

IRME

(Identificatie en lexicale Representatie van Multiwoord-Expressies)

In het IRME-project werken UiL-OTS (Utrecht), Alfa-Informatica (Groningen) en Van Dale Lexicografie (Utrecht) samen op het gebied van de zogenaamde multiwoord expressies (MWE's).

JAN ODIJK

M

MWE's zijn woordcombinaties met eigenschappen die niet voorspelbaar zijn uit de eigenschappen van de individuele woorden of uit de normale grammaticaregels, en die daarom opgenomen moeten worden in een woordenboek. Een combinatie van woorden kan bijvoorbeeld een onvoorspelbare betekenis hebben door een idiomatische interpretatie (*de boeken neerleggen*), of een specifieke technische betekenis

“Onderzoek is in het algemeen te veel gericht op individuele woorden en op generieke taal.”

(*aandelen aan toonder*), beperkte gebruiksmogelijkheden (*met vriendelijke groet als afsluiting van een brief*), of een onvoorspelbare vertaling (*nuclear plantkerncentrale*).

De huidige state-of-the-art natuurlijke-taalverwerkende systemen, zoals auteur-systemen, vertaalsystemen en intelligente zoeksystemen, werken nog steeds het best als ze afgestemd zijn op een specifiek domein. Daarom is het noodzakelijk dat een generiek taalverwerkend systeem snel aangepast kan worden aan zo'n specifiek domein. Ieder domein brengt echter zijn eigen vocabulaire mee, en met name ook veel MWE's die alleen in dat domein voorkomen.

Daarom is technologie die snel en zo automatisch mogelijk MWE's en hun eigenschappen kan identificeren in bestaande documenten, van groot belang om taaltechnologie nuttig in te kunnen zetten.

En dit is precies waar één deel van het IRME-project zich mee bezig houdt: het onderzoeken en ontwikkelen van innovatieve methodes en bijbehorende tools om MWE's en hun eigenschappen automatisch of semi-automatisch te identificeren in tekstcorpora.

Om de geïdentificeerde MWE's te kunnen gebruiken is het van belang dat dergelijke expressies lexicaal gerepresenteerd worden op een manier die toelaat ze efficiënt te integreren in een willekeurig taalverwerkend systeem. En dat is waar het andere deel van het IRME-project zich mee bezighoudt: het onderzoeken en ontwikkelen van een methode die MWE's zo theorie- en systeemafhankelijk mogelijk representeert en, onlosmakelijk daarmee verbonden, het ontwikkelen van methodes om aldus gerepresenteerde expressies te integreren in een willekeurig taalverwerkend systeem.

Het probleem van de MWE's wordt door de taaltechnologische industrie als een belangrijk probleem gezien. De *Technologieverkenning Nederlandstalige taal- en spraaktechnologie* van M&I/Partners en Montemore, die de laatste barrière voor het opstarten van het STEVIN-programma weggenomen heeft, vermeldt dat NOTaS erop wijst dat het onderzoek in het algemeen te veel gericht is op individuele woorden en op generieke taal, terwijl de aandacht meer zou moeten liggen op multiwoord-expressies en domeinspecifieke taal. Ook Gregor Thurmair, van Linguatex uit Duitsland, dat onder andere automatische vertaalsystemen maakt, heeft hier bij verschillende gelegenheden op gewezen. Met het IRME-project trachten we tegemoet te komen aan deze wens.

JAN ODIJK
COÖRDINEERT
HET IRME-PROJECT

IRME is een project uit de eerste ronde. Het project loopt tot eind mei 2007. Zie: <http://www-uilots.let.uu.nl/irme/>.