

Nederlandse Samenvatting

Men neemt algemeen aan dat er in de orde van vijf- tot zesduizend talen zijn. Afgezien van het Engels, Frans of het Spaans, bestaat er voor veel talenparen $\langle X, Y \rangle$ niet een woordenboek $X \rightarrow Y$ of $Y \rightarrow X$. Men moet het dan meestal doen met woordenboeken $X \rightarrow$ Engels/Frans/Spaans en Engels/Frans/Spaans $\rightarrow Y$. Toch is er een maatschappelijke behoefte aan vertaalwoordenboeken die de leden van een paar direct met elkaar in een vertaalrelatie brengen zonder de tussenkomst van een klein aantal West-Europese talen met een koloniaal verleden. Ook op theoretische gronden is een dergelijke behoefte te verdedigen.

Het maken van een kwalitatief goed woordenboek vergt veel tijd, en daar er uit de vijf- tot zesduizend talen zo'n 25 tot 30 miljoen talenparen zijn, is het van belang een database te hebben, op grond waarvan directe vertalingen tussen talen mogelijk worden gemaakt. Het proefschrift brengt enkele problemen in kaart die zich bij zo'n onderneming voordoen, tracht enkele daarvan op te lossen en van andere aan te tonen dat de weg niet begaanbaar is.

Een bekend probleem is dat woorden uit verschillende talen moeilijk op elkaar te passen zijn: woorden in verschillende talen hebben vaak niet hetzelfde bereik aan betekenissen, niet alle woorden uit de ene taal hebben een equivalent in een andere, etc. In dit proefschrift geef ik een aanzet tot de opzet van een database waarin een groot deel van deze problemen opgelost wordt. Cruciaal in deze opzet is de structurering van de tussentaal, waarmee in de database niet-corresponderende betekenissen toch op gestructureerde wijze aan elkaar gerelateerd kunnen worden. De structuur van deze tussentaal wordt geleverd door een logisch raamwerk, onder de naam Formele Begripsanalyse. Met deze opzet kan onder meer voor woorden waarvoor geen directe vertaling is in de doeltaal toch een omschrijvende vertaling gegenereerd worden. Daarmee wordt het werk van een lexicograaf die een vertaalwoordenboek voor een talenpaar moet maken vergemakkelijkt.

De wijze waarop dit proefschrift is opgebouwd is als volgt. In hoofdstuk 1 van dit proefschrift wordt een beschouwing gegeven op de vraag aan welke eisen een lexicaal gegevensbestand moet voldoen om in staat te zijn niet-corresponderende betekenissen op gestructureerde wijze aan elkaar te

verbinden. Hierbij wordt een aantal bestaande lexicale gegevensbestanden onder de loep genomen om te kijken om welke reden dit in die systemen op dit moment niet mogelijk is.

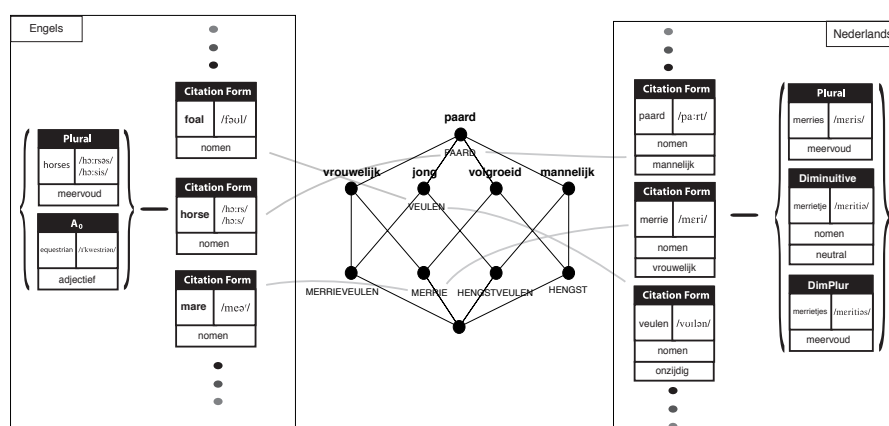
In hoofdstuk 2 wordt de opzet van het SIMuLLDA-systeem geschetst. In deze opzet worden de verschillende talen in het lexicale gegevensbestand aan elkaar gekoppeld door middel van een tussentaal (interlingua). En zoals gezegd is de structuur van deze tussentaal het hart van het SIMuLLDA-systeem. De tussentaal is een gestructureerde eenheid, bestaande uit een traliestructuur. De knopen van deze tralie bestaan uit paren van taalonafhankelijke betekenissen en eigenschappen van deze betekenissen die *definitionele attributen* genoemd worden.

De taalonafhankelijke betekenissen zijn alle betekenissen die door een van de woorden (of eigenlijk: lexemen) uit een van de talen worden uitgedrukt. Dus ieder lexeem uit iedere taal drukt een of meer betekenissen uit. Het omgekeerde is echter niet waar: niet iedere betekenis is in elke taal gelexicaliseerd. Stel bijvoorbeeld dat er in het Italiaans een woord is voor *wegen die naar Rome leiden*, en dat er geen enkele andere taal is met een dergelijk woord. Dan is deze betekenis nog steeds een taalonafhankelijke betekenis, echter een waarvoor alleen in het Italiaans een concreet woord bestaat. De andere talen hebben dan wat heet een *lexicale leemte* ten aanzien van deze betekenis, c.q. dit Italiaanse woord.

De definitionele attributen zijn de eigenschappen die de taalonafhankelijke betekenissen vastleggen. Deze definitionele attributen vinden hun herkomst in de *differentiae specificae* in monolinguale woordenboeken. Om een voorbeeld te geven: het woord *hengstveulen* is gedefinieerd in een woordenboek als een *jong mannelijk paard*. In deze definitie worden **jong** en **mannelijk** opgevoerd als kenmerkende eigenschappen van de betekenis van het woord *hengstveulen*. Het zijn deze kenmerken die gelden als definitionele attributen in SIMuLLDA. De *genus proximum* in deze definitie (*paard*) duidt geen definitioneel attribuut aan, maar verwijst naar een andere betekenis in het woordenboek, waar weer nieuwe definitionele attributen bijhoren.

Als we even afzien van de doorverwijzing naar **paard** en **paard** wel degelijk beschouwen als een definitioneel attribuut, krijgen we een structuur als weergegeven in Figuur 1.

Definitionele attributen zijn, zoals gezegd, eigenschappen die de betekenissen in SIMuLLDA vastleggen. Maar daar deze betekenissen taalonafhankelijk zijn, kunnen deze attributen zelf nooit taalspecifiek zijn. Derhalve zijn ook definitionele attributen taalonafhankelijke elementen van de gestructureerde tussentaal, die in elk van de talen in het lexicale gegevensbestand gelexicaliseerd kunnen worden. Daarbij is er ook een lexicalisatie voor een definitioneel attribuut als dit attribuut in de gegeven taal geen rol speelt.



Figuur 1: Opzet van het SIMuLLDA-Systeem

Om terug te keren op het eerder genoemde voorbeeld: het deel *die naar Rome leiden* zal een Nederlandse lexicalisatie van een definitioneel attribuut zijn (**naar Rome voerend**). En dan specifiek een definitioneel attribuut dat bij de bepaling van geen enkel Nederlands woord een rol speelt.

De opzet in Figuur 1 maakt het mogelijk vertalingen voor woorden te geven: het lexeem *horse* is gekoppeld aan de betekenis HORSE en de betekenis HORSE is weer gekoppeld aan het Nederlandse woord *paard*. Dus *paard* en *horse* zijn rechtstreekse vertalingen of 'vertalingssynonymen' van elkaar.

Door de structuur van de tussentaal wordt het echter ook mogelijk omschrijvende vertalingen te geven voor woorden waarvoor geen directe vertaling bestaat. Een voorbeeld aan de hand van de tussentaal in figuur 1: het Nederlandse woord *hengstveulen* kent geen rechtstreekse vertaling in het Frans: er is wel een woord voor *merrieveulen* (*poulliche*) en een algemener woord voor *veulen* (*poulain*), maar er is geen woord voor *hengstveulen* als zodanig.

Door de plaatsing van de betekenis HENGSTVEULEN in de tralie kunnen we echter wel van alles zeggen over deze betekenis. Allereerst hangt de knoop waarbij deze betekenis hoort onder de knoop van de betekenis VEULEN, en is de betekenis VEULEN wel gelexicaliseerd in het Frans: *poulain*. Dus vanuit de tralie kunnen we stellen dat *poulain* een redelijke, zij het iets te algemene vertaling is voor *hengstveulen*. Dat deze vertaling te algemeen is komt doordat HENGSTVEULEN meer definitionele attributen heeft dan VEULEN: het heeft een *definitioneel surplus*. Dit definitionele surplus bestaat uit precies één definitioneel attribuut: **mannelijk**. Dus wat mist in de *poulain*-vertaling is **mannelijk**, wat in het Frans kan worden uitgedrukt met *mâle*. De complete betekenis van *hengstveulen* in het Frans is de combinatie van deze twee: *poulain mâle*.

Gegeven de manier waarop lexicale leemten worden opgevuld is de notie van *differentiae specificae* in SIMuLLDA geheel op het niveau van de tussentaal vastgelegd: HENGSTVEULEN = VEULEN + **mannelijk**. Het is ook mogelijk de rechterzijde van deze vergelijking weer terug in het Nederlands te vertalen. Dit levert een lexicale definitie op: **hengstveulen** - *mannelijk veulen*. Dus met de SIMuLLDA-opzet is het mogelijk zowel lexicale definities te genereren als vertalingen voor tweetalige woordenboeken, ook in het geval er geen vertalingssynonym bestaat.

Tenslotte worden in hoofdstuk 2 ook nog enige logische eigenschappen besproken van het systeem dat ten grondslag ligt aan SIMuLLDA: Formele Concept Analyse. FCA zorgt ervoor dat de relatie tussen definitionele attributen en taalonafhankelijke betekenissen de traliestructuur opleveren die is weergegeven in Figuur 1. Als onderdeel hiervan wordt ook een internet-applicatie geïntroduceerd die ten behoeve van dit proefschrift ontwikkeld is: JaLaBA, een programma dat de omzetting van tabellen naar de traliestructuur automatisch uitvoert, en een 3-dimensionale weergave van de tralie geeft. Deze applicatie is, net als alle andere aan dit proefschrift gerelateerde zaken, te vinden op de web-site van dit proefschrift: <http://maarten.janssenweb.net/simullda>.

In de hiervoor beschreven opzet van SIMuLLDA zit een aantal verborgen premissen: ieder woord drukt een vast aantal betekenissen uit, woorden uit verschillende talen kunnen dezelfde betekenis uitdrukken en betekenissen worden vastgelegd door middel van definitionele attributen. Al deze premissen zijn echter problematisch. Met name de laatste doet sterk denken aan de reeds lang verworpen stelling dat woordbetekenis terug te voeren is op een kleine set aangeboren semantische primitieven. Om desondanks de opzet te hanteren zoals hierboven beschreven is het daarom van groot belang precies aan te geven wat de status van de verschillende onderdelen van het systeem is: wat woorden, talen, betekenissen en definitionele attributen precies geacht worden te zijn. En met name ook welke claims niet worden gemaakt door het systeem.

Een diepgaande beschouwing van alle onderdelen wordt gegeven in hoofdstuk 3: talen worden gedefinieerd als willekeurige verzamelingen lexemen, waarbij lexemen de ingangen in een woordenboek zijn. Lexemen bestaan op hun beurt weer uit verzamelingen woordvormen, te weten de vervoegingen en verbuigingen van het woord. En de woordvormen zijn representaties van abstracte entiteiten, bestaande uit een uitspraak, een spelling, een woordklasse en evt. een geslacht. Lexemen zijn geen woorden in de standaardzin van het woord: lexemen kenmerken zich niet doordat ze door spaties gescheiden worden; lexemen kunnen ook uit meerdere woorden bestaan (bv. bij idiomatische uitdrukkingen) of kleinere delen omvatten (morfemen).

Bij de karakterisering van de betekenissen en definitionele attributen is het vooral van belang aan te geven wat deze niet zijn: betekenissen zijn

niet denotationeel van karakter. Dat wil zeggen, betekenissen zijn niet gelijk aan noch worden bepaald door de verzameling van objecten die onder het begrip vallen; betekenissen leggen ook niet vast welke objecten er precies onder vallen, noch stelt de betekenis je in staat van ieder object eenduidig vast te stellen of het onder dat begrip valt of niet. Daarnaast zijn de definitionele attributen die de betekenissen vastleggen geen zwaar fundamentele atomen zoals Katz & Fodor hebben voorgesteld: ze zijn niet aangeboren, er is niet een van God gegeven aantal attributen, en definitionele attributen leggen niet alles vast wat we doorgaans onder woordbetekenis laten vallen. Van veel begrippen weten we hoe het kenmerkende element ervan er uit ziet (bv. wat een typisch ontbijt is), maar dergelijke prototypen zijn niet taalafhankelijk en worden ook niet vastgelegd door definitionele attributen. Definitionele attributen zijn niet meer en niet minder dan de *differentiae specificae* uit woordenboeken.

De opzet van dit proefschrift is het leveren van een werkbaar model dat het werk van lexicografen verlicht. Daarom is het van groot belang te laten zien dat de hierboven beschreven opzet in praktijk ook daadwerkelijk bruikbaar is. Hoewel de echte bruikbaarheid slechts blijkt als het systeem op grote schaal gebruikt zou worden, wordt in hoofdstuk 4 van dit proefschrift een beperkte empirische test beschreven: als we alle woorden voor wateren (meren, rivieren, plassen, baaien, etc.) in een zestal talen in oogschouw nemen, is het mogelijk deze met behulp van SIMULLDA aan elkaar te koppelen. Dit levert dan daadwerkelijk de gewenste situatie waarbij voor ieder woord in iedere taal een vertaling in iedere taal gegenereerd kan worden. Er is een tweede, kleinere empirische test voor alle namen van zeilen op een vijf-master. Dit laatste voornamelijk om een vergelijking te trekken met de analyse die gegeven wordt in het proefschrift van Marc van Campenhoudt. Hoewel er zich bij deze empirische tests enige problemen voordoen, zijn deze problemen (op een na) met kleine aanpassingen van het systeem op te lossen.

Het enige probleem dat niet op te lossen is binnen het systeem is het volgende: een woordenboekdefinitie omschrijft de betekenis van een woord door (naast de *differentiae specificae*) te verwijzen naar het genus proximum. Dit genus is de *betekenis* van een van de andere woorden in het woordenboek. Nu komt het echter om verschillende oorzaken voor dat deze genusbetekenis zelf niet in het woordenboek staat; niet door een leemte, maar doordat het ondergespecificeerd is wat de genus betekenis precies is. De reden hiervoor is dat een basale aanname die impliciet achter woordenboeken schuil gaat, in feite incorrect is: woorden hebben geen welomschreven aantal onderscheiden betekenissen. Een goed voorbeeld hiervan is dat bv. het woord *raam* twee betekenissen heeft die onderscheidbaar maar toch ook hetzelfde zijn: het is een opening in de zijkant van een huis, maar ook het glas dat daarin besloten zit. Als we hiervoor twee betekenissen opnemen, moeten we ook van alle hyponymen van *raam* aannemen dat

ze ambigu zijn: een *dakraam* is een raam in deze beide betekenissen samen. De stelling is ook dat binnen een systeem waarin woordenboek definities serieus worden genomen, dit probleem niet opgelost kan worden.

In zijn basale opzet behelst het systeem slechts een beperkt gedeelte van de in woordenboeken aanwezige informatie: alleen de semantische definities en dan ook nog alleen van nomina. Om een volledige lexical gegevensbestand te zijn dienen de andere delen van woordenboekinformatie echter ook een plaats te krijgen in het systeem. Dit wordt ten dele in hoofdstuk 5 opgelost. In dat hoofdstuk wordt beschreven hoe labels, collocaties, voorbeeldzinnen en morfologische derivaties kunnen worden gemodelleerd in het systeem, deels gebruik makend van de lexicale functie uit de *Meaning*⇔*Text Theory*. Ook wordt kort besproken hoe het systeem zich verhoudt tot andere woordklassen dan nomina, zoals werkwoorden en adjectieven en worden enkele aspecten beschreven van een eventueel voor dit systeem te ontwikkelen toepassing.