

Nederlandse Samenvatting van:
TRANSIT:
a linguistically motivated Dutch-Turkish
MT-system

Dr. A.M. Stoop

4 december 1995

**Een wetenschappelijke proeve op het gebied van de Letteren
in het bijzonder de Taalwetenschap**

PROEFSCHRIFT

**ter verkrijging van de graad van Doctor
aan de Katholieke Universiteit te Nijmegen,
volgens besluit van het College van Decanen
in het openbaar te verdedigen
op maandag 4 december 1995.**

Door

A. M. Stoop

geboren op 1 november 1957 te Utrecht

Promotor: Prof. Dr. J.J.A. van Bakel

Co-promotor: Dr. P.A.J.M. Coppens

Co-promotor: Dr. E.J. Zürcher

Samenvatting

In deze studie wordt onderzocht of de semantische taal **SELANCA**, ontworpen door Jan van Bakel, op adequate wijze de semantische inhoud van Nederlandse zinnen weergeeft. De gedachte is dat wanneer **SELANCA** als een scharnier in een automatisch vertaalsysteem kan fungeren, waarbij de semantische inhoud van de resulterende doeltaalzinnen overeenkomt met de semantische inhoud van de als uitgangspunt gekozen brontaalzinnen, dit scharnier **SELANCA** dus blijkbaar ook deze semantische inhoud bezit. Het ontwikkelde vertaalsysteem heet **TRANSIT**. Het Nederlands fungeert als brontaal; het Turks is als doeltaal gekozen.

Hoofdstuk 1 presenteert naast een globale plaatsbepaling van **TRANSIT** ten opzichte van andere automatisch vertaalsystemen in Nederland (§1.1) een overzicht van de aard en de mogelijke toereikendheid van **SELANCA** (§1.2). Duidelijk wordt gemaakt op welke manier de term *semantiek* gehanteerd wordt binnen het model en welke *semantische noties* hun plaats vinden in de semantische taal.

De analyse van het Nederlands, resulterend in **SELANCA**-expressies, wordt gepleegd door het **AMAZON-CASUS**-systeem dat Jan van Bakel in de jaren 70 en 80 aan de Universiteit van Nijmegen ontwikkeld heeft. In hoofdstuk 2 wordt samenvattenderwijs in drie stappen uitgelegd hoe na lexicalisering van de inputstring (§2.2), via een syntactische parsing van het resultaat van de lexicalisering in §2.3 een semantische interpretatie van een subset van Nederlandse zinnen opgeleverd wordt in §2.4.

Hoofdstuk 3 is gewijd aan de vertaling naar het Turks. In willekeurig welk vertaalsysteem men ook werkt, altijd zal, voordat men aan het genereren van een doeltaal kan beginnen, een diepgaande analyse van de structuur van de doeltaal gepleegd moeten worden. Omdat er computationele hulpmiddelen voorhanden waren, is een analyse van het Turks gepleegd op een wijze die parallel loopt aan de analyse van het Nederlands door **AMAZON-CASUS**. Al snel wordt duidelijk dat door de agglutinerende aard van de taal de morfologische analyse (§3.2.1) van het Turks een groter gewicht krijgt dan de lexicalisering van het Nederlands. De syntactische analyse in §3.2.2, gevolgd door een semantische analyse in §3.2.3 leveren **SELANCA**-expressies op die een positieve afloop van het onderzoek met betrekking tot **SELANCA** doen vermoeden.

Nadat inzicht is verkregen over de wijze waarop de analyse van het Turks gepleegd wordt, wordt in §3.3 in een comparatieve analyse van het Nederlands en het Turks inhoudelijk naar verschillen en overeenkomsten tussen beide talen gekeken. Onderdelen uit deze analyse zijn: de enkelvoudige zin, de NP, de AP, niet-relatieve ingebedde zinnen, copula-constructies, operatoren, vraagzinnen, passieve constructies, relatieve bijzinnen en het possessieve werkwoord “hebben”.

In §3.4 wordt een vertaalmodel gepresenteerd. Via het bottom-up doorlopen van cycli in de **SELANCA**-expressie, wordt in theorie iedere sub-boom die gedomineerd wordt door een cyclische knoop zodanig bewerkt in een drietrapsmodel (lexicon, syntaxis, semantiek), dat een Turkse oppervlakte-structuur

opgeleverd wordt, waarin alleen de Turkse suffixen nog in een diepte-representatie voorkomen. Deze diepte-representaties worden in een afsluitende fonotactische component bewerkt tot oppervlakte-representaties. De wijze van vertalen van de bij de comparatieve analyse opgesomde onderdelen wordt per onderdeel theoretisch behandeld, waarna tenslotte de benodigde regels in de fonotactische component besproken worden. Er wordt een model ontworpen waarmee de vocaalharmonie en de consonantalternering in het Turks op transformationele wijze beregeld kan worden.

In hoofdstuk 4 wordt de theorie uit hoofdstuk 3 geïmplementeerd in een computermodel. Nu pas blijkt hoe gecompliceerd de implicaties van transformationele regels kunnen zijn, wanneer deze regels samengevoegd worden in één systeem. Tevens komen nu de onvolkomenheden van **SELANCA** boven water drijven. De implementatie wordt aan de hand van de eerder genoemde onderdelen besproken.

Hoofdstuk 5 geeft de discussie weer van het eerder besproken onderzoek en mogelijkheden voor vervolgonderzoek. De discussie wordt in twee delen gesplitst. Allereerst wordt de claim besproken dat **SELANCA** voldoet als semantische taal. Geconcludeerd wordt dat op basis van de linguïstische definitie van semantiek door Van Bakel, **SELANCA** inderdaad voldoet. Problemen kunnen in de toekomst verwacht worden bij de uitwerking van de overige semantische kernen, met name de *N*.

Wordt gekeken naar de distinctieve features die gehanteerd worden om tot een **SELANCA**-expressie te komen, dan blijkt dat deze onvoldoende en ad hoc zijn. Alleen met gebruikmaking van semantische noties blijkt automatisch vertalen niet mogelijk; de distinctieve features bij nomina en verba spelen een belangrijke rol, waarbij gebleken is dat een onafhankelijke keuze van deze features (zoals dat gebeurt is in CASUS), onvermijdelijk leidt tot vertaalproblemen. Een aantal noodzakelijke, maar zeker niet alle, features kunnen in de **SELANCA**-expressie geïntroduceerd worden door opname bij Turkse lexicale items in het lexicon.

Ten tweede wordt een evaluatie gepleegd op basis van technolinguïstische criteria. Nu blijkt dat met betrekking tot de vier gehanteerde theorieën in deze studie deze alle (de analyserende, de genererende, de fonotactische en de vertaaltheorie) consequent doorgevoerd zijn in de implementatie en als zodanig ook herkenbaar.