/instituut voor de
Nederlandse taal/

Meerjarenbeleidsplan 2023-2027

Instituut voor de Nederlandse Taal

Beknopte samenvatting

Het Instituut voor de Nederlandse Taal (INT) heeft zich in de afgelopen beleidsperiode succesvol omgevormd naar een breed opgezet kennisinstituut voor het Nederlands. Die transitie werd vastgelegd in het vorige meerjarenbeleidsplan 2018-2022 en in goed overleg met de Taalunie uitgevoerd. Dankzij nauwgezette projectplanning en begrotingsopvolging zijn de doelstellingen voor de vorige beleidsperiode dan ook over het geheel genomen verwezenlijkt. Dat heeft zich in 2021 vertaald in een positieve beoordeling van het INT door de externe visitatiecommissie. In dit nieuwe meerjarenbeleidsplan voor de periode 2023-2027 zet het INT uiteen hoe het zijn rol als het kennisinstituut voor de Nederlandse taal, met een focus op de digitale taalinfrastructuur, verder wil uitbouwen. Dit beleidsplan neemt de aanbevelingen en inzichten uit het visitatie- en zelfevaluatierapport ter harte en sluit eveneens aan bij de krachtlijnen van de Taalunie rond digitalisering, internationalisering en inclusie/diversiteit.

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 en zo praktische uitvoering te geven aan een aantal centrale verdragstaken van de Taalunie[1]. Concreet betekent dit dat het INT enerzijds zelf corpusdata, linguïstische databanken en taalsoftware maakt 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 ook de doelstelling de kennis en expertise over taalinfrastructuur verder uit te bouwen door eigen wetenschappelijk onderzoek. Daarbij neemt het ook deel aan extern gefinancierde nationale en internationale onderzoeks- en infrastructuurprojecten. Globaal is deze opdracht te vergelijken met andere instituten in Europa[2] waarmee het INT intensief samenwerkt in Europese projecten en netwerken. Net als gelijkaardige instituten in binnen- en buitenland, en zoals aanbevolen door de visitatiecommissie, hoopt het INT op vrij korte termijn rechtstreekse toegang te krijgen tot competitieve NWO-middelen

In de plannen voor de komende vijf jaar zal de focus liggen op de verdere integratie van een aantal componenten in de digitale taalinfrastructuur en het daarvoor noodzakelijke onderzoek:

- Van twee naar één gebruiksvriendelijke repository voor Nederlandse taaldata en taalsoftware met bijbehorend expertisecentrum en maximaal ingebed in de Europese taalinfrastructuur;
- Van aparte lexicale databanken naar één centrale, modulair opgebouwde kennisbank voor de Nederlandse woordenschat van waaruit eindproducten (verder) ontwikkeld worden;
- Van losse websites over grammatica naar één grammaticaportaal;
- Van aparte databanken naar één dialect- en streektaalplatform.

In de komende beleidsperiode blijft het INT taalinfrastructurele ondersteuning en dienstverlening bieden aan specifieke doelgroepen, zoals taalprofessionals (terminologen en vertalers), het onderwijsveld, onderzoekers in de digital humanities of gebarentaligen. Ten slotte zal het INT zich blijven inzetten om de digitale taalinfrastructuur van het Nederlands maximale zichtbaarheid te verlenen met gerichte communicatie naar diverse doelgroepen zoals onderzoekers, actoren uit de taalsector en het brede taalgeïnteresseerde publiek.

[1] zie Taalunieverdrag, Hoofdstuk 1, Artikelen 2, 3, 4 en 5
[2] o.a. Leibniz-Institut für Deutsche Sprache, Dansk Sprognævn, Eesti Keele Instituut, Norsk Språkrådet

Inhoudsopgave

1 Het INT als kennisinstituut voor het Nederlands

Als kennisinstituut focust het INT op de ontwikkeling, het onderzoek, de duurzame beschikbaarstelling en de ondersteuning van de digitale taalinfrastructuur voor het Nederlands. Het instituut geeft daarmee concrete uitvoering aan de verdragstaken van de Taalunie (Taalunieverdrag, hoofdstuk 1) door het ontwikkelen en beschikbaar stellen van lexica met de officiële spelling (art. 4b), woordenboeken, woordenlijsten en grammatica's met oog voor taalvariatie (art. 4b en 4d), databanken voor terminologie (art. 4c en 5e), door de ondersteuning van onderwijs (art. 5b en 5c) en het wetenschappelijk onderzoek over de Nederlandse taal (art. 5a), zowel binnen het taalgebied zelf als in het buitenland (art. 3d) en door de participatie in Europese taalgerelateerde initiatieven (art. 4f en 4g). In wat volgt, bespreken we eerst wat een moderne taalinfrastructuur precies inhoudt, wat de rol van het INT is en hoe die de activiteiten en de onderzoeksfocus van het instituut bepaalt.

Het onderzoek naar het Nederlands en de ontwikkeling van taalapplicaties voor het Nederlands hoeven gelukkig niet van nul te beginnen. Net als in andere kennissectoren bouwt de R&D in de taalsector ook voort op een onderliggende digitale *infrastructuur*. Die taalinfrastructuur bestaat uit:

1. **primaire data**: taalgebruik dat representatief is voor bepaalde taalvariëteiten, afgeleid is uit ruwe data (tekst/audio/video), in corpora is verzameld, van metadata is voorzien en taalkundig is verrijkt
2. **secundaire data**: gesystematiseerde informatie en kennis over taal, typisch afgeleid uit primaire data en in de vorm van gestructureerde[3] linguïstische databanken (lexicaal, grammaticaal…)
3. **Software**: tools en andere software om primaire en secundaire data te compileren, te beheren, te verrijken, te doorzoeken, er informatie uit te extraheren en verder te ontsluiten.
4. **Online services** die de taaldata en taalsoftware terugvindbaar en duurzaam beschikbaar maken en er ondersteuning voor bieden.

Wat precies onder digitale taalinfrastructuur valt, verandert met de tijd: elektronische corpora en computationele lexica bestaan al sinds de jaren zestig terwijl deep-learning-gebaseerde taalmodellen dan weer vrij recent zijn. Essentieel is wel dat deze data, software en services niet op zichzelf beschouwd worden maar als schakels in een waardeketen binnen een ruimer ecosysteem van taalgerelateerde R&D. Taaldata en taaltechnologie functioneren pas als deel van een *infrastructuur* wanneer anderen er vlot gebruik van kunnen maken om zelf aan onderzoek en ontwikkeling te doen.

Binnen dit ecosysteem speelt het INT enerzijds een van de vele actoren die elk in hun specifieke deeldomein aan taalinfrastructuur-opbouw doen. Voor het INT gaat het daarbij, in uitvoering van de verdragstaken van de Taalunie, voornamelijk om de aanleg van corpora en lexicale databanken voor de historische en hedendaagse variëteiten van het Nederlands en de ontwikkeling van de daarvoor nodige taaltechnologie en software. Anderzijds bekleedt het instituut als structureel gefinancierd taalinfrastructuurinstituut echter ook een unieke positie: terwijl andere kennisinstellingen binnen de taalsector typisch aan datacompilatie en softwareontwikkeling doen in het kader van tijdelijke projecten, heeft het INT vanuit de verdragstaken van de Taalunie ook de opdracht om de taalinfrastructuur voor het Nederlands op de lange termijn te verzekeren. Een aantal basisonderdelen

[3]Gestructureerd verwijst hier naar het verschil dat binnen de datawetenschap gemaakt wordt tussen ongestructureerde en gestructureerde data: corpora bevatten ongestructureerde data in de zin dat informatie en kennis over taal alleen impliciet aanwezig. Gestructureerde data daarentegen maakt informatiecategorieën expliciet.

van de infrastructuur, zoals een up-to-date corpus hedendaags Nederlands of de inventarisatie van nieuwe woorden, vergen noodzakelijkerwijze een continue uitbouw i.p.v. een tijdelijke en projectgebaseerde. Ook de duurzame beschikbaarheid van de taalinfrastructuur kan alleen door een structureel gefinancierd kennisinstituut als het INT gegarandeerd worden. Meer algemeen veronderstelt expertiseopbouw inherent een langetermijnperspectief en dit geldt bij uitstek ook voor de expertise over taalinfrastructuur. Tabel 1 toont hoe de activiteiten van het INT binnen dit taalinfrastructurele kader in een vijftal categorieën uiteenvallen. De kolommen geven aan hoe die activiteiten gefinancierd worden, in welke fase van de infrastructuuropbouw ze gesitueerd zijn, wat de rol van het INT bij de productie is, en welk type R&D ze vertegenwoordigen.

Tabel 1: Taalinfrastructurele activiteiten van het INT

#	**Activiteitencategorie**	**Financiering**	**Fase**	**Rol INT**	**R&D-type**
1	Structurele en continue ontwikkeling van specifieke basiscomponenten van de Nederlandse taalinfrastructuur (o.a. corpusdata, lexicale databanken, corpustools)	structureel	productie nieuwe taalinfra-structuur	producent	toege-past
2	(Deelname in) *projectgebaseerde* taalinfrastructuur-opbouw (data en tools)	tijdelijk, op projectbasis	productie nieuwe taalinfra-structuur	(co)-producent	toege-past
3	Duurzame terbeschikkingstelling van de taalinfrastructuur voor het Nederlands via Language Resources Repository's	structureel	verspreiding bestaande producten	depot en distri-buteur	toege-past
4	Uitbouw van expertise over taalinfrastructuur (o.a. datamodellen, standaarden voor interoperabiliteit en kwaliteitscontrole, schaalbaarheid, interconnectiviteit met internationale taalinfrastructuur, opvolgen van state of the art in taaltechnologie)	gemengd: structureel en op projectbasis	voorbereiden toekomstige infrastructuur	expertise-centrum	basis-onder-zoek

5	Dienstverlening gebaseerd op de taalinfrastructuurexpertise	gemengd: structureel en op projectbasis	ondersteu-ning van productie en publicatie door derden	consul-tancy en service-provider	toege-past

Er is een sterke wisselwerking tussen de verschillende activiteitencategorieën maar ze hebben telkens ook specifieke eigen kenmerken. Zo komen de data en software die binnen categorie 1 en 2 ontwikkeld worden in de repository (categorie 3) terecht. Daar worden ze samen met taalproducten die door derden ontwikkeld werden ter beschikking gesteld aan gebruikers en meteen ook ingebed in de Europese taalinfrastructuur. De expertise over taalinfrastructuur (cat. 4) vloeit voor een groot deel voort uit de ervaring met taalinfrastructuurontwikkeling (cat. 1 en 2) en terbeschikkingstelling (cat. 3), maar wordt ook actief uitgebouwd door eigen onderzoek en publicaties en door deelname in nationale en internationale netwerken en onderzoeksprojecten over taalinfrastructuur (zie paragraaf 8). Die expertise betreft niet alleen de werkprocessen om kwaliteitsvolle taaldata en taalsoftware te maken, maar ook de randvoorwaarden waaronder data en software daadwerkelijk als een performante *infrastructuur* kunnen functioneren (terugvindbaarheid, interoperabiliteit, schaalbaarheid etc.). Het is deze gecombineerde expertise die het INT weer toepast in de eigen productie (cat.1), inbrengt in projecten (cat.2), aanwendt in de repository-activiteiten (cat.3) en waarop door derden als dienstverlening een beroep kan worden gedaan (cat.5).

In de volgende jaren zal het INT zijn voor het Nederlandse taalgebied unieke expertise systematisch verder uitbouwen door eigen wetenschappelijk onderzoek naar verschillende aspecten van de taalinfrastructuuropbouw. Die wetenschappelijke focus wordt ondersteund door continue bijscholing en interne kennisuitwisseling en zal zich weerspiegelen in het personeelsbeleid voor de komende jaren. De wetenschappelijke activiteiten worden deels structureel gefinancierd. Deze structurele financiering wordt aangewend om de belangrijke verdragstaken van de Taalunie uit te voeren. Daarnaast participeert het INT in nationale en internationale projecten en onderzoeksnetwerken via competitief verworven onderzoeksmiddelen (zie paragraaf 8). Op die manier wil het INT een significante bijdrage leveren tot het blijvend verzekeren van een kwalitatief hoogstaande taalinfrastructuur voor het Nederlands. De benchmark hiervoor is dubbel. Enerzijds is er de positie van het Nederlands tegenover andere talen, waar het INT meewerkt aan de Europese doelstellingen voor taalgelijkheid in het digitale tijdperk (resolutie 2018/2028 van het Europees Parlement) en mee de situatie voor het Nederlands in kaart gebracht heeft in een beleidsvoorbereidend rapport (Steurs et al. 2022) voor het project European Language Equality (zie paragaaf 8.1). Hiermee geeft het INT ook uitvoering aan de verdragstaken van de Taalunie met betrekking tot de positie van het Nederlands op Europees niveau (Taalunieverdrag art. 4f en 4g). Anderzijds is er ook de diversiteit aan taalinfrastructurele noden *binnen* de Nederlandse taalgemeenschap die afgedekt moet worden. Daar sluit het INT aan bij de diversiteits- en inclusiedoelstellingen van de Taalunie en wil het instituut een brede waaier aan doelgroepen bereiken met taalinfrastructuur en bijhorende dienstverlening op maat.

In de volgende paragrafen worden de verschillende taalinfrastructurele activiteiten per deeldomein in detail behandeld en worden telkens de plannen voor de komende beleidsperiode voorgesteld. Paragraaf 2 behandelt eerst de Language Resource Repository-activiteiten die de voor taalinfrastructuur essentiële duurzame beschikbaarheid realiseren. Paragraaf 3 bespreekt de ontwikkeling van primaire taaldata (corpora). Paragrafen 4 tot 7 beschrijven de plannen voor de ontwikkeling van secundaire datatypes op het gebied van woordenschat, dialecten en terminologie en

grammatica alsook de bijhorende dienstverlening. Paragraaf 8 behandelt de nationale en internationale netwerken en projecten waarbinnen het INT aan infrastructuur- en expertiseopbouw doet en geeft ook een overzicht van de infrastructurele dienstverlening die het INT verzorgt. Paragraaf 9 vat samen welke interne infrastructuuruitbreiding (hardware, software en IT-expertise) het INT voorziet om zijn taalinfrastructuuropdracht te kunnen waarmaken. Paragrafen 10 en 11 bespreken de activiteiten die niet strikt taalinfrastructureel van aard zijn maar die wel voortbouwen op de taalinfrastructuuropdracht, namelijk onderwijs en pr & communicatie. Paragraaf 12 rondt af met een overzicht in tabelvorm van de geplande werkzaamheden en deliverables.

2 Language Resources Repository en expertisecentrum

Een essentiële rol die het INT vervult voor de taalinfrastructuur van het Nederlands is het garanderen van de duurzame beschikbaarheid van corpusdata, linguïstische databanken en taaltechnologie, of die nu ontwikkeld werden door het INT zelf of door andere kennisinstellingen. Deze structureel gefinancierde functie als Language Resources Repository heeft een dubbele ontstaansgeschiedenis. Enerzijds is de *Taalmaterialen*-catalogus (https://taalmaterialen.ivdnt.org/) ontstaan uit de overdracht van de TST-Centrale (Taal- en SpraakTechnologie) van de Taalunie naar het INT. Anderzijds is het INT ook een CLARIN-B-centrum voor Nederland en Vlaanderen met een CLARIN-portal waarlangs taalsoftware en taaldata binnen het Europese CLARIN-netwerk beschikbaar gemaakt kunnen worden (https://portal.clarin.ivdnt.org/). De CLARIN-centra zijn ontstaan uit het European Research Infrastructure Consortium CLARIN (Common Language Resources and Technology Infrastructure) om een duurzame terugvindbaarheid en beschikbaarheid van taaldata en software te waarborgen voor onderzoek op Europees niveau (zie ook paragraaf 8 over de Europese netwerken). Deze twee repository's, Taalmaterialen en CLARIN Portal, richten zich deels op een ander doelpubliek: Taalmaterialen is er zowel voor onderzoekers, bedrijven als het brede publiek terwijl CLARIN zich vooral op onderzoekers richt. De aangeboden materialen vertonen echter ook een sterke overlap, waardoor soms verwarring ontstaat voor gebruikers. In de komende beleidsperiode zullen beide repository's dan ook sterker geïntegreerd worden en vanuit één gebruikersinterface met betere zoekmogelijkheden toegankelijk gemaakt worden. Daarbij zal wel verzekerd worden dat de dienstverlening aan de verschillende doelgroepen bewaard blijft en het zal met vernieuwde en duidelijkere licenties, die maximaal aansluiten bij internationale standaarden, nog inzichtelijker worden voor gebruikers wat men al dan niet met de data en tools mag doen.

Aan de toeleveringskant zullen de depositieprocedures voor leveranciers van taaldata en taalsoftware verder gestroomlijnd en vereenvoudigd worden. Ook hier zullen de depositieprocessen voor Taalmaterialen en het CLARIN Portal maximaal geïntegreerd worden met behoud van dienstverlening aan de diverse doelgroepen. Op het vlak van hardware (storage, memory en processing power) zullen de repository's voorbereid worden op het toenemend belang van multimediale content en tools die meer rekenkracht vergen (zie paragraaf 9). Bij de uitbreiding van de repository's met nieuwe materialen neemt het INT niet louter een passieve houding aan. Er wordt actief contact gelegd met onderzoekers om na te gaan of interessante datasets via de repository's beschikbaar gesteld kunnen worden, of aangelegd kunnen worden door het INT in diverse projecten. Hierbij geeft het INT speciale aandacht aan de diversiteits- en inclusiedoelstellingen van de Taalunie met gerichte initiatieven naar resources voor de ondersteuning van gebarentaal, eenvoudige taal[4] en tweedetaalsprekers of voor de studie en remediëring van taalbeperkingen (dyslexie, afasie, laaggeletterdheid).

[4]Ook wel duidelijke taal genoemd.

Naast de uitbouw van de repository's die het INT zelf beheert, zijn er de komende jaren ook ontwikkelingen te verwachten in de nationale en Europese taalinfrastructuur (zie paragraaf 8 voor een uitgebreidere beschrijving van deze netwerken) waarop het INT zal inspelen. Naast CLARIN zorgt het INT via European Language Grid en European Language Resource Coordination er nu al voor dat de Nederlandse taalinfrastructuur aanwezig is op diverse Europese platformen. In de komende jaren wordt door de Europese Commissie ook de Common European Language Data Space uitgerold en het INT zal opvolgen hoe de Nederlandse taalinfrastructuur hierbinnen ingebed en ondersteund kan worden. Daarnaast, op nationaal Nederlands niveau, zal CLARIAH-NL vanaf het najaar 2022 alle resources samenbrengen in het nieuwe INEO-platform. Bij al deze initiatieven zal het INT de verantwoordelijkheid op zich nemen om de consistente, crossplatform terugvindbaarheid en beschikbaarheid van de door het instituut beheerde data en software voor het Nederlands maximaal te bevorderen.

Uit de vorige alinea mag blijken dat het landschap van language resource repository's en taalinfrastructuurnetwerken vrij complex is. Het INT biedt daarom 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, ook geïntegreerd worden met wat het INT onder de Europese CLARIN-paraplu aan ondersteuning en advies biedt. Sinds de zomer van 2021 is het INT immers door CLARIN erkend als expertisecentrum voor het Nederlands (CLARIN Knowledge Centre for Dutch). Op de portaalsite K-Dutch (https://kdutch.ivdnt.org) 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 ook als expert deel uitmaken van het CLARIN Knowledge Centre for Lexicography, dat na de afloop van het Europese ELEXIS-project in 2022 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 en door als steunpunt te fungeren voor onderzoekers en studenten met specifieke taalinfrastructurele noden. Op die manier draagt het INT significant bij aan de bevordering van de studie van, en het wetenschappelijk onderzoek naar, de Nederlandse Taal in binnen- en buitenland (Taalunieverdrag art. 3d en 5a)

3 Corpusinfrastructuur

Corpora vormen een essentieel onderdeel van de infrastructuur voor het Nederlands. Ze bevatten 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 wordt in eerste instantie uitgevoerd ten behoeve van de verdere uitbouw van de centrale kennisbank voor de Nederlandse woordenschat (zie paragraaf 4) 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, infrastructurele ondersteuning biedt voor het bouwen, het gebruik dan wel het ter beschikking stellen van het corpusmateriaal (zie paragraaf 8.3 voor een overzicht van lopende en geplande projecten).

3.1 Corpuswerkzaamheden ten behoeve van de centrale kennisbank

In deze paragraaf gaan we nader in op de corpuswerkzaamheden die het INT zelf initieert en als onderdeel ziet van haar kerntaak om de Nederlandse taal te monitoren en te documenteren. Voor de

uitbouw van die corpora onderhoudt het INT langetermijnrelaties met leveranciers van taaldata, zoals uitgeverijen van kranten en tijdschriften. Op die manier biedt het INT onderzoekers toegang tot taaldata, die uitgeverijen vanuit een bekommernis over ongecontroleerde verspreiding liever niet ter beschikking stellen van kortlopende onderzoeksprojecten.

Er volgt een beschrijving van de beoogde samenstelling van deze corpora. Voor een gedetailleerde beschrijving van wat corpusbouw inhoudt, hoe corpusdata toegankelijk worden gemaakt en welke werkzaamheden daarvoor voorzien zijn in de komende beleidsperiode, verwijzen we naar bijlage I van dit document.

Voor het aanleggen van corpusmateriaal heeft het INT zich tot nog toe gefocust op geschreven taal of transcripties van gesproken taal. Daarvoor gebruikt het INT goed gemetadateerd (ruw) bronmateriaal, dat van oorsprong digitaal is of het resultaat van digitalisering. Voor gedigitaliseerd materiaal beperken we ons tot materiaal van de hoogst mogelijke kwaliteit qua tekstdigitalisering. Het INT heeft inmiddels behoorlijk wat expertise opgebouwd op het gebied van digitalisering (dataficatie) en is in dat verband betrokken bij het IMPACT Centre of Competence (vergelijk paragraaf 8.1).

Het hedendaags Nederlandse materiaal komt terecht in het *Corpus Hedendaags Nederlands* (CHN). Het CHN bevat hedendaags, goed gedateerd taalmateriaal uit kranten, boeken, tijdschriften, internetfora etc., voornamelijk uit de laatste twee decennia, onder meer afkomstig uit corpora die in het verleden door het instituut zijn samengesteld, of verzameld ten behoeve van specifieke projecten. De kern van het CHN echter is een monitorcorpus[5] van kranten dat na de recente samenwerking met DPG Media[6] nu de belangrijkste landelijke kranten van Vlaanderen, Nederland, Suriname en het Caribisch gebied bevat. Het corpus is online beschikbaar[7] en wordt maandelijks geüpdatet.

Voor het historisch Nederlands zijn de afgelopen jaren diverse corpora afzonderlijk online beschikbaar gemaakt, van zowel het INT als van derden[8]. Daarnaast is een begin gemaakt met het ontwikkelen van een groot diachroon corpus, samengesteld uit verschillende bestaande tekstverzamelingen en corpora, met als belangrijkste randvoorwaarde dat het tekstmateriaal goed gedateerd is en van uitstekende tekstkwaliteit[9]. Het materiaal is reeds ingezet ten behoeve van het ontwikkelen en testen van contextgevoelige word embeddings[10] waarmee bijvoorbeeld semantische verandering automatisch gedetecteerd zou kunnen worden.

De focus van de komende beleidsperiode blijft het verder uitbouwen van het monitorcorpus van kranten van het CHN. Het krantenmateriaal zal verder aangevuld worden tot een min of meer evenwichtig monitorcorpus voor dit millennium. Daarnaast zal aan het CHN materiaal toegevoegd worden dat via samenwerkingen in extern gefinancierde projecten beschikbaar komt, zoals bijvoorbeeld parlementaire debatten (ParlaMint-project) of hoorcolleges (SaBeD) (vgl. paragraaf 8.3).

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

[5]Een monitorcorpus is een corpus van een vaste samenstelling dat continu wordt aangevuld met nieuw materiaal waarmee taalverandering bestudeerd kan worden.
[6]De Pers Groep Media (www.dpgmediagroup.com).
[7]https://chn.ivdnt.org
[8]Zie daarvoor https://ivdnt.org/historisch-nederlands/.
[9]Hierin onderscheidt dit corpus zich van Nederlab waar tekstkwaliteit en datering niet overal toereikend is.
[10]Contextgevoelige word embeddings verwijst naar de taaltechnologie om woorden te representeren in een multidimensionale semantische ruimte, waarbij de betekenis van woorden bepaald wordt afhankelijk van de context waarin ze voorkomen en waarbij woorden met gelijkaardige betekenis dichtbij elkaar staan.

corpus zal, net zoals het CHN, taalkundig verrijkt worden met woordsoort en lemma en als geheel in één userinterface online doorzoekbaar gemaakt worden.

Ten slotte zal onderzocht worden of het monitorcorpus van kranten verder naar het verleden kan worden uitgebreid. De uitdaging hierbij wordt niet alleen de omvang, maar ook de kwaliteit van de digitalisering van de oudere kranten.

Zo wil het INT stapsgewijs toewerken naar één diachrone corpuscollectie van historisch tot hedendaags.

3.2 Overige corpuswerkzaamheden

De samenwerking met externe partijen biedt het INT de mogelijkheid om de corpusinfrastructuur qua data en tooling verder uit te bouwen en nieuwe wegen te verkennen. Voorbeelden uit de afgelopen beleidsperiode zijn het online brengen van het OpenSoNaR-pluscorpus, met voor het Corpus Gesproken Nederlands de koppeling tussen tekst en geluid en het aanleggen van corpora van eenvoudig Nederlands zoals WAI-NOT en Wablieft of een corpus van de Start!-krant ten behoeve van het project Eenvoudig Communiceren (vgl. paragraaf 8.3).

Dergelijke samenwerkingen zullen ook in de komende beleidsperiode verder worden gezet, en waar mogelijk uitgebreid. In de context van CLARIAH+ (zie paragraaf 8.3) wordt de corpus search engine uitgebreid om met parallelle corpora om te kunnen gaan en met syntactisch geannoteerd corpusmateriaal. In het project SignON (zie paragraaf 8.3) wordt bijgedragen aan de totstandkoming van multimediaal corpusmateriaal voor gebarentaal. Corpora van gesproken taal komen tot stand in het ParlaMintproject, waarin het INT verantwoordelijk is voor de data van het Belgisch federaal parlement (zie paragraaf 8.3), het SABeD-project, waar een corpus van gesproken academisch Belgisch-Nederlands wordt samengesteld, en het GCND-project, waarin het INT verantwoordelijk is voor het online toegankelijk maken van transcripties en geluid van gesproken dialectmateriaal (zie paragraaf 8.3).

4 Beschrijving van de woordenschat door de eeuwen heen: naar een centrale kennisbank.

Tot de centrale verdragstaken van de Taalunie behoren het bepalen van de officiële spelling (Taalunieverdrag art. 4b) en het opzetten van initiatieven voor de ontwikkeling van woordenboeken en woordenlijsten (art. 4d). Het INT geeft hieraan als toegepast wetenschappelijk instituut een concrete uitvoering door de ontwikkeling van een brede waaier aan lexicale databanken voor het Nederlands. Binnen de digitale taalinfrastructuur (zie overzicht in paragraaf 1) vormen lexicale databanken een secundair datatype: ze worden typisch op basis van primaire corpusdata gecompileerd en ze bevatten gesystematiseerde en gestructureerde informatie en kennis over de woorden. In die zin zijn ze de moderne, digitale versie van (spellings)lexica, woordenboeken en thesauri. De wetenschappelijke beschrijving van de Nederlandse woordenschat in al zijn facetten was altijd al en blijft ook in de komende jaren een van de kerntaken van het INT, zij het met een vernieuwde aanpak die tegemoetkomt aan de veranderende noden en eisen. Lexicale informatie die vroeger typisch in aparte eindproducten ondergebracht en raadpleegbaar was, moet nu makkelijk centraal bevraagbaar, flexibel hercombineerbaar en aan de primaire corpusdata gekoppeld zijn om op maat gemaakte taalapplicaties mogelijk te maken en data-gedreven en interdisciplinair onderzoek te

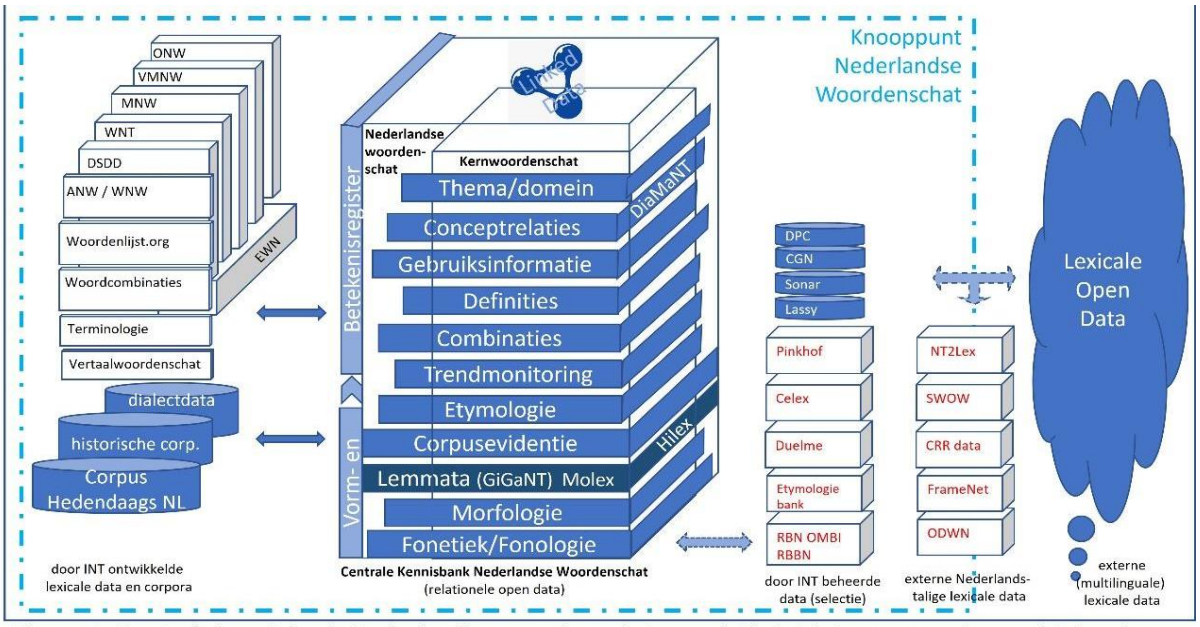
ondersteunen. Net als vergelijkbare instituten in het buitenland[11] zal het INT daarom in de komende jaren nog sterker inzetten op twee belangrijke vernieuwingen voor de lexicale taalinfrastructuur die in de vorige beleidsperiode voorbereid werden:

- De integratie van alle componenten van de woordenschatbeschrijving in één centrale, modulair georganiseerde, relationele kennisbank van de Nederlandse woordenschat door de eeuwen heen;
- De relatie tussen de primaire corpusdata en de afgeleide lexicale data wordt versterkt en geëxpliciteerd, zowel in het lexicografische compilatieproces, als in de resulterende databanken.

In wat volgt gaan we kort in op wat die twee vernieuwingen precies inhouden. Daarna wordt kort geschetst hoe vanuit de centrale kennisbank bestaande en nieuwe lexicografische eindproducten en diensten verder ontwikkeld worden. De concrete stappen en werkzaamheden om tot de centrale kennisbank te komen worden toegelicht in bijlage II. De onderliggende wetenschappelijke visie en motivatie voor deze integratie van de woordenschatbeschrijving is te vinden in bijlage III.

4.1 Integratie van de woordenschatbeschrijving

De integratie van de woordenschatbeschrijving in één kennisbank zal in een eerste fase gebeuren voor de algemene hedendaagse en historische woordenschat en op basis van de bestaande lexicale databanken. Op langere termijn zullen ook de terminologiebanken, de dialectdatabanken en de vertaalwoordenschat geïntegreerd en gekoppeld worden. De woordenschatbeschrijving samenbrengen binnen één kennisbank i.p.v. in aparte databanken heeft meerdere voordelen. Enerzijds leidt de integratie tot behoorlijke efficiëntiewinsten omdat dezelfde informatie, die vroeger over meerdere woordenboeken verspreid zat, nu aan elkaar gekoppeld is en centraal beheerd, bewerkt en op consistentie gecontroleerd kan worden. Anderzijds laat een geïntegreerde kennisbank, naast de bestaande functie als naslagwerk, ook nieuwe types van gebruik toe, met name voor datagedreven wetenschappelijk onderzoek en voor de ontwikkeling van taalapplicaties waarin woordinformatie op maat wordt aangeboden.



Figuur 1: Centrale kennisbank Nederlandse woordenschat en relatie tot interne en externe databanken

[11]Gelijkaardige integratieprojecten zijn al aan de gang voor o.a. het Deens (Central OrdRegister for Dansk), Duits (Digitales Wörterbuch der Deutschen Sprache), Sloveens (Slovenscina) en Pools (https://lab.dariah.pl/).

Figuur 1 toont de geplande opbouw van de geïntegreerde kennisbank (centraal in de figuur) en de relaties tot andere interne (links) en externe (rechts) databanken. In de afgelopen jaren werd met GiGaNT (het Groot Geïntegreerd Lexicon van de Nederlandse Taal) al het fundament van de kennisbank gelegd met een lexicon dat alle lemmata en hun vormvarianten uit de hedendaagse (Molex) en historische (Hilex) lexicale databanken van het INT centraal aan één uniek super-lemma-ID koppelt. Dat laat nu al een gecentraliseerde bewerking van uitspraakinformatie en morfologie toe. Bovenop GiGaNT zal in de komende jaren een *betekenisregister* (sense inventory) gebouwd worden waarin verschillende types lexicale informatie ook op betekenisniveau aan elkaar gekoppeld worden. DiaMaNT (het Diachroon seMantisch lexicon van de Nederlandse Taal) koppelt nu al definities uit verschillende historische woordenboeken op een schematisch niveau aan elkaar. Het betekenisregister zal dit uitbreiden naar de hedendaagse lexicale databanken en naar andere types lexicografische informatie zoals collocatiegedrag, conceptrelaties (synonymie, hyperonymie etc.) of, op termijn, vertaalequivalentie. Alle informatie over woorden – of ze nu hedendaags, historisch, etymologisch, dialectologisch, fonologisch, morfologisch, syntactisch, semantisch, pragmatisch, variatielinguïstisch, ontologisch of thematisch van aard is, informatie die tot nog toe verspreid zat – wordt op die manier op woord/betekenis-niveau bijeengebracht en gecombineerd doorzoekbaar en bruikbaar gemaakt voor nieuwe types onderzoek en ontwikkeling. Door deze integratie zal de workflow voor de historische en hedendaagse woordenschat geharmoniseerd worden zodat de unieke sterkte van de lexicografie aan het INT, namelijk haar diachrone diepgang, maximaal tot zijn recht kan komen.

4.2 Versterking van de relatie tussen corpusdata en lexicale data

De corpusgebaseerde lexicografische workflow zal geleidelijk aan gemodulariseerd worden om nieuwe content voor de componenten van de kennisbank aan te leveren, content die van meet af aan voldoet aan de vereisten van gestructureerde relationele data. Binnen elk van de modules zal telkens het lexicografische proces zelf geherorganiseerd worden om de koppeling tussen primaire corpusdata en lexicografische beschrijving sterker en explicieter te maken. Het datamanagement zal zo georganiseerd worden dat de tussenresultaten[12] van het lexicografisch proces als op zichzelf waardevolle datasets opgeslagen worden. Dit is essentieel om deze processen aan de hand van geavanceerde methodes uit de domeinen van Natural Language Processing (NLP) en Artificial Intelligence (AI) te modelleren zodat nieuwe data-analysetechnieken zoals Word Sense Induction in het lexicografische proces kunnen worden geïntegreerd. Bovendien zullen deze tussenresultaten als datasets ter beschikking komen van externe onderzoekers en ontwikkelaars. In de centrale kennisbank zullen corpusdata en lexicografische beschrijving in twee richtingen aan elkaar gekoppeld worden zodat nieuwe zoekmogelijkheden ontstaan, zoals in het corpus zoeken op kenmerken uit de lexicografische beschrijving of, omgekeerd, vanuit de kennisbank alle corpusvoorbeelden van een lexicografische beschrijving ophalen.

4.3 Lexicografische eindproducten, API's, datasets en Open Data

De huidige eindproducten, zoals woordenlijst.org, ANW, WNW, Woordcombinaties en de historische woordenboeken, zullen als afgeleide producten uit de kennisbank verder ontwikkeld en verbeterd worden. Daarnaast biedt de kennisbank nieuwe mogelijkheden. Allereerst zal het eenvoudiger zijn via API's de data aan te spreken en ter beschikking te stellen voor onderzoekers en ontwikkelaars. Uit de kennisbank zijn eenvoudig datasets af te leiden die als taalmaterialen ter beschikking gesteld zullen worden. De applicaties van de huidige producten kunnen dankzij de API's nieuwe mogelijkheden

[12]Het gaat dan om door lexicografen statistisch en taalkundig geanalyseerd taalgebruik zoals collocaties, concordanties en contextuele gebruikspatronen.

krijgen zoals het direct ophalen van corpusvoorbeelden of lexicografische informatie uit andere modules van de kennisbank. Dat zal betekenen dat de interfaces van de bestaande lexicografische producten aangepast zullen worden, of eventueel volledig herzien. Daarnaast wordt de kennisbank ook het uitgangspunt voor de ontwikkeling van compleet nieuwe lexicografische producten en diensten zoals voor de ondersteuning van historische tekststudie in de digital humanities of de ontwikkeling van didactisch materiaal voor woordenschatverwerving.

De geïntegreerde kennisbank zelf zal extern beschikbaar komen als *relationele open data* voor onderzoek en ontwikkeling via een API en een gebruikersinterface en zal maximaal ingebed worden in de nationale en internationale taalinfrastructuurnetwerken. Omdat de kennisbank unieke en persistente identifiers bevat voor zowel lemma's als betekenissen wordt het dan ook mogelijk om externe lexicale resources en datasets eenduidig te linken aan de kennisbank die op langere termijn het potentieel heeft om uit te groeien tot hét *digitale knooppunt* voor alle woordenschatgerelateerde hulpbronnen van het Nederlands. Dit past volledig in de visie van het INT om als het centrale taalinfrastructuurinstituut voor het Nederlands te functioneren.

5 Beschrijving van de Nederlandse dialecten: naar een dialect- en streektaalportaal

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 zoals de *Database van de Zuidelijk-Nederlandse dialecten (DSDD)* en de *elektronische Woordenbank van de Nederlandse dialecten (eWND)*. Het doel is om uiteindelijk een *dialect- en streektaalportaal* te kunnen bieden voor alle dialectvarianten van het Nederlands. Dit betekent in de eerste plaats dat de bestaande Database van de Zuidelijk-Nederlandse dialecten zal worden uitgebreid met dialectwoordenboeken uit het hele Nederlandse taalgebied.

In de beleidsperiode 2023-2027 worden hiervoor de eerste stappen gezet. Samen met de Radboud Universiteit wordt onderzocht of de kleinere onomasiologische dialectwoordenboeken van het Gelders, het Achterhoeks en het Liemers kunnen worden geïntegreerd. Daarnaast zal ook gekeken worden of semasiologische bronnen kunnen worden toegevoegd aan de van oorsprong onomasiologisch opgezette DSDD die zo veralgemeend wordt tot een *Database van de Nederlandse dialecten* (DDD). Dat zal een eigen strategie behoeven die zal worden uitgewerkt aan de hand van een aantal lokale dialectwoordenboeken uit de woordenbanken van Nederland (eWND) en Vlaanderen (Woordenbank.be[13]).

Intussen zal worden verder gewerkt aan de aanvulling en uitbreiding van de huidige DSDD en de eWND met hulp van vrijwilligers en stagiairs.

In samenwerking met de UGent zal gekeken worden of uit de dialectwoordenboeken een database van idiomen (spreekwoorden, uitdrukkingen) kan worden afgeleid. Dergelijke informatie is niet in de grote regionale woordenboeken te vinden.

Op de lange termijn zal ook een koppeling met de centrale kennisbank gerealiseerd worden.

[13]Een project van Variaties vzw. Koepelorganisatie voor dialecten en oraal erfgoed in Vlaanderen.

Ten slotte zal het INT, in samenwerking met de Taalunie, ook infrastructuur bieden aan streektaalorganisaties om hun talig cultureel erfgoed te kunnen beschrijven en beschikbaar stellen. Het ontwerp en de uitwerking van de infrastructuur gebeurt met behulp van externe financiering.

6 Terminologie: het expertisecentrum voor Nederlandstalige vaktaal

Het INT is in de afgelopen jaren een belangrijk centrum voor terminologie geworden. In uitvoering van de verdragstaken van de Taalunie i.v.m. terminologie (Taalunieverdrag art. 4c en 5e) kreeg het instituut de opdracht het Expertisecentrum Nederlandstalige Terminologie (ENT) op te richten en uit te bouwen. Dat is in de afgelopen jaren succesvol gerealiseerd. Er zit veel kennis en expertise in dit centrum en het is een blijvende taak om het ENT van nieuwe informatie te voorzien.

Dit vertaalt zich voor de komende beleidsperiode in klassieke veldondersteuning in de vorm van verdere updates van de website en de productie en distributie van vier nieuwsbrieven Terminologie per jaar.

Daarnaast blijft het ENT permanent het overzicht van de hedendaagse termenlijsten updaten en uitbreiden. Analoog hiermee wordt in de lijn van het INT en zijn historische woordenboeken en corpora ook een gedeelte voor lijsten met historische termen ingericht.

Er wordt ook de nodige aandacht besteed aan tooling t.b.v. terminologisch werk. Er wordt geïnvesteerd in de ontwikkeling van een nieuwe termtreffer[14], die aansluit bij de huidige corpusinfrastructuur, en er wordt een editingomgeving ter beschikking gesteld voor het bewerken van terminologische data. Deze kerntaken worden vanuit de lumpsum gefinancierd.

Het INT werkt nauw samen met externe partijen:

- NL-Term (samenwerking voor de jaarlijkse TiNT-dag)
- Termraad (samenwerking voor het bepalen van nieuwe termen voor overheid en EU-instellingen)
- Termraad Academy (opleiding en stages)
- European Association for Terminology (samenwerking op Europees niveau)

Het INT wil op termijn aansluiten bij initiatieven zoals de Federated eTranslation Termbank[15].

Naast deze permanente ondersteuning en uitbouw van het ENT ligt de focus op enkele belangrijke domeinen voor het versterken en ontwikkelen van de Nederlandstalige vaktaal.
Voor deze projecten moet externe financiering gezocht worden.
Het betreft:

- Overheidsterminologie (in het bijzonder de terminologie van het hoger onderwijs)
- Zorgterminologie
- Juridische terminologie
- Nederlands als wetenschapstaal

Overheidsterminologie:

De harmonisering en het in kaart brengen van de hogeronderwijsterminologie voor Nederland en Vlaanderen wordt voortgezet via stages en scripties. De groei van deze databank is uiteraard

[14]Een applicatie voor het extraheren van terminologie uit corpora.
[15]https://www.eurotermbank.com/

afhankelijk van de stages die bij het ENT worden aangevraagd en de begeleiding daarvan. Begeleiding van studenten blijft belangrijk en het stageaanbod voor terminologiewerk wordt blijvend gepromoot. Het is ook belangrijk voor de netwerkpositie van het ENT. Verder wil het INT vanuit dit project internationale samenwerkingen opzetten zoals met de gelijkaardige projecten bij EURAC (Bolzano) en de Universitat Autònoma de Barcelona. Dit project is waardevol genoeg om blijvend in te zetten op extra financiering en internationale samenwerking.

Zorgterminologie:

Een eerste versie van het *Pinkhof Geneeskundig woordenboek* werd in 2021 online gezet via een door het INT ontwikkelde applicatie. Dit woordenboek wordt bijgewerkt met als primair doel een verklarend hedendaags medisch woordenboek en een taalbaak voor medisch Nederlands te vormen. Om dit te realiseren wordt er samengewerkt met de Stichting Beheer Pinkhof-database. Er is een raad van advies opgericht om afgeleide medische terminologieprojecten te begeleiden en ons te adviseren wat prioritaire wensen en noden van de samenleving zijn.

Juridische terminologie:

Het *Juridisch woordenboek Diccionario juridico* van (M.C. Oosterveld-Egas Reparaz en J. Vuyk-Bosdriesz) werd in 2022 online gezet via een door het INT ontwikkelde applicatie en er wordt verder gewerkt aan het updaten en uitbreiden van het bestand. In samenspraak met dr. Karl Hendrickx (Universiteit Antwerpen en KU Leuven) wordt in 2023 een studie gemaakt om de mogelijkheid te onderzoeken Belgische equivalenten aan het juridische woordenboek toe te voegen. Deze studie zal worden uitgevoerd op basis van bestaande lijsten, maar met een vergelijkend onderzoek naar equivalenten en semi-equivalenten. Hiervoor is extra budget nodig.

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. In dit verband werkt het INT samen met het Proefproject Nederlands als wetenschapstaal - van corpora naar terminologielijsten. Dit project is een samenwerking tussen Stichting Nederlands / Vlaams Platform Taalbeleid Hoger Onderwijs, KULeuven, UGent en het INT. De motivatie voor dit project is breed: het past bij “Nederlands in de aansluiting” tussen het middelbaar en hoger onderwijs, en het streven om de drempel voor het hoger onderwijs te verlagen. Tevens is het een ondersteuning van het gebruik van het Nederlands als wetenschapstaal. In 2021 werden door het INT aangepaste terminologielijsten ontwikkeld voor de vakken scheikunde en wiskunde (niveau BA1 NL en VL). Deze lijsten worden via een speciaal ontwikkelde zoekapplicatie ontsloten en zullen in 2023 online raadpleegbaar zijn. Ook hier is verder onderzoek nodig voor andere vakken in BA1. De ontwikkeling van een corpus Gesproken Academisch Belgisch-Nederlands (project SABeD KU Leuven zie 8.3) en de ontwikkeling van een woordenlijst academisch Nederlands worden in de komende jaren (2023-24) verdergezet. Zo wordt een verband gelegd tussen de vaktaal en het algemene lexicon.

Op de lange termijn is het de bedoeling om de beschrijving van de vaktaal verder te integreren in de centrale kennisbank voor de Nederlandse woordenschat.

7 Grammatica: naar een grammaticaportaal

Sinds 2020 valt de grammaticabeschrijving, als een van de centrale verdragstaken van de Taalunie (Taalunieverdrag art. 4b en 4d), ook binnen de structurele basisopdracht van het INT. Dit betekent dat het INT zorgt voor de ontwikkeling, het beheer en de beschikbaarstelling van verschillende digitale

grammaticaproducten. Op termijn wil het INT toewerken naar een grammaticaportaal, een centraal digitaal platform dat de toegang moet vormen tot deze producten.

In de periode 2023-2027 worden hiervoor de eerste stappen gezet. Er zal een eerste versie van het grammaticaportaal worden gebouwd, met verwijzingen en centrale zoekmogelijkheden in Taalportaal, Algemene Nederlandse Spraakkunst (e-ANS) en Taaladvies.net. Vervolgens zal het portaal stapsgewijs worden uitgebreid met o.a. een centrale grammaticale termendatabank en informatie over lopende projecten. Het INT voorziet hiervoor een samenwerking met de Northwest University, Zuid-Afrika. Ten slotte doet het INT vooronderzoek om te komen tot een inventaris van taalkundige constructies, en integratie hiervan met de centrale kennisbank voor de woordenschat en het project Woordcombinaties.

Daarnaast zal er worden gewerkt aan de e-ANS en het Taalportaal:

- De e-ANS zal verder worden herzien op inhoudelijk en technisch vlak, met coördinatie vanuit het INT, en in samenwerking met externe auteurs. Daarnaast wordt een didactische laag toegevoegd: een onderwijsmodule met samenvattingen, begrippenlijsten, oefeningen en video's, bedoeld om de inhoud van de ANS toegankelijker te maken voor een breder publiek in het algemeen, en voor neerlandici extra muros in het bijzonder. Voor al deze werkzaamheden is het INT afhankelijk van financiering (van de Taalunie) en de beschikbaarheid van externe auteurs.
- Voor het Taalportaal zijn de komende jaren data-updates en technisch onderhoud gepland: nieuwe materialen, waaronder een nieuw deel van de Syntax of Dutch, zal door het INT worden verwerkt en door het INT worden gepubliceerd en de webapplicatie zal worden geüpdatet. Ook zal er, in samenwerking met het Zuid-Afrikaanse SaDilar, kruisgewijze zoekfunctionaliteit met het Zuid-Afrikaanse language portal worden toegevoegd.

8 Nationale en Internationale Samenwerkingsverbanden

Als toegepast wetenschappelijk instituut beweegt het INT zich in het veld van onderzoekers en taalkundigen, die tegelijk ook een belangrijke doelgroep voor de door het INT beheerde taalinfrastructuur zijn. Bestaande contacten met onderzoekers uit binnen- en buitenland, verbonden aan wetenschappelijke instituten en universiteiten, worden onderhouden in netwerken, netwerkprojecten en onderzoeks- en infrastructuurprojecten en waar mogelijk geïntensiveerd en uitgebreid. Daarnaast verleent het INT ook langduriger infrastructurele ondersteuning voor het veld.

8.1 Netwerken

CLARIN

De CLARIN-centra zijn ontstaan uit het European Research Infrastructure Consortium CLARIN (Common Language Resources and Technology Infrastructure) om een duurzame terugvindbaarheid en beschikbaarheid van taaldata en -software te waarborgen voor onderzoek op Europees niveau.

Zoals besproken in paragraaf 2 is het INT als CLARIN-B-centrum en als CLARIN Knowledge Centre een van de knooppunten in dit Europees netwerk. Binnen het CLARIN-netwerk is het INT coördinator voor CLARIN-België, sinds de formele goedkeuring hiervan door de Belgische overheid in de zomer van 2021. Het INT is ook de vertegenwoordiger van Vlaanderen en België in de CLARIN National Coordinators Forum. Vanuit het CLARIN-netwerk is het INT ook betrokken bij specifieke onderzoeks- en

infrastructuurprojecten in Nederland (CLARIAH+) en Vlaanderen (CLARIAH-Vlaanderen). Deze worden in paragraaf 8.3 verder voorgesteld.

European Language Grid en European Language Equality (European Language Data Space)

Het INT is National Competence Centre (NCC) voor Nederland voor het Horizon 2020-project European Language Grid. In dit project is een platform ontwikkeld waar zowel publieke als commerciële partners taalkundige resources kunnen publiceren en ter beschikking stellen van anderen. Hoewel het project afloopt in 2022 wordt het platform verdergezet binnen het European Language Data Space-initiatief van de Europese Commissie. Vanuit zijn functie als repository en expertisecentrum (zie paragraaf 2) wil het INT ook hier zijn rol blijven spelen als NCC voor Nederland.

Het INT is consortiumpartner in het Horizon 2020 European Language Equality (ELE) project, het zusterproject van European Language Grid, dat eveneens afloopt in 2022 maar voortgezet wordt binnen de European Language Data Space. Het vervolgproject European Language Equality 2 (ELE2), waarin het INT betrokken is via consortiumpartner EFNIL, sluit aan op het ELE-project en heeft een looptijd van 1 jaar. Beide projecten (ELE en ELE2) hebben als doel het ontwikkelen van een agenda en roadmap voor het bereiken van volledige taalgelijkheid in Europa tegen 2030. Het INT heeft meegewerkt aan deze agenda en zal zich ook verder inzetten om de taalgelijke behandeling van Nederlandstaligen online zoveel mogelijk te bevorderen.

ELRC

Het INT is betrokken bij ELRC (European Language Resource Coordination), en vervult binnen dit project samen met Jan Odijk van de Universiteit van Utrecht de rol van Technical National Anchor Point voor Nederland. ELRC heeft als doel op grote schaal taaldata te verzamelen die gebruikt kunnen worden voor het ontwikkelen van automatische vertaalsystemen en andere taaltechnologische tools voor publieke diensten in alle EU-lidstaten, zodat er beter tegemoetgekomen kan worden aan de dagelijkse noden van deze publieke diensten. Dergelijke tools zijn van groot belang om taalbarrières in Europa te doorbreken en de positie van talen zoals het Nederlands te waarborgen. Het INT zal zijn rol in de organisatie van ELRC-workshops blijven spelen en zo het belang van het verzamelen van Nederlandse taaldata verder blijven promoten, zowel binnen Nederland als binnen Europa.

DARIAH

Via de CLARIAH-projecten in Nederland en Vlaanderen en het IMPACT Centre of Competence is het INT ook betrokken bij DARIAH (Digital Research Infrastructure for the Arts and Humanities), een Europees onderzoeksinfrastructuurconsortium dat specifiek focust op de digital humanities.

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 sedert 2019 ook CLARIN Knowledge Centre en organiseert de DATECH-conferences[16]. De werkzaamheden m.b.t. digitalisering die onder andere in de context van CLARIAH+ worden uitgevoerd, worden in samenwerking met het Centre uitgevoerd.

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

Nederlands/Vlaams Platform Taalbeleid Hoger Onderwijs

Het INT werkt samen met het Nederlands/Vlaams Platform Taalbeleid Hoger Onderwijs; dit is een platform voor ontmoeting en uitwisseling van kennis en ervaringen. Door middel van kennisdeling en kennisvorming wil het Platform bijdragen aan initiëring en verdere ontwikkeling van een breed taalbeleid in het hoger onderwijs, met name in de hogescholen in Vlaanderen en Nederland. In 2020 startte de samenwerking met het Platform in de vorm van projecten rond "Nederlands als wetenschapstaal". In dit verband zal het INT ook in komende jaren terminologielijsten van bepaalde opleidingsonderdelen voorbereiden en online beschikbaar stellen. In dit kader loopt ook het project rond Academisch Nederlands, waarbij via cursusmateriaal een selectie van de academische woordenschat wordt geselecteerd en van definities voorzien (zie ook paragraaf 6).

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 als doel de Nederlandse activiteiten in AI te stimuleren, te ondersteunen en waar nodig te organiseren. Het INT is als werkgroeplid bij dit initiatief betrokken.

Elexis Association

De ELEXIS Association bouwt voort op het netwerk dat tijdens het ELEXIS-project ontstaan is en heeft als doel om verdere onderzoeksinitiatieven en -activiteiten over lexicografie te bevorderen en te coördineren.

8.2 Netwerkprojecten

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

Het INT is partner in het Europese onderzoeksnetwerk (COST) Nexus Linguarum[17], dat taalkundigen, computerwetenschappers, terminologen en andere belanghebbenden uit het bedrijfsleven en de samenleving samenbrengt om de verdere uitbouw van *Linguistic Data Science* als onderzoeksdomein binnen de datawetenschap te bevorderen. Het netwerk beoogt een ecosysteem uit te bouwen van meertalige en semantisch interoperabele linguïstische data om de systematische taaloverschrijdende ontdekking, exploratie, exploitatie, uitbreiding, curatie en kwaliteitscontrole van linguïstische data te bevorderen. Als partner garandeert het INT dat ook de taalinfrastructuur voor het Nederlands in dit ecosysteem van relationele open data ingebed is.

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

Het INT neemt deel aan het Europese onderzoeksnetwerk (COST) UniDive[18] (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.

[17]COST (Cooperation in Science and Technology) action 18209. Looptijd 28/10/2019 tot 27/10/2023 https://nexuslinguarum.eu

[18]COST action 21167. Looptijd 23/09/2022 tot 22/09/2026 (https://www.cost.eu/cost-action/universality-diversity-and-idiosyncrasy-in-language-technology/).

8.3 Onderzoeks- en infrastructuurprojecten

CLARIAH+ Nederland (2019-2023)

Daar waar het CLARIAH (Common Lab for Research in the Arts and Humanities) CORE-project erop gericht was een gemeenschappelijke infrastructuur tot stand te brengen voor data-intensief wetenschappelijk onderzoek in de geesteswetenschappen, richt het vervolgproject CLARIAH-PLUS, zich nog meer op het concreet ondersteunen van de onderzoeker door middel van het tot stand brengen van (virtuele) onderzoeksomgevingen.

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

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.

CLARIAH Vlaanderen (2021-2024)

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

SignON (2021-2024)

Het INT is momenteel consortiumpartner in het SignON-project, een door Horizon 2020 gefinancierde Research and Innovation Action. Dit project loopt tot eind 2023. Het doel van dit project is om een applicatie te bouwen die het mogelijk maakt automatisch te vertalen tussen verschillende gebarentalen (waaronder Vlaamse Gebarentaal en Nederlandse Gebarentaal) en gesproken talen (waaronder het Nederlands). De rol van het INT in dit project bestaat hoofdzakelijk uit twee aspecten. Het eerste aspect is het opzetten van de nodige infrastructuur om de backend van de app te draaien. Deze taak biedt ons de ruimte om met externe financiering ervaring op te doen in het opzetten van infrastructuur voor state-of-the-art neurale *transformer*-modellen, die gebruik maken van specifieke hardware. Het tweede aspect is het verzamelen en beschikbaar stellen van datasets voor het trainen van de machine-learningalgoritmes die deze vertaaltaak moeten uitvoeren. Deze taak sluit aan bij onze

doelstelling om in het Language Resource Centre de beschikbare datasets uit te breiden naar multimediale data, parallelle data en gebarentalen. Het INT zal ook het initiatief nemen voor het opzetten van vervolgprojectaanvragen m.b.t. automatische vertaling van gebarentalen.

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

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

De opnames werden volgens een nieuw ontwikkeld transcriptieprotocol getranscribeerd en taalkundig verrijkt met bestaande tools. Het INT zal de audio, de transcripties en de annotaties vanaf 2024 vrij online beschikbaar en doorzoekbaar maken en duurzaam bewaren. Het INT zal in overleg met UGent verder onderzoeken welke stappen gezet moeten worden om de dialectopnames en de erbij horende transcripties te ontsluiten voor wetenschappers en het grote publiek.

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 new(s) (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, vanuit de hypothese dat kranten in de verbreiding van de Nederlandse standaardtaal een cruciale rol hebben gespeeld, als het eerste massamedium dat door alle sociale klassen werd gelezen. 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. Het project voorziet ook in een educatieve module via een website (apart gefinancierd), getiteld ‘Krantentaal vroeger en nu’. Op deze website zullen eenvoudige tools aangeboden worden aan scholieren waarmee ze oude, 17e-eeuwse krantenberichten zullen kunnen doorzoeken en vergelijken met modern nieuws. Deze module zal worden gehost bij het INT en worden vervaardigd in samenwerking met docenten uit het middelbaar onderwijs.

SABeD (2021-2023)

Voor het al eerder (zie paragraaf 3.2) genoemde SABeD-project faciliteert het INT het werk rond corpusdesign en annotatieprocedures en biedt praktische ondersteuning. Het resulterende corpus zal blijvend beschikbaar gesteld worden via de INT-infrastructuur.

ClaSABeD (2022-2023)

Dit CLARIAH-project (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.

ParlaMint II (december 2021 – mei 2023)

ParlaMint is een project, gefinancierd door CLARIN ERIC, dat bijdraagt aan de totstandkoming van vergelijkbare en uniform geannoteerde meertalige corpora van parlementaire zittingen. ParlaMint I creëerde corpora en maakte ze 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.

8.4 Overige infrastructurele dienstverlening

Vertaalwoordenschat

Aansluitend bij de verdragstaken van de Taalunie omtrent woordenboeken (Taalunieverdrag art. 4d) biedt het INT infrastructuur voor het ter beschikking stellen en onderhouden van de tweetalige woordenboeken die in de afgelopen decennia, onder meer in opdracht van de Commissie Lexicologische Vertaalvoorzieningen (CLVV, 1993-2003), zijn gemaakt voor taalparen die voor de Nederlandstalige gebruiker wel relevant zijn, maar op de commerciële markt niet spontaan aan bod kwamen. Om ervoor te zorgen dat deze woordenboeken beschikbaar blijven voor gebruikers en dat vertalingen tussen diverse talenparen blijvend worden ondersteund, nu er geen nieuwe gedrukte woordenboeken meer van worden gemaakt, is de Vertaalwoordenschat ontwikkeld. Het INT zal de Vertaalwoordenschat in de komende beleidsperiode blijven uitbreiden met nieuwe taalparen. Daarnaast zullen de woordenboeken inhoudelijk worden bijgewerkt en zal er gekeken worden naar de mogelijkheden om extra informatie die nuttig is voor gebruikers (zoals vervoegingen) aan de applicatie toe te voegen. De woordenboeken hebben minimaal een koppeling op lemmaniveau met de lexicale kennisbank.

Etymologiebank

Het INT is sinds 2020 verantwoordelijk voor het hosten van de *Etymologiebank*. 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. 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.

Taaladvies.net

Een belangrijke verdragstaak van de Taalunie is het bevorderen van het verantwoorde gebruik van het Nederlands (Taalunieverdrag art. 3b). De webapplicatie Taaladvies.net geeft daar een concrete uitvoering aan. Het INT ontwikkelde de webapplicatie voor Taaladvies.net en sinds 2021 wordt deze ook door het instituut gehost en up-to-date gehouden. Hoewel de verantwoordelijkheid voor de inhoud bij de samenwerkende taaladviesorganen blijft, is er door deze nieuwe opzet nauw contact tussen de taaladviseurs en het INT. Zo helpt het INT met technische vragen en wensen van de taaladviseurs, en voorzien de taaladviseurs het INT van input bij de herziening van de ANS. Tussen de

ANS, de spellingsbank (woordenlijst.org) en Taaladvies.net wordt veelvuldig naar elkaar verwezen, en het is dan ook de bedoeling om de komende jaren te werken aan de verbetering van koppelingen, en inhoudelijke samenhang tussen beide sites.

Neerlandistiek

Aansluitend bij de verdragstaken van Taalunie met betrekking tot de studie van het Nederlands (Taalunieverdrag art. 3b en 3d) levert het INT sedert 2021 technische ondersteuning voor de website van het online tijdschrift Neerlandistiek. Neerlandistiek is een elektronisch tijdschrift voor de Nederlandse taalkunde, letterkunde en taalbeheersing dat dagelijks informeert over ontwikkelingen in het vakgebied voor iedereen die er belangstelling voor heeft. Het tijdschrift heeft ook een dagelijkse nieuwsbrief.

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.

Pallas

Een internationaal team van onderzoekers werkt aan een geannoteerde digitale editie van het achttiende-eeuwse Russische Сравнительный словарь всѣхъ языковъ и нарѣчій, по азбучному порядку расположенный, of *Vergelijkend woordenboek van alle talen en dialecten, in alfabetische volgorde*, dat op verzoek van Catharina de Grote in 1790-1791 is samengesteld door de Duitse onderzoeker Peter Simon Pallas. Het INT host de database met de vertaling in 311 talen van 296 begrippen of een deel daarvan (het totale woordenboek bevat bijna 62.000 ingangen). De komende periode zal dit woordenboek inclusief verrijkingen zoals wetenschappelijke transliteratie en toegevoegde moderne spellingen en Engelse vertalingen op het internet beschikbaar worden gemaakt, en er zal een edited volume verschijnen, geschreven door internationale specialisten in de verschillende taalfamilies, met inhoudelijke informatie over de gegevens uit het woordenboek.

9 Investering in eigen IT-capaciteit

De interne computationele infrastructuur van het INT dient in de nodige opslagcapaciteit en rekenkracht te voorzien voor de verschillende taalinfrastructurele opdrachten zoals de repository, de eigen productontwikkeling en de infrastructuurondersteuning aan derden. Aangezien deep learning in steeds meer werkprocessen zal figureren, is structurele beschikbaarheid van GPU's[19] van belang. Daarbij moet (vooral voor het trainen van modellen) ook worden gedacht aan cloud-voorzieningen. De IT-infrastructuur moet ook voorbereid worden op het toenemend belang van multimediale content.

Daarnaast is van groot belang dat wordt geïnvesteerd in relevante kennis bij het IT-team. Met name de kennis van en praktische ervaring met deep learning zal worden uitgebreid. Een apart overzicht van de financiële implicaties van de groei in software en hardware wordt toegevoegd aan de meerjarenbegroting.

10 Onderwijs

Vanuit zijn unieke expertise als taalinfrastructuurinstituut verleent het INT ook een aantal specifieke diensten aan het onderwijs. Het dichtst aansluitend bij de wetenschappelijke activiteiten als toegepast

[19] GPU: Graphics Processing Unit.

onderzoeksinstituut is de dienstverlening naar universiteiten en hogescholen. Voor universitaire studenten biedt het INT zowel in Leiden als in Leuven een collegereeks over computationele en corpusgebaseerde lexicografie aan. Daarnaast begeleiden de gepromoveerde taalkundigen van het instituut ook masterproeven en promovendi en dit zal ook in de komende jaren voortgezet worden. Voorts zal het INT zich blijven inzetten om stageplaatsen aan te bieden voor studenten die belangstelling hebben voor lexicografie, terminologie, etymologie, dialectologie, corpuslinguïstiek, artificiële intelligentie en digital humanities. Zo doen de potentiële lexicografen, terminologen en (computationele) linguïsten van de toekomst praktijkervaring op met het opbouwen, inventariseren en analyseren van databanken en andere taalinfrastructuur. Binnen die actieve bijdrage aan het hoger talenonderwijs past ook de participatie in het Nederlands/Vlaams Platform Taalbeleid Hoger Onderwijs (zie paragraaf 8.1).

Door de toenemende integratie van de taalinfrastructuur op Europees niveau (zie paragraaf 2) ligt het voor de hand dat er ook initiatieven genomen worden om hierover gespecialiseerd wetenschappelijk onderwijs op Europees niveau te organiseren. Het INT is nu al 'observer' bij de bestaande European Master in Lexicology[20] (EMLex) en zal in de komende jaren onderzoeken of het een onderwijsopdracht binnen dit of een gelijkaardig initiatief kan opnemen.

Naast de directe betrokkenheid bij het wetenschappelijk onderwijs, zet het INT zich op verschillende manieren actief in om de door het instituut beheerde taalinfrastructuur zo toegankelijk en bruikbaar mogelijk te maken voor educatieve doeleinden. Ten eerste wordt aan een aantal databanken door het instituut zelf een didactische laag toegevoegd. Bij de Algemene Nederlandse Spraakkunst (ANS) gaat het om didactisch materiaal bij een aantal hoofdstukken met als primaire doelgroep neerlandici extra muros, in overeenstemming met de verdragstaak van de Taalunie om de studie van de Nederlandse taal in het buitenland te bevorderen (Taalunieverdrag art. 3d). Bij de databank met woordcombinaties is er een speciale online tool ontwikkeld die het leren van het Nederlands als vreemde taal ondersteunt. Voor beide projecten zal bijkomende financiering en samenwerking gezocht worden om in de komende jaren de didactische mogelijkheden verder uit te breiden. Daarnaast speelt de taalinfrastructuur van het INT een voorname rol bij de ontwikkeling door derden van educatieve materialen voor het secundair of NT2[21]-onderwijs. Het instituut levert daarbij niet rechtstreeks leermaterialen voor deze doelgroepen, maar ondersteunt de ontwikkeling ervan en werkt hiervoor actief samen met de relevante spelers in de educatieve sector. Zo worden de lexicale kennisbank en de corpusinfrastructuur van het INT ingezet bij onder andere het samenstellen van verrijkte woordenlijsten voor specifieke taalvaardigheidniveaus (zoals de CEFR[22]-niveaus). Corpusmateriaal wordt gebruikt om semi-automatisch taaloefeningen te genereren. Met de laatste toepassing is ervaring opgedaan in een pilootproject in samenwerking met Stichting Expertisecentrum Oefenen.nl. Het doel is om dit verder uit te bouwen tot een generieke infrastructuur die dienstig is aan alle mogelijk geïnteresseerde externe partners bij de ontwikkeling van nieuw educatief materiaal voor uiteenlopende doelgroepen binnen verschillende toepassingen. Bij de verdere ontwikkeling van de omgeving zal ook worden aangesloten bij de behoefte om zowel de Woordcombinaties-applicatie als de ANS te verrijken met oefenmateriaal. Andere voorbeelden van taalinfrastructurele ondersteuning voor educatieve doeleinden zijn het Spread the new(s)-project (zie paragraaf 8.3) en de ontwikkeling van academische woordenlijsten in de marge van het SABeD-project (zie paragraaf 8.3). Ook in de komende jaren zal het instituut dit type dienstverlening binnen de mate van het mogelijke blijven

[20] https://www.emlex.phil.fau.eu/

[21] NT2: Nederlands als tweede taal.

[22] CEFR: Common European Framework of Reference, een uniform systeem van taalniveau-indeling dat gebruikt kan worden voor alle Europese talen.

aanbieden. Ten slotte zat het INT het educatieve potentieel van zijn taalinfrastructuur blijven promoten door aanwezig te zijn op de relevante fora voor het secundair en NT2-onderwijs (Week van het Nederlands, HSN[23]-conferentie) en door zelf voorbeelden van lesmateriaal aan te bieden op een speciale onderwijspagina van de website (zie ook volgende paragraaf 11).

[23]HSN: Het Schoolvak Nederlands.

11 PR en communicatie

Het INT wil met zijn communicatie diverse doelgroepen zo goed mogelijk bereiken. Door verschillende activiteiten aan te bieden toegespitst op de afzonderlijke doelgroepen kan het INT een ruim publiek bedienen en informatie en ondersteuning op maat geven. Hiervoor zal het INT in de komende jaren gebruikersenquêtes ontwikkelen om meer inzicht te krijgen in de doelgroepen en hun behoeften.

Om de doelgroepen in de toekomst nog beter te bereiken, werkt het INT samen met de Taalunie en genootschappen zoals Onze Taal en De Buren om gemeenschappelijk naar buiten te komen. In de komende tijd wordt onderzocht hoe de krachten op het vlak van communicatie kunnen worden gebundeld.

Doelgroepen:

- Algemeen publiek
- Taalprofessionals (docenten, vertalers, copywriters etc.)
- Onderzoekers en ontwikkelaars

De website is nu uitermate geschikt en toegankelijk voor het algemene publiek en de taalprofessionals. Voor de komende beleidsperiode zal het INT kijken hoe het via de website onderzoekers en ontwikkelaars beter kan informeren over de rol als taalinfrastructuurinstituut en over de dienstverlening.

Jarenlang was Twitter het voornaamste socialemediakanaal waarop het INT actief was voor het delen van nieuws, evenementen en kennis over taal. Met de komst van de vernieuwde website eind 2020 is er een plug-in geïntegreerd waarmee alle socialemediakanalen vanuit één plek bijgehouden kunnen worden. Daarmee is de activiteit van het INT op LinkedIn en Facebook aanzienlijk toegenomen, met zichtbaar resultaat. Ook is het INT gestart met Instagram, waar de huisstijl van het instituut goed tot zijn recht komt. De komende jaren blijven sociale media belangrijk om het publiek te bereiken. Het INT blijft monitoren wat voor het instituut goed werkt op welk kanaal, welke ontwikkelingen er zijn op het gebied van sociale media en hoe het het bijhouden van alle kanalen zo goed mogelijk kan stroomlijnen.

Een van de grootste aandachttrekkers op de website is de populairwetenschappelijke rubriek ‘Nieuw woord van de week’, over neologismen. Inmiddels is ‘Terug in de taal’ over historisch Nederlands ook een vaste waarde geworden en zijn er nieuwe rubrieken toegevoegd waaronder ‘Uit de streek’ over dialect, ‘Uitgelichte term’ en ‘Taalmateriaal uitgelicht’. Deze en mogelijk nieuwe rubrieken zullen ook in de komende beleidsperiode op de website van het instituut terug te vinden zijn.

Elk jaar publiceert het INT aan het einde van het jaar een populairwetenschappelijke uitgave die als nieuwjaarsgeschenk verstuurd wordt aan relaties. Ook voor de komende periode wordt deze traditie voortgezet. Met die publicatie sluit het INT aan bij een terugkerend thema in dat jaar, voortvloeiend uit de werkzaamheden van het INT.

Als gevolg van de coronapandemie en het zoeken naar nieuwe vormen van communicatie met de doelgroepen, heeft het werken met digitale media bij het INT een vlucht genomen. Zo heeft het instituut twee succesvolle podcastreeksen ontwikkeld, goedbezochte webinars en digitale en hybride evenementen georganiseerd en een start gemaakt met het ontwikkelen van educatieve video’s. Er werd onder andere ook een korte animatievideo gemaakt waarin verbeeld wordt hoe je gericht in de historische woordenboeken kunt zoeken. Deze nieuw ingeslagen weg is verfrissend en biedt talloze mogelijkheden. Komende jaren zal het INT daarom ook blijven inzetten op het produceren van podcasts, organiseren van webinars en de uitbreiding van videomateriaal.

12 Planning

Deze paragraaf bevat een overzicht van de geplande werkzaamheden die niet onder de externe projecten vallen. De timing (eerste helft dan wel tweede helft beleidsperiode) moet gezien worden als een globale indicatie van wanneer het zwaartepunt van de werkzaamheden ligt.

Activiteit	Deliverables	Timing
Repository	Licentiemodellen internationaal	eerste helft beleidsperiode
	Vernieuwde opzet repository	tweede helft beleidsperiode
	Stroomlijning met andere Europese platforms	doorlopend
Kenniscentrum	Inhoud K-Dutch updaten	doorlopend
	Servicedesk	doorlopend
	Promotie taalinfrastructuur	doorlopend
Corpusinfrastructuur	Monitorcorpus (uitbreiding samenstelling)	doorlopend
	Diachroon corpus (uitgebreid en met betere metadata)	doorlopend
	Trainings- en evaluatiemateriaal voor taalkundige verrijking	eerste helft beleidsperiode
	Vernieuwde basisverrijkingslaag met state-of-the-arttechniek	eerste helft beleidsperiode
	Lemmatisering met koppeling aan het lexicon	tweede helft beleidsperiode
	Syntactische annotatie van het monitorcorpus	eerste helft beleidsperiode
	Herinrichting corpusbuildingworkflow en corpusopslag	eerste helft beleidsperiode
Centrale kennisbank: datamodel	Datamodel kennisbank, uitgebreid met nieuwe componenten en gestroomlijnd	eerste helft beleidsperiode
Centrale kennisbank: GiGaNT	GiGaNT-ONW	eerste helft beleidsperiode
	GiGaNT-VMNW	eerste helft beleidsperiode
	Workflow neologismen	eerste helft beleidsperiode
	Workflow nieuwe oude woorden (lacunes)	tweede helft beleidsperiode
	Morfologische analyse: principes en dataset	tweede helft beleidsperiode
	Morfologische analyse: infrastructuur en analysesoftware	tweede helft beleidsperiode
	Verbeterde parsering historische woordenboeken	doorlopend
	Bewerkingsomgeving om de codering van de woordenboeken te verbeteren	eerste helft beleidsperiode
	GiGaNT: data-updates	doorlopend
	GiGaNT: datastructuur conform nieuw datamodel kennisbank	eerste helft beleidsperiode

Centrale kennisbank, betekenisregister: ondersteuning lexicografische bewerking	Omgeving opbouw betekenisinventaris	eerste helft beleidsperiode
	Omgeving linken koppelbetekenissen aan corpusdata	eerste helft beleidsperiode
	Omgeving bewerking onbeschreven woorden	eerste helft beleidsperiode
	Omgeving analyse en beoordeling van combinaties	doorlopend
	Opbouw datasets (kernwoordenschat)	doorlopend
Centrale kennisbank, betekenisregister: informatie-extractie	Blacklab: ondersteuning grotere corpora, performanceverbetering door gedistribueerde architectuur	eerste helft beleidsperiode
	Blacklab en frontend: zoeken op syntactische verrijking	eerste helft beleidsperiode
	Blacklab en frontend: uitgebreide statistische componenten corpus-frontend	eerste helft beleidsperiode
	Blacklab en frontend: extractie combinatiegedrag van woorden	eerste helft beleidsperiode
	Blacklab en frontend: selectie corpusattestaties	tweede helft beleidsperiode
	WSD en WSI	tweede helft beleidsperiode
Centrale kennisbank: lexicografische eindproducten API's en datasets	Woordenlijst.org: updates	doorlopend
	ANW: data-updates	doorlopend
	WNW: data-updates	doorlopend
	Historische woordenboeken: data-updates	doorlopend
	Woordcombinaties	doorlopend
	Vernieuwde applicatie ANW/WNW	tweede helft beleidsperiode
	Vernieuwde applicatie WNW	tweede helft beleidsperiode
	Vernieuwde applicatie online historische woordenboeken	tweede helft beleidsperiode
Dialecten	DSDD-update WVD	eerste helft beleidsperiode
	DSDD-update Zeeuws (inhoud en technisch)	eerste helft beleidsperiode
	DDD-update onomasiologische wdb	doorlopend
	DDD-update semasiologische wdb	tweede helft beleidsperiode
	infrastructuur dialecten	eerste helft beleidsperiode
Terminologie	ENT-activiteiten (update website, Nieuwsbrief Termenlijsten	doorlopend

	Termtreffer: productieversie	afhankelijk van externe financiering
	Terminologisch werk: zorgterminologie, juridische terminologie, Nederlands als wetenschapstaal	afhankelijk van externe financiering
Grammatica	e-ANS-update (data en technisch)	doorlopend
	Update didactische laag e-ANS	doorlopend
	Update Taalportaal (inhoud en/of technisch)	eerste helft beleidsperiode
	Grammaticaportaal versie 1 (basismodules)	tweede helft beleidsperiode
	Grammaticaportaal versie 2 (termen/projecten)	tweede helft beleidsperiode
	Constructicon-onderzoeksrapport	tweede helft beleidsperiode
Nationale en internationale samenwerkingsverbanden	Onderhouden en uitbreiden netwerk	doorlopend
	Projecten	doorlopend
	Overige infrastructurele dienstverlening	doorlopend
IT-capaciteit	In stand houden en uitbreiden hardware-ondersteuning	doorlopend
	In stand houden en uitbreiden kennis en opleiding	doorlopend
Onderwijs	Wetenschappelijk onderwijs	doorlopend
	Infrastructurele ondersteuning	doorlopend
PR & Communicatie	vernieuwing website t.b.v. academisch publiek	eerste helft beleidsperiode
	Overige activiteiten (social media, nieuwsbrieven, podcasts, video's, webinars, evenementen etc.)	doorlopend

Verklarende lijst van termen en afkortingen

Term/afkorting	Verklaring
AI	Artificial Intelligence
ANW	Algemeen Nederlands Woordenboek
API	Application Programming Interface
BA1	Bachelor 1
BlackLab	de corpuszoeksoftware van het INT
CEFR	Common European Framework of Reference, een uniform systeem van taalniveau-indeling dat gebruikt kan worden voor alle Europese talen.
CELEX	multilinguale lexicale databank ontwikkeld door het *Centre for Lexical Information* van het Max Planck Instituut voor Psycholinguïstiek
CHN	Corpus Hedendaags Nederlands
CLARIAH+	Common Lab Research Infrastructure for the Arts and Humanities
CLARIN ERIC	Common Language Resources and Technology Infrastructure - European Research Infrastructure Consortium
CLARIN-B-centrum	*Service Providing Centre* binnen het CLARIN-netwerk
COST	Cooperation in Science and Technology
CRR-datasets	psycholinguïstische lexicale datasets ontwikkeld door het *Center for Reading Research* aan de Universiteit Gent
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DATECH	Digital Access to Textual Cultural Heritage
DDD	Database of Dutch Dialects
DiaMaNT	Diachroon seMantisch lexicon van de Nederlandse Taal
DPC	Dutch Parallel Corpus
DSDD	Database of Southern Dutch Dialects
DuELME	Dutch Electronic Lexicon of Multiword Expressions
e-ANS	elektronische Algemene Nederlandse Spraakkunst
EFNIL	European Federation of National Institutions for Language
ELE	European Language Equality
ELEXIS	European Lexicographic Infrastructure: Horizon 2020 Research Infrastructure project
ELRC	European Language Resource Coordination
ENT	Expertisecentrum Nederlandstalige Terminologie
EWN	Etymologisch Woordenboek van het Nederlands

eWND	elektronische Woordenbank van de Nederlandse Dialecten
FrameNet	type lexico-grammaticale databank, voor het Nederlands momenteel in ontwikkeling aan de Vrije Universiteit Amsterdam
FWO	Fonds voor Wetenschappelijk Onderzoek - Vlaanderen
GCND	Gesproken Corpus van de Zuidelijk-Nederlandse Dialecten
GiGaNT	Groot Geïntegreerd Lexicon van de Nederlandse Taal
GLAD	Global Anglicism Database Network
GPU	Graphics Processing Unit
GTB	Geïntegreerde Taalbank
Hilex	Historisch lexicon
Horizon 2020	Programma voor onderzoeksfinanciering van de Europese Commissie
HSN	Het Schoolvak Nederlands
IMPACT	IMProving ACcess to Text
IPR	Intellectual Property Rights
K-Dutch	CLARIN Knowledge Centre Dutch
LASSY	Large Scale Syntactic Annotation of Written Dutch
LLOD	Linguistic Linked Open Data
MNW	Middelnederlandsch Woordenboek
Molex	Modern lexicon
NL-Term	veldorganistatie ter bevordering van Nederlandstalige Terminologie
NLP	Natural Language Processing
NT2	Nederlands als tweede taal
NT2Lex	lexicale databank met CEFR-informatie voor het NT2, ontwikkeld door het *Centre de traitement automatique du langage* aan de UC Louvain
NWO	Nederlandse Organisatie voor Wetenschappelijk Onderzoek
ODISSEI	Open Data Infrastructure for Social Science and Economic Innovations
ODWN	Open Dutch WordNet, ontwikkeld aan de Vrije Universiteit Amsterdam
OMBI	Omkeerbare Bilinguale Bestanden
ONW	Oudnederlands Woordenboek
PR	Public Relations
R&D	Research & Development
RBBN	Referentiebestand Belgisch Nederlands
RBN	Referentiebestand Nederlands
SABeD	Spoken Academic Belgian Dutch: corpuscompilatieproject

SaDilar	South African Centre for Digital Language Resources
SoNaR	STEVIN Nederlandstalig Referentiecorpus
SSHOC-NL	Social Science and Humanities Open Cloud for the Netherlands
SWOW	Small World of Words: databank van woordassociaties ontwikkeld door de ConCat-onderzoeksgroep aan de KU Leuven
TDN	Tagset Diachroon Nederlands
TEI	Text Encoding Initiative
TiNT	Terminologie in het Nederlandse Taalgebied
UniDive	Universality, diversity and idiosyncrasy in language technology (COST-actie)
VMNW	Vroegmiddelnederlands Woordenboek
WNT	Woordenboek der Nederlandsche Taal
WNW	Woordenboek van Nieuwe Woorden
WSD	Word Sense Discrimination
WSI	Word Sense Induction

Bijlage I: Werkzaamheden voor de corpusbouw en corpusexploratie

Corpusbouw

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 de komende beleidsperiode hebben zowel betrekking op de verrijking als op de workflow.

Verrijking met woordsoort en lemma

In grote lijnen zullen de werkzaamheden gericht zijn op een harmonisering van de taalkundige verrijking in de diverse corpora en op een kwalitatieve verbetering van de verrijking.

De woordsoorttoekenning voor hedendaags en historisch Nederlands wordt geharmoniseerd door ook het hedendaags Nederlands taalmateriaal te verrijken conform de richtlijnen van de TDN[24]. Bij de lemmatisering zal het GiGaNT-lexicon ingezet worden met toekenning van GiGaNT-lemma-id's om zo de koppeling met de lexicale infrastructuur te vereenvoudigen

Er zal ook ingezet worden op een hogere accuraatheid van de taalkundige verrijking door het inzetten van state-of-the-art op deep learning gebaseerde technieken met behulp van omvangrijker trainingsmateriaal. Dat betekent dat het INT gaat investeren in de conversie en curatie van bestaand trainingsmateriaal en dat het met name voor historisch Nederlands ook nieuw trainingsmateriaal zal bijmaken[25]. Door gebruik te maken van domeinadaptatie[26] hoopt het INT het manuele werk zoveel mogelijk te beperken.

Syntactische annotatie

Om de mogelijkheden voor het extraheren van informatie uit corpusdata te verbeteren (onder andere woordcombinaties) zal het INT voor het hedendaags Nederlands het monitorcorpus voorzien van syntactische verrijking. 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+).[27] Er zal onderzocht worden 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. Daar waar mogelijk wil het INT de metadata voor historisch en modern corpusmateriaal nog

[24] Tagset voor Diachroon corpusmateriaal van het Nederlands https://ivdnt.org/wp-content/uploads/2021/05/TDN_INT_WP_1.pdf

[25] 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.

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

[27] 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.

verder uniformeren. Het metadatamodel van de corpora zal worden geformaliseerd en gepubliceerd. Daarnaast wil het INT voor het hedendaags Nederlands onderzoeken hoe het domein- en onderwerpsclassificatie kann toepassen op het monitorcorpus om semi-automatisch de corpusmetadata te verrijken.

Workflow voor conversie, verrijking en datamanagement

De infrastructuur voor corpusconversie en corpusdatamanagement dient te worden aangepast aan de vereisten die de steeds grotere hoeveelheid data en de toevoeging van nieuwe annotatielagen stellen.

Voor het hedendaags Nederlands corpusmateriaal is daarvoor een omgeving ingericht voor opslag en dataprocessing, waarbij het proces van ruwe digitale data naar geïndexeerde verrijkte data zoveel mogelijk is geautomatiseerd. Aan dat proces moeten de bovenvermelde nieuwe annotatielagen worden toegevoegd. Daarnaast zal gewerkt worden aan een verbeterde schaalbaarheid, door middel van optimalisatie van de dataprocessing en databasestructuur en zo nodig gedistribueerde opzet. De verbeterde infrastructuur voor modern Nederlands wordt het model voor de in te richten infrastructuur voor historisch Nederlands.

Corpusexploratie

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[28] en de corpus-frontend[29], 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. Deze werkzaamheden worden nader gespecificeerd in bijlage II onder “Informatie-extractie uit Corpora”.

[28] https://github.com/INL/BlackLab
[29] https://github.com/INL/corpus-frontend

Bijlage II: Werkzaamheden voor de uitbouw van de kennisbank

De verdere uitbouw van de lexicografische infrastructuur is een programma dat de komende vijf beleidsjaren overschrijdt. Het biedt vele mogelijkheden om op een efficiëntere manier te werken, maar niet zonder de nodige uitdagingen. We noemen er een paar.

Het modulair bewerken van de kennisbank en de omschakeling van een grotendeels hiërarchisch en artikel-gebaseerd datamodel naar een meer modulaire en relationele opzet, geeft meer mogelijkheden om binnen de modules efficiëntere en beter technisch ondersteunde bewerkingstrajecten in te zetten. Echter, het maakt de structuur van de kennisbank meer complex vanwege de vele interdependenties. Het vereist bovendien het volledig herdenken van een deel van de databewerking en het impliceert dat de relatie tussen kennisbank en eindproducten voor diverse doelgroepen in sommige opzichten complexer is.

We willen de beschrijving van historisch en modern Nederlands meer naar elkaar toetrekken, alleen hebben we te maken met een verschillende uitgangssituatie. Voor hedendaags Nederlands hebben we uitgebreide corpora en naar behoren functionerende technieken voor automatische verrijking terwijl deze punten voor historisch Nederlands nog veel aandacht vergen. Daarnaast is er voor het hedendaags Nederlands gewerkt aan lexicografische eindproducten zoals het ANW, het Woordenboek van Nieuwe Woorden (WNW) en Woordcombinaties, terwijl voor het historisch Nederlands reeds geïnvesteerd is in de uitbouw van een lexicografische infrastructuur, gebaseerd op de bestaande historische woordenboeken, met als resultaat het computationeel lexicon GiGaNT en het diachroon semantisch lexicon DiaMaNT. Bovengenoemde aspecten vinden hun weerslag in de geplande werkzaamheden voor de komende vijf jaar die hierna beschreven worden.

Datamodel, formaten en annotatierichtlijnen

Voor het succesvol uitwerken van een complexe data-infrastructuur voor de kennisbank voor de Nederlandse Taal is een doordacht datamodel essentieel dat rekening houdt met reeds bestaande onderdelen in de huidige infrastructuur. Naast de samenhang van de modules van de kennisbank behoeft het model op een aantal punten verdere uitwerking en verfijning:

- De classificatie en modellering van meerwoordsexpressies vergt nog nadere uitwerking
- De modularisering van de kennisbank moet strakker worden uitgewerkt en toegepast

Behalve het formele aspect van het datamodel, i.e. de beschrijving van datacategorieën, de relationele structuur van de koppeling tussen de datacategorieën en het annotatievocabulaire, is het van groot belang dat ook de annotatierichtlijnen goed gedocumenteerd zijn. Voor de Tagset Diachroon Nederlands (TDN) is reeds een uitgebreide beschrijving beschikbaar en ook de lemmatiseerprincipes zijn reeds gepubliceerd. Beide documenten zullen naar verwachting een update krijgen.

Zoals in de inleiding al aangegeven met de term “Relationele Open Data”, heeft het uit te werken datamodel veel gemeen met de voor de LLOD-cloud vigerende Linguistic Linked Open Data Cloud (LLOD-cloud), en zal ook zeker een mapping naar dit framework gemaakt worden. We zullen echter om de aansluiting van de in de bewerkingsomgevingen gebruikte relationele databases transparant te houden dichter bij het relationele model blijven.

GiGaNT : diachroon computationeel lexicon

Het fundament van de kennisbank is het computationeel lexicon GiGaNT, dat toelaat om lexicale databanken op lemmaniveau te koppelen. De ontwikkeling van GiGaNT gebeurt modulair. Er is simultaan gewerkt aan de historische component (Hilex) en de moderne component (Molex) en er is

een eerste koppeling tussen de twee componenten gerealiseerd. Voor Hilex lag de focus op de integratie van de data van het *Middelnederlandsch Woordenboek* en het *Woordenboek der Nederlandsche Taal.* Voor Molex lag de focus enerzijds op het aanleveren van de nodige data voor *woordenlijst.org* en anderzijds op de koppeling van Molex met interne lexicale data en integratie van Molex in de workflow voor de producten *ANW*, *WNW* en *Woordcombinaties*.

Hilex heeft op woordvormniveau een koppeling met attestaties (bewijsplaatsen voor de woordvormen die bij ieder lemma opgenomen zijn). Het zijn de citaten uit de voor de woordenboeken gebruikte corpora. De lemmata uit Molex zijn gebaseerd op zowel corpusdata als lexicale bronnen. Corpusevidentie ontbreekt echter nog. Het uiteindelijke doel is dat digitaal corpusmateriaal (historisch en hedendaags Nederlands) gescreend wordt op lacunes en nieuw te beschrijven woordenschat, en dat de relevante woorden met een link naar de nodige corpusevidentie opgenomen worden in GiGaNT.

Tot op heden is de morfologische informatie in GiGaNT zeer beperkt aanwezig. Het uitbouwen van een morfologische component is belangrijk om verschillende redenen: a. morfologische analyse kan helpen bij het controleren van de consistentie van de toekenning van woordgeslacht en paradigma-uitbreiding; b. morfologische productiviteit is een indicator die relevant is voor de overlevingskans van nieuwe woorden en c. het biedt, zowel diachroon als synchroon, informatie over woordvorming.

Voor de komende beleidsperiode wil het INT, naast reguliere werkzaamheden ten behoeve van de diverse producten die een koppeling met GiGaNT hebben, de volgende zaken aanpakken:

- Het uitbreiden van GiGaNT met de data uit het *Oudnederlands Woordenboek* (ONW) en het *Vroegmiddelnederlands Woordenboek* (VMNW) zodat GiGaNT alle taalfasen van het Nederlands bevat
- Het uitwerken van de infrastructuur met een workflow om corpusmateriaal systematisch te screenen op lacunes en op nieuwe woorden, maar ook met een omgeving om deze data makkelijk te analyseren, structureren en bewerken. Focus wordt hierbij het hedendaags Nederlands. Het doel is om het CHN dat wekelijks geüpdatet wordt systematisch te screenen ten behoeve van de uitbreiding van Molex, en daarbij frequentie-informatie aan Molex toe te voegen als eerste vorm van corpusevidentie.
- De verdere uitbouw van de morfologische component van GiGaNT, en de ontwikkeling van de nodige tooling daarvoor.
- De verdere verbetering van de parsering van de data van de historische woordenboeken, die niet alleen ten goede komt van GiGaNT maar ook van de semantische component. Daarvoor zal een nieuwe werkomgeving ingericht worden.
- Het overeenstemmen van GiGaNT met het datamodel van de kennisbank

Betekenisregister

De koppeling van lexicale data op *betekenisniveau* is de belangrijkste uitdaging voor de realisatie van de kennisbank. Betekenisinformatie is aanwezig in de diverse lexicografische producten van het INT. Voor de woordenschat tot en met ca. 1976 zijn dat de historische woordenboeken van het INT. De woordenboekdata zijn grotendeels geparseerd, waardoor betekenissen, combinaties en citaten (elk met persistent identifiers) uit de woordenboekdata gehaald kunnen worden. Betekenisinformatie voor het hedendaags Nederlands zit verder in de lopende projecten als het ANW, WNW, Woordcombinaties en in beperkte mate in Molex (glossen bij homoniemen) en in lexicografische producten als het RBN en de vertaalwoordenboeken.

AI en met name deep-learning-technieken hebben geleid tot verbetering van de state of the art van corpus-gebaseerde semantische technieken als sense alignment, word sense disambiguation en word

sense induction. We zullen onderzoeken hoe deze mogelijkheden de werkprocessen kunnen ondersteunen.

De eerste stap daarin is het opbouwen van een betekenisinventaris voor een kernvocabulaire. We bouwen die op vanuit het hedendaags Nederlands en gaan daarbij onder andere uit van de definities uit bronnen die we afgelopen beleidsperiode op lemmaniveau aan GiGaNT-Molex hebben gekoppeld (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. Net zoals DiaMaNT (het Diachroon seMantisch lexicon van de Nederlandse Taal) nu al definities koppelt uit verschillende historische woordenboeken op een schematisch niveau, zal het betekenisregister worden uitgebreid naar de hedendaagse lexicale databanken en naar andere types lexicografische informatie. Zo wordt voor de kernwoordenschat alvast ook een koppeling gemaakt met corpusdata. Het doel hiervan is enerzijds om corpusevidentie toe te voegen aan de koppelbetekenissen, maar ook om het geannoteerde corpus als trainings- en evaluatiemateriaal in te zetten voor taken als word sense induction/disambiguation.

Hieruit volgen een paar concrete probleemstellingen:

- Hoe vergelijken we precies de omschrijvingen uit verschillende lexicale bronnen? Kunnen automatische sense-alignmenttechnieken ons daarbij helpen, zoals bijvoorbeeld geïmplementeerd in het in ELEXIS verder ontwikkelde NAISC[30]?
- Hoe relateren we informatie uit andere modules (woordcombinaties, corpusattestaties, semantische relaties) aan de koppelbetekenissen? Kunnen moderne benaderingen voor word sense disambiguation, die minder trainingsmateriaal veronderstellen, hierbij behulpzaam zijn?
- Kunnen WSI- (*word sense induction*) gerelateerde technieken de werkprocessen ondersteunen om te komen tot een voorsortering van corpusattestaties voor woorden waarvoor nog helemaal geen betekenisomschrijvingen zijn vastgesteld?
- Hoe combineren we de automatische heuristieken en handmatig werk in een werkomgeving?

Naast de beschrijving van het kernvocabulaire zal verder gewerkt worden aan de beschrijving van neologismen die gepubliceerd worden in het WNW. Nieuw wordt het toekennen van definities aan historische woorden die nog geen beschrijving hebben gehad in de historische woordenboeken. Bij de beschrijving daarvan zal meteen een koppeling met de betekenisinventaris gemaakt worden. Op termijn zal ook de werkwijze van het WNW hieraan aangepast worden.

Conceptrelaties vormen ook een onderdeel van het betekenisregister. Het ANW bevat informatie over conceptrelaties en conceptrelaties vormen een belangrijk onderdeel van de DiaMaNT-module van de infrastructuur voor historisch Nederlands. Daar waar DiaMaNT tot nog toe alleen met de data van de historische woordenboeken werkte, zal voor de verdere uitbouw van het lexicon een koppeling gemaakt worden met het hedendaags Nederlands via de koppelbetekenissen. Daarbij zal de focus liggen op het vastgestelde kernvocabulaire.

Tot slot zal gewerkt worden aan de combinatiecomponent van het betekenisregister, en dat vooralsnog voor het hedendaags Nederlands. Combinaties en meerwoordexpressies vormen de kern van het project Woordcombinaties maar ook het ANW en de diverse historische woordenboeken besteden aandacht aan de beschrijving ervan. Er zal onderzocht worden hoe de beschrijving meer gecentraliseerd kan worden, waarbij een belangrijke rol weggelegd is voor het project

[30] NAISC is een tool voor datasetlinking op basis van textual similarity; https://github.com/insight-centre/naisc.

Woordcombinaties. Verder lexicografisch werk aan combinaties zal gebeuren door het screenen van het monitorcorpus op combinaties die vervolgens in een daartoe geschikte omgeving beoordeeld en gelabeld kunnen worden. Uiteindelijk zullen de combinaties gelinkt worden aan koppelbetekenissen.

Ondersteuning van de lexicale bewerking

Het modulair bewerken van de kennisbank en de omschakeling van een grotendeels hiërarchisch en artikel-gebaseerd datamodel naar een meer relationele opzet vereist aangepaste bewerkingsomgevingen. Voor deze werkomgevingen zal Lex'it[31] een belangrijke rol spelen. Omgevingen die gerealiseerd worden zijn:

- Omgeving voor het opbouwen van de betekenisinventaris
- Omgeving voor het linken van de koppelbetekenissen aan corpusdata
- Omgeving voor het bewerken van onbeschreven woorden
- Omgeving voor de analyse en beoordeling van combinaties uit corpusdata.

Informatie-extractie uit corpora

Voor de extractie van relevante informatie ten behoeve van de lexicografische beschrijving van het Nederlands zullen we de corpusretrievalomgeving verder uitbouwen. Daarnaast zal onderzoek gedaan worden naar de inzetbaarheid van distributionele technieken en AI om semantische informatie uit de corpusdata te halen. Hierna komt een gestructureerde opsomming van de geplande activiteiten.

Voor wat betreft de backend van de corpusretrievalomgeving (BlackLab, BlackLab Server) staan de volgende doelstellingen voorop:

- 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.[32] Deze werkzaamheden worden mede in de context van CLARIAH+ uitgevoerd.
- Uitbreiding van de functionaliteit om efficiënt statistieken uit het materiaal te extraheren. Hierbij denken we met name aan diachrone frequentieprofielen en andere taalvariationele (lectale) distributies en het uitbreiden van de resultaatweergavemogelijkheden om het combinatiegedrag van woorden inzichtelijk te maken.
- Ondersteuning van de extractie van meerwoordspatronen op basis van de syntactische verrijking (relationele collocatiecomponent) en samenbrengen van die informatie in een profiel van het combinatiegedrag (vergelijkbaar met, maar niet identiek aan de Wordsketch-functionaliteit van Sketch Engine of het Wortprofil van het Digitales Wörterbuch der deutschen Sprache), op zijn minst van een lemma, maar waar mogelijk ook voor een specifieke betekenis.
- Functionaliteit om representatieve en informatieve attestaties te selecteren ("Good Dictionary Examples") maar ook met aandacht voor de representatie van het betekenis- en combinatieprofiel van een woord.

Voor wat betreft de userinterface zal geïnvesteerd worden in:

[31]Lex'it is een op PostgreSQL en datatables.net gebaseerd rapid application development platform, ontwikkeld op het INT, dat ingezet wordt voor de ontwikkeling van omgevingen voor de bewerking van gestructureerde data.

[32]We zijn ons bewust van het bestaan van zoekmachines voor diepe syntactische bomen, zoals GrETEL, PaQu en PML-Tree Query maar deze zijn uiterst langzaam voor het doorzoeken van heel grote hoeveelheden data.

- 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.
- Query-building voor syntactische retrieval.
- Verbeteren van de mogelijkheid om met voorgedefinieerde en door de gebruiker gedefinieerde subcorpora te werken.

Voor de efficiëntere extractie van semantische informatie uit corpora zal onderzocht worden hoe distributionele technieken en grootschalige voorgetrainde taalmodellen in de corpusanalyse-workflow geïntegreerd zouden kunnen worden. We zullen ons hierbij richten op:

- Automatische word (sense) disambiguation (WSD) voor woorden waarvoor reeds een betekenisprofiel bestaat. Behalve de lemmatisering met lemma-id's (zie bijlage I), die eigenlijk al een grove WSD van homoniemen impliceert, onderzoeken we ook de mogelijkheden om fijnmazigere betekenisonderscheidingen te herkennen, om het koppelen van attestaties aan betekenissen te ondersteunen en betekenisspecifieke woordcombinatieprofielen te maken.
- WSI- (word sense induction) gerelateerde technieken, om te komen tot een voorsortering van corpusattestaties. Daarbij wordt niet zozeer gedacht aan een volautomatische indeling, maar meer aan een werkomgeving die de lexicograaf helpt het betekenisprofiel van een woord te verkennen.

Bijlage III: Wetenschappelijke visie op lexicografie

De beschrijving van de Nederlandse woordenschat in al zijn facetten is en blijft een van de kerntaken van het INT: van de vroegste tijden tot heden, van standaardtaal via groepstalen tot dialecten, van basiswoordenschat via academisch taalgebruik tot vaktaal, in eerste instantie monolinguaal maar ook multilinguaal met het Nederlands als spil. Het INT heeft een lange traditie van wetenschappelijk onderbouwde, evidentie-gebaseerde woordenschatbeschrijving en die resulteert zowel in zorgvuldig samengestelde en verrijkte corpora als in kwalitatief hoogstaande lexicografische producten. Daarmee komt het instituut tegemoet aan de noden van een uiterst divers doelpubliek, gaande van wetenschappelijke onderzoekers en taalleerders tot ontwikkelaars van software-applicaties. Het INT biedt toegang tot die materialen op verschillende manieren, gaande van online doorzoekbare corpora en raadpleegbare naslagwerken tot downloadbare datasets en API's (Application Programming Interfaces) voor geautomatiseerde toepassingen. De corpora en lexicale databanken van het INT zijn daarmee nu al een cruciaal onderdeel van de digitale taalinfrastructuur voor het Nederlands. Echter, de versnelde digitalisering in alle geledingen van de kennismaatschappij heeft de eisen aan die taalinfrastructuur in de laatste jaren grondig veranderd. Vooral voor Onderzoek en Ontwikkeling (R&D) door kennisinstellingen en bedrijven is het belang van op zich staande taalmaterialen en lexicografische producten relatief verminderd ten gunste van geïntegreerde kennisbanken waaruit makkelijk specifieke datasets voor R&D-doeleinden geëxtraheerd kunnen worden. Enerzijds zijn onderzoekers, ook in de humane wetenschappen, immers alsmaar meer computationeel onderlegde *datawetenschappers*. Ook andere taalgebruikers verwachten steeds meer gepersonaliseerde taalondersteuning die naadloos geïntegreerd is binnen andere applicaties eerder dan via online raadpleegbare naslagwerken. Dit stelt behoorlijk wat uitdagingen voor de lexicografie als geheel. Het INT is zowel door zijn interne projecten van de afgelopen jaren als door zijn vooraanstaande rol in Europese projecten, zoals het onderzoeksinfrastructuurproject ELEXIS,[33] goed geplaatst om aan die veranderende noden en eisen aan de taalinfrastructuur tegemoet te komen. Op het vlak van corpusuitbouw en woordenschatbeschrijving zal het instituut in de komende 5 jaar dan ook nóg sterker inzetten op twee strategisch belangrijke langetermijnontwikkelingen: (1) de versterking van de relatie tussen empirische evidentie (corpora) en woordenschatbeschrijving en (2) de integratie van alle componenten van de woordenschatbeschrijving in één geïntegreerde, relationele kennisbank van de Nederlandse woordenschat door de eeuwen heen. Daarmee beoogt het instituut drie dingen: (a) een versterking van de wetenschappelijke onderbouwing en kwaliteit van de taaldocumentatie, (b) een optimalisatie van de eigen dataverwerkingsprocessen door efficiëntiewinsten en een sterkere integratie van AI en taaltechnologie en (c) een verbreding van de toepassingsmogelijkheden van de taalinfrastructuur zowel voor externe R&D als voor de verdere ontwikkeling van eigen eindproducten. In wat volgt wordt besproken wat de motivatie is voor een geïntegreerde kennisbank, hoe die zal opgebouwd worden binnen een gemodulariseerd lexicografisch proces, en hoe een relatie tussen corpusdata en woordenschatbeschrijving versterkt zal worden. Tenslotte wordt een toekomstperspectief voor de geïntegreerde kennisbank geschetst.

Lexicografische modules in een geïntegreerde kennisbank

In een traditionele lexicografische benadering werden de verschillende types informatie over de woordenschat ondergebracht in verschillende, op zichzelf staande eindproducten, d.w.z. woordenboeken en lexica, die in eerste instantie als naslagwerk bedoeld waren voor een welomschreven doelgroep. Zo zijn definities en gebruiksinformatie te vinden in een hedendaags woordenboek voor een breed publiek, frequentie-informatie in een frequentiewoordenboek voor een

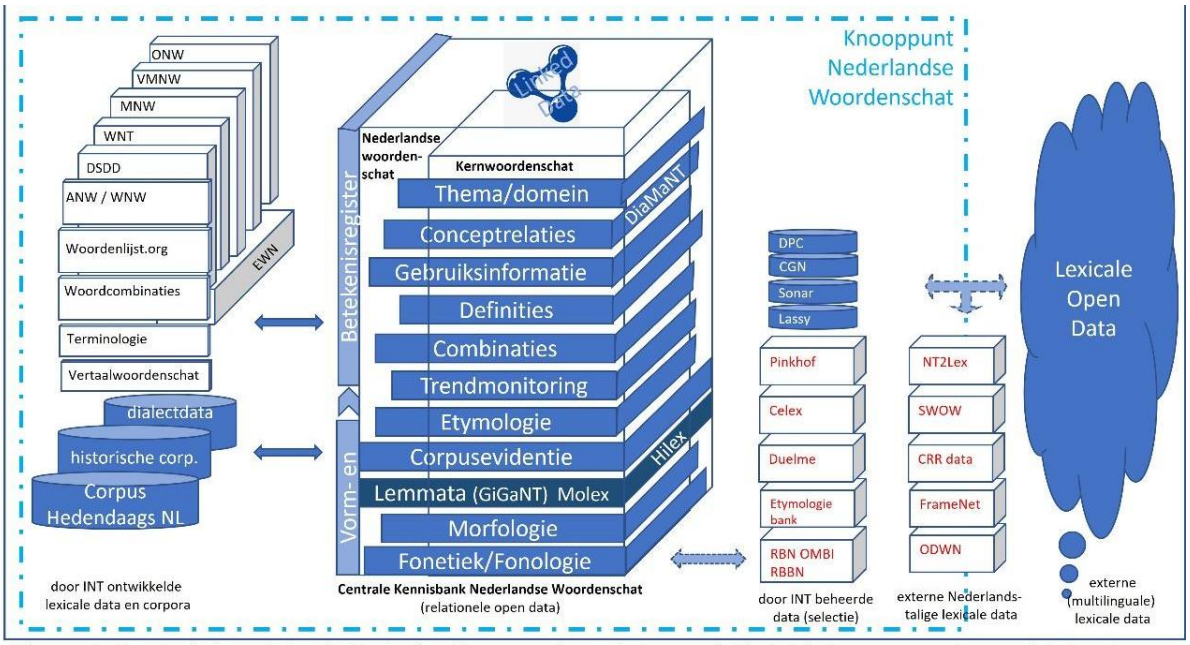
[33] https://elex.is/

statistisch geïnteresseerd publiek, collocaties en verbindingen in een combinatiewoordenboek voor educatieve doeleinden, synoniemwoordenboeken en spellingslexica bieden schrijfhulp, de vroegere taalstadia worden beschreven in historische woordenboeken, de etymologie in een etymologisch woordenboek enzovoort. Ook de elektronische versies van deze woordenboeken bleven veelal geconcipieerd als een online raadpleegbaar en op zichzelf staand eindproduct. In de laatste decennia is er in de internationale lexicografie echter de consensus gegroeid dat de toekomst van de lexicografie niet ligt in het beschrijven van de woordenschat in op zichzelf staande eindproducten maar veeleer in het aanleggen van een geïntegreerde kennisbank waarin alle informatie en kennis over woorden wordt samengebracht en onderling gekoppeld en waaruit dan weer diverse eindproducten voor specifieke doelgroepen en toepassingen kunnen worden afgeleid. Enerzijds maakt zo'n integratie behoorlijke efficiëntiewinsten mogelijk omdat dezelfde informatie die vroeger over meerdere woordenboeken verspreid zat, nu aan elkaar gekoppeld is en centraal beheerd, bewerkt en op consistentie gecontroleerd kan worden. Anderzijds laat een geïntegreerde kennisbank ook nieuwe types van gebruik toe, met name voor wetenschappelijk onderzoek en voor de ontwikkeling van applicaties waarin woordinformatie op maat aangeboden wordt eerder dan in een apart naslagwerk. De meerwaarde van een geïntegreerde kennisbank wordt nog groter als die ook nog eens *intern* gekoppeld is aan de corpusdata die aan de taalbeschrijving ten grondslag ligt, en *extern* aan andere lexicale databanken, die bijvoorbeeld meertalige, encyclopedische of psycholinguïstische informatie bevatten over woorden en de concepten waarnaar ze verwijzen.

Als toonaangevend lexicografisch instituut op Europees vlak is het INT een van de wegbereiders van deze geïntegreerde benadering. Er is in de afgelopen jaren al sterk geïnvesteerd in een centraal lexicon, GiGaNT (Groot Geïntegreerd Lexicon van de Nederlandse Taal), met morfologische en spellingsinformatie voor alle woorden van het Nederlands van vroeger (Hilex) tot nu (Molex). Nu al kunnen de vormelijke ontwikkelingen van elk (etymologisch verschillend) lemma door alle taalstadia van het Nederlands gevolgd worden. Ondertussen zijn bijna alle hedendaagse en historische woordenboeken en lexica die het INT beheert *op lemma-niveau* gekoppeld aan GiGaNT en zo dus ook aan elkaar. Nu al wordt alle vorm- en flexie-informatie centraal aangemaakt en beheerd en komt dan vanuit GiGaNT in de verschillende lexicografische eindproducten van het INT terecht. Bovendien is GiGaNT ook al deels beschikbaar als downloadbare dataset en API-service, die bijvoorbeeld door de Koninklijke Bibliotheek gebruikt wordt om historische teksten beter te ontsluiten. Op deze centrale databank voor woordvormen en lemmata wil het INT de komende jaren verder bouwen om geleidelijk aan ook alle woordenschatbeschrijving op betekenisniveau te integreren en beheren in één centrale kennisbank[34] van de Nederlandse woordenschat. In eerste instantie zullen de woordenboeken voor het algemeen hedendaags en het historisch Nederlands in de kennisbank samengebracht worden. Op termijn zullen ook de terminologiebanken, de dialectdatabanken en eventuele lexicon-georiënteerde grammaticabeschrijvingen (*constructicon*) geïntegreerd en gekoppeld worden. De centrale kennisbank zal enerzijds als basis dienen voor het continueren van bestaande en het ontwikkelen van nieuwe INT-eindproducten, en anderzijds, in de mate van het mogelijke,[35] ter beschikking gesteld worden als *relationale open data*[36] voor externe O&O. Lexicografische gegevens koppelen op het abstractere betekenisniveau is een grotere uitdaging dan op vormniveau, maar ook hiermee heeft het instituut in de afgelopen jaren al behoorlijk wat ervaring opgedaan: In het project DiaMaNT (Diachroon seMantisch lexicon van de Nederlandse Taal) worden definities en citaten uit de verschillende historische woordenboeken nu al op een schematisch niveau

[34] De technische implementatie kan verschillende vormen aannemen: De kennisbank is niet noodzakelijk één databank, maar kan ook de vorm aannemen van een "datawarehouse" met gekoppelde databanken.
[35] Met inachtneming van IP-restricties.
[36] Verschillende implementaties zijn mogelijk, o.a. publicatie in de llod-cloud.

aan elkaar gekoppeld. Bij de totstandkoming van de Databank van Zuid-Nederlandse Dialecten (DSDD), was het INT verantwoordelijk voor het koppelen van de verschillende dialectwoordenboeken op conceptniveau. In het Europese onderzoeksinfrastructuurproject ELEXIS werden technologie, platformen en testdatasets ontwikkeld om lexicografische databanken zowel binnen één taal als tussen talen onderling op betekenisniveau te linken. Bovendien heeft het INT intense samenwerking met andere Europese instituten die gelijkaardige kennisbanken aan het uitbouwen zijn.[37] Desalniettemin is deze integratieve benadering nog volop in ontwikkeling. Er is nog heel wat O&O nodig zowel binnen het INT als in samenwerking met nationale (Nederlandse en Vlaamse) en internationale partners. Sommige aspecten van de opbouw van de kennisbank zitten daarom inherent nog in experimentele fase. Vanuit een beredeneerde risicoanalyse formuleert het meerjarenbeleidsplan daarom zowel relatief snel haalbare minimumdoelstellingen als meer uitdagende innovaties die weliswaar realistisch zijn maar nog verdere ontwikkeling behoeven. In de volgende paragrafen wordt kort geschetst hoe de macrostructuur en onderdelen van de centrale kennisbank geconcipieerd zijn.



Figuur 1: Centrale kennisbank Nederlandse woordenschat en relatie tot interne en externe databanken

GiGaNT als uitgangspunt op lemma-niveau

Zoals figuur 1 toont, biedt GiGaNT op lemma-niveau al een stevig fundament voor de centrale kennisbank. GiGaNT onderscheidt woorden op etymologische basis, bevat informatie over vormvariatie en heeft voor elk Nederlands lemma een unieke en persistente identifier. In een eerste iteratie is GiGaNT ontwikkeld in twee aparte componenten: Molex voor het modern Nederlands, met als uitgangspunt de spellingsdatabank, en Hilex voor het historisch Nederlands, met als vertrekpunt de historische woordenboeken (WNT, VMNW, MNW, ONW). Eind 2022 zullen beide componenten helemaal aan elkaar gekoppeld zijn via een “super-lemma-ID”. De hedendaagse woordenboeken (ANW, Woordcombinaties en Vertaalwoordschat) zijn nu al gelinkt aan GiGaNT-Molex. Nieuwe woorden, zowel neologismen als bestaande onbeschreven woorden, worden centraal in GiGaNT

[37] Gelijkaardige projecten zijn al aan de gang voor o.a. het Deens (Central OrdRegister for Dansk), Duits (Digitales Wörterbuch der Deutschen Sprache), Sloveens (Slovenscina) en Pools (https://lab.dariah.pl/)

toegevoegd en van vorminformatie voorzien. Vanop deze basis wil het INT in de komende jaren de verschillende modules van de kennisbank verder uitbouwen. Op lemma-niveau gaat het dan om de volgende modules (zie ook figuur 1):

[Fonetiek/Fonologie] Uitspraakinformatie voor de standaardvariëteiten van het Nederlands wordt in de vorm van IPA-transcripties in de loop van 2022 aan Molex toegevoegd en zal nadien standaard ook aan nieuwe woorden toegevoegd worden.

[Morfologie] Morfologische informatie is op dit moment al beperkt aanwezig in GiGaNT, maar in de komende jaren wordt die uitgebouwd tot een volwaardige module waarin elke morfeem van het Nederlands een eigen uniek ID krijgt en gekoppeld is aan alle lemmata waarin het voorkomt. Dit laat intern niet alleen toe om flexievarianten consistenter te behandelen maar biedt ook nieuwe mogelijkheden voor onderzoek naar morfologische productiviteit bij de woordvorming in het Nederlands doorheen de verschillende taalstadia.

[Corpusevidentie] Eveneens op lemma-niveau zullen we geleidelijk meer citaten aan de lemmata toevoegen.[38] Voor historische data komen de citaties in eerste instantie uit de historische woordenboeken en deze zijn gekoppeld via Hilex. Op termijn kunnen corpusvoorbeelden uit de historische corpora ingevoegd worden. Voor het hedendaagse Nederlands zullen corpusvoorbeelden uit de hedendaagse woordenboeken hergebruikt worden en zullen ze verder binnen het vernieuwde, modulaire lexicografische proces toegevoegd worden.

[Etymologie] Omdat GiGaNT woorden op etymologische basis onderscheidt, kan ook etymologische informatie grotendeels op lemma-niveau gekoppeld worden. In eerste instantie komt die data uit het Etymologisch Woordenboek van het Nederlands (EWN), maar op termijn kunnen ook de andere woordenboeken van de Etymologiebank ingebracht worden

[Trendmonitoring] Een trendmonitoring-module zal informatie over de ontwikkelingen in de Nederlandse woordenschat in kaart brengen. Op lemma-niveau gaat het dan o.a. over (relatieve) woordfrequenties of de productiviteit van samenstellingvormende woorden, en dit zowel door de tijd heen als in verschillende variëteiten van het Nederlands. Die trendmonitoring gebeurt in eerste instantie op basis van het monitorcorpus van kranten van het Corpus Hedendaags Nederlands van het INT. Het CHN bevat op dit moment continue en voortdurend aangevulde krantendata vanaf 1999 en ook blogmateriaal. Op middellange termijn is het de bedoeling om het corpus aan te vullen met meer diverse types taalgebruik en gedigitaliseerde data van vóór 1999. Op langere termijn zal het CHN aangesloten worden op de historische corpora die via het INT ter beschikking staan zodat er één continu diachroon monitorcorpus ontstaat.

Een betekenisregister met beschrijvingsmodules als laag bovenop GiGaNT

Zoals verticale balkjes in figuur 1 tonen, zullen de meeste beschrijvingsmodules in de kennisbank niet op lemmaniveau maar op betekenisniveau georganiseerd zijn. De belangrijkste innovatie binnen de kennisbank zal daarom bestaan uit de aanleg van een *betekenisregister* bovenop GiGaNT. Dat moet toelaten om informatie uit uiteenlopende lexicografische datasets en de toekomstige beschrijvingsmodules te koppelen op betekenisniveau. Een betekenisregister, of *sense inventory,* bevat een overzicht van de hoofdbetekenissen van elk woord. De betekenissen in het register zijn geen traditionele betekenisindelingen, zoals die als definities in een woordenboek te vinden zijn, maar veeleer een pragmatisch werktuig, om uiteenlopende types lexicografische informatie te kunnen

[38] Met citaten worden hier geattesteerde voorbeelden bedoeld die illustratief zijn voor gebruik en betekenis (zogenaamde “knowlegde rich contexts”) die binnen een computationeel ondersteund lexicografisch redactieproces uit corpora en andere bronnen geselecteerd zijn.

koppelen op betekenisniveau. Het is net die koppeling die de meerwaarde van de kennisbank realiseert tegenover *aparte* woordenboeken voor hedendaags en historisch, synoniemen, frequentie-informatie, woordcombinaties, vertaling enzovoort. Deze *koppelbetekenissen* zijn dus containers[39] die enerzijds ruim genoeg moeten zijn om lexicografische databanken met een uiteenlopende semantische granulariteit aan elkaar te kunnen koppelen, maar anderzijds beperkend en distinctief genoeg om relevante betekenisgerelateerde woordkenmerken te kunnen vatten. Bij die laatste horen ook conceptrelaties zoals synonymie of hyperonymie zodat de koppelbetekenissen inherent ook een onomasiologische organisatie in de kennisbank binnenbrengen, zonder zich evenwel vast te leggen op een specifiek cognitieve of ontologische invulling van concepten. Zoals gezegd zijn koppelbetekenissen een pragmatisch werktuig om lexicale databanken te linken.

Om koppeling op betekenisniveau mogelijk te maken moeten de bestaande lexicografische datamodellen grondig worden uitgebreid. De koppelbetekenissen in een betekenisregister zullen, net als de lemma's in GiGaNT, een uniek en persistent ID krijgen en worden geïdentificeerd aan de hand van een geprefereerde definitie en/of hun positie in een semantisch netwerk. Voor een woord als *ezel* zijn de koppelbetekenissen dan enerzijds gelinkt aan de betekenissen en bijhorende definities uit ANW en WNT maar anderzijds ook bepaald door hun hyperonymie-relaties met o.a. (betekenissen van) *dier*, *persoon* en *voorwerp*. Koppelbetekenissen zijn echter niet strak afgebakend en in het datamodel zal expliciet de mogelijkheid voorzien zijn dat lexicografische bronnen en corpusvoorbeelden niet altijd restloos één-op-één koppelbaar zijn.

In de centrale kennisbank zullen de verschillende beschrijvingsmodules dus niet alleen op lemma-niveau aan GiGaNT, maar ook op (koppel)betekenisniveau aan het betekenisregister gelinkt zijn, en op die manier dus ook onderling aan elkaar. Het gaat om volgende modules (zie figuur 1):

[Corpusevidentie] Citaten en corpusvoorbeelden, gedesambigueerd naar hoofdbetekenissen[40]
[Trendmonitoring] Mits er voldoende gedesambigueerde corpusvoorbeelden zijn, kunnen trends ook op betekenisniveau in kaart gebracht worden
[Combinaties] Collocaties en meerwoorduitdrukkingen gekoppeld aan de respectieve hoofdbetekenissen van de lemmata
[Definities] Betekenisbeschrijvingen afkomstig uit verschillende lexicografische bronnen
[Gebruiksinformatie] Geografische, situationele, sociolinguïstische of pragmatische kenmerken van woordbetekenissen)
[Conceptrelaties] Synonymie, taxonomische relaties (hyperonymie, hyponymie) en ontologische relaties (heeft_functie, heeft_kenmerk)
[Thema/domein] Indelingen van de woordenschat op een macroniveau naar thema, vakdomein, leerdersniveau, etc..

Dit belet natuurlijk niet dat in elke module ook informatie op een meer fijnmazig betekenisniveau kan opgenomen zijn. Ook gedetailleerde woordstudies zoals die in het WNT of ANW zullen zo gekoppeld zijn aan informatie uit andere bronnen en modules.

Eens operationeel, zal de centrale kennisbank de inhoud aanleveren voor de bestaande lexicografische eindproducten maar ook toelaten om eenvoudiger nieuwe eindproducten te ontwikkelen, al dan niet in samenwerking met externe partners. Voor elk van die eindproducten zal een redactie een selectie van de lexicografische gegevens uit de centrale kennisbank kunnen maken die het meest geschikt is

[39] In zekere zin zijn ze de reïncarnatie van de sorteerbakjes uit de pre-digitale lexicografie waarin definities uit oudere woordenboeken en citaten uit bronmateriaal grofmazig georganiseerd werden in hoofdbetekenissen.
[40] Hierdoor ontstaat dan ook een deelcorpus met betekenisdesambiguering vergelijkbaar met SemCor-corpora en de integratie van het bestaande DutchSemCor (Vossen et al. 2012) zal ook onderzocht worden

voor een bepaald doelpubliek of een applicatie. Als tijdens die productontwikkeling de behoefte aan ontbrekende informatie geformuleerd wordt, kan de kennisbank daarmee uitgebreid worden zodat de nieuwe informatie meteen ook voor andere of toekomstige toepassingen beschikbaar is. Bij wijzigingen aan bestaande informatie zal de "backward compatibility" gewaarborgd worden door gebruik van labels als "current" en "deprecated".

Kennisbankcompilatie

De opbouw van de kennisbank zal uitgevoerd worden in verschillende stappen, deels in parallel en met geleidelijke aanpassing van de huidige workflows. De content in de modules zal in eerste instantie uit de bestaande woordenboeken geëxtraheerd worden en daarna doorlopend aangevuld worden binnen de vernieuwde lexicografische workflows. We lichten beide stappen in de compilatie van de kennisbank hieronder verder toe.

In een eerste fase zullen uit de huidige woordenboeken verschillende lexicografische informatiecategorieën geëxtraheerd worden om als *gestructureerde data* in de modules van de kennisbank ondergebracht te worden. Wat de historische woordenschat betreft, zijn de vrije tekstvelden uit de oorspronkelijk in druk verschenen woordenboeken in vorige projecten al deels geanalyseerd. In de komende jaren wordt nog een verdere gerichte parsering van de data en metadata in de woordenboekartikelen voorzien om specifieke types informatie zoals afgeleide samenstellingen (opnoemers) en vaste verbindingen te kunnen extraheren en in de modules van de centrale kennisbank te integreren. Voor de hedendaagse woordenschat hebben we vanuit het Algemeen Nederlands Woordenboek (ANW) en de Vertaalwoordenschat al vrij goed gestructureerde data, al zijn er hier ook nog vrije tekstvelden (definitie, collocaten, conceptrelaties) waarvoor we zullen onderzoeken of ze geparseerd en in gestructureerde koppelbare data omgezet kunnen worden. ANW en Vertaalwoordenschat hebben echter een duidelijk verschil in granulariteit van de betekenisbeschrijving en de uitdaging situeert zich hier dus vooral bij de koppeling op betekenisniveau. In eerste instantie zullen we voor een testset de definities uit ANW en Vertaalwoordenschat aan het centraal betekenisregister koppelen. Deze pilot zal een vergelijkingsstandaard creëren voor de verdere ontwikkeling en evaluatie van automatische koppelingssoftware. Voor de overige hedendaagse woordenschat zal de betekenisinformatie uit de verschillende woordenboeken alvast op lemma-niveau bij elkaar gebracht worden om die in een tweede fase door een combinatie van (verbeterde) automatische koppeling en post-editing geleidelijk ook op betekenisniveau aan elkaar te linken. Ten slotte wil het INT ook een koppeling tussen historisch en hedendaags Nederlands realiseren. Omdat dit de grootste uitdaging vormt, zal dit eerst uitgeprobeerd voor een kernwoordenschat die zal bestaan uit een representatieve steekproef van lemmata met verschillende kenmerken (woordsoort, frequentie, polysemie, periode van eerste attestatie etc.). Hoewel deze linking-testcase kan voortbouwen op de ervaring met koppelen uit DiaMaNT, de DSDD en de ontwikkeling van linking-technologie in het ELEXIS-project, wordt toch nog een reële investering in Onderzoek en Ontwikkeling voorzien. Desalniettemin zullen, waar mogelijk, al *gedeeltelijke* koppelingen gerealiseerd worden zodat die relatief snel beschikbaar komen voor externe gebruikers.

In parallel met het overhevelen van bestaande lexicografische data naar de kennisbank, zal ook de workflow voor de beschrijving van nieuwe woorden geleidelijk aan gemodulariseerd worden om nieuwe content aan te leveren voor de centrale kennisbank, waarbij die content van meet af aan maximaal aan de vereisten van gestructureerde en gerelateerde data moet voldoen. Binnen elk van de modules zal telkens het lexicografische proces zelf ook geherorganiseerd worden om de koppeling tussen corpusdata en lexicografische beschrijving sterker te maken en zo ook de integratie van geavanceerde NLP- en AI-technieken te vergemakkelijken (daarover meer in de volgende paragraaf). Zo zullen in de centrale kennisbank corpusdata en betekenisbeschrijving in twee richtingen aan elkaar

gekoppeld worden. Aan lemmata met bestaande koppelbetekenissen zal bijkomende corpusgebaseerde informatie worden toegevoegd zoals citaten, combinatiemogelijkheden, trends of gebruikslabels. Daarbij kan het betekenisregister in het licht van nieuwe corpusevidentie ook geüpdatet worden, zij het met garanties voor terugwerkende compatibiliteit. Omgekeerd zullen nieuwe woorden (neologismen of onbeschreven bestaande woorden) die in onze corpora geattesteerd worden, meteen zowel een lemma in GiGaNT als een koppelbetekenis in het betekenisregister toegewezen krijgen zodat de verdere beschrijving van het woord in de verschillende kennisbankmodules hieraan gelinkt kan worden. Belangrijk is dat die beschrijving nu modulair kan gebeuren en er niet meer eerst per woord een volledig woordenboekartikel afgewerkt moet worden. Nu al worden meerwoordige combinaties voor het hedendaags Nederlands systematischer behandeld dankzij een aparte workflow voor woordcombinaties waarvan de resultaten dan nadien terug aan de beschrijving van de individuele woorden gekoppeld worden. Dat zal ook mogelijk zijn voor andere aspecten zoals betekenisrelaties of lexicale kenmerken die zich systematischer op een hoger structureel niveau van de woordenschat laten beschrijven eerder dan telkens apart per woordenboekartikel. De modulaire opbouw en koppelingen tussen modules zal het ook makkelijker maken om de consistentie zowel per module als door de hele kennisbank heen te controleren en te bewaren. De huidige werkomgevingen, die primair bedoeld zijn om woordenboekartikel per woordenboekartikel af te werken, zullen dan ook geleidelijk vervangen worden door een verdere uitbouw van de Lex'it-omgeving, die door het INT specifiek ontwikkeld werd om lexicografische relationele databanken aan te maken en te beheren. Op die manier zal de workflow voor de historische en hedendaagse woordenschat op termijn geharmoniseerd worden zodat de unieke sterkte van de lexicografie aan het INT, namelijk haar diachrone diepgang, maximaal tot zijn recht kan komen.

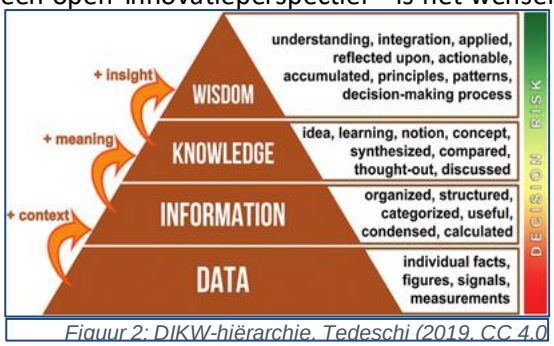
Van data over informatie naar kennis

De geleidelijke integratie van de woordenschatbeschrijving in één centrale modulaire kennisbank gaat hand in hand met een tweede strategische doelstelling, namelijk de versterking van het data-gedreven karakter van de woordenschatbeschrijving. Het INT heeft al een lange traditie en internationaal erkende expertise in de *evidentiegebaseerde* creatie van lexicografische kennis. Die expertise wil het INT verder uitbouwen, toepassen binnen onze lexicografische workflows en ter beschikking stellen aan alle gebruikers van onze taalinfrastructuur.

Vanuit kennisbeheer-oogpunt volgt de evidentiegebaseerde lexicografie het model van een data-informatie-kennishiërarchie:[41] Empirische taaldata worden kwalitatief en kwantitatief geanalyseerd en de resulterende informatie over woorden wordt verder bewerkt tot gestructureerde kennis over de woordenschat. Die gestructureerde kennis vormt op zijn beurt een kennisinfrastructuur waarvan andere kennisprocessen op intelligente wijze gebruik kunnen maken. Het instituut beschikt over representatieve, zorgvuldig gemetadateerde en verrijkte corpora die dit kenniscreatieproces maximaal ondersteunen. Die data worden met gevalideerde, corpuslinguïstische methoden geanalyseerd en de resulterende kennisrepresentatie is onderbouwd door wetenschappelijk-lexicografische principes. In de komende jaren zal de relatie tussen data, informatie en kennis binnen de lexicografische workflow nog verder versterkt worden. Op dit moment is die workflow vooral op het afleveren van hoogwaardige eindproducten gericht maar is het kenniscreatieproces zelf minder goed gedocumenteerd. Zo komt bijvoorbeeld van de vele concordanties die geanalyseerd en beoordeeld worden om tot een woordbeschrijving te komen uiteindelijk slechts een handjevol in het woordenboek terecht zonder dat de tussenresultaten van het voorafgaande kenniscreatieproces systematisch worden opgeslagen. De meerwaarde van een kennisinfrastructuur wordt echter groter naarmate de structurele en functionele relaties tussen onderliggende data, data-analyse en resulterende kennis

[41] https://en.wikipedia.org/wiki/DIKW_pyramid

sterker en explicieter worden en alle tussenstappen goed gedocumenteerd en als relationele (open) data beschikbaar gemaakt worden. (1) Vanuit wetenschappelijk oogpunt wordt op die manier de kwaliteit van kennis vergroot omdat het proces van kennisopbouw meer transparant, reconstrueerbaar en ook repliceerbaar wordt. (2) In een open-innovatieperspectief[42] is het wenselijk dat (een selectie van) tussenresultaten uit het kenniscreatieproces ook publiek beschikbaar gesteld worden en door anderen voor onderzoek en ontwikkeling gebruikt kunnen worden. Met onze corpora doen we dat nu al en dat zou bijvoorbeeld ook kunnen voor alle concordanties die tijdens het lexicografische proces statistisch geanalyseerd, geannoteerd en gedesambigueerd worden. (3) Een systematische koppeling tussen kennis, informatie en data laat ook toe om de lexicografische kennisbank als een gestructureerde toegang tot de onderliggende informatie en data te gebruiken. Zo zouden uit zo'n gekoppelde kennisbank makkelijk alle woorden die ook beroepsaanduidingen zijn en minstens één synoniem hebben samen met alle relevante corpusvoorbeelden geëxtraheerd kunnen worden. Als dan meteen ook relatieve frequenties van alle synoniemparen in zowel het Belgisch als Nederlands Nederlands eraan gelinkt zijn, dan is zo'n dataset bijzonder nuttig voor een bedrijf dat automatisch vacatures verwerkt, zijn software met machinelearning wil lokaliseren voor de Belgische en Nederlandse markt en daarbij met de verschillen in beroepsaanduidingen moet rekening houden. (4) Ten slotte, wat veruit het belangrijkst is voor het INT zelf, geldt dat het lexicografische kennisextractieproces alleen maar computationeel gemodelleerd kan worden als zoveel mogelijk tussenresultaten en analyses opgeslagen worden en aan elkaar gelinkt kunnen worden. Dit is essentieel om nieuwe NLP- en AI-technieken in het lexicografisch proces te integreren en verder te optimaliseren. Voor het trainen van bijvoorbeeld een Word Sense Induction-module die de lexicograaf ondersteunt bij de betekenisanalyse, is het handjevol citaten dat nu het woordenboek haalt niet genoeg, maar is de hele geanalyseerde set van concordanties en collocaties nodig. De ontwikkeling van die nieuwe modules voor automatische analyse- en kennisextractie kan deels intern gebeuren, maar zal toch vooral tot stand komen via samenwerking binnen de bredere NLP- en AI-community. Daarvoor is het wenselijk dat het INT datasets met trainings- en testmateriaal (gold standards) kan aanleveren en eventueel via zogenaamde "shared tasks" mee de O&O-agenda kan aansturen.



Figuur 2: DIKW-hiërarchie, Tedeschi (2019, CC 4.0

De vernieuwing van de lexicografische workflow om via een doorgedreven databeheer en verdere automatisering de opbouw van een modulaire kennisbank te ondersteunen heeft implicaties voor de interne infrastructuur van het instituut, de corpusuitbouw en de analyse- en bewerkingsomgevingen. Als trends in de woordenschat of gebruiksbijzonderheden primair uit de corpora afgeleid worden, dan moeten die corpora een tijdsdimensie bevatten, representatief zijn voor verschillende taalvariëteiten en vooral goed gemetadateerd zijn. Als we meerwoordige uitdrukkingen en constructies willen extraheren, dan moeten de corpora ook syntactisch geparseerd worden. Als we de manuele en statistische corpusanalyses willen documenteren om er nieuwe NLP- en AI-modules voor te ontwikkelen, en die nadien ook in de workflow te pluggen, dan moet onze eigen corpuszoekomgeving (BlackLab) verder uitgebouwd worden om corpusanalyses en datalogging toe te laten die aan de kennisbank gelinkt kunnen worden. Ook in de andere richting moet de informatie uit de kennisbank op termijn als bijkomende annotatie in de corpora terecht komen: pas als de woorden in het corpus geannoteerd zijn met hun GiGaNT-lemma of koppelbetekenis uit het betekenisregister, kan de meerwaarde van corpora en kennisbank als relationele open data volledig gerealiseerd worden. Dit is

[42] https://nl.wikipedia.org/wiki/Open_innovatie

een ontwikkeltraject met een ruimere tijdshorizon dan dit meerjarenbeleidsplan, maar waarin we de komende vijf jaar al wel belangrijke stappen willen zetten.

Toekomstperspectief: Een digitaal knooppunt voor de Nederlandse woordenschat

De geïntegreerde kennisbank zal in de mate van het mogelijke ook extern beschikbaar komen als *relationele open data* voor Onderzoek en Ontwikkeling. Omdat de kennisbank unieke en persistente identifiers bevat voor zowel lemmata als koppelbetekenissen wordt meteen ook de mogelijkheid gecreëerd om bijkomende externe lexicale resources en datasets eenduidig te linken aan de kennisbank. Het INT zal dit in eerste instantie proefgewijs uitproberen voor de lexicale taalmaterialen die nu al door het instituut beheerd worden (RBN, RBBN, e-lex, Verschueren Groot Encyclopedisch Woordenboek) zodat deze via de repository niet alleen apart downloadbaar zijn, maar binnen de kennisbank gecombineerd doorzoekbaar worden met andere lexicale resources en bovendien ook makkelijker de eigen woordenschatbeschrijving van het INT kunnen ondersteunen. Voor andere externe lexicale resources (bv. Open Dutch WordNet, Dutch FrameNet van de VU, Small World of Words van KU Leuven, de Lexicon Projects van de Universiteit Gent of de Wiktionary voor het Nederlands) zullen we, waar mogelijk, samenwerking zoeken met de ontwikkelaars van deze resources om ze te koppelen zodat de kennisbank kan uitgroeien tot hét *digitale knooppunt* voor alle woordenschatgerelateerde hulpbronnen van het Nederlands (zie figuur 1). Bovendien blijven we binnen het bestaande Europese netwerk van lexicografische instituten verder participeren in de koppeling van lexicografische data over talen heen, bijvoorbeeld via de Matrix-dictionary uit het ELEXIS-project. Zo zal het INT zijn rol blijven waarmaken als vitale schakel in de taalinfrastructuur voor het Nederlands en op Europees niveau.