



TAAL IN HET TIJDPERK VAN DE KUNSTMATIGE INTELLIGENTIE. WAT MET DE BEVEILIGING VAN ONZE TAALDATA?



Frieda Steurs is hoogleraar aan de KU Leuven, faculteit Letteren, campus Antwerpen, en was campusdecaan tot 2015. Zij doet onderzoek op het vlak van computerondersteund vertalen, terminologiebeheer en meertalig documentbeheer en is lid van de onderzoekseenheid QLVL van de KU Leuven. Sinds 1 september 2016 is zij de directeur van het Instituut voor de Nederlandse taal in Leiden. Het Instituut voor de Nederlandse Taal staat in voor het aanmaken, archiveren en onderhouden van corpora, lexica, woordenboeken en grammatica's voor het Nederlands, en levert daarmee alle bouwstenen voor de bestudering van de Nederlandse taal (www.ivdnt.org). Zij is internationaal actief als research fellow aan de universiteit van Oranje Vrijstaat in Bloemfontein, Zuid-Afrika, en was jarenlang geassocieerd professor aan de Université Catholique de L'Ouest in Angers, Frankrijk. Tevens is zij President van TermNet, een internationaal terminologienetwerk gesticht door UNESCO en sinds 2018 Secretaris-Generaal van CIPL (het internationaal comité van linguïsten).

Taal- en spraaktechnologie

Spraaktechnologie wordt ingezet bij spraakherkenning. Het is een deelgebied van de informatica en de taalkunde waarbinnen methoden worden

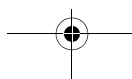
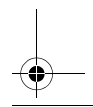
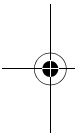
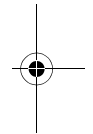




onderzocht en ontwikkeld die het mogelijk maken om computers het gesproken woord te laten herkennen en verwerken. Moderne spraakherkenning evolueert nu razendsnel: van statistische modellen van spraak en taal naar intelligente systemen die zelf kunnen leren. We horen de laatste tijd meer en meer over kunstmatige neurale netwerken (ofwel: *artificial intelligence* (AI)): *deep learning* is een vorm van *machine learning* die een computermodel gebruikt dat geïnspireerd is door de structuur van onze hersenen. Door deze technologie kunnen we een computer trainen en steeds beter laten presteren.

Spraak is nu eenmaal de meest directe en natuurlijke vorm van communicatie. De digitale trends zijn duidelijk: tech-giganten zoals Apple en Samsung integreerden de afgelopen jaren al digitale assistenten in de iPad of smartphone. Zo vind je via de Google Assistent al veel generieke toepassingen, zoals nieuws, file-informatie en weerberichten. Een nieuwe trend is ook dat de digitale stem gepersonaliseerd wordt. Geen gerobotiseerde stemgeluiden meer! Een authentieke stem met sterke dialogen en de juiste toon of stemkwaliteit is belangrijk voor de klant. Nu gesproken taal een belangrijk onderdeel wordt van de klantservice, richt men zich daar ook op beveiliging. Enerzijds zal het belang van stem-authenticatie (het nagaan of de stem werkelijk is van wie die beweert te zijn) toenemen. Om vergissingen en fraude tegen te gaan, wil een organisatie immers zeker weten of iemand daadwerkelijk de persoon is die hij of zij beweert te zijn. Spraakherkenning en spraakgeneratie nemen ook een vlucht in het zakelijke domein. Het vereenvoudigen van operationele processen via digitale spraak zal bij veel organisaties leiden tot gemak en efficiëntie en dus ook tot kostenbesparing. Denk aan de verkoper die op weg naar huis zijn gespreksverslag alvast vastlegt en een vervolgspraak inplant. Of de manager die vroeger zelf in spreadsheets moest duiken, maar in de toekomst zijn digitale assistent vraagt om een update van de dag.

De toepassingsmogelijkheden van spraakherkenning in de economie en de samenleving lijken onbegrensd. Op heel veel plekken waar mensen praten, kan een spraakherkenner een nuttige rol vervullen. Zo ook in de klantenservice van een onderneming. Selfservice via de telefoon is in veel opzichten aantrekkelijk. Er is geen wachttijd voor de klant en er zijn geen personeelskosten voor het bedrijf. Postcode, kentekens, data en namen zijn lastig in te voeren met de toetsen van je telefoon. Spraakherkenning op basis van een 'gesloten grammatica' is hiervoor de meest gebruikte oplossing. Bij een gesloten grammatica wordt precies aangegeven wat men op welke manier kan zeggen. Als mensen precies binnen dit beperkte menu blijven, is de herkenning uitstekend.



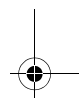
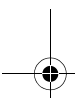
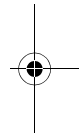
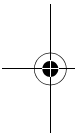


Spraakherkenning wordt gebruikt in tal van professionele omgevingen: juristenkantoren gebruiken dit al vrij systematisch, omdat het ook juridische terminologie ondersteunt. Woorden worden drie keer sneller in tekst omgezet, dan wanneer men het toetsenbord gebruikt. Een interessante toepassing van spraakherkenning is een product dat wordt ingezet bij de rechtbanken. Het leven van de griffier wordt er iets makkelijker door. Met een speciaal spraakherkenningssysteem, 'RechtSpraakHerkenning', kunnen rechters en griffiers makkelijk opzoeken wat er is gezegd tijdens een zitting. Notoir lastige items als namen van de verdachten, delicten en locaties kunnen dan goed herkend worden.

De combinatie van goede geluidsopnamen, sprekersspecifieke akoestische modellen en een taalmodel (hoe een taal werkt qua grammatica en lexicon) resulteert in goede herkenning. Dat, gecombineerd met een zoekstelsel dat het mogelijk maakt niet alleen in de 'best herkende' zin maar ook in mogelijke alternatieven te zoeken, levert een geavanceerd systeem voor het zoeken in rechtszittingen. Een toepassing die bijvoorbeeld voor politieverhoren interessant kan zijn. Ook in callcenters wordt er gewerkt met spraakherkenning, zodat men meteen kan achterhalen wat de beller wil weten, om dan juist door te schakelen. En veel auto's hebben een spraakherkenning ingebouwd in het navigatiesysteem, zodat je opdrachten kan geven en handenvrij de route kan zoeken.

Is spraakherkenning en sprekersherkenning hetzelfde? Neen, ze hebben wel met elkaar te maken. Stemherkenning of sprekersherkenning is een techniek waarbij de stem van een specifieke persoon wordt herkend, bijvoorbeeld ten behoeve van zijn identificatie. Forensische taalkundigen doen onderzoek naar deze automatisering van de herkenning van stemmen.

Taaltechnologie maakt gebruik van woordenboeken, grammatica's, teksten en van talloze andere vormen van taalkennis die mensen hanteren wanneer ze elkaar schrijven of met elkaar spreken. Het grote probleem daarbij is dat menselijke taal veel ingewikkelder in elkaar zit dan computertalen en vol zit met dubbelzinnigheden, uitzonderingen en betekenisnuances. Toch kan taaltechnologie zinvol worden ingezet in vele toepassingen om bedrijfsprocessen in kennis- en informatiemanagement te verbeteren, zo bijvoorbeeld in automatische classificatie: dit is een grote hulp bij het organiseren van een documentsysteem. In tegenstelling tot systemen als Google, die trefwoorden uit de teksten zelf halen, hoeven bij automatische classificatie die trefwoorden helemaal niet in de teksten voor te komen. Zo zal bijvoorbeeld in een krantenartikel over een nieuw model auto de term





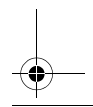
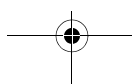
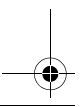
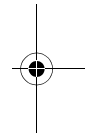
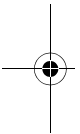
automobielinindustrie hoogstwaarschijnlijk helemaal niet voorkomen. Het kan echter belangrijk zijn het trefwoord toe te kennen, omdat iemand die speciaal op zoek is naar alle artikelen over het onderwerp *automobielinindustrie*, dan met één druk op de knop alle artikelen over dat onderwerp direct beschikbaar heeft.

Een ander voorbeeld is klantgericht onderzoek: bedrijven houden graag de vinger aan de pols en willen weten hoe de klanten denken over een nieuw, pas gelanceerd product. Marktonderzoek is vaak heel kostbaar en neemt veel tijd in beslag. Gelukkig zijn er goedkopere en eenvoudiger methoden. Zo bestaan er programma's die automatisch alle berichten op sociale media analyseren. Zo analyseert men iedere dag volledig automatisch duizenden tweets, posts, blogs en nieuwsberichten en ontdekt men automatisch of er positief, negatief of neutraal wordt gesproken over het bedrijf, de producten of de diensten ervan; dit noemt men 'sentimentanalyse'. Op die manier kan een bedrijf snel en zonder inspanning een beeld krijgen van de impact van zijn producten of diensten in de samenleving.

Maar er is ook een keerzijde: recente schandalen rond Facebook illustreren hoe gevaarlijk het is om onbewaakt gegevens te delen op sociale media en andere online platformen. Meer en meer informatie wordt gedeeld via allerlei online platformen waarbij de gebruiker geen idee heeft van wat er allemaal met die gegevens wordt gedaan. Websites en applicaties als Google Maps, Whatsapp, Bol.com, enz., analyseren het koopgedrag van klanten en geven op basis daarvan tips voor eventuele volgende aankopen. We geraken gewend aan een digitale assistent die ons de hele tijd begeleidt in onze keuzes, maar die ook verbazingwekkend veel weet over onze interesses en ons koopgedrag.

Het gevaar is dat in onze digitale economie de uitruil van informatie als een centraal element gebruikt wordt in de marktwerking. Data over gedrag en voorkeuren van mensen zijn voor digitale platformen als AirBnB en Uber van doorslaggevend belang bij het gericht aanbieden (dus ook onthouden) en de prijszetting van producten. Die platformen brengen dankzij grote datastromen gebruiker en aanbieder direct bij elkaar. Google, Facebook, Linked-In en Vliegwinkel zijn niet gratis; de gebruiker betaalt met persoonlijke informatie. Naast het product waarom hij vraagt krijgt hij als bonus te maken met ondoorzichtige pogingen om zijn koopgedrag te beïnvloeden. Dat is gewoon manipulatie.

Hoe kan dit worden opgelost? Mensen zijn niet gelijk aan hun digitale gegevens! Er moet een ethische reflectie komen over de wijze waarop innova-



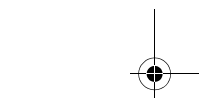
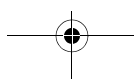
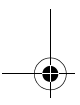
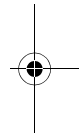
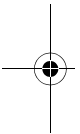


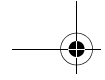
tieve processen ingrijpen in het persoonlijke leven. Gevoelige punten in het gebruik van digitale toepassingen zijn vooral het gebruik van software-interfaces die eigenlijk *dark patterns* zijn: een soort zwarte dozen waarvan de gebruiker niet weet of ze wel echt correct omgaan met de gegevens van hun klanten, en de metadata die aan de gegevens hangen. Die metadata vertellen heel veel over de klanten, en de klant is zich niet altijd bewust van hoe die gegevens worden gebruikt. Een nieuw soort *consent management* zal ingang moeten vinden waarbij de gebruiker zijn eigen data veel beter kan beheren. Een nieuwe uitdaging zal zijn om de AVG (Algemene verordening gegevensbescherming) toe te passen door een nieuw 'dataprofiel' van de gebruiker te creëren dat door de eigenaar van de data zelf kan worden beheerd.

Vertaaltechnologie: de vertaler krijgt hulp

Het is nog niet zo lang geleden dat vertalers gewoon met pen en papier aan de slag gingen. Het digitale tijdperk, en vooral de komst van de pc, bracht een eerste ondersteuning. De teksten werden opgeslagen, konden makkelijk worden aangepast, de zoekfunctie bracht de mogelijkheid een term op te sporen en meteen overal in de tekst te veranderen en nog meer klassieke hulpmiddelen, zoals spelling- en grammaticacontroles werden verwelkomd door de professionele vertaler.

Een volgende doorbraak kwam er met de massale groei van het internet. Grote tekstbestanden werden online gezet. Zo kon bijvoorbeeld iedere juridische vertaler zijn voordeel doen met heel veel juridische teksten die via het internet beschikbaar zijn, van het staatsblad tot Eur-lex, de databank van juridische documenten van de EU. Het werd allemaal via één muisklik toegankelijk. Vertalers die deze krachtige materiaalverzameling intelligent wilden benutten, leerden snel hoe je via geavanceerde dataminingstechnieken (het zoeken in grote gegevensverzamelingen) betrouwbare resultaten kon vinden op het internet. Ook de hoeveelheid (gratis) beschikbare online woordenboeken en termenlijsten groeide voortdurend. Pas daarna kwam de volgende stap in de vertaalprocessen. Dat wil zeggen dat de menselijke vertaler echt wordt geholpen met professionele computertoepassingen. Dat gebeurt met de zogenaamde CAT-tools, softwarepakketten die computerondersteund vertalen mogelijk maken. Let wel: in deze vorm van vertalen vertaalt de mens nog altijd de tekst, maar het vertaalproces wordt ondersteund door professionele software. Wat moeten we ons bij deze softwarepakketten voorstellen? Het gaat in de eerste plaats om intelligente databanksyste-





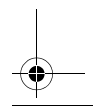
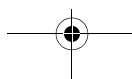
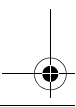
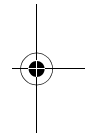
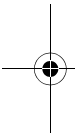
men, die terminologie en volledige zinnen met vertaling kunnen opslaan en een snelle zoekfunctie hebben. De vertaler kan de databank aanpassen per klant en kan binnen zijn werkomgeving de databank in een apart venster open laten staan.

Eén trend is duidelijk: veel speelt zich nu in de cloud af. De onderliggende principes zijn 'SaaS': *Software as a Service*. De klant moet de software niet meer aankopen en installeren op de eigen computers, maar betaalt een maandelijkse bijdrage om de software te mogen gebruiken. De software blijft bij de aanbieder, die alle updates en verdere ondersteuning aanbiedt. Het is een online dienstverlening en het bespaart de klant veel werk. Bij vele online vertaaldiensten en vertaalbedrijven zien we dat alles draait rond snelheid van opleveren. *Time is money* en dus moeten vertalingen snel kunnen worden geleverd, desnoods door verschillende vertalers samen te laten werken en door de revisoren en proeflezers ook sneller aan het werk te krijgen.

Het modewoord is *collaborative translation* ofwel in het Nederlands: samenwerking van alle deelnemers aan het vertaalproces. Dat kan de vertaalcyclus met 50 procent korter maken. Proeflezers kunnen daardoor veel sneller aan de slag en hoeven niet te wachten tot alles vertaald is. De communicatie tussen de vertaler, de proeflezer en de klant gebeurt via online platformen. Een vertaler kan in Boston aan het werk zijn, de proeflezer werkt in Los Angeles, en de klant zelf heeft zijn hoofdkantoor in Londen. Er hoeven ook geen documenten meer via e-mail verstuurd te worden. Alles blijft gecentraliseerd in de cloud en is online beschikbaar.

Gevoelige documenten vertalen? Nee, liever niet online!

Gebruik maken van de diensten van een cloud-provider betekent de controle over de IT-infrastructuur uit handen geven. Veel klanten hebben daar moeite mee en vertalers krijgen soms het verbod van hun klanten of bedrijf om online te werken. Gevoelige documenten, bijvoorbeeld van financiële of juridische aard, ziet men niet graag ongecontroleerd verdwijnen van de eigen servers. Cloud-providers moeten het beheer en onderhoud transparanter en controleerbaarder maken om het voor klanten gemakkelijker te maken deze stap te zetten. Online diensten en SAAS zijn een trend die niet meer te stoppen is. Een heleboel IT-bedrijven hebben hun applicaties en infrastructuur in de cloud. Toch blijven er ernstige vragen over de beveiliging. Af en toe bereiken ons geruchten over informatiehacking en cyber-





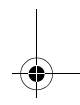
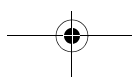
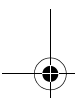
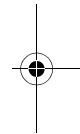
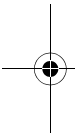
crime. Dus volgt de vraag: zijn onze gegevens veilig afgeschermd? Het blijft een feit dat bedrijven in de cloud computergebruik aanbieden aan meerdere gebruikers tegelijk. Dat betekent dat die gegevens van verschillende klanten in een gedeelde infrastructuur worden opgeslagen. Bedrijven moeten zich dan ook goed informeren alvorens de overstap naar de cloud te maken en een sluitend beveiligingsbeleid uitwerken. Dat is een belangrijk aspect van een goed algemeen bestuur.

Machinevertaling. Neemt de computer het vertalen over?

Sinds halverwege de twintigste eeuw wordt er al gedroomd van het ultieme computersysteem dat de menselijke taal kan decoderen en vertalen. De menselijke taal blijkt echter een heel moeilijk algoritme om via een computer te analyseren. De eerste grote projecten van automatisch vertalen waren dan ook gedoemd om te mislukken. Vanaf de jaren tachtig werd er veel genuanceerder gedacht over machinevertaling en werden de eerste successen geboekt.

De huidige systemen van machinevertaling die regelgebaseerd of statistisch gebaseerd zijn, blijven hun beperkingen hebben. Bij regelgebaseerde systemen moet de brontaal én de doeltaal volledig geanalyseerd zijn in alle componenten (regels) van de grammatica en het lexicon; bij statistische systemen kent de computer de talen totaal niet, maar gaat hij uit van de probabiliteit (de statistische kans) dat een vertaling juist is, op basis van wat op het internet aanwezig is. Dat lukt het beste voor talen waarvoor er heel veel materiaal online beschikbaar is (big data).

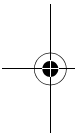
‘Neurale Machine Vertaling’ is een recente toepassing, die gebruik maakt van kunstmatige neurale netwerken, modellen waarbij het menselijke brein wordt nagebootst in een computer. Waar voorgaande vertaalsystemen veelal de meest voorkomende woordvormen kiezen, kiest de nieuwe methode de juiste woordvorm door de zinsopbouw in de brontaal te analyseren. Het neurale netwerk is in staat zelf grammaticale functies van woorden af te leiden, zonder dat het expliciete kennis van grammatica heeft. Onderzoeksgroepen ontwikkelen nu grote neurale netwerken om de vertaalkwaliteit te verhogen. Het onderzoek gaat verder, en de nieuwe technologieën zullen steeds meer mogelijkheden bieden om de vertaalkwaliteit te optimaliseren, zo onder meer met de ‘transformer neurale netwerken’.



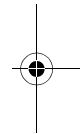


Een concreet voorbeeld is eBay. Dit online verkoopplatform wilde ook de Russische markt veroveren, maar stond voor een onmogelijke vertaalslag. Hoe kunnen alle aanbiedingen op eBay ook meteen in het Russisch worden vertaald, en omgekeerd, hoe kunnen klanten uit andere landen een vertaling krijgen van de Russische informatie? Dat kon alleen door een slimme vertaalmotor te gebruiken. De informatie op eBay wisselt namelijk van uur tot uur, afhankelijk van het aanbod dat de klanten doen, en de verkoop. Geen menselijke vertaler kan dit soort werk aan. De *Translator for Worldwide*-toepassing voor eBay was een groot succes. Alle aanbiedingen, miljoenen per rubriek, worden onmiddellijk vertaald naar het Russisch. En wat de Russische klanten aanbieden, wordt automatisch vertaald naar talen zoals het Engels, Spaans, Duits en Chinees. Het gebruik van die vertaalmotor vergde een aanzienlijke investering, maar de winsten zijn nog veel en veel groter.

Jibbiggo, Facebook en EU-Bridge. Nieuwe manieren van vertalen en communiceren. Wat biedt de toekomst?



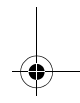
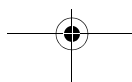
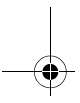
Nieuwe realisaties die spraaktechnologie koppelen met machinevertaling zijn er ook. Er zijn bedrijven zoals Jibbiggo.⁴⁶ Wat bieden zij? Het gaat om een heel innovatieve manier om offline te vertalen, gebaseerd op gesproken taal. Het product werd ontwikkeld door een onderzoeksteam van Carnegie Mellon University. Eenvoudig gezegd gaat het over een mobiele toepassing om op je smartphone gesproken taal om te zetten in een vertaling. Het project is succesvol en is inmiddels verkocht aan Facebook. Je opent de app, drukt op een knop en spreekt een zin in. De zin wordt meteen vertaald als tekst in beide talen, en wordt uitgesproken in de doeltaal. Een handige toepassing, die ook offline kan worden gebruikt. De toepassing werd in 2011 uitgeroepen tot de slimste ontwikkeling ter wereld.



Verdere toepassingen van deze gecombineerde spraakherkenning en machinevertaling vinden we in het project EU-Bridge⁴⁷, dat automatische transcriptie en vertaaldiensten levert. Mogelijkheden van dit project zijn onder meer de automatische ondertiteling van televisieprogramma's zoals weerberichten, Euronews-uitzendingen en YouTube filmpjes. Webinars, colleges die online worden gegeven, kunnen op deze wijze ook meteen ondertiteld worden in een andere taal. Er is ook een toepassing om

46. www.jibbiggo.com

47. <https://www.eu-bridge.eu/>





universiteitslezingen te ondertitelen voor anderstalige studenten. Dat gebeurt ‘live’: een professor geeft bijvoorbeeld een lezing in het Duits aan een Duitse universiteit en de studenten die bij de lezing aanwezig zijn kunnen via het softwareprogramma onmiddellijk de transcriptie zien van wat de prof gezegd heeft in het Duits, mét vertaling in bijvoorbeeld het Engels. Die toepassing is mogelijk voor alle talen en talencombinaties, op voorwaarde dat er voldoende digitaal taalmateriaal voorhanden is.

Clarín-Eric, taalmaterialen en gegevensbeveiliging

Omdat Europa grote ambities heeft wat betreft de meertaligheid en het aan de top blijven staan van de digitale taalrevolutie, werd in 2012 het netwerk CLARIN opgericht: *Common Language Resources and Technology Infrastructure*⁴⁸. CLARIN is een Europese gemeenschap van wetenschappers van verschillende disciplines en een netwerk van instellingen. De wetenschappers gebruiken CLARIN als een forum voor gezamenlijke ontwikkelingen en uitwisseling van middelen. Ze wisselen ook informatie uit over normen en procedures voor het langdurig archiveren van onderzoeksgegevens. Het instituut voor de Nederlandse Taal (INT) is het Clarín-centrum voor Vlaanderen en levert diensten op het vlak van digitale taaldata voor het Nederlands. Het hoofddoel van CLARIN is het opzetten en onderhouden van een veilige en duurzame infrastructuur voor taalmaterialen (teksten, woordenboeken enz.), zodat die eenvoudig gedeeld, (her)gebruikt, geanalyseerd en/of verrijkt kunnen worden.

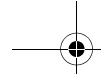
In CLARIN staat de vindbaarheid en/of de raadpleging van taalmaterialen centraal. Onderzoekers kunnen materialen zoeken in de CLARIN-catalogus, de Virtual Language Observatory (VLO). De VLO bevat verwijzingen naar meer dan 1,6 miljoen materialen in zowel nationale als regionale talen. Specifiek voor het Nederlands vindt de gebruiker alles ook via de website van het INT, onder het kopje ‘Taalmaterialen’⁴⁹.

Een groot voordeel van CLARIN is de schaalvergroting: algemene infrastructuurdiensten kunnen over grenzen heen gebruikt worden, waardoor de kost van het opzetten en onderhouden van die structuren gedeeld kan worden tussen landen. De toegang tot de enorme rijkheid aan CLARIN-materialen zelf kan ook leiden tot meer geavanceerd onderzoek, tot dalende ontwikkelingskosten en tot nieuwe samenwerkingsmogelijkheden

48. <https://www.clarin.eu/>

49. <https://ivdnt.org/taalmaterialen>





over grenzen en vakgebieden heen, wanneer onderzoekers van over heel Europa verder kunnen bouwen op elkaars resultaten.

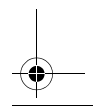
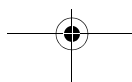
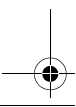
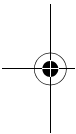
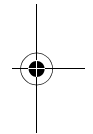
Het CLARIN-lidmaatschap draagt bij aan een betere positie van het Nederlands in een Europese (en internationale) context doordat de Nederlandse gegevens zichtbaarder en eenvoudiger toegankelijk worden voor een breder publiek, net zoals het culturele erfgoed dat inherent deel uitmaakt van taalmaterialen in de CLARIN-infrastructuur.

Besluit

Taal- en spraaktechnologie evolueren razendsnel en laten de gebruiker toe om via allerlei innovatieve manieren te communiceren, gebruik te maken van diensten, aankopen te doen en dit in meerdere talen. Voor het Nederlands is het belangrijk dat onderzoek ertoe blijft bijdragen dat de taal digitaal ontsloten blijft en dat alle toepassingen ook in onze taal beschikbaar zijn. Tegelijk is waakzaamheid geboden op het vlak van privacy en beschikbaar stellen van taalmaterialen.

IN HET KORT

Taal- en Spraaktechnologie (TST) is een heel ruim begrip. We verwijzen ermee naar de technologie waarmee geprobeerd wordt de menselijke, talige communicatie na te bootsen. Het omvat zowel Spraak-naar-Tekst als Tekst-naar-Spraak, analyse van teksten, vertaling en nog vele andere toepassingen. De taaltechnologie heeft in de laatste decennia een grote vlucht genomen. Waar men lang gedacht heeft dat menselijke taal niet of moeilijk na te bootsen was door een computer, en dat een computer taal niet kon interpreteren, is het inzicht heel snel aan het veranderen: de recente ontwikkelingen in de kunstmatige intelligentie en het gebruik van sociale media zorgen voor een ware omwenteling in het gebruik van talige materialen. Dit leidt tot onvermijdelijke vragen op het gebied van de beveiliging van deze data. Het Instituut voor de Nederlandse Taal (INT, Leiden) is een Vlaams-Nederlands instituut dat de digitale taalmaterialen voor het Nederlands verzamelt en beschikbaar stelt. In dit artikel bespreken we ook de noodzakelijke investeringen die voor elke taal moeten gebeuren om deze digitaal 'levend' te houden.





Dat betekent dat de nodige gegevens (teksten, spraak, woorden en termen) beschikbaar moeten zijn op digitale dragers, anders kan de bedrijfs-wereld de nieuwe toepassingen niet in die taal op de markt brengen. We belichten kort de nieuwe trends in de vertaaltechnologie en het leren van talen en beantwoorden de vraag: neemt de computer het vertalen en andere talige communicatie over? Tot slot: wat zijn de gevolgen op het vlak van privacy en het gebruik of misbruik van onze talige data?

LITERATUUR

- Eynikel, J. (2017). *Robot aan het stuur: over de ethiek van techniek*. Lan-
noo Campus.
- Kaptein, M. (2018). *Hallo wereld, hallo computer: wat iedereen zou moe-
ten weten over digitale technologie*. Amsterdam: Atlas Contact.
- Mayer-Schönberger, V. & Cukier, K. (2013). *De big data-revolutie: Hoe de
data-explosie al onze vragen gaat beantwoorden*. Amsterdam: Maven
Publishing.
- Steurs, F. (2016). *Taal is Business. Taal, de turbo naar economisch succes!*.
Schiedam: Scriptum.
- Tegmark, M. (2017). *Life 3.0 Mens zijn in het tijdperk van kunstmatige
intelligentie*. Amsterdam: Maven Publishing.

Websites

- [https://www.wrr.nl/publicaties/rapporten/2016/04/28/big-data-in-
een-vrije-en-veilige-samenleving](https://www.wrr.nl/publicaties/rapporten/2016/04/28/big-data-in-een-vrije-en-veilige-samenleving) Rapport van de wetenschappelijke
raad voor het regeringsbeleid, Nederland, 2015
- <https://ivdnt.org/taalmaterialen>

