# Gebruik van taal en spraak in een informatie-systeem

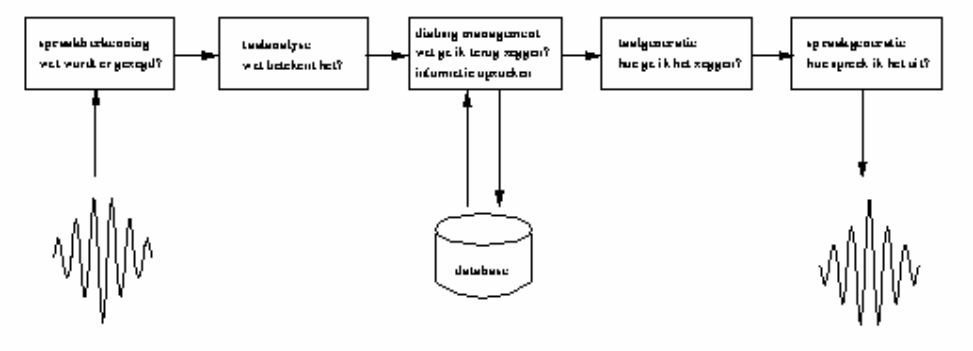
*Mariët Theune en Esther Klabbers*

Tegenwoordig is er een enorme hoeveelheid informatie beschikbaar, meestal opgeslagen in computerbestanden. Voorbeelden hiervan zijn reisgegevens bij een reisbureau, openbaar vervoersinformatie van de NS, informatie over telefoonnummers, bioscoopinformatie, etc. Er zijn verschillende manieren om toegang te krijgen tot die informatie. Soms is het mogelijk om via internet de informatie op te vragen. Meestal echter wordt de informatie beschikbaar gemaakt via een tussenpersoon, bijvoorbeeld een medewerker van een reisbureau of een telefoniste van OVR. Het voordeel van zo'n tussenpersoon is dat je het systeem niet fysiek hoeft te bedienen. Je kunt op afstand en in je eigen woorden vertellen welke informatie je zoekt. Een nadeel voor de informatieverstrekker is dat het werken met een tussenpersoon arbeidsintensief en duur is. Er wordt bijvoorbeeld jaarlijks zo'n 17 miljoen keer naar Openbaar Vervoer Reisinformatie gebeld. Dat zijn veel meer telefoontjes dan de ruim 400 telefonistes aankunnen. In maar liefst 7 miljoen gevallen is de wachttijd zo lang dat bellers ophangen. Een mogelijke betaalbare oplossing is om te proberen een deel van de taak van de tussenpersoon te automatiseren, zonder het voordeel van het gebruik van gesproken natuurlijke taal kwijt te raken. Menselijke tussenkomst blijft noodzakelijk en kan dan ingezet worden voor de beantwoording van meer complexe informatie vragen. Het voordeel van automatisering is dat er veel meer mensen geholpen kunnen worden, omdat een automatisch systeem meerdere bellers tegelijk kan bedienen, en omdat het systeem 24 uur per dag beschikbaar is. Bovendien is het maken van een dergelijk systeem een goede aanleiding om diverse taalkundige inzichten toe te kunnen passen. In het NWO Prioriteitsprogramma Taal- en Spraaktechnologie werken we daarom aan een systeem voor het opvragen van reisinformatie.

*Taken van een "talig" informatiesysteem*

Taal is het meest natuurlijke communicatiemiddel en een gesprek (dialoog) gaat ons meestal moeiteloos af. Toch is het voeren van een dialoog een complexe taak. Er zijn een vijftal basistaken te onderscheiden die wij onbewust uitvoeren. (1) We worden geconfronteerd met een spraaksignaal (geluidsgolven) van onze dialoogpartner en hieruit onderscheiden wij woorden, (2) waarvan we de betekenis bepalen. (3) Op grond van wat we gehoord hebben bedenken we vervolgens iets om terug te zeggen, (4) we kiezen op wat voor manier we dat gaan zeggen en (5) spreken het vervolgens uit (geluidsgolven). In een dialoogsysteem op de computer zijn het meestal aparte modules (grotendeels zelfstandig opererende computer programma's) die verantwoordelijk zijn voor deze verschillende activiteiten. Bij het bouwen van deze modules wordt duidelijk hoeveel taalkundige kennis er nodig is voor het uitvoeren van bovenstaande taken.

De dialoogtaken die wij als mens haast ongemerkt uitvoeren, blijken lang niet zo gemakkelijk te simuleren door de computer. Het probleem begint voor de computer al bij de spraakherkenning, waarbij op grond van het spraaksignaal van de menselijke dialoogpartner hypothesen gevormd worden over de woorden die zijn uitgesproken. Deze taak is complex doordat spraak een continue stroom van geluidsgolven is, waar woordgrenzen niet in gemarkeerd zijn. Bovendien kunnen op elkaar lijkende woorden makkelijk worden verwisseld. Daarom levert de spraakherkenningsmodule meestal meerdere hypothesen op. De taalanalyse module moet vervolgens een keuze maken uit die hypothesen. Hierbij is kennis van syntaxis nodig (welke woordcombinaties vormen een grammaticale zin?), van semantiek (welke woordcombinaties zijn te combineren tot een betekenisvolle zin?) en van de dialoog context (welke zin sluit het beste aan bij het voorgaande dat is gezegd?).




Na bepaling van de meest waarschijnlijke interpretatie van wat er gezegd is, moet de dialoog management module beslissen wat er teruggezegd moet worden, en eventueel de daarvoor benodigde gegevens uit de database halen. Hierbij is o.a. kennis nodig over het verloop van dialogen, bijv. dat er na een vraag een antwoord moet volgen. Daarnaast moet rekening gehouden worden met het feit dat er geen *zekerheid* is over wat er gezegd is. Vaak zal een reactie van het systeem daarom moeten bestaan uit een verificatie van wat het systeem denkt dat de gebruiker gezegd heeft, bijvoorbeeld "Zei u dat ... ?" De precieze vorm waarin deze verificatie (of andere boodschappen van het systeem) gegoten zal worden, wordt bepaald in de taalgeneratiemodule. Welke woorden zijn het meest geschikt om de boodschap uit te drukken, en wat voor zinsvorm kan het best gebruikt worden? Daarna kunnen de gegenereerde zinnen worden uitgesproken door de spraakgeneratie module, die ervoor moet zorgen dat de gebruiker van het systeem ze goed kan verstaan en aangenaam vindt klinken.
In de afgelopen jaren is er veel onderzoek gedaan naar gesproken dialoogsystemen. Echter, de nadruk lag meestal op de spraakinvoerkant omdat hier de problemen het grootst leken. Tegenwoordig is de kwaliteit van spraakherkenning goed genoeg voor toepassing in commerciële dialoogsystemen, zodat er ook steeds meer andacht kan worden besteed aan de spraakuitvoerkant, die in feite even belangrijk is. Aangezien de spraakuitvoer het enige onderdeel is van het systeem dat de gebruiker bewust waarneemt, zal hij hier het hele systeem op beoordelen. Onnatuurlijke spraak en vreemd geformuleerde zinnen zullen het gebruikersoordeel negatief beïnvloeden. Bovendien kan een slechte kwaliteit van de spraakuitvoer tot gevolg hebben dat een gebruiker zijn spreekstijl aanpast, wat negatieve gevolgen heeft voor de resultaten van de spraakherkenning. Het is dus van groot belang om de kwaliteit van de uitvoer te maximaliseren.

*Taalgeneratie*
Nadat de dialoog management module van het informatiesysteem heeft bepaald *wat* het systeem moet gaan zeggen, bepaalt de taalgeneratie module *hoe* het gezegd gaat worden: de talige "verpakking" van de boodschap. Dezelfde boodschap kan vaak op verschillende manieren worden uitgedrukt; er moeten dus keuzes gemaakt worden. We zullen er een paar bespreken aan de hand van voorbeelden uit ons treininformatiesysteem, dat als doel heeft om reisgegevens aan een gebruiker te verstrekken. Om te achterhalen welke informatie de gebruiker wil, stelt het systeem vragen en probeert het zijn interpretatie van de antwoorden op die vragen te verifiëren. De taalgeneratie kan bijvoorbeeld de opdracht krijgen van de dialoog management module om een systeemuiting te formuleren die de volgende elementen uitdrukt: (1) een vraag naar het gewenste vertrekstation; (2) verificatie van *Utrecht* als aankomststation en (3) verificatie van *morgen om 10 uur* als tijdstip van vertrek.
De eerste keuze die nu gemaakt moet worden betreft het opdelen van de boodschap in zinnen. In principe kan de hele boodschap in één zin worden uitgedrukt, bijv. (i) *Van waar uit wilt u morgen om 10 uur naar Utrecht vertrekken?* De te verifiëren informatie wordt hierbij opgenomen in de vraag, in de verwachting dat de gebruiker het systeem zal corrigeren als het een fout heeft gemaakt. De vraag in (i) bevat echter zoveel informatie, dat het voor de gebruiker moeilijk zal zijn alles te controleren. Beter is het om de verificaties en de vraag in aparte zinnen onder te brengen, zodat de gebruiker niet in een keer alles hoeft te verwerken: (ii) *U wilt dus morgen om 10 uur naar Utrecht vertrekken. Van waar uit wilt u vertrekken?* Wanneer er is gekozen voor een bepaalde opdeling van de boodschap in zinnen, volgen er nog veel meer beslissingen op het gebied van woordkeus, woordvolgorde etc.

Hiervan zullen we er nog een bespreken, namelijk de beslissing welke woorden een accent moeten krijgen en welke niet. In natuurlijke taal worden bepaalde woorden in de zin met meer nadruk uitgesproken dan andere; deze woorden zijn *geaccentueerd*. Welke woorden geaccentueerd moeten worden is afhankelijk van informatie die aanwezig is in de taalgeneratie module (maar niet in de spraakgeneratie module), bijvoorbeeld informatie over de structuur van de zin en over wat er al eerder in de dialoog gezegd is. Woorden die informatie uitdrukken die als "bekend" wordt verondersteld, worden in het algemeen niet geaccentueerd. In (ii) zal bijvoorbeeld het woord *vertrekken* aan het einde van de vraagzin geen accent krijgen, omdat de voorafgaande zin ook over een vertreksituatie ging. Echter, als de voorafgaande zin een verificatie van de *aankomst* tijd had bevat, had *vertrekken* beslist wel een accent moeten hebben: (iii) *U wilt dus morgen om 10 uur in Utrecht aankomen. Van waar uit wilt u VERTREKKEN?* In dit geval geeft de accentuering van *vertrekken* een contrast aan. Om te kunnen bepalen welke informatie als bekend of contrastief beschouwd moet worden, houdt de taalgeneratie module een zgn. "discourse model" bij. Hierin wordt bijgehouden welke informatie genoemd is in de dialoog tot zover, welke woorden gebruikt zijn, etc. Nadat de taalgeneratie bepaald heeft welke woorden geaccentueerd moeten worden, is het de taak van de spraakgeneratie om deze accenten op de juiste wijze in spraak te realiseren.

*Spraakgeneratie*

Bij het genereren van spraakuitvoer is het van belang om een balans te vinden tussen flexibiliteit en natuurlijkheid. De hoogste kwaliteit wordt bereikt door alle zinnen die het systeem moet kunnen uitspreken vantevoren op te nemen. Omdat iedere nieuwe zin die gegenereerd wordt, opnieuw moet worden opgenomen, kan deze methode alleen gebruikt worden wanneer het aantal zinnen beperkt is en de zinsconstructie niet varieert. De hoogste mate van flexibiliteit wordt verkregen door het gebruik van spraaksynthese, d.w.z. volledig door de computer gegenereerde spraak. De meest gebruikte vorm van spraaksynthese is difoonconcatenatie. Een difoon is een opgenomen spraaksegment dat bestaat uit de overgang tussen twee klanken. Het woord *taal* bestaat uit de difonen /stilte-t/ /t-a/ /a-l/ en /l-stilte/. Voor het Nederlands zijn er zo'n 1500 difonen nodig om alle klankovergangen te representeren. Daarmee kan iedere willekeurige tekst worden uitgesproken. De prosodische eigenschappen, zoals de duur van klanken en de zinsmelodie, worden door middel van regels aan de achter elkaar geplakte difonen toegekend. Het resultaat is een spraakuitvoer die goed verstaanbaar is maar die qua natuurlijkheid nog te wensen over laat. Toch is het nodig om deze methode te gebruiken in gevallen waarbij de woordenschat groot is of waarbij er veel variatie is in zinsconstructie en accentuering. Daarom is het een goede zaak dat er onderzoek wordt verricht naar verbetering van de kwaliteit van synthetische spraak zodat die op den duur net als spraakherkenning *wel* in commerciële systemen kan worden gebruikt.

Een compromis tussen deze twee extremen is *frase-concatenatie*. Dit bestaat uit het aan elkaar plakken van woorden en frasen (zinsdelen), die met één spreker zijn opgenomen. Hiermee kan een hoge kwaliteit spraakuitvoer worden bereikt, mits er bij de opnamen rekening wordt gehouden met de context waarin de eenheden worden uitgesproken. Wanneer alle eenheden los worden uitgesproken zal het spreektempo te veel variëren en klinkt de zinsmelodie van de geconcateneerde zin niet natuurlijk. Daarom worden de woorden en frasen opgenomen terwijl ze zijn ingebed in zinnen die de contexten weergeven waarin de betreffende woorden kunnen voorkomen. De frase-concatenatie geeft het beste resultaat wanneer bepaalde woorden in meerdere contexten worden opgenomen, afhankelijk van (1) de positie in de zin waarin het woord kan voorkomen, en (2) de accentuering, die door de taalgeneratie module wordt berekend. Op die manier kan er optimaal gebruik worden gemaakt van de kennis die de taalgeneratie levert.

De frase-concatenatie methode kan in het geval van een treininformatiesysteem heel goed gebruikt worden omdat de uitvoer bestaat uit een beperkt aantal zinnen waarin stationsnamen, data en tijden kunnen worden ingevuld. Wanneer het aantal zinnen drastisch toeneemt, zal de hoeveelheid opnamen die gemaakt moeten worden ook erg toenemen, wat frase-concatenatie erg onpraktisch maakt. Bovendien is het belangrijk dat de woordenschat constant is. Wanneer er steeds nieuwe informatie bijkomt, moeten er steeds nieuwe opnamen bijgemaakt worden met dezelfde spreker. Daarmee loop je het risico dat de spreker op een gegeven moment niet meer beschikbaar is, of dat er verschillen tussen verschillende opnamesessies hoorbaar worden.

De inzichten die voortkomen uit het huidige onderzoek naar spraakuitvoergeneratie in het treininformatiesysteem, kunnen ook worden toegepast in andere applicaties. Dit kunnen andere dialoogsystemen zijn, maar ook systemen die opgeslagen data direct omzetten in taal, bijv. in de vorm van gesproken weer- of beursberichten, sportverslagen, etc. Hierdoor zou een groeiend aantal databestanden ontsloten kunnen worden voor een groot publiek.