Vertaalwoordenschatbestanden in de centrale infrastructuur.

## 2.5. Grammatica

Vanaf 2020 valt grammatica binnen de structurele basistaken van het INT. Dit betekent dat het INT zorgt voor de ontwikkeling, het beheer en de beschikbaarstelling van de verschillende digitale grammaticaproducten. Onder deze producten vallen vooralsnog het Taalportaal en de e-ANS. Voor de langere termijn zijn er ook plannen om de taaladviezen van Taaladvies.net aan deze producten te koppelen. Afgezien van de werkzaamheden aan de e-ANS en het Taalportaal (zie hieronder), zal gewerkt worden aan een functioneel en technisch ontwerp van een geïntegreerd grammaticaportaal, een webpagina die de spil moet gaan vormen van alle grammaticaonderdelen en zal fungeren als ontvangstpagina voor geïnteresseerde gebruikers, met informatie over projecten en producten, een zoekfunctie voor alle producten en een loket voor vragen.

### 2.5.1. e-ANS

In 2022 gaat de herziening van de e-ANS verder. Het werk aan deze herziening bestaat uit een aantal componenten: contact onderhouden met externe auteurs, werving nieuwe auteurs, redactie en eindredactie van nieuw herziene hoofdstukken. We organiseren door het jaar heen een aantal publicatiemomenten, waarbij we niet alleen de nieuwe hoofdstukken beschikbaar maken voor het publiek, maar daar ook aandacht aan schenken via diverse kanalen. Daarnaast zal ook de webapplicatie op enkele punten verder ontwikkeld worden. Naast de herziening van de ANS werken we in 2022 ook verder aan de didactische laag, in de vorm van een onderwijsmodule voor het in 2021 gepubliceerde hoofdstuk over klankleer.

De plannen voor de versnelling van 2022 zijn: de inhoudelijke herziening van hoofdstuk 2: Het werkwoord en hoofdstuk 4: Het lidwoord, en daarnaast de didactische laag van hoofdstuk 1, de klankleer.

### 2.5.2. Taalportaal

In 2022 zal de SoD-conversietool zijn geüpdatet, en zal hiermee het nieuwe syntaxis-deel 'Coordination' van Hans Broekhuis naar XML worden geconverteerd en gepubliceerd op Taalportaal.org. De auteursomgeving Oxygen zal in samenwerking met Zuid-Afrika worden geüpdatet, zodat deze toekomstbestendig is, en ook de webapplicatie zal worden bijgewerkt.

### 2.5.3. Grammaticaportaal

Ten slotte starten we in de loop van 2022 met de ontwikkeling van een interne testversie van het grammaticaportaal, die als startpagina moet gaan dienen voor de verschillende grammatica-applicaties binnen het INT: de e-ANS, het Taalportaal, Taaladvies.net en Woordcombinaties.

# 3. Historisch Nederlands

## 3.1. Infrastructureel

### 3.1.1. Historisch woordenboekenportaal

De beschrijving van de historische woordenschat is te vinden in de historische woordenboeken van het INT. Deze woordenboeken zijn online beschikbaar in het historische woordenboekenportaal (gtb.ivdnt.org). In die applicatie 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 hierna). Verbeteringen aan de woordenboekdata worden momenteel uitgevoerd in de context van GiGaNT-Hilex. Op termijn zullen deze terugvloeien naar onderliggende data in het woordenboekportaal.

De afronding van de conversiewerkzaamheden volgens de meeste recente versie (P5) van de TEI-coderingsrichtlijnen van de woordenboekdata moest in 2021 worden uitgesteld, maar wordt opgepakt in 2022. Daarnaast zal ook het werk aan de vertalingen van de betekenisomschrijvingen van het ONW in het Engels en Duits worden afgerond en aan het ONW-bestand worden toegevoegd.

### 3.1.2. GiGaNT-Hilex

GiGaNT is het centrale lexicon waaraan alle lexicale producten van het INT gekoppeld worden. Het historische onderdeel GiGaNT-Hilex is gebaseerd op de historische woordenboeken. Het huidige Hilex bevat het materiaal uit het *Woordenboek der Nederlandsche Taal* en het *Middelnederlandsch Woordenboek.* De koppeling met de online woordenboeken is behouden. Het lexicon is voor onderzoekers en applicatieontwikkelaars beschikbaar via de lexiconservice via maandelijkse updates.

Omdat binnen GiGaNT-Hilex met striktere lemmatiseringsprincipes wordt gewerkt dan de woordenboeken waarop het lexicon is gebaseerd, vinden er allerlei herschikkingen plaats van de lexiconinhoud. Deze werkzaamheden worden ook in 2022 voortgezet. Daarnaast zal verder gewerkt worden aan koppeling met de moderne component van het centrale lexicon.

In 2022 publiceren we de eerste formele TDN-conforme release van het GiGaNT-Hilex lexicon met de modules MNW-WNT, zowel als dataset als in de lexicon service.

### 3.1.3. Semantisch lexicon DiaMaNT

DiaMaNT bouwt een betekenislaag op GiGaNT-Hilex en heeft als doel een hulpmiddel te bieden bij tekstontsluiting en bij het onderzoek naar begrippen door de eeuwen heen. Ook voor DiaMaNT vormen de historische woordenboeken de belangrijkste bron. In de afgelopen jaren is gewerkt aan het opzetten van een infrastructuur voor de lexiconbouw op basis van woordenboek- en corpusdata. De in het lexicon opgenomen data zijn nu merendeels gebaseerd op het oplossen van de synoniemverwijzingen in de MNW- en WNT-data naar uit het WNT afkomstige lemmata en betekenissen. Het resultaat is, mede in het kader van CLARIAH, gepubliceerd als linked open data en gepubliceerd in een eerste versie van een applicatie voor het bredere publiek.

In 2022 zal verder gewerkt worden het expliciteren van betekenisrelaties ten behoeve van de semasiologische component en het fijnmaziger koppeling van synoniemen ten behoeve van de onomasiologische component. Daarnaast toetsen we, onder andere in samenwerking met het MacBERTh[2] project, nieuwe technologieën om met behulp van conventionele en contextafhankelijke wordembeddings semantische informatie uit corpora af te leiden.

### 3.1.4. Historische corpora

In 2020 is verder geïnvesteerd in optimalisering van het corpus frontend. Naast taalkundig verrijkt materiaal kan er ook niet-verrijkt corpusmateriaal beter doorzoekbaar worden gemaakt door inzet van de lexiconservice. Zes corpora zijn inmiddels online gezet. Mede in het kader van het CLARIAH+-project is een nieuw voorstel voor een flexibele tagset voor diachroon Nederlands uitgewerkt (“TDN”). In 2021 zijn de beschikbare historische corpora omgezet overeenkomstig TDN.
In 2022 zal, mede in het kader van CLARIAH-PLUS gewerkt worden aan een betere taalkundige verrijking van het historisch Nederlands. Daarvoor zal een grote hoeveelheid trainingsdata worden ontwikkeld, waarbij gebruik wordt gemaakt van een in het CLARIAH-PLUS project verbeterde versie van de door het INT ontwikkelde COBALT-tool voor de taalkundige annotatie met woordsoort en lemma.

[2] https://pdi-ssh.nl/en/2020/06/funded-projects-2020-call/

Het werk aan het corpus historische kranten dat door Nicoline van der Sijs in samenwerking met het Meertens Instituut en een groep vrijwilligers gedigitaliseerd is, zal worden afgerond. Verder zal het Gekaapte Brievencorpus worden gecureerd en gepubliceerd in de corpusapplicatie.

### 3.1.5. Etymologie

Sinds 2020 zijn de Etymologiebank (etymologiebank.nl), en de Uitleenwoordenbank (uitleenwoordenbank.ivdnt.org) in het beheer van het INT. Voor 2022 zijn inhoudelijke uitbreidingen voorzien voor de Etymologiebank. Verder zal het INT verantwoordelijk zijn voor het hosten van de data van het Global Anglicism Database Network; hiervoor zal een online bewerkingsapplicatie met behulp van het Lex'it platform worden ontwikkeld.

# 4. Beschrijving van de Nederlandse dialecten

Sedert 2020 hebben de dialecten van het Nederlands een plek gekregen op het INT. Na de lancering van de Database van de Zuidelijk-Nederlandse Dialecten in 2020, kreeg in 2021 de lancering van de *Atlas van het dialect in Vlaanderen*, waar het INT aan meewerkte de nodige aandacht in de pers.

## 4.1. Elektronische Woordenbank van de Nederlandse dialecten (eWND)

In 2021 is de hosting en het onderhoud van het eWND-portaal overgenomen door het INT. Dit wordt in 2022 uitgebreid met nieuwe dialectwoordenboeken.

## 4.2. Database van de Zuidelijk-Nederlandse Dialecten (DSDD)

Eind 2021 bevatte de Database van de Zuidelijk-Nederlandse Dialecten (DSDD) ongeveer 20.000 concepten. De ontbrekende landbouwwoordenschat uit de drie regionale woordenboeken die ten grondslag liggen aan de database is grotendeels toegevoegd in 2021. Aan de UGent worden in de eerste helft van 2022 nog ontbrekende lemmata uitgewerkt (vooral over de oogst en het paard) die in de loop van 2022 zullen worden toegevoegd.

In 2022 wordt verder gewerkt aan het vullen van de database met ontbrekend materiaal, namelijk de woordenschat van de vaktalen, die slechts beperkt vertegenwoordigd was in het pilotproject (bv. de bakker, de smid, de metaalverwerkende beroepen). Met hulp van vrijwilligers zal ook worden nagegaan of er nog lacunes zijn, die met het oog op een gelijke verdeling kunnen worden aangevuld.

Verder zullen de gegevens voor het Zeeuwse taalgebied (ook een Zuidelijk-Nederlands dialect) worden toegevoegd. De uitdaging hierbij is dat het Zeeuws alleen een semasiologische beschrijving heeft. Het DSDD-portaal is onomasiologisch van opzet. Om tot een goede strategie te komen voor het toevoegen van semasiologische bronnen aan het portaal is bij de analyse ook een aantal lokale

dialectwoordenboeken uit de woordenbanken van Nederland en Vlaanderen (eWND, gehost door INT, en woordenbank.be) betrokken. Op lange termijn beogen we immers een dialectplatform voor het hele Nederlandse taalgebied te realiseren.

In dat kader zal in 2022 gekeken worden of de gedigitaliseerde gegevens van het *Woordenboek van de Achterhoekse en Liemerse Dialecten* (WALD) kunnen worden geïntegreerd in een dialectplatform voor het Nederlandse taalgebied.

### 4.3. Gesproken Corpus van de Zuidelijk-Nederlandse Dialecten

Het INT is partner in het project Gesproken Corpus van de Zuidelijk-Nederlandse Dialecten, een project dat loopt van 2020 tot 2024 en wordt gerealiseerd aan de UGent. Het project beoogt de ontsluiting van een collectie van dialectopnames uit 768 plaatsen in België, Frankrijk en het zuiden van Nederland, opgenomen tussen 1963 en 1976 (te beluisteren via www.dialectloket.be en op de Nederlandse dialectenbank: https://www.meertens.knaw.nl/ndb/).

De opnames worden volgens een nieuw ontwikkeld transcriptieprotocol getranscribeerd om vervolgens met bestaande tools taalkundig verrijkt te worden. Het INT zal de audio, de transcripties en de annotaties op termijn vrij online beschikbaar en doorzoekbaar maken en duurzaam bewaren. Het INT heeft de eerste twee jaar van het project een adviserende rol. In de tweede helft van 2022 zal in overleg met UGent verder onderzocht worden welke stappen gezet moeten worden om de dialectopnames en de bijbehorende transcripties te ontsluiten.

## 5. Taalmaterialen

Een structurele taak van het INT is het beheren en ter beschikking stellen van taalmaterialen, te vinden op de in 2020 vernieuwde website. Onderdeel van de beheertaak is het informeren en adviseren van gebruikers. De kennis die hierbij is opgedaan zal verzameld worden en gepubliceerd ten behoeve van het geplande CLARIN Knowledge centre voor de Nederlandse taal (zie hierna). In 2022 wordt voorts onderzocht voor welke datasets de licenties kunnen worden aangepast volgens internationaal gebruikelijke licentiemodellen voor (open) data zoals GPL, Creative Commons en Apache Licence.

### 5.1. CLARIN-ERIC; het INT als CLARIN-centrum

Het INT is al jaren CLARIN-B centrum voor Nederland. In 2020 werd het INT betrokken bij de indiening van CLARIN-VL, als derde partij, bij de Vlaamse FWO/EWI-call met betrekking tot International Research Infrastructures. Dit project werd goedgekeurd met als resultaat de oprichting van CLARIN-België in september 2021. Voor 2022 houdt dit in dat Vincent Vandeghinste voor het INT de rol vervult als Nationaal Coördinator voor CLARIN op Europees niveau, als

vertegenwoordiger voor België in het National Coordinators Forum. Daarnaast neemt Jesse de Does vanuit het INT de vertegenwoordiging van België op in het Standing Committee on CLARIN Technical Centres. Hierdoor zorgt het INT voor twee van de drie CLARIN-België vertegenwoordigers op Europees niveau. Het INT is, eveneens als derde partij, betrokken bij CLARIAH-VL, het door FWI/EWI gefinancierde project waarin het grootste deel van de Vlaamse CLARIN taken gefinancierd wordt. Het INT zal zo een actieve rol vervullen als liaison tussen de Belgische onderzoekers en de Europese CLARIN infrastructuur. We voorzien ook een nauwe samenwerking tussen de Vlaamse onderzoekers in CLARIAH-VL en het INT als CLARIN-B centrum voor België. Daarnaast bestaan de voorziene taken voor het INT in 2022 uit het organiseren van zogeheten User Involvement events, waarbij CLARIN onder de aandacht gebracht wordt van onderzoekers in de Digital Humanities; uit het opnemen van tools en datasets gemaakt door Vlaamse onderzoekers en de integratie hiervan in de Europese CLARIN-infrastructuur; en uit het delen van tools en modellen met Vlaamse (en andere) onderzoekers.

In 2021 werd het INT een CLARIN Knowledge centre voor het Nederlands. Dit houdt in dat het INT een expertisecentrum geworden is voor CLARIN-gebruikers die informatie en raad willen omtrent het Nederlands, zijn bronnen en tools. Dit behelst het onderhouden en uitbreiden van Engelstalige webpagina's waarin internationale onderzoekers hun weg vinden met hun vragen over het Nederlands, naast het voorzien van een helpdesk waarbij vragen omtrent het Nederlands naar best vermogen beantwoord worden. Zie ook https://kdutch.ivdnt.org.

## 5.2. European Language Resources Coordination Initiative (ELRC)

Het INT is betrokken bij het ELRC-initiatief. Dit is een doorlopend initiatief dat als doel heeft tekstdata te verzamelen in alle EU-lidstaten, IJsland en Noorwegen, die gebruikt kunnen worden om CEF eTranslation (de automatische vertaaldienst van de Europese Commissie) verder te ontwikkelen. 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. In 2022 zal het INT in het kader van het ELRC-initiatief het belang van het verzamelen van Nederlandse taaldata verder blijven promoten zowel binnen Europa als binnen Nederland.

## 5.3 European Language Grid (ELG)

In 2022 loopt het project European Language Grid (ELG) ten einde. Het INT is hiervoor National Competence Centre (NCC) voor Nederland, en zorgde samen met de Universiteit Antwerpen (NCC voor België) voor de gegevens omtrent het Nederlands. In 2022 wordt nog een eindverslag van de werkzaamheden van ELG gepland en het bijwonen en eventueel bijdragen aan de conferentie META-

FORUM die plaatsvindt te Brussel van 8 tot 10 juni 2022, en die het slotevent vormt voor het ELG-project.

## 5.4 European Language Equality (ELE)

In 2022 loopt het project European Language Equality (ELE) ten einde. Het project loopt parallel aan ELG (zie hierboven). Ook hier is het INT het National Competence Centre (NCC) voor Nederland, en zorgde samen met de Universiteit Antwerpen (NCC voor België) voor de gegevens omtrent het Nederlands. Er werd een onderzoek ingesteld naar de stand van de digitale taalmaterialen voor het Nederlands, de database werd up-to-date gemaakt en in 2022 schrijven we een verslag over de situatie van de digitale ondersteuning van het Nederlands. Ook hier wordt tijdens de conferentie META-FORUM in juni 2022 verslag van uitgebracht.

## 5.5. Impactcentrum en digitization.eu

Het INT is voorzitter van het IMPACT Centre of Competence (www.digitisation.eu). Dit is een non-profitorganisatie bestaande uit publieke en commerciële organisaties met als doel de digitalisering van historisch materiaal “beter, sneller, en goedkoper” te maken. Het centrum voorziet in data, tools, services en expertise op het gebied van document imaging, taaltechnologie en het verwerken van historisch tekstmateriaal. Het IMPACT Centre of Competence is sedert 2019 ook CLARIN Knowledge centre en wordt sedert het najaar van 2021 geleid door Sally Chambers (UGent en KBR). Voor 2022 staat onder andere de organisatie van DATeCH op het programma.

De werkzaamheden m.b.t. digitalisering die in de context van CLARIAH plus worden uitgevoerd, worden in samenwerking met het Centre uitgevoerd. In het najaar van 2021 is het INT betrokken bij de aanvraag voor een Europees project *Editio*[3] om te voorzien in services voor digitalisering.

# 6. Overige infrastructuur- en netwerkprojecten

## 6.1. European Lexicographic Infrastructure (ELEXIS)

Het INT is partner in ELEXIS (https://elex.is), een Europees Horizon 2020-project, dat loopt van 1 februari 2018 tot eind juli 2022.

Het doel van het project is om een infrastructuur voor e-lexicografie op te zetten. ELEXIS streeft ernaar de lexicografische inspanningen binnen Europa zoveel mogelijk te harmoniseren door best practices te formuleren, conversietools te ontwikkelen en, belangrijker nog, door de bestaande

[3] in het kader van de HORIZON-INFRA-2021-EOSC-01 (Enabling an operational, open and FAIR EOSC ecosystem (2021)); topic HORIZON-INFRA-2021-EOSC-01-04 ; type of Action: HORIZON-RIA

lexicografische bronnen aan elkaar te koppelen, zodat ze kunnen worden gebruikt om nieuwe data, technologieën, producten en diensten te ontwikkelen. Tevens zal de ELEXIS-infrastructuur door training en educatie bijdragen aan het verkleinen van verschillen in expertise tussen lexicografen in Europa.

Het INT leidt het werkpakket ‘Lexicographic data and workflow’. Daarnaast werkt het INT mee aan andere werkpakketten, met name de werkpakketten ‘Interoperability and Linked (Open) Data’, ‘Lexicographic data for NLP’, ‘NLP for lexicography’ en ‘Training and Education’. In de eerste helft van 2022, zullen de taken in deze werkpakketten worden afgerond en zal de infrastructuur worden opgeleverd.

## 6.2. CLARIAH+ Nederland

Het CLARIAH (Common Lab for Research in the Arts and Humanities) CORE-project (2015-2018) was erop gericht een gemeenschappelijke infrastructuur tot stand te brengen voor data-intensief wetenschappelijk onderzoek in de geesteswetenschappen. Vanaf begin 2019 tot en met 2023 loopt het vervolgproject CLARIAH-PLUS, waarin het accent nog meer gericht is op het concreet ondersteunen van de onderzoeker door middel van het tot stand brengen van (virtuele) onderzoeksomgevingen.

Het INT houdt zich onder andere bezig met een verbetering van de infrastructuur voor historisch Nederlands, uitbreiding op de corpuszoekmachine BlackLab naar parallelle corpora en treebanks, hulpmiddelen voor het aanbrengen van persistente gebruikersannotaties in corpuszoekresultaten, een gebruikersvriendelijker digitalisatieworkflow en curatie van dialectwoordenboekdata.

Het werk zal zich in 2022 richten op het grotendeels afronden van het werk aan de infrastructuur voor historisch Nederlands, verder ontwikkelen van de infrastructuur voor digitalisatie en conversie, voortgaan met de implementatie van het doorzoeken van treebanks en parallelle corpora in de BlackLab-corpusomgeving. Voorts zal worden bijgedragen aan de use cases, onder andere door werk aan het 17e-eeuwse krantencorpus, waarbij de deelcollectie die reeds in tekst is omgezet gecureerd wordt, en een tweede deelcollectie na de automatische tekstherkenning door vrijwilligers zal worden gecorrigeerd.

## 6.3. CLARIAH Vlaanderen

Het INT was in 2020 betrokken bij de indiening van CLARIAH-VL: Advancing the open humanities service infrastructure, als derde partij, bij de Vlaamse FWO/EWI-call met betrekking tot International Research Infrastructures. Dit project werd gehonoreerd en de hoofdtaak van het INT is het voorzien van de benodigde infrastructuur voor het opzetten van het Digital Text Analysis Dashboard & Pipeline. Het doel van deze infrastructuur is om onderzoekers uit de Digital Humanities toe te staan

teksten van automatische annotaties te voorzien, zonder van hen een technische achtergrond te verwachten, en dit d.m.v. een cloud-based systeem waarbij teksten geüpload kunnen worden. Hiervoor is het noodzakelijk om, in samenwerking met de Vlaamse CLARIN/CLARIAH-partners tools zoals taggers en parsers te benchmarken, zodat de beste tools ter beschikking gesteld kunnen worden. Er wordt in het kader van CLARIAH-VL ook verder gewerkt aan een pilotproject in samenwerking met de Vlaamse Super Computer (VSC), waarbij het plan is om een contextueel taalmodel (cf. BERT en BART modellen) te trainen op basis van de corpora hedendaags Nederlands waarover het INT beschikt. Dit project dient als test voor zowel de VSC als CLARIAH-VL om de gebruiksvriendelijkheid van de toegang tot de supercomputers te verbeteren, zodat deze ook makkelijker bruikbaar worden voor onderzoekers in de Digital Humanities.

## 6.4. CLARIN-België / CLARIN-Vlaanderen

Het INT was in 2020 betrokken bij de indiening van CLARIN-VL , als derde partij, bij de Vlaamse FWO/EWI-call met betrekking tot International Research Infrastructures. Dit project werd goedgekeurd met als resultaat de oprichting van CLARIN-België in september 2021. Voor 2022 houdt dit in dat Vincent Vandeghinste voor het INT de rol vervult als Nationaal Coördinator voor CLARIN op Europees niveau, als vertegenwoordiger voor België in het National Coordinators Forum. Daarnaast neemt Jesse de Does vanuit het INT de vertegenwoordiging van België op in het Standing Committee on CLARIN Technical Centres. Hierdoor zorgt het INT voor twee van de drie CLARIN-België vertegenwoordigers op Europees niveau.

Daarnaast bestaan de voorziene taken binnen CLARIN-VL voor het INT in 2022 uit het organiseren van zogeheten User Involvement events, waarbij CLARIN onder de aandacht gebracht wordt van onderzoekers in de Digital Humanities; uit het opnemen van tools en datasets gemaakt door Vlaamse onderzoekers en de integratie hiervan in de Europese CLARIN-infrastructuur; en uit het delen van tools en modellen met Vlaamse (en andere) onderzoekers.

## 6.5. SignOn-project

Het INT is als consortium betrokken bij het SignON-project, dat vanaf voorjaar 2021 voor drie jaar gefinancierd wordt binnen het kader van het Horizon 2020 programma van de Europese Commissie. Het hoofddoel van dit project is het opzetten van automatische vertaalservices tussen gebarentalen en zogenaamde gesproken talen. De gebarentalen die bovenaan de agenda staan van deze Research and Innovation Action zijn Vlaamse Gebarentaal, Nederlandse Gebarentaal en Ierse Gebarentaal. Gesproken talen zijn in eerste instantie het Nederlands en het Engels. In latere fase wordt ook het Spaans en Spaanse Gebarentaal toegevoegd. Het consortium van dit project heeft een sterk Belgisch-Nederlandse component, met als consortiumpartners uit België: VRT, KU Leuven, UGent, Vlaams

Gebarentaalcentrum en European Union for the Deaf. Vanuit Nederland nemen deel: INT, de Taalunie, Radboud Universiteit Nijmegen, Tilburg University, en als derde partij Beeld en Geluid. Het project wordt geleid door Dublin City University.

De taak van het INT bestaat hoofdzakelijk uit het opzetten van de infrastructuur voor dit onderzoek. Er wordt eveneens gewerkt aan het verzamelen van gebarentaalcorpora, zowel voor VGT als voor NGT, en de inspanningen om corpora van de VRT ter beschikking te krijgen worden verdergezet. Een andere taak van het INT is om de infrastructuur op te zetten om Vertalen-als-een-service aan te kunnen bieden, die dan aangesproken kan worden binnen de Android- en iPhone-apps die ontwikkeld worden in de use cases, die in samenspraak met de doelgroepen ontwikkeld worden.

## 6.6. SABeD: Spoken Academic Belgian Dutch

Het industrieel onderzoeksfonds KU Leuven heeft in 2020 het project Spoken Academic Belgian Dutch goedgekeurd, dat twee jaar duurt. Het project werd aangevraagd door Elke Peters van het Centrum voor Taal en Onderwijs, in samenwerking met twee onderzoeksgroepen die deel uitmaken van Leuven.AI: het Centrum voor Computerlinguïstiek en de ESAT-PSI Speech groep. Het INT is in deze aanvraag derde partij, en zal zorgen voor de opname van het corpus in de CLARIN-infrastructuur, zowel als download voor onderzoek als online doorzoekbaar, op gelijkaardige wijze als nu het geval is voor het Corpus Gesproken Nederlands in de OpenSoNaR-toepassing.

Hoorcolleges zijn typisch voor het hoger onderwijs. In hoorcolleges leren studenten nieuwe lesinhouden in een taalregister waarmee ze weinig vertrouwd zijn, academisch Nederlands. Het doel van dit project is (1) om een corpus academisch gesproken Nederlands te compileren en (2) hierbij de effectiviteit van spraaktechnologie te onderzoeken voor automatische transcriptie van gesproken teksten, (3) om nadien een woordfrequentielijst academisch gesproken Nederlands en (4) een woordenschattoets academisch gesproken Nederlands te kunnen ontwikkelen. De compilatie van dit corpus laat toe leermateriaal en toetsen voor instromers te creëren. Het corpus zal een belangrijk hulpmiddel zijn voor zowel onderzoekers als beleidsmakers.

## 6.7. European network for Web-centered linguistic data science

Het INT neemt deel aan de NexusLinguarum COST-actie. Het thema van deze actie is ‘linguistic data science’, een deelgebied binnen de opkomende ‘data science’. Taalkundige data vormen een specifiek geval en zijn tot nu toe nog grotendeels onontgonnen in een big data-context.
Het hoofddoel van NexusLinguarum is om taalkundigen, computerwetenschappers, terminologen en andere belanghebbenden in één netwerk bij elkaar te brengen om zo samenwerking en kennisdeling op het gebied van ‘linguistic data science’ te bevorderen. De actie is eind oktober 2019 van start gegaan en heeft een looptijd van 4 jaar.

De activiteiten van de actie voor 2022 omvatten werkvergaderingen, conferenties en workshops, training schools, STSM's (Short Term Scientific Missions) en andere evenementen.