Hetzelfde, maar dan anders:
Semantische overlap detecteren en gebruiken (DAESO)

Emiel Krahmer, Erwin Marsi, Walter Daelemans,
Maarten de Rijke & Jakub Zavrel[1]

Er zijn veel manieren om hetzelfde te zeggen. Vergelijk bijvoorbeeld de volgende twee openingszinnen, uit respectievelijk het NRC en de Telegraaf van 11 september 2006:

> *De 44-jarige Steve Irwin - bekend door zijn tv-programma's over dieren - stierf maandagmiddag (plaatselijke tijd) nadat hij tijdens het duiken voor de Australische noordoostkust bij Port Douglas in zijn borstkas werd gestoken door een giftige pijlstaartrog.*

> *Steve Irwin, de Australische televisiepresentator die bekend is als The Crocodile Hunter, is maandag overleden nadat hij tijdens een duikexpeditie was gestoken door een pijlstaartrog.*

Hoewel deze twee zinnen dezelfde gebeurtenis beschrijven, doen ze dit in grotendeels verschillende bewoordingen. Dit fenomeen wordt wel **semantische overlap** genoemd. Vanuit een taaltechnologisch perspectief vormt het automatisch detecteren van semantische overlap een hele uitdaging.

Neem een **information retrieval** (IR) toepassing: een gebruiker die zoekt naar informatie over *de dood van de Crocodile Hunter* wil waarschijnlijk zowel het NRC als het Telegraaf artikel lezen, hoewel de term *Crocodile Hunter* in het ene tekstfragment wel en in het andere niet voorkomt. Een ander voorbeeld kan ontleend worden aan **automatische vraag-antwoordsystemen** (QA). Stel dat een dergelijk systeem de vraag *"Op welke leeftijd is de Crocodile Hunter overleden?"* wil beantwoorden. Het antwoord op deze vraag is niet direct te vinden in één van de fragmenten, maar wel wanneer deze gecombineerd worden. Ook voor een **automatische multi-document samenvatter**, d.w.z. een systeem dat meerdere teksten over hetzelfde onderwerp kan samenvatten, zou het nuttig zijn om te weten dat de beide zinnen grotendeels hetzelfde uitdrukken. Op die manier kan voorkomen worden dat het systeem ze allebei aan de samenvatting toevoegt, wat in het algemeen tot minder redundantie zal leiden. Helemaal ideaal zou het zijn wanneer de samenvatter niet alleen kan vaststellen dat twee zinnen grofweg dezelfde informatie bevatten, maar tevens in staat zou zijn om de inhoud van de gerelateerde zinnen samen te voegen tot één nieuwe zin die als het ware beide zinnen combineert. Dit is een vorm van taalgenerering die bekend staat als **zinsfusie** (sentence fusion).

Voor deze en andere taaltechnologische toepassingen zou het dus heel nuttig zijn wanneer automatisch bepaald kon worden of, en in hoeverre, twee zinnen semantisch overlappen. Hoe dit gedaan kan worden is de centrale onderzoeksvraag van het Stevin **DAESO** project. DAESO (de afkorting staat voor *Detecting and Exploiting Semantic Overlap*) kan beschouwd worden als een *spin-off* van het Imogen-IMIX project (zie Dixit 4.1: 12-15), en loopt van 1 oktober 2006 tot 1 oktober 2009.

Het DAESO project bestaat uit drie fases. Allereerst zal een **monolinguaal parallel corpus** worden ontwikkeld (1M woorden) bestaande uit Nederlandse tekstparen die vergelijkbare informatie bevatten (denk aan verschillende nieuwsberichten die dezelfde gebeurtenis beschrijven, en aan verschillende modern-Nederlandse vertalingen van één brontekst). In eerste instantie zullen overlappende woorden en frases handmatig worden opgelijnd. Mede op basis van dit corpus zullen **software tools** voor het detecteren van semantische overlap worden ontwikkeld. Belangrijk hierbij is dat niet alleen

[1]Affiliaties: Emiel Krahmer en Erwin Marsi: Universiteit van Tilburg; Walter Daelemans: Universiteit van Antwerpen; Maarten de Rijke: Universiteit van Amsterdam; Jakub Zavrel: TextKernel.

gekeken zal worden naar welke frases uit de paralelle teksten gerelateerd zijn, maar ook naar de wijze *waarop* ze samenhangen. Zo is in het eerder genoemde voorbeeldpaar de frase *maandagmiddag (plaatselijke tijd)* in de eerste zin *specifieker* dan de corresponderende term *maandag* in de tweede zin. In de derde en laatste fase van het project, de **externe evaluatie**, zal gekeken worden of de tools daadwerkelijk bijdragen tot een verbetering van taaltechnologische applicaties (IR, QA en multi-document summarization).

De resultaten van het DAESO project zijn potentieel interessant voor organisaties die met grote hoeveelheden textuele data werken (uitgeverijen, nieuws- en persbureau's en taaltechnologiebedrijven). Wanneer u op de hoogte wilt blijven van het verloop en de resultaten van DAESO kunt u zich aanmelden voor de gebruikergroep d.m.v. een email naar: `E.J.Krahmer@uvt.nl`.