Towards the automatic detection of syntactic differences

Kroon, M.S.

Samenvatting

De zinsbouw – of syntaxis – van natuurlijke taal is een systeem van combinatorische regels waarmee uit woorden en morfemen[1] complexe hiërarchische structuren worden gebouwd, zoals zinnen en woordgroepen. Het inzicht dat de woorden in een zin niet alleen lineair maar ook hiërarchisch zijn geordend, namelijk als woordgroepen die woordgroepen bevatten, staat centraal in de moderne taalwetenschap (cf., e.g., Berwick and Chomsky 2016).

Oppervlakkige vergelijking van verschillende talen lijkt erop te duiden dat hun syntaxis in grote mate verschilt: bijvoorbeeld, variatie in woordvolgorde, variatie in de aan- of afwezigheid van een morfeem, woord of woordgroep, het verdubbelen van grammaticale kenmerken, of variatie in de morfosyntactische expressie van grammaticale relaties zoals congruentie tussen het onderwerp en de persoonsvorm. Niettemin kan men in de syntactische literatuur talloze argumenten vinden voor de hypothese dat alle menselijke talen dezelfde abstracte set syntactische principes delen. Het hoofddoel van theoretisch comparatief-syntactisch onderzoek is om de syntactische variatie tussen natuurlijke talen in kaart te brengen door hun structuren te vergelijken en de syntactische overeenkomsten en verschillen te beschrijven, om vervolgens een alomvattende, taaloverstijgende theorie te kunnen formuleren die deze variatie beschrijft en verklaart (Cinque and Kayne 2005). Het vakgebied probeert antwoorden te vinden op vragen als: wat is een (on)mogelijke natuurlijke taal, welke syntactische eigenschappen zijn universeel en welke zijn taalspecifiek, en is de syntactische variatie een eigenschap van de module van de mentale grammatica die hiërarchische structuren bouwt of is het mogelijk om syntactische variatie te herleiden tot eigenschappen van andere modules van de grammatica, zoals het lexicon of de module die zorgt voor de fonologische *spellout* en linearisatie?

Het is gebruikelijk voor syntactici om hun eigen moedertaal met andere talen te vergelijken door het bestuderen van gedetailleerde grammatica's en

[1] Een morfeem is de kleinste lexicale eenheid die betekenis draagt in taal. Bijvoorbeeld, het woord *draaglijk* bestaat uit de morfemen *draag* en *-lijk*, die zelf niet verder kunnen worden opgedeeld in kleinere, betekenisdragende stukjes.

andere taalkundige literatuur en de (on-)grammaticaliteit van zinnen te toetsen door ze voor te leggen aan vakgenoten en andere proefpersonen. Maar door het enorme aantal natuurlijke talen en dialecten, de hoge mate van variatie die ze vertonen (zelfs tussen nauw verwante talen of dialecten) en het technisch gezien oneindige aantal mogelijke zinnen per taal of dialect die de taalkundige kan onderzoeken is systematische vergelijking een vrijwel onmogelijke opgave.

Het gevolg hiervan is dat syntactici veel verschillen en associaties tussen die verschillen over het hoofd kunnen zien en dat formele beschrijvingen van taal onvolledig blijven. Het vakgebied zou daarom sterk gebaat zijn bij een (gedeeltelijke) automatisering van het proces, omdat dit het onderzoek zou versnellen en het grootschaliger, systematischer en reproduceerbaarder zou maken. Een computer kan veel meer data in veel meer talen veel systematischer verwerken en analyseren, wat het waarschijnlijker maakt dat nieuwe variatie in syntaxis kan worden ontdekt en correlaties tussen variabelen die terug te voeren zijn tot abstractere, onderliggende kenmerken kunnen worden gevonden. De vraag blijft echter: **kunnen syntactische verschillen tussen talen automatisch ontdekt worden en, zo ja, hoe dan?** Dit is de vraag die in dit proefschrift centraal staat.

Hoofdstuk 1 is een uitgebreide inleiding, waarin onder andere de literatuur over dit vraagstuk besproken wordt en de aard van de data wordt beschreven. De data waarmee in dit proefschrift wordt gewerkt bestaan uit parallelle tekstcorpora: grote tekstverzamelingen waarbij elke zin gelinkt is aan een vertaling in een tweede taal. Specifiek wordt gebruik gemaakt van de Bijbel in het Engels, Hongaars en Nederlands, en van een Engels, Nederlands en Tsjechisch fragment van het Europarl corpus (Koehn 2005), wat bestaat uit de vergaderingen van het Europees Parlement, die vertaald worden in alle talen van de Europese Unie. Deze tekstcorpora worden voor de analyse door middel van bestaande algoritmes automatisch syntactisch geannoteerd, waar ze dat nog niet waren.

In Hoofdstuk 2 komt het probleem van syntactisch onvergelijkbare zinsparen aan bod. In parallelle corpora zijn zinsparen op syntactisch niveau lang niet altijd vergelijkbaar: twee zinnen die hetzelfde betekenen kunnen zeer verschillende constructies bevatten of vrije vertalingen betreffen. Wanneer dit soort zinnen worden gebruikt om syntactische verschillen te ontdekken, worden er vele verschillende soorten verschillen gevonden die niet informatief zijn voor de taalkundige. “Vrije” vertalingen moeten daarom verwijderd worden uit de dataset, maar door de omvang van gebruikte datasets is het veelal onmogelijk om dit met de hand te doen. Er is daarom een methode en kwantitatieve maat nodig om zinsparen automatisch uit de dataset te kunnen filteren die niet syntactisch vergelijkbaar zijn.

In dit hoofdstuk worden vier manieren om dit te bewerkstelligen verkend en geëvalueerd met datasets met Engelse, Nederlandse en Duitse parallelle zinsparen. Voor de zinsparen is van tevoren met de hand gedetermineerd of ze wel of niet syntactisch vergelijkbaar zijn. Het eerste filter is gebaseerd op de Levenshteinafstand op POS-tags (woordsoortlabels), een bekend algoritme dat het minimaal aantal bewerkingen berekent dat nodig is om de ene sequentie in

de andere sequentie te veranderen (Levenshtein 1966). Het tweede filter maakt gebruik van de zinslengteratio tussen de twee zinnen onder de aanname dat als een zin significant langer is dan zijn vertaling, het zinspaar waarschijnlijk te sterk van elkaar verschilt of zelfs foutief aan elkaar is verbonden. Het derde filter is gebaseerd op de bewerkingsafstand tussen de syntactische dependentiebomen van de twee zinnen. Deze bewerkingsafstand is equivalent aan de Levenshteinafstand, maar toegepast op hiërarchische structuren in plaats van lineaire sequenties. Het laatste filter combineert de vorige drie filters in een logistisch regressiemodel.

De resultaten van Hoofdstuk 2 laten vooral zien dat filteren op syntactische vergelijkbaarheid een moeilijke opgave is, deels omdat syntactische vergelijkbaarheid lastig te definiëren is. Niettemin zijn de filters bruikbare tools voor de automatische selectie van syntactisch vergelijkbare zinsparen uit een parallel corpus. De beste resultaten kunnen worden behaald met het filter dat gebruik maakt van een logistisch regressiemodel, terwijl de filters die gebruik maken van de Levenshteinafstand en de bewerkingsafstand tussen de syntactische bomen gebruikt kunnen worden met redelijk resultaat.

In Hoofstuk 3 presenteer ik een systematische methode om mogelijke syntactische verschillen te detecteren en hypotheses erover te rangschikken voor verder onderzoek door gebruik te maken van parallelle data en het *Minimum Description Length*-principe (MDL). MDL biedt een elegant paradigma voor het ontdekken van structuur in data. Het formaliseert het idee dat elke regelmatigheid in de data kan worden gebruikt om de data te comprimeren (among others Grünwald 2007; Barron, Rissanen and Yu 1998). Deze regelmatigheden kunnen dan worden beschouwd als karakteristieke bouwstenen onderliggend aan de data. Ik maak hierbij gebruik van het SQS-algoritme ('Summarising event seQuenceS'; Tatti and Vreeken 2012) – een algoritme ontwikkeld om patronen in sequentiële data te ontdekken met MDL – om 'typische' sequenties van POS-tags te *minen* voor elke taal die wordt onderzocht. SQS produceert inderdaad lijsten met daarin verwachte patronen van POS-tags die men als karakteristiek voor een taal zou beschouwen. Uit deze lijsten wordt een lijst van mogelijke syntactische verschillen geproduceerd op basis van het aantal parallelle zinnen waar een patroon voorkomt in de ene taal maar niet in de andere. Met behulp van een statistische test worden dan hypotheses gegenereerd over waar er syntactische verschillen kunnen worden gevonden tussen het taalpaar in kwestie. In het hoofdstuk wordt de methode toegepast op parallelle corpora van het Engels, Nederlands en Tsjechisch en ik onderzoek het effect van het filter van Hoofdstuk 2 op de resultaten. De resultaten laten zien dat de methode veelbelovend is in zowel het *minen* van karakteristieke bouwstenen van een taal, alsook het ontdekken van bruikbare syntactische verschillen tussen talen.

Waar de methode van Hoofdstuk 3 aanneemt dat er POS-taggers (programma's die automatisch woorden voorzien van een woordsoortlabel) beschikbaar zijn voor beide talen die worden onderzocht en dat beide talen zijn geannoteerd met dezelfde set labels en volgende dezelfde conventies, is dit niet

altijd het geval. Sterker nog, hoewel het Universal Dependencies-programma (UD; Nivre et al. 2016) streeft naar consistente tagging en annotatie van syntactische dependentiebomen tussen talen,[2] kunnen de richtlijnen van taal tot taal significant verschillen (waarvoor altijd goed onderbouwde redenen zijn).

In Hoofdstuk 4 wordt daarom een andere methode onderzocht om syntactische verschillen te ontdekken, die niet afhankelijk is van de beschikbaarheid van annotatietools voor beide talen. De hoofdvraag van het hoofdstuk is of het mogelijk is om in parallelle tekst volledig geannoteerde tekst in de ene taal (die we de brontaal noemen) te gebruiken om grammaticale eigenschappen van een andere, minder goed beschreven taal (die we de doeltaal noemen) te ontdekken, en verschillen tussen de twee talen.

Hiertoe wordt gebruik gemaakt van *word alignment*, het automatisch oplijnen van woorden die elkaars vertaling zijn binnen twee zinnen. Aan de hand van *word alignment* wordt de annotatie van woorden van de brontaal op woorden van de doeltaal geprojecteerd, met het doel om syntactische eigenschappen van de doeltaal en verschillen tussen de bron- en doeltaal in kaart te brengen door deze projecties semi-automatisch te analyseren. Er zijn drie algoritmes ontwikkeld om de met *word alignment* opgelijnde data te analyseren: de Data Grouper for Attribute Exploration (DGAE), waarmee handige overzichten worden gegeven van de frequentie van annotaties en eigenschappen binnen groepen woorden; de Generalization Tree Inducer (GTI), waarmee de data wordt gestructureerd op basis van de entropie van de annotaties in een poging om te generaliseren over woordklassen; en de Affix-Attribute Associator (AAA), waarmee hypotheses worden gegenereerd over welke tekenreeksen, of *strings*, mogelijk affixen zijn in de doeltaal door ze te associëren met morfosyntactische eigenschappen van woorden in de brontaal. Deze drie tools zijn geëvalueerd op het taalpaar Engels-Hongaars. Zonder enige kennis te hebben van het Hongaars heb ik de tools gebruikt om 43 hypotheses te vormen aangaande morfosyntactische eigenschappen van het Hongaars of verschillen met het Engels. Deze hypotheses zijn onafhankelijk gecontroleerd door een moedertaalspreker en een expert van het Hongaars en zijn syntaxis en zijn getoetst aan een lijst van karakteristieke verschillen tussen het Hongaars en het Engels die van tevoren onafhankelijk door dezelfde expert was samengesteld. De conclusie luidt dat de tools zeer effectief gebruikt kunnen worden om veel correcte hypotheses te vormen over verschillen tussen de talen, verspreid over meerdere syntactische domeinen. Met behulp van de tools heb ik zelfs twee hypotheses gevormd waarvan het vooralsnog onbekend is of ze correct zijn of niet, wat de kracht van de tools in de zoektocht naar syntactische verschillen tussen talen louter onderstreept.

De dissertatie wordt afgesloten met een uitgebreide discussie in Hoofdstuk 5, waarin alle observaties van de voorgaande hoofdstukken bijeengebracht worden en aan elkaar worden verbonden, hetgeen leidt tot nieuwe, overkoepelende observaties en conclusies. Daarin is de belangrijkste conclusie dat het

[2] `universaldependencies.org`

mogelijk is om automatisch syntactische verschillen te ontdekken. De tools die zijn ontwikkeld in het kader van dit onderzoek werken goed en kunnen een taalkundige aanzienlijk helpen in de zoektocht naar verschillen of overeenkomsten. Niettemin werken de tools niet perfect en zijn ze bijvoorbeeld afhankelijk van de kwaliteit van de data en de annotaties: het proces is daarom, vooralsnog, wellicht niet zo gedetailleerd, geautomatiseerd of objectief als men zou willen, maar de tools bieden een goed uitgangspunt voor vervolgonderzoek.