Raad voor de Nederlandse Taal en Letteren
Nederlandse Taalunie

https://www.raadntl.org

Telefoon
+31 70 750 87 01
E-mail
raad@taalunie.org

Adres
Paleisstraat 9
2514 JA Den Haag
Nederland

Digitale taalinfrastructuur en AI: aanbevelingen voor een verantwoord en duurzaam beleid

Inleiding
Met dit advies wil de Raad voor de Nederlandse Taal en Letteren bijdragen aan een verantwoord en toekomstbestendig beleid voor de digitale taalinfrastructuur voor het Nederlands. Het advies is een reactie op de opmars van artificiële intelligentie (AI) op basis van taalmodellen, en de roep om Europese soevereiniteit van kennis en data. De onderliggende visie is dat het Nederlands, in al zijn variëteit, structureel verankerd moet zijn in de Europese AI-ontwikkeling. Een taal die niet meesurft op de golf van artificiële intelligentie, belandt in de marge. Duurzame investering in technologie die behalve up-to-date, ook transparant is en meerstemmig, is daarom voor de directe en indirecte ondersteuning van de diverse sociaal-economische sectoren en groepen gebruikers van het Nederlands een strategische noodzaak.

Het advies aan het Comité van Ministers richt zich op de volgende punten:

1. beleid en structurele investeringen gericht op de uitbouw van de taalinfrastructuur voor het Nederlands;
2. een evaluatieprogramma dat inzicht geeft in de performance van taalmodellen en AI-toepassingen, voor de volle breedte van de taalvariëteit in Nederland, Vlaanderen en Suriname;
3. een kenniscentrum dat informatie en expertise rond taal en AI bundelt en toegankelijk maakt.

Context en achtergrond
Kansen en risico's
De in de afgelopen decennia sterk toegenomen volumes aan taaldata kunnen in combinatie met AI worden gebruikt als basis voor het 'trainen' van krachtige taalmodellen (LLM's, van 'large language models'). De beschikbaarheid van LLM's heeft voor veel categorieën taalgebruikers, organisaties en bedrijven het aanbod aan digitale hulpmiddelen, variërend van vertaaltools, spraakgeneratie, tekstoptimalisering, zoekmachines tot chatbots (hierna: AI-voor-taal), aanzienlijk verbreed. De risico's van deze ontwikkeling zijn echter maar ten dele bekend. Gesignaleerde zorgen en problemen zijn onder meer: applicaties die onjuiste informatie genereren, juridische en ethische vraagstukken rondom eigendom van de onderliggende trainingsdata en de dominantie van bepaalde taalvarianten (ook wel aangeduid als 'bias'), de impact op de menselijke cognitie en communicatiecultuur in brede zin, en de afhankelijkheid van buitenlandse private ondernemingen. Verder zijn veel taalmodellen primair ontwikkeld voor het Engels of combinaties van talen. (De Europese Unie investeert in zulke brede of generieke LLM's via projecten als LLMs4EU). Voor adequate ondersteuning van producten en diensten voor het Nederlands is bijstelling ('finetuning') nodig van die LLM's. Om de aansluiting van het Nederlands (zowel in geschreven als gesproken vorm) bij het aanbod van diensten

voor het Engels en andere grotere talen te behouden zijn ruimere investeringen nodig in de taalinfrastructuur voor het Nederlands, zowel voor de technologische basis, als de coördinatie van de kennisinfrastructuur. Investeringen in rekenfaciliteiten die al op stapel staan (zoals het Vlaams Supercomputer Centrum en de AI Factory in Nederland) en algemene beleidskaders (zoals de Nederlands Digitaliseringstrategie en het Vlaams Beleidsplan AI) dragen bij aan de benodigde rekencapaciteit.

De basis: data van hoge kwaliteit
Voor effectief beleid rondom AI-voor-taal moet de primaire focus liggen op de kwaliteit, herkomst en representativiteit van de onderliggende taaldata voor training. Cruciaal zijn de beschikbaarheid van Nederlands taalmateriaal, in adequate volumes voor alle taalvarianten, en ethisch en juridisch verantwoord verzameld en verwerkt. Het vanuit Nederland opgezette project GPT-NL en recente vergelijkbare Vlaamse initiatieven zijn belangrijke bouwstenen, maar voor de succesvolle uitbouw van de infrastructuur voor digitale taaldata zijn grotere investeringen nodig. Voor de duurzame inbedding van wat AI de gebruikers van het Nederlands kan bieden moet de bestaande federatieve infrastructuur van datacentra en archieven voor het beheer van taalmateriaal versterkt worden. Er zijn al initiatieven in voorbereiding bij de nationale bibliotheken KB en KBR om grote volumes hedendaagse en historische openbare bronnen beschikbaar te stellen voor taalmodellering. Verdere standaardisering en samenwerking zijn essentieel, naast optimale aansluiting bij lopende infrastructurele initiatieven op Europees niveau zoals CLARIN ERIC, ALT-EDIC en de LDS, waarin publieke en private organisaties samenwerken.

Naast kwantiteit en kwaliteit van data: balans
Naast de omvang en kwaliteit van gebruikte data is balans in taalvariëteit bepalend voor de bruikbaarheid van taalmodellen en de mogelijkheid van finetuning. Denk aan formeel versus informeel taalgebruik, regionale variatie, registers voor specifieke domeinen (bijvoorbeeld zorg, rechtspraak, media en economie). Ook voor AI-tools voor het leren van het Nederlands als tweede taal en andere educatieve contexten is adequate selectie van trainingsdata essentieel. Multilinguale (of generieke) LLM's hebben door hun omvang vaak een betere performance en minder bias dan monolinguale (of taalspecifieke) LLM's. Om adequate toepassingen voor specifieke gebruikscontexten te realiseren is finetuning van generieke modellen nodig ('small' LM's). Dat vraagt om datasets die afgestemd zijn op de relevante taalvariant(en). Voor contextspecifieke validatie van LLM's zijn bovendien specifieke evaluatieparadigma's van belang. Dat alles vraagt om data-infrastructuur met een gebalanceerd aanbod aan taaldata dat de taalvariatie in het Nederlandse taalgebied voldoende representeert.

Bestaande expertise en de zichtbaarheid ervan
In het Nederlandse taalgebied is er brede expertise op het terrein van AI-voor-taal aanwezig, zoals hierboven geïllustreerd door de betrokkenheid van Nederland en Vlaanderen bij diverse initiatieven. Ook het Instituut voor de Nederlandse Taal (INT) heeft AI-expertise in huis, plus een lange traditie in verantwoorde opbouw, beheer, en metadatering van taalkundig verrijkte datasets. Bij diverse organisaties (in het academisch domein en daarbuiten) zijn er LLM's in ontwikkeling (zoals GPT-NL, en VLAM bij IMEC en VRT). Daarnaast is er de Vlaamse AI Academie (VAIA), en vormt het stelsel van

bibliotheken een rijke bron van kennis over relevante datasets voor AI-tools (zowel generatieve als analytische), en de impact ervan op gebruikersgroepen. Wat nog versterking behoeft, is de zichtbaarheid van al deze kennis en expertise voor alle relevante maatschappelijke sectoren. De aanwijzing van een verbindende regisseur kan ertoe bijdragen dat de partijen die verantwoordelijk zijn voor het taalbeleid voor het Nederlands meer gecoördineerd te werk gaan als het gaat om de aan AI-voor-taal gerelateerde uitdagingen.

Adviezen

De Raad adviseert het Comité van Ministers drie strategische beleidslijnen te overwegen:

1. *Open en toegankelijke data*

De Raad adviseert het opzetten van stimulerend en ondersteunend beleid (inclusief structurele investeringen) gericht op de uitbouw van de taalinfrastructuur voor het Nederlands. In lijn met het meerjarenbeleidsplan kan een coördinerende rol worden toebedeeld aan de Taalunie om de relevante actoren in een permanente overlegstructuur te verenigen.

Regie is nodig voor het vergroten van het volume aan representatieve datasets voor de finetuning van generieke modellen voor toepassing in Nederlandse, Vlaamse en Surinaamse context. Om tot een efficiënte benutting van de middelen hiervoor te komen is nauwe afstemming nodig tussen de verschillende datacentra. Ook afstemming van de juridische kaders voor het verzamelen en delen van taaldata is van belang om tot een voldragen systeem van licenties en vergoedingen te komen voor uitgevers en andere aanbieders van data. Dit alles in lijn met de binnen LDS uitgewerkte kaders en met aansluiting bij lokale initiatieven zoals GPT-NL en VLAM. Voor de toegang tot materialen kan verder het federatieve model van CLARIN als basis dienen: een gedistribueerd netwerk van repositories voor data die voldoen aan de FAIR-principes, waaronder het INT dat een leidende rol zou kunnen spelen in de uitbouw van de technische infrastructuur.

2. *Transparantie, evaluatie en cultuurspecifieke benchmarking*

De Raad bepleit investeringen in het opzetten van een evaluatieprogramma dat inzicht geeft in de performance van taalmodellen en AI-toepassingen binnen specifieke gebruikscontexten en voor de volle breedte van de taalvariëteit in Nederland, Vlaanderen en Suriname. Zo'n validatieprogramma moet rekening houden met culturele en professionele diversiteit in taalgebruik, de score bepalen van AI-tools (benchmarking) voor sectorspecifieke workflows, en onderscheid maken tussen toepassingen voor bijvoorbeeld onderwijs, media, zorg en het bedrijfsleven in brede zin. Een dergelijk programma moet ook de transparantie kunnen meewegen met betrekking tot de herkomst van trainingsdata en uitlegbaarheid van conclusies op basis van AI-toepassingen. De expertise van het INT kan hierbij als basis dienen. Regie op de afstemming tussen Nederland, Vlaanderen en Suriname kan worden belegd bij de Taalunie als supranationale verdragsorganisatie.

3. *Publieke kennisdeling over taal en AI*

Er is behoefte aan **een kenniscentrum dat informatie en expertise rond taal en AI** bundelt en toegankelijk maakt voor professionals, beleidsmakers en het brede publiek. Dit centrum kan worden opgebouwd rond de Taalunie en het INT, in samenwerking met partners zoals CLARIN, KB, KBR en VAIA en in afstemming op de bestaande

dienstverlening voor het (taal)onderwijs. Het centrum kan opleidingen, nascholing en publiekscommunicatie verzorgen, en bijdragen aan een bewust en verantwoord gebruik van AI in talige contexten. **De Raad adviseert om de haalbaarheid en de financiële voorwaarden voor zo'n centrum te onderzoeken, met betrokkenheid van alle relevante partijen.**

Kortweg
De Raad voor de Nederlandse Taal en Letteren bepleit versterking en adequate financiering van de taalinfrastructuur als duurzame basis voor de optimale ondersteuning van de gebruikers van het Nederlands binnen de nieuwe realiteit die zich met de opkomst van AI heeft aangediend in bijna alle vormen van persoonlijke en professionele communicatie.

Appendix: afkortingen en verdere referenties

Gebruikte afkortingen

AI	Artificiële Intelligentie
ALT-EDIC	Alliance for Language Technologies - EDIC (link)
CLARIN ERIC	Common LAnguage Research INfrastructure ERIC (link)
EDIC	European DIgital Infrastructure Consortium
ERIC	European Research Infrastructure Consortium
FAIR	Findable, Accessible, Interoperable, Reusable
GPT-NL	Generative Pre-trained Transformer - NL (link)
IMEC	Interuniversitair Micro-Electronica Centrum (link)
INT	Instituut voor de Nederlandse Taal (link)
KB	Nationale Bibliotheek | KB, Nederland (link)
KBR	Koninklijke Bibliotheek, België (link)
LDS	Language Data Space (link)
LLMs4EU	LLMs for Europe, EU-project (link)
LLM's	Large Language Models
VAIA	Vlaamse AI Academie (link)
VLAM	Vlaams LAnguage Model
VRT	Vlaamse Radio- en Televisieomroeporganisatie (link)

Verdere referenties

AI Coalitie for NL (link)
AI Factory Nederland (nog geen stabiele link beschikbaar)
Nederlandse Digitaliseringsstrategie (link)
STOA Panel, over Europese soevereiniteit van kennis en data (link)
Vlaams AI-onderzoeksprogramma (link)
Vlaams Beleidsplan AI (link)
Vlaams Supercomputer Centrum (link)