/instituut voor de
Nederlandse taal/

Beleidsplan 2024

22 november 2023

Inhoudsopgave

1. Het INT als kennisinstituut voor het Nederland

Het Instituut voor de Nederlandse Taal (INT) heeft zich in de afgelopen jaren succesvol omgevormd tot een breed opgezet kennisinstituut voor het Nederlands. Die transitie werd vastgelegd in het meerjarenbeleidsplan 2018-2022 en in goed overleg met de Taalunie uitgevoerd. Dankzij nauwgezette projectplanning en begrotingsopvolging zijn de doelstellingen voor de afgelopen beleidsperiode over het geheel genomen verwezenlijkt. Dat heeft zich in 2021 vertaald in een positieve beoordeling door de externe visitatiecommissie. Voortbouwend op de aanbevelingen van de commissie heeft het INT in het najaar van 2022 zijn meerjarenbeleidsplan voor de periode 2023-2027 neergelegd. Daarin werd uiteengezet hoe het INT zijn rol als het kennisinstituut voor de Nederlandse taal, met een focus op de digitale taalinfrastructuur, verder zal uitbouwen, en hoe het daarbij aansluit bij de krachtlijnen van de Taalunie rond digitalisering, internationalisering en inclusie/diversiteit. Goedgekeurd door de Taalunie en de Raad van Toezicht, vormt het meerjarenbeleidsplan 2023-2027 dan ook het uitgangspunt voor het beleidsplan 2024. In 2023 werd een aanvang gemaakt met het opzetten van een groot woordregister, zoals uitgewerkt in het meerjarenbeleidsplan en dit wordt voortgezet in 2024.
In 2024 neemt opnieuw een lid van de raad van toezicht afscheid: Gertine van der Vliet beëindigt haar mandaat en wordt vervangen door Erik Boels. In de raad van advies zijn eveneens wisselingen, enkele leden namen afscheid en een nieuw lid werd aangetrokken waardoor de verankering met de Universiteit Leiden sterker wordt.

Het INT heeft als structureel gefinancierd kennisinstituut een unieke positie en opdracht om voor het hele Nederlandse taalgebied (Nederland en de Caribische rijksdelen, Vlaanderen en Suriname) op een wetenschappelijk verantwoorde wijze de digitale taalinfrastructuur uit te bouwen. Het INT voert daarbij een aantal taken uit het Taalunieverdrag uit. We verwijzen hier naar hoofdstuk 1 uit het Taalunieverdrag, artikelen 2, 3, 4 en 5. Het INT ontwikkelt enerzijds zelf corpusdata, linguïstische databanken en taalsoftware voor een aantal specifieke domeinen, of de ontwikkeling ervan ondersteunt, en anderzijds dat het INT ook taalmaterialen en taalsoftware van andere kennisinstellingen verzamelt en samen met de eigen taalmaterialen duurzaam ter beschikking stelt via repository’s, websites, API’s en als open source software. Het INT promoot die taalinfrastructuur bij onderzoekers, ontwikkelaars en het brede publiek om zo Research & Development en andere activiteiten rond de Nederlandse taal te stimuleren en te ondersteunen. Daarnaast heeft het INT als toegepast onderzoeksinstituut de doelstelling de kennis en expertise over taalinfrastructuur verder uit te bouwen door eigen wetenschappelijk onderzoek. Om de onderzoeksactiviteiten van het INT meer in de kijker te zetten is er aan de website een specifiek gedeelte toegevoegd zodat onderzoekers onze activiteiten goed kunnen terugvinden. Daar is een overzicht te vinden van deelname aan extern gefinancierde nationale en internationale onderzoeks- en infrastructuurprojecten. De opdracht van het INT is te vergelijken met andere taalinstituten in Europa waarmee het intensief samenwerkt in Europese projecten en netwerken. Zoals door de

visitatiecommissie al werd aangegeven, is tegelijk wel duidelijk dat het INT erg klein is in vergelijking met deze buitenlandse tegenhangers en dat de financiering achter blijft. Uit het meerjarenbeleidsplan 2023-2027 spreekt desalniettemin de ambitie om de taalinfrastructuur voor het Nederlands op hoog niveau te houden, maar de activiteiten in het huidige beleidsplan 2024 blijven noodgedwongen eerder conservatief ingepland omdat er niet voldoende mensen en middelen zijn.

In de volgende paragrafen wordt uiteengezet hoe het INT in 2024 uitvoering zal geven aan zijn vele taken en hoe daarbij wordt gestreefd naar zoveel mogelijk samenhang en synergie in de planning van die taken en de uitvoering ervan. Deze kerntaken worden uitgevoerd vanuit de structurele financiering van de Taalunie. Wat de taalinfrastructuur voor het Nederlands zelf betreft, wordt in paragrafen 2 t.e.m. 7 toegelicht hoe het INT in 2024 met een aantal integratie-operaties de efficiëntie bij de opbouw en het beheer wil verhogen en tegelijk met nieuwe initiatieven de blijvend hoge kwaliteit van het aanbod voor gebruikers wil waarborgen. Paragraaf 8 geeft een overzicht van hoe het INT als onderzoeksinstituut binnen nationale en internationale projecten zijn expertise ook in 2024 ter beschikking zal stellen en verder uitbouwen. Paragraaf 9 rondt af met een overzicht van de activiteiten waarmee het INT in het komende jaar de taalinfrastructuur voor het Nederlands bij een breed publiek bekend wil maken.

2. Language Resources Repository en CLARIN-centrum

Zoals in het meerjarenbeleidsplan 2023-2027 gesteld vervult het INT een essentiële rol voor de taalinfrastructuur voor het Nederlands: enerzijds het duurzaam beschikbaar stellen van corpora, linguïstische databanken en taaltechnologische software, ontwikkeld door het INT zelf of door andere kennisinstellingen, in een language resource repository en anderzijds deze taalinfrastructuur verspreiden en bekend maken via verschillende platformen, en dan vooral het Europese CLARIN. Dit is in lijn met artikelen 4g en 5e van het Taalunieverdrag.

Language Resources Repository

Momenteel zijn er twee elkaar overlappende catalogi, nl. de Taalmaterialen-catalogus en het CLARIN-portaal. In het meerjarenbeleidsplan wordt de integratie van beide catalogi in het vooruitzicht gesteld. In 2023 werd de eerste fase van deze werkzaamheden opgestart, d.w.z. een exploratieve studie van de mogelijkheden die er zijn om deze geïntegreerde implementatie uit te voeren. Deze exploratie wordt verdergezet. De geschiktheid van bestaande open source software platformen, zoals de LINDAT D-space, CLARIN Virtual Language Observatory, de European Language Grid (ELG) of de ELRC-SHARE en van eventuele Repository-as-a-Service platformen wordt nagegaan. Het INT kijkt hierbij naar verdere evoluties in het Europese taalinfrastructuurlandschap, zoals de European Open Science Cloud, de Europese Language Data Space, of op Nederlands niveau het INEO-platform, waarin alle

CLARIAH-NL resources samengebracht worden. Hierbij wordt gelet op de consistente, crossplatform terugvindbaarheid en beschikbaarheid van de door het INT beheerde resources.

Aan de toeleveringskant zal eveneens gekeken worden wat de mogelijkheden zijn voor het gebruik van bestaande depositieprocedures voor leveranciers van taaldata, zodat dit zo automatisch mogelijk kan gebeuren. Hierbij wordt expliciet gekeken naar wat er binnen CLARIN-ERIC gebeurt en aangeraden wordt.

In 2024 wordt verder onderzocht voor welke datasets de licenties kunnen worden aangepast volgens internationaal gebruikelijke licentiemodellen voor (open) data zoals GPL, Creative Commons en Apache Licence. Het INT bekijkt of hiervoor gebruikgemaakt kan worden van de door CLARIN aangeraden licence selection tools.

Het INT biedt als expertisecentrum ondersteuning en advies aan geïnteresseerde gebruikers van de (Nederlandse) taalinfrastructuur. Eerder bestond al de servicedesk voor vragen over Taalmaterialen (servicedesk@ivdnt.org), maar in de komende jaren zal deze dienstverlening, net als de repository, geïntegreerd worden met wat het INT onder de Europese CLARIN-paraplu aan ondersteuning en advies biedt. Het INT is immers door CLARIN erkend als expertisecentrum voor het Nederlands. Op de portaalsite K-Dutch wordt de expertise van het INT omtrent het Nederlands gepresenteerd voor een internationaal publiek en wordt een servicedesk aangeboden voor concrete vragen. Het INT was betrokken partij bij het indienen voor een aanvraag tot CLARIN Knowledge Centre for Lexicography in 2023. In 2024 leidt dit hopelijk tot de oprichting van dit K-Centre dat de expertise over de lexicografische infrastructuur op Europees niveau op een duurzame manier zal coördineren. Ten slotte zal het INT als expertisecentrum de bekendmaking van de taalinfrastructuur bij onderzoekers in de sociale en humane wetenschappen en het ruimere publiek verder zetten door middel van lezingen, seminaries in lessenreeksen, en hands-on workshops, de zogenaamde User Involvement events. Zo wil het INT zich nog verder profileren als steunpunt voor onderzoekers en studenten in Vlaanderen en Nederland die meer behoefte hebben aan taalmaterialen uit CLARIN. Ook dit is een taak uit het Taalunieverdrag, met name artikel 2b dat gaat over de bevordering van de kennis van de Nederlandse Taal en artikel 2d dat verwijst naar de bevordering van de studie en verspreiding van de Nederlandse Taal.

Het INT als CLARIN-centrum

Het INT is al jaren CLARIN-B-centrum voor Nederland en blijft dat in 2024. Ook in 2024 blijft het INT het enige CLARIN-B-centrum voor België. Dit houdt in dat het INT instaat voor de technische infrastructuur voor Belgisch CLARIN-onderzoek en als Nationaal Coördinator voor CLARIN-BE optreedt en België vertegenwoordigt in het National Coordinators Forum. Daarnaast neemt 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-vertegenwoordigingen op Europees niveau. Het INT is ook vertegenwoordigd in het Knowledge Infrastructure Committee, dat een coördinerende rol vervult voor de Knowledge Centres. Het INT vervult eveneens de rol van Program Committee Chair voor de CLARIN Annual Conference 2024, die vermoedelijk in Bilbao zal plaatsvinden.

Het INT is, als derde partij, betrokken bij CLARIAH-VL, het door FWO/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. Het instituut voorziet 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 2024 uit medewerking aan het organiseren van zogeheten User Involvement Events (cf. infra), waarbij CLARIN onder de aandacht gebracht wordt van onderzoekers; 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.

3. Corpusinfrastructuur

Zorgvuldig samengestelde en wetenschappelijk onderbouwde corpora vormen een essentieel onderdeel van de taalinfrastructuur voor het Nederlands. Ze bevatten immers de primaire taaldata op basis waarvan de Nederlandse taal gedocumenteerd kan worden en taalapplicaties ontwikkeld kunnen worden. De corpusinfrastructuur van het INT omvat naast corpora een arsenaal aan gereedschappen voor dataprocessing en ontsluiting. De werkzaamheden worden voor een belangrijk deel uitgevoerd ten behoeve van de verdere uitbouw van de kennisbank maar resulteren ook in een corpusinfrastructuur voor de brede onderzoeksgemeenschap.. Daarnaast is het INT betrokken in diverse projecten waarin corpora worden gebouwd, waarbij het INT, naast expertise, infrastructurele ondersteuning biedt voor het bouwen, het gebruik dan wel het ter beschikking stellen van het corpusmateriaal.

Monitorcorpus

In principe wordt het Corpus Hedendaags Nederlands verder geüpdatet met het krantenmateriaal dat we wekelijks ontvangen. Het doel voor het hedendaags Nederlands blijft om het krantenmateriaal in het Corpus Hedendaags Nederlands (CHN) aan te vullen om tot een min of meer evenwichtig monitorcorpus voor dit millennium te komen. Daarvoor zullen de nodige acquisitiewerkzaamheden verricht worden. Mits er financiële ruimte is, en daardoor de beschikking over een dataprocessingspecialist, zal aan het CHN nieuw materiaal toegevoegd worden dat via samenwerkingen in extern gefinancierde projecten beschikbaar komt, zoals bijvoorbeeld de het ondertitelingsmateriaal van SignOn. Daarnaast wordt verder gewerkt aan de uitbreiding van het materiaal uit Caribisch-Nederland.

Voor het historisch Nederlands zal er, naast het afzonderlijk online publiceren van diverse corpora, gewerkt worden aan de verdere uitbouw van het in het meerjarenbeleidsplan genoemde groot diachroon corpus.

Taalkundige Verrijking

De INT-corpora worden samengesteld uit bestaand digitaal materiaal of, waar nodig, door digitalisering. De brondata worden geconverteerd naar eenzelfde XML-standaard (TEI) en zorgvuldig van metadata voorzien en daarna automatisch taalkundig verrijkt. Metadata en taalkundige verrijking bieden een nadrukkelijke meerwaarde om zinvolle informatie uit de corpora te kunnen extraheren. De werkzaamheden voor het komend beleidsjaar worden hieronder gespecificeerd.

HedendaagsNederlands

Het is de bedoeling om de huidige pipeline voor taalkundige verrijking te vervangen door een state-of-the-art universal dependencies pipeline die zowel woordsoort, lemma als syntactische annotatie zal aanbrengen. Om de interoperabiliteit van modern en historisch materiaal te behouden zal de relatie onderzocht worden tussen de ontwikkelde TDN-annotatierichtlijnen voor woordsoort en lemma en de overeenkomstige UD guidelines, en zal er een mapping worden gemaakt.

Om de mogelijkheden voor het extraheren van informatie over lexicaal combinatiegedrag uit corpusdata te verbeteren, wil het INT op termijn het Corpus Hedendaags Nederlands syntactisch verrijken. Hierbij denken we minimaal aan verrijking volgens het Universal Dependenciesmodel waarmee het INT aansluit bij internationale standaarden en wat in het technisch bereik ligt van de corpuszoekmachine BlackLab (de syntactische uitbreiding is geïmplementeerd in CLARIAH+).[1] In 2023 is naar voren gekomen dat een aantal taalspecifieke extensies op de dependenties nodig zijn om het verrijkte materiaal optimaal in te kunnen zetten binnen de informatie-extractie ten behoeve van de kennisbank. In 2024 stemt het INT dit af met de UD-community en werkt het de richtlijnen voor de syntactische annotatie uit. Het INT onderzoekt hoe het tot geschikt trainingsmateriaal kan komen.

Historisch Nederlands

In het kader van CLARIAH+ is het beschikbare trainingsmateriaal voor historische verrijking uitgebreid en zijn diverse opties verkend om, met behulp van dit materiaal, op deep learning gebaseerde technieken in te zetten voor deze taak. In 2024 rondt het INT deze werkzaamheden af en wordt het resultaat hiervan in het GALAHAD-platform, dat dan ook in productie zal gaan, geïntegreerd.

[1] Deze uitbreiding van BlackLab is nadrukkelijk bedoeld als basisfaciliteit voor grote hoeveelheden materiaal en pretendeert vanzelfsprekend niet de gesofisticeerde zoekmogelijkheden van gespecialiseerde treebank-engines als GrETEL, PaQu en PML-Tree Query te vervangen.

Metadata

De corpora hebben een gemeenschappelijk metadataformaat, met ruimte voor subcorpus-specifieke metadata. In 2023 is het INT tot een verdere uniformering van het metadatamodel voor historisch en modern corpusmateriaal gekomen, waardoor beide op termijn als één doorlopend diachroon corpus op metadatacategorieën doorzoekbaar kunnen worden.

In 2024 wordt onderzocht hoe het ontwikkelde metadataschema efficiënt kan worden geïmplementeerd als afzonderlijke metadatabase-component in de corpusworkflow.

Informatie-extractie

Het corpusmateriaal wordt toegankelijk gemaakt via een applicatie waarmee in de corpora gezocht kan worden. Wanneer de IPR (Intellectual Property Rights) het toelaten, wordt het corpusmateriaal ook als dataset beschikbaar gesteld in de language resource repository dan wel in de vorm van n-grammen. De software is open source beschikbaar.

Voor de ontwikkeling van de corpusapplicatie, die bestaat uit de search engine BlackLab[2] en de corpus frontend,[3] ligt de prioritering bij de ondersteuning van de diverse INT-taken. Deze werkzaamheden worden ondersteund door het samenwerken met diverse partijen die de software gebruiken, en door de mogelijkheden die externe projecten bieden om deze software verder te ontwikkelen. Voor wat betreft de backend van de corpusretrievalomgeving (BlackLab, BlackLab Server) staat voornamelijk de ondersteuning van steeds grotere corpora door middel van optimalisaties en gedistribueerd zoeken op het programma. Het onderzoek en de werkzaamheden daarvoor zijn in 2022 gestart en worden gefaseerd uitgevoerd. In 2024 wordt gestart met de laatste fase.

Voor de userinterface wordt verder onderzoek gedaan naar gebruikersvriendelijke querybuilding, onder andere linguïstisch gemotiveerde query-sjablonen en example-based query constructie. De in 2023 onderzochte uitgebreidere mogelijkheden voor extractie en visualisatie van distributie van lexicale variabelen, met name het mogelijk maken van groepering op meerdere (metadata)kenmerken, weergave van de groepering op queryonderdelen en visualisatie van trends, ook van meerdere variabelen samen, wordt geïmplementeerd.

4. Beschrijving van de woordenschat door de eeuwen heen

De wetenschappelijke, corpusgebaseerde beschrijving van de Nederlandse woordenschat in al zijn facetten blijft in de komende jaren een van de kerntaken van het INT (zie verdragstaak 4d van het

[2] https://github.com/INL/BlackLab
[3] https://github.com/INL/corpus-frontend

Taalunieverdrag). Zoals uiteengezet in het Meerjarenbeleidsplan 2023-2027, zal het INT in de komende jaren versterkt inzetten op twee belangrijke vernieuwingen voor de lexicale taalinfrastructuur:

- De integratie van alle componenten van de woordenschatbeschrijving in één centrale, modulair georganiseerde, relationele kennisbank van de Nederlandse woordenschat door de eeuwen heen, van waaruit bestaande en nieuwe lexicografische producten verder ontwikkeld zullen worden.
- Een versterking en explicitering van het corpusgebaseerde lexicografische proces, resulterend in een bidirectionele linking tussen primaire corpusdata en afgeleide lexicale data.

In 2023 werd een begin gemaakt met beide vernieuwingen waarbij de focus lag op het conceptuele ontwerp van de infrastructuur en het opstellen van de vereisten voor de databanken, de workflows, de software en de hardware. Er werden daarbij een aantal pilotstudies uitgevoerd en prototypes geïmplementeerd waarvoor in 2024 en productieversie geïmplementeerd wordt om ze vervolgens binnen de woordenschatbeschrijvingsworkflow in gebruik te nemen De reguliere werkzaamheden (updates en nieuwe content) aan de bestaande infrastructuur en lexicografische producten worden voortgezet.

Datamodel voor de centrale kennisbank van de woordenschat

Voor het succesvol uitwerken van een complexe data-infrastructuur voor de centrale kennisbank is een doordacht datamodel essentieel. Dat model bouwt voort op de al bestaande onderdelen in de huidige infrastructuur, met name het centrale lexicon GiGaNT (cf. infra), de koppeling op lemma-niveau met de lexicografische databanken en het Diachroon seMantisch lexicon van de Nederlandse Taal (DiaMaNT). In het uitgebreidere datamodel zullen volgende aspecten verder worden uitgewerkt, geëxpliciteerd en geformaliseerd:

- De verschillende modules van de kennisbank die telkens één lexicografische datacategorie behandelen. Vooral voor het betekenisregister (cf. infra) zal dit uitgebreid studiewerk vragen;
- De interacties en samenhang tussen de modules;
- De types linking met corpusdata;
- De manier waarop compatibiliteit met bestaande databanken, datamodellen, en back-compatibility bij toekomstige wijzigingen gegarandeerd wordt.

Behalve het formele aspect van het datamodel, d.w.z. de beschrijving van datacategorieën, de relationele structuur van de koppeling tussen de datacategorieën en het beschrijvingsvocabulaire, zullen ook richtlijnen en procedures voor compilatie, ontleding van bestaande lexicografische bestanden en wijzigingen aan de kennisbank uitgebreid en precies gedocumenteerd worden. In 2023 ging de aandacht naar de classificatie en modellering van de meerwoordsexpressies en naar de compatibiliteit van zowel de hedendaagse als historische lexicale beschrijving met de Tagset Diachroon Nederlands.

In 2024 wordt verder gewerkt aan de herziening van het datamodel van GiGaNT, met name waar het de consistentie en synchronisatie van gekoppelde data in meerdere deelmodules betreft. Op basis van de ervaringen uit pilotstudies van 2023 wordt daarenboven het datamodel voor het centrale betekenisregister verder uitgewerkt en in een eerste proefversie in gebruik genomen.

Het Centrale Lexicon (GiGaNT)

In 2024 zet het INT de reguliere werkzaamheden aan het centrale lexicon voort. Dat houdt in: het onderhoud aan de bestaande modules, het verder verwerken van de integratie van GiGaNT-Hilex en GiGaNT-Molex en de uitbreiding van GiGaNT-Molex ten behoeve van de diverse lexicografische producten die een koppeling met GiGaNT hebben, zoals woordenlijst.org (cf. infra). Daar waar mogelijk wordt al gewerkt aan de implementatie van de herzieningen van het datamodel (zie hierboven).

In 2023 is een eerste eerst testcase opgezet voor het uitrollen van de vernieuwde corpusgebaseerde lexicografische workflow waarbij het de bedoeling is om corpusmateriaal systematisch te screenen op lacunes en op nieuwe woorden, aangevuld met een omgeving om deze data makkelijk te analyseren, te structureren en te bewerken. Focus hierbij is het hedendaags Nederlands (Molex). Het doel is om het CHN, dat wekelijks geüpdatet wordt, systematisch te screenen ten behoeve van de uitbreiding van Molex. In 2024 wordt onderzocht hoe frequentie-informatie zoals tijdreeksen en regionale distributies aan Molex toegevoegd kunnen worden als kwantitatieve corpusevidentie.

Betekenisregister

Zoals uiteengezet in het Meerjarenbeleidsplan 2023-2027 is de koppeling van verschillende lexicografische databanken op betekenisniveau de belangrijkste innovatie en grootste uitdaging voor de woordenschatbeschrijving in de komende jaren. Binnen het datamodel voor de centrale kennisbank werd daarom in 2023 een begin gemaakt met het uitwerken van een betekenisregister dat die koppeling op betekenisniveau moet mogelijk maken. Het ontwerp en het op punt stellen van het datamodel is noodzakelijkerwijze een proces van trial and error. In 2023 werd daarom in parallel met het opstellen van een eerste datamodelconcept, ook een eerste proefversie van het betekenisregister gemaakt. Die proefversie is opgebouwd op basis van de lexicografische bronnen die tijdens de afgelopen beleidsperiode op lemmaniveau aan GiGaNT-Molex gekoppeld zijn (ANW, WNW, Referentiebestand Nederlands (RBN) en de Vertaalwoordenschat. Deze proefversie van het betekenisregister wordt in 2024 verder aan praktijktesten onderworpen om zowel bestaande betekenisbeschrijvingen te koppelen en nieuwe toe te voegen om zo het datamodel en de implementatie van het betekenisregister verder te verfijnen. Voor de koppeling van de historische woordenboeken aan het betekenisregister wordt in 2024 met een kleinere testset geëxperimenteerd,. De problemen en uitdagingen die bij de experimenten naar boven komen zullen dan de uitwerking van de diachrone aspecten van het datamodel informeren.

Lexicografische eindproducten, API's en datasets

Zoals het Meerjarenbeleidsplan 2023-2027 aangeeft, worden op termijn de huidige lexicografische eindproducten, zoals woordenlijst.org, ANW, WNW, Woordcombinaties en de historische woordenboeken, samen met eventuele nieuwe eindproducten, als afgeleide producten vanuit de centrale kennisbank verder ontwikkeld en verbeterd. Hoewel de huidige productspecifieke workflows ook in 2024 nog deels aangehouden worden, zullen in de loop van het komende jaar al wel de eerste stappen gezet worden om lexicografische content vanuit de centrale kennisbank in de eindproducten te publiceren.

Woordenlijst.org

Woordenlijst.org (de lijst met de officiële spelling van de Taalunie) wordt door het INT voortdurend uitgebreid en up-to -date gehouden, zoals vastgelegd in het Taalunieverdrag artikel 4b. Voor woordenlijst.org is die publicatie vanuit de centrale kennisbank nu al gerealiseerd, met name dan vanuit GiGaNT als de module voor de woordenschatbeschrijving op lemma-niveau. In 2024 zullen de systematische updates van woordenlijst.org verdergezet worden en beschikbaar zijn via de in 2023 vernieuwde applicatie.

Algemeen Nederlands Woordenboek (ANW)

Zoals in de vorige paragraaf aangegeven, is in 2023 binnen de centrale kennisbank een eerste proefversie van het betekenisregister gemaakt aan de hand van een aantal hedendaagse lexicografische bronnen. Dat zal dan toelaten om in de loop van 2024 een eerste set van betekenisbeschrijvingen en andere lexicografische informatie uit de centrale kennisbank in het ANW te publiceren. Daarbij wordt gelet op de eigenheid en consistentie van de betekenisbeschrijving binnen het ANW.

Woordenboek van Nieuwe Woorden (WNW)

Het WNW houdt in 2024 nog grotendeels zijn productspecifieke workflow aan maar wordt daarnaast ook al gevoed vanuit de eerste proefversie van de centrale kennisbank dankzij de vernieuwde corpusgebaseerde workflow voor de identificatie van nieuwe woorden binnen GiGaNT-molex (cf. supra). De lexicografische behandeling van neologismen wordt verder in overeenstemming gebracht met de modulaire opbouw en het datamodel van de centrale kennisbank en het betekenisregister.

Woordcombinaties

Woordcombinaties is de online taaltool die leerders van het Nederlands als vreemde taal en moedertaalgebruikers ondersteunt bij het gebruiken van woorden in context. Dit project inventariseert en beschrijft systematisch combinaties (collocaties, idiomen en patronen) in het Nederlands.
In 2024 wordt verder gewerkt aan de lexicografische beschrijving van de combinatiemogelijkheden van werkwoorden en substantieven. Dankzij de aangepaste werkomgeving zullen ook idiomen en formules een belangrijkere rol gaan spelen binnen het project.

Historische woordenboeken

In 2023 is een verregaande uniformering van de codering van de historische woordenboekdata uitgevoerd. Dat resulteerde in een versie van de data in TEI 5 van waaruit het onderhoud aan de woordenboekdata verder uitgevoerd kan worden. Voor dat onderhoud is een eerste versie van de infrastructuur ingericht waarmee de data van ONW en VMNW verder geüpdatet kunnen worden om uiteindelijk in GiGaNT geïntegreerd te kunnen worden. De infrastructuur wordt verder uitgebouwd om synchronisatie tussen de woordenboeken en GiGaNT te kunnen waarborgen.

Vertaalwoordenschat

De Vertaalwoordenschat is een platform waarop tweetalige bestanden worden ontsloten die in de jaren 90 en begin 2000 ontwikkeld werden door de Taalunie voor commercieel niet afgedekte taalparen. Inmiddels staan het Nederlands-Nieuwgrieks/Nieuwgrieks-Nederlands, het Nederlands-Portugees/Portugees-Nederlands, het Nederlands-Estisch en het Nederlands-Fins/Fins-Nederlands online en sinds eind 2023 ook het Nederlands-Deens.

In 2024 wordt ruimer ingezet op het aanbrengen van inhoudelijke verbeteringen aan de verschillende taalparen die inmiddels online staan (m.n. Fins, Deens, …) of al voorbereid zijn (zoals het Arabisch) door externe redacteuren met ondersteuning van het INT.

API's en datasets

De API voor de centrale kennisbank wordt in 2024 verder ontwikkeld vanuit de al bestaande lexiconservice, waarbij vooral gefocust wordt op het waarborgen van de beschikbaarheid en betrouwbaarheid van de dienst. Diezelfde informatie komt telkens ook in de nieuwe release van GiGaNT-Molex als dataset in de language repository ter beschikking.

5. Beschrijving van de Nederlandse dialecten

Als logische uitbreiding van de opdracht om de Nederlandse woordenschat in al haar facetten te beschrijven, hebben de dialectwoordenboeken uit het Nederlandse taalgebied sinds 2020 een plek gekregen op het INT. Het INT heeft de verantwoordelijkheid gekregen voor het beheer, de ontwikkeling en de beschikbaarstelling van diverse dialectproducten.

Elektronische Woordenbank van de Nederlandse dialecten (eWND)

In 2021 is de hosting en het onderhoud van het eWND-portaal overgenomen door het INT. Het eWND wordt in 2024 uitgebreid met nieuwe dialectwoordenboeken met behulp van vrijwilligers, die het materiaal voorbereiden.

Database van de Zuidelijk-Nederlandse Dialecten (DSDD)

Eind 2023 bevatte de Database van de Zuidelijk-Nederlandse Dialecten (DSDD) ongeveer 30.000 concepten. Het is de bedoeling om semasiologische woordenboekdata aan de DSDD toe te voegen. Een eerste woordenboek is het heel uitgebreide woordenboek van de Zeeuwse dialecten. In 2023 zijn de data van dit woordenboek geschikt gemaakt voor verdere verwerking in de DSDD. In 2024 wordtgewerkt aan de koppeling met de DSDD.

Aan de hand van het Overijssels woordenboek (zie verder) wordt in 2024 onderzocht hoe het platform verder uitgebouwd kan worden tot een dialectplatform voor het hele Nederlandse taalgebied.

Digitale infrastructuur voor dialecten en streektalen

Het INT zal, in samenwerking met en op vraag van de Taalunie, infrastructuur bieden aan streektaalorganisaties om hun talig cultureel erfgoed te beschrijven en beschikbaar te stellen. Het ontwerp en de uitwerking van de infrastructuur gebeurt met behulp van externe financiering. In 2023 werden de eerste stappen gezet met het pilootproject voor het Bildts (zie verder), In 2024 wordt er een vervolg gegeven

Digitale infrastructuur voor het Bildts

In 2023 heeft het INT een digitale infrastructuur opgezet voor streektalen en dialecten van het Bildts. Deze infrastructuur biedt het Bildts Aigene de nodige ondersteuning om hun lexicale data te bewerken en een omgeving om de data via een applicatie doorzoekbaar te maken. Afhankelijk van verdere financiering zal de infrastructuur verder uitgebreid worden met een component voor het bouwen en doorzoekbaar maken van corpusdata en manieren om audio en video toe te voegen.

Digitale infrastructuur voor het Overijssels

Vanaf eind 2023 wordt i.s.m. de Overijsselacademie een digitale infrastructuur ontworpen om kleinere onomasiologische woordenboeken zoals het Woordenboek van de Overijsselse dialecten (WOD) in een DSDD-achtige structuur om te zetten met zoekfaciliteiten en een kaartmodule. Deze infrastructuur kan dienen om later het Woordenboek van het Gelders en het Woordenboek van de Achterhoekse en Liemerse dialecten, die een vergelijkbare structuur hebben als het WOD te integreren in de infrastructuur. Daarnaast is het de bedoeling om in 2024 kleinere (semasiologische) woordenlijsten en woordenboeken van een dorp/stad of een regio gemakkelijk te kunnen bewerken en online te brengen. Op die manier wordt een eerste stap gezet voor een uitgebreidere digitale structuur.

LT2-lesmateriaal voor het Limburgs duurzaam digitaal toegankelijk voor iedereen

Op vraag van de Taalunie en De Raod veur 't Limburgs is bij de provincie Limburg een project aangevraagd om LT2-lesmateriaal voor het Limburgs duurzaam digitaal toegankelijk te maken. In

Limburg worden momenteel vier cursussen aangeboden (Echt, Heerlen, Roermond en Venlo) met behulp van audio- en videomateriaal. Cursisten krijgen de mogelijkheid deze bestanden via een website op te roepen en te beluisteren. het INT ontwikkelt in 2024 een platform om dit cursusmateriaal toegankelijk te maken. Dit onderdeel is een eerste stap naar een groter dialectplatform met informatie over de diverse dialecten en streektalen in Nederland en Vlaanderen.

6. Expertisecentrum voor Nederlandstalige 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. Dit bevordert het gebruik van Nederlandstalige terminologie bij het bredere publiek, in het onderwijs en bij de vakexperten. (zie artikel 4c en 5e van het Taalunieverdrag).

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. Daarom worden de hedendaagse termenlijsten op de ENT-webpagina's permanent geüpdatet en uitgebreid. Analoog hiermee worden in de lijn van het INT en zijn historische woordenboeken en corpora lijsten met historische termen verzameld. Hiervoor wordt de internationale Library of Congress-classificatie gebruikt. Een webrubriek zal op die basis worden toegevoegd wanneer een voldoende substantieel aantal historische termenlijsten beschikbaar is.

Tools

Qua terminologietools is in 2023 een nieuwe tool als pilot ontwikkeld en geïmplementeerd. Het gaat hierbij om de integratie van de bestaande termextractor TermTreffer en de termbank-editor TermBeheerder, zodat één tool ontstaat die geschikt is voor de invoer, de verwerking en het beheer van termmateriaal. Het resultaat dient een applicatie te zijn die in een breed, generiek verband en afgestemd op het Nederlands en zijn taaleigenschappen, de termextractiemodule en de editeerfunctionaliteiten van beide systemen combineert. Deze versie wordt door het INT zelf volledig ontwikkeld, rekening houdend met de vereisten van de gebruikers. Het tweede luik van dit project wordt in 2024 afgerond, zodat een volledig operationele applicatie beschikbaar zal zijn. Die zal mee worden getest door de bestuursleden van NL-TERM. De nieuwe tool krijgt de naam TermWerk.

Veldondersteuning

Klassieke veldondersteuning in de vorm van updates van de bestaande websiterubrieken, vooral de opleidingen in Nederland en Vlaanderen, blijft aan de orde. In 2024 worden vier uitgebreide

nieuwsbrieven Terminologie gedistribueerd en wordt de evenementenrubriek voor terminologie continu bijgehouden.
Het pilootproject over hogeronderwijstermen in de HOTNeV-termbank is afgerond, maar toch wordt de termbank doorlopend aangevuld via stages en scripties. De groei van deze databank is dus afhankelijk van de stages die bij het ENT worden aangevraagd. Afstudeerscripties van studenten kunnen eveneens bijdragen aan de verdere invulling van de HOTNeV-termbank. Meer algemeen blijft de begeleiding van studenten belangrijk en het stageaanbod voor terminologiewerk wordt blijvend gepromoot. Dit draagt immers ook bij aan de netwerkpositie van het ENT.
Verdere samenwerking met de Termraad wordt in 2024 gerealiseerd. Via dit overlegplatform voor terminologische afstemming binnen het Nederlandse taalgebied werken EU-terminologen samen met Belgische en Nederlandse partners uit overheidsdiensten, terminologieverenigingen en vertaalopleidingen. Het INT participeert aan de driemaandelijkse bijeenkomsten met het oog op de verzameling, beschrijving en uniformering van terminologie in specifieke vakgebieden. Op Europees niveau is het INT toegetreden tot het consortium dat een nieuwe COST-actie voorbereidt: COST Action Proposal OC-2022-1-26011 "**Collaborative Terminology Network**". Het voorstel werd ingediend op 20 oktober 2022 en in eerste fase niet goedgekeurd. Een herziening en herindiening wordt nu voorbereid.

Daarnaast werkt het INT verder aan drie speerpunten die in het verlengde liggen van de krachtlijnen van de Taalunie. De voortgang van deze projecten is echter wel afhankelijk van bijkomende financiering.
Medische vaktaal: een eerste versie van het Pinkhof geneeskundig woordenboek staat online via een applicatie die bij het INT werd ontwikkeld. Dit woordenboek wordt bijgewerkt met als primaire gebruiksfunctie een verklarend hedendaags medisch woordenboek en een taalbaak voor medisch Nederlands. Om dit te realiseren wordt er samengewerkt met de Stichting Beheer Pinkhof-database. Er werd een raad van advies opgericht om afgeleide medische terminologieprojecten te begeleiden.
Juridische vaktaal: wat betreft het juridisch woordenboek van M.C. Oosterveld-Egas Reparaz en Johanna Vuyk-Bosdriesz (Nederlands Recht) wordt verder gewerkt aan de updating en uitbreiding van het bestand. Het juridische woordenboek werd op 4 oktober 2023 gepresenteerd via een workshop in het kader van de week van het Nederlands. Het product is via een nieuwe applicatie gebouwd bij het INT. In 2024 probeert het INT – mits extra werkingsmiddelen – een uitbreiding van deze datacollectie te realiseren. Ook een uitbreiding met de terminologie van het Belgische recht is zeer wenselijk.
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. Het INT werkt 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.

7. Grammatica

Sinds 2020 valt grammatica binnen de structurele basistaken van het INT zoals ook vastgelegd in het Taalunieverdrag artikel 4d. Dit betekent dat het INT zorgt voor de ontwikkeling, het beheer en de beschikbaarstelling van de verschillende digitale grammaticaproducten. Hieronder vallen het Taalportaal en de e-ANS. Daarnaast maakt het INT als infrastructurele partner deel uit van het samenwerkingsverband achter Taaladvies.net, in lijn met artikel 5c in het Taalunieverdrag. Afgezien van de werkzaamheden aan de e-ANS en het Taalportaal (cf. infra), wordt gewerkt aan een geïntegreerd grammaticaportaal, een webpagina die de spil moet vormen van alle grammaticaonderdelen en fungeert als ontvangstpagina voor geïnteresseerde gebruikers, met informatie over projecten en producten, een zoekfunctie voor alle producten en een loket voor vragen.

e-ANS

In 2024 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. Door het jaar heen worden één à twee publicatiemomenten georganiseerd, waarbij de herziene hoofdstukken beschikbaar gemaakt worden voor het publiek. Daarnaast zal de webapplicatie op enkele punten doorontwikkeld worden, en wordt verder gewerkt aan de zogenaamde didactische laag, een module van de ANS waarin de inhoud op een toegankelijkere manier wordt gepresenteerd.

Taalportaal

In 2024 zal de update van Taalportaal worden doorgezet. De webapplicatie wordt verder bijgewerkt. Op inhoudelijk gebied wordt er met de betrokken auteurs gewerkt aan een update van de Nederlandse syntaxis, het Fries en de nieuw toegevoegde taal het Saterfries.

Grammaticaportaal

In 2024 wordt de ato-versie van het Grammaticaportaal verder uitgewerkt. Deze site zal als startpagina dienen voor de verschillende grammatica-applicaties binnen het INT: de e-ANS, het Taalportaal, Taaladvies.net, maar ook voor projecten als Dagenta (cf. infra) en Woordcombinaties (cf. supra)

8. Nationale en internationale samenwerkingsverbanden

Netwerken

EFNIL

Het INT is lid van EFNIL, de Europese federatie van nationale taalinstituten en participeert actief in dit netwerk. In deze federatie komen gelijkaardige taalinstellingen uit alle Europese lidstaten samen. EFNIL biedt de instellingen een forum om informatie over hun werk uit te wisselen en om informatie over taalgebruik en taalbeleid binnen de Europese Unie te verzamelen en te publiceren.

Daarnaast stimuleert de Federatie het bestuderen van de officiële Europese talen en een gecoördineerde aanpak van het leren van moeder- en vreemde talen, als middel om de taalkundige en culturele diversiteit binnen de Europese Unie te bevorderen. Ieder jaar is er een algemene vergadering en wordt er een workshop gehouden rond een actueel thema in verband met de taalinfrastructuur. EFNIL is tevens het aanspreekpunt binnen de European Language Data Space en Meta-Net.

IMPACT Centre of Competence

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 sinds 2019 ook CLARIN Knowledge Centre.

European Language Data Space (voorheen ELRC en ELG)

Het INT is het nationale aanspreekpunt binnen de European Language Data Space. Dit is de verderzetting van het vroegere ELRC-initiatief. In 2024 zal het INT in het kader van de European Language Data Space een nationale workshop organiseren.

ALT-EDIC (Alliance for Languages Technologies European Data Infrastructure Consortium)

Het INT verleent haar medewerking aan het ALT-EDIC voorstel en de in-kind bijdrage van Vlaanderen voor de ALT-EDIC door het aanleveren van data voor het trainen van grote taalmodellen e.d. Meer info over het ALT-EDIC voorstel is te vinden op https://european-language-equality.eu/wp-content/uploads/2023/07/MF2023_Session6-2_EDIC.pdf

Elexis Association

In 2023 is het INT lid geworden van de ELEXIS Association die uit het H2020 ELEXIS-project (2018-2022) is voortgekomen. De ELEXIS Association heeft als doel verdere onderzoeksinitiatieven en -activiteiten over lexicografie te bevorderen en te coördineren. Het INT neemt in 2024 deel aan overleggen en initiatieven van de Association.

Nederlandse AI Coalitie

De Nederlandse AI Coalitie is een publiek-private samenwerking, waarbij overheid, bedrijfsleven, onderwijs- en onderzoeksinstellingen en maatschappelijke organisaties samenwerken. De coalitie heeft tot doel de Nederlandse activiteiten in AI te stimuleren, te ondersteunen en waar nodig te organiseren. Het INT is als werkgroeplid bij dit initiatief betrokken en neemt deel aan de overleggen.

Netwerkprojecten

European network for Web-centered linguistic data science (NexusLinguarum, 2019-2024)

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.

Deze COST-actie loopt in het eerste kwartaal van 2024 af. Er wordt nog een grote bijeenkomst georganiseerd.

Universality, diversity and idiosyncrasy in language technology (UniDive, 2022-2026)

Het INT neemt deel aan het Europese onderzoeksnetwerk (COST) UniDive (Universality, diversity and idiosyncrasy in language technology). Het doel van deze COST-actie is om te onderzoeken hoe taaltechnologie verbeterd kan worden door betere kennis van wat talen gemeenschappelijk hebben en van waarin ze zich onderscheiden. Met de actie beoogt men aan de theoretische kant een beter begrip van taaluniversalia te krijgen en, aan de praktische kant, de beschikking te zullen hebben over taaltechnologie die om kan gaan met een grotere verscheidenheid van taalverschijnselen in een groot aantal talen, waaronder talen met weinig middelen en bedreigde talen. Het INT is nauw betrokken bij de activiteiten van WG 2 die zich richt op de lexicon-corpus interface.

Onderzoeks- en infrastructuurprojecten

CLARIAH-Vlaanderen (2021-2025)

Het INT is als derde partij betrokken bij het Vlaamse research infrastructure project *CLARIAH-VL: Advancing the open humanities service infrastructure.* 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.

CLARIAH+ Nederland (2019-2024)

Het vervolgproject van CLARIAH (Common Lab for Research in the Arts and Humanities) loopt van 2019 tot en met 2023. 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 dependency treebanks, hulpmiddelen voor het aanbrengen van persistente gebruikersannotaties in corpuszoekresultaten, een gebruikersvriendelijkere digitalisatieworkflow en curatie van dialectwoordenboekdata. De werkzaamheden zijn grotendeels afgerond in 2023, enige taken die vertraging hebben opgelopen zullen tegen medio 2024 afgerond zijn.

SSHOC-NL (2024-2029)

Het instituut neemt deel aan het SSHOC-NL (Social Science and Humanities Open Cloud for the Netherlands) project (2024-2029). Dit vervolgproject van CLARIAH+ beoogt te komen tot een consortium van onderzoeksinfrastructuren, gericht op het creëren van een ecosysteem van diensten, gegevens en instrumenten voor de sociale en menswetenschappen. Het consortium wordt geleid door ODISSEI, de Nederlandse nationale infrastructuur voor sociale wetenschappen en CLARIAH, de Nederlandse nationale infrastructuur voor geesteswetenschappen. Binnen dit project zal het INT zich onder andere richten op de infrastructuur voor het inzetten van machine learning en AI voor dataverrijking.

SABeD (2021-2024)

Het project Spoken Academic Belgian Dutch, gefinancierd door de KU Leuven, wordt gefinaliseerd in 2024. 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.

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.

Gesproken Corpus van de Zuidelijk-Nederlandse Dialecten (2020-2024)

Het INT is partner in het project Gesproken Corpus van de Zuidelijk-Nederlandse Dialecten, een project dat loopt van 2020 tot 2024 en dat 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/). In 2024 zal in overleg met UGent verder gewerkt worden aan de ontsluiting van de dialectopnames en de bijhorende transcripties.

Projectaanvraag Geparset Corpus van de gesproken Nederlandse Dialecten + (GCND+) (2024-2028)

Het INT is partner in het project *Geparset Corpus van de gesproken Nederlandse Dialecten*, een project dat uitgevoerd zou moeten worden tussen 2024 en 2028 en gerealiseerd wordt aan de UGent. Het project is een vervolg op het hierboven genoemde Gesproken Corpus van de Zuidelijke Nederlandse Dialecten en voorziet in een aanzienlijke geografische uitbreiding waardoor het corpus het hele Europees-Nederlandse gebied zal bevatten. Het project zal de bestaande audio-gealigneerde transcripties en annotaties uit de eerste fase van het GCND gebruiken om ASR- en NLP-tools te hertrainen, om de transcriptie en annotatie van de nieuwe data te versnellen. De rol van het INT wordt om de voor het GCND gebouwde infrastructuur om de data te ontsluiten hierop aan te passen en uit te breiden.

Pilootproject Duidelijke Taal (2023-2024)

Het pilootproject Duidelijke Taal, dat van start gegaan is eind 2023 wordt verdergezet in 2024. Het gaat hier om het aanmaken van datasets omtrent omzettingen van complexe zinnen naar duidelijke taal, die gevalideerd zijn aan de hand van menselijke beoordeling. Het INT doet dit door synthetische datasets aan te maken door middel van prompting voor grote voorgetrainde taalmodellen, zoals ChatGPT en GPT-3.5/4. De uitvoer van deze systemen worden beoordeeld door middel van crowdsourcing, waarbij de crowd zinnen kan beoordelen op drie dimensies: accuraatheid, vlotheid, en complexiteit. Het INT kan deze beoordelingen correleren met linguïstische proxies. Daarnaast krijgen de deelnemers ook de mogelijkheid om complexe zinnen te melden.

Spread the News (2020-2025)

In het onderzoeksproject Spread the new(s). Understanding standardization of Dutch through 17th-century newspapers, gefinancierd door NWO Open Competition SSH en uitgevoerd aan de Radboud Universiteit en het INT, wordt onderzocht welke (socio)linguïstische factoren bepalend zijn bij de functionele implementatie van een standaardtaal. Het INT verzorgt de technische voorzieningen voor dit project, waaronder de ontsluiting en verrijking van een corpus van 17e-eeuwse kranten dat door vrijwilligers is gedigitaliseerd.

Projectaanvraag Signify (2024-2028)

Er werd eind september 2023 een SBO aanvraag ingediend bij het FWO, waar het INT, samen met KU Leuven, UGent en het Vlaamse Gebarentaalcentrum, het consortium vormt. Het doel van Signify is het opzetten van een online collaboratieve multimedia corpusannotatieomgeving om video's met Vlaamse GebarenTaal te transcriberen, annoteren en vertalen naar het Nederlands. Dit door middel van het post-editeren van automatische suggesties geleverd door automatische gebarentaalherkenning en -vertaling. Eventuele goedkeuring en start van het project zijn voorzien in de loop van 2024. Deze aanvraag geldt als Vlaams vervolg op het afgelopen SignON project.

Projectaanvraag SSHOC-VL (2025-2028)

Het INT is betrokken als derde partij bij het nieuw in te dienen Vlaamse voorstel voor een International Research Infrastructuur voor de menswetenschappen. Deze aanvraag geldt als vervolgproject op CLARIAH-VL.

Overige infrastructurele dienstverlening

Etymologiebank

Het INT is verantwoordelijk voor het hosten van de etymologiebank van Nicoline van der Sijs. Het werk aan de etymologiebank wordt voortgezet, vooral met behulp van stagiairs en vrijwilligers van universiteiten in Nederland en België: de etymologiebank wordt met nieuwe woordenboeken en datasets verrijkt en daarnaast wordt gewerkt aan de betere en verdere ontsluiting van de bestaande gegevens. Het is de bedoeling om op termijn de koppeling te maken met de centrale kennisbank.

GLAD

GLAD (Global Anglicism Database Network) is een internationaal project waarin de Engelse invloed op talen wereldwijd wordt geïnventariseerd. Het INT host de database van dit project en op termijn ook de website en levert technische en inhoudelijke bijdragen aan het project.

The digital Palles

The Digital Pallas is een internationaal project waarin 300 concepten in 311 talen zijn opgenomen, gebaseerd op de tweede druk van het Russischtalige Comparative dictionary of all languages and dialects, in alphabetical order, gepubliceerd in 1790-1791 in vier delen en samengesteld door de Pruisische geleerde Peter Simon Pallas. Het INT host sinds 2023 de database van dit project.

Sofeer

Sofeer (https://sofeer.ivdnt.org/) is een digitaal woordenboek van Hebreeuwse en Jiddisje woorden in het Nederlands. Het geeft informatie over spelling, betekenis, uitspraak en herkomst. Het INT host de database van dit project.

DaGeNTa

Vanaf 2023 wordt de website van DaGeNTa (Database Geschiedenis Nederlandse Taalkunde) gehost door INT. Deze database stelt zich als doel historische werken over de Nederlandse taal te ontsluiten. De gegevens kunnen worden verbonden aan de grammaticale websites van het INT. De DaGeNTa-website zal met de hulp van stagiairs en vrijwilligers verder worden uitgebreid.

EenvoudigNL

In samenwerking met Stichting Expertisecentrum Oefenen.nl is in een pilootproject een infrastructuur ontwikkeld om semi-automatisch taaloefeningen te genereren uit corpusmateriaal. De infrastructuur zal afhankelijk van bijkomende financiering verder worden uitgebouwd.

Wat je zegt ben je zelf

In 2023 is door NWA Wetenschapscommunicatie 2021/22 subsidie toegekend voor de uitvoering van het project Wat je zegt ben je zelf. Het INT is verantwoordelijk voor het ontwerpen van de website en een enquêteplatform. De werkzaamheden worden uitgevoerd in 2023 en 2024.

Jiddisch woordenboek

In 2024 zal het INT in het kader van de verduurzaming met subsidie van onder andere Maror het Jiddische woordenboek gaan hosten. De Stichting Jiddische Lexicografie Amsterdam is verantwoordelijk voor dit woordenboek.

9. Disseminatie: onderzoek, onderwijs en het algemene publiek

Aansluitend bij de algemene doelstellingen van het Taalunieverdrag, en in nauwe samenwerking met de Taalunie, levert het INT ook in 2024 zijn bijdrage aan de bevordering van de kennis over de Nederlandse taal en neemt daarbij initiatieven om de taalinfrastructuur voor het Nederlands, en het onderzoek rond de opbouw ervan, bij een ruim publiek bekend te maken.

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 in 2024 zowel binnen samenwerkingsprojecten als door geregelde contacten onderhouden en waar mogelijk geïntensiveerd en uitgebreid. De medewerkers van het INT makenook in 2024 het onderzoek aan het instituut bekend via wetenschappelijke congressen en publicaties. Voor universitaire studenten verzorgt het INT twee verschillende collegereeksen over computationele lexicografie. Aan de Universiteit Leiden gaat het om het vak *Corpus Lexicography* binnen de research master Linguistics en aan de KU Leuven om het vak *Computationele Lexicografie* binnen de master Taalkunde. Er worden ook masterproeven begeleid binnen deze masteropleidingen en binnen de Master of Artificial Intelligence aan de KU Leuven.

De communicatie naar andere onderzoekers en onderzoeksinstellingen wordt in 2024 verder op punt gezet door op de website van het INT een aparte sectie in te richten met informatie op maat voor andere onderzoekers uit het Nederlandse taalgebied en daarbuiten. Op deze webpagina's zal in het Nederlands en het Engels een overzicht gegeven worden van het toegepaste onderzoek aan het INT en van de verschillende manieren waarop andere onderzoekers met het instituut kunnen samenwerken in het kader van nationale of Europese projecten of via onderzoeksstages.

Voor het secundair en het tertiair onderwijs blijft het INT zijn taalmaterialen nog beter toegankelijk maken. In dat verband is het INT aanwezig op en profileert het zich op beurzen, conferenties (HSN-conferentie) en evenementen (Neerlandistiekdagen). 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 in 2024 weer bijgehouden en geregeld uitgebreid.

Ook het algemene publiek wordt niet uit het oog verloren. Net zoals in de voorgaande jaren zal in 2024 minimaal zes keer per jaar een algemene nieuwsbrief verstuurd worden aan geïnteresseerden. Daarnaast is er een nieuwsbrief terminologie die vier keer per jaar verschijnt en die informatie geeft over vaktaal. In 2024 worden er weer regelmatig publieksevenementen georganiseerd, zowel live als digitaal. Met webinars, livestreams van evenementen en berichten op de sociale media Instagram, Facebook, LinkedIn en Twitter brengt het INT in 2024 ook online (de werkzaamheden van) het instituut bij alle doelgroepen onder de aandacht.

Vanaf 2023 heeft het INT niet alleen de technische verzorging maar ook de hosting van Neerlandstiek.nl op zich genomen; tevens wordt het INT in de redactie van Neerlandistiek vertegenwoordigd door Nicoline van der Sijs.

Op vraag van de Taalunie zullen ook in 2024 enkele taalmaterialen worden belicht in het kader van het project ‘benutting van digitale infrastructuur’.