Col: 0703.20

Date: 11 maart 2007
From: Peter-Arno Coppen <P.A.Coppen@let.ru.nl>
Subject: Col: 0703.20: Linguïstisch Miniatuurtje CXVII: De Smurf van Columbus

Linguïstisch Miniatuurtje CXVII: De Smurf van Columbus

Het lijkt erop dat nu dan toch eindelijk de oplossing in zicht is voor het zich al jaren voortslepende onderzoek naar het automatisch vertalen. Talloze onderzoeksprojecten zijn in het verleden op niets uitgelopen. Methoden uit de taalwetenschap, de informatica, de statistiek, in welke combinatie dan ook, ze hebben allemaal hun hoge verwachtingen niet kunnen waarmaken. Maar nu lijkt een Nederlandse onderzoeksgroep gestuit te zijn op het gouden idee, de linguïstische Steen der Wijzen, waarmee de fundamentele beperkingen van de bestaande methoden definitief tot het verleden lijken te behoren.

Volgens het hoofd van de International Machine Translation Group, die voor deze doorbraak verantwoordelijk is, de Tsjechische professor A. Antic, is de tijd niet ver meer dat vertaalproblemen binnen de Europese Unie tot het verleden zullen behoren. Zij schat dat binnen vijf tot tien jaar de vele miljarden die nu nog worden besteed aan het vertalen van documenten, vrijkomen voor zinvoller doelen.

Het idee is even eenvoudig als briljant. De meeste linguïstisch georiënteerde systemen uit het verleden werken met het zogeheten Analyse-Transfer-Synthesemodel. Dat is een model waarin de tekst uit de brontaal via een ingewikkeld analyseproces wordt omgezet in een abstracte betekenisrepresentatie, zeg maar een soort ontleding. Deze ontleding is nog steeds brontaalafhankelijk, maar een stuk minder dan de brontaal zelf natuurlijk. Deze abstracte betekenisrepresentatie wordt vervolgens in een Transfer-stap omgezet naar een soortgelijke representatie, maar nu een die hoort bij de doeltaal. De derde stap, de Synthese, zet ten slotte die representatie weer om in een tekst van de doeltaal.

De meest problematische stap in dit model is de Transfer-stap, omdat daar de overstap naar de andere taal plaatsvindt. Al jaren geleden zijn theoretici erachter gekomen dat de Transferstap kleiner wordt, naarmate de analyse dieper is. Hoe verfijnder je analyseert, hoe minder taalafhankelijk je representatie wordt. In het theoretische geval zou bij een oneindige diepte van analyse de Transferstap helemaal kunnen wegvallen. Dat is het idee van de *interlingua*.

De International Machine Translation Group stelt nu voor om als interlingua de zogeheten *Smurfentaal* in te zetten. Volgens professor Antic heeft de smurfentaal een aantal interessante eigenschappen die uitstekend passen bij een interlingua. Daar bestaan overigens veel misverstanden over. Wie *smurfentaal* in de Wikipedia opzoekt, vindt daar een beknopte fragmentarische uiteenzetting, én de vermelding dat een aantal jaren geleden de jongerentaal denigrerend met de term smurfentaal werd aangeduid. Antic windt zich daar zichtbaar over op: "In de Wikipedia staat dat Smurfentaal een *fictieve taal* is. Maar Smurfentaal is niet één taal, want de Franse Smurfentaal is heel anders dan Nederlandse Smurfentaal. In dat opzicht lijkt het meer een *geheimtaal* dan een *kunsttaal*." Ook de grammatica komt er in de Wikipedia bekaaid vanaf. Die zou 'subtiele diepere regels' bevatten, en voornamelijk bestaan uit de vervanging van 'onbelangrijke woorden en woorddelen' door het woord *smurf*. Antic: "Met een dergelijke half-serieuze behandeling wordt de Smurfentaal natuurlijk geen recht gedaan."

Het inzicht dat Smurfentaal iets zou kunnen betekenen voor het automatisch vertalen brak bij Antic op een onverwacht moment door: "Ik was mijn haren aan het wassen en opeens zag ik het. Ik had net mijn kinderen een smurfenverhaaltje voorgelezen, en ik realiseerde me dat ik *intuïties* had over de kwaliteit van de smurfentaalzinnetjes. Zo vond ik de vertaling *Smurfe koppen* voor *Koppige koppen* slecht. Ik dacht: dat moet *Smurfige koppen* zijn. Maar waarom? Omdat in Smurfentaal alleen *vrije morfemen* worden vervangen door het woord *smurf*. De gebonden morfemen (suffixen, prefixen) blijven onaangetast. Maar die vrije morfemen, dat zijn precies de problematische elementen bij de automatische vertaling!"

"De volgende morgen heb ik meteen een onderzoeksgroep op deze kwestie gezet. Al snel bleek dat de conventionele Smurfentaal een beetje halfslachtig was. Zo lijkt het erop dat je per *informatie-eenheid* maar één morfeem kunt vervangen door *smurf*. Dus wel *smurfentaart* of *appelsmurf*, maar niet *smurfensmurf*. Dat is voor kleine kinderen -de doelgroep- blijkbaar net iets te moeilijk. Maar wij konden een stapje verder gaan."

"Zo ontwikkelden wij het *Totaalsmurfs*, waarin *alle* vrije morfemen vervangen worden door het morfeem *smurf*. Als we dat toepassen in het Analyse-Transfer-Synthesemodel, verdwijnt een groot deel van het Transfer-probleem. Immers, de omzetting van de suffixen en prefixen naar een andere taal is vrij eenvoudig. Het betreft hier betekenisarme taalelementen van een eindige woordklasse. Het echte vertaalprobleem zit in de vrije morfemen, en dat probleem verdwijnt."

Maar wordt het probleem zo niet verschoven naar de Synthese-fase? Antic: "Dat zou kunnen, maar wij hopen dit probleem in de komende jaren met Europese subsidie te kunnen onderzoeken. Op voorhand kan echter al geconstateerd worden dat dit natuurlijk ook gewenst is. Het vertaalprobleem kan alleen worden opgelost door de eindgebruiker en niet door de opsteller van de brontekst. De eindgebruiker moet immers iets met de tekst doen."

De International Machine Translation Group is niet de enige onderzoeksgroep die de Smurfentaal ontdekt heeft. Ook het Nederlandse Bureau Smurf ijvert voor de vereenvoudiging van Europese documenten. Directeur Smit: "Het zou beter zijn voor de leesbaarheid als alle Europese documenten zouden worden omgezet in Smurfentaal. Omdat Smurfentaal het interpretatieprobleem bij de lezer legt, kan die daar veel meer kanten mee op. Uit onderzoek is gebleken dat meer dan 60 procent van de Europeanen iets kunnen met een document dat vertaald is naar Smurfentaal."

Is het dan geen probleem dat iedere Europeaan Smurfentaal zou moeten leren? Smit: "Voor Smurfentaal bestaat maar een kleine leercurve. Voor kinderen is het al helemaal geen probleem. Er zijn geen 'moeilijke woorden' meer in Smurfentaal. En enkele onbegrijpelijke grammaticale eigenaardigheden vallen compleet weg. Neem bijvoorbeeld het grammaticale geslacht. In het Nederlands moet je van ieder zelfstandig naamwoord bepalen of het mannelijk, vrouwelijk of onzijdig is. Zelfs in Nederlandse Smurfentaal verdwijnt dat verschil. Een uitdrukking als *Het ei van Columbus* is in Smurfentaal niet *Het Smurf van Columbus*, maar *De Smurf van Columbus*. Het mooie is dat iedereen die twee stripboeken over de Smurfen gelezen heeft, die intuïtie al heeft. Daar heb je helemaal geen universitaire opleiding voor nodig."

Peter-Arno Coppen