/instituut voor de
Nederlandse taal/

jaarverslag

2023

verslag aangaande de werkzaamheden, gebeurtenissen enz. in het afgelopen jaar; jaarlijks verslag

(Woordenboek der Nederlandsche Taal)

Jaarverslag 2023

Instituut voor de Nederlandse Taal

Inhoudsopgave

1. Inleiding: het jaar 2023

Het eerste meerjarenbeleidsplan (2018-2022) werd succesvol uitgevoerd; aansluitend werd er een nieuw meerjarenplan geschreven dat in 2023 van start ging en dit in nauw overleg met de Nederlandse Taalunie (NTU).

De internationale visitatie van 2021 was een goede toets om te zien hoe het hervormde instituut de waaier van taken had ingevuld. Het eindrapport beoordeelde het INT als excellent en beschreef ons als hét kennisinstituut voor de Nederlandse taal. Ook het Comité van Ministers nam kennis van het rapport en sprak lof uit voor onze activiteiten.

De eerder gemelde zorgen van de visitatiecommissie blijven bestaan: zij stelden dat de prestaties van het INT indrukwekkend zijn, maar zij maken zich zorgen of dit duurzaam kan zijn nu de financiering achterblijft. De structurele financiering en het weerstandsvermogen zijn te laag waardoor een verdere ontwikkeling van het instituut problematisch kan worden. Tegenslagen in projecten kunnen nauwelijks worden opgevangen en de belasting van het personeel kent haar grenzen ondanks dat de commissie vol lof is over de betrokkenheid, bevlogenheid en deskundigheid van het personeel. De aanbeveling om de taken van het personeel goed te bewaken wordt ter harte genomen en de planning van de projecten nauwkeurig gevolgd.

Het lexicografische werk dat op deze infrastructuur voortbouwt, heeft zich met succes gediversifieerd in kleinere projecten die op verschillende manieren aan elkaar gekoppeld zijn. Door deze nieuwe benadering en modulaire aanpak kunnen hedendaagse, historische, dialectologische, morfologische, syntactische en semantische informatie enz. voor een bepaald woord bijeengebracht worden in verschillende maar nauw samenwerkende projecten en kan deze informatie afhankelijk van de doelgroep op diverse manieren online aangeboden worden. Op deze manier kan het INT zijn opdracht vervullen en moderne wetenschappelijk onderbouwde lexicografische producten blijven aanbieden.

In 2023 werden de statuten van het INT aangepast; hierin werd meer aandacht besteed aan wetenschappelijke activiteiten die onze rol als kennisinstituut benadrukken. Tevens werd de samenwerkingsovereenkomst met de Taalunie geactualiseerd. Ook is er toegang gecreëerd tot de projectmiddelen van NWO. Bij FWO Vlaanderen kunnen we als buitenlandse partij ook meedingen voor bepaalde opdrachten (meer bepaald bij Strategisch Basisonderzoek).

Het INT heeft als structureel gefinancierd kennisinstituut een unieke positie en opdracht om voor het hele Nederlandse taalgebied (Nederland en de Caribische rijksdelen, Vlaanderen en Suriname) op een wetenschappelijk verantwoorde wijze de digitale taalinfrastructuur uit te bouwen. Het INT voert daarbij een aantal taken uit het Taalunieverdrag uit. We verwijzen hier naar hoofdstuk 1 uit het Taalunieverdrag, artikelen 2, 3, 4 en 5. Het INT ontwikkelt enerzijds zelf corpusdata, linguïstische databanken en taalsoftware voor een aantal specifieke domeinen of ondersteunt de ontwikkeling ervan; anderzijds verzamelt het INT ook taalmaterialen en taalsoftware van andere kennisinstellingen en stelt het deze samen met de eigen taalmaterialen duurzaam ter beschikking via repository’s, websites, API’s en als opensource software.

Het INT promoot die taalinfrastructuur bij onderzoekers, ontwikkelaars en het brede publiek om zo onderzoek en andere activiteiten rond de Nederlandse taal te stimuleren en te ondersteunen. Daarnaast heeft het INT als toegepast onderzoeksinstituut ook de doelstelling de kennis en expertise over taalinfrastructuur verder uit te bouwen door eigen wetenschappelijk onderzoek.

Wat betreft het Taalunieverdrag draagt het INT bij aan de ontwikkeling en bevordering van de kennis van het Nederlands, en ook de studie en verspreiding ervan. We ontwikkelen en updaten de spellingapplicatie Woordenlijst.org en zijn verantwoordelijk voor de Algemene Nederlandse Spraakkunst. Door onze rol als CLARIN-instituut voor Nederland én Vlaanderen moedigen we het wetenschappelijk onderzoek aan. Verder wordt door het Expertisecentrum Nederlandstalige Terminologie gewerkt aan de promotie van het Nederlands als wetenschapstaal en de verspreiding ervan in databanken en woordenlijsten. Daarnaast participeert het INT in projecten op nationaal en internationaal niveau en houdt het zo de vinger aan de pols wat betreft de ontwikkelingen in de lexicografie en de computationele taalkunde.

Bij de aanvang van 2023 werkten er 33 werknemers (27,25 fte). Eind december 2023 waren dat 34 personeelsleden (28,15 fte). Er waren enkele verschuivingen door pensionering van één werknemer en er kwamen twee nieuwe medewerkers: een managementassistente (40% fte) en een systeemontwikkelaar (100% fte).

In de raad van toezicht beëindigde zowel Jan Cerfontaine zijn mandaat op 1 mei 2023 als Gertine van der Vliet op 1 december 2023. Zij werden vervangen door respectievelijk Frank Judo en Erik Boels.

Het INT is hét kennisinstituut voor het Nederlands en neemt een centrale positie in voor het hele Nederlandse taalgebied op het vlak van het wetenschappelijk verantwoord ontwikkelen, bewaren en duurzaam beschikbaar stellen van taalmateriaal. Het INT streeft ernaar om het best gesorteerde wetenschappelijk instituut te zijn op het gebied van de Nederlandse taal en de woordenschat, gecombineerd met hoge kwaliteit en goede toegankelijkheid. Het instituut ontwikkelt en levert data voor woordenboeken, (computationele) lexica, corpora en tools. De woordenboeken zijn online te raadplegen. Software en computerlinguïstische tools zijn opensource beschikbaar. Het instituut speelt in op de nieuwe ontwikkelingen in de geesteswetenschappen, met name op het terrein van de digital humanities. Om deze rol te kunnen vervullen beheert en onderhoudt het INT een digitale infrastructuur voor het Nederlands, met aandacht voor taalvariatie (terminologie, dialecten etc.). Zowel academische als niet-academische partijen kunnen gebruik maken van deze infrastructuur.

Er stonden in 2023 een heleboel activiteiten op het programma. Er werd vooral ingezet op het verbeteren van de digitale infrastructuur van het Nederlands, met de koppeling van het hedendaagse met het historische lexicon. Verder werd er hard gewerkt aan een nieuwe API voor de spelling, wat resulteerde in een nieuwe applicatie voor Woordenlijst.org. In samenwerking met de Taalunie werd uitvoerig aandacht besteed aan een aantal taalmaterialen, met name de grammatica (e-ANS), het Corpus Hedendaags Nederlands, Woordcombinaties en de historische woordenboeken. Dit gebeurde via webinars en kennisclips.

Het thema van het jaar was “meertaligheid” wat resulteerde in een aantal activiteiten en een boek van de hand van Nicoline van der Sijs *Daar is geen woord Frans bij*.

Op Europees vlak werd het ELRC-consortium overgeheveld naar de European Language Data Space, waar wij de verantwoordelijkheid voor Nederland nemen, en waren we verantwoordelijk voor het

ELE-project (European Language Equality). We werden ook betrokken in twee nieuwe COST-netwerken.

Internationaal werden de banden met het Caribisch gebied aangehaald in de vorm van een samenwerking met het pas opgerichte Nationale Taalinstituut in Curaçao. We bereidden een intensieve samenwerking voor om te helpen het Papiamento/u te beschrijven en te documenteren zoals we dat ook voor het Nederlands doen. Het officiële samenwerkingsakkoord wordt in 2024 ondertekend.

De lijn van webinars en online presentaties die we sinds de pandemie hebben uitgewerkt werd nog versterkt door een nieuwe reeks podcasts in samenwerking met Onze Taal, waarin we experts aan het woord laten over één thema binnen de taalkunde, zoals dialectologie, neologismen, terminologie, grammatica, etc. Deze podcasts worden gretig beluisterd door een breed publiek, ook door studenten en docenten Nederlands in het buitenland.

Februari 2024
prof. dr. Frieda Steurs
wetenschappelijk directeur/bestuurder

2. Language Resources Repository en CLARIN-centrum

Language Resources Repository

Via de website taalmaterialen.ivdnt.org stelt het Instituut voor de Nederlandse Taal (INT) bronnen, data en tools beschikbaar voor taalkundig onderzoek en taal- en spraaktechnologie (TST) binnen het hele Nederlandse taalgebied.

Een aantal van die tools en resources zijn ook beschikbaar vanuit het INT CLARIN portal, op https://portal.clarin.inl.nl/. Deze site kreeg 3900 site visits van 3000 bezoekers in 2023.

In oktober 2023 is voor de komende drie jaar opnieuw het Core Trust Seal (CTS) toegekend aan het Instituut voor de Nederlandse Taal (INT). Deze heeft betrekking op de kwaliteit van beheer en dienstverlening van het archief. Dit certificaat is van belang voor het verlengen van de status van het INT als CLARIN B-Centre.

Het totaal aantal downloads voor zowel commerciële als niet-commerciële producten is gestegen in vergelijking met het voorgaande jaar:

- Niet-commerciële downloads 816 (522 in 2022)
- Commerciële downloads 20 (5 in 2022)

Nieuwe producten

In het afgelopen jaar hebben we de volgende producten toegevoegd aan de catalogus:

- De META-COVID Ontology verbindt 30 interdisciplinaire COVID-onderwerpen met 203 specifieke concepten vanuit wetenschappelijke ontologieën. Deze ontologie is ontwikkeld binnen het EOSC Futureproject als onderdeel van het wetenschappelijk proefproject "COVID-19 metadata findability and interoperability in EOSC (META-COVID)".
- SoNaR Character n-grams. Uit het SoNaR-corpus versie 1.2 (SONAR500) zijn n-grammen van lettertekenreeksen met lengtes 1, 2 en 3 afgeleid. Van de originele bestanden werden tekstbestanden gemaakt in UTF-8. Op basis van die bestanden werden met een Perlscript – dat meegeleverd wordt – de n-grammen berekend, die vervolgens werden weggeschreven naar een tab-gescheiden bestand. Hoofdletters werden omgezet in kleine letters en werden dus niet apart geteld.
- Hotel Review Corpus in Nederlandse Gebarentaal (NGT_HoReCo) - inmiddels vervangen door een update. Een multimodaal parallel corpus met de talen Nederlands en Nederlandse Gebarentaal (NGT). 283 geschreven hotelbeoordelingen werden vertaald uit het Nederlands in NGT door 6 professionele, dove vertalers. Elke beoordeling is vertaald door slechts 1 vertaler. Het aantal woorden in de beoordelingen varieerde tussen 15 en 400. De duur van de NGT-video's varieerde tussen 10 seconden tot ongeveer 4 minuten. Het resulterende corpus bevat 21.825 woorden in het Nederlands en ruim 3,5 uur aan NGT-videomateriaal.

- The LiLaH Emotion Lexicon of Greek, Kurdish, Turkish, Spanish, Farsi and Chinese. Dit product bevat een uitbreiding van het NRC-emotielexicon. Het bevat een lijst met woorden in het Grieks, Koerdisch, Turks, Spaans, Farsi en Chinees (traditioneel en vereenvoudigd) en hun associaties met acht basisemoties (boosheid, angst, verwachting, vertrouwen, verrassing, verdriet, vreugde en afkeer) en twee sentimenten (negatief en positief). De annotaties zijn automatisch gegenereerd en handmatig gecontroleerd door moedertaalsprekers. Aanvullende talen (waaronder Nederlands) kunnen hier gevonden worden.
- Woordenboek Vlaamse Gebarentaal (Woordenboek VGT). Dit product bevat het videomateriaal uit het online Woordenboek Vlaamse Gebarentaal. In de 10.025 video's is per video een gebaar vastgelegd. Het online woordenboek is te vinden op woordenboek.vlaamsegebarentaal.be.

Updates van bestaande producten

- Belgian Covid Sign Language Corpus (BeCoS Corpus). Het Belgische Federale COVID-19-corpus, genaamd het BeCoS-corpus (Belgian Covid Sign language corpus), bestaat uit het volledige archief van officiële persconferenties van de Belgische federale overheid betreffende de COVID-19-pandemie. De sprekers spreken meestal Nederlands of Frans en een enkele keer Duits, en bijna alle spraak wordt getolkt door een dove gebarentaaltolk die live tolkt wat er wordt gezegd. De data is beschikbaar als ELAN-bestanden en is voorverwerkt met automatische detectie van sprekerverandering, Belgisch-Nederlandse spraakherkenning, taalidentificatie, interpunctievoorspelling en detectie gebarentalige verandering. In de video's is keypointherkenning toegepast op de gebarentaaltolken. Versie 1.1 bevat aanvullende *tiers* in de ELAN-bestanden betreffende taalidentificatie en spraakherkenning
- Lassy Groot-corpus. Het Lassy Groot-corpus is een corpus van ongeveer 700 miljoen woorden met automatisch gegenereerde syntactische annotaties. De lemma's en POS-tags werden automatisch toegevoegd aan het corpus m.b.v. Tadpole (nu: Frog). De syntactische dependentiestructuren werden toegevoegd m.b.v. Alpino.
- Lassy Klein-corpus. Het Lassy Klein-corpus is een corpus van ongeveer 1 miljoen woorden met manueel geverifieerde syntactische annotaties. Lemma’s en POS-tags werden automatisch toegevoegd aan het corpus m.b.v. Tadpole (nu: Frog). De syntactische dependentiestructuren werden toegevoegd m.b.v. Alpino. De lemma’s, POS-tags en syntactische boomstructuren werden geverifieerd en gecorrigeerd. Het corpus is beschikbaar in zowel XML- als in Dact-formaat en de zoeksoftware Dact wordt meegeleverd in het downloadbestand. De download bevat daarnaast ook frequentielijsten.

Geografische verspreiding van gebruikers

Gebruikers dienen zich te registreren alvorens ze producten kunnen downloaden. Daarbij wordt ook hun e-mailadres vastgelegd. Aan de hand van de extensie van die adressen kan een beeld worden verkregen vanuit welke landen belangstelling bestaat voor onze taalmaterialen. De meesten komen uit Nederland (131), daarna België (71) en verder Duitsland (12) en de Verenigde Staten (.edu) (10), en lager.

De e-mailadressen geven niet een volledig beeld van waar de gebruikers zitten, omdat veel een ambigue extensie hebben als .com of .org. Daarom gebruiken we ook de IP-adressen die gebruikt worden voor het downloaden. Door een veranderde configuratie van de Taalmaterialen-server is het helaas niet meer mogelijk om rechtstreeks die IP-adressen in te zien. Het is nog wel mogelijk om de statistieken van de downloadpagina op te vragen. Dat leverde het volgende overzicht op: Nederland (52%), België (27%), Verenigde Staten (3%), Duitsland (3%), Finland (2%) en een groot aantal lagere scores.

Een volledig overzicht van downloads per product is in bijlage 4 te vinden.

Het INT als CLARIN-centrum

Voor 2022 vervulde Vincent Vandeghinste voor het INT de rol van Nationaal Coördinator voor CLARIN op Europees niveau, als vertegenwoordiger voor België in het National Coordinators Forum (NCF). Vincent trad ook toe tot het Knowledge Infrastructure Committee van CLARIN-ERIC. Daarnaast nam Jesse de Does vanuit het INT de vertegenwoordiging van België op zich in het Standing Committee on CLARIN Technical Centres. Het INT vervulde een actieve rol als liaison tussen Belgische onderzoekers en de Europese CLARIN-infrastructuur en tussen de Vlaamse onderzoekers in CLARIAH-VL en de CLARIN-infrastructuur van het INT als CLARIN-B centrum voor België. Het INT fungeerde ook als lokale organisator voor de jaarlijkse CLARIN-conferentie die in 2023 in Leuven plaatsvond, en Vincent werd door het NCF verkozen tot Programme Committee Chair voor de volgende CLARIN-conferentie. Samen met de UGent werd MATEO gereleased, een app voor online evaluatie van Machine Translation output. Hiervoor werd een webinar georganiseerd. K-Dutch, het CLARIN Knowledge Centre for Dutch, werd verder uitgebouwd, waarbij de website verder werd uitgebreid en verfijnd.

3. Corpusinfrastructuur

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

Monitorcorpus

Het Corpus Hedendaags Nederlands (CHN) bevat al het hedendaags Nederlandse corpusmateriaal waarover het INT beschikt. Het corpus wordt continu uitgebreid met materiaal uit Nederland, Vlaanderen, Suriname en de Caribische rijksdelen. Er is een interne versie van het CHN, die wekelijks geüpdatet wordt, en een externe versie die elke maand een update krijgt.

Afgezien van de reguliere updates voor het CHN is er met name gewerkt aan realiseren van een min of meer evenwichtig monitorcorpus van kranten voor dit millennium. De aandacht ging daarbij in 2023 uit naar het Vlaams krantenmateriaal. Dat is zeer substantieel uitgebreid. Voor de kranten *De Gazet van Antwerpen, Het Laatste Nieuws* en *Het Belang van Limburg* is er nu ook materiaal dat teruggaat van 2020 tot begin van dit millennium. Voor de krant *De Standaard* is de lacune voor het jaar 2012 opgelost.

Daarnaast is er verder gewerkt aan het uitbreiden van het CHN met nieuwe bronnen. Een van de nieuwe bronnen is de *Parbode* uit Suriname. Daarvoor is het materiaal verwerkt van 2016-2022. Al het hiervoor genoemde materiaal is beschikbaar in het externe CHN. Het interne corpus bevat sedert 2023 ook materiaal van *OpenSubtitles.* Op termijn zal dit materiaal ook aan het externe CHN worden toegevoegd.

Daar waar het interne corpus eind 2022 ruim 2,4 miljard tokens en ruim 5,3 miljoen documenten bevatte, bevat het op 2 februari 2024 ruim 4 miljard tokens en ruim 9,9 miljoen documenten. Het externe CHN bevatte ruim 1,15 miljard woorden en ruim 3 miljoen documenten eind 2022. Dat is eind 2023 9,2 miljoen documenten geworden en ruim 2,8 miljard tokens.

In samenwerking met de Taalunie is er aandacht besteed aan het CHN door middel van een webinar over het corpus en de corpusapplicatie, en een kennisclip.

Historische corpora

Ook voor wat betreft het historisch Nederlands is veel aandacht geschonken aan krantenmateriaal. Nadat in 2022 het Couranten Corpus met 17e-eeuwse kranten online was gegaan, is een proces in gang gezet om de vervolgset te digitaliseren (zie CLARIAH Plus). In 2023 hebben vrijwilligers van de Stichting Vrijwilligersnetwerk Nederlandse Taal met behulp van het programma Transkribus daarbij

grote vorderingen gemaakt. Maar liefst 40% van de meer dan 5200 pagina's van de Amsterdamse Courant is dit jaar gecontroleerd en van metadata voorzien door een groep van circa 25 enthousiaste vrijwilligers. Toen 50% van de Amsterdamse Courant door hen was voltooid, is het online gelegenheidswoordenboekje *Oud nieuws* gepubliceerd.

Nadat de metadata ervan waren gecureerd, er nieuwe metadata aan waren toegevoegd en er een nieuwe website en zoekmachine voor de data was ontwikkeld, is het corpus Gekaapte Brieven online gepubliceerd. Voor deze corpusinstantie is voor het eerst gebruik gemaakt van een IIIF-server voor de afbeeldingen en een daarvoor geschikte image viewer.

Met een presentatie op de HSN Conferentie Onderwijs Nederlands in Rotterdam is aandacht geschonken aan de mogelijkheden die het Couranten Corpus biedt om ingezet te worden voor onderwijsdoeleinden.

Verrijking

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

Verrijking met woordsoort en lemma

Voor de verrijking van historisch Nederlands is (mede in het CLARIAH Plus-project, zie aldaar) gewerkt aan de ontwikkeling van gold standard datasets voor de verrijking met woordsoort en lemma volgens de TDN-richtlijnen. Inmiddels zijn ruim 250.000 tokens, gespreid over de 15e tot 19e eeuw, volledig nagekeken. Op basis van dit materiaal zijn deep learning-gebaseerde taggers en lemmatisers getraind met behoorlijk resultaat. Het materiaal zal begin 2024 worden aangevuld door harmonisatie van reeds eerder verrijkte bestanden zoals *Corpus Gysseling*, *Corpus van Reenen-Mulder* en *Brieven als Buit*.

Syntactische annotatie

Er loopt onderzoek naar de vraag of de huidige set UD-dependentierelaties voor het Nederlands voldoende aanknopingspunten biedt voor de toepassing binnen met name het project Woordcombinaties, of dat wellicht een aantal specifieke extensies (bijvoorbeeld voor het onderscheiden van maatcomplementen, locaal-directionele complementen en voorzetselvoorwerpen) nodig zijn. Op een aantal punten zijn extensies voorgesteld; de haalbaarheid van de automatische implementatie wordt in 2024 verder verkend.

Metadata

De corpora hebben een gemeenschappelijk metadataformaat, met ruimte voor subcorpus-specifieke metadata. Er is gewerkt aan een verdere uniformering van het metadatamodel voor historisch en modern corpusmateriaal zodat beide op termijn als één doorlopend diachroon corpus op metadatacategorieën doorzoekbaar kunnen worden.

Informatie-extractie

Het corpusmateriaal wordt toegankelijk gemaakt via een applicatie waarmee in de corpora gezocht kan worden. Wanneer de IPR (Intellectual Property Rights) het toelaten, wordt het corpusmateriaal ook als dataset beschikbaar gesteld in de language resource repository. De software is opensource beschikbaar.

Voor de ontwikkeling van de corpusapplicatie die bestaat uit de corpuszoekmachine BlackLab en de corpus frontend, ligt de prioritering bij de ondersteuning van de diverse INT-taken. Deze werkzaamheden worden ondersteund door het samenwerken met diverse partijen die de software gebruiken, en door de mogelijkheden die externe projecten bieden om deze software verder te ontwikkelen. Voor wat betreft de backend van de corpusretrievalomgeving (BlackLab, BlackLab Server) is gewerkt aan:

- 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. Er is een uitbreiding van de CQL-zoektaal gedefinieerd en geïmplementeerd die het zoeken op hiërarchische relatiepatronen mogelijk maakt.
- Uitbreiding van de functionaliteit om efficiënt statistieken uit het materiaal te extraheren, in eerste instantie ten behoeve van diachrone frequentieprofielen.

Voor wat betreft het user interface is gewerkt aan een nieuwe versie van de functie waarmee zoekresultaten binnen de corpusapplicatie gegroepeerd kunnen worden. Deze moet meer mogelijkheden voor analyse geven, met name voor syntactische fenomenen. Vanwege een gewijzigde prioritering ten behoeve van het ontwikkelen van een single sign on voor het INT, is het werk aan een query-builder voor syntactische retrieval uitgesteld.

4. Beschrijving van de woordenschat door de eeuwen heen

Datamodel voor de centrale kennisbank van de woordenschat

Zoals uiteengezet in het Meerjarenbeleidsplan 2023-2027 is de koppeling van verschillende lexicografische databanken op zowel vorm- als betekenisniveau de belangrijkste innovatie en de grootste uitdaging voor de woordenschatbeschrijving in huidige beleidsperiode. Voor het succesvol uitwerken van een complexe data-infrastructuur voor de centrale kennisbank is in 2023 een begin gemaakt met de uitbreiding van het datamodel en het opzetten van een betekenisregister.

Voor de uitbreiding van het datamodel werd voortgebouwd op de al bestaande onderdelen in de huidige infrastructuur, met name het centrale lexicon GiGaNT (cf. infra), de koppeling op lemmaniveau met de lexicografische databanken en het Diachroon seMantisch lexicon van de Nederlandse Taal (DiaMaNT). Er werd in kaart gebracht welke bijkomende datacategorieën nodig zijn, welke koppelingen binnen de kennisbank wenselijk zijn en welke daarbuiten (bv. met corpusdata). Bijzondere aandacht ging in 2023 daarenboven naar de classificatie en modellering van meerwoordsexpressies en naar compatibiliteit van zowel de hedendaagse als historische lexicale beschrijving met de Tagset Diachroon Nederlands (TDN).

In wat volgt beschrijven we de activiteiten die in 2023 voor de verschillende onderdelen van de centrale kennisbank en de eraan gekoppelde lexicografische producten uitgevoerd werden.

Het Centrale Lexicon (GiGaNT)

In 2023 zijn in eerste instantie de reguliere werkzaamheden aan het centrale lexicon voortgezet. Deze houden in: het onderhoud aan zowel de historische component (GiGaNT-Hilex) als de moderne component (GiGaNT-Molex), het verder werken aan de integratie van deze componenten en de uitbreiding van GiGaNT-Molex ten behoeve van de diverse lexicografische producten die een koppeling met GiGaNT hebben, zoals Woordenlijst.org (cf. infra). Bijzondere aandacht is besteed aan de classificatie en modellering van meerwoordsexpressies. Daarbij is uitgegaan van de moderne lexiconcomponent. Het datamodel van GiGaNT is aangepast en de database en de data zijn overeenkomstig aangepast. Daarnaast zijn voor zowel de histoirsche als de moderne lexiconcomponent de woordsoorttoekenningen waar nog nodig conform de Tagset Diachroon Nederlands (TDN) toegekend.

Daarnaast is er een eerste testworkflow opgezet om de corpusgebaseerde bewaking van neologismen te koppelen aan GiGaNT-Molex. Wekelijks worden de nieuwe woordvormen in het CHN geïdentificeerd en vergeleken met de woordvormen die al in Molex aanwezig zijn. Kandidaat-neologismen worden in een aparte databank bijgehouden en opgevolgd totdat ze aan criteria voor opname in olex en eventuele verdere lexicografische behandeling voldoen (bv. toename in frequentie of verdere verspreiding over het taalgebied). Deze testworkflow zal in 2024 in de reguliere Molexworkflow geïntegreerd worden.

Betekenisregister

In parallel met het opstellen van het datamodel voor de centrale kennisbank werd ook een eerste proefversie van het betekenisregister gemaakt. Die proefversie is opgebouwd vanuit het hedendaags Nederlands en gaat daarbij uit van de definities uit lexicografische bronnen die op vormniveau reeds aan het lexicon GiGaNT-Molex gekoppeld zijn (ANW, WNW, Referentiebestand Nederlands (RBN), Woordcombinaties en Vertaalwoordenschat). Het samenbrengen van alle betekenisbeschrijvingen uit deze bronnen, zoweldefinities als glossen), heeft enerzijds toegelaten om een aantal inconsistenties tussen de verschillende bronnen op te sporen en anderzijds de verdere uitwerking van het datamodel te informeren. Met een centrale koppeling van lexicografische bronnen op betekeninsniveau zet het INT een belangrijke stap verder in het uitvoeren van de opdracht in het Taalunieverdrag om de woordenschat van het Nederlands systematisch en gestructureerd te beschrijven.

Lexicografische eindproducten, API's en datasets

Woordenlijst.org

Voor de website Woordenlijst.org is een vernieuwde applicatie ontwikkeld, met een verbeterde suggestiemodule en uitgebreide zoekmogelijkheden. De presentatie van de zoekresultaten is verbeterd, zodat het nu mogelijk is alle zoekresultaten door te lopen en te filteren op woordsoort. Ook de weergave van het eindresultaat is aangepast. De weergave van de trefwoorden is inzichtelijker gemaakt, en er is bijzondere aandacht besteed aan de paradigmata bij voornaamwoorden.

De informatie bij trefwoorden is met de volgende zaken uitgebreid:

- Er is een begin gemaakt met het toevoegen van uitspraakinformatie, met als focus homografen en woorden met uitspraakproblemen.
- Aan het werkwoordparadigma zijn de gebiedende en aanvoegende wijs toegevoegd.
- Werkwoorden hebben features gekregen (transitiviteit, reflexiviteit, onpersoonlijkheid) en het bijbehorende hulpwerkwoord.

Het aantal trefwoorden is uitgebreid met homoniemen, vormvarianten en verkleinwoorden van trefwoorden die al online stonden. Op 24 januari 2024 is de nieuwe woordenlijst online gegaan, met 223.043 trefwoorden (eerder 220.067) en 830.385 woordvormen (eerder 777.042).

Tot slot heeft het INT ook dit jaar weer ondersteuning geboden aan de Commissie Spelling bij haar werkzaamheden.

Algemeen Nederlands Woordenboek (ANW)

Het *Algemeen Nederlands Woordenboek* (ANW) is een corpusgebaseerd, digitaal, multimediaal woordenboek van het eigentijdse Nederlands in Nederland en de Caribische rijksdelen, Vlaanderen en Suriname. De taalperiode die het ANW bestrijkt, loopt van 1970 tot heden. Naast de woorden uit de kernwoordenschat van de genoemde taalgebieden wordt in het ANW ook een selectie van door het INT verzamelde neologismen (nieuwe woorden, nieuwe verbindingen, nieuwe uitdrukkingen, nieuwe betekenissen van al bestaande woorden) opgenomen. Het beschrijvingsproces binnen het

ANW is dan ook sterk geïntegreerd met dat binnen het Woordenboek van Nieuwe Woorden (WNW, zie onder): de neologismen die ingeburgerd geraken, worden na behandeling in het WNW ook opgenomen in het ANW. Beide woordenboeken gebruiken dezelfde bewerkingsomgeving en zijn bovendien allebei gekoppeld aan GiGaNT-Molex, het centrale lexicon van het hedendaags Nederlands dat ook de basis vormt van de Woordenlijst Nederlandse Taal en het Groene Boekje.

Met de toevoegingen van het afgelopen jaar zijn in de huidige versie van het ANW ruim 286.000 woorden opgenomen die behandeld zijn in ruim 86.000 trefwoorden. De trefwoorden bevatten in totaal ruim 45.000 betekenissen.

In 2023 heeft het ANW de voorheen gebruikte set domeinnamen beperkt tot 56 hoofddomeinen, die grofweg gebaseerd zijn op de classificatie van de Library of Congress. Deze indeling wordt binnen het INT tevens gebruikt door het Expertisecentrum voor Nederlandse Terminologie bij het klasseren en opstellen van termenlijsten. Deze stroomlijning zal een verdere integratie van de woordenschatbeschrijving in de volgende jaren vergemakkelijken.

Behalve aan het ANW werkte de redactie ook mee aan het Woordenboek van Nieuwe Woorden (WNW) en aan de voorbereidingen voor de update van GiGaNT-Molex, o.a. vanuit het perspectief van de inhoudelijke consistentie tussen ANW, WNW en het centrale lexicon.

Woordenboek van Nieuwe Woorden (WNW)

Het INT onderzoekt de vorming van nieuwe woorden en houdt die woorden onder andere bij in woordenboeken. Sinds 2018 wordt er op het INT gewerkt aan het *Woordenboek van Nieuwe Woorden* (WNW): een online woordenboek waarin woorden die vanaf het jaar 2000 zijn ontstaan, worden beschreven. Het WNW neemt niet alleen de beklijvende nieuwe woorden op (bijvoorbeeld *app*, *selfie*, *keuzestress* en *ontspullen*), maar juist ook veel nieuwe woorden die kortere tijd bestaan (zoals *appaire* en *stookschaamte*). Ook in 2023 is er verder gewerkt aan dit woordenboek. Het WNW bevatte eind 2023 ruim 14000 woorden. Het WNW wordt dagelijks geüpdatet.

Alle woorden in het WNW zijn op dit moment digitaal gekoppeld aan de bijbehorende lemma's in het centrale lexicon GiGaNT-Molex en daarmee ook aan de bijbehorende ingangen in de Woordenlijst Nederlandse Taal. Naast GiGaNT-Molex werkte de redactie van het WNW ook mee aan het ANW (zie boven).

Ook afgelopen jaar is er iedere week een neologisme behandeld in de rubriek 'Nieuw woord van de week'. Verder publiceerde de WNW-redactie enkele artikelen over nieuwe woorden, hield de redactie lezingen, webinars en interviews (op radio en tv) over dit onderwerp en begeleidde de redactie diverse profielwerkstukken over neologismen, alsook een stagiaire en een masterstudent taalkunde van de opleiding Nederlands aan de Universiteit Leiden.

Woordcombinaties

Woordcombinaties is een online taaltool in ontwikkeling die leerders van het Nederlands als vreemde taal ondersteunt bij het gebruiken van woorden in context. Het project integreert de beschrijving van collocaties, idiomen en syntactische patronen en behandelt de combinatoriek van werkwoorden en substantieven. Het project is corpusgebaseerd en gebruikt de Sketch Engine als

corpusquerytool. Woordcombinaties toont hoe woorden gebruikt worden in goede voorbeeldzinnen, welke woorden met elkaar gecombineerd worden (collocaties of combinatiemogelijkheden) en hoe valentiepatronen en andere patronen samen met collocaties gebruikt worden voor het bouwen van zinnen. Taalleerders leren niet alleen woorden kennen, maar ook woorden gebruiken in context.

In het verslagjaar zijn patronen bewerkt van werkwoorden uit de schooltaalwoordenschat en het *Frequentiewoordenboek Nederlands* (Tiberius, C. en T. Schoonheim (2013)). Daarnaast zijn ook voorbeeldzinnen en combinaties bewerkt van werkwoorden en substantieven. Het datamodel voor de idiomen is getest en verder ontwikkeld en zal in 2024 gebruikt kunnen worden. Er is gewerkt aan richtlijnen voor consistentere lemmatisering van idiomen. In 2023 kwam er een update online, waaraan ook suggesties voor oefeningen in het onderwijs zijn toegevoegd. In samenwerking met de Taalunie gaf de redactie een webinar die op veel belangstelling kon rekenen.

Historische woordenboeken

De beschrijving van de historische woordenschat van het Nederlands is te vinden in de verschillende historische woordenboeken van het INT. Deze woordenboeken zijn online beschikbaar in het historische woordenboekenportaal (gtb.ivdnt.org): het *Oudnederlands Woordenboek* (ONW), het *Vroegmiddelnederlands Woordenboek* (VMNW), het *Middelnederlandsch Woordenboek* (MNW) en het *Woordenboek der Nederlandsche Taal* (WNT).

Om de rijkdom van deze historische woordenboeken en van de webapplicatie waarin ze te raadplegen zijn nog meer onder de aandacht te brengen, werd net als in voorgaande jaren in de webrubriek ‘Terug in de taal’ een twintigtal bijdragen rond aansprekende historische woorden en begrippen gepubliceerd. Daarnaast zijn er in 2023 drie historische gelegenheidswoordenboekjes online gekomen: *Het toilet van Couperus*, *Het hijgend hert en andere uitdrukkingen uit Het Boek der Psalmen* en *Oud nieuws*.

In samenwerking met de Taalunie zijn de Historische Woordenboeken – evenals drie andere taalbronnen van het INT – in de schijnwerpers geplaatst. Zo zijn er kennisclips gemaakt en is er een webinar over het gebruik van de historischewoordenboekenapplicatie georganiseerd, dat door deelnemers in een enquête positief werd beoordeeld. Daarnaast zijn tijdens presentaties en webinars de woordenrijkdom van deze woordenboeken en de onderzoeksmogelijkheden ervan voor het voetlicht gebracht. Uit de loggegevens én uit de reacties en vragen die we regelmatig zowel van specialistische als niet-specialistische gebruikers ontvangen, blijkt dat de webapplicatie ondertussen een veelgebruikt werkinstrument geworden is.

De woordenboekdata vormen de kern van de historische lexiconmodule van GiGaNT (GiGaNT-Hilex). In GiGaNT-Hilex vinden we de volgende woordenboekelementen: historisch lemma, modern Nederlands equivalent, woordsoort van een lemma, geattesteerde woordvormen van het lemma in de citaten en de citaten zelf. De werkzaamheden die ook in 2023 in dit kader zijn uitgevoerd, leveren correcties op die voor de online woordenboeken eveneens relevant zijn.

In 2023 is de XML-codering van de woordenboekdata grondig gereviseerd. Het datamodel is gereviseerd en de serialisatie naar TEI P5 is uitgevoerd. Er is een ontwerp gemaakt van een bewerkingsomgeving, die in 2024 gebouwd wordt om op een betrouwbare manier correcties in de TEI-XML van de woordenboekbestanden te kunnen uitvoeren. Hiermee komt een einde aan een

ingewikkeld conversietraject dat voorheen nodig was om de woordenboekdata te kunnen updaten voor het woordenboekportaal.

Citatenmateriaal van het WNT vormt ook onderdeel van het in ontwikkeling zijnde trainingsmateriaal voor het verrijken van historisch corpusmateriaal (zie boven).

Vertaalwoordenschat

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

Inmiddels staan het Nederlands-Nieuwgrieks / Nieuwgrieks-Nederlands, het Nederlands-Portugees / Portugees-Nederlands, het Nederlands-Estisch en het Nederlands-Fins / Fins-Nederlands online. De woordenboeken zijn gratis raadpleegbaar op de website en via de app Vertaalwoordenschat.

In 2023 is ook het taalpaar Nederlands-Deens toegevoegd aan de Vertaalwoordenschat. Samen met een data-update van het Nieuwgrieks-Nederlands / Nederlands-Nieuwgrieks is dit in de tweede week van januari 2024 officieel gereleased. Deens-Nederlands staat op de testserver, maar vereist eerst nog de nodige structurele en inhoudelijke aanpassingen (die voornamelijk handmatig dienen te gebeuren door oud-redacteuren) voordat het taalpaar officieel gereleased kan worden (gepland voor eind 2024). Daarnaast is in 2023 de onlineversie van de woordenboeken Noors voorbereid. Hiervoor is net als voor het Deens-Nederlands een apart conversietraject opgezet aangezien het om minimaal gestructureerde (Nederlands-Noors) en ongestructureerde (Noors-Nederlands) inputdata gaat.

De externe redacteuren die meegewerkt hebben aan de pilot met de editor in 2022, hebben in 2023 doorgewerkt aan correcties en aanvullingen in het Nieuwgrieks-Nederlands / Nederlands-Nieuwgrieks. Voor de andere taalparen is door externe redacteuren nog niet gebruikgemaakt van de mogelijkheid om de bestanden te verbeteren in de editor.

API’s en datasets

Ook in 2023 werd de lexiconservice waarmee GiGaNT-Hilex toegankelijk wordt gemaakt voor gebruik in bijvoorbeeld corpusapplicaties als Nederlab, Delpher, het Corpus Middelnederlands of de Gekaapte Brieven maandelijks geüpdatet.

Een update van GiGaNT-Molex naar versie 3.0 wordt gereleased in 2024.

5. Beschrijving van de Nederlandse dialecten

Elektronische Woordenbank van de Nederlandse dialecten (eWND)

In 2023 zijn met behulp van twee vrijwilligers enkele nieuwe woordenboeken toegevoegd aan de elektronische Woordenbank van de Nederlandse Dialecten (eWND). Verder is op 28 maart 2023 het e-WGD (elektronisch Woordenboek Gelderse Dialecten) gelanceerd o.l.v. Henk van den Heuvel, Roeland van Hout en Nicoline van der Sijs.

Database van de Zuidelijk-Nederlandse Dialecten (DSDD)

In 2023 zijn in de Database van de Zuidelijk-Nederlandse Dialecten (DSDD) geen nieuwe concepten toegevoegd maar er zijn wel correcties aangebracht. Vrijwilligers hebben in 2023 ongeveer 500 concepten nagekeken op tik- of koppelfouten. Die werden in de loop van 2023 gecorrigeerd.

Eind 2023 is een reeks nieuwe gegevens uit de Landbouwterminologie aangeleverd uit het Woordenboek van de Vlaamse Dialecten (het paard, aardappeloogst, landbouwvoertuigen, …). Die worden begin 2024 toegevoegd en gekoppeld aan de gegevens in de DSDD.

In 2023 is het *Woordenboek der Zeeuwse dialecten* geanalyseerd, en is een deel van de gegevens bewerkt en opgenomen in een proefomgeving waar ze gekoppeld zijn aan de DSDD. In een eerste test zijn er ongeveer 350 concepten gekoppeld.

Digitale infrastructuur voor het Bildts en andere taalvariëteiten

Het INT zetelt in de werkgroep contactvariëteiten, een werkgroep die bekijkt hoe dialectgemeenschappen (met als pilot het Bildts) eigen informatie kunnen samenbrengen en delen. In 2023 werd het pilotproject ‘Taalinfrastructuur voor de Bildtse taal’ opgestart. Na overleg met het Bildts Aigene is een applicatie gebouwd waarin de Bildtenaren hun gegevens kunnen inlezen en vergelijken met het al bestaande woordenboek van Buwalda. In de loop van 2024 zullen de gegevens via een bestaande infrastructurele voorziening (WoordWaark) beschikbaar gemaakt worden voor het grote publiek. Op basis van die pilot is in 2023 overleg opgestart met de Overijsselacademie i.v.m. het beschikbaar stellen van Overijsselse gegevens die moeten dienen om een zakwoordenboek voor het West-Overijssels te realiseren.

In het kader van CLARIAH Plus (zie aldaar) neemt het INT deel aan een deelproject ten behoeve van de integratie van dialectwoordenboeken in de CLARIAH-infrastructuur. In dat kader werden de data van het Woordenboek van de Gelderse Dialecten verder opgeschoond, zodat die later in een DSDD-omgeving kunnen worden ingelezen.

Overig

In 2023 verscheen tweewekelijks ‘Uit de Streek’, een rubriek over dialectwoorden uit de DSDD en de woordenbank. Het INT heeft in 2023 ook Limburgs cursusmateriaal beschikbaar gesteld via de website Limburgsluisteren.ivdnt.org. op verzoek van Levende Talen Limburgs.

6. Expertisecentrum voor Nederlandstalige Terminologie

Terminologie

Veldondersteuning en positionering in een terminologisch netwerk zijn de traditionele kerntaken van het Expertisecentrum Nederlandstalige Terminologie (ENT). In 2023 leidde dat tot de volgende realisaties.

Termenlijsten

Terminologie is een van de noodzakelijke aspecten in de ondersteuning en uitbouw van de communicatiepositie van de Nederlandse taal. Terminologie voorziet in de communicatiebehoeften in specialistische contexten, en de begripsomschrijvingen alsook de vertaalsuggesties in het geval van meertalige terminologie zijn bijzonder dienstbaar voor de vele taalgerichte beroepsbeoefenaars. Het ENT verzamelt daarom digitale termenlijsten in een eigen rubriek die gestaag wordt uitgebreid. Gezien de specifieke taakstelling van het INT houden ze steeds, eentalig of meertalig, verband met het Nederlands. Om de grote variatie in de onderwerpen en hun vaktaal te ordenen, worden ze ingedeeld volgens de onderwerpencatalogus van de Library of Congress, een internationaal systeem dat ook vele Nederlandse universiteitsbibliotheken hanteren om hun collectie te classificeren. Naast onderhoud en actualisering werden in de update 63 nieuwe termenlijsten toegevoegd.

Betreft het voorgaande online beschikbare een- of meertalige termverzamelingen waarvoor het INT de links inventariseert en beschikbaar stelt, dan worden daarnaast ook kansen benut die zich voordoen door bestanden te hosten. In dit opzicht is een samenwerkingsovereenkomst aangegaan voor het online beschikbaar stellen van een viertalig bestand met ruim 15.000 financieel-economische begrippen (Nederlands, Engels, Frans en Duits) en een aansluitend meertalig bestand met meer dan 1700 frequente afkortingen die het begrip van teksten van financiële analisten en commentatoren bevorderen alsook de vertaling van financiële documenten. Bestandsanalyse, aansluitende aanpassingen en omzetting in TBX-formaat vonden reeds plaats en een zoekinterface voor viertalige ingangen is ontwikkeld. Tests en publicatie zijn beoogd in 2024.

Tools

In 2023 zijn werkzaamheden verdergezet om een nieuwe terminologietool te ontwikkelen. Ze vormen een eerste implementatie van de beslissing om twee bestaande taaltechnologische hulpmiddelen te moderniseren en te integreren in een nieuwe online werkomgeving, te weten de voormalige TermTreffer, geschikt voor termextractie en bepaalde editeermogelijkheden, en de TermBeheerder, ontworpen voor het management van termverzamelingen. Ter voorbereiding is in 2023 eerst een alfa-versie ontwikkeld die bedoeld is voor het uittesten van een aantal basisfunctionaliteiten. Binnen een applicatietestomgeving (ato) is deze tool toegankelijk gemaakt voor een beperkte groep testgebruikers. Deze versie is ook ontwikkeld om te onderzoeken of de bestaande maar moeilijk te onderhouden TermTreffer opnieuw kan worden geïmplementeerd binnen de bredere toolset van het Instituut voor de Nederlandse Taal met het oog op een duurzame

verdere ontwikkeling door het INT. De TermTreffer2-α laat gebruikers toe om hun eigen Nederlandstalige teksten als een domeinspecifiek corpus te compileren en te doorzoeken, vervolgens uit dit corpus automatisch een lijst van kandidaattermen te extraheren, samen met hun frequentie, termhood-scores en lemmatisering, en daarna deze lijst te exporteren als Excel- of TBX-bestand. De TermTreffer2-α is in de lente en zomer van 2023 extern uitgetest en vergeleken met de oude TermTreffer door studenten van het vak ICT en Terminologie uit de bachelor Toegepaste Taalkunde aan de KU Leuven (campus Antwerpen) en daarna ook door 2 werkstudenten in een door de Taalunie gefinancierd project rond medische terminologie, met name de uitbreiding van het Pinkhof-woordenboek met COVID-19-gerelateerde termen. Deze evaluaties gaven aan dat de TermTreffer2-α zowel op het vlak van termextractie als gebruiksvriendelijkheid even goed of zelfs lichtjes beter scoort dan de oude TermTreffer, maar dat een uitbreiding met nieuwere sequentiële extractietechnieken, zoals D-Terminer, een duidelijke meerwaarde zou zijn. Hiervoor wordt dan ook samengewerkt met de academische onderzoekers in dit domein.

Op basis van al deze inzichten is in het najaar van 2023 begonnen met de verdere uitwerking van een bètaversie en een productieversie waarbij de termextractiefuncties en de editeeromgeving voor termenbanken in één platform TermWerk samengebracht worden. Als eerste stap is er een nieuwe module ontwikkeld voor gebruikersauthenticatie, zodat zowel bestaande gebruikers van de oude TermTreffer/TermBeheerder als nieuwe gebruikers (via een CLARIN-login) van TermWerk gebruik zullen kunnen maken. De realisatie van de geïntegreerde omgeving is voorzien voor de eerste helft van 2024.

Veldondersteuning

Voor de veldondersteuning werden in 2023 opnieuw de websites van het INT en ENT ingezet, evenals projecten die door hun thematiek en online raadpleegbare materiaalverzamelingen een toegepaste waarde hebben.

Zo werden vier nieuwsbrieven verstuurd en vervolgens op de website gepubliceerd en gearchiveerd. Traditioneel bestaan ze uit een aantal vaste rubrieken over voorbije evenementen, nieuwe ontwikkelingen en publicaties, tools en bronnen, berichten van NL-Term en een agenda. Als veldondersteuning fungeerde ook de evenementenrubriek op de ENT-webpagina met geregelde updates over terminologiecongressen en -symposia in Europa.

Op de webrubriek Terminologieprojecten kon het HOTNeV-project over hogeronderwijsterminologie verder worden uitgebreid door de resultaten van een stagebegeleiding. Een update om de resultaten van termanalyses en -beschrijvingen online beschikbaar te stellen, heeft weer plaatsgevonden.

Medische vaktaal, juridische vaktaal en Nederlands als wetenschapstaal

Speerpunt 1: de medische vaktaal

De database van *Pinkhof Geneeskundig woordenboek*, sinds 2021 online raadpleegbaar bij de INT-taalmaterialen, gaat terug op de editie van 2012 met aanvullingen en verbeteringen in de loop van de tijd. Het streven is om de huidige versie te herzien. Daarvoor is een samenwerkingsverband opgericht tussen het INT en de Stichting Beheer Database Pinkhof Geneeskundig Woordenboek, die finaal de beslissingen neemt over de aanpassingen. Zo'n herziening stelt eisen in velerlei opzicht. Het INT verzocht de Stichting om een eerste verkenning inzake inhoudelijke richtlijnen voor het schrijven en opnemen van nieuwe lemmata, mede met het oog op beoogde doelgroepen. E.e.a. resulteerde in een document van de Stichting met opvattingen voor de realisatie van een Pinkhof 2.0. Naar aanleiding hiervan heeft het INT medio februari 2024 overleg gepland met de Raad van Advies medische terminologie, inz. over inhoud, samenwerking, en financiering. Het INT zette ook de verkenning naar een geschikte lexicografische dan wel terminologische editeeromgeving verder.

Ter ondersteuning van het Pinkhofproject heeft ook een pilootproject over coronatermen zijn beslag gekregen. Doel van dit coronaproject was het ontwikkelen en testen van een strategie voor het toevoegen van medische termen – artsenjargon én lekentermen – aan het *Pinkhof Geneeskundig woordenboek* en zo te bepalen hoe dit medische woordenboek in de toekomst ook op andere deeldomeinen van de geneeskunde verder kan worden aangevuld of verbeterd op een systematische, beredeneerde manier. Als eindresultaat van dit pilootproject zijn twee doelen gerealiseerd. Het eerste betrof een lijst met kandidaattermen uit de coronaterminologie met verwijzing naar de concept-id's uit de internationale medische ontologie SNOMED CT. Dit doel werd gerealiseerd door vergelijkenderwijs een coronagerelateerde lijst van kandidaattermen samen te stellen, deels op basis van bestaande termenlijsten, deels door extractie van nieuwe termen uit een samen te stellen corpus van teksten uit het publieke domein, en deze te toetsen aan het Pinkhofbestand en het SNOMED-bestand. Het tweede doel had betrekking op het verkrijgen van meer inzicht in de precisie van termextractie uit een corpus van coronagerelateerde medische artikelen door de oude versie van TermTreffer, de nieuwe versie en D-Terminer. Na bepaling van de manifeste termen in de eerste 300 kandidaattermen van enkel- en meerwoordige lijsten werd de precisie van de termextractietools berekend en konden de tools met elkaar worden vergeleken. Nadien werd nagegaan of de als echte termen gekwalificeerde kandidaten al in het *Pinkhof Geneeskundig woordenboek* stonden. Termen die nog niet in het woordenboek stonden, werden gekoppeld aan de code van de overeenkomende term uit de Nederlandse – en dus niet de Belgisch-Nederlandse – SNOMED CT.

Speerpunt 2: de juridische vaktaal

In het kader van het Expertisecentrum Nederlandstalige Terminologie dat het INT in de lijn van taakstellingen van het Taalunieverdrag van 1980 uitbouwt, passen ook terminologiecollecties over onder meer het juridische domein. Met het hosten van dergelijke bestanden kan het INT kennis ondersteunen die bijdraagt aan de internationale context van rechtsvergelijking en juridische vertaling. Na voorbereidingen in 2022 vond in 2023 de online publicatie plaats van het *Juridisch Woordenboek Nederlands-Spaans met register Spaans-Nederlands voor rechtspraktijk, handel en bedrijfscommunicatie* van M.C. Oosterveld-Egaz Repáraz en J.B. Vuyk-Bosdriesz. Het behandelt de

terminologie van het Nederlandse rechtssysteem wat betreft de uitgangstermen, en de rechtstermen van het Spaanse rechtssysteem als doeltaal. Hierbij is een methode ontwikkeld die gelijktijdig een lexicografische en rechtsvergelijkende benadering toepast en die resulteerde in een praktijkgericht en wetenschappelijk verantwoord woordenboek voor een beoogd publiek van juridische vertalers en internationaal werkende juristen. Het materiaal is in samenwerking met IT'ers en computerlinguïsten van het INT beschikbaar gesteld in een zoekinterface. Tijdens de Week van het Nederlands presenteerden het INT en de Stichting Rechtstaal – Lingua Iuris op 4 oktober 2023 dit woordenboek in een academisch symposium te Leiden, waar juridische experten de verdere uitbouw naar de Belgisch-Nederlandse rechtstaal hebben bepleit.

Speerpunt 3: Nederlands als wetenschapstaal

Binnen een samenwerkingsverband tussen de Stichting Nederlands / Vlaams Platform Taalbeleid Hoger Onderwijs, de Taalunie, KU Leuven, UGent en het INT leverde destijds het 'Proefproject Nederlands als wetenschapstaal – van corpora naar terminologielijsten' drie termenbestanden op in de bètarichtingen. Vaktalige begrippen uit de scheikunde, wiskunde en natuurkunde werden er voorzien van definities, een Engelse vertaling, toelichtingen en voorbeelden. Het doel was studenten in de overstap naar het eerste bachelorjaar de vaktalige ondersteuning te geven die uit onderzoek wenselijk was gebleken, en om de positie van het Nederlands als wetenschapstaal te versterken. Het scheikundebestand is in 2023 ter beschikking gesteld met een online zoekapplicatie. Het wiskundebestand is na verdere controle omgezet in een TBX-versie, en aanvulling van semantische relaties zoals synoniemen, hyperoniemen en hyponiemen heeft plaatsgevonden. Het natuurkundebestand heeft in 2023 een tekstuele correctieslag ondergaan en is vervolgens omgezet naar TBX voor verdere digitale optimalisering. Publicatie van zowel het wiskundebestand als het natuurkundebestand is beoogd voor 2024.

Terminologisch netwerk

Deelname aan of bijwonen van congressen en symposia droeg ook in 2023 bij aan de positionering in het terminologische netwerk. In dit verband vermelden we: European Association for Terminology 11th Terminology Summit (Barcelona), Terminologie diachronique: un bilan, des perspectives (Lyon). In het geval van het internationale congres in 'terminology: domain loss and gain' (Brussel) was het INT mede-initiatiefnemer, samen met NL-Term, EAFT en Infoterm. Traditioneel nam het ENT opnieuw deel aan de terminologievergaderingen die de Termraad Nederlands belegde voor Nederlandstalige (ver)taaldiensten van EU-instellingen en van Nederlandse, Belgische en Vlaamse overheden, en participeerde het ENT in een stagevoorlichting van de Termraad Academy. Mede werd bijgedragen aan het overleg van de Focusgroep Begrijpelijke Taal Nictiz. Stagebegeleiding vond ditmaal plaats voor een masterstudent van de universiteit van Straatsburg.

7. Grammatica

e-ANS

In 2023 is het vernieuwde hoofdstuk 7 Het telwoord gepubliceerd. Het is door Kathy Rys inhoudelijk in overeenstemming gebracht met de huidige inzichten in dit onderdeel van de Nederlandse grammatica. Het hoofdstuk is daarnaast op verschillende manieren verrijkt, volgens de doelstellingen van de herziening van de e-ANS. Zo zijn er uitgebreide literatuurverwijzingen toegevoegd, koppelingen naar taaladviesbronnen, en voorbeelden uit het Corpus Hedendaags Nederlands. De voorbeelden zijn voorzien van informatie over variatie, zoals gebruiksverschillen tussen Nederland en België, en of een vorm vooral in informeel of juist formeel taalgebruik voorkomt.

Bovendien zijn in 2023 onderwijsmodules toegevoegd aan hoofdstuk 7 Het telwoord en hoofdstuk 12 Woordvorming en woordstructuur. Deze onderwijsmodules zijn speciaal ontwikkeld voor docenten en leerders van de Nederlandse grammatica, voor wie het niveau van de ANS vaak net te hoog ligt. De modules bevatten samenvattingen, termenlijsten, veelgestelde vragen en lesideeën.

Op technisch gebied is er regulier onderhoud gepleegd aan de ANS, en zijn o.a. geluidsfragmenten toegevoegd aan het hoofdstuk over de klankleer. Werkzaamheden aan het Grammaticaportaal, een overkoepelende portal-pagina voor zowel de e-ANS als Taalportaal en Taaladvies, zijn doorgeschoven naar het volgende jaar.

Taalportaal

De Taalportaalapplicatie is dit jaar grondig herzien, waarbij de website in lijn is gebracht met de huisstijl van het INT. Daarnaast is in samenwerking met de Fryske Akademie een vierde taal toegevoegd aan de site, het Saterfries. De teksten over de morfologie en de syntaxis zijn op het INT geconverteerd naar XML en gepubliceerd.

Taaladvies

Aan Taaladvies.net, een zeer druk bezochte site die onder beheer is bij het INT, is afgelopen jaar regulier onderhoud gepleegd. Er is onderling contact tussen de redacties van Taaladvies.net en ANS om de kruisverwijzingen bij herziene hoofdstukken te corrigeren.

8. Nationale en internationale samenwerkingsverbanden

Netwerken

IMPACT Centre of Competence

Het INT is voorzitter van het IMPACT Centre of Competence (www.digitisation.eu). Dit is een non-profitorganisatie bestaande uit publieke en commerciële organisaties met als doel de digitalisering van historisch materiaal “beter, sneller, en goedkoper” te maken. Het centrum voorziet in data, tools, services en expertise op het gebied van document imaging, taaltechnologie en het verwerken van historisch tekstmateriaal. Het IMPACT Centre of Competence is sedert 2019 ook CLARIN Knowledge centre. De werkzaamheden m.b.t. digitalisering die in de context van CLARIAH Plus worden uitgevoerd, worden in samenwerking met het Centre uitgevoerd.

In 2023 is de website van het IMPACT Centre of Competence vernieuwd. Er is gewerkt aan een nieuwe omgeving om de IMPACT ground truth dataset toegankelijk te maken en die omgeving wordt in 2024 gereleased. Tot slot is er verder gewerkt aan een white paper rondom het thema Sharing and Sustaining Digitisation Knowledge. De doelgroep van de white paper zijn onderzoekers en erfgoedinstellingen en de focus ligt op de uitdagingen die het veranderende digitaliseringslandschap met zich meebrengt.

European Language Data Space (voorheen ELRC en ELG)

In januari 2023 is het European Language Resources Coordination-initiatief (ELRC) overgegaan in de European Language Data Space. Ook binnen dit initiatief is het INT het nationale aanspreekpunt. In 2023 is een subcontract getekend voor het organiseren van een ‘country workshop’ waar alle stakeholders geïnformeerd worden over de European Language Data Space. Deze workshop staat gepland voor 2024. In het kader van de European Language Data Space wordt er momenteel een European Digital Infrastructure Consortium (EDIC) voorbereid onder de naam Alliance for Language Technologies. Het INT volgt deze ontwikkeling mee op als Technical National Anchor Point voor ELRC in Nederland.

Elexis Association

Het INT was partner in ELEXIS, een Europees Horizon 2020-2022-project, waarbinnen een infrastructuur voor e-lexicografie is opgezet. Om de duurzaamheid van de infrastructuur te waarborgen is een nieuwe associatie in het leven geroepen, de ELEXIS Association. Doel van de associatie is het organiseren en coördineren van lexicografisch gerelateerde activiteiten en NLP-activiteiten voor zover die relevant zijn voor de lexicografie. Het INT is lid geworden van de associatie. Carole Tiberius is gekozen als Deputy van de President van de ELEXIS Association.

Netwerkprojecten

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

Het INT neemt deel aan de COST-actie NexusLinguarum. Het thema van deze actie is ‘linguistic data science’, een deelgebied binnen de opkomende 'data science'. Taalkundige data vormen een specifiek geval en zijn tot nu toe nog grotendeels onontgonnen in een bigdatacontext. Het hoofddoel van NexusLinguarum is om taalkundigen, computerwetenschappers, terminologen en andere belanghebbenden in één netwerk bij elkaar te brengen om zo samenwerking en kennisdeling op het gebied van ‘linguistic data science’ te bevorderen. De actie is eind oktober 2019 van start gegaan en heeft een looptijd van 4 jaar die eindigt op 27 april 2024.

Carole Tiberius is MC voor Nederland en Kris Heylen neemt deel aan WG3 (Use cases and applications). Carole en Kris organiseerden in samenwerking met Jelena Kallas (EKI) en Ilan Kernerman (Lexicala), en ondersteund door de NexusLinguarum COST-actie, de workshop *Linking Lexicographic and Language Learning Resources* (4LR), op het congres LDK 2023 (Wenen, 13 september 2023).

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

Het INT neemt deel aan het Europese onderzoeksnetwerk UniDive (Universality, diversity and idiosyncrasy in language technology). Het doel van deze COST-actie is om te onderzoeken hoe taaltechnologie verbeterd kan worden door te focussen op universaliteit, maar tegelijkertijd ook de diversiteit van talen in het oog te houden. Dit om te voorkomen dat talen met weinig middelen en bedreigde talen niet buitengesloten worden in de digitale wereld. Op 16-17 maart 2023 vond de eerste algemene bijeenkomst van deze COST-actie plaats in Parijs. Het INT is actief binnen twee werkgroepen (Corpusannotatie en Lexicon-Corpus Interface).

European Network On Lexical Innovation (ENEOLI, 2023-2027)

Het INT neemt deel aan het Europese onderzoeksnetwerk (COST) European Network On Lexical Innovation (ENEOLI). De belangrijkste doelstellingen van het netwerk kunnen als volgt worden samengevat: 1) Een gemeenschappelijke, meertalige kernterminologie voor lexicale innovatie definiëren van 2) Digitale methodes en instrumenten in kaart brengen die gebruikt worden om lexicale innovaties in Europese talen te identificeren en te verklaren; 3) Vergelijkende studies uit te voeren naar lexicale innovaties in Europese talen, met speciale aandacht voor ontleningen en hun equivalenten; 4) Specifieke training in neologie aan te bieden aan vertalers, redacteuren, journalisten, technisch schrijvers en docenten. Op 6 oktober 2023 vond de eerste bijeenkomst van het Management Committee van deze COST-actie plaats in Brussel.

Onderzoeks- en infrastructuurprojecten

CLARIAH-Vlaanderen (2021-2024)

Binnen het FWO-IRI-project 'CLARIAH-VL: Advancing the open humanities service infrastructure' is de hoofdtaak van het INT het voorzien van de benodigde infrastructuur voor het opzetten van de Digital Text Analysis Dashboard & Pipeline. Door personeelswijzigingen aan de KU Leuven heeft dit vertraging opgelopen en werd er nog geen beroep gedaan op INT-infrastructuur.

Er werd in het kader van CLARIAH-VL ook verder gewerkt aan een pilotproject in samenwerking met de Vlaamse Super Computer (VSC), met als doel een contextueel taalmodel te trainen op basis van de corpora hedendaags Nederlands waarover het INT beschikt. Het INT stelde voor de onderzoekers van de UGent data rechtstreeks ter beschikking op de HPC (High Performance Cluster).

CLARIAH Plus-Nederland (2019-2023)

Het vervolgproject van CLARIAH (Common Lab for Research in the Arts and Humanities) loopt van 2019 tot en met 2023, met een extensie tot medio 2024. 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.
In 2023 is de syntactische module van BlackLab afgerond, en is het annoteren van corpuszoekresultaten in de corpus-frontend-userinterface geïntegreerd. Een en ander kan in productie worden genomen als de authenticatiemodule afgerond is. De module voor parallele corpora wordt in 2024 afgerond.

Het GaLAHaD-platform voor betere en gebruikersvriendelijke taalkundige verrijking van historisch Nederlands is goed gevorderd en zal begin 2024 worden afgerond. Het deel van de trainingcorpora voor historisch Nederlands dat niet op bestaande verrijkte corpora is gebaseerd, is grotendeels afgerond; op basis van dit materiaal zijn taggers en lemmatisers getraind en opgenomen in GaLAHaD.

Binnen de infrastructuur voor digitalisatie en conversie is gewerkt aan het beoogde platform voor het delen en publiceren van datasets en modellen die in lopende en afgelopen digitaliseringsprojecten zijn ontwikkeld. Het platform bevat een uitgebreide uitleg van het digitaliseringsproces en heeft mede als doel de keuzes die bij het digitaliseringsproces gemaakt moeten worden te ondersteunen.

Het werk wordt begin 2024 voltooid.

SSHOC-NL (Social Science and Humanities Open Cloud for the Netherlands)

Dit vervolgproject van CLARIAH Plus beoogt te komen tot een consortium van onderzoeksinfrastructuren, gericht op het creëren van een ecosysteem van diensten, gegevens en instrumenten voor de sociale wetenschappen en menswetenschappen. Het consortium wordt geleid door ODISSEI, de Nederlandse nationale infrastructuur voor sociale wetenschappen en CLARIAH, de

Nederlandse nationale infrastructuur voor geesteswetenschappen. Binnen het project zal het INT zich onder andere richten op de infrastructuur voor het methodologisch verantwoord inzetten van machine learning en AI voor dataverrijking.

ParlaMint II (2021-2023)

ParlaMint is een door CLARIN-ERIC gefinancierd project, dat bijdraagt aan de totstandkoming van vergelijkbare en uniform geannoteerde meertalige corpora van parlementaire zittingen. ParlaMint I heeft voor 17 talen corpora opgeleverd. ParlaMint II heeft het dataschema verfijnd, onder meer met extra metadata, de bestaande corpora uitgebreid tot juli 2022, en corpora voor nieuwe talen toegevoegd. Het INT was verantwoordelijk voor de data van het Belgisch federaal parlement en de taalkundige verrijking van de Nederlandse parlementaire data.

SignON (2021-2024)

Het INT was als consortiumpartner betrokken bij het SignON-project, dat vanaf voorjaar 2021 voor drie jaar gefinancierd werd binnen het kader van het Horizon 2020-programma van de Europese Commissie. Het hoofddoel van dit project is het opzetten van automatische vertaalservices tussen gebarentalen en zogenaamde gesproken talen. De gebarentalen die bovenaan de agenda staan van deze Research and Innovation Action zijn Vlaamse Gebarentaal (VGT), Nederlandse Gebarentaal (NGT) en Ierse Gebarentaal. Gesproken talen zijn in eerste instantie het Nederlands en het Engels. Het consortium van dit project heeft een sterk Belgisch-Nederlandse component, met als consortiumpartners uit België: VRT, KU Leuven, UGent, Vlaams Gebarentaalcentrum en European Union for the Deaf. Vanuit Nederland nemen deel: INT, de Taalunie, Radboud Universiteit Nijmegen, Tilburg University, en als derde partij Beeld en Geluid. Het project wordt geleid door Dublin City University.

De taak van het INT bestond hoofdzakelijk uit het opzetten van de infrastructuur voor dit onderzoek. Er werd eveneens gewerkt aan het verzamelen van gebarentaalcorpora, zowel voor VGT als voor NGT. Samen met Universiteit Tilburg werd extra financiering vanuit European Language Equality gevonden om een dataset te maken waarbij online hotelrecensies door dove vertalers vertaald werden naar NGT. Het NGT-HoReCo-project werd uitgevoerd in 2023, en uitgebreid met VGT. Resultaten zijn beschikbaar. Daarnaast werd voor een andere dataset samen met de Universiteit Tilburg extra financiering vanuit de European Association for Machine Translation gevonden voor het vertalen naar het Nederlands van reeds bestaande video's met VGT als brontaal werd eveneens gehonoreerd. Dit project werd grotendeels uitgevoerd in 2023, en zal beschikbaar komen in het voorjaar van 2024.

Een andere taak binnen SignON was om de infrastructuur op te zetten om Vertalen-als-een-service aan te kunnen bieden, die dan aangesproken kan worden binnen de Android- en iPhone-apps die ontwikkeld worden in de use cases, die in samenspraak met de doelgroepen ontwikkeld worden. Die infrastructuur werd onderhouden en verder uitgebouwd in 2023.

Gelet op het aflopen van SignON op 31/12/2023 werd met de Vlaamse partners UGent, KU Leuven en VGTC een projectaanvraag ingediend bij het FWO in het Strategisch Basisonderzoeksprogramma. In het Signify-project wordt ook financiering voor INT voorzien voor het ontwikkelen van een collaboratieve online omgeving voor de annotatie van videocorpora. In de loop van 2024 wordt duidelijk of deze aanvraag gehonoreerd wordt.

SABeD (2021-2024)

Het INT werkte mee aan de supervisie van de ontwikkeling van het corpus Spoken Academic Belgian Dutch van de KU Leuven, en bouwde eveneens een aantal tools om met gesproken data om te gaan. De eerste twee batches van dit corpus (i.e. 170 lezingen) werden manueel getranscribeerd en door het INT verder geprocest. De finale oplevering van dit corpus is voorzien voor 2024.

ClaSABeD (2022-2023)

In dit project passen we CLARIAH-Plus- en CLARIAH_NL-tools toe op het SABeD-corpus. Het SABeD-corpus werd geanalyseerd met Frog en beschikbaar gesteld voor gebruikers van het corpus (op verzoek) via Autosearch. Er werd ook onderzocht of de PICCL- of TICCL-tools geschikt zouden zijn om automatische post-editing van spraakherkenning te doen, maar dat bleek niet het geval. Het eindrapport werd opgeleverd aan CLARIAH-NL en is beschikbaar op aanvraag.

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

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

In 2023 is een conversie- en indexeringstraject ontwikkeld dat door UGent aangeleverde database eerst omzet naar – zo goed mogelijk bij het bestandsformaat van het Corpus Gesproken Nederlands aansluitende – FoLiA-XML-bestanden, en deze bestanden vervolgens indexeert in de corpuszoekmachine BlackLab.

Tevens is een eerste prototype voor de voorziene corpusretrieval-applicatie opgeleverd.

Pilootproject Duidelijke Taal (2023-2024)

In dit project wordt een dataset gevalideerd die als goudstandaard kan dienen bij automatische tekstvereenvoudiging. Bestaande data en automatisch gegenereerde data laten we beoordelen door middel van crowdsourcing, waarbij de crowd zinnen kan beoordelen op drie dimensies: accuraatheid, vlotheid en complexiteit. In 2023 werd de 1e fase van dit project opgestart, waarbij er data van verschillende bronnen verzameld werden, waaronder overheidsteksten.

Spread the News (2020-2025)

Het project ‘Spread the new(s). Understanding standardization of Dutch through 17th-century newspapers’, dat wordt uitgevoerd met subsidie van NWO Open Competition SSH, heeft ook in 2023 goede vorderingen gemaakt. Promovenda Machteld de Vos, die gedeeltelijk op het INT en gedeeltelijk bij de Radboud Universiteit werkt, heeft o.a. een artikel geschreven met Ulrike Vogl van de Gentse universiteit en ze heeft enkele weken in Leuven doorgebracht en daar met technische ondersteuning van Freek Van de Velde gewerkt aan een ander artikel. In de tweede helft van 2023 lag het promotiewerk stil omdat de promovenda een half jaar als tijdelijk Programma-/

beleidsmedewerker aangesteld was bij NWO. Daarom is de einddatum van het project uitgesteld tot 14 juni 2025.

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. Het project, voorzien voor 2023, is uitgesteld tot begin 2024.

Overige infrastructurele dienstverlening

Etymologiebank

Sinds 2020 zijn de etymologiebank (etymologiebank.nl), en de uitleenwoordenbank (uitleenwoordenbank.ivdnt.org) in het beheer van het INT. In 2023 hebben vier stagiairs een bijdrage geleverd aan de verrijking van de etymologiebank.

GLAD

Verder is het INT verantwoordelijk voor het hosten van de data van het Global Anglicism Database Network (GLAD); hiervoor is een online bewerkingsapplicatie met behulp van het Lex'it-platform ontwikkeld. Twee stagiaires, waaronder een Russische, hebben aan GLAD meegewerkt. De publieksversie met Engelse leenwoorden in 17 talen is in 2023 gelanceerd op https://glad.ivdnt.org/. De Nederlandse gegevens van GLAD zijn in 2023 uitgebreid tot 9600 trefwoorden.

DaGeNTa

Van de 'Database Geschiedenis Nederlandse Taalkunde' (DAGENTA) is een uitgebreide versie gepubliceerd in juni 2023 op https://dagenta.ivdnt.org/; aan de uitbreiding is meegewerkt door een stagiair.

Pallas

Aan de ontsluiting van de Digitale Pallas, een 18e-eeuws Russisch meertalig woordenboek, hebben in 2022 twee stagiaires meegewerkt. Het resultaat is op 12 september 2023 in open access beschikbaar gekomen als 'The Digital Pallas'. Inhoud van Peter Simon Pallas, Comparative Dictionary of All Languages and Dialects (1790-1791): https://pallas.ivdnt.org/.

9. Disseminatie

Doelgroepenbeleid (inclusief onderwijs)

Het INT richt zich als toegepast-wetenschappelijk instituut traditioneel op onderzoekers en taalkundigen. Bestaande contacten met onderzoekers, al dan niet verbonden aan wetenschappelijke instituten en universiteiten, worden onderhouden en waar mogelijk geïntensiveerd en uitgebreid. Aan de KU Leuven is het vak Computationele Lexicografie gegeven in de Master Taalkunde en Taal- en Letterkunde en aan de Universiteit Leiden de collegereeks Corpus Lexicography en Computational Corpus Analysis. Daarnaast heeft het INT zijn werkterrein actief verbreed naar docenten en studenten. In dat verband was het Instituut aanwezig op en profileerde het zich op beurzen, conferenties (HSN-conferentie), festivals (Week van het Nederlands, Letterlijk Leiden) en wetenschappelijke bijeenkomsten. Ook heeft het INT opnieuw een opgave gemaakt voor de jaarlijkse Taalkundeolympiade van de Universiteit Leiden.

Op de website van het INT is een apart menu-item voor onderwijs ingericht, onderverdeeld in informatie voor docenten, leerlingen en studenten. Het INT heeft in 2023 in samenwerking met de Taalunie webinars georganiseerd over werken met online taalbronnen.

Ook het algemene publiek wordt niet uit het oog verloren. Zo heeft het INT in het kader van het festival Letterlijk Leiden opnieuw de talige wandeling 'Weg van woorden' aangeboden, die ook als audiotour beschikbaar is. Op onze website verschenen wekelijks populairwetenschappelijke rubrieken van eigen medewerkers over grammatica, dialectwoorden, nieuwe woorden en historische woorden. We zijn actief op verschillende sociale media en maken op regelmatige basis nieuwe afleveringen van onze podcastserie *Over taal gesproken* (die ook gebruikt worden als lesmateriaal). Tien keer per jaar werden twee soorten nieuwsbrieven verstuurd aan geïnteresseerden. Daarnaast hielden medewerkers voordrachten, waren zij te horen in radioprogramma's en droegen zij bij aan publicaties voor een algemeen publiek dat belangstelling heeft voor taal in het algemeen en Nederlands in het bijzonder. Zie daarvoor het uitgebreide overzicht in bijlage 3.

Communicatiemiddelen

Website

De website is vaak het eerste wat het publiek ziet van het instituut, en dat maakt het een van de belangrijkste communicatiemiddelen. Veel communicatie-uitingen zijn erop gericht om bezoekers naar de website te trekken. Sinds 2023 worden de statistieken bijgehouden met Plausible. Daarvoor werd Google Analytics gebruikt om statistieken te verzamelen. In 2023 telde ivdnt.org 570.699 paginaweergaven (378.782 unieke bezoekers), een stijging ten opzichte van het voorgaande jaar (half miljoen paginaweergaven in 2022).

Het grootste deel van de bezoekers (225.000) komt via Google op de website terecht. 76.200 bezoekers tikken ivdnt.org direct in de browser in, 55.300 bezoekers komen via diverse nieuwsbrieven (eigen nieuwsbrief, Taalpost, Taalunienieuwsbrief) en 9.800 via de sociale media: de

meesten daarvan vinden de website via X (voorheen Twitter), gevolgd door Facebook, LinkedIn en Instagram.

Van de statische content is de pagina van het *Woordenboek der Nederlandsche Taal* (WNT) het vaakst bezocht met 26.243 paginabezoeken (21.080 unieke bezoekers). De best bezochte actuele rubriek blijft 'Nieuw woord van de week', met in totaal 122.186 bezoeken (87.942 unieke bezoekers). Gemiddeld blijven de bezoekers zo'n 3 minuten op de pagina. Ook de (sinds 2020 niet meer bijgehouden) rubriek 'Woordbaak' wordt nog regelmatig geraadpleegd met in 2023 96.704 bezoeken in totaal (87.764 unieke bezoekers).

Het onderdeel Terminologie kreeg dit jaar 6.077 bezoeken (4.281 unieke bezoekers), een lichte stijging ten opzichte van 2022; zo'n 1.100 van die bezoeken waren voor de termenlijsten.

De columns van Ewoud Sanders en Ludo Permentier zijn met ingang van 2023 gestopt en er is sinds maart een nieuwe actuele rubriek bijgekomen: Grammaticasafari, met tweewekelijks weetjes uit de Algemene Nederlandse Spraakkunst.

Totaal aantal paginabezoeken per categorie

Actueel	332.722
Woordenboeken	118.553
Spelling & grammatica	7.251
Corpora & lexica	6.721
Over ons	6.562
Terminologie	6.077
Themapagina Historisch Nederlands	3.657
Onderwijs	2.841
Themapagina Hedendaags Nederlands	2.570
Onderzoek & projecten	1.914
Taalmaterialen	1.824

Totaal aantal paginabezoeken Woorden van de week

Nieuw woord van de week	122.186
Woordbaak	96.704
Uit de streek	19.203
Terug in de taal	12.733
Grammaticasafari (vanaf maart 2023)	10.333

48% van de bezoekers raadpleegt ivdnt.org op een smartphone, 27% maakt gebruik van een desktop, 21% van een laptop, en 4% van een tablet. Sinds augustus 2023 wordt bijgehouden hoeveel bezoekers doorscrollen naar het einde van de homepage: in een half jaar tijd geldt dat voor zo'n 5% van de bezoeken.

Aan de achterkant van de website wordt continu gewerkt aan verbetering. Zo worden fouten hersteld, kapotte links verwijderd of vervangen, wordt de veiligheid verbeterd, en wordt op verschillende manieren gewerkt aan het versnellen van de laadtijd en zoekmachineoptimalisatie (SEO). In de SEO zijn grote stappen gemaakt: wanneer je via een incognitovenster in Google zoekt op termen als 'Nederlandse taal' (tweede resultaat), 'nieuwe woorden' (eerste resultaat), 'hedendaags Nederlands' (tweede resultaat), 'historisch Nederlands' (tweede resultaat) of 'terminologie' (tweede resultaat), verschijnt de INT-website steeds in de bovenste zoekresultaten.

Huisstijl

Een team bestaande uit de webmaster, softwareontwikkelaars en de communicatieadviseur heeft een standaardwebapplicatie ontwikkeld die als vertrekpunt dient voor de ontwikkeling van nieuwe INT-applicaties. De gelijkvormigheid en herkenbare INT-huisstijl zorgen ervoor dat alle door het INT ontwikkelde applicaties door de gebruikers ook als zodanig herkend worden. Die vormgeving wordt nu zo veel mogelijk geïmplementeerd in nieuwe en bestaande INT-websites. In 2023 is dat onder andere gebeurd voor DAGENTA en Sofeer.

Nieuwsbrieven & persberichten

In 2023 zijn er zes algemene INT-nieuwsbrieven verschenen en vier nieuwsbrieven over terminologie. De algemene nieuwsbrief telde eind 2023 4.490 abonnees, van wie per editie 49% de nieuwsbrieven opent; van de 3.357 geadresseerden van de nieuwsbrief terminologie doet 39% dat.

Er zijn drie persberichten verstuurd: ‘Samenwerking INT en DPG Media, Invloed van het Engels op verschillende talen vergelijken met de Global Anglicism Database (GLAD), Koosjer & Sjoa: online woordenboek van Hebreeuwse en Jiddisje woorden in het Nederlands.

Sociale media

Alle berichten die verschijnen op ivdnt.org worden met behulp van Blog2Social direct doorgeplaatst op X (5.591 volgers), LinkedIn (3.596 volgers) en Facebook (802 vind-ik-leuks). Voor Instagram (1.522 volgers) wordt voor elke post een aparte afbeelding opgemaakt in de huisstijl. Nieuw toegevoegd in 2023 is Mastodon (435 volgers): een opensource socialemediaplatform dat vergelijkbaar is met X.

De laatste tijd verliest X aan populariteit. Ook de activiteit op Facebook loopt terug. Daarentegen groeit het aantal volgers juist op LinkedIn en Instagram. Om die reden is in 2023 besloten vooral in te zetten op deze twee platforms. Het nadeel van Instagram is dat volgers op het platform blijven en dat het medium weinig traffic genereert naar ivdnt.org of andere INT-websites. Het format voor de Instagramposts is daarom aangepast en uitgebreid met o.a. een foto (bij een post over de podcast) en een tweede beeld met extra uitleg of een definitie, om de volgers zo veel mogelijk te bedienen op het platform zelf.

Podcasts

In 2023 zijn er tien afleveringen verschenen van de podcast ‘Over taal gesproken’, die het INT samen met Onze Taal maakt. Laura van Eerten (INT) en Raymond Noë (Onze Taal) spreken elke aflevering een deskundige over het Nederlands. In totaal zijn er nu drieëntwintig afleveringen verschenen van ‘Over taal gesproken’, en de drie best beluisterde in de eerste zeven dagen na publicatie zijn ‘Wat zegt je naam over jou?’ (2.438 downloads), ‘Grammatica en gevoel’ (2.309 downloads) en ‘Hoe vangen we geur in woorden?’ (2.015 downloads). Over alle afleveringen verspreid waren er in 2023 in totaal zo’n 60.000 downloads van de podcast: een verdubbeling ten opzichte van het voorgaande jaar. 73% van de luisteraars komt uit Nederland en 13% uit België. ‘Over taal gesproken’ staat in de top 5% (1.123 downloads in de eerste zeven dagen) van alle podcasts die gehost worden op het platform Buzzsprout. ‘Over taal gesproken’ heeft eind 2023 5.310 abonnees op Spotify en 1.750 abonnees op Apple Podcasts. Een fragment is opgenomen in een educatieve uitgave voor de Vlaamse markt (Campus Nederlands) van Pelckmans Uitgevers.

Er zijn in 2023 geen nieuwe afleveringen verschenen van ‘Waar komt pindakaas vandaan?’. Toch waren er maar liefst 36.000+ downloads van de podcast. Daarnaast is een fragment van de podcast opgenomen in een lesmethode van uitgeverij KleurRijker. Voor 2024 staat een nieuw seizoen gepland.

Luisteronderzoek

Om meer te weten te komen over de luisteraars en om feedback te verzamelen verspreidden we in 2023 een enquête onder de volgers van de podcast. 59 personen vulden de enquête in. De leeftijd van de respondenten lag voor het grootste deel tussen de 30-50 jaar (32%) en de 50-70 jaar (34%), 90% is hooggeschoold en de meeste respondenten gaven aan te luisteren vanwege persoonlijke interesse. Meer dan de helft vond de podcast via Onze Taal, 20% maakte via een podcastapp kennis met 'Over taal gesproken'. Suggesties voor verbetering gingen met name over de (te aanwezige) muziek en (het gebrek aan) transcripten. Verder werden er enkele interessante onderwerpen aangedragen.

Gebruikersenquêtes

Niet alleen onder de podcastluisteraars werd een enquête verspreid, ook voor het Corpus Hedendaags Nederlands, Woordcombinaties, de e-ANS en de Historische Woordenboeken online zijn gebruikersenquêtes uitgezet in 2023. Die enquêtes zijn gemaakt in het kader van een project van de Taalunie, waarbij vier taalbronnen van het INT werden uitgelicht om de bekendheid en het gebruik van deze bronnen te vergroten. Uit de enquêtes kwam naar voren dat er veel docenten, wetenschappers en vertalers deelnamen aan de door het INT georganiseerde webinars, en weinig studenten. De meeste deelnemers vonden een webinar een geschikte manier om een introductie te krijgen in werken met een bepaalde bron, ze waren onder de indruk van de vele mogelijkheden, en sommigen zouden een meer diepgaande (live) workshop als vervolg prettig vinden.

Andere populairwetenschappelijke activiteiten

Er zijn in 2023 vier gelegenheidswoordenboekjes gepubliceerd: 'Het toilet van Couperus', 'Het hijgend hert en andere uitdrukkingen uit Het Boek der Psalmen', 'Datingwoordenboekje' en 'Oud nieuws (Couranten Corpus)'.

In samenwerking met de Taalunie zijn vier taalbronnen van het INT extra uitgelicht om het bereik en gebruik van de bronnen te vergroten. Voor het Corpus Hedendaags Nederlands, Woordcombinaties, de e-ANS en de Historische Woordenboeken zijn er o.a. kennisclips gemaakt en webinars georganiseerd.

Tijdens de Week van het Nederlands in oktober vond de aftrap plaats van de webinarreeks. Ook werd het online Juridisch Woordenboek Nederlands-Spaans II (JWSII) gepresenteerd in Leiden en publiceerden we een podcastaflevering over het thema 'emotie in taal'. Aan het einde van het jaar brachten we weer een populairwetenschappelijke publicatie uit. Dit keer het boek *Daar is geen woord Frans bij* van Nicoline van der Sijs. Tijdens de festivalmaand Letterlijk Leiden presenteerde Nicoline haar boek, gaf Roland de Bonth de stadswandeling 'Weg van woorden' en maakten we een postertentoonstelling over de nieuwe woorden van 2023.

Het INT was al verantwoordelijk voor de techniek achter de website Neerlandistiek.nl en zal in 2024 de hosting en het onderhoud op zich nemen.

Bijlage 1: Raad van Toezicht en Raad van Advies

Raad van Toezicht

Mr. Mieke Zaanen (voorzitter)

Prof. mr. Jan Cerfontaine (tot 1 mei 2023)

Drs. Erik Boels

Mr. Frank Judo

Drs. Gertine van der Vliet (tot 1 december 2023)

Raad van Advies

Prof. dr. Hans Bennis (voorzitter)

Prof. dr. Dirk Geeraerts

Prof. dr. Veronique Hoste

Prof. dr. Reinhild Vandekerckhove

Lic. Wim Vanseveren

Dr. Eric Mijts

Prof. dr. Lisa Cheng

Prof. dr. Willy Vandeweghe

Jan Jaap Knol (tot oktober 2023)

Prof. dr. Jack Hoeksema (tot oktober 2023)

Bijlage 2: Medewerkers

Prof. dr. Frieda Steurs – directeur/bestuurder

Dr. Maaike Beliën – onderzoeker/taalkundige

Marjolijn van Bennekom – taalkundig assistent

Dr. Bob Boelhouwer – computerlinguïst

Dr. Roland de Bonth – onderzoeker/taalkundige

Tim Brouwer – softwareontwikkelaar

Lic. Lut Colman – onderzoeker/taalkundige

Lic. Griet Depoorter – onderzoeker/taalkundige

Lic. Katrien Depuydt – onderzoeker/taalkundige

Lic. Veronique De Tier – onderzoeker/taalkundige

Dr. Jesse de Does – computerlinguïst

Job van Doeselaar – webmaster

Laura van Eerten MA – communicatieadviseur

Drs. Mathieu Fannee – systeemontwikkelaar

Drs. Thomas Haga – taalkundig assistent

Dr. Kris Heylen – onderzoeker/taalkundige

Ruud de Jong – systeemontwikkelaar

Dr. Dirk Kinable – onderzoeker/taalkundige

Uma Kraus - managementassistent

Drs. Marco van der Laan – systeembeheerder

Dr. Frank Landsbergen – computerlinguïst

Koen Mertens – systeemontwikkelaar

Drs. Jan Niestadt – systeemontwikkelaar

Vincent Prins - systeemontwikkelaar

Dr. Kathy Rys – onderzoeker/taalkundige

Prof. dr. Nicoline van der Sijs – onderzoeker/taalkundige

Paulette Tacx – managementassistent (pensioen per november 2023)

Dr. Carole Tiberius – computerlinguïst

Dr. Vincent Vandeghinste – computerlinguïst en CLARIN-coördinator

Lic. Katrien Van pellicom – onderzoeker/taalkundige

Drs. Boukje Verheij – onderzoeker/taalkundige

Rayvano van Vliet – systeembeheerder

Machteld de Vos, MPhil – onderzoeker in opleiding

Drs. Vivien Waszink – onderzoeker/taalkundige

Karin van Weerlee – managementassistent

Laetitia de Winter – administrateur

Bijlage 3: publicaties, lezingen, media, prijzen etc.

Publicaties

Andree, M., Waszink, V & Eerten van, L. van (2023). *Datingwoordenboekje.* Op: ivdnt.org. link

Bonth, R. de (2023). 'De Verdam-krijgt-een-kastquiz'. Op: *Neerlandistiek* (19 december). link

Bonth, R. de (2023). 'Smakelijk ette!' Op: *Neerlandistiek* (5 oktober). link

Bonth, R. de (2023). *Het toilet van Couperus* [gelegenheidswoordenboekje vanwege de 100e sterfdag van Louis Couperus]. Op: ivdnt.org. link

Bonth, R. de (2023). *Oud nieuws* [gelegenheidswoordenboekje omdat de helft van de 17e-eeuwse Amsterdamse Courant is getranscribeerd en gecontroleerd]. Op: ivdnt.org. link

Bonth, R. de (2023). 'Achtergrondzangerettes'. Op: *Neerlandistiek* (5 januari). link

Bonth, R. de (2023). 'Alleen in -ette wil ik wonen'. Op: *Neerlandistiek* (6 juli). link

Bonth, R. de (2023). 'Bij de les'. Op: *Neerlandistiek* (20 september). link

Bonth, R. de (2023). 'Bijzondere woorden uit 17e-eeuwse couranten'. Op: *Neerlandistiek* (11 september). link

Bonth, R. de (2023). 'Blij dat GLAD er is'. Op: *Neerlandistiek* (18 oktober). link

Bonth, R. de (2023). 'Bungalette'. Op: *Neerlandistiek* (2 maart). link

Bonth, R. de (2023). 'De eerlijke vinderette'. Op: *Neerlandistiek* (6 april). link

Bonth, R. de (2023). 'De geschiedenis van het woord [vul in]'. Op: *Neerlandistiek* (17 juni). link

Bonth, R. de (2023). 'De staart is eraf: -ette wordt -et'. Op: *Neerlandistiek* (3 augustus). link

Bonth, R. de (2023). 'De voortreffelicke Ceneton, ofte 't gheluckigh tooneelspeelen; een zomerquiz'. Op: *Neerlandistiek* (18 juli). link

Bonth, R. de (2023). 'Docent of leerling? Over forensische taalkunde'. Op: *Neerlandistiek* (17 januari). link

Bonth, R. de (2023). 'Geef acht Dibbetz' *Groot Militair Woordenboek*!' In Sterkenburg, P. van, Bonth, R. de & Heylen, K., *In termen van taal. Liber amicorum Frieda Steurs*. Scriptum, pp. 48-56.

Bonth, R. de (2023). 'Het nette komt door -ette'. Op: *Neerlandistiek* (2 november). link

Bonth, R. de (2023). 'Het toilet van Couperus'. Op: *Neerlandistiek* (20 maart). link

Bonth, R. de (2023). 'Literaire ontwikkeling toetsen'. Op: *Neerlandistiek* (16 maart). link

Bonth, R. de (2023). 'Mannelijk, vrouwelijk en onzijdig. Pieter Langendijk schaaft aan *Het wederzyds huwelyksbedrog*'. In: *Accolade* XLI (december 2023), pp. 130-145.

Bonth, R. de (2023). 'Op je voeten -lette'. Op: *Neerlandistiek* (2 februari). link

Bonth, R. de (2023). 'Openbare toilETTEn'. Op: *Neerlandistiek* (7 december). link

Bonth, R. de (2023). 'Silhouette'. Op: *Neerlandistiek* (5 januari). link

Bonth, R. de (2023). ‘Slaap zacht met -ette!’ Op: *Neerlandistiek* (7 september). link

Bonth, R. de (2023). ‘Taalvrouw’. Op: *Neerlandistiek* (19 oktober). link

Bonth, R. de (2023). ‘Vragen over ’t aapenland’. Op: *Neerlandistiek* (15 februari). link

Colman, L. & Tiberius, C. (2023). Meerwoordexpressies in Woordcombinaties: de gordiaanse knoop. In: Sterkenburg, P. van, Bonth, R. de & Heylen K. (red.), *In termen van taal*. Liber amicorum Frieda Steurs. Scriptum, pp. 69-81.

Cornips, L., Craenenbroeck, J. van, Sijs, N. van der & Swanenberg, J. (2023). ‘Wat riekt het hier zwallig! Geurwoorden in het Nederlands en zijn dialecten’, in: Leemans, I. & Verbeek, C. (red.), *Neuswijzer. Geuratlas van de Lage Landen*. Amsterdam, pp. 163-173.

De Sisto, M. et al (among others Vandeghinste, V.) (2023). GoSt-ParC-Sign: Gold Standard Parallel Corpus of Sign and spoken language. *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. pp. 501–502, Tampere, Finland, June 2023. pp. 503-504. link

De Tier, V. (2023). Lukken en strienen tussen Kerstdag en dertienavond. In: *Diejalektgazette Bachtn de Kuupe*, jg. 18, nr. 1, pp. 7-9.

De Tier, V. (2023). 'De Zeeuwse dialectwoordenschat gaat mee met zijn tijd'. In: *Nehalennia*, 219, voorjaar 2023, pp. 17-19.

De Tier, V. (2023). 'Zeeuwse dialectwoorden op het internet'. In: *Zeeuws Erfgoed*, jg. 22, 1, pp. 26-27. link

De Tier, V. (2023). 'Zilt en zoet in de Zeeuwse dialecten'. In: *Zeeuws Erfgoed*, jg. 22, 3, pp. 14-15. link

De Tier, V. (2023). ''t Is kerremesse in d’helle: Zeeuwse woorden voor regen'. In: *Nehalennia*, 220, najaar 2023, pp. 13-14.

De Tier, V. (2023). ‘Gezocht voor nieuwgeborene: lokkedijzen, kindjeskak en ander klein spul’. In: *Diejalektgazette Bachtn de Kuupe*, jg. 19, nr. 1, pp. 7-9.

De Tier, V. (2023). ‘Het Instituut voor de Nederlandse Taal leert nu ook dialect’, in: Sterkenburg, P. van, Bonth, R. de & Heylen, K. (red.), *In termen van taal.* Liber amicorum Frieda Steurs. [Schiedam], Scriptum, 2023, pp. 317-323.

De Tier, V. (2023). ‘Nieuwloopte bendige leevaards contra hoofdige astrante boffers’. In: *Diejalektgazette Bachtn de Kuupe*, jg. 18, nr. 2, pp. 7-9.

De Tier, V. (2023). ‘Pas mar op dat joen buuk nie e versjchoepert’. In: *Diejalektgazette Bachtn de Kuupe*, jg. 18, nr. 3, pp. 7-9.

De Tier, V. (2023). ‘Spin op de kaart. Zelf kaarten maken met de DSDD’. Op: *Neerlandistiek* (30 september). link

Depuydt, Katrien & Jesse de Does (2023). *Heden en verleden in een infrastructuur van medische termen - een verkenning.* In: Piet van Sterkenburg, Roland de Bonth en Kris Heylen (red.), *In termen van taal. Liber Amicorum Frieda Steurs.* [z.p.]

Eerten, L. van & R. Noë (2023). 'Taal tijdens het strijken. De beste taalpodcasts op een rijtje'. In *Onze Taal*, 6 - 2023, p. 24-27. link

Eerten, L. van (2023). ‘Waar gaat Onze Kees naartoe?’. In: *Onze Kees*, p. 22.

Erjavec, Tomaž, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darǵis, Orsolya Ring, Ruben van Heusden, Maarten Marx & Darja Fišer (2023). The ParlaMint corpora of parliamentary proceedings. *Lang Resources & Evaluation* 57, 415–448.

Fannee, M. (2023). *Atlas van Laatmiddeleeuws Warmond – Reconstructie van het dorp en zijn omgeving (1250-1574)*. Historisch Genootschap Warmelda, Warmond. (tweede verbeterde en uitgebreide druk, 468 pp).

Fannee, M. (2023). *Warmonds toponymisch woordenboek – Middeleeuwen & Vroegmoderne tijd*. Historisch Genootschap Warmelda, Warmond. (tweede verbeterde en uitgebreide druk, 264 pp).

Fannee, M. (2023). “De oudste schout van Warmond”, in: *De Hekkensluiter*, Historisch Genootschap Warmelda, jaargang 20, nr. 2.

Geeraerts, D. et al. (onder wie Heylen, K.) (red.) (2023). *Lexical Variation and Change. A Distributional Semantic Approach*. Oxford, England: Oxford University Press.

Heuven, V. van & Sijs, N. van der (2023). ‘Zeg eens “kaas”’, in *Kaas = NL? Melk, koe, ras, kolonie, taal, kunst, mest en meer*, red. L. Cornips, M. Hendriksen en G. Mak, pp. 126-138.

Heylen, K. (2023). 'Een levensloop van woorden. Een taalinfrastructurele behoefteanalyse voor de diachrone lexicologie van het Nederlands. In: Sterkenburg, P. van, Bonth, R. de & Heylen, K. (red.) *In termen van taal*. Schiedam: Scriptum Uitgeverij, 168-177.

Huyssteen, G.B. van & Tiberius, C. (2023). Towards a lexical database of Dutch taboo language. In: Medveď, M. et al. (eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*. Brno, 27–29 June 2023. Brno: Lexical Computing CZ s.r.o. 53-74. link

Kaltenböck et al. (among others Vandeghinste, V.) (2023). Deep Dive Data and Knowledge. In: Rehm, G., Way, A. (eds) *European Language Equality. Cognitive Technologies*. Springer, Cham. link

Kinable, D. (2023). 'Lexicografische definities en hun terminologische raakvlak volgens ISO 704'. In: Sterkenburg, P. van, Bonth, R. de & Heylen, K. (red.), *In termen van taal*. Schiedam: Scriptum Uitgeverij, 178-187.

Konontsjoek, O. & Sijs, N. van der (2023).), ‘De stelselmatige onderdrukking van het Oekraïens’, in: *Onze Taal* 3, 50-52.

Kruyt, T. & Bonth, R. de (2023). ‘3. Wörter und ihre Bedeutung’. In: Boonen, U.K. & Harmes, I., *Niederländische Sprachwissenschaft. Eine Einführung*. 2., vollständig überarbeitete Auflage. Tübingen: Narr Francke Attempto Verlag, 2023, pp. 69-95.

Landsbergen, F. (2023). ‘Richting het emeritaat’. In: Sterkenburg, P. van, Bonth, R. de & Heylen, K. (red.), *In termen van taal. Liber amicorum Frieda Steurs*. [Schiedam], Scriptum, 2023, pp. 204-213.

Medveď, M. et al (onder anderen Tiberius, C.) (eds.) (2023). *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*. Brno, 27–29 June 2023. Brno: Lexical Computing CZ s.r.o.

Montes, M. & Heylen, K. (2023). 'Parameters and Procedures for Token-Based Distributional Semantics'. In: Geeraerts et al. (red.), *Lexical Variation and Change. A Distributional Semantic Approach*. Oxford, England: Oxford University Press, 65-87.

Norré, M., Cardon, R., Vandeghinste, V. & François, T. (2023). Word Sense Disambiguation for Automatic Translation of Medical Dialogues into Pictographs. In *Recent Advances in Natural Language Processing.* link

Schuurman, I. et al. (onder wie Vandeghinste, V.) (2023). Are there just WordNets or also SignNets? *Proceedings of the 13th Global WordNet Conference*. Donastia, Spain. link

Sijs, N. van der (2023).), 'Etymologica: Starnakel en straalbezopen', in: *Neerlandistiek.* link

Sijs, N. van der (2023). *Daar is geen woord Frans bij. Het beeld van vreemde talen in Nederlandse uitdrukkingen*. Scriptum.

Sijs, N. van der (2023). Over de regenboogbrug', in: *Onze Taal* 2, 34.

Sijs, N. van der (2023). Sijs, Nicoline van der (2023), 'Twee zielen, één taal: leve de latrelatie!', in: *Het grote misverstaan. Nadenken over de hereniging van de Nederlanden*, red. Karel Luyckx, Sterck & De Vreese, pp. 118-125.

Sijs, N. van der (2023). Vijfhonderd jaar Nederlandse Wat & Hoe-gidsen. Op: ivdnt.org. link

Sijs, N. van der (2023). 'Charivari', in: *Onze Taal* 1, 34.

Sijs, N. van der (2023). 'De overlevingsgraad van Engelse leenwoorden', in: Sterkenburg, P., Bonth, R. de & Heylen, K. *In termen van taal. Liber amicorum Frieda Steurs*, pp. 286-296.

Sijs, N. van der (2023). 'Dé l'amour libre et des âmes délicates', in *Septentrion* 8, 153-157.

Sijs, N. van der (2023). 'Een korte cultuurgeschiedenis van Nederlandse kaasnamen', in Kaas = NL? Melk, koe, ras, kolonie, taal, kunst, mest en meer, red. L. Cornips, M. Hendriksen en G. Mak, pp. 109-118.

Sijs, N. van der (2023). 'Een wondermiddel. Sint-janskruid (Hypericum perforatum)', in: *Flora Batava 1800-1934 De wilde planten van Nederland*, red. Esther van Gelder & Norbert Peeters, Lannoo, p. 302.

Sijs, N. van der (2023). 'Een 'boom' aan Engelse leenwoorden', in: *Onze Taal* 1, 20-23. link

Sijs, N. van der (2023). 'Etymologica: Biks', in: *Neerlandistiek*. link

Sijs, N. van der (2023). 'Etymologica: Gesnopen!', in: *Neerlandistiek*. link

Sijs, N. van der (2023). 'Etymologica: Groente', in: *Neerlandistiek*. link

Sijs, N. van der (2023). 'Etymologica: Kokkels', in: *Neerlandistiek*. link

Sijs, N. van der (2023). 'Etymologica: Korenwolven en hamsters', in: *Neerlandistiek*. link

Sijs, N. van der (2023). 'Etymologica: Korenwolven en hamsters', in: *Neerlandistiek*, 3-9-2023.

Sijs, N. van der (2023). ‘Etymologica: Mandelig, een Fries-Saksische rechtsterm’, in: *Neerlandistiek*. link

Sijs, N. van der (2023). ‘Etymologica: Recordveel’, in: *Neerlandistiek*. link

Sijs, N. van der (2023). ‘Etymologica: Starnakel en straalbezopen’, in: *Neerlandistiek*. link

Sijs, N. van der (2023). ‘Etymologica: Statiegeld’, in: *Neerlandistiek*. link

Sijs, N. van der (2023). ‘Etymologica: Vuurwerk te land, te water en in de lucht’, in: *Neerlandistiek*. link

Sijs, N. van der (2023). ‘Geuren en verstanken. Hoe rijk is onze historische geurtaal?’, in: Leemans, I. & Verbeek, C. (red.), *Neuswijzer. Geuratlas van de Lage Landen*, Amsterdam, pp. 137-146.

Sijs, N. van der (2023). ‘Glamour’, in: *Onze Taal* 3, 54.

Sijs, N. van der (2023). ‘Het hijgend hert en andere uitdrukkingen uit Het Boek der Psalmen (1773)’, een gelegenheidswoordenboekje. link

Sijs, N. van der (2023). ‘Jo en Joshua. Een voetnoot bij de geschiedenis van de taalkunde’, in: *Neerlandistiek*. link

Sijs, N. van der (2023). ‘Le néerlandais standard est moins influencé par le brabançon et le flamand qu’on ne le croit’, in: *Les Plats Pays*. link

Sijs, N. van der (2023). ‘Maak je geen zorgen over het Nederlands’, in *De Telegraaf* 15 februari 2023, p. 24.

Sijs, N. van der (2023). ‘Nondeju’, in: Jansen, J. & Jansen, D., *Als we wisten wat we deden, heette het geen onderzoek*. Amsterdam, pp. 69-72.

Sijs, N. van der (2023). ‘Oorlogstaal’, in: *Uitzwaaiboek voor Johan*, pp. 63-67.

Sijs, N. van der (2023). ‘Op het einde komt het altijd goed’, in: *Onze Taal* 4, 40.

Sijs, N. van der (2023). ‘Op het einde komt het altijd goed’, in: *Onze Taal* 4, 40.

Sijs, N. van der (2023). ‘Over hijgende herten en de wondere wereld der naamvallen’, in: *Neerlandistiek*. link

Sijs, N. van der (2023). ‘Quand révolution sexuelle rime avec révolution lexicale’, in: *Les Plats Pays*. link

Sijs, N. van der (2023). ‘Reuring voor Kees’, in: *Onze Kees*, p. 55.

Sijs, N. van der (2023). ‘Rieken en ruiken’, in: Leemans, I. & Verbeek, C. (red.), *Neuswijzer. Geuratlas van de Lage Landen*. Amsterdam, pp. 1537-155.

Sijs, N. van der (2023). ‘Stank voor dank. Geurgezegden’, in: Leemans, I. & Verbeek, C. (red.), *Neuswijzer. Geuratlas van de Lage Landen*. Amsterdam, pp. 175-187.

Sijs, N. van der (2023). ‘Stereotype’, in: *Onze Taal* 6, 56.

Sijs, N. van der (2023). ‘Stijlbloempjes. Egelantier (Rosa rubiginosa)’, in: *Flora Batava 1800-1934 De wilde planten van Nederland*, red. Esther van Gelder & Norbert Peeters, Lannoo, p. 414.

Sijs, N. van der (2023). ‘Van k*twoorden en kl*tetaal’, in: *De Lage Landen* 66, november 2023, 43-47.

Sijs, N. van der (2023). ‘Waar komt die regel toch vandaan?’, in: *Onze Taal* 3, 35.

Sijs, N. van der (2023). ‘Zoetgevooisd’, in: *Onze Taal* 5, 36.

Sterkenburg, P. van, Bonth, R. de & Heylen, K. (red.) (2023). *In termen van taal*. Schiedam, Nederland: Scriptum Uitgeverij.

Steurs, F. (2023). On the (un)translatability of legal texts: can technology help in harmonizing inequivalences in different legal concepts and terms? In: Palumbo, G., Peruzzo, K. & Pontrandolfo, G. (eds.) *What’s Special about Specialised Translation? Essays in honour of Federica Scarpa*. Peter Lang, Bern, pp.201-217.

Steurs, F. & Lewandowska-Tomaszczyk, B. (2023). Collaboration and Crowdsourcing Applications. In: Lewandowska-Tomaszczyk, B. & Trojszczak, M. (eds.), *Language in Educational and Cultural Perspectives*. Springer Verlag, 119-131.

Steurs, F., Vandeghinste, V. & Daelemans, W. (2023). Language Report Dutch. In: Rehm, G., Way, A. (eds) *European Language Equality. Cognitive Technologies*. Springer, Cham. pp. 123-126. link

Tempelaars, R. (2023). ‘Een revolutionaire verandering in de Nederlandse woordenboeken. Een kanttekening bij een (vermeende) mijlpaal’, in: Piet van Sterkenburg, Roland de Bonth en Kris Heylen (red.), *In termen van taal*. Liber amicorum Frieda Steurs. [Schiedam], Scriptum, 2023, pp. 317-322.

Tempelaars, R. (2023). ‘Hallo, hallo, wie meurt daar zo?’, in: Inger Leemans en Caro Verbeek (red.), *NeusWijzer. Geuratlas van de Lage Landen*. Amsterdam, Boom, 2023, p. 169.

Tempelaars, R. (2023). ‘Stinken als een bunzing. Reukvergelijkingen’, in: Inger Leemans en Caro Verbeek (red.), *NeusWijzer. Geuratlas van de Lage Landen*. Amsterdam, Boom, 2023, pp. 141-143.

Tempelaars, R. & Smelik, K.A.D. (2023). ‘Etty Hillesum schijft een brief op het strand van Knokke’, in: Klaas A.D. Smelik (eindred.), *Etty Hillesum over God en Lot*. Cahiers Etty Hillesum, deel 3. [Antwerpen/’s-Hertogenbosch], Gompel & Svacina, [2023], pp. 153-161.

Tempelaars, R. en Leemans, I. (2023). ''Geheime geurwapens', in: Leemans, I. & Verbeek, C. (red.), *NeusWijzer. Geuratlas van de Lage Landen*. Amsterdam, Boom, 2023, pp. 251-253.

Tiberius, C. (2023). Country Profile The Netherlands. In: Marra, E. et al., *AI for a Multilingual Europe: Why language data matters*. ELRC White Paper.172-176. link

Tiberius, C. et al. (2023). A Lexicographic Practice Map of Europe. In: *International Journal of Lexicography* link

Tiberius, C. et al. (2023). An insight into Lexicographic Practices in Europe. Results of the Extended ELEXIS Survey on User Needs. In: Klosa-Kückelhaus, A. et al. (eds), *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*. Mannheim: IDS-Verlag, 509-521. link

Vandeghinste, V, & Guhr, O. (2023). FullStop: Punctuation and Segmentation Prediction for Dutch with Transformers. *Language Resources and Evalution*. Springer. link

Vandeghinste, V. & Guhr, O. (2023). FullStop: Punctuation and Segmentation Prediction for Dutch with Transformers. Preprint op ArXiv link

Vandeghinste, V. et al. (2023). A Spoken Academic Belgian Dutch Corpus. In: *Proceedings of the CLARIN Annual Conference*. Leuven. video link

Vandeghinste, V. et al. (2023). SignON: Sign Language Translation. Progress and challenges. *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. pp. 501–502, Tampere, Finland, June 2023. link

Waszink, Vivien (2023). 'Over wokewappies, wokewashing en wokisme' (2023). In: Piet van Sterkenburg, Roland de Bonth, Kris Heylen. *In termen van taal. Liber amicorum Frieda Steurs*, pp. 366-373. Uitgeverij Scriptum.

Waszink, Vivien (2023). 'De houdbaarheid van geurwenswoorden. Hardnekkig of snel vervliegend?'. In: *Neuswijzer, Geuratlas van de Lage Landen*. Boom Amsterdam.

Wijnands, A. & Sijs, N. van der (2023). 'Taalverandering. Hoe maak je leerlingen bewust van hedendaagse en historische taalveranderingen?', in: Vaardig met vakinhoud. Handboek vakdidactiek Nederlands, red. Jeroen Dera, Joyce Gubbels, Janneke van der Loo en Jimmy van Rijt, Bussum: Coutinho, 233-245.

Zingano Kuhn, T. et al. (onder anderen Tiberius, C.) (2023). Data preparation in crowdsourcing for pedagogical purposes: the case of the CrowLL game. Slovenščina 2.0, 10(2): 62-100. link

Lezingen en presentaties

Candela, Gustavo; Cuper, Mirjam; Vodopivec, Ines; Depuydt, Katrien; Romein, Christel Annemieke; Chambers, Sally; de Does, Jesse; Martínez-Sempere, Isabel; Parkoła, Tomasz; Antonacopoulos, Apostolos, *Sharing and Sustaining Digitisation Knowledge: a White Paper on written cultural heritage digitisation.* DARIAH Annual Event 2023.

Colman, Lut. *Word Combinations*. Demo op eLex 2023, Brno, 27-29 juni 2023.

Colman, Lut*. Woordcombinaties*. Webinar i.s.m. Nederlandse Taalunie, 12 oktober 2023.

Depuydt, K., V. De Tier, V. & C. Krottje: Demonstratie Bildts Platform. Streektaalconferentie 22 september 2023, Franeker.

Depuydt, Katrien, Nicoline van der Sijs, Jesse de Does, Ruud de Jong, Roland de Bonth, Mathieu Fannee, Annemieke Romein, Joris van Zundert, *Content providers, Researchers, Technology and the Crowd: Discovering the Best Possible Collaborative Strategies for Datafication and Publication of a Dutch Historical Newspaper Corpus.* Poster gepresenteerd door Ruud de Jong op DH 2023, 10-14 juli, Graz.

Depuydt, Katrien, Nicoline van der Sijs, Jesse de Does, Ruud de Jong, Roland de Bonth, Mathieu Fannee, Annemieke Romein, Joris van Zundert, *Digitising a Dutch Historical Newspaper Corpus. Discovering the Best Possible Strategies for Datafication and Publication.* Clariah Annual Conference, 30 november 2023, Utrecht.

Depuydt, Katrien, Jesse de Does, *Towards an infrastructure for the semi-automatic development of corpus-based language exercises.* Japan Association for English Corpus Studies (JAECS) spring workshop, 13 mei 2023.

Depuydt, Katrien. *Corpus Hedendaags Nederlands.* Webinar in de serie Werken met online taalbronnen. 5 oktober 2023.

De Tier, V. Does, J. de, Depuydt, K, Mertens K. e.a. The Southern Dutch dialects in the digital age: from concept to realization. Xth Congress of the International society for dialectology and geolinguistics. 4th-8th September 2023, Bucharest, Romania.

De Tier, V. & G. Dehaes. Van tweetalige straatnaamborden tot dialectadvertenties. Dialect in het straatbeeld in Vlaanderen en Brussel. Streektaalconferentie 22 september 2023, Franeker.

De Tier, V. (2023) ‘De Taalsector presenteert: de dialectoloog’ online meeting 20 januari.

De Tier, V. (2023) ‘Schelden in dialect’, lezing voor de Orde van den Prince, afd. Bergen-op-Zoom, 3 maart.

De Tier, V. (2023) ‘De Oost-Vlaamse dialecten in de DSDD’, Melsen, 12 maart.

De Tier, V. (2023) ‘De Zeeuwse Woordenbank’, Biervliet 23 maart.

De Tier, V. (2023) ‘Streektaal en meertaligheid’, Doetinchem, 29 maart (symposium).

De Tier, V. (2023), Ontstaan en afbakening van de dialecten in Nederland en Vlaanderen, Sassenheim, Rijnlandse geschiedenis, 20 april.

De Tier, V. (2023) ‘Kedjoarkies veur de virkies. De geschiedenis van de dialecten in de streek van Oudenaarde. De Haagschool 28 april.

De Tier, V. (2023) ‘Zeeuwse woordenschat op de kaart’, Biervliet, 27 juni.

De Tier, V. (2023) ‘Online dialectbronnen’, De Taalsector, 11 september.

De Tier, V. (2023) ‘Cursus spreekwoorden en zegswijzen in dialect’, Koksijde, 2 en 9 oktober.

De Tier, V. (2023) ‘Op de schouders van onze voorgangers. Het woordenboek der Zeeuwse Dialecten op de kaart.’, Zeeuwse dialectdag, Kapelle, 21 oktober.

De Tier, V. (2023) ‘Familienamen en dialect’, Hulst 7 en 14 november.

De Tier, V. (2023) ‘Op de grens van twee dialectgebieden’, Burst, 9 november.

De Tier, V. (2023) ‘Gents of boers’, Markant Mariakerke-Wondelgem, 13 november.

De Tier, V. (2023) ‘Op de schouders van onze voorgangers. Het woordenboek der Zeeuwse Dialecten op de kaart.’, Biervliet 23 november.

De Tier, V. (2023) ‘Peetie doe tekkie toe’ Dialect in Heurne en omgeving’, Heurne, 13 december.

Heylen, Kris, Ilan Kernerman en Carole Tiberius. *Linking CEFR-based learner profiles to lexicographic data*. Presentatie op Workshop on Profiling second language vocabulary and grammar. Göteborg, 20 Apil 2023.

Heylen, Kris, Ilan Kernerman, en Carole Tiberius. 2023. *Linking lexicographic and CEFR resources*. Presentatie op Japan Association for English Corpus Studies (JAECS) Spring Forum, online, 13 May 2023.

Heylen, Kris, Ilan Kernerman, Jelena Kallas, en Carole Tiberius*. An infrastructure for lexicography and CEFR*. Presentatie op ELEX-workshop *Lexicography and CEFR: Linking lexicographic resources and language proficiency levels,* Brno, 29 juni 2023.

Heylen, Kris, Ilan Kernerman, Jelena Kallas, Carole Tiberius. *Linking Lexicographic and Language Learning Resources*. Presentatie op LDK 2023-workshop *Linking Lexicographic and Language Learning Resources (4LR),* Wenen, 13 september 2023.

Heylen, Kris, Ruud de Jong en Jesse de Does. *Towards an integrated web application for terminology extraction and termbase editing.* Posterpresentatie op The 33rd Meeting of Computational Linguistics in The Netherlands (CLIN 33), Antwerpen, 22 september, 2023.

Heylen, Kris en Carole Tiberius. Presentatie op de SAILS Workshop: AI and LLMs: Keeping the Linguist in the Loop, 8 december 2023.

Klosa Kückelhaus, Annette & Carole Tiberius. *The lexicographic process revisited.* Presentatie op eLex 2023. Brno. 27 juni 2023.

Kumar Thirukokaranam Chandrasekar, K., Chambers, S. De Tier, V. Does, J. de, Depuydt, K. When AI and Dialect Data meet: crossing-borders between dialectology and data science: an exploration for the Southern Dutch Dialects. DH Benelux, Brussel 31 mei-2 juni 2023.

Landsbergen, Frank. De Algemene Nederlandse Spraakkunst in de 21e eeuw. Presentatie op de Neerlandistiekdagen. Utrecht, 14 april 2023.

Landsbergen, Frank. Algemene Nederlandse Spraakkunst (e-ANS). Webinar in de serie Werken met online taalbronnen. 7 november 2023.

Vandeghinste, Vincent. 2023. Challenges with Sign Language Datasets. Keynote at AT4SSL 2023. Tampere, Finland.

Vandeghinste, Vincent. 2023. Challenges in Sign Language Translation. CENTAL Seminars. UCLouvain. Louvain-la-Neuve.

Vandeghinste, Vincent. 2023. Artificial Intelligence and NLP: Opportunities and Pitfalls, for the course Tomorrow in Psychological Science. Seminar for Masters of Theoretical Psychology. KU Leuven.

Vandeghinste, Vincent. 2023. SAILS workshop, Leiden: Invited panel member.

Vandeghinste, Vincent. 2023. What can we expect from Large Language Models? TEKOM.BE 2023, Antwerpen.

Charlotte Van de Velde, Bram Vanroy & Vincent Vandeghinste. 2023 Automatic sentence-level simplification for Dutch. CLIN 2023. Antwerpen.

Theresa Seidl, Vincent Vandeghinste & Tim Van de Cruys. 2023. Controllable Sentence Simplification in Dutch. CLIN 2023. Antwerpen.

Ineke Schuurman, Bram Vanroy, Vincent Vandeghinste, Caro Brosens, Margot Janssens, Thierry Declerck & Sam Bigeard. 2023. ODWN, OMW: issues when dealing with spoken languages, but especially also with sign languages. CLIN 2023. Antwerpen

Sijs, Nicoline van der (2023), 'Hoeveel dialecten kende het negentiende-eeuwse Amsterdams?', Orde van den Prince afdelingen Nijmegen, 10 mei 2023; en 25 april afd Leiden en 5-10-2022 afd Utrecht.

Sijs, Nicoline van der (2023), 'Historische taalkunde en medische terminologie', cursus bij Trefpunt Medische Geschiedenis Nederland, Urk, 17 maart 2023.

Sijs, Nicoline van der (2023), 'Nederlandse leenwoorden en uitleenwoorden', gastcollege in de collegereeks Nederlands in contact voor bachelorstudenten Taal- en Letterkunde te Brussel, 13 maart 2023.

Van Huyssteen, Gerhard B. & Carole Tiberius. *Towards a lexical database of Dutch taboo language.* Presentatie op eLex 2023. Brno. 28 juni 2023.

Waszink, Vivien. *Dat mag je óók (al niet meer) zeggen*. Lezing over inclusief taalgebruik op diverse plaatsen, o.a. Universiteit Leiden en deBuren Brussel.

Congressen en workshops

Colman, Lut. Deelname eLex 2023, Brno, 27-29 juni 2023.

Colman, Lut. Deelname *Workshop on lexicography and CEFR*, Brno, 29 juni 2023.

Colman, Lut. Deelname SAILS Workshop: *AI and LLMs: Keeping the Linguist in the Loop*, Leiden, 8 december 2023.

De Tier, Veronique. Organisatie Streektaalconferentie 22 september 2023: 'Streektaal in zicht: Zichtbaarheid van streektalen in de openbare ruimte' in Franeker: Stichting Nederlandse Dialecten i.s.m. de Nederlandse Taalunie, provincie Fryslân en gemeente Waadhoeke.

Tiberius, Carole, Ilan Kernerman, Jelena Kallas, Kris Heylen. Organisatie van de workshop *Lexicography and CEFR: Linking lexicographic resources and language proficiency levels,* op het congres ELEX, Brno, 29 juni 2023.

Heylen, Kris, Jelena Kallas, Ilan Kernerman, Carole Tiberius. Organisatie van de workshop Linking Lexicographic and Language Learning Resources (4LR), op het congres LDK 2023, Wenen, 13 september 2023.

Onderwijs

Colman, Lut. Presentatie over Woordcombinaties en SKEMA voor studenten KU Leuven, 17 april 2023.

Heylen, Kris & Vincent Vandeghinste. Collegereeks Computationele lexicografie (B-KUL-F0SV1A) binnen de Master Taalkunde aan de KU Leuven: februari-mei 2023.

Tiberius, Carole. MA-collegereeks Corpus Lexicography. Universiteit Leiden, februari-mei 2023.

Tiberius, Carole. MA-collegereeks Computational Corpus Analysis. Universiteit Leiden, april-mei 2023.

Waszink, Vivien, samen met Jaap de Jong en Peter Burger. Collegereeks MA-scriptie inclusieve taal, MA Nederlandse Taal en Cultuur, december 2022-april 2023.

In de media

Eerten, Laura van. Over o.a. de woorden van het jaar en de podcasts *Over taal gesproken* en *Waar komt pindakaas vandaan?* in het programma *De nacht van…* NPO Radio 1. 18 december 2023.

Eerten, Laura van. Over 'waterige compromissen' en de podcast *Over taal gesproken* in het artikel 'De VVD wil een nieuw hoofdstuk maar gebruikt oude woorden', *NRC*, 10 november 2023.

Sijs, Nicoline van der. De Taalstaatmeester 2023 Linda Nooitmeer, als lid panel. 30 december 2023.

Sijs, Nicoline van der. Blow the cobwebs out of your brain with 'uitwaaien', The World. 8 december 2023.

Sijs, Nicoline van der. 'Zoveel uitdrukkingen uit andere talen, je zou het er Spaans benauwd van krijgen. RTL Nieuws, 20 november 2023.

Sijs, Nicoline van der. ‘Invloed van de Nederlandse taal op Noord-Amerika’, NPO Radio 1. 23 augustus 2023.

Sijs, Nicoline van der.‘Feit of Fictie: Gebruiken we meer Engelse leenwoorden dan Franse leenwoorden?’, NPO Radio 1. 11 augustus 2023.

Sijs, Nicoline van der. ‘Het Nederlands kan gemakkelijker’, in Villa VdB, NPO Radio 1. 22 mei 2023.

Sijs, Nicoline van der. Waarom zijn er zoveel gezegdes over een klok? en Tierelier: Podcast *Alledaagse vragen*. 2 mei 2023.

Vandeghinste, Vincent (2023). Interview in AIAIAI Podcast. Van auteur naar AI-teur?

Vandeghinste, Vincent (2023). Interview met onder meer Vincent. ChatGPT: een hallucinerende babbelaar. Over taal gesproken.

Waszink, Vivien. 'Diergezondheidsfonds' meest verhullende woord uit dierenindustrie’. Interview over eufemismen voor Editie.nl en RTL Nieuws, 10 januari 2023..

Waszink, Vivien (2023). Interview in *Nederlands Dagblad* over ontlezing jongeren, januari 2023.

Waszink, Vivien (2023). ‘Doubleren in plaats van zittenblijven. Taalwetenschappers geven het weinig kans’. Interview in *de Volkskrant*, 12 april 2023

Waszink, Vivien. ‘Het woord ‘zittenblijven’ is negatief, dus gebruikt het LAKS liever ‘doubleren’. Maar taaladviezen zijn geen taalgeboden’. Interview in *NRC*, 12 april 2023.

Waszink, Vivien. ‘Online in het nieuwe normaal. Internet voegt woorden toe aan onze taal’. Interview over onlinetaal voor Kennislink, 14 juni 2023.

Waszink, Vivien. ‘Esma en lawa, jongeren gebruiken afkortingen: 'Willen eigen taal''. Interview over afkortingen voor Editie.nl en RTL Nieuws, 4 juli 2023.

Waszink, Vivien. ‘Hiphop is geen bedreiging voor, maar een verrijking van de taal’. Interview over 50 jaar hiphop in: *De Standaard*, 12 augustus 2023.

Waszink, Vivien. ‘Trauma’s, triggers en toxische sfeer: therapeutentaal is in ons dagelijkse gesprek beland. Wat vinden experts daarvan?’. Interview over therapietaal in *de Volkskrant*, 5 september 2023.

Waszink, Vivien. ‘T.z.t, brb en VrijMiBo ken je misschien wel: maar jeugd voert hele gesprekken in afkortingen’. Interview over afkortingen in *Algemeen Dagblad*, 4 november 2023.

Waszink, Vivien (2023). Sarah in Wokeland, interview over inclusief taalgebruik. VRT, 5 december 2023.

Waszink, Vivien (2023). ‘Op zo’n manier samengevat kun je weinig tegen woke hebben’. Recensie in *Parool* over ‘Sarah in wokeland’, in *Parool*, 6 december 2023.

Waszink, Vivien. ‘Graaiflatie is woord van het jaar’. Interview over Woord van het Jaar voor radioprogramma *Spraakmakers*, NPO Radio 1, 19 december 2023.

Waszink, Vivien. Interview over inclusief taalgebruik, gebruik in Digitale Methode, een lesmethode voor het Vlaams secundair onderwijs.

Waszink, Vivien. 'Mag je met dat woke taalgebruik nou niets meer zeggen?'. Interview in de podcast van *Van Dale*.

Waszink, Vivien. Interview over inclusief taalgebruik in de de podcast Damn Honey.

Diversen

Colman, Lut

- Reviewer eLex 2023

De Tier, Veronique

- Voorzitter Stichting Nederlandse Dialecten
- Bestuurslid Variaties vzw. Koepelorganisatie voor dialecten en oraal erfgoed in Vlaanderen
- Bestuurslid Brusseleir!
- Bestuurslid Sociolinguistics Circle
- Lid werkgroep infrastructuur Fries-Nederlandse contactvariëteiten
- Redactielid Tijdschrift Nehalennia
- Begeleiding stagiair Palmyra De Nil (dialectologie)

Heylen, Kris

- ENEOLI COST-actie (CA22126): Core Group member, Working Group 3 co-leader en Management Committee Member voor Nederland
- NexusLinguarum COST-actie (CA18209): Working group member

Sijs, Nicoline van der

- Begeleiding stagiairs: Sara Schreuders (DAGENTA), Gemma van Dam (etymologiebank), Romy Regina Ricci (etymologiebank), Marcella Faltas (etymologiebank), Jos Verkroost (etymologiebank, eWND), Carmen Been (eWND), Francesca Tammaro (GLAD), Elizaveta Zaitseva (GLAD), Valentina Casini (The Digital Pallas), Maud van Es (The Digital Pallas).
- Lid manuscriptcommissie Charlotte Verheyden (2023), *"Ge moet zien dat gij uw vlaamsche taal niet vergeet". Een historisch-sociolinguïstisch onderzoek naar de invloed van het Frans op het Zuidelijke Nederland*, PhD Brussel
- Lid manuscriptcommissie Brenda Assendelft (2023), *Verfransing onder de loep – Nederlands-Frans taalcontact (1500–1900) vanuit historisch-sociolinguïstisch perspectief*, PhD Leiden
- Mentor for the University of New Europe (UNE)
- Lid van het Taalstaatpanel voor de verkiezing van Taalstaatmeester

- Bestuurslid Stichting Nederlandse Dialecten (SND)
- Secretaris van de Orde van den Prince, afdeling Utrecht
- Voorzitter Stichting beheer database Pinkhof geneeskundig woordenboek
- Redacteur van Internationale Neerlandistiek
- Redacteur van Neerlandistiek
- Hoofdredacteur, met Peter-Arno Coppen en Marc van Oostendorp, van de nieuwsbrief Neerlandistiek voor de klas
- Lid beoordelingscommissie van het Onderwijsfonds van de Maatschappij der Nederlandse Letterkunde
- Voorzitter van het bestuur van de Kiliaanstichting ter bevordering van het etymologisch onderzoek in Nederland en België
- Voorzitter van het bestuur van de Stichting Jiddische Lexicografie Amsterdam
- Curator namens de Universal Esperanto Association van de bijzondere leerstoel Esperanto en Interlinguïstiek bij de Universiteit van Amsterdam
- Redacteur van Trefwoord, tijdschrift op het gebied van lexicografie en lexicologie

Tiberius, Carole

- Board of Trustees Adam Kilgarriff Prize
- National Anchor Point LDS (European Language Data Space)
- Secretary-Treasurer EURALEX (European Association of Lexicography)
- NexusLinguarum COST-actie (CA18209): Management Committee Member voor Nederland
- UniDive COST-actie (CA21167):Management Committee Member voor Nederland

Vandeghinste, Vincent

- Lid jury doctoraat J. Sibeko. Measuring text readability in Sesotho. North-West University. Zuid-Afrika

Waszink, Vivien

- Auteur lesmethode *Nieuw Nederlands*, Noordhoff, methodes vmbo 3 gt, 4 havo, 5 havo, 4 vwo, 5 vwo en 6 vwo

- Stagebegeleider Milou Andree, masterstudent taalkunde aan de Universiteit Leiden, februari 2023-juli 2023

- Adviseur voor een onderzoek naar discriminerend taalgebruik, uitgevoerd door het Kennisplatform Inclusief Samenleven (KIS)

- Lid Taalexplorium, een samenwerkingsverband van diverse organisaties die educatieve activiteiten rond taal organiseert in o.a. bibliotheken

Bijlage 4: Taalmaterialen

Overzicht downloads commercieel

In totaal is er twintig keer een product afgenomen voor commerciële toepassingen.

CGN	4
Basilex Lexicon	4
Gigant-Molex	3
Basilex Corpus	1
CHN-ngrams	1
DID-NS	1
DuOMAn	1
E-lex	1
Lassy Groot	1
RBN	1
SoNaR	1
SumNL	1

Overzicht downloads niet-commercieel

Er is in totaal 816 maal een product gedownload voor niet-commercieel gebruik.

Corpus Gesproken Nederlands (CGN)	117
SoNaR-corpus	57
CGN-annotaties	49
Frequentielijsten corpora	32
GiGaNT-Molex	31
Dutch Parallel Corpus (DPC)	30
SoNaR Nieuwe Media Corpus	28
CHN N-grams	28
CELEX-2 Dutch	28
Lassy Klein-corpus	26
AI-Trainingset - Tag de Tekst voor Named Entity Recognition (NER)	22
Hoger Onderwijs Terminologie in Nederland en Vlaanderen (HOTNeV)	21
Referentiebestand Nederlands (RBN)	20
INT Historische Woordenlijst	19
Cd-rom Middelnederlands	17

BLISS Dialogue Summaries	16
DuOMAn Subjectivity Lexicon	16
Medische Termen Belgisch-Nederlands (MedTermBN)	15
Corpus Middelnederlands (Data)	15
Corpus Pathologische en Normale Spraak (COPAS)	14
Corpus Ondertitelde UVN-Colleges (COUC)	11
Corpus Gysseling (Data)	11
Belgian Covid Sign Language Corpus (BeCoS Corpus)	11
Medische Pilot (MedPilot)	10
BLISS Spoken Dialogue Dataset	9
Annotated Corpora for Term Extraction Research (ACTER)	8
Eindhoven-corpus	8
e-Lex	8
SoNaR Character n-grams	8
Referentiebestand Belgisch-Nederlands (RBBN)	7
PAROLE-lexicon	7

WAI-NOT Corpus	7
DuELME	7
Dutch Idiom Database: Native Speakers (DID-NS)	6
IFA Dialoog-videocorpus	6
IFA-corpus	6
RND Woordenlijsten	6
Wablieft-corpus	6
The LiLaH Emotion Lexicon of Greek, Kurdish, Turkish, Spanish, Farsi and Chinese	5
Brieven als Buit - Gouden Standaard	5
Jasmin	4
Woordenboek Vlaamse Gebarentaal (Woordenboek VGT)	4
Hotel Review Corpus in Nederlandse Gebarentaal (NGT_HoReCo)	4
Paco-MT Parallelle Corpora	4
D-TUNA-corpus	4
Afrikaans Custom Dictionary for Government Domain	4
Basilex Corpus	3

Moroccorp	3
COREA-coreferentiecorpus	3
Corpus Vlaamse Gebarentaal (Corpus VGT)	3
DAESO-corpus: parallelle Nederlandstalige monolinguale treebank	3
INT IMPACT NE-lexicon	3
Lwazi Sesotho Pronunciation Dictionary	2
CombiLex	2
Basilex Lexicon	1
Chorec	1
Lwazi Afrikaans Pronunciation Dictionary	1
Xitsonga Genre Classification Corpus	1
AUTONOMATA-namencorpus	1
AUTONOMATA-POI-corpus	1
Autshumato Sesotho sa Leboa-English Translation Memory	1
Lwazi Setswana Pronunciation Dictionary	1
Lwazi English Pronunciation Dictionary	1
OMBI Nederlands-Indonesisch	1

OMBI Arabisch-Nederlands	1
NAMES Corpus	1
Lwazi isiXhosa ASR Corpus	1
Lwazi Xitsonga Pronunciation Dictionary	1
Lwazi Xitsonga ASR Corpus	1
Lwazi Tshivenda Pronunciation Dictionary	1
Lwazi Siswati Pronunciation Dictionary	1