Title: De zijnsstatus van de afhankelijke V1-constructie in het Nederlands

English title: The ontological status of the dependent verb-first construction in Dutch

Author: Barend Beekhuizen[1]

Abstract

Dependent verb-first clauses have received a fair amount of interest from a functionalist perspective. In this paper, I argue that their conceived unity as a grammatical construction might be overstated. I do so on the basis of both categorical and probabilistic differences in the distributions of the various types that I discern. The shared functional properties of the various dependent verb-first clauses are best seen as non-conventional tendencies rather than as a conventional symbolic function, that may have driven the development historically, but is not a part of the inventory of linguistic knowledge of a language user synchronically.

[1] Ik ben de redactie van de Dag van de Nederlandse Zinsbouw en de anonieme beoordelaar erkentelijk voor hun opmerkingen en commentaar.

Keywords: asyndetic subordinate clauses, verb-first, conditionals, Dutch, exemplar-based grammar, construction grammar, memory-based learning, conventionality

1 Inleiding

Naast de gebruikelijke ondergeschikte bijzin met het vervoegde werkwoord op een zinsfinale positie en een inleidend onderschikkend voegwoord, kent het Nederlands een voegwoordloze, werkwoordsinitiële bijzin (afhankelijke V1-constructie, of AV1 in het vervolg), geïllustreerd in zinnen (1)-(3)

(1) Wordt de sluiting losgetrokken, dan maakt die het karakteristieke scheurende geluid
(2) Werden in 1942 vijfentachtig onderzeeboten tot zinken gebracht, in 1943 was dit aantal gestegen tot tweehonderdenzevenentachtig
(3) Zit je voor jaren in de cel, moet je nog verhuizen ook.

In zin (1) zien we een voorwaardelijke relatie tussen de bijzin en matrixzin en een parafrase met *als* ligt voor de hand (*Als de sluiting wordt losgetrokken*). Zin (2) geeft een contrast tussen de bij- en matrixzin aan: beide zinnen worden geasserteerd door de spreker, in tegenstelling tot (1), waar het de voorwaardelijke relatie is die geasserteerd wordt. De parafrase die bij zin (2) voor de hand ligt, is

er één met *waar* of *terwijl* (*Waar/Terwijl in 1942 vijfentachtig onderzeeboten tot zinken gebracht werden*). Het type in zin (3), ten slotte, lijkt op zowel het voorwaardelijke type (een parafrase met *dan* of *en dan* is niet geheel ondenkbaar) als het contrastieve (beide zinnen in (3) zijn geasserteerd), maar drukt daarnaast de sterke verwondering van de spreker uit.

De AV1 heeft, zowel in de neerlandistische taalkunde als daarbuiten, vooral op aandacht mogen rekenen door zijn woordvolgorde. Het patroon deelt immers het vormelijke kenmerk van een vooropgeplaatst vervoegd werkwoord met (o.m.) vraagzinnen en de karakteristieke inleidingszin van moppen (*Komt een man bij de dokter*). Dit leidt – synchroon, en vanuit een functionalistisch perspectief – tot de vraag wat de werkwoordsinitialiteit van een zin signaleert, d.w.z. wat de functie ervan is, en – diachroon – tot de vraag hoe het AV1-patroon ontstaan is.

Functionalisten zoals Daalder (1983), Van der Horst (1995), en Diessel (1997; voor het Duits) menen dat het vooropgeplaatste werkwoord (in AV1-patronen maar ook in vraagzinnen, imperatieven, etc.) een bepaalde functie signaleert, dat wil zeggen: dat werkwoordsplaatsing een taalteken met een vorm- en een

betekeniskant is. Volgens Daalder behelst de betekenis van werkwoordsplaatsing de mate van ‘deiktische spanning’ tussen inhoud van de zin enerzijds en het hier-en-nu van de gesprekssituatie, inclusief de wederzijds gedeelde kennis van de participanten anderzijds. Werkwoordsinitiële zinnen markeren dan dat deze spanning hoog is: er wordt aandacht gevraagd voor de spanning tussen de inhoud van de zin en het op dat moment door de sprekers gekende. Van der Horst meent dat de plaatsing van de persoonsvorm een symbool is dat de mate van aandacht voor de ‘attitudinele verhouding’ van de spreker tot de inhoud uitdrukt. Dit wil zeggen dat met de werkwoordsplaatsing de spreker signaleert dat zij wil dat de hoorder een bepaalde mate van aandacht schenkt aan de illocutieve (cf. Searle 1979) of modale status van het predikaat. Met werkwoordsinitiële zinnen wil de spreker zeggen dat haar verhouding tot de inhoud bijzonder is en bovengemiddeld veel aandacht van de hoorder verdient. Diessel, ten slotte, bedt de betekenis van werkwoordsplaatsing (voor het Duits) in in Searles (1979) taalhandelingstheorie: werkwoordsplaatsing opereert op de *fit* van de taalhandeling. Waar zinnen met het werkwoord op de tweede plaats ongemarkeerd zijn in het uitdrukken van een wereld-naar-woorden

fit, ontkennen werkwoordsinitiële zinnen deze ongemarkeerde *fit*: de *fit* is dus 'niet wereld naar woorden'.[2]

Wat de diachrone ontwikkeling van het AV1-patroon betreft, vinden we recente inzichten in Leuschner & Van den Nest (2015) voor het Duits en Engels. Als verschillende gesuggereerde bronconstructies van de AV1 noemen zij het polaire-vraagpatroon (cf. Jespersen 1940) en het niet-afhankelijke V1-declaratiefpatroon (cf. Hopper 1975). Uitgaande van de hypothese dat het polaire-vraagpatroon de bronconstructie is, laten Leuschner & Van den Nest zien dat Engelse AV1s meer afwijken van Engelse polaire vraagzinnen dan Duitse AV1s van Duitse polaire vraagzinnen in het gebruik van hulpwerkwoorden en de tempus- en moduspatronen. Dit suggereert dat de Engelse AV1 verder gegrammaticaliseerd is dan de Duitse, hoewel de Duitse AV1 op hetzelfde grammaticalisatiepad lijkt te zitten als de Engelse.

Al de functionele studies beschouwen het AV1-patroon (en zelfs breder: het werkwoordsinitiële patroon) als een categorie die enige realiteit ofwel op taalstructureel niveau, ofwel op individueel-cognitief niveau heeft. Het is die aanname die ik in deze bijdrage wil

[2] Wat, zoals een anonieme beoordelaar terecht opmerkt, tot de interessante situatie leidt dat vraagwoordvragen een andere *fit* hebben dan polariteitsvragen. De mogelijke gevolgen van deze implicatie vallen buiten het bereik van dit artikel.

nuanceren. Zowel de ‘harde’ distributionele gegevens als de meer graduele wijzen erop dat de AV1 beter als een zwak gerelateerde familie van patronen kan worden gezien. Omdat die lokalere patronen wel gedeelde functionele kenmerken hebben, rijst de vraag waar deze vandaan komen. Mijn voorgestelde oplossing behelst het loskoppelen van het conventionele (cognitieve en interne) taalsysteem en niet-conventionele interpretatieneigingen van taalgebruikers. De AV1 is geen taalteken, maar het vooropplaatsen van het vervoegde werkwoord is tegelijkertijd geen betekenisloze daad.

2 Bestaat de AV1?

De centrale vraag is of AV1-patronen als groep een zinnig voorwerp van beschrijving vormen. Deze vraag kan toegespitst worden op de verschillende zijnsniveaus waarop dit voorwerp zich bevindt: beschrijven we een verzameling talige conventies in een gemeenschap of de mentale weerslag daarvan in een (mogelijk geïdealiseerde) taalgebruiker; en beschrijven we taalkennis (in de vorm van de weerslagen van de talige conventies in een gemeenschap) of bepaalde verwerkingsvermogens die niet als taalkennis maar meer als niet-conventionele ‘taalkunnens’ (d.w.z.,

elementen van het procedurele i.p.v. declaratieve geheugen) gelden? Deze nadere indeling wordt echter pas relevant als we er een gebruiksgebaseerde visie op taal op nahouden.

In de gebruiksgebaseerde visie op taal (Langacker 1988; zie Verhagen 2005 voor een Nederlandstalig overzicht) wordt aangenomen dat de individuele taalkennis uit de verwerkte ervaring van taalgebruik bestaat. Uit deze aanname volgt dat deze verwerkte ervaringen aspecten van het signaal (de klankvorm, in het geval van gesproken taal) en het gesignaleerde (een conceptualisatie en een communicatieve intentie) bevat. Deze conclusie ligt aan de basis van de verwante benadering van de constructiegrammatica (zie Goldberg 2003 en voor een Nederlandstalig overzicht wederom Verhagen 2005), die stelt dat alle taalkennis bestaat uit paren van signalerende vormen en gesignaleerde betekenissen: Goldberg (2003) spreekt in dit kader van “Constructions all the way down”.

De precieze structuur van de cognitieve representatie die dit oplevert en de mechanismes waarmee dit gebeurt, zijn het voorwerp van voortschrijdend inzicht. In het algemeen wordt aangenomen dat taalgebruikers het vermogen om nieuwe gebruiksgevallen te vormen en begrijpen, baseren op die gebruiksgevallen door te abstraheren over gemeenschappelijke vormelijke en inhoudelijke kenmerken van

verschillende verwerkte taalervaringen. Deze abstractie wordt in de traditie niet gezien als het daadwerkelijk cognitief 'onttrekken' van regels aan de concrete gebruiksgevallen maar veeleer als de gedeelde neurale activatiepatronen tussen gebruiksgevallen (Langacker 1988 spreekt in dit kader van 'immanentie').[3]

Het gebruik van de activatiepatronen in het taalgebruik zorgt vervolgens voor hun verdere versterking. Of een 'abstract' patroon versterkt wordt, hangt af van de frequentie waarmee het gebruikt wordt om nieuwvormingen te vormen en begrijpen. In abstracto: stel dat er een patroon [*ab*X] verworven is, waarin *a* en *b* concrete elementen zijn en X een open positie, gevormd op grond van de ervaringen [*abc*] en [*abd*], dan zal [*ab*X] alleen verder versterkt worden als het gebruikt wordt om nieuwvormingen (zoals *abe*) te analyseren, en niet als het gebruikt wordt om verdere gevallen als [*abc*] of [*abd*] te analyseren. Langacker (1988) spreekt in dit verband van de voorrang van het concrete op het abstracte: als een

[3] In de constructiegrammatica wordt de metafoor waarin de grammatica in een 'netwerk' georganiseerd is, veel gebruikt. In deze metafoor zijn abstracties de 'ouderknopen' van 'kindknopen' (d.w.z. concretere instantiaties van die abstracties). Deze metafoor suggereert (m.i. ten onrechte) dat abstracties aparte cognitieve entiteiten zijn. Zoals elke metafoor plaatst het bepaalde aspecten van het begrip van de organisatie van de grammatica op de voorgrond, terwijl andere aspecten op de achtergrond blijven of zelfs conflicteren met de metafoor.

taalgebruiker zowel [*abc*] als [*ab*X] heeft, en hij komt de uiting *abc* tegen, dan zal hij eerder [*abc*] dan [*ab*X] gebruiken omdat dit minder cognitieve energie kost: [*abc*] zit conceptueel dichter bij *abc* dan [*abX*]. Dit inzicht kan worden vertaald in Bybees (2006) opmerking dat het de typefrequentie van een patroon is (dus: het aantal 'nieuwe' unieke gevallen dat ermee gevormd of begrepen wordt) die de genesteldheid ('entrenchment') van dat patroon bepaalt.[4]

Taalkennis, de weerslag van de ervaring met de talige conventies van een gemeenschap, speelt een belangrijke rol in het interpretatie- en productieproces. Het is echter belangrijk om te beseffen dat deze visie niet impliceert dat er geen andere cognitieve processen werkzaam zijn in het verwerken van taaluitingen: bekende fenomenen als het links-rechtsprincipe (ANS: 21.1.2.1) en lineaire modificatie (Bolinger 1957) kunnen moeilijk als taalkennis worden gezien: beide zijn veeleer processen die de interpretatie sturen dan conventionele inhoudelijke informatiebrokken en voor beide kan worden volgehouden dat het geen aangeleerde vermogens zijn, maar

[4] Vooral op syntactisch gebied is de vraag wat een type is en wat een token niet eenduidig te beantwoorden. Voor een conceptuele unificatie van Bybees perspectief met Langackers opvatting over taalverwerking en abstractie en de effecten daarvan in een computationele simulatie van kindertaalverwerving, zie Beekhuizen (2015; sectie 6.3).

dat ze volgen uit het lineaire verwerken van (talige of niet-talige) informatie. Dergelijke processen spelen, naast de beperkingen opgelegd door een verzameling conventies, een belangrijke rol in het verwerken van taal. Op dezelfde manier zijn er zowel aan de signalerende (meestal fonologische) als de gesignaleerde (conceptuele) kant neigingen die de mogelijkheden en waarschijnlijkheden van de klankvorm en conceptuele inhoud van taal beperken. Klankvormen worden logischerwijs beperkt door ons auditieve en vocale vermogen, maar binnen de mogelijkheden zijn er neigingen tot categorievorming die universeel lijken te zijn (zie bv. Schwartz et al. 2005 en Tsuji et al. 2015).

Ons vermogen tot conceptualisatie wordt ook op vergelijkbare wijzen beperkt en gestuurd: op een heel basaal niveau laat dit fenomeen zich illustreren met kleurterminologie. We kunnen maar een beperkt gedeelte van het kleurspectrum zien, en dus conceptualiseren. Binnen dat te conceptualiseren domein zijn er universele voorkeuren voor conceptuele onderverdelingen die gegrond zijn in de saillantie van de golflengte van het licht t.o.v. de kegeltjes in het oog (Kay & McDaniel 1978) en die over historische tijd het ontstaan van kleurterminologiesystemen beïnvloeden. Dit is sturend, maar niet determinerend in het historisch ontstaan van

kleurterminologie: het interageert bijvoorbeeld met de ecologie van de taalgemeenschap en de communicatieve noden die daaruit voortvloeien. Kort door de bocht: als er weinig blauws is om te beschrijven, zal een aparte term voor ‘blauw’ waarschijnlijk niet snel ontstaan in een gemeenschap (Roberson 2006). Op vergelijkbare wijze zijn er verwerkingsmechanismes die bepaalde talige fenomenen waarschijnlijker maken dan andere. Op grond van een voorstelling van talige communicatie als een informatietheoretisch proces waarbij informatie over een ‘ruizig’ kanaal moet worden gecommuniceerd (een *noisy channel*, waarbij door de ruis interpretatiefouten van het signaal kunnen ontstaan), komen Gibson et al. (2013) tot de conclusie dat SOV en SVO de optimale woordvolgordestrategieën zijn om geen verwarring over de rollen van de participanten te laten ontstaan en ze verklaren daarmee de typologische dominantie van deze twee patronen.

We zijn nu in de positie beland om iets te zeggen over de AV1. Aan de ene kant is er de vraag hoe sterk en cognitief relevant de overkoepelende categorie AV1 is. Het bestaan ervan kunnen we trivialiter aannemen. Een gedeeld activatiepatroon tussen alle zinnen die we onder de AV1 scharen is immers eenvoudig te vinden: het zijn ondergeschikte zinnen, en het werkwoord staat voorop. Wellicht dat

zelfs de gedeelde abstracte functionele of conceptuele eigenschappen van de verschillende zinnen geregistreerd worden door het taalverwerkend brein. Maar met Bybee kunnen we ons dan afvragen wat de mate van genesteldheid van dit abstracte patroon is. Gegeven de 'voorrang van het concrete boven het abstracte', kunnen we ons voorstellen dat het vooral concretere patronen dan 'AV1' zijn, die door het gebruik aangesterkt raken. In het corpusonderzoek in de volgende paragrafen bespreek en beargumenteer ik welke meer concrete patronen dit zouden kunnen zijn.

Aan de andere kant roept de conventies-en-neigingen benadering de vraag op wat de juiste analytische plaats is van de functies zoals Daalder, Van der Horst, en Diessel die aan de AV1 toeschrijven. Als taalgebruikers geen sterk 'AV1' patroon hebben, blijft het zo dat voor alle drie de benaderingen wel wat te zeggen is: de functies komen wel min of meer overeen met het type functies dat AV1-zinnen (en V1-zinnen in het algemeen) vervullen. Mijn voorstel is dat we dit echter bij de groep 'neigingen' plaatsen, en niet bij de 'taalkennis' (of conventies, of taaltekens). Dit idee steunt enerzijds op de visie dat de taalkennis rond zinnen die 'AV1' genoemd kunnen worden zeer locaal is georganiseerd, zoals ik in de komende paragrafen uiteen zal zetten, en anderzijds op het gemak waarmee de

functie, zoals voorgesteld door de drie bovengenoemde auteurs, kan worden verbonden aan het lineair verwerken van taal. We kunnen mijns inziens de deiktische spanning, attitudinele verhouding, of specifieke fit van werkwoordinitiële patronen, en dus van de AV1, zien als een voortvloeisel van het vooropstellen, of 'topicaliseren', wellicht 'focaliseren' van het vervoegde werkwoord, en dus van een neiging om lineair te interpreteren. De grondvesting (*grounding*) van de werkwoordelijke betekenis in de gedeelde ervaring van spreker en hoorder wordt dus als het ware in de schijnwerpers gezet. Suggestieve evidentie hiervoor vinden we in de kindertaalverwerving, waar voor zowel het Engels (Sadock 1982) als het Hongaars (MacWhinney 1985) is beargumenteerd dat kinderen (in die talen) ongrammaticale patronen met het werkwoord voorop vormen om optatieve functies uit te drukken (*Eat Benny now* in de betekenis 'dat Ben ete', bijvoorbeeld). Ik laat deze stelling voor deze bijdrage als hypothese staan, die ik niet verder zal proberen te onderbouwen. Nader onderzoek is hiervoor noodzakelijk.

Wel biedt deze aanname een interessante diachrone invalshoek op het ontstaan en ontwikkelen van AV1-patronen: werkwoordsinitialiteit kan gezien worden als een functionele 'niche' (vanuit de 'neigingen') of iconisch principe waarbinnen

conventionele taalpatronen zich kunnen ontwikkelen: ze trekt als het ware patronen aan die (in Daalders termen) deiktische spanning van het predikaat tot de gedeelde kennis van spreker en hoorder uitdrukken.

Samenvattend: mijn stelling is dat, op dat abstracte niveau, de AV1 van beschrijving geen zinnig voorwerp van onderzoek is *als taalteken*, maar wél als een gevolg van de neiging om in de lineaire presentatie van informatie bepaalde keuzes te maken. Het juiste beschrijvingsniveau, of: de zijnsstatus van de hypothetische functies van de AV1 is dus in de 'neigingen' en niet in de 'taaltekens' te vinden. De zinnen die onder de AV1 lijken te vallen, zijn op een lokaler niveau georganiseerd dan 'werkwoordsinitiële voegwoordloze bijzin', wat ik in de komende paragrafen zal bespreken.

3 Corpusonderzoek naar de AV1: corpus en methodes

In de volgende paragrafen laat ik zien hoe de zinnen die onder de AV1 geschaard kunnen worden, het beste beschreven kunnen worden in termen van een verzameling clusters, dus als lokaal georganiseerde patronen. Ik houd me in deze secties grotendeels verre van de

functionele kant van de patronen, maar bekijk vooral de distributionele eigenschappen van de AV1 zinnen.

Met de recente beschikbaarheid van grootschalige syntactisch geannoteerde corpora is de mogelijkheid om de AV1 te onderzoeken sterk toegenomen. De AV1 heeft geen lexicaal gespecificeerde elementen, wat het zoeken op lineair niveau moeilijk maakt. Voor deze verkenning is gekozen voor het Wikipediagedeelte van SoNaR-500 (te vinden op http://lands.let.ru.nl/projects/SoNaR/). SoNaR-500 is een automatisch syntactisch geannoteerd corpus van ongeveer 500 miljoen woorden dat hedendaags geschreven Nederlands bevat.

Om AV1-zinnen te onttrekken aan het corpus kunnen we de tag ‘sv1’ gebruiken, die toegekend is aan AV1-zinnen. Door de automatische annotaties zijn er echter miscategorisaties in het corpus: alle verwijzingen naar andere wikipediapagina’s met *zie tevens <Wikipedia-pagina>* zijn opgemerkt als AV1-zinnen. Deze miscategorisaties zijn automatisch weggelaten door de verdere restrictie op de zoekopdracht te plaatsen dat de knoop van de syntactische boomstructuur die met ‘sv1’ is aangemerkt verder een subject als dochter moet bevatten en in een relatie ‘sat’ (sateliet), ‘mod’ (modificeerder) of ‘--‘ (geen geannoteerde relatie) tot de bovenliggende knoop (de matrixzin) moet staan. Dan nog bevatten de

resulterende zinnen een aanzienlijke hoeveelheid vals-positieve ruis, die handmatig is verwijderd. Dit leverde een steekproef van 4370 AV1-zinnen op.

De drie in de introductie genoemde categorieën zijn vervolgens handmatig toegekend.[5] Ik maakte de verdeling tussen de voorwaardelijke en contrastieve types door te kijken of beide zinnen afzonderlijk asserteerbaar waren in de context (wat het geval is voor het contrastieve type, maar niet voor het voorwaardelijke). Gevallen van het derde type werden niet aangetroffen: deze verwondering-uitdrukkende constructie lijkt een spreektalig fenomeen te zijn.[6]

Het contrastieve type wordt in het vervolg aangeduid met 'ctr-AV1'. Het voorwaardelijke type kunnen we verder uitsplitsen naar de voorwaardelijke AV1 met de verleden tijd van het hulpwerkwoord *mogen* ('mocht-AV1'), die met het hulpwerkwoord *willen* ('wil-AV1'), en de overige gevallen (voor het gemak de 'lexicale' voorwaardelijke AV1 ('vrw-AV1') genoemd, hoewel ook hier een

[5] De data zijn beschikbaar via https://github.com/dnrb/publications/blob/master/data_AV1_zinnen.csv.

[6] In mijn eerdere studie naar de AV1 (Beekhuizen 2009), dat het Twente Nieuws Corpus gebruikte kwamen dergelijke AV1-zinnen opmerkelijk genoeg wel voor (voornamelijk in de weergave van andere sprekers in de directe rede) en maakten ongeveer 5% van alle AV1-zinnen uit. Het genre van krantentaal biedt blijkbaar meer ruimte voor mirativiteit dan Wikipedia.

hulpwerkwoord vooraan kan staan). De eerste twee patronen gedragen zich in hun kwalitatieve distributie anders dan de overige voorwaardelijke AV1-zinnen, bv. door de mogelijkheid tot achteropplaatsing (4), en tot onderschikking aan ondergeschikte zinnen (5).

(4) Hij moet even bellen (*komt hij te laat / mocht hij te laat komen / wil hij nog mee kunnen)

(5) … dat hij (*komt hij te laat / mocht hij te laat komen / wil hij nog mee kunnen) even moet bellen.

Met deze onderverdeling in vieren zien we de frequentieverdeling in Tabel 1. Merk op dat ik deze onderverdeling voor paragrafen 5-7 aanneem, maar in paragraaf 8 juist weer nuanceer.

type	**frequentie**
vrw-AV1	3156
mocht-AV1	506
wil-AV1	249
ctr-AV1	459
totaal	**4370**

Tabel 1. Frequentieverdeling van de vier aangetroffen subtypes in het Wikipediagedeelte van SoNaR.

Een verdere handigheid van de annotaties in het SoNaR corpus is dat andere kenmerken van de zin automatisch onttrokken kunnen worden. In de kwantitatieve verkenning van de data zal gebruik gemaakt worden van (deelverzamelingen van) deze types. Ik heb gekozen de kenmerken in Tabel 2 te onttrekken. Merk hierbij op dat ik de term *matrixzin gebruik* om de zin te benoemen die volgens de SoNaR annotaties de AV1-zin regeert, dus de zin waaraan de AV1 direct ondergeschikt is. Dit hoeft dus niet de matrixzin van de hele, complexe, zin te zijn, zoals we weldra zullen zien.

kenmerk	**mogelijke waardes**
positie	{ vooraan, achteraan, middenveld }
resumptiepatroon	{ integratief, resumptief, niet-integratief, niet van toepassing }
onderwerp AV1	1 of meer woorden
onderwerp matrixzin	0 of meer woorden
zelfstandig werkwoord AV1	1 woord
zelfstandig werkwoord matrixzin	1 woord
hulpwerkwoorden AV1	0 of meer woorden
hulpwerkwoorden	0 of meer woorden

matrixzin	
tempus AV1	{ tegenwoordig, verleden }
tempus matrixzin	{geen, tegenwoordig, verleden }
argumentstructuur AV1	zie documentatie SoNar
argumentstructuur	zie documentatie SoNaR
matrixzin	
eerste postverbaal	{ subject, anders }
element AV1	

Tabel 2. Automatisch onttrokken kenmerken van de AV1-zinnen

Deze kenmerken vormen een enigszins arbitraire collectie. Omdat er over de kwantitatieve tendensen van de verschillende AV1-zinnen nog niets bekend is, leken dit me enigszins voor de hand liggende kandidaten.

Voor 'positie' heb ik de volgorde van elementen in de SoNaR-annotatie gevolgd. Merk daarbij op dat als een element het laatste is, het niet noodzakelijk *niet* in het middenveld staan: bij een ontbrekende tweede zinspool is zonder verdere distributionele toetsen niet uit te maken of de lineair laatste constituent vóór of na de lege tweede zinspool staat. Bij 'resumptiepatroon' volg ik Renmans en Van Belles (2003) codeerschema voor voorwaardelijke *als*-zinnen, waarbij bijzinnen op de eerste zinsplaats van de matrixzin 'integratief' worden genoemd, bijzinnen gevolgd door *dan*

‘resumptief’, en bijzinnen die voorafgaan aan een door een ander element gevulde eerste zinsplaats van de matrixzin ‘niet-integratief’. Dit kenmerk is uiteraard niet van toepassing als de ‘positie’ niet ‘vooraan’ is.

De onderwerpen van matrix- en bijzin zijn reeksen woorden. In Tabel 2 is aangegeven dat het onderwerp van de bijzin altijd minstens één woord bevat, en dat het onderwerp van de matrixzin ook leeg kan zijn. Dit is het geval wanneer de matrixzin een niet-finiete zin is (bijv. *Hij besloot haar te bellen, mocht ze nog later komen*). Bij niet-finiete matrixzinnen is de ‘tempus matrixzin’ tevens ‘geen’. De kenmerken ‘hulpwerkwoorden AV1’ en ‘hulpwerkwoorden matrixzin’ bestaan uit minstens nul, en hoogstens een (theoretisch) oneindig aantal hulpwerkwoorden, zoals deze door de SoNaR-annotatie zijn aangemerkt.

Ten slotte is het kenmerk ‘eerste postverbaal element AV1’ onttrokken. In Beekhuizen (2009) merkte ik op dat bij het contrastieve type vaak niet-subjectsconstituenten op de eerste positie na het vervoegde werkwoord voorkomen.

Alle kenmerken zijn automatisch onttrokken uit automatisch syntactisch geannoteerde taalmateriaal en de annotatie bevat derhalve fouten. Een informele steekproef leert dat de kwaliteit van een heel

degelijk niveau is, maar de fouten die blijven (en waarvoor de manuele correctie nog steeds een omvangrijk project zou behelzen) zullen voor enige ruis in de data zorgen.

4 De verdergaande integratie van mocht-AV1 en wil-AV1

Renmans & Van Belle (2003) stellen dat de mate van vormelijke incorporatie van een bijzin in de matrixzin iconisch is voor de mate van semantische integratie. Hierbij wordt semantische integratie opgevat als de nauwheid van de conceptuele band tussen bijzins- en matrixzinsinhoud. Inhoudelijke voorwaardelijkheid (bv. *Als het regent, worden de straten nat*) wordt daarin als nauwer gezien dan inferentiële voorwaardelijkheid (bv. *Als de auto voor de deur staat, dan zal Jan wel thuis zijn*) die op zijn beurt weer nauwer is dan taalhandelingsvoorwaardelijkheid (bv. *Als je bier wil, dan ligt er een sixpack in de koelkast*). Renmans & Van Belle operationaliseren'vormelijke incorporatie' voor voorwaardelijke *als*-zinnen dan, in navolging van König & Van der Auwera (1988) als de mate waarin de (vooropgeplaatste) bijzin deel uitmaakt van de matrixzin. Wanneer een bijzin de eerste zinsplek van de matrixzin inneemt ('integratief'), is de vormelijke incorporatie het grootst. Iets

minder vormelijke incorporatie vinden we bij een ‘resumptief’ patroon, waarbij de bijzin gevolgd wordt door een resumptief element zoals *dan*. Bij ‘niet-integratief’ ten slotte, wordt de bijzin gevolgd door een eerste zinsplaats van de matrixzin. Voorbeelden uit het corpus, alle van het vrw-AV1 type, zijn hieronder te vinden:

(6) Wordt de Militair daadwerkelijk geraakt, zal een indringende pieptoon hoorbaar zijn [integratief]
(7) Is de lens gedeeltelijk ondoorzichtig, dan spreekt men van onvolledig cataract [resumptief]
(8) Had Benoit het op voorhand geweten, hij had nooit zijn Rubenscantate gecomponeerd [niet-integratief]

Renmans en Van Belle vinden op grond van hun corpus dat de mate van semantische integratie samenhangt met de mate van vormelijke incorporatie. Nu is het zo dat in hun corpus, dus voor *als*-zinnen, de resumptieve patronen een kleine minderheid van alle gevallen vormen. Zoals we in Tabel 3 kunnen zien, is dit anders voor de voorwaardelijke AV1-zinnen.

positie	integratie-patroon	vrw-AV1	mocht-AV1	wil-AV1	ctr-AV1	**totaal**
voor	integratief	112	42	14	7	**175**
		4%	*8%*	*6%*	*2%*	*4%*
	resumptief	3037	278	113	16	**3444**
		96%	*55%*	*45%*	*3%*	*79%*
	niet-integratief	3		1	436	**440**
		0%		*0%*	*95%*	*10%*
midden	n.v.t.		13	6		**19**
			2%	*2%*		*0%*
achter	n.v.t.	4	173	114		**291**
		0%	*34%*	*46%*		*7%*
onduidelijk	n.v.t.			1		**1**
				0%		*0%*
totaal		**3156**	**506**	**249**	**459**	**4370**

Tabel 3. Verdeling van de vier aangetroffen types over de verschillende AV1-posities en integratiepatronen in de matrixzin.

Dit gegeven is opmerkelijk. Het lijkt (in ieder geval anekdotisch) zo te zijn dat voorwaardelijke AV1-zinnen niet minder vaak inhoudelijke voorwaardelijkheidsrelaties uitdrukken dan voorwaardelijke *als*-bijzinnen. Erg is dit echter niet voor Renmans & Van Belles hypothese: als mate van vormelijke integratie iconisch is voor de

mate van inhoudelijke integratie, dan kan het natuurlijk goed zo zijn dat deze 'neiging' ondergeschikt is aan andere neigingen of aan conventies. Voor voorwaardelijke AV1-zinnen zou een andere (wellicht tot zwakke conventie verworden) neiging kunnen zijn dat er geen fonologisch aanwezige markering van de voorwaardelijkheidsrelatie is, en dat het resumptieve element *dan* die rol moet vervullen.[7]

We zien verder dat de status aparte van mocht-AV1 en wil-AV1 gewaarborgd is op grond van hun distributie: beide types kunnen op andere posities dan vooraan voorkomen,[8] en hebben, wanneer wel voorop, een iets grotere voorkeur voor een integratieve volgorde dan vrw-AV1 (13% en 11% van de vooropgeplaatste mocht-

[7] Een verdere reden waarom de AV1-zinnen geen weerlegging van Renmans & Van Belle's idee vormen, zoals een anonieme beoordelaar terecht opmerkt, is dat het onderzoek de twee semantische types (inhoudelijke en inferentiële voorwaardelijkheidsrelaties) als uitgangspunt neemt en daar de iconische uitingen (zoals de integratie tussen matrix- en bijzin) van onderzoekt. Dat deze anders aan de oppervlakte komen bij AV1-zinnen dan bij *als*-zinnen is dan redelijkerwijs te verwachten. Evenwel blijft het opmerkelijk dat integratieve AV1-zinnen zelden voorkomen, gegeven dat ook met dit zinstype frequent inhoudelijke verbanden worden uitgedrukt. Er moet dus verdere redenen zijn waarom dit niet gebeurt (zoals mijn voorgestelde 'vermijd ondermarkering').

[8] De Algemene Nederlandse Spraakkunst (ANS; paragraaf 21.3.1.2) merkt dit overigens alleen op voor de mocht-AV1.

AV1 en wil-AV1, respectievelijk).[9] Dit laatste kan weer verbonden worden aan de neiging een semantische relatie expliciet te maken: *mochten* en *willen* signaleren hier, als frequente hulpwerkwoorden in deze constructie, mogelijk (tevens) het feit dat het om een conditionele zin gaat.

Met deze status aparte van de mocht-AV1 en wil-AV1 zien we voorts het werkwoordsinitialiteit-als-niche idee verder geïllustreerd. De inhouden van wil-AV1s lijken functioneel eerder een doelaanduidende rol ten opzichte van de matrixzinsinhoud dan een voorwaardelijke te vervullen. Dit zien we ook aan het feit dat echt voorwaardelijke AV1s met *willen* als hulpwerkwoord niet achteropgeplaatst kunnen worden (voorbeelden (9)-(10)). Voorts heeft het werkwoord *wil* hier een functie waarin de volitionaliteit van het onderwerp op z'n minst op de achtergrond is komen te staan, zo niet weggevallen is (getuige wil-AV1s met onwillige en niet-bezielde onderwerpen – vb. (11)-(12))

[9] Opgemerkt dient te worden dat de mocht-AV1 in het Wikipediacorpus hiermee een andere distributie van matrixzinsintegratie vertoont dan de mocht-AV1 in gesproken Nederlands (cf. Boogaart 2007), waar de mocht-AV1 (op een enkele, voorgelezen, uitzondering na) niet zonder resumptief element voorkomt.

(9) Wil ze wel paren, dan laat ze het mannetje in haar hol

(9') *Ze laat het mannetje in haar hol, wil ze paren

(10) Wil je een paard vangen, dan moet je een lasso meenemen.

(10') Je moet een lasso meenemen, wil je een paard vangen.

(11) Er moet nog flink wat bewijs gevonden worden, wil de verdachte daadwerkelijk veroordeeld worden.

(12) Wil een boom een mooie houtkwaliteit leveren dan moet zij beschermd staan, bijvoorbeeld in een bos

Hetzelfde geldt m.m. voor mocht-AV1s. Wanneer *mocht(en)* een deontische lezing heeft (d.w.z. 'het is toegestaan dat'), dan is achteropplaatsing niet mogelijk, getuige voorbeelden (13)-(14):

(13) Mocht hij zijn zoontje zien, dan was dat altijd onder toezicht.

(13') *Het was altijd onder toezicht, mocht hij zijn zoontje zien

(14) Mocht hij te laat komen, dan wordt hij ontslagen

(14') Hij wordt ontslagen, mocht hij te laat komen

Van de wil-AV1s en mocht-AV1s lijken de laatste nog verder geïntegreerd te zijn, getuige voorbeelden als (15), waarin de mocht-AV1 een NP lijkt te modificeren. In (15) is de voorwaardelijke relatie

er niet één tussen het ‘genoodzaakt zijn Italië te ontvluchten’ en ‘het vragen om bescherming’, maar tussen het ‘genoodzaakt zijn Italië te ontvluchten’ en de ‘bescherming’: een parafrase zou zijn ‘ze vroeg of hij haar, als ze genoodzaakt was Italië te ontvluchten, zou beschermen’ en niet ‘als ze genoodzaakt was Italië te ontvluchten, vroeg ze of hij haar zou beschermen’, d.w.z.: de gehele propositie ‘als ze genoodzaakt was Italië te ontvluchten, zou hij haar beschermen’ is hetgeen gevraagd wordt. Dezelfde semantische verhoudingen vinden we terug in voorbeeld (15).

(15) […] ze vroeg om bescherming mocht ze genoodzaakt zijn Italië te ontvluchten

We zien dus dat wil-AV1 en mocht-AV1 zich anders gedragen dan vrw-AV1. De distributie is, zowel kwalitatief als kwantitatief, anders dan de vrw-AV1, en de hulpwerkwoorden *willen* en *mochten* vertonen betekenisspecialisatie in de constructie. Daarnaast zouden we kunnen beargumenteren dat beide patronen als geheel ook een andere betekenis vertonen dan de vrw-AV1. De vrw-AV1 wordt vaak gekenmerkt door een bepaalde spanning tot het voorgaande discourse: bijvoorbeeld van het type ‘als A, dan B, als niet-A, dan C’,

waarbij de ‘als niet-A’ dan als vrw-AV1 wordt uitgedrukt, zoals in voorbeeld (16) (zie verder Beekhuizen 2009 en Verheij 2011 voor besprekingen van de discoursekarakteristieken van vrw-AV1).

(16) Als het lukt, sta je in vijf minuten weer buiten. Lukt het niet, dan moet je even bellen.

Mocht-AV1 en wil-AV1 lijken deze functie minder uitgesproken te vertonen (hoewel dit speculatief is, en, totdat deze hypothese degelijk geoperationaliseerd is, slechts een intuïtie is). Wil-AV1 heeft een sterke doelaanduidende smaak, en mocht-AV1 lijkt een vrij typische ‘hypothetische situatie’ uit te drukken, zonder dat dat contrast met het voorgaande discourse sterk aanwezig is. In dit kader is het maar de vraag of de aanname van een abstracte representatie over alle voorwaardelijke AV1-types (vrw-AV1, mocht-AV1, en wil-AV1) een zinnige is: de patronen met *mochten* en *willen* hebben behoorlijk andere eigenschappen en zijn (zeker vanwege de distributionele eigenschappen) hoogstwaarschijnlijk apart van de lexicale AV1 opgeslagen.

In tabel 3 is te zien dat er vier voorbeelden van achteropgeplaatste vrw-AV1 zinnen aangetroffen zijn. Dit betreffen

echter allemaal gevallen van het Belgisch-Nederlandse voorwaardelijke AV1-patroon met *moesten* (cf. Boogaart 2007; een voorbeeld is te vinden in (17)). Vermoedelijk is het ofwel zo dat minder Wikipediaredacteurs Belgisch zijn, ofwel dat Belgische Wikipediaredacteurs de gestigmatiseerde AV1 met *moesten* vermijden.

(17) In “De lachende wolf” (1952) zegt ze dat ze niet naar Canada zou gaan moest Lambik haar ten huwelijk vragen.

Over de ctr-AV1 zijn ook nog enige descriptieve zaken op te merken. De Algemene Nederlandse Spraakkunst (ANS; paragraaf 21.8.2.2) meent dat de ctr-AV1 (aldaar als ‘toegevende bijzin’ aangemerkt) verplicht een niet-integratief patroon heeft. In de SoNaR-data komen we evenwel 23 gevallen tegen, goed voor 5% van alle ctr-AV1 zinnen, met integratieve of resumptieve patronen (waarbij vooral het resumptieve patroon een historische verwantschap met het voorwaardelijke type suggereert en wellicht een ‘bridging context’ tussen vrw-AV1 en ctr-AV1 zonder resumptief element vormt).

(18) Heeft Nederland een spanning van 1500 volt gelijkspanning, hebben de Duitse draden 15 kV op 162/3 Hz wisselstroom. [integratief]

(19) Waren er in 1930 nog 81,2% Nederlandstaligen dan was dit aantal in 1947 tot 42,9% geslonken. [resumptief]

Een verdere brug tussen de ctr-AV1 en de voorwaardelijke types vormen zinnen als (20), die vormelijk en functioneel typische ctr-AV1s zijn maar wel *mocht(en)* als hulpwerkwoord in de bijzin hebben, in een vergelijkbare epistemische functie als de mocht-AV1. Van dit type trof ik vijf gevallen in het corpus aan.

(20) Mocht de adel als juridisch en staatkundig begrip dan niet meer bestaan, zij heeft wel een maatschappelijke rol van betekenis

5 Distributionele profilering met TiMBL

5.1 Motivatie en opzet

In de voorgaande sectie heb ik verscheidene distributionele kenmerken van de verschillende typen laten zien. Voor sommige van de kenmerken (zoals mogelijkheid tot achteropplaatsing en onderschikking) lijken er categorische verschillen tussen de typen te zijn, voor andere kenmerken is het verschil bijna categoriaal (resumptiepatronen, bijvoorbeeld), en voor weer andere, hier nog niet besproken, is het zeer gradueel maar is er wel een tendens waar te nemen (tempus van het werkwoord in matrix- en bijzin, bijvoorbeeld).
Als we uitgaan van de gebruikgebaseerde benadering, zoals geschetst in paragraaf 2, is het echter ook goed mogelijk dat er lokale regelmatigheden zijn. Wat we willen weten is of er inderdaad genoeg lokale regelmaat in de vier types zit om ze van elkaar te onderscheiden. Dit zou ondersteuning bieden voor het idee dat de clusters taalstructureel een eigen leven leiden.

Als methode om dit te doen kunnen we gebruik maken van regressiegebaseerde modellen. Deze modellen gaan er echter van uit

dat de kenmerken in principe onafhankelijk van elkaar opereren (het toevoegen van interacties is mogelijk, maar het toevoegen van alle interacties tussen alle kenmerken maakt het model schier oninterpreteerbaar). Als een methode die recht doet aan het gebruiksgebaseerde inzicht dat lokale interacties van kenmerken tot generaliseerbare patronen leiden, gebruik ik het computationele categorisatiemodel *Memory-Based Learning*, geoperationaliseerd in het programma *TiMBL* (Daelemans & van den Bosch 2005). Dit model is een *categorisatiemodel*, wat wil zeggen dat het op grond van een corpus van gebruiksgevallen nieuwe gebruiksgevallen probeert te categoriseren. Met deze benadering kunnen we een aantal zaken onderzoeken. Ten eerste kunnen we de stabiliteit van de categorieën (i.c. de vier types) onderzoeken: kunnen instantiaties van een categorie goed voorspeld worden? Een goede voorspelling betekent dat er (lokale) regelmaat bestaat die het model gebruikt om tot die juiste categorisatie te komen. Daarnaast kan er binnen de resultaten nader gekeken worden naar de interne structuur van de categorieën. Aangezien het model categoriseert op grond van de gelijkenis van een gebruiksgeval met omringende gebruiksgevallen, kan het zijn dat bepaalde van die omringende gebruiksgevallen vaak gebruikt worden als bron van categorisatie. Als dat gebeurt, zouden we kunnen

spreken van een prototype van een categorie: een beste geval dat als model[10] dient voor vergelijkbare, maar niet identieke zinnen. In paragrafen 5.2 en 5.3 kijk ik naar de prototypestructuren van de verschillende categorieën.

Een volledige bespreking van MBL valt buiten het bereik van dit artikel – ik verwijs de geïnteresseerde lezer naar Daelemans & Van den Bosch (2005) voor een volledige bespreking. De intuïtie achter het model is de volgende: alle tegengekomen gebruiksgevallen worden opgeslagen in het geheugen van het model als punten in een hoogdimensionale ruimte, waar elk kenmerk een dimensie vormt. Wanneer een nieuw gebruiksgeval wordt aangetroffen waarvan de categorie nog niet bekend is, kan een voorspelling van de categorie worden gedaan op grond van alle geziene gebruiksgevallen. TiMBL categoriseert het nieuwe gebruiksgeval op grond van de *k* dichtstbijzijnde gebruiksgevallen in de ruimte. De afstand, of nabijheid, van de gebruiksgevallen wordt berekend door de afstand

[10] Opgemerkt dient te worden dat met de pure categorisatie nog geenszins de productie- of begripstaak is volbracht. Een vollediger computationeel model zou verschillende structuren met elkaar moeten combineren om tot ofwel een volledige taaluiting ofwel een interpretatie te komen. Een categorisatiemodel bestrijkt maar een klein gedeelte van dit proces en is geen volledig *generatief* model. Voor een vollediger, exemplar-gebaseerd computationeel model van grammatica, zie Beekhuizen et al. (2014) en Beekhuizen (2015).

op elke dimensie (kenmerk) te nemen, gewogen naar de saillantie van die dimensie. De afstand op een dimensie is het absolute verschil tussen de waardes van de twee gebruiksgevallen op die dimensie, als het een numerieke dimensie betreft (zoals zinslengte), of een binaire waarde die 1 is bij identiteit en 0 bij verschil, als het een categorische dimensie betreft (zoals tempus). De saillantie van een dimensie of kenmerk wordt door het model bepaald op grond van de Gain Ratio, een informatietheoretische maat die de bijdrage van dat kenmerk aan het correct classificeren van de tot dan toe geziene gebruiksgevallen meet. In de hier gepresenteerde experimenten gebruiken we telkens $k = 1$, dat wil zeggen: een categorisatie van een getest voorbeeld op grond van het enkele, dichtstbijzijnde gebruiksgeval.

5.2 Globaal categorisatiegedrag: de stabiliteit van de subtypes

Laten we eerst kijken naar de globale accuratesse van het model. We meten de globale accuratesse door voor alle zinnen het model te trainen op alle zinnen behalve die ene, en dan de categorie van die ene zin te voorspellen (de zgn. *leave-one-out cross-validation* procedure). Omdat we de waargenomen categorie van die zin in het corpus hebben, kunnen we deze vergelijken met de voorspelde

categorie. Op deze manier verkrijgen we inzicht in het gedrag van het model: als het model de categorieën beter voorspelt dan op grond van kans verwacht zou zijn, 'herkent' het blijkbaar een latente structuur in het samenspel van kenmerken. Wat deze structuur inhoudt, zal dan verder onderzocht moeten worden, wat ik in secties 5.3 en 5.4 zal doen.

Voor dit experiment neemt TiMBL alle kenmerken uit tabel 2 mee, behalve positie en integratiepatroon, en negeert het hulpwerkwoord in de AV1 als dit *willen* of *mogen* is (het meenemen van deze hulpwerkwoorden zou een (bijna) circulaire categorisatietaak tot gevolg hebben). De hulpwerkwoorden zijn gevectoriseerd, wat wil zeggen dat voor de aan- of afwezigheid van elk hulpwerkwoord een aparte kolom is opgenomen, met voor elk gebruiksgeval een waarde van 1 als het hulpwerkwoord aanwezig is, en 0 als het afwezig is.

Voordat we de naar de maten van accuratesse kijken, is het interessant te zien hoe zinnen binnen de waargenomen categorieën gecategoriseerd worden. Dit is weer te geven in een zgn. waarheidsmatrix (*confusion matrix*), waar op de rijen de voorspelde categorieën staan en in de kolommen de waargenomen categorieën. Elke cel geeft dan het aantal gevallen aan waarin die combinatie van

waargenomen en voorspelde categorie voorkomt. Tabel 4 geeft de waarheidsmatrix voor de onderhavige verzameling zinnen, waarbij de correcte categorisatieoordelen in een vette letter zijn weergegeven.

	waargenomen			
voorspeld	ctr-AV1	vrw-AV1	mocht-AV1	wil-AV1
ctr-AV1	**239**	139	75	6
vrw-AV1	112	**2895**	68	81
mocht-AV1	85	101	**307**	13
wil-AV1	4	104	9	**132**

Tabel 4: Waarheidsmatrix voor de vier hypothetische categorieën in het SoNaR corpus.

Wat uit deze waarheidsmatrix af te leiden valt, is dat 3573 van de 4370 gevallen correct gecategoriseerd worden (82%). Verder zien we dat de vaakst voorkomende fouten (in absolute zin) het miscategoriseren van een ctr-AV1 als vrw-AV1 en vice versa zijn. Merk op dat het model geen beschikking had over het kenmerk ‘resumptiepatroon’ omdat we geïnteresseerd zijn in de profielen van de werkwoorden en onderwerpen in matrix- en bijzin, en niet hun (zo goed als verplichte) grammaticale eigenschappen. Voegen we ‘resumptiepatroon’ toe als kenmerk, dan stijgt de accuratesse naar 91%, en verdwijnt dit type fouten vrijwel geheel, aangezien bijna alle

V1-con zinnen een niet-integratief resumptiepatroon hebben, en bijna alle V1-vrw zinnen een resumptief resumptiepatroon.

Hoe ‘goed’ wordt nu elke categorie voorspeld? Ik kwantificeer het globale gedrag van het model met twee maten per categorie, de *precisie* (‘precision’) en de *volledigheid* (‘recall’). De *precisie* voor een categorie *c* geeft weer welke proportie van de gevallen die door het model als *c* gecategoriseerd zijn, ook daadwerkelijk *c* zijn. De *volledigheid*, complementair, geeft de proportie van daadwerkelijke gevallen van *c* die ook als *c* gecategoriseerd zijn. We vergelijken de gemeten *precisie* en *volledigheid* met de *precisie* en *volledigheid* zoals die zouden zijn op grond van een bijna willekeurige gok. De kans dat deze bijna willekeurige gok goed is, berekenen we als volgt. Als we in een categorisatietaak alleen de frequentie van de verschillende categorieën zouden kennen, zouden we op grond van die frequentie gokken. Als een categorie *c* dus 70% van de data uitmaakt, hebben we dus in 70% van de als *c* gecategoriseerde gevallen een juiste gok en categoriseren we 70% van de daadwerkelijke gevallen van *c* correct. We doen dit om een ‘grondlijn’ te hebben: een voorspelde precisie of volledigheid is betekenisloos in isolatie: als een categorie 90% van alle gebruiksgevallen uitmaakt, is het betrekkelijk

gemakkelijk te voorspellen (met een bijna willekeurige gok voorspellen we ook al 90% goed), maar als een categorie zeldzamer is, is deze moeilijker te voorspellen. Tabel 5 geeft de precisie- en volledigheidsscores voor elke categorie, zowel voor het model (TiMBL) als voor de 'bijna willekeurige gok' (kans)

	precisie		***volledigheid***	
	TiMBL	**kans**	**TiMBL**	**kans**
V1-con	0,54	0,12	0,52	0,12
V1-vrw	0,89	0,72	0,92	0,72
V1-mocht	0,67	0,11	0,61	0,11
V1-wil	0,57	0,05	0,53	0,05

Tabel 5. Precisie en volledigheid voor de vier subgroepen.

We zien in tabel 5 dat de categorisatie aardig boven het niveau van kans uitstijgt. Dat het verre van perfect is, doet niet zo veel ter zake: de ruime afname van het aantal foute categorisaties ten opzichte van het kansniveau betekent dat er structuur zit in de (hulp)werkwoorden en onderwerpen van de matrix- en bijzinnen die het model mogelijk maakt om de groepen van elkaar te scheiden. Waar dat aan ligt, en hoe dit de categoriestructuur reflecteert, bekijken we in de volgende sectie.

5.3 Nadere analyse: discriminerende kenmerken

Een eerste verkenning naar waar de kwantitatieve verschillen tussen deze types in zitten begint bij de Gain Ratios, eerder al besproken als de gewichten van de dimensies. De Gain Ratio vertelt ons hoe belangrijk het kenmerk achter een dimensie is, en is als zodanig ook informatief in de analyse. De tien kenmerken met de hoogste Gain Ratio zijn in Tabel 6 weergegeven, waar voor elke waarde het percentage aanwezigheid van dat kenmerk voor elk van de vier groepen is genoemd.

kenmerk	**Gain Ratio**	**waarde**	**con**	**vrw**	**mocht**	**wil**
tempus AV1	0,369	tgw. verl.	14% 86%	83% 17%	0% 100%	78% 22%
hulpwerkwoord	0,229	1	0%	0%	0%	0%
matrixzin = *laten*						
hulpwerkwoord matrixzin = *moeten*	0,209	1	1%	4%	5%	53%
	0,144	tgw. verl.	41% 59%	85% 15%	53% 33%	69% 19%

		geen	0%	0%	14%	12%
tempus hulpwerkwoord matrixzin	0,104	1	0%	0%	0%	0%
matrixzin = *hoeven* hulpwerkwoord	0,101	1	0%	1%	0%	0%
AV1 = *blijven* onderwerp	0,090	zie Tabel 7				
matrixzin onderwerp AV1 hulpwerkwoord	0,089 0,085	1	1%	1%	1%	10%
AV1 = *kunnen* hulpwerkwoord	0,063	1	0%	4%	0%	2%
AV1 = *zullen*						

Tabel 6. De tien kenmerken met de hoogste Gain Ratio's, met voor elke waarde van dat kenmerk het percentage van aanwezigheid van dat kenmerk voor elke subgroep.

type	zin	1	2	3
con	bijzin			
	matrixzin	*het* (18)		
vrw	bijzin	*men* (11)	*we* (8)	*een speler* (5)
	matrixzin	*men* (24)	*we* (16)	*deze* (4)
mocht	bijzin	*iets* (11)	*hij* (9)	*het* (6)
	matrixzin	GEEN (76)	*die* (8)	*hij* (6)
wil	bijzin	*men* (37)	*ze* (12)	*hij* (12)
	matrixzin	GEEN (27)	*je* (5)	*men* (4)

Tabel 7. Significant distinctieve collexemen op de subjectspositie van matrix- en bijzin met een frequentie groter dan of gelijk aan 5. Tussen haakjes de distinctief-collexemische aantrekkingskracht, afgerond op het gehele getal.

Zoals we kunnen zien, is de tempus van de AV1-zin het informatiefste kenmerk. Dit is vrij triviaal, aangezien mocht-AV1s altijd een verleden tijd in de bijzin vertonen. Interessanter is echter dat het contrastieve type ook een sterke voorkeur voor verleden tijden vertoont, zowel in de bij- als matrixzinnen (zie de vierde rij: tempus matrixzin). Het voorkomen van *laten* als hulpwerkwoord in de bijzin is een artefact van het model: aangezien dit kenmerk maar één keer aanwezig is, is de waarde van het kenmerk flink overtrokken. Hetzelfde geldt voor de kenmerken '*hoeven* als hulpwerkwoord van de matrixzin', '*blijven* als hulpwerkwoord in de bijzin', en '*zullen* als

hulpwerkwoord in de bijzin'. Dit geldt echter niet voor de kenmerk '*moeten* als hulpwerkwoord in de matrixzin' en '*kunnen* als hulpwerkwoord in de matrixzin'. Hier zien we dat wil-AV1 zich anders gedraagt dan de drie andere types door de sterke voorkeur voor *moeten* en *kunnen* in de matrixzin bij dit type.

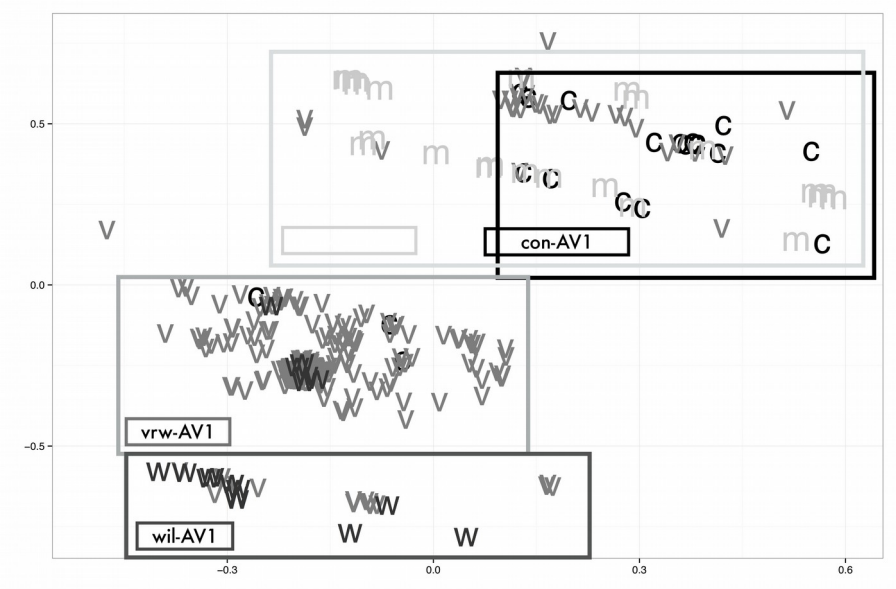
Voor de kenmerken van de onderwerpen zijn er zo veel waardes, dat een zinniger benadering is om een distinctieve collexeemanalyse (Stefanowitsch & Gries 2003) uit te voeren. Deze analyse stelt vast of elementen op bepaalde positie van een patroon beduidend vaker voorkomen dan op dezelfde positie in concurrerende patronen. Hoe hoger de waarde die uit deze analyse volgt, hoe meer aantrekking er is tussen het patroon en het element dat die bepaalde positie vult. In Tabel 7 zien we een aantal opmerkelijke resultaten. Allereerst heeft de contrastieve AV1 vrijwel geen significant aangetrokken collexemen op de subjectposities. Dit komt door de contrastieve aard: vaak staan er op de subjectsposities van bij- en matrixzin volle NPs die twee verschillende entiteiten uitdrukken. De variatie in subjecten is hiermee veel groter dan bij de andere types, waar we een vrij te verwachten reeks voornaamwoorden aantreffen. Ten tweede is het ontbreken van *men* bij de mocht-AV1 opmerkelijk. Waar de andere voorwaardelijke patronen (vrw-AV1 en wil-AV1)

men als significant aangetrokken collexeem op onderwerpspositie van bij- en matrixzin hebben, is dit niet zo voor de mocht-AV1. Wellicht bevatten de inhouden van zinnen met een mocht-AV1 meer specifieke voorwaardelijke verbanden terwijl wil-AV1 en vrw-AV1 vaak over generieke verbanden gaan. Een nadere analyse hiervan laat ik graag aan toekomstig onderzoek over. Ten slotte zien we dat bij de wil-AV1 en mocht-AV1 het onderwerp in de matrixzin vaak ontbreekt: dit is het gevolg van het feit dat deze types aan bijzinnen, inclusief niet-finiete bijzinnen, kan worden ondergeschikt.

5.4 Nadere analyse: categoriestructuur en beste gebruiksgevallen

De analyse van de Gain Ratios geeft ons een inkijk in de voorspellende kenmerken waarmee de vier groepen uit elkaar te houden zijn. Hoe de verschillende categorieën gestructureerd zijn, wordt er evenwel nog niet mee weergegeven. Om hier visueel inzicht in te krijgen, kunnen we alle gebruiksgevallen in een tweedimensionaal vlak projecteren. Omdat de data hoogdimensionaal zijn, moeten we hiertoe een dimensiereductietechniek gebruiken, i.c. Multi-Dimensional Scaling (MDS). De afstanden tussen de

gebruiksgevallen zijn op dezelfde manier berekend als in de TiMBL-analyse, dus met inbegrip van de Gain Ratio's als gewichtsfactoren op de afstand tussen de gebruiksgevallen. Omdat we met meer dan vierduizend gevallen te maken hebben, plotten we alleen die gevallen die minstens tweemaal gebruikt worden als Nearest Neighbor in de categorisatietaak (hiermee vangen we dus de exemplars die een 'voorbeeldrol' vervullen in minstens twee gevallen).



Figuur 1. Multi-Dimensional Scalingrepresentatie van de gebruiksgevallen die twee keer of meer als Nearest Neighbor gebruikt worden in de classificatietaak. V = vrw-AV1, C = ctr-AV1, M = mocht-AV1, W = wil-AV1

Wat deze MDS-representatie laat zien, is dat de vier groepen elk een eigen plek in de ruimte innemen, en dus kenmerken hebben (onder de (hulp-)werkwoorden, tempi, en onderwerpen) die ze onderscheidbaar maken. De omkaderingen geven de vier groepen aan. Van rechtsboven naar linksbeneden zien we het ctr-AV1, mocht-AV1, vrw-AV1 en wil-AV1 cluster.

Als we ons richten op de interne structuur van de categorieën, dan valt op dat de mocht-AV1 een betrekkelijk grote spreiding van gebruiksgevallen heeft, meer zelfs dan het lexicale type. Dit betekent dat (o.m.) de onderwerpen en lexicale werkwoorden van matrix- en bijzin diverser zijn bij de mocht-AV1 dan bij de vrw-AV1. Verbazingwekkend is dit niet, als we de suggestie in herinnering brengen dat vrw-AV1 een specifiekere discourserol vervult dan mocht-AV1s. Mocht-AV1s hebben een minder nauwe voorkeur voor de inhoudswoorden dan de vrw-AV1.

Het mocht-AV1 cluster lijkt drie centrale leden te hebben, d.w.z. gebruiksgevallen die vaak worden gebruikt om andere gevallen mee te categoriseren, hieronder weergegeven in (21)-(23). Opmerkelijk is dat de eerste twee gevallen van onderschikking aan een finiete resp. niet-finiete bijzin zijn.

(21) … dat de wet van toepassing zou zijn op jezelf, mocht men ooit de minnaar worden van een getrouwde vrouw.
(22) … om de ruwe randjes te polijsten en eventueel te fluorideren mocht dat nodig zijn.
(23) Mocht het een sauropode zijn, dan is het de eerste die buiten Engeland is ontdekt.

Ook voor het ctr-AV1 cluster vinden we twee gebruiksgevallen die het frequentst gebruikt worden om andere gevallen mee te categoriseren, weergegeven in (24)-(25). Bij deze gevallen zien we juist dat het resumptiepatroon is dat ze scheidt. Het type met niet-integratieve volgorde is natuurlijk veel frequenter, maar, gegeven deze data, vormt het type met resumptief *dan* toch ook een lokaal prototype.

(24) Was hij bij de aanvang eerder liberaal gezind, dan evolueerde hij over de jaren heen naar het conservatieve katholieke kamp

(25) Was hij in Europa een sensatie van wie iedereen het fijne wilde weten, in eigen land kreeg hij na ' Stockholm ' te maken met racisme

Wat we uit deze analyse kunnen opmaken, is dat de subcategorieën verschillen in hun coherentie (mocht-AV1 is minder coherent dan vrw-AV1) en dat verscheidene types nog nadere onderverdelingen laten zien. Over de nadere onderverdelingen op grond van de hulpwerkwoorden in de vrw-AV1 is nog wel meer te zeggen, hetgeen we in de volgende paragraaf zullen doen.

6 Latente clusters: de hulpwerkwoordelijke subtypes

Leuschner & Van den Nest bespreken hoe het Engelse AV1-patroon, beperkt tot enkele hulpwerkwoorden, verder gegrammaticaliseerd is dan het Duitse. De diachrone grammaticalisatie van de hulpwerkwoordelijke types is natuurlijk alleen mogelijk als die types

aanvankelijk kwantitatief, en later kwalitatief afwijken van het hoofdtype. We zien dit natuurlijk het duidelijkst voor de wil-AV1 en mocht-AV1, die hier als aparte subtypes zijn aangemerkt. We kunnen ons echter ook afvragen of andere hulpwerkwoorden soortgelijk gedrag vertonen. Als dat zo zou zijn, hebben we verdere evidentie voor de lokale organisatie van het AV1-patroon in het Nederlands.

6.1 Een experiment met TiMBL

Ook deze vraag kunnen we weer met het categorisatiemodel van TiMBL beantwoorden. In dit geval is de categorisatietaak een nog wat kunstmatigere: kunnen we het model laten herkennen of een zin een bepaald hulpwerkwoord bevat of niet. Dat wil zeggen: we geven het model, voor elk hulpwerkwoord dat in de AV1 voorkomt, de taak om elk geval op grond van alle andere *n*-1 voorbeelden te voorspellen. Als het model dit goed doet, betekent dit dat er in de overige kenmerken (andere hulpwerkwoorden, hulpwerkwoorden in de matrixzin, zelfstandige werkwoorden, onderwerpen, integratiepatroon en positie) informatie zit die de ‘categorieën’

onderscheidbaar maakt. Natuurlijk is het kenmerk van het betreffende hulpwerkwoord in de AV1 zelf weggelaten (dit zou wederom een circulaire taak tot gevolg hebben).

hulpwerkwoord	***n***	**precisie**		**volledigheid**	
		kans	**MBL**	**kans**	**MBL**
hebben	129	0,03	0,50	0,03	0,45
kunnen	64	0,02	0,13	0,02	0,09
mogen	511	0,13	0,84	0,13	0,77
moeten	36	0,01	0,04	0,01	0,03
willen	241	0,06	0,62	0,06	0,60
worden	467	0,12	0,75	0,12	0,52
zullen	149	0,04	0,59	0,04	0,54
zijn	226	0,06	0,49	0,06	0,25

Tabel 8. Classificatie van de hulpwerkwoorden met TiMBL.

Tabel 8 presenteert de resultaten voor de hulpwerkwoorden die meer dan 30 maal voorkomen. Het kansniveau geeft, net als in het vorige experiment, telkens aan wat de verwachte accuratesse is, als we op grond van de frequentie zouden categoriseren. Voor *mogen* en *willen* wordt, zoals te verwachten is op grond van de analyses in de voorgaande paragrafen, een flinke vooruitgang ten opzichte van het kansniveau behaald. Dit geldt echter ook voor de gevallen met *hebben, zijn, zullen,* en *worden*, waarvoor we ook behoorlijke

verbeteringen ten opzichte van het kansniveau zien. Dit betekent dat deze (tentatieve) categorieën in hun kenmerken zodanig van de overige vrw-AV1 afwijken dat een model ze kan herkennen zonder te weten dat het hulpwerkwoord in kwestie er in staat. De mogelijkheid dat er uit deze lokaal coherente clusters verder gegrammaticaliseerde types ontstaan, is denkbaar.

6.2 Verdere evidentie

En deze mogelijkheid is niet alleen denkbaar, maar zelfs marginaal reëel. Wanneer we op google zoeken, vinden we voor elk hulpwerkwoord wel een voorbeeld van achteropplaatsing van de AV1 met dat hulpwerkwoord (voorbeelden (26)-(30)). Dit betekent dat de vrw-AV1 met hulpwerkwoord zich anders gedraagt dan de vrw-AV1 met een vervoegd lexicaal werkwoord. Mijn impressie van de voorbeelden op google is dat dit type vooral in het Belgisch Nederlands, en dan vooral met voorwaardelijke onderwerpszinnen voorkomt,[11] wat de interessante situatie zou suggereert dat de AV1 in

[11] Een Belgische informant achtte al de zinnen (26)-(30) acceptabel, maar de achteropplaatsing van een vrw-AV1 met een lexicaal vervoegd werkwoord (bv. *Het is jammer lukt het niet*) niet. Meerdere Nederlandse informanten, waaronder een beoordelaar van dit stuk, voelen ook een contrast in aanvaardbaarheid van zinnen

het Belgisch Nederlands verder gegrammaticaliseerd is dan in het Nederlands Nederlands. Dit idee resoneert met Boogaarts (2007) analyse dat zowel de AV1 met *moesten* als die met *mochten* in het Belgisch Nederlands een hogere mate van zinsintegratie en functiespecialisatie vertonen dan de mocht-AV1 in het Nederlands Nederlands.

(26) Maar het zou inderdaad triest zijn, was het echt geweest. [http://www.dumpert.nl/mediabase/6637990/01c53152/me_er_vs_gevaarlijke_hooligan.html]

(27) Het zou bizar zijn zou het 4k zijn. [http://tweakers.net/nieuws/100519/lg-introduceert-219-monitor-met-amd-freesync-voor-gamers.html]

(28) Het zou beter zijn kon het papier en karton droog gemaald worden en dan gemengd worden met het houtzaagsel. [http://www.oaza.be/index.php/startpagina]

(29) maar het had mooi geweest had dat ergens gestaan.

---

van het type (26)-(30) en achteropgeplaatste vrw-AV1s met een lexicaal vervoegd werkwoord. Voor mij zijn ze simpelweg niet grammaticaal: er lijkt dus, in ieder geval, een schaal te zijn van de mate waarin achteropgeplaatste vrw-AV1s aanvaardbaar gevonden worden, waarbij in het Belgisch Nederlands de taalgebruikers zinnen van het type (26)-(30) acceptabeler vinden dan taalgebruikers van het Nederlands Nederlands.

[http://www.chatnrun.nl]

(30) Maar ja, het zou mooi geweest zijn werden ze in die tijd niet meer dan eens duchtig vervalst ... .
[http://belgiangourmand.typepad.com/]

6.3 De niches van de hulpwerkwoordelijke voorwaardelijke AV1

De combinatie van de discrimineerbaarheid van de hulpwerkwoordelijke subtypes met het bestaan van gevallen waarin die subtypes een vrijere distributie ten opzichte van de matrixzin vertonen, suggereert dat de voorwaardelijke AV1-types op een lokaler niveau worden opgeslagen door taalgebruikers dan als 'voorwaardelijke AV1'. Immers, zonder een afwijkende distributie van de subtypes zou de vrijere distributie zoals in sectie 6.2 aangegeven nooit kunnen zijn ontstaan.

Het lijkt er daarnaast op dat de hulpwerkwoorden in specifieke functies gespecialiseerd raken in de AV1: *hebben* en *zijn* in verleden tijden als tegenfeitelijke conditionele zinnen, *zou* in een tegenfeitelijke lezing of één vergelijkbaar met *mocht.* Of *worden* al functiespecialisatie vertoont, was mij niet direct duidelijk, maar gegeven de hoge mate van precisie en volledigheid waarmee dit

patroon voorspeld kan worden, lijkt het me dat er nog eens goed naar AV1s met *worden* moet worden gekeken.

7 Slot

In deze bijdrage heb ik laten zien dat de AV1 geen (sterk) taalteken is. De zinnen die onder de AV1 lijken te vallen, zijn op een lokaler niveau georganiseerd dan ‘werkwoordsinitiële voegwoordloze bijzin’. Hiervoor heb ik de verschillen in mate van zinsintegratie (paragraaf 3), de verschillende distributies van de vier bekende subtypes (paragrafen 4 en 5), en de aanwezigheid van latente clusters van kenmerken rond de gebruikte hulpwerkwoorden in de AV1 (paragraaf 6) aangedragen. De gradueel afwijkende distributies leiden tot systemen van organisatie die door een computationeel categorisatiemodel redelijk goed herkend kunnen worden, wat suggereert dat er in de geproduceerde gebruiksgevallen, zoals aangetroffen in het corpus, lokale coherentie zit. Het is dan ook goed denkbaar dat ook de taalgebruiker de verschillende patronen zo opgeslagen heeft.

Houdt dit in dat de voorstellen van functionalisten als Daalder (1983), Van der Horst (1995), en Diessel (1997) ongegrond zijn? Ik

meen van niet, maar het is wel zaak de juiste plaats voor deze functie te vinden. Vanuit het perspectief dat in het lineair presenteren van informatie de eerste zinsplek een prominente rol inneemt, is het denkbaar de drie analyses te zien als iconische 'neigingen' tot interpretatie.[12] Werkwoordsinitialiteit zou dan diachroon een niche zijn die constructies aantrekt en waarin constructies zich ontwikkelen die deiktische spanning, gemarkeerde attitudinele verhoudingen, en niet-wereld-naar-woorden-*fit* uitdrukt. Het verschil tussen het Belgisch Nederlands en Nederlands Nederlands, zoals getoond in Boogaarts (2004) analyse van de AV1 met *moesten* en *mochten*, en de voornamelijk Belgisch Nederlandse gevallen van achteropplaatsing van een voorwaardelijke AV1 met een hulpwerkwoord laten zien dat we werkwoordsinitialiteit als signaal beter in de hoek van de niet-conventionele 'neigingen' dan in de hoek van de conventionele taaltekens kunnen zien.

[12] Hoewel Van der Horst de positie van het vervoegde werkwoord als taalteken opvat, lijkt hij hier toch niet zeker van te zijn, getuige Van der Horst (1995: 274, vn. 76). Daar legt Van der Horst het verband tussen V1 en de topicalisatie van constituenten in V2-zinnen. Beide maken, aldus Van der Horst, datgene wat voorop wordt gezet het psychologisch onderwerp. Aangezien die topicalisatie in Van der Horsts opvatting geen taalteken is, kan het betwijfeld worden of V1 dat ook is. Ik dank Freek Van de Velde voor deze observatie.

Auteursinformatie

Barend Beekhuizen
Department of Computer Science
University of Toronto
barend@cs.toronto.edu