Te zwak of niet te zwak: natuurlijke taal als modelleergereedschap

De SIKS bijeenkomst van 4 mei j.l. in Maastricht kenmerkte zich door een geanimeerde discussie. Een van de hete hangijzers was de stelling van Han Schouten dat "Iedere getuigenis van menselijke kennis, ongeacht de vorm waarin deze is vastgelegd, met behoud van alle kennis, kan worden verwoord als zinnen in een natuurlijke taal".

Voor veel mensen die werkzaam zijn in de informatie-modelleer-wereld is natuurlijke taal een hekel punt. Aan de ene kant zijn er talloze methoden en technieken die ervan uitgaan dat natuurlijke taal (Natural Language, NL) een geschikt middel is om niet-technische mensen bij het ontwerpproces te betrekken. NL is immers de gemeenschappelijke noemer van (vrijwel) alle mensen. Maar aan de andere kant is NL lastig te "vangen" in formele termen die noodzakelijk zijn voor een formeel systeemontwerp, zodat er ook een grote groep mensen is die nadrukkelijk kiest voor formeel gereedschap en schema-technieken.

Maar is deze zwart/wit-redenering wel op zijn plaats?

Een van de bezwaren tegen Han's stelling dat NL de ultieme kennistransporteur is luidde als volgt: "Probeer dan maar eens het Nederlandse woord *gezellig* in het Engels te vertalen." Een bekende valkuil, want de combinatie van gesloten gordijnen, schemerlampen en koffie met gesopte koekjes is buiten ons land nauwelijks bekend. Een vertaling, of zelfs een omschrijving van *gezellig* stuit op moeilijkheden. *Gezellig* is niet een uniek identificeerbaar entiteittype, of een formeel definieerbare property. Veel, zo niet alle mensen zullen behoorlijk moeten nadenken en een flink stuk tekst produceren in een poging *gezellig* te definiëren.

Wat zegt Wolters' hierover? "**Gezellig** *bn, bw* 1 aangenaam, knus; 2 onderhoudend; 3 huiselijk, een huiselijk gevoel gevend, vriendelijk aandoend, pretentieloos". Daar komen we niet zoveel verder mee. Een trap dieper dan maar: "**Aangenaam** *bn, bw* behaaglijk, prettig"; "**Knus** *bn, bw* 1 aardig, lief; 2 gezellig". Daar begint de cirkel al. De Engelse versie is volgens van Dale "enjoyable, pleasant, companionable, social"; duidelijk niet direct vertaalbaar (zoals met zovele niet-entiteittypen). Frans dan maar? "Sociable, liant, agréable, charmant" en nog een tiental andere zeer zwierig aandoende uitdrukkingen.

Na al deze omschrijvingen van gezellig te hebben gelezen beginnen we toch wel redelijk gevoel te krijgen voor het gebruik van het woord. Voeg er nog een paar voorbeeldzinnen aan toe en we zijn er al bijna. Maar hoe komt dat? Waarom "voelen" we de betekenis van een term zoals *gezellig* aan door onszelf te confronteren met nog meer niet-gedefinieerde termen?

Hier raken we de kern van het probleem met NL. Natuurlijke taal heeft geen contextvrije basis-semantiek. NL is een onvoorstelbaar complex netwerk van onderling gerelateerde begrippen, die geen van allen uiteengerafeld zijn in basisprimitieven. De zinsstructuur van

een taal is inmiddels redelijk formeel beschrijfbaar, en een groot deel van de beschikbare logica's heeft tot doel de semantiek van taaluitingen te formaliseren. Tot op zekere hoogte lukt dat ook. Maar al deze hulptechnieken gaan voorbij aan het feit dat wijzelf, als mensen, hoogstwaarschijnlijk niet werken met formele definities.

Mensen zijn contextwezens. Wij *zijn* onze context. Ons hele leven lang doen we niets anders dan het opnemen van prikkels en het daarmee bijstellen van onze perceptie van de context waarin wij leven. We hebben een mentaal model van de wereld in ons hoofd, waarvan we hopen dat het voldoende nauwkeurig is om voorspellingen mee te doen. Vaak gaat dat goed, soms niet. Bij een conflict tussen ons model en de werkelijkheid stellen we ons model bij.

Omdat mensen sociale wezens zijn zullen we zeer veel ervaringen gedurende ons leven delen met anderen. Een zeer groot deel van ons mentaal model zal dus ook overeenstemmen met dat van onze directe omstanders. Hoe verder uit elkaar mensen leven, hoe meer hun mentale modellen zullen verschillen. Maar we blijven Mensen op Aarde, en een deel van onze context wordt daardoor bepaald. De mentale modellen van alle mensen zullen altijd een gemeenschappelijk deel hebben. In hoeverre erfelijke factoren hierbij meespelen is nog een onderwerp van discussie.

Het gemeenschappelijk deel—hoe groot of klein ook—is overigens niet exact gelijk. Het is juist voldoende gelijk om met elkaar te communiceren onder de aanname dat het gelijk is. Uit overwegingen van efficientie en wegens praktische redenen is een communicatie-mechanisme dat zo weinig mogelijk gegevens transporteert voor een zo groot mogelijk effect te prefereren. De natuur blijkt dit principe op zeer veel plaatsen te hanteren, en we hebben geen redenen om aan te nemen dat natuurlijke taal ervan afwijkt (ook al zitten er wel degelijk redundanties in de syntaxis). NL is zo de drager van aanpassingen (deltas) op het mentaal model van de ontvanger, gezien vanuit het model van de zender. Het mechanisme werkt alleen correct bij voldoende overeenstemming tussen beide mentale modellen.

NL is daarom ontzettend contextgevoelig. Het gaat zover dat dezelfde taaluiting in een andere context iets compleet verschillends kan betekenen. Voor informatie-modellering lijkt dat een ramp. Wij informatici werken immers het liefst met formeel gedefinieerde basisconcepten waarin we alle andere zaken uitdrukken. Dat modelleert lekker en geeft ook behoorlijke voordelen bij het ontwerp en de implementatie. Toch kunnen we vaak niet om die context heen; verderop in dit betoog kom ik erop terug.

De contextgevoeligheid van NL is ook een van de redenen dat we het woord *gezellig* niet zo één-twee-drie "formeel" uit kunnen drukken. We hebben een paar jaar nodig gehad voordat we het woord voldoende vaak tegengekomen waren om het in te koppelen in ons al bestaande model, en de enige manier die we hebben om het weer over te brengen naar anderen is het stelselmatig opsommen van al die koppelingen, hopend dat er voldoende koppelingen overeen zullen stemmen met het mentaal model van de toehoorder. Waar wij als mens jaren over doen om te leren is waarschijnlijk niet in enkele regels formele logica te vangen.

Er zijn nog meer manieren om het probleem van *gezellig* te tackelen. We stuiten daarbij op het onderscheid tussen de uitdrukkingskracht van NL ten opzichte van kunstmatige formele talen, en de uitdrukkingskracht van NL in het algemeen. Onder de formele talen scharen we nu even alle kunstmatige vormen van communicatie, zowel symbolisch als grafisch, die zijn ontwikkeld voor toepassing in een beperkt domein en die gefundeerd zijn op begrippen uit de logica of de wiskunde. Modelleringstalen voor informatiesystemen vallen hier onder.

De stelling dat iedere notie die in willekeurig welke formele taal uitdrukbaar is, ook kan worden uitgedrukt in NL, is zeer wel te verdedigen. NL is en blijft het meest krachtige gereedschap dat wij tot onze beschikking hebben voor het overbrengen van ideeën en kennis. Formele talen kunnen in een specifiek domein natuurlijk wel veel korter zijn dan de equivalente NL-uitdrukking. Maar alleen al het feit dat iedere formele taal eerst aan de gebruikers ervan moet worden uitgelegd zegt al iets. Een formele taal moet namelijk worden uitgelegd in NL.

De stelling dat NL iedere vorm van kennis kan representeren hangt vast aan onze definitie van kennis. Als we kennis definiëren als onze volledige verzameling ervaringen, dan lopen we vast. Niet al onze ervaringen zijn uitdrukbaar als een concept. Zo zijn er bijvoorbeeld ook *percepten*, zoals kleuren, emoties en muziek, die in geen enkele taal volledig uitdrukbaar zijn (ook niet in formele talen; λ=670 nm is geen kleuraanduiding). *Gezellig* is ook zo'n percept. In dergelijke gevallen is een woord hoogstens te beschouwen als een pijl in de goede richting. Maar voor iemand die het betreffende percept niet kent zal zo'n pijl weinig helpen; hij of zij dient dan meerdere malen geconfronteerd te worden met dat deel van de werkelijkheid dat gevangen wordt door de betreffende uitdrukking.

Wanneer we kennis nauwer definiëren en ons beperken tot de gangbare definitie *kennis is de verzameling overdraagbare ervaringen*, dan kunnen we wèl stellen dat NL iedere vorm van kennis kan overdragen. Ook hier geldt weer dat gespecialiseerde formele talen efficiënter kunnen zijn, maar nooit krachtiger.

Maar zelfs met deze definitie van kennis is NL soms een onhandig gereedschap omdat voor sommige concepten geen woord bestaat. In zo'n geval heeft kennelijk niemand nog ooit de behoefte gevoeld om er een woord voor te maken. Als die noodzaak er dan om wat voor reden dan ook wel komt, moet er soms heel wat denkwerk, inclusief het invoeren van hulpbegrippen, gedaan worden voordat de definitie kan slagen. De ontwikkeling van terminologieën, inclusief wetenschappelijk taalgebruik, valt onder deze noemer.

Terug naar het grensgebied tussen formele talen en NL als gereedschap voor het modelleren van informatiesystemen. Zou het misschien ook werken om de contextvrije formele semantiek vaarwel te zeggen en systemen te bouwen die gebaseerd zijn op de context die we al kennen? Zodat we niet telkens vanaf nul alle begrippen moeten definiëren in de kleinst mogelijke primitieven, maar direct kunnen starten vanuit bekend veronderstelde complexere concepten?

Inderdaad, dit is glibberig gebied. Hoe weten we precies welke concepten al bekend zijn, en welke niet? Het antwoord is dat we dat gewoonweg *niet* weten. We kunnen er slechts een gooi naar doen. Net zoals in de werkelijke wereld zullen we vaak raak gooien, en soms volledig mis. Maar een misser is helemaal niet erg—zolang we ons systeem eenvoudig en snel kunnen aanpassen aan de "nieuwe" werkelijkheid.

Dergelijke op NL gebaseerde beschrijvingsmethoden zijn geen vervanger van bestaande formele technieken, omdat er veel minder eenvoudig een deterministisch systeem mee te bouwen valt. Voor sommige (technische, administratief/bureaucratische) toepassingen is een dergelijk deterministisch gedrag een vereiste. Maar door ons te concentreren op formele beschrijvingen sluiten we ons af van de enorme wereld van de informele, inter-menselijke communicatie. Taalkundig verantwoorde systemen die werken met concepten binnen een context van gelijkgezinden kunnen wellicht behoorlijk wat betekenen voor minder formele omgevingen, waar nu de automatisering zich nog beperkt tot E-mail, Lotus Notes en Word Perfect. En voor het ontwerpen van formele systemen is een informeel, talig voortraject misschien ook een aantrekkelijke aanvulling op de huidige schematechnieken. Veel methoden hebben al zo'n talig voortraject, maar er zijn er maar heel weinig die ook werkelijk de taalkundige informatie verwerken in het eindresultaat.

Een groepje binnen SIKS rond Reind van de Riet, Jan Dietz, Robert Meersman en Hans Weigand is al enige tijd bezig dit idee verder uit te werken en zo te komen tot een nuttige inzet van een Lexicon, een coherente verzameling van begrippen uit de natuurlijke taal die in de dagelijkse praktijk worden gebruikt. Het Lexicon is geen simpele woordenlijst met (NL) definities, zoals een Data Dictionary. Het Lexicon is meer. Het bevat juist voldoende algemene taalkundige (woorden)kennis om voor specifieke domeinen de begrippen te kunnen uitdrukken in gangbare *common sense* kennis. Binnen dat domein is de inhoud van het Lexicon net zo formeel als de traditionele beschrijvingstechnieken. Het moet dus mogelijk zijn om op deze manier een informatie- en communicatiesysteem formeel te beschrijven met gebruikmaking van informele begrippen. Informeel omdat we ze niet verder uitdiepen. En zoals hierboven beargumenteerd, informeel is niet hetzelfde als vaag.

Maar wat is vaag?

Jeroen Hoppenbrouwers
ITK/Infolab
Katholieke Universiteit Brabant