/instituut voor de
Nederlandse taal/

Beleidsplan 2023

najaar 2022

Inhoudsopgave

1. Het INT als kennisinstituut voor het Nederlands.

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 dan ook 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 2023. Dat geeft een overzicht van de activiteiten die voor het eerste jaar van de nieuwe beleidsperiode gepland zijn.
In 2023 komt er een wisseling van de wacht bij de Raad van Toezicht: voorzitter Paul Rüpp beëindigt zijn mandaat en wordt vervangen door Mieke Zaanen. Tegelijk wordt Gert Oostindie die afscheid neemt wegens emeritaat, in de Raad van Advies vervangen door Eric Mijts, verbonden aan de Universiteit van Aruba en daardoor onze contactpersoon voor het Caraïbisch gebied.

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 ondersteunt de ontwikkeling ervan; anderzijds verzamelt het INT ook taalmaterialen en taalsoftware van andere kennisinstellingen en stelt deze samen met de eigen taalmaterialen duurzaam ter beschikking 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 onderzoek en andere activiteiten rond de Nederlandse taal te stimuleren en te ondersteunen. Daarnaast heeft het INT als toegepast onderzoeksinstituut ook de doelstelling de kennis en expertise over taalinfrastructuur verder uit te bouwen door eigen wetenschappelijk onderzoek. Daarbij neemt het deel aan extern gefinancierde nationale en internationale onderzoeks- en infrastructuurprojecten. Deze opdracht is te vergelijken met andere taalinstituten in Europa waarmee het INT ook 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 2023 zijn noodgedwongen eerder conservatief ingepland omdat er niet voldoende mensen en middelen zijn.

In de volgende paragrafen wordt uiteengezet hoe het INT in 2023 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 het Comité van Ministers. Wat de taalinfrastructuur voor het Nederlands zelf betreft, wordt in paragrafen 2 t.e.m. 7 toegelicht hoe het INT in 2023 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 2023 ter beschikking zal stellen en verder uitbouwen onder meer dankzij bijkomende financiering uit competitief verworven middelen. 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 gaat het INT de eerste fase van deze werkzaamheden in, d.w.z. een exploratieve studie van de mogelijkheden die er zijn om deze geïntegreerde implementatie uit te voeren. 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 ook 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 2023 wordt 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 gebruik gemaakt 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. Daarnaast zal het INT als expertisecentrum deel uitmaken van het CLARIN Knowledge Centre for Lexicography, dat vermoedelijk in 2023 opgericht wordt en 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 verderzetten 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 nood 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 2023. Ook in 2023 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, 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 2023 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. Een te verwachten toolkit voor 2023 is *MATEO (MAchine Translation Evaluation Online),* die door UGent en KU Leuven ontwikkeld wordt in het kader van een CLARIN-project.
CLARIN-België is gevraagd om eventueel de CLARIN Annual Conference 2023 te organiseren, en kandidaat hiervoor is de Universiteit Gent. Het INT verleent, als coördinerende partij voor CLARIN-België hieraan vanzelfsprekend zijn medewerking.

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. Een belangrijk deel van de werkzaamheden aan de corpusinfrastructuur in 2023 wordt in eerste instantie uitgevoerd ten behoeve van de verdere uitbouw van de centrale kennisbank voor de Nederlandse woordenschat maar resulteert ook in corpusinfrastructuur voor de brede onderzoeksgemeenschap. Daarnaast is het INT betrokken in diverse projecten waarin corpora worden gebouwd, waarbij het INT, naast expertise, ook infrastructurele ondersteuning biedt voor het bouwen, het gebruik dan wel het ter beschikking stellen van het corpusmateriaal.

Monitorcorpus

Voor het hedendaags Nederlands ligt de focus het komende jaar op het aanvullen van krantenmateriaal in het Corpus Hedendaags Nederlands (CHN) om tot een min of meer evenwichtig monitorcorpus voor dit millennium te komen. Voorts zal aan het CHN nieuw materiaal toegevoegd worden dat via samenwerkingen in extern gefinancierde projecten beschikbaar komt, zoals bijvoorbeeld parlementaire debatten (ParlaMint-project) of hoorcolleges (SaBeD). Daarnaast zal in samenwerking met de Taalunie gewerkt worden aan de uitbreiding van het materiaal uit Suriname en de Antillen.

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

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 en worden deels mede gefinancierd door CLARIAH+.

Verrijking met woordsoort en lemma

De focus zal liggen op de verbetering van de taalkundige verrijking van historisch Nederlands. Daarvoor zal worden onderzocht hoe state-of-the-art, op deep-learning-gebaseerde technieken met behulp van omvangrijker trainingsmateriaal ingezet kunnen worden. Dat betekent dat geïnvesteerd wordt in de conversie en curatie van bestaand trainingsmateriaal en dat ook nieuw trainingsmateriaal zal worden bijgemaakt[1]. Door gebruik te maken van domeinadaptatie[2] zal worden getracht om het manuele werk zoveel mogelijk te beperken.

Syntactische annotatie

Om de mogelijkheden voor het extraheren van informatie over lexicaal combinatiegedrag uit corpusdata te verbeteren zal het Corpus Hedendaags Nederlands syntactisch verrijkt worden. Hierbij denken we minimaal aan verrijking volgens het Universal Dependenciesmodel waarmee we aansluiten bij internationale standaarden en wat in het technisch bereik ligt van onze corpuszoekmachine BlackLab (syntactische uitbreiding zal worden geïmplementeerd in CLARIAH+).[3] Er zal worden onderzocht of de huidige set dependentierelaties voor het Nederlands voldoende aanknopingspunten biedt, of wellicht een aantal taalspecifieke extensies (bijvoorbeeld voor maatcomplementen) nodig zijn.

Metadata

De corpora hebben een gemeenschappelijk metadataformaat, met ruimte voor subcorpus-specifieke metadata. Het INT wil komend jaar toewerken naar een verdere uniformering van het metadatamodel voor historisch en modern corpusmateriaal zodat beide op termijn als één doorlopend diachroon corpus op metadatacategorieën doorzoekbaar kunnen worden.

[1] Uit de ervaringen met het Nederlabproject bleek de noodzaak voor niet alleen een voor diachroon corpusmateriaal geschikte tagset, maar ook voor evaluatie- en trainingsmateriaal om de verrijking van historisch materiaal substantieel te kunnen verbeteren

[2] Domeinadaptatie is het aanpassen van een al beschikbaar, op een bepaalde tekstsoort getraind model aan andersoortig materiaal.

[3] 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.

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. De software is open source beschikbaar.

Voor de ontwikkeling van de corpusapplicatie die bestaat uit de search engine BlackLab[4] en de corpus frontend,[5] 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) staan de volgende werkzaamheden gepland:

- Ondersteuning van steeds grotere corpora door middel van optimalisaties en gedistribueerd zoeken. Dit is de voortzetting van werkzaamheden die in 2022 zijn opgestart.
- Ondersteuning van zoeken met syntactische verrijking[6]
- Uitbreiding van de functionaliteit om efficiënt statistieken uit het materiaal te extraheren. Hierbij denkt het INT met name aan diachrone frequentieprofielen en andere taalvariationele distributies.

Voor wat betreft de userinterface zal worden gewerkt aan:

- Query-building voor syntactische retrieval.
- Onderzoek naar uitgebreidere mogelijkheden voor 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.

4. Beschrijving van de woordenschat door de eeuwen heen

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

[4] https://github.com/INL/BlackLab
[5] https://github.com/INL/corpus-frontend
[6] We zijn ons bewust van het bestaan van zoekengines voor diepe syntactische bomen, zoals GrETEL, PaQu en PML-Tree Query maar deze zijn uiterst langzaam voor het doorzoeken van heel grote hoeveelheden data.

- 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 zal een begin gemaakt worden met beide vernieuwingen waarbij de focus ligt op het conceptuele ontwerp van de infrastructuur en het opstellen van de vereisten voor de databanken, de workflows, de software en de hardware. De reguliere werkzaamheden (updates en nieuwe content) aan de bestaande infrastructuur en lexicografische producten worden in dit eerste jaar van het Meerjarenbeleidsplan voortgezet. Daarnaast worden tests voor de vernieuwde en geïntegreerde lexicografische processen opgezet. In wat volgt bespreken we kort (a) het geplande conceptuele werk aan het datamodel voor de centrale kennisbank, (b) de geplande vernieuwingen in de corpusgebaseerde workflow voor het centrale lexicon, (c) de plannen voor een eerste koppeling op betekenisniveau via een betekenisregister en ten slotte (d) de manier waarop de verdere ontwikkeling van lexicografische eindproducten alvast voor een aantal beschrijvingscomponenten vanuit de centrale kennisbank zal gebeuren.

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. Bijzondere aandacht gaat in 2023 naar de classificatie en modellering van meerwoordsexpressies en naar compatibiliteit van zowel de hedendaagse als historische lexicale beschrijving met de Tagset Diachroon Nederlands (TDN). Voor

de TDN is al een uitgebreide beschrijving beschikbaar en richtlijnen voor de morfosyntactische verrijking van het GiGaNT-lexicon en de lemmatiseerprincipes zijn al gepubliceerd. Deze documenten zullen in 2023 een update krijgen.

Het Centrale Lexicon (GiGaNT)

In 2023 worden in eerste instantie de reguliere werkzaamheden aan het centrale lexicon voortgezet. Deze houden in: het onderhoud aan de bestaande modules, het verder werken aan 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 nodig zal GiGaNT in overeenstemming gebracht worden met het datamodel van de kennisbank.

Daarnaast zal vanuit GiGaNT een eerst testcase opgezet worden voor het uitrollen van de vernieuwde corpusgebaseerde lexicografische workflow. Met name wordt een infrastructuur uitgewerkt 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 wordt hierbij het hedendaags Nederlands (Molex). Het doel is om het CHN, dat wekelijks geüpdatet wordt, systematisch te screenen ten behoeve van de uitbreiding van Molex, en daarbij frequentie-informatie zoals tijdreeksen en regionale distributies aan Molex toe te voegen 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 woordenschatsbeschrijving in de komende jaren. Zoals hierboven aangeven zal bij het opstellen van het datamodel voor de centrale kennisbank in 2023 veel aandacht gaan naar het betekenisregister dat die koppeling op betekenisniveau moet mogelijk maken. Het ontwerp en het op punt stellen van het datamodel zal noodzakelijkerwijze een proces van trial and error zijn. Daarom zal in 2023, parallel met het opstellen van het datamodel, ook een eerste proefversie van het betekenisregister gemaakt worden voor een testset van woorden. Die proefversie wordt opgebouwd vanuit het hedendaags Nederlands en gaat daarbij onder andere uit van de definities uit bronnen die tijdens de afgelopen beleidsperiode op lemmaniveau aan GiGaNT-Molex gekoppeld zijn (ANW, WNW, Referentiebestand Nederlands (RBN), Vertaalwoordenschat en, via de koppeling met Hilex, het Woordenboek der Nederlandsche Taal (WNT)). Op grond hiervan worden koppelbetekenissen vastgesteld, met persistent identifier, die worden gebruikt als aanknopingspunt voor koppeling op semantisch niveau binnen de kennisbank. Voor de hedendaagse lexicografische bronnen zal een koppeling van woorden en betekenissen uit de testset uitgeprobeerd worden om de mogelijkheid van bulkoperaties te onderzoeken en in te schatten. Voor de koppeling van de historische woordenboeken aan het betekenisregiser zal met een kleinere testset geëxperimenteerd worden, vertrekkend vanuit DiaMaNT (het Diachroon seMantisch lexicon van de

Nederlandse Taal). De problemen en uitdagingen die bij de experimenten naar boven komen zullen dan de uitwerking van het datamodel informeren.

Daarnaast zal een eerste workflow opgezet worden om de beschrijving van neologismen, die inherent een eenvoudigere semantische structuur hebben, meteen aan het betekenisregister te koppelen en dus binnen de centrale kennisbank te integreren. Van hieruit kan deze lexicografische informatie dan gepubliceerd worden in het Woordenboek van Nieuwe Woorden.

Lexicografische eindproducten, API's en datasets

Zoals het Meerjarenbeleidsplan 2023-2027 aangeeft, zullen 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 worden. Hoewel de huidige productspecifieke workflows in 2023 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 (lijst met de officiële spelling) 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 2023 zullen de systematische updates van woordenlijst.org hervat worden en beschikbaar zijn via de vernieuwde applicatie.

Algemeen Nederlands Woordenboek (ANW)

Zoals in de vorige paragraaf aangegeven, zullen in 2023 voor een testset een aantal hedendaagse lexicografische bronnen op betekenisniveau aan elkaar gekoppeld worden. Dat zal dan toelaten om in de loop van 2023 uit te proberen hoe die gekoppelde betekenisbeschrijvingen en andere lexicografische informatie uit de centrale kennisbank in het ANW gepubliceerd kunnen worden. Daarbij zal met name gekeken worden hoe de consistentie van de betekenisbeschrijving in het ANW gegarandeerd kan worden.

Woordenboek van Nieuwe Woorden (WNW)

Het WNW zal in eerste instantie nog de productspecifieke workflow aanhouden maar in de loop van 2023 zal deze wel gevoed kunnen worden vanuit de centrale kennisbank dankzij de vernieuwde corpusgebaseerde workflow voor de identificatie van nieuwe woorden (cf. supra). Bovendien zal de lexicografische behandeling van neologismen in de loop van het jaar in overeenstemming gebracht worden met de modulaire opbouw en het datamodel van de centrale kennisbank, zodat het mogelijk wordt om het WNW in de komende jaren als afgeleid product vanuit de centrale kennisbank verder te ontwikkelen.

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 2023 zal het project uitgebreid worden met voorbeeldzinnen en combinaties voor substantieven uit het frequentiewoordenboek. Daarnaast zullen meer patronen beschreven worden bij werkwoorden. Met een aangepaste werkomgeving zal het INT idiomen met substantieven en werkwoorden efficiënter kunnen bewerken. Ook zal de nodige aandacht besteed worden aan het toegankelijker maken van de idiomen in de gebruikersapplicatie.

Historische woordenboeken

Voor de historische woordenboeken zal de focus liggen op de verdere verbetering en ontleding van de gedigitaliseerde woordenboekdata, wat niet alleen ten goede komt aan GiGaNT maar ook aan het betekenisregister. Voor het onderhoud en de synchronisatie met de diverse modules in de kennisbank zal een nieuwe werkomgeving ingericht worden.

Vertaalwoordenschat

De Vertaalwoordenschat is een platform waarop tweetalige bestanden worden ontsloten die in de jaren 90 en begin 2000 ontwikkeld werden 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.

In 2022 zijn in een pilot inhoudelijke verbeteringen aangebracht in de bestanden Nieuwgrieks-Nederlands/Nederlands-Nieuwgrieks die in een grote data-update van de Vertaalwoordenschat in het eerste kwartaal van 2023 officieel wordt gereleaset samen met de lancering van het taalpaar Nederlands-Deens/Deens-Nederlands. Daarnaast zal in 2023 de lancering van een volgend taalpaar (Nederlands-Noors/Noors-Nederlands) worden voorbereid en zal de informatie in de gebruikslabels in de vertaalwoordenschatbestanden worden geharmoniseerd. Verder zullen de bestanden van de taalparen die al online staan, klaargezet worden in de redactieomgeving die in 2022 is opgeleverd, zodat redacteuren op vrijwillige basis inhoudelijke verbeteringen kunnen aanbrengen. Ten slotte zullen de Nederlandse glossen uit de vertaalwoordenschat meegenomen worden in de tests voor het koppelen op betekenisniveau in de centrale kennisbank.

API's en datasets

De API voor de centrale kennisbank zal in 2023 verder ontwikkeld worden vanuit de al bestaande lexiconservice, die nu al toegang biedt tot informatie in GiGaNT. In eerste instantie zal worden onderzocht of nieuwe informatie uit GiGaNT (lemma-IDs en frequentie-informatie) via de API op een schaalbare manier opvraagbaar gemaakt kan worden. Diezelfde informatie zal ook in de nieuwe release van GiGaNT-Molex als dataset in de language repository ter beschikking komen.

5. Beschrijving van de Nederlandse dialecten

Als logische uitbreiding van de opdracht om de Nederlandse woordenschat in al haar facetten te beschrijven, hebben ook de dialectwoordenboeken uit het Nederlandse taalgebied sedert 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 2023 uitgebreid met nieuwe dialectwoordenboeken met behulp van vrijwilligers, die het materiaal voorbereiden.

Database van de Zuidelijk-Nederlandse Dialecten (DSDD)

Eind 2022 bevatte de Database van de Zuidelijk-Nederlandse Dialecten (DSDD) ongeveer 30.000 concepten. In 2023 worden met de hulp van vrijwilligers nog enkele lacunes voor de Vlaamse, Brabantse en Limburgse dialecten aangevuld waar mogelijk. Daarnaast zal verder gewerkt worden aan een strategie om semasiologische woordenboeken te kunnen toevoegen aan de onomasiologisch opgezette DSDD aan de hand van het woordenboek van de Zeeuwse dialecten. Tot slot zal er onderzocht worden hoe het platform verder uit te bouwen is tot een dialectplatform voor het hele Nederlandse taalgebied.

Digitale infrastructuur voor het Bildts

In 2023 zal het INT beginnen met het opzetten van een digitale infrastructuur voor streektalen en dialecten aan de hand van een pilotproject voor het Bildts. Deze infrastructuur biedt streektaalorganisaties/dialectverenigingen de nodige ondersteuning om 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.

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 gebruik. Een webrubriek zal op die basis worden toegevoegd wanneer een voldoende substantieel aantal historische termenlijsten beschikbaar is.

Tools

Qua terminologietools dient in 2023 een nieuwe tool te worden ontwikkeld en geïmplementeerd. Er is in 2022 al gewerkt aan een prototype, dat begin 2023 kan worden getest en dan verder uitgewerkt. 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. Ook het ENT-advies wordt meegenomen.

Veldondersteuning

Klassieke veldondersteuning in de vorm van updates van de bestaande websiterubrieken, vooral de opleidingen in Nederland en Vlaanderen, blijft aan de orde. De evenementenrubriek voor terminologie zal continu worden bijgehouden en ook in 2023 worden vier uitgebreide nieuwsbrieven Terminologie gedistribueerd.

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. Verder houdt het INT de koppeling van dit project aan internationale samenwerking, zoals met EURAC (Bolzano) in overweging. 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 2023 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 een kennisgeving van goedkeuring of afwijzing volgt in het voorjaar van 2023.

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 ook 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 in oktober 2022 online beschikbaar gesteld via een nieuwe applicatie gebouwd bij het INT.
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), zal gewerkt worden aan een geïntegreerd grammaticaportaal, een webpagina die de spil moet 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.

e-ANS

In 2023 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 ook 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 2023 zal de conversie van het nieuw syntaxis-deel ‘Coordination’ van Hans Broekhuis afgerond worden en er komt een update van het Taalportaal. Hierbij zal zowel de inhoud als de webapplicatie worden bijgewerkt.

Grammaticaportaal

Ten slotte wordt in de loop van 2023 een eerste publieke testversie van het Grammaticaportaal gelanceerd, de site die als startpagina moet gaan dienen voor de verschillende grammatica-applicaties binnen het INT: de e-ANS, het Taalportaal, Taaladvies.net, maar ook projecten als Dagenta (cf. infra) en Woordcombinaties (cf. supra).

8. Nationale en internationale samenwerkingsverbanden

Netwerken

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. Het Centre organiseert in 2023 een nieuwe DATECH-conference[7]. De werkzaamheden m.b.t. digitalisering die onder andere in de context van CLARIAH+ worden uitgevoerd, gebeuren in samenwerking met het IMPACT Centre of Competence.

European Language Data Space (voorheen ELRC en ELG)

Het INT was betrokken bij het ELRC-initiatief (European Language Resource Coordination) dat tot doel had 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. In 2023 wordt ELRC opgeheven (18 januari 2023) en op 19 januari gaat het initiatief over in de European Language Data Space. Ook binnen dit initiatief blijft het INT het nationale aanspreekpunt. Eenzelfde integratie is gepland voor de data die binnen het European Language Grid (ELG) verzameld werd en waar het INT eveneens bij betrokken was.

Elexis Association

De ELEXIS Association bouwt voort op het netwerk dat tijdens het ELEXIS-project (2018-2022) ontstaan is en heeft als doel verdere onderzoeksinitiatieven en -activiteiten over lexicografie te bevorderen en te coördineren. Het INT neemt in 2023 deel aan het overleg om deze associatie vorm te geven.

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 in 2023 deel aan de overleggen.

[7] https://datech.digitisation.eu/

Netwerkprojecten

European network for Web-centered linguistic data science (NexusLinguarum, 2019-2023)

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 2023 omvatten werkvergaderingen, conferenties en workshops, training schools, STSM’s (Short Term Scientific Missions) en andere evenementen.

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. De activiteiten van de actie voor 2023 omvatten werkvergaderingen, conferenties en workshops, training schools, STSM’s (Short Term Scientific Missions) en andere evenementen

Onderzoeks- en infrastructuurprojecten

CLARIAH-Vlaanderen (2021-2024)

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.

Via CLARIAH-VL is het INT betrokken als dataleverancier bij een Tier-I project bij het Vlaams Supercomputer Centrum om contextuele taalmodellen (o.a. SpanBERT) te trainen op basis van de corpora hedendaags Nederlands waarover het INT beschikt.

CLARIAH+ Nederland (2019-2023)

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 worden afgerond in 2023.

SSHOC-NL (aangevraagd)

Het instituut heeft meegewerkt aan de SSHOC-NL (Social Science and Humanities Open Cloud for the Netherlands) infrastructuuraanvraag. 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, indien gehonoreerd, zal het INT zich onder andere richten op de infrastructuur voor het inzetten van machine learning en AI voor dataverrijking.

ParlaMint II (2021-2023)

ParlaMint is een door CLARIN-ERIC gefinancierd project, dat bijdraagt aan de totstandkoming van vergelijkbare en uniform geannoteerde meertalige corpora van parlementaire zittingen. ParlaMint I creëerde en maakte corpora voor 17 talen beschikbaar. ParlaMint II zal het XML-schema en de validatie verbeteren, de bestaande corpora uitbreiden tot ten minste juli 2022, corpora voor nieuwe talen toevoegen, de corpora verder verbeteren met extra metadata en de bruikbaarheid van de corpora verbeteren. Het INT is verantwoordelijk voor de data van het Belgisch federaal parlement. In 2023 wordt het werk aan dit corpus afgerond.

SignON (2021-2024)

Het INT is als consortiumpartner 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, Spaanse, Britse en Ierse Gebarentaal. Gesproken talen zijn het Nederlands, het Spaans, het Iers en het Engels. 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, en het verzamelen van gebarentaalcorpora. Dit heeft al geleid tot een aantal publicaties en resources die beschikbaar gesteld worden, zowel voor VGT als voor NGT. Deze inspanningen worden in 2023 verdergezet.

SABeD (2021-2023)

Het project *Spoken Academic Belgian Dutch*, gefinancierd door de KU Leuven wordt gefinaliseerd in 2023. 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. Verder verleent het INT ook in 2023 zijn medewerking bij het annotatieproces (expertise, voorbewerking video-data, omzetten van pdf's en slideshows naar tekstbestanden e.d.).

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 2023 zal in overleg met UGent gewerkt worden aan de ontsluiting van de dialectopnames en de bijhorende transcripties.

Pilootproject Duidelijke Taal (2023-2024)

Op vraag van de Taalunie wordt in 2023 een pilootproject opgezet omtrent automatische omzetting van documenten naar eenvoudige taal. De pilot leidt tot een demo-systeem waarbij gebruik gemaakt wordt van state-of-the-art technieken uit de artificiële intelligentie. Dit project zal lopen vanaf het najaar 2023 tot begin 2024.

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.

Clasabed (2022-2023)

Het CLARIAH-project Clasabed (Clariah.nl tools in SABeD) behelst de evaluatie van diverse tools uit de CLARIN-infrastructuur op hun inzetbaarheid voor de annotatie en analyse van de corpusdata van het hierboven vermelde project SABeD.

Using CoBaLT and GaLAHaD for historical corpus annotation (2023)

In het CLARIAH-project Using CoBaLT and GaLAHaD for historical corpus annotation zullen CoBaLT, een tool voor interactieve corpusannotatie, het GaLAHaD-platform voor taalkundige annotatie van historisch Nederlands, en diverse tools voor het taggen en lemmatiseren van historische teksten geëvalueerd worden.

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.

DaGeNTa

In 2023 zal INT de website van DaGeNTa (Database Geschiedenis Nederlandse Taalkunde) hosten. 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.

9. Disseminatie: onderzoek, onderwijs en het algemene publiek

Aansluitend bij de algemene doelstellingen van het Taalunieverdrag, en in nauwe samenwerking met de Taalunie, zal het INT ook in 2023 zijn bijdrage leveren aan de bevordering van de kennis over de Nederlandse taal en daarbij initiatieven nemen 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 ook in 2023 zowel binnen samenwerkingsprojecten als door geregelde contacten onderhouden en waar mogelijk geïntensiveerd en uitgebreid. De medewerkers van het INT zullen ook in 2023 het onderzoek aan het instituut bekendmaken 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 master-opleidingen en binnen de Master of Artificial Intelligence aan de KU Leuven.

Voor het secundair en het tertiair onderwijs zullen de taalmaterialen van het INT in 2023 nog beter toegankelijk gemaakt worden. 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. 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 ook in 2023 bijgehouden en geregeld uitgebreid.

Ook het algemene publiek wordt niet uit het oog verloren. Net zoals in de voorgaande jaren zal in 2023 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. Ook in 2023 zullen er regelmatig publieksevenementen georganiseerd worden, zowel live als digitaal. Met webinars, livestreams van evenementen en berichten op de sociale media Instagram, Facebook, LinkedIn en Twitter brengt het INT ook in 2023 online (de werkzaamheden van) het instituut bij alle doelgroepen onder de aandacht.