Aspecten van automatisch vertalen. Resultaten – problemen

Mei 2001 (rev 7, December 2003)

Steven Krauwer
Utrechts instituut voor Linguïstiek UiL OTS
http://www-sk.let.uu.nl
steven.krauwer@let.uu.nl

1 Overzicht

In dit stuk proberen we een overzicht te geven van de stand van zaken met betrekking tot het automatisch vertalen. We zullen daartoe achtereenvolgens naar de volgende aspecten kijken:

- De vertaalcomputer
- Geschiedenis
- Problemen met oplossingen
- Vertaalstrategieën
- Problemen zonder oplossingen
- Waar staan we, en hoe nu verder?

Het is niet het doel van deze tekst om een technische verhandeling te geven van alle ins en outs van het automatisch vertalen. Daar bestaat veel geschiktere literatuur voor. We geven in deze tekst geen literatuurverwijzingen. Wie meer wil weten kan een bezoek brengen aan de website http://www-sk.let.uu.nl/ond/mt99.htm, die een aantal verwijzingen geeft naar literatuur en andere websites over automatisch vertalen.

2 De vertaalcomputer

2.1 Wat we er mee bedoelen

2.1.1 Teksten, geen spraak

We gebruiken de term automatisch vertalen in de regel voor het vertalen van geschreven teksten door een computer. We gaan er daarbij van uit dat de tekst wordt aangeboden in een vorm die de computer gemakkelijk kan verwerken, bijvoorbeeld de output van een tekstverwerker (Microsoft Word, WordPerfect), teksten die per elektronische mail worden aangeboden, of teksten die op een andere manier via het internet worden verkregen.

Dat betekent dat we niet in de eerste plaats denken aan het vertalen van gedrukte teksten die op papier worden aangeboden of het vertalen van handgeschreven teksten. We laten ook uitdrukkelijk het vertalen van gesproken tekst buiten beschouwing.

Voor wat betreft gedrukte teksten houdt dit eigenlijk geen beperking in: wie niet al te lang geleden voor zijn PC thuis of op het werk een scanner heeft aangeschaft (een apparaat om

papieren teksten en plaatjes om te zetten in een vorm die voor de computer hanteerbaar is), zal uit ervaring weten dat er vaak uitstekende programma's bijgeleverd worden die met verrassend grote precisie een als plaatje ingescande tekst in echte tekst (d.w.z. behandelbaar door een tekstverwerker) kunnen omzetten.

Het herkennen van handgeschreven teksten (d.w.z. omzetten in verwerkbare tekst) staat helaas nog in de kinderschoenen, en werkt eigenlijk alleen op heel beperkte schaal bevredigend (lezen van postcodes, lezen van bank- en giroformulieren). Hetzelfde geldt (zij het in iets mindere mate) voor gesproken taal: het omzetten van gesproken in geschreven tekst (geschikt voor vertaling) is een probleem op zich, en het zal nog wel even duren voordat de problemen daarbij goed genoeg opgelost zijn om deze technologie te kunnen gebruiken om gesproken taal direct te vertalen.

Niettemin is dit laatste natuurlijk wel wat we uiteindelijk zouden willen: we telefoneren direct met een Japanse collega, en beide partijen spreken hun eigen taal, en horen dat wat de ander zegt in de eigen taal, en liefst ook nog met de stem van de spreker.

2.1.2 Zakelijke teksten, geen romans of gedichten

Wanneer je met mensen praat over de moeilijkheid van automatisch vertalen, beamen ze dat meestal snel, verwijzend naar het vertalen van gedichten en romans. Dat getuigt van een wat optimistische kijk op de zaak. In feite is zelfs het vertalen van saaie teksten, zoals bijvoorbeeld de gebruiksaanwijzing van een mixer al een probleem waar we nog niet helemaal uit zijn. Het vertalen van poëtische of gevoelige teksten ligt nog heel ver bij ons vandaan. Ook de menselijke vertaler heeft daar moeite mee, en zelfs een vertaler met een perfecte beheersing van de vreemde taal en de moedertaal zal dat soort teksten alleen kunnen vertalen wanneer hij zelf een beetje dichter of romanschrijver is.

Het automatisch vertalen beperkt zich op dit moment geheel tot zakelijke, technische teksten (gebruiksaanwijzingen, beursrapporten, bijsluiters van medicijnen, patentaanvragen, weerrapporten, etc).

2.2 Waarom we het eigenlijk doen

Er zijn een aantal redenen waarom het aantrekkelijk is om naar automatisch vertalen te kijken. We noemen er hier een paar van:

- Economisch: drukken van de vertaalkosten. Vertalen is mensenwerk, en dus duur. Voor vertaling tussen de gangbare Europese talen ligt het tarief van een professionele vertaler al gauw tussen de f 90 en f 150 per bladzijde van 300 woorden (afhankelijk van het talenpaar), en voor andere talenparen kan het nog aanzienlijk hoger liggen. Bij een simpel apparaat, zoals een printer voor een computer, zit al gauw een handleiding van 500 pagina's (vertaalkosten ca f 50 000), en als de fabrikant zijn apparaten wereldwijd wil afzetten, met voor elk taalgebied een handleiding in de eigen taal (volgens onze eigen Consumentenbond een *must*), moeten we dit bedrag in elk geval met een factor 100 vermenigvuldigen: 5 miljoen gulden voor alleen de documentatie. Voor complexere apparaten, zoals legertanks of vliegtuigen, is de hoeveelheid documentatie per exemplaar nog eens honderd keer zo groot. Het is duidelijk dat we het hier hebben over een groot

economisch belang, en zelfs een reductie van de vertaallast van een paar procent kan al aanzienlijke bedragen opleveren.

- Praktisch: verhogen van de snelheid. Vertalen gaat niet snel. Een vertaler heeft voor een pagina al gauw 1 a 2 uur nodig. Die tijd zit niet alleen in het bedenken en opschrijven van de vertaling, maar ook in het lezen van achtergronddocumentatie, en het achterhalen (en soms zelfs bedenken) van de juiste technische termen, vaak een tijdrovende klus. De printerhandleiding alleen al kost de vertaler 3 a 4 maanden. Dat is voor bepaalde bedrijfstakken, waar producten elkaar snel opvolgen vaak te traag. Daarnaast zijn er ook vertalingen die, ongeacht hun omvang, toch snel geleverd moeten kunnen worden. Denk bijvoorbeeld aan medische of juridische rapporten, die onmiddellijk nodig zijn voor verdere behandeling van de zaak.
- Volume: er is meer werk dan vertalers aankunnen. Het aantal gekwalificeerde vertalers is relatief klein, vooral wanneer het gaat om minder gebruikelijke talenparen. Het beeld wordt vaak vertroebeld doordat er in deze branche veel gebeunhaasd wordt. Wie een vreemde taal goed spreekt is daarmee nog niet noodzakelijk een goede vertaler, maar er is voor zo iemand geen enkele barrière om zich te afficheren als vertaler.
- Ideologisch: opheffen van taalbarrières. Wie in een groot taalgebied woont (bv in de Verenigde Staten) heeft gemakkelijk toegang tot grote hoeveelheden informatie. Wie in een klein taalgebied woont (bv in Nederland of Griekenland), heeft alleen toegang tot al die informatie wanneer hij een vreemde taal heeft geleerd. Ook is de informatie in de eigen taal niet toegankelijk voor mensen van buiten die taalgemeenschap. Politici, en ook burgers, zijn vaak van mening dat iedereen, ongeacht zijn taal of opleiding, gelijke toegang moet hebben tot alle informatie die er maar is. Vertaalsystemen kunnen hierbij in elk geval het probleem van de taalbarrières oplossen.
- Nieuwsgierigheid: hoe doe je het? Een heel ander soort overweging komt voort uit wetenschappelijke nieuwsgierigheid. Vertalen door de mens is een uiterst complex en nog nauwelijks begrepen proces, en dat maakt het tot een uiterst boeiende onderneming om na te gaan hoe je ditzelfde proces (of een ander proces dat tot hetzelfde resultaat leidt) met een computer kunt uitvoeren.

Ongeacht het achterliggende motief, het probleem blijft moeilijk en interessant, en men dient zich steeds voor ogen te houden dat het helemaal niet zeker is dat er echt een enkele oplossing voor dit hele probleem bestaat. De vraag dient dan ook niet steeds te zijn: ‘hebben we het vertaalprobleem opgelost’, maar liever ‘hoeveel dichter kunnen we bij een oplossing komen’.

3 Geschiedenis

De geschiedenis van het automatisch vertalen, kun je indelen in drie periodes:

- 1946-1966: Koude oorlog
- 1978-1992: Internationalisatie
- 1998-????: Internet

Alle drie de perioden kun je beschrijven aan de hand van drie voorwaarden die (als er aan voldaan is) een goed startpunt vormen voor wetenschappelijke vooruitgang op enig terrein:

- Een (militair of maatschappelijk) probleem
- Een technologische ontwikkeling
- Een financieringsbron

3.1 Periode 1946-1966

Het probleem van deze tijd was de Koude oorlog. De Amerikanen wilden voortdurend op de hoogte blijven van de bewegingen van de Sovjet Unie, op militair, technologisch en wetenschappelijk terrein. Daartoe moesten grote hoeveelheden Russische documenten vertaald worden, en gezien het volume, lag het voor de hand daar de computer bij in te schakelen.

Voor en vooral gedurende de tweede wereldoorlog was er grote vooruitgang geboekt op het gebied van cryptografie (coderen en decoderen van militaire berichten), signaalverwerking (radio, televisie, telegrafie, telefonie), en de opkomst in dezelfde periode van de computer gaf aanleiding tot een groot optimisme om het vertaalprobleem realistisch aan te kunnen pakken.

De grote financieringsbron voor al deze activiteiten was het Amerikaanse ministerie van defensie, dat gedurende die hele periode een van de grote financiers van de hele Amerikaanse wetenschap is geweest.

Ondanks alle inspanningen was het resultaat met betrekking tot het automatisch vertalen uiterst teleurstellend, culminerend in het verschijnen in 1966 van het zogeheten "ALPAC Rapport", waarin door een belangrijke adviescommissie van de Amerikaanse overheid geconcludeerd werd dat het op dat moment geen zin had verder te investeren in onderzoek en ontwikkeling op het gebied van automatisch vertalen. De kosten waren hoog, en er was geen uitzicht op enig interessant resultaat.

3.2 Periode 1978-1992

In de periode 1966 tot 1978 was er weinig activiteit op het gebeid van het automatisch vertalen. Enkele centra in de VS, die niet van overheidsfinanciering afhankelijk waren gingen gewoon door, en ook in Europa waren er hier en daar centra (vooral in Frankrijk, Duitsland en Engeland) die bleven werken aan de ontwikkeling van het automatisch vertalen.

Rond 1978 kwam er een kentering, getriggerd door het opdoemen van nieuwe problemen, nieuwe technologieën, en nieuwe financieringsbronnen.

De toenemende internationalisering, zowel van handel en economie, alsook op politiek gebied, creëerde geheel nieuwe taalproblemen. Handel en industrie gingen over de grenzen werken, en ook de politieke en economische samenwerking in Europa (denk aan de groei van de Europese Unie, maar daarbuiten ook aan de NAVO, en de Verenigde Naties) leidden tot een enorme groei van informatie-uitwisseling over grenzen (en dus ook taalgrenzen) heen.

Daarbij kwam de explosieve stijging van de loonkosten op alle gebieden, en het feit dat de toename van het aantal gekwalificeerde vertalers geen gelijke tred hield met de groeiende behoefte.

Technologisch gezien werd deze periode gekenmerkt door de opkomst van snellere, grotere en vooral ook goedkopere computers, die door de opkomst van betere programmeertalen veel

beter inzetbaar werden voor het oplossen van niet-numerieke problemen dan daarvoor het geval was.

Ook binnen de taalkunde hadden zich inmiddels grote ontwikkelingen voorgedaan, die het gemakkelijker maakten taalkundige kennis en inzichten vast te leggen op een manier die zich leende voor computertoepassingen.

Omdat de taalbarrières ook op politiek niveau duidelijk gevoeld werden, waren organisaties zoals de Europese Commissie ook bereid fors te investeren in onderzoek en ontwikkeling op het gebied van het automatisch vertalen. Het politieke besluit om alle officiële talen van de lidstaten uit te roepen tot officiële werktalen, noodzaakte de EU om alle documenten in principe in elke taal beschikbaar te stellen. De hieruit voortvloeiende vertaallast bleek enorm, zowel financieel alsook praktisch, in verband met de vertragingen die dat opleverde voor de dagelijkse operaties. De Commissie besloot dan ook al in 1978 over te gaan tot de aanschaf van een bestaand vertaalsysteem (SYSTRAN), en tot het starten van een eigen vertaalproject, EUROTRA, dat in samenwerking tussen alle lidstaten uitgevoerd zou moeten worden.

Ook in Nederland bleven overheid en bedrijfsleven niet achter bij het investeren in het automatisch vertalen. Drie grote projecten werden (mede) door de Nederlandse overheid gesteund:

- EUROTRA: het vertaalproject van de Europese Commissie, dat beoogde om een hoge kwaliteit vertaalsysteem te maken tussen alle werktalen (in die tijd oplopend van 6 tot 9, en nu 11 talen)
- ROSETTA: het vertaalsysteem van het Philips Natuurkundig Laboratorium in Eindhoven
- DLT: het vertaalsysteem van het toenmalige Utrechtse softwarebedrijf BSO

Het resultaat van al deze activiteiten (zowel binnen als buiten Nederland) was in grote lijnen teleurstellend. Geen van de drie projecten leverde het resultaat op dat verwacht werd. Het EUROTRA project had veel last van zijn grote omvang (in de topperiode ca 20 centra over heel Europa, met ca 300 onderzoekers aan het werk). De benadering van het ROSETTA project was inhoudelijk en conceptueel misschien wel de mooiste in zijn soort, maar het Philips concern zag er op korte termijn geen brood in, en zette het project stop. Het DLT project timmerde erg aan de weg, en sprak vooral aanvankelijk erg tot de verbeelding van het publiek door zich te baseren op het gebruik van Esperanto, maar zag later (toen deze benadering allang weer verlaten was) geen kans financiers te vinden voor verdere ontwikkeling van het systeem tot een commercieel product.

Deze teleurstellingen leidden tot een nieuwe dip in het geloof in en financiering voor het automatisch vertalen, niet alleen in Nederland, maar in heel Europa.

Niettemin is deze periode heel belangrijk geweest voor de ontwikkeling van de Europese taaltechnologie, vooral in landen waar met name het EUROTRA project leidde tot het opzetten van onderwijs- en onderzoekscentra voor automatisch vertalen en taaltechnologie in het algemeen. Kortom: de impact was enorm, maar een oplossing voor het vertaalprobleem kwam er niet uit.

3.3 Periode 3: 1998-????

Na een vrij korte, maar wel duidelijk merkbare dip als gevolg van de vorige periode, is er, sinds 1998 sprake van een nieuwe revival van het automatisch vertalen.

Weer zijn er problemen, technologische ontwikkelingen en financieringsbronnen aan te wijzen die hiertoe geleid hebben.

Als eerste is daar de verdere globalisering van handel en industrie. Markten voor producten en diensten worden steeds internationaler, zowel binnen de EU als daarbuiten. Daarnaast is er de explosieve opkomst van het internet, dat in principe geen grenzen kent. Elektronische handel en informatie zijn binnen ieders bereik, en het bestaan van taalbarrières werpt grote blokkades op. In een land als Nederland wordt het belang hiervan gemakkelijk onderschat. Wij zijn opgevoed in een traditie waarin het kennen van een of meer andere Europese talen als noodzakelijk en vanzelfsprekend wordt voorgesteld, maar in de wat grotere taalgebieden om ons heen blijken zowel burger als overheid veel meer ingesteld op gebruik en kennis van geen andere taal dan de eigen taal. Dit, gevoegd bij de steeds sterker wordende overtuiging dat het de plicht van de overheden is om alle burgers in gelijke mate toegang te geven tot informatie, leidt er toe dat het opheffen van taalbarrières een belangrijke prioriteit geworden is. Automatisch vertalen hoeft natuurlijk niet de enige oplossing te bieden (meer aandacht voor het onderwijs in de vreemde talen is natuurlijk ook nuttig), maar het blijft een belangrijke mogelijkheid.

De technologische en wetenschappelijke ontwikkelingen spelen een wezenlijke rol. De PC is krachtiger geworden dan de grote mainframes uit de zeventiger jaren, en is door zijn lage prijs voor vrijwel iedereen binnen ons deel van de wereld bereikbaar. In de aanpak van problemen zoals het automatisch vertalen is ook een duidelijke kentering gekomen. Waar in het verleden vaak gekozen werd voor een specifiek type benadering van het probleem, is nu te zien dat er een neiging bestaat naar meer hybride aanpakken, waarbij van verschillende benaderingswijzen de beste ingrediënten worden gekozen, om die samen te smeden tot een meer effectieve en meer efficiënte methode.

Ook de visie op de taak van het automatisch vertalen is geëvolueerd. Waar in het begin het hoogste doel van onderzoekers en ontwikkelaars was het bereiken van een zo getrouw mogelijke kopie van het gedrag van de menselijke vertaler, bestaat er nu steeds meer een neiging om het automatisch vertaalprobleem als een collectie van situatiegebonden deelproblemen te zien, waarbij elke situatie zijn eigen oplossing vereist.

Financiers zijn er ook. Niet alleen de Europese Commissie is overtuigd van het belang van het wegnemen van de taalbarrières (zowel uit praktisch, politiek als uit commercieel oogpunt), maar ook zijn er talloze bedrijven die het commerciele belang inzien van maatregelen om taalbarrières te slechten. We denken hierbij o.a. aan internet- en telecombedrijven.

4 Problemen met oplossingen.

4.1 Wat helemaal niet moeilijk is

Bij het leren en begrijpen van vreemde talen zijn er belangrijke verschillen tussen mens en computer. Voor mensen vormt het gebruik van vreemde lettertekens (Grieks, Russisch, Chinees) al snel een belangrijk obstakel. Voor een computer is dit geen enkel probleem. Ieder teken is even gemakkelijk te leren of te onthouden (al moeten we niet uit het oog verliezen dat niet iedere tekstverwerker met dit soort problemen even gemakkelijk omgaat). Ook de lees- en schrijfrichting (van links naar rechts, van rechts naar links, van boven naar onder) is op geen enkele wijze een probleem. Wij moeten er zelf erg aan wennen, maar de computer heeft daar geen probleem mee.

Ook het voorkomen van veel onregelmatigheden (Latijnse, Griekse, of Franse onregelmatige werkwoorden, om maar wat jeugdtrauma's noemen) vormen op geen enkele wijze een probleem. De computer leert snel en heeft een feilloos en onbegrensd geheugen, en laat zich niet in de war brengen. Grote hoeveelheden woorden vormen evenmin een probleem. Een computer vindt het leren van een miljoen woorden niet gemakkelijker of moeilijker dan het leren van duizend woorden, zelfs als ze veel op elkaar lijken of vol onregelmatigheden zitten (denk aan de Duitse meervouden, of Engelse werkwoorden).

4.2 Wat wel moeilijk is

Als wij een zin lezen in de eigen taal of in een taal die wij goed kennen, kost het ons meestal geen moeite om te begrijpen wat de zin betekent, welke woorden er in voorkomen, welke betekenis ze hebben, en hoe de zin grammaticaal in elkaar zit.

Hoe moeilijk het is om vervolgens te bepalen hoe de zin vertaald moet worden, hangt voornamelijk af van onze kennis van de taal waar we naar toe moeten vertalen.

Voor de computer ziet dit beeld er totaal anders uit. We zullen dit in het onderstaande illustreren. We bekijken hiertoe eerst hoe het ons zou vergaan wanner we met behulp van enkel een woordenboek vanuit een totaal onbekende taal naar het Nederlands toe moeten vertalen.

4.2.1 Vertalen met een woordenboek

We bekijken de volgende zin (die misschien wat geforceerd aandoet, maar die we goed kunnen gebruiken om het probleem te illustreren):

Ik was de was weer aan het wassen

Wanneer we de verschillende woorden in het woordenboek opslaan, vinden we het volgende (we vermelden steeds het woord, de woordsoort [werkwoord, zelfstanding naamwoord, etc], het aantal vermelde betekenissen, en in telegramstijl de betekenis zoals we die in een niet al te uitgebreid woordenboek aantreffen):

- ***ik (2)****: "ik", "het ik"*
- ***was (zn) (5):*** *"wasproces", "bijenwas", "boenwas", "stijging", "wasgoed"*
  was (ww) (7): *"zijn", "kleren schoonmaken", "de afwas doen", "erts wassen", "dieren wassen", "kaarten schudden", "groeien"*
- ***de (2)****: "de fiets", "een gulden de meter"*
- ***was (12)****: zie boven*
- ***weer (werkwoord) (1):*** *"afweren"*
  weer (zelfst. naamwoord) (4): *"hamel", "weersgesteldheid", "afweer", "keerdam"*
  weer (bijwoord) (1): *"wederom"*
- ***aan:*** *meer dan* ***10*** *betekenissen*
- ***het: (3):*** *"het huis", "hij/zij/het", "appels voor een gulden het stuk"*
- ***wassen (12):*** *zie boven*

Het resultaat is dat we op basis van alleen het opzoeken van de woorden in een woordenboek al zo'n 2 x 12 x 2 x12 x 6 x 10 x 3 x 12 = 1 244 160 mogelijke keuzes hebben voor hoe de zin qua gebruikte woorden in elkaar zit.

Een explosie aan mogelijkheden dus, en het mag eigenlijk een wonder heten dat we hier in het dagelijkse leven zo weinig last van lijken te hebben.

4.2.2 De taalkunde als redder

Met een beetje oppervlakkige taalkundige kennis, kunnen we al flink in de mogelijkheden gaan snoeien. We weten uit ervaring (en dat kunnen we gelukkig ook vrij gemakkelijk aan een computer duidelijk maken) dat niet alle woordopeenvolgingen mogelijk zijn, zelfs als je alleen maar kijkt naar de woordsoorten (lidwoorden , zelfstandige naamwoorden, werkwoorden, etc).

We zullen dit hier niet in detail nagaan, maar een simpele beschouwing op deze basis reduceert het aantal keuzemogelijkheden voor de woorden in onze voorbeeldzin al aanzienlijk (we geven steeds aan hoeveel van de eerdergenoemde mogelijkheden er overblijven op grond van wat er na elkaar kan voorkomen):

woord	**woordsoort**	**mogelijkheden eerst**		**mogelijkheden op basis van woordsoorten**
ik	(pers. vnwd)	2	=>	1
was	(werkwoord)	12	=>	7
de	(lidwoord)	2	=>	1
was	(zelfst. nwd)	12	=>	5
weer	(bijwoord)	6	=>	1
aan	(voorzetsel)	10	=>	5
het	(lidwoord)	3	=>	1
wassen	(werkwoord)	12	=>	6

Het resultaat is dat er nu nog maar 1050 mogelijkheden over zijn. Een reductie van het probleem met een factor duizend dus – maar helaas nog niet genoeg.

4.2.3 Nog wat taalkunde

Onze taalkundige kennis (die we ook vrij gemakkelijk op computers kunnen overbrengen) stelt ons in staat het aantal mogelijke interpretaties van de zin nog verder te reduceren. Ter illustratie:

- *ik was* kan hier alleen van *zijn* komen (niet van schoonmaken of groeien)
- *de was* kan nog steeds 5 betekenissen hebben
- *aan* kan alleen van *aan het --- zijn* komen
- *het wassen* kan nog op 5 soorten van wassen slaan (maar niet op *groeien,* omdat daar geen lijdend voorwerp bij kan)

Dit brengt het aantal mogelijkheden terug tot 25. Maar er is nog meer mogelijk. We kunnen in het woordenboek nog wat extra informatie toevoegen, zoals:

- bij zelfstandige naamwoorden: *mens, dier, instrument, vloeibaar, delfstof, voertuig, abstract, telbaar, ...*
- bij werkwoorden: *onderwerp moet mens zijn, lijdend voorwerp vloeibaar, er moet een tijdsbepaling bij, een plaatsbepaling, ...*
- bij voorzetsels: *met een zelfstandig naamwoord dat een tijd aanduidt is het een tijdsbepaling, met plaats een plaatsbepaling, ..., etc*

Eigenlijk gaat het hier om verkapte betekenisinformatie, maar we kunnen dat (ondanks dat we eigenlijk niet zo goed weten wat betekenis eigenlijk is) heel goed toevoegen alsof het om objectieve taalkundige kennis gaat, net als woordsoorten. Ter illustratie:

Jan kocht bloemen voor

- *... Marie*
- *... half zeven*
- *... zijn laatste geld*
- *... moederdag*
- *... de ingang van het CS*

wat *voor* hier betekent is vast nog wel op te lossen door extra woordenboekinformatie, bijvoorbeeld door vast te leggen dat *half zeven* een tijdsbepaling is, *de ingang van het CS* een plaatsbepaling, etc.

4.2.4 Het resultaat

Met de extra informatie kunnen we dan de laatste problemen wel uit de weg ruimen:

- *was* hoort niet tot de categorie *serviesgoed, dier, erts* of *kaartspel*
- *bijenwas* behoort niet tot de categorie zaken die je kunt wassen, dus we houden nog maar een mogelijkheid over

Voorlopige conclusie

- met een woordenboek alleen is het niet mogelijk uit te maken welke zin je feitelijk moet vertalen (te veel mogelijkheden)
- de grammatica (woordsoorten en zinsdelen) beperkt het aantal problemen
- en wat betekenisinformatie er bij geeft het laatste zetje

De tussenstand: we hebben nu een (impressionistisch) beeld geschetst van een proces

- dat ons in staat stelt vast te stellen wat eigenlijk de zin is die we zouden willen vertalen
- dat berust op objectief beschrijfbare kennis (woordenboeken, grammaticaregels)
- dat een noodzakelijke eerste stap is op weg naar het vertalen
- dat redelijk goed te automatiseren lijkt (objectief vast te stellen, goed te coderen in woordenboeken of grammatica's)

5 Basisingrediënten van een vertaalsysteem

We kunnen nu gaan kijken hoe een vertaalsysteem in principe opgebouwd zou kunnen zijn. We pretenderen niet dat dit de enig mogelijke oplossing is, maar als je kijkt naar state-of-the-art vertaalsystemen, dan kom je de onderstaande ingrediënten altijd wel in de een of andere vorm tegen.

5.1 De basisbouwstenen

De taal van de ter vertaling aangeboden tekst noemen we *brontaal*, de taal waarin de tekst vertaald moet worden de *doeltaal*. We zien in een vertaalsysteem meestal de volgende basisbouwstenen:

- grammaticale analyseregels die je in staat stellen de brontaalzinnen te ontleden of *analyseren*
- vertaalregels: ze vertalen geanalyseerde brontaalzinnen in doeltaal-halffabrikaten (de grammaticale hoofdstructuur is correct, maar de vorm en de volgorde van woorden en zinsdelen hoeven nog niet in orde te zijn)
- regels die uit de bovengenoemde doeltaal-halffabrikaten correcte doeltaalzinnen produceren (syntheseregels)
- woordenboeken (in brontaal, in doeltaal, en tweetalig), met veel extra gegevens over woordsoort, geslacht, verbuiging, vervoeging, etc
- computerprogramma's die deze kennis toepassen; we gaan op dit onderdeel hier niet in, maar we vermelden slechts dat de taak van de computer hier goed te vergelijken is met die van een middelbare scholier die begint met het vertalen van Griekse of Latijnse teksten: gewapend met de kennis van de regels ga je op zoek naar het werkwoord, dan naar het onderwerp, en vervolgens naar de zinsdelen die daar nog bij horen (lijdend voorwerp, bepalingen, etc); hiervoor bestaan heel adequate computertechnieken; als de grammatica goed is, zal de computer er in het algemeen geen moeite mee hebben om de grammaticale kennis correct toe te passen.

5.2 Drie strategieën

Globaal gezien kunnen we in bestaande vertaalsystemen drie verschillende strategieën aantreffen. Soms zijn ze geheel of gedeeltelijk gecombineerd, maar in principe komen ze op het volgende neer:

5.2.1 Directe systemen

Kenmerk van deze systemen (historisch de oudste) is een vergaande vervlechting tussen allerlei soorten kennis. Ze dateren uit een periode dat het programmeren van de computer het belangrijkste probleem was, en dat het nog niet in zwang was om een scheiding aan te

brengen tussen kennis (geformuleerd in een computeronafhankelijk formalisme) en de programma's die deze kennis moesten toepassen. Kenmerk van deze benadering is dat analyse-, vertaal- en syntheseregels in elkaar verweven zitten in één enkele regelcomponent, en er zijn geen aanwijsbare tussenstadia in het vertaalproces.

Een belangrijk nadeel van deze benadering is dat hij erg ingewikkeld is en dus foutgevoelig. Daarnaast kan hij tot veel overbodig werk leiden. Als je een vertaalsysteem maakt vanuit één taal naar meer talen doe je steeds hetzelfde (je zult toch altijd wel moeten vaststellen wat het onderwerp van de zin is, of je nu naar het Frans, Duits of Grieks toe wilt), maar net even anders (namelijk direct gekoppeld aan de vertaalregels naar de desbetreffende talen). Dat brengt een hoop duplicatie van werk met zich mee, en ieder nieuw inzicht in bv de grammatica van het Nederlands moet in alle vertaalcomponenten afzonderlijk ingecodeerd worden

5.2.2 Transfersystemen

Deze systemen voeren het vertaalproces uit in drie stappen:
- analyseer de aangeboden zin of tekst (in termen van een syntactische of semantische abstracte representatie, bv een ontleding in zinsdelen met hun functies)
- vertaal de uit de analyse verkregen representatie van de brontaal in een doeltaalrepresentatie (bijvoorbeeld ook een ontledingsstruktuur)
- synthetiseer uit de doeltaalrepresentatie de juiste zinnen

Voordeel: je doet analyse en synthese meer een keer per taal (of je nu naar het Russisch, Chinees of Swahili toe wilt), maar je hebt wel veel vertaalcomponenten (een voor elk talenpaar)

5.2.3 Tussentaalsystemen (ook wel: interlingua-systemen)

Hier gebruiken we maar twee stappen:
- analyse, resulterend in een taalonafhankelijke betekenisrepresentatie
- synthese vanuit deze representatie, resulterend in de uiteindelijke doeltaaltekst

Dit is veruit de elegantste methode. Je hebt voor elke taal precies een analysecomponent nodig, en een synthesecomponent. Toevoeging van een nieuwe taal vergt ook niet meer dan toevoeging van deze twee componenten voor de nieuwe taal. Ideaal dus.

Helaas heeft deze benadering ook een grote zwakte: tot dusverre is niemand in staat geweest een geschikte tussentaal (*interlingua*) te construeren die het mogelijk maakt met deze methode te werken. Het is verleidelijk om hierbij te denken aan bijvoorbeeld logische systemen, of kunsttalen zoals het Esperanto, maar helaas blijkt dat deze voorstellen meer problemen creëren dan oplossen. De wezenlijke problemen van het automatisch vertalen (zoals we die in de volgende secties zullen demonstreren) worden er niet door opgelost.

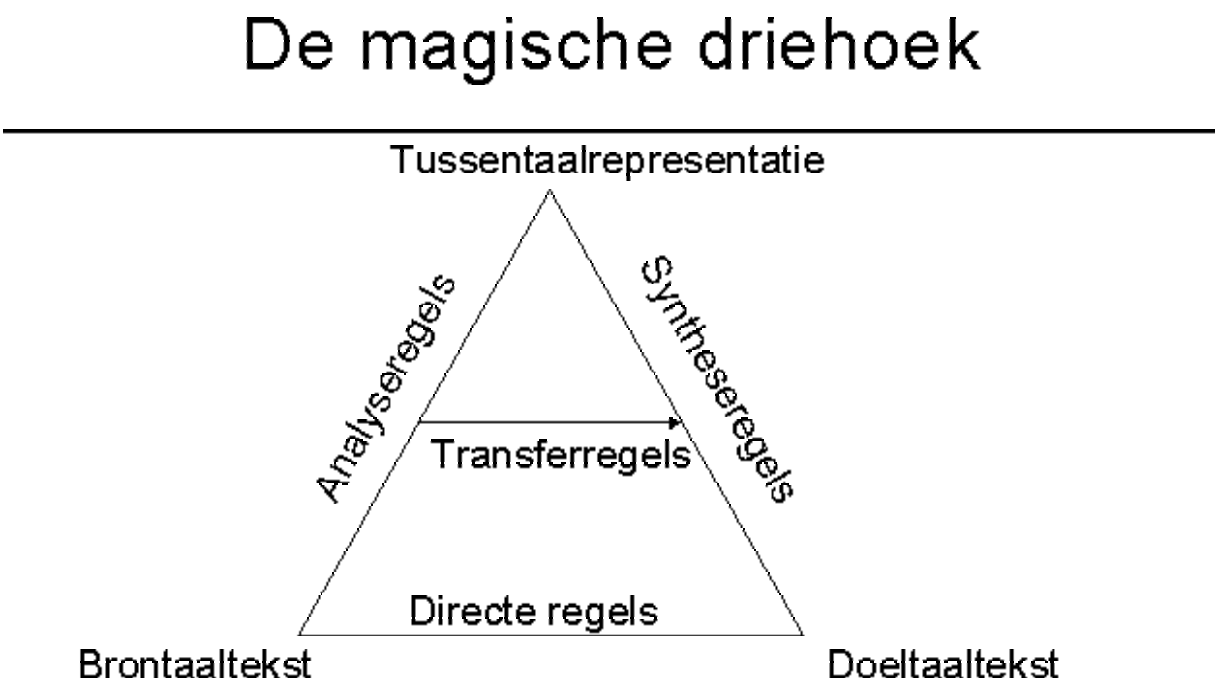
5.3 Kosten en gebruik van de drie typen

Als je kijkt naar het aantal componenten dat bij ieder van de benaderingen nodig is, kun je de volgende rekensom maken:

- Directe systemen: bij 12 talen 12 x1 1 complexe componenten
- Transfersystemen: 12 + 12 + 12 x 11 componenten
- Tussentaalsystemen: 12 + 12 componenten

Het aantal componenten op zich hoeft natuurlijk niet doorslaggevend te zijn. Minstens even belangrijk is de complexiteit van elke component afzonderlijk, de kosten van het onderhoud (zowel voor correctie als uitbreiding), en de kosten van het toevoegen van nieuwe talen. Vroeger waren directe systemen het meest in zwang. Daarna heeft men erg veel energie gestoken in het ontwerpen van tussentaalsystemen, maar op dit moment voeren de transfersystemen de boventoon.

In de onderstaande figuur geven we de drie strategieën schematisch weer.



6 Problemen zonder oplossingen

We hebben het hierboven gehad over problemen die geen echte problemen waren (vreemde lettertekens, onregelmatigheden), en over problemen die we met wat taalkunde wel konden oplossen. We gaan nu kijken naar de echte problemen, waar eigenlijk niemand nog een echt goede oplossing voor heeft. De vier voornaamste zijn:

- Ambiguïteit
- De mismatch tussen de talen
- Het compromis van het vertalen
- Robuustheid

Hieronder leggen we ze uit.

6.1 Ambiguïteit

We spreken van ambiguïteit wanneer een woord of zin verschillende mogelijke interpretaties heeft. We noemen hier vier hoofdsoorten:

- Woordambiguïteit
- Aanhechtingsambiguïteit
- Relatieambiguïteit
- Verwijzingsambiguïteit

6.1.1 Woordambiguïteit

We spreken van woordambiguïteit wanneer woorden van dezelfde grammaticale categorie (werkwoorden, zelfstandige naamwoorden, etc) verschillende betekenis kunnen hebben. Voorbeelden:

- *was*: bijenwas of vuil goed
- *bank*: zitbank of geldbank
- *paard*: rij-, gym- of schaakpaard

Mensen hebben zelden last van dit probleem, omdat ze over het algemeen begrijpen waar een tekst over gaat. Voor vertaalsystemen is dit een stuk moeilijker, maar het is wel noodzakelijk om dit probleem op te lossen, omdat in de regel de vertaling van een woord afhankelijk is van zijn betekenis.

Twee methoden om dit probleem te lijf te gaan zijn:

- Tevoren vaststellen (en aan het vertaalsysteem vertellen) dat een tekst over een bepaald onderwerpsgebied (domein) gaat
- Statistische methoden gebruiken om de meest waarschijnlijke interpretatie te kiezen.

Geen van beide methoden is feilloos: in een tekst over schaken waarin verteld wordt dat Karpov om zijn tegenstander te intimideren te paard naar het schaaktoernooi kwam, zullen beide methoden waarschijnlijk niet de juiste keuze maken.

6.1.2 Aanhechtingsambiguïteit

We spreken van aanhechtingsambiguïteit wanneer een zinsdeel op verschillende plaatsen in de zin aan een ander zinsdeel gehecht kan worden. Voorbeelden:

- *Ik keek naar de hond met de verrekijker*
- *Ik keek naar de hond met de lange staart*

of:

- *De toespraak van de minister van gisteren*
- *De toespraak van de minister van verkeer*

Ook hier geldt dat wij er zelden moeite mee hebben, maar dat een vertaalsysteem dit wel moeten weten om goed te vertalen, maar het niet uit zijn taalkundige kennis kan afleiden.

Drie mogelijke remedies (geen van alle perfect) zijn:

- extra informatie in het woordenboek (bv verrekijker is een kijkinstrument, honden hebben staarten)
- vaste strategieën (bv hecht altijd aan dichtstbijzijnde zinsdeel)
- statistische benaderingen (meet over een grote hoeveelheid teksten hoe het daar in de praktijk meestal gaat)

6.1.3 Relatieambiguïteit

We spreken van relatie-ambiguïteit wanneer we wel weten welke zinsdelen bij elkaar horen, maar niet hoe ze bij elkaar horen, d.w.z. welke rol ze ten opzichte van elkaar vervullen. Weer wat voorbeelden ter illustratie:

- *Jan kocht bloemen voor half zeven (tijdsbepaling)*
- *Jan kocht bloemen voor het CS (plaatsbepaling)*
- *Jan kocht bloemen voor een tientje (waardebepaling)*

Nog veel interessanter (en in het Nederlands heel populair) zijn de samengestelde woorden:

- *tarwemeel, (meel van tarwe)*
- *vismeel, (meel van vis)*
- *aardappelmeel, (meel van aardappelen)*
- *pannenkoekenmeel, (meel om pannenkoeken mee te bereiden)*
- *kindermeel, (meel voor kinderen)*

Voor vertaling naar bv het Duits, dat dezelfde mogelijkheden heeft voor het bouwen van samengestelde woorden als het Nederlands, levert dit geen problemen op. Wel als je bv naar het Frans gaat, waar de meeste samengestelde woorden worden weergegeven met voorzetselconstructies: *wasmachine / machine à laver* ; het voorzetsel is afhankelijk van de relatie die er tussen beide woorden bestaat, en die relatie moet je dus kennen.

Ook hier geldt dat het woordenboek en de statistiek de meest gebruikte (maar lang niet altijd goed werkende) oplossingsmethoden zijn. Je kunt bv wel veel samengestelde woorden met

juiste vertaling en al in het woordenboek opnemen, maar helaas is in het Nederlands de taalgebruiker vrij om nieuwe samengestelde woorden te bouwen, die door iedereen onmiddellijk begrepen worden.

6.1.4 Verwijzingsambiguïteiten

Woorden zoals persoonlijke voornaamwoorden ontlenen hun betekenis niet aan wat ze zelf betekenen, maar aan andere eenheden in een zin of tekst waarnaar ze verwijzen. Ook hier geldt dat de mens er zelden moeite mee heeft, maar het vertaalsysteem des te meer. Voorbeelden:

- *"De politieagenten vuurden op de demonstrerende verpleegsters omdat ze revolutie wilden"*
- *"De politieagenten vuurden op de demonstrerende verpleegsters omdat ze revolutie vreesden"*

De taalkunde kan hier de juiste beslissing niet nemen. Alleen kennis van hoe het in onze maatschappij normaal toegaat stelt ons in staat de juiste keuze te maken.

Een ander voorbeeld van hetzelfde:
(1) De soldaten schoten op de vrouwen.
(2) Ze vielen dood neer.
(2') Ze ontlaadden hun geweren.

Zin 1 kan gevolgd worden door zin 2 of zin 2'. Grammaticaal is er tussen de beide vervolgzinnen niet zo'n groot verschil, maar in het eerste geval slaat *ze* terug op de vrouwen, en in het tweede geval op de soldaten.

Binnen de taalkunde hebben we eigenlijk geen middelen om dit probleem aan te pakken. Ook het woordenboek biedt ons geen goede oplossing, en statistische methoden evenmin. Toch is een oplossing van deze problemen noodzakelijk om bv goed naar het Frans te kunnen vertalen.

Alleen een benadering waarbij we in staat zijn om onze kennis van hoe het in onze wereld toegaat in te computer onder te brengen zou uitkomst kunnen bieden, maar helaas bestaan daar op dit moment geen algemeen bruikbare voorstellen voor. Alleen bij het werken binnen beperkte domeinen, waar vastligt wat er wel en niet kan gebeuren, kunnen deze problemen op dit ogenblik aangepakt worden.

6.1.5 Kennis van de wereld

Bij een aantal ven de bovengenoemde problemen lijkt een oplossing gevonden te kunnen worden wanneer we bij het vertalen de computer gebruik kunnen laten maken van wereldkennis: inzicht in wat er zoal kan gebeuren en hoe bepaalde gebeurtenissen samenhangen.

We worden hierbij geconfronteerd met drie grote problemen:

- hoe verzamel je die kennis
- hoe leg je die vast
- hoe consulteer je die

We hebben daar, zoals gezegd, geen goede oplossingen voor. Voor de menselijke spreker of vertaler vormen deze zaken nauwelijks een probleem:

- hij beschikt over veel van die kennis
- hij weet waar en hoe te zoeken wanneer hij deze kennis niet paraat heeft

6.2 De mismatch tussen de talen

Talen zeggen niet alles op dezelfde manier. Waar de ene taal een enkel woord gebruikt, gebruikt de andere een hele constructie. Waar de ene taal iets als een concept beschouwt en er dus ook maar een woord voor heeft, beschouwt een andere taal iets als twee afzonderlijke concepten, met verschillende woorden. Waar de ene taal iets met een bijwoord uitdrukt, gebruikt de andere er een bijzin voor. Wat voorbeeldjes:

- *schimmel (een woord) / grey horse (constructie)*
- *runway (een concept) / landingsbaan, startbaan (twee concepten)*
- *ik zwem graag (zin met bijwoord) / I like to swim (zin met bijzin)*
- *er werd gedanst (passieve vorm) / on dansait (actieve vorm)*

Deze problemen zijn van een andere orde dan de bovengenoemde, want we kunnen ze stuk voor stuk formuleren en oplossen, maar ze kunnen (vooral wanner ze in combinatie voorkomen) leiden tot moeilijk controleerbare interactie tussen regels, en ook problemen opleveren die niet gemakkelijk door het vertaalsysteem zelf te detecteren zijn (denk aan "*de piloot veroorzaakte een ongeluk doordat hij de startbaan aanzag voor de landingsbaan*").

6.3 Het compromis van het vertalen

Vertalen houdt in dat een tekst wordt omgezet van de ene taal in een andere taal. Er verandert dus nogal wat aan de tekst, maar er moet ook iets behouden blijven. Niet alleen de betekenis, maar ook allerlei andere eigenschappen. We noemen er hier een paar ter illustratie:

- *betekenis* (maar welke: logisch gezien betekent 1+1=2 hetzelfde als 2=1+1, maar zou je dit ooit willen veranderen als je van Nederlands naar Engels gaat?)
- *boodschap* (maar zou je 'wilt u alstublieft stil zijn' vertaald willen zien als 'koppen dicht'?)
- *(on)waarheid* (maar zou je een aperte onwaarheid in een tekst, bv 1+1=3 moeten laten staan?)
- *stijl* (maar moet je in een vertaling vanuit bv het Japans de hoogdravende stijl waarmee collega's communiceren moeten vervangen door de stijl waarmee dezelfde collega's hier met elkaar zouden communiceren?)

- *effect* (moet je vertalen wat mensen tegen elkaar zeggen of moet je een vertaling zoeken die hetzelfde effect teweeg zou brengen?)
- *vaagheid* (moet je bij het vertalen een vaagheid wegnemen, of zou je deze, bv in politieke teksten, juist moeten behouden?)
- *humor* (moet je elk grapje of woordgrapje overnemen in de vertaling?)
- *compactheid* (moet de vertaling net zo compact zijn als het origineel, of mag je sommige dingen met wat meer omhaal vertalen, bijvoorbeeld wanneer je het woord voordeurdeler of Melkertbaan moet vertalen naar een cultuur waar dat niet bestaat)
- *discriminatie* (moet je discriminerende formuleringen vertalen door minder discriminerende varianten, of juist niet)

Zoals uit het bovenstaande moge blijken, is het ambacht van de vertaler verre van gemakkelijk. Voortdurend wordt hij geconfronteerd met keuzes, en zelfs als hij de keuzes weet te maken, is het nog niet zeker dat hij een tekst kan maken die aan alle keuzes recht doet. Een alledaags voorbeeld zijn de woordgrapjes die in de ondertiteling verdwenen blijken te zijn, omdat het eenvoudigweg niet mogelijk is om die naar net Nederlands om te zetten.

De vertalershandboeken staan vol met beschouwingen over dit soort zaken, en geven tal van aanwijzingen over hoe met de verschillende problemen om te gaan, maar helaas is onze kennis hierover nog verre van volledig (vertalen lijkt in dat opzicht meer een kunst dan een ambacht), en de kennis die er is, is niet beschikbaar op een manier die zich leent voor formalisatie in een regelsysteem dat voor de computer bruikbaar zou kunnen zijn.

6.4 Robuustheid

Onder robuustheid van een vertaalsysteem verstaan we het vermogen van systemen om te reageren op onverwachte input. Deze onverwachte input kan bestaan uit nieuwe woorden, die (al dan niet toevallig) nog niet in het woordenboek van het systeem zitten, maar ook uit het gebruik van grammaticaregels die wel correct zijn, maar waarin de systeemmaker nog niet voorzien had. Dit kan heel gemakkelijk gebeuren, omdat er voor geen enkele taal een ook maar bij benadering volledige grammatica geschreven is.

Daarnaast kan de input van een systeem ook bestaan uit zinnen of woorden die gewoon fout zijn: verkeerde woorden, verbuigingen, regels, combinaties, of wat dan ook.

Een in de praktijk bruikbaar systeem moet met deze onverwachte gebeurtenissen om kunnen gaan, net zo goed als je mag verwachten dat een vliegtuig niet bij elke storing meteen neerstort. In de praktijk gebruikte remedies zijn:

- fouten vooraf uitfilteren
- interactie met gebruiker
- statistisch verantwoord gokken

Geen van de methoden biedt een 100% garantie, maar in bepaalde situaties kunnen ze gedeeltelijke oplossingen bieden.

7 Waar staan we nu

We kunnen de stand van zaken binnen het automatisch vertalen als volg samenvatten:

- Onze kennis schiet nog in alle opzichten tekort voor het maken van goede, volautomatische vertaalsystemen die net zo betrouwbaar zijn als bv de calculator
- De huidige vertaalsystemen zijn meestal gebaseerd op woordenboeken, grammaticale kennis, en gebruiken aanvullende statistische gegevens om problemen op te lossen
- De kwaliteit is over het algemeen uiterst matig (minder dan VWO)
- De grootste doorbraak tot nu toe is het toevoegen van statistische methoden voor het maken van keuzes geweest

Van de kwaliteit van state-of-the-art vertaalsystemen kun je op dit moment een goede indruk krijgen door op het World-wide Web te kijken naar de vertaalfaciliteit van bv de zoekmachine AltaVista: http://www.altavista.com. De kwaliteit die daaruit komt is vaak heel aardig, maar soms ook echt ridicuul. De kwaliteit is goed genoeg om een idee te krijgen waar het over gaat, maar absoluut niet voor zakenbrieven of handleidingen, of in het algemeen informatie die je naar een breder publiek zou willen verspreiden.

De conclusie dat automatisch vertalen dus nutteloos en hopeloos is, ligt dan ook erg voor de hand. Het is echter geen juiste conclusie. Vanuit het oogpunt van de onderzoeker is het natuurlijk heel vervelend dat we nog niet echt weten hoe het moet (maar het houdt hem wel nog even van de straat), en ook een gewone burger die een Japanse brief wil schrijven of een IJslandse krant wil lezen heeft er eigenlijk niet zo veel aan. In een bedrijfsmatige omgeving ligt het echter heel anders: daar zijn vragen zoals *'is de vertaling goed'* of '*hebben we het vertaalprobleem eindelijk begrepen'* volstrekt irrelevant. Daar bestaat maar een enkel criterium dat alles bepaalt: het geld. De vraag die men zich daar stelt is dan ook: *'kunnen we door gebruikmaking van automatisch vertalen onze kosten reduceren of onze winst vergroten'*. Met name in grote vertaalbedrijven of in bedrijven die over eigen vertaaldiensten beschikken blijkt dat kostenbesparingen van 40% moeiteloos te realiseren zijn door teksten eerst door een vertaalsysteem (imperfect) te laten vertalen, om ze vervolgens door een menselijke vertaler te laten corrigeren. In vergelijking met volledig menselijke vertaling (waar in een professionele omgeving elke vertaling ook alltijd door een revisor gecontroleerd en gecorrigeerd wordt, is het eerste deel van het traject (de machinale vertaling) dramatisch veel goedkoper, en het tweede deel (de correctie) een stukje duurder (door de slechtere kwaliteit). Maar op het gehele traject (inclusief de eerste investering in de aanschaf en aanpassing van het vertaalsysteem wordt een aanzienlijke besparing bereikt.

8 Waar moeten/gaan we naar toe?

Na de sombere beschouwingen in het voorafgaande kun je je de vraag stellen of automatisch vertalen überhaupt mogelijk is. Die vraag wil ik hier niet stellen. Niet alleen omdat ik het antwoord niet weet, maar ook omdat het een improductieve, filosofische vraag is, waarvan ook het kennen van het antwoord je geen stap verder brengt.

Een van de belangrijke redenen waarom het automatisch vertalen vaak als een mislukte en tot mislukken gedoemde onderneming gezien wordt, is het feit dat men zich lang heeft blindgestaard op het idee dat een goed vertaalsysteem een perfecte (en liefst zelfs verbeterde) afspiegeling van de menselijke vertaler zou moeten zijn. Maar waarom zouden we dat eigenlijk willen?

We hebben hierboven al gezien dat ook met de huidige vertaalkwaliteit al geld verdiend of bespaard kan worden, en financiele motieven spelen in onze wereld een zodanige rol dat je eigenlijk gewoon van een commercieel succes kunt spreken, al blijven de voordelen er van beperkt tot een vrij kleine kring van gebruikers.

Voor een normale burger die niet 24 uur per dag gedreven wordt door economische motieven zien het probleem en dus ook de mogelijke oplossingen er heel anders uit: de enige reden waarom wij automatische vertaalsystemen willen hebben is dat we taalbarrières willen doorbreken, en er is geen enkele reden om aan te nemen dat de beste benadering daarvoor is het maken van imitaties van de menselijke vertaler.

De vraag of automatisch vertalen mogelijk is, kunnen we daarom beter vervangen door de vraag hoe ver we kunnen komen met het verwijderen van de taalbarrières. Er is geen enkele noodzaak om met een enkel hulpmiddel in een klap alle barrières weg te nemen. Ook een verzameling verschillende hulpmiddelen die elk op hun eigen manier en in hun eigen context een deel van de barrières wegnemen brengt ons dichter bij het doel.

De hele notie ‘succes’ voor vertaalsystemen (traditioneel gemeten in termen van het aantal fouten ten opzichte van de menselijke vertaler) krijgt daardoor een geheel andere inhoud. Succes wordt niet gemeten in termen van vertaalfouten, maar in termen van het al dan niet slagen van de beoogde communicatie. Een geslaagde communicatie over een taalbarrière heen is een succes, onafhankelijk van het aantal gemaakte taal- of vertaalfouten. Revolutionair is dat overigens niet: ook wanneer wij zelf een vreemde taal spreken zullen wij ons succes daarbij meestal afmeten aan de mate waarin wij daarmee succesvol kunnen communiceren.

9 Strategieën voor de toekomst

Als we naar de toekomst kijken, dan zien we op dit ogenblik de volgende ontwikkelingen en bewegingen:

- Betere integratie van bestaande benaderingen (taalkundig, statistisch, kunstmatige intelligentie)
- Samenwerking tussen verschillende modaliteiten (taal, spraak en beeld), die ieder op zich misschien niet tot perfecte communicatie hoeven te leiden, maar dat in combinatie wel doen (als we in een Franse winkel de naam van een soort groente of een vis niet weten, lossen we dat immers ook op door er gewoon op te wijzen).
- Verdeel en heers: gespecialiseerde typen systemen voor specifieke toepassingen, domeinen en gebruikersgroepen

Van dit laatste geven we nog enkele voorbeelden:

- De vertaler wil geen vertaalsysteem (dat kan hij zelf net zo goed), maar wel een vertaalgeheugen dat onthoudt hoe hij een gelijke of gelijksoortige tekst eerder heeft vertaald, of een goed terminologiesysteem waarin hij snel de vertaling van vaktermen kan vinden
- De elektronische toerist hoeft geen perfecte vertaling, maar wil wel snel weten waar een website over gaat
- De hotelhouder wil het dagelijkse weerbericht in 6 talen op het prikbord hangen, en heeft daarvoor niet perse een vertaalsysteem nodig, maar een programma dat hem in staat stelt om eenvoudige mededelingen over het weer in veel talen weer te geven.

10 Slotopmerkingen

Automatisch vertalen is moeilijk, en heeft nog een lange weg te gaan. Een aantal fundamentele problemen is nog niet opgelost, maar de huidige kennis en technologie maken het wel mogelijk op dit ogenblik al veel locale taalbarrières te overbruggen. De notie 'succes' dient niet uitsluitend gemeten te worden in vertaalfouten, maar in termen van geld *(wat kan ik er mee besparen of verdienen)* of (meer ideeel gezien) van succesvolle communicatie over taalgrenzen heen.

Op mijn webpagina http://www-sk.let.uu.nl/ond/mt.html kun je een aantal verwijzingen vinden naar vertaalsystemen (waarvan sommige on-line), vertaalprojecten, vertaalsites, rapporten, etc. Helaas zijn de vertaalvoorzieningen voor het Nederlands gering: er bestaan wat on-line en electronische woordenboeken, waarvan sommige pretenderen vertaalsystemen te zijn. De output er van laat duidelijk zien dat hier van vertalen geen sprake is. Sinds kort biedt SYSTRAN vertaalsystemen aan van het Engels en Frans naar het Nederlands en omgekeerd. Er is op dit moment een gratis demo beschikbaar op het web. Het adres is http://www.systransoft.com Voor wie helemaal geen Engels of Frans spreekt is het natuurlijk moeilijk om de kwaliteit er van te beoordelen, maar als je een willekeurige Engelse of Franse webpagina in het Nederlands laat vertalen, kun je je wel een indruk vormen van de kwaliteit van het systeem en de bruikbaarheid er van voor je eigen doelen. We moeten hier overigens bij aantekenen dat de makers van vertaalsystemen ook altijd bereid zijn hun systemen aan de specifieke behoeften van gebruikers aan te passen (grotere en betere woordenboeken, gebaseerd op het taalgebruik binnen het bedrijf), maar dat is (net als bij de goedkope inkjetprinters en de dure inktpatronen) hun eigenlijke handel, en dus niet goedkoop.