Regelmaat in een regelloos systeem. De Nederlandse superlatief

Folgert B. Karsdorp & Barend F. Beekhuizen

Trefwoorden: analogie, Memory-Based Learning, superlatief, collocationele aantrekkingskracht, alternantie

Samenvatting

In this paper, we show that analogy is a possible cognitive mechanism behind linguistic categorization. Given a formalized stochastic model that shows preference for certain forms as analogues over other forms, the argument that analogy is a void notion because it is unrestricted (Chomsky 1986) no longer applies. As a test case we tried to predict the alternation between the two patterns of forming the superlative in Dutch. Using the Memory-Based Learning model (Daelemans & Van den Bosch 2005), we can predict 67% of the variation on the basis of three features. In a follow-up experiment, the effect of a fourth factor, collocational attraction between the adjective and the noun, was investigated. Starting from a different level of exemplar granularity, this model accounted for 93% of the cases in the dataset.

1 Regels of analogie?

In de geschiedenis van de taalkunde is het een prominente opvatting dat analogie het onderliggende mechanisme is van taalproductie en taalverwerking. Analogie is dan het cognitieve mechanisme waarmee een taalgebruiker talige elementen in zijn geheugen vergelijkt en op basis van deze vergelijking nieuwe woorden vormt of interpreteert. Het traditionele formalisme van analogie, dat we bijvoorbeeld al terugvinden bij Hermann Paul (1920), is proportionele analogie (in de Engelse literatuur ook wel four-part analogy genoemd). In de proportionele analogie worden analogieën genoteerd als onder 1.

(1) $A : B :: C : D$

Deze notatie wil zeggen dat A zich verhoudt tot B zoals C tot D. Zo verhoudt *verspelen* zich bijvoorbeeld tot *spelen* zoals *verbellen* tot *bellen*. Het is de proportionele verhouding tussen de vormen waardoor de vormovereenkomst, in dit geval *ver-*, betekenis krijgt (vgl. Hüning (1998, 24)). Proportionele analogie hoeft zich niet te beperken tot twee paren van woorden, maar kan zich ook uitbreiden naar grotere analogische verzamelingen woorden, zoals in (2) en in (3):

(2) $A : B :: C : D :: E : F$

(3) $A : B : C :: D : E : F$

Bij de uitbreiding in (2) kunnen we denken aan een toename van het aantal voorbeelden: aan de vergelijking '*spelen* staat tot *verspelen* zoals *bellen* staat tot *verbellen*', voegen we toe dat op dezelfde manier ook *verdobbelen* zich tot *dobbelen* verhoudt. Bij de uitbreiding in (3) gaat het om de toevoeging van een derde vergelijkingspunt tussen de woorden. Een voorbeeld is dat *dobbelen* zich tot *verdobbelen* en *gedobbel* verhoudt zoals *bellen* tot *verbellen* en *gebel*.

De algemeenheid en flexibiliteit van de proportieformule maken het tot een maken het tot een aantrekkelijk hulpmiddel om verbanden tussen woorden bloot te leggen. Als voorspellend model is proportionele analogie echter problematisch omdat het model geen restricties oplegt aan mogelijke proporties. Neem een werkwoord als *wijken*. Op basis van de proportieformule onder (4) zou een taalgebruiker tot de verleden tijd weken kunnen komen:

(4) $\left\{\begin{array}{ll} \text{blijken} & : \text{bleken} \\ \text{lijken} & : \text{leken} \\ \text{kijken} & : \text{keken} \\ \text{strijken} & : \text{streken} \end{array}\right\} = \text{wijken} : \text{weken}$

De verleden tijd *weken* kan echter ook gebaseerd zijn op een verzameling modelwoorden als in (5):

(5) $\left\{\begin{array}{ll} \text{krijgen} & : \text{kregen} \\ \text{zwijgen} & : \text{zwegen} \\ \text{stijgen} & : \text{stegen} \\ \text{splijten} & : \text{spleten} \\ \text{slijten} & : \text{sleten} \end{array}\right\} = \text{wijken} : \text{weken}$

De verzameling woorden onder (4) die eindigen op *-ijk* lijkt een overtuigender groep modelwoorden voor de verleden tijd weken dan die onder (5) omdat de

werkwoorden uit deze groep meer overeenkomst met *wijken* vertonen. Het probleem is echter dat het model deze mate van overeenkomst helemaal niet nodig heeft maar ook niet beloont (vgl. Albright (2009, 187-190)). Het model beschikt, met andere woorden, niet over de formele middelen om waarschijnlijke modelwoorden van minder waarschijnlijke te onderscheiden.

Het belangrijkste bezwaar vanuit regelgebaseerde theorieën tegen het idee van analogie is dan ook dat het te weinig voorspellende kracht zou hebben (zie bijvoorbeeld Albright & Hayes (2003, 119)). Een treffend citaat in dit verband is afkomstig van Becker (1990, 23): "[...] es gibt offenbar keine unmöglichen Analogiebildungen!" Elke analogievorming is dus een mogelijke. Precies in deze onbegrensdheid zit het probleem. Als elke analogie mogelijk is, hoe kunnen we dan verklaren dat bepaalde fenomenen juist niet optreden? Chomsky (1986, 32) komt daarom tot de conclusie: "This idea [analogie] is not wrong but rather vacuous [...]". Toch is het de moeite waard om te onderzoeken of alle analogiemodellen deze onbegrensdheid kennen. Het evidente voordeel van een analogiemodel is namelijk de manier waarop het met uitzonderingen kan omgaan. In een analogiemodel vormt en begrijpt de taalgebruiker met één mechanisme zowel de regelmatige als de onregelmatige gevallen. In een traditioneel regelmodel daarentegen zijn er verschillende mechanismes nodig om de uitzonderingen en de regelmatige gevallen te verklaren. De uitzonderingen worden geplaatst in het mentale lexicon, terwijl de regelmatige gevallen op basis van grammaticale regels worden gevormd (zie bijvoorbeeld Pinker (1999)).

De uitdaging voor een model van analogie ligt er dus in een verklaring te vinden voor het feit dat sprekers sommige generalisaties wel en andere niet maken (Albright & Hayes 2003). Welke woorden kunnen wel als modelwoord dienen voor een analogie en welke niet? In de morfologie zijn de afgelopen twee decennia verschillende computationele modellen van analogische inferentie ontwikkeld die een oplossing proberen te vinden voor de onbegrensdheid van analogie in de traditionele formalismes. Voorbeelden daarvan zijn het model van Analogical Modeling of Language van Skousen (1989) en het Memory-Based Learning model van Daelemans & Van den Bosch (2005). Deze modellen verdisconteren de ongelijke waarschijnlijkheid van bepaalde analogievormingen. Een belangrijk voordeel van deze computationele modellen is in de eerste plaats dat er, gegeven een aantal geformaliseerde aannames, objectievere toetsing mogelijk is van de werking van analogie in morfologische systemen. Daarnaast dienen de modellen een descriptief doel, omdat op empirische basis fijne morfologische patronen zichtbaar gemaakt kunnen worden.

De ontwikkeling van de genoemde modellen heeft tot verscheidene studies naar de werking van analogie geleid, met name in synchrone fenomenen. Zo onderzoeken Krott, Baayen & Schreuder (2001) de invloed van analogie op de keuze tussen de verschillende bindfonemen in Nederlandse samenstellingen. An-

dere voorbeelden zijn Keuleers, Sandra, Daelemans, Gillis, Durieux & Martens (2007), die de invloed van analogie op meervoudsvorming in het Nederlands proberen te modelleren en Chapman & Skousen (2005), waarin de werking van analogische processen in taalverandering wordt onderzocht.

De meeste toepassingen van de genoemde modellen betreffen fonologische en morfonologische fenomenen. Dergelijke fenomenen kenmerken zich doorgaans door een hoge mate van regelmaat. Ze zijn daarom bij uitstek geschikt als casus voor een emergerend morfologisch systeem. In dit artikel willen wij laten zien dat het idee van analogie als voorspellend model van taalproductie ook voor talige patronen die minder regelmatig zijn, een vruchtbaar perspectief biedt. Als casus voor ons betoog nemen wij de Nederlandse superlatief. In het Nederlands wordt de superlatief gevormd met behulp van het suffix *-st(e)* of met een omschrijving met *meest*. We laten zien dat de keuze voor een morfologische of een perifrastische superlatief te voorspellen is op basis van analogie.

De opbouw van het artikel is als volgt. In paragraaf 2 geven we een beschrijving van de superlatief in het Nederlands en bespreken we enkele eerder genoemde factoren die mogelijk de keuze tussen de twee superlatieven beïnvloeden. Omdat de tendensen in de data erg gradueel zijn, is een probabilistisch categorisatiemodel een waarschijnlijke kandidaat om nieuwvormingen mee te voorspellen. In paragraaf 3 bespreken we het gekozen analogiemodel, Memory-Based Learning van Daelemans & Van den Bosch (2005). Paragraaf 4 geeft de resultaten van het onderzoek. We sluiten af in paragraaf 5 met onze belangrijkste bevindingen en conclusies.

2 De Nederlandse superlatief

Het Nederlands beschikt over twee manieren om de superlatief te vormen: door middel van het derivationele *-st(e)* (*het oudste boek*) en een omschrijving met *meest* (*de meest populaire politicus*). In de weinige literatuur die dit onderwerp bespreekt gaat de aandacht voornamelijk uit naar de morfologische superlatief met *-st(e)*. De perifrastische overtreffende trap komt slechts marginaal aan bod. De gebruiksfrequentie van de twee superlatieven in de componenten *a* tot en met *n* van Corpus Gesproken Nederlands[1] lijkt deze aandachtsverdeling te rechtvaardigen: de morfologische superlatief komt ongeveer 10 keer zo vaak voor als de perifrastische.[2] Kijken we echter naar het aantal unieke gevallen van een bepaalde

[1] Omdat we in het volgende alleen nog de componenten *a* tot en met *n* uit het Corpus Gesproken Nederlands gebruiken, korten we deze benaming af tot *het corpus*.

[2] Het Nederlands bevindt zich in dit opzicht tussen het Engels en het Duits in. In het Engels wordt veel meer van de omschrijvende superlatief gebruik gemaakt, terwijl in het Duits uitsluitend de morfologische superlatief mogelijk is.

superlatief	tokens	types
morfologisch	2041	154
perifrastisch	236	159

Tabel 1: Token- en typefrequentie attributief gebruikte superlatieven in het corpus.

superlatief, de typefrequentie, dan ontstaat een geheel ander beeld waarin juist de perifrastische superlatieven de meerderheid vormen.

De verhouding tussen tokens en types kunnen we beschrijven in termen van de waarschijnlijkheid op een nieuw type naarmate we meer tokens tegen zijn gekomen. Het mag duidelijk zijn dat de waarschijnlijkheid van een nieuwe perifrastische superlatief vele malen hoger is dan die van een nieuwe morfologische superlatief. In navolging van o.a. Baayen (1991) kunnen we ook zeggen dat de potentiële productiviteit van de perifrastische superlatief hoger is dan de morfologische. [3]

De twee superlatieven hebben een zeer vergelijkbare functie. Op het eerste gezicht is niet duidelijk wat het verschil is tussen de twee of welke factoren de keuze tussen een morfologische en een perifrastische superlatief determineren. In de E-ANS (6.4.3.1.ii) wordt het gebruik van de perifrastische superlatief als ondergeschikt beschouwd aan de morfologische. De uitgangssituatie is dat een morfologische superlatief wordt gebruikt, tenzij aan een bepaalde conditie wordt voldaan. Daarbij moeten we denken aan bepaalde codas van het adjectief. Als een adjectief eindigt op *-st*, *-s*, *-sk* of *-de*, gaat de voorkeur uit naar een omschrijvende overtreffende trap:

(6) a *de meest robuuste oplossing* (naast ?*de robuustste oplossing*)

b *de meest problematische buurt* (naast ?*de problematischste buurt*)

c *het meest groteske schepsel* (naast ?*het groteskste schepsel*)

d *het meest stupide antwoord* (naast ?*het stupideste antwoord*)

Een volgende factor die het gebruik van de morfologische superlatief tegenwerkt, is het aantal lettergrepen van het adjectief. De E-ANS spreekt over de neiging om een omschreven overtreffende trap te gebruiken naarmate het aantal

[3] Baayen's (1991) productiviteitsindex P berekenen we door het aantal woorden met een frequentie van 1 te delen door de totale tokenfrequentie. Hoe hoger de index, hoe hoger de potentiële productiviteit. In ons corpus is dat voor de morfologische superlatief $P = \frac{65}{2041} = 0.032$ terwijl voor de perifrastische superlatief de index $P = \frac{113}{236} = 0.479$.

lettergrepen van een adjectief toeneemt. We zeggen niet zo snel (7a) maar wel (7b):

(7) a ?*de meest warme dag* (naast *de warmste dag*)

b *de meest spectaculaire stunt* (maar k *de spectaculairste stunt*)

Tot slot noemt de E-ANS nog dat een omschrijving met meest gebruikt kan worden om nadruk te leggen:[4]

(8) a *de méést complete krant van Nederland.*

b *Het is wel de meest arrogante kwal die ik ken.*

2.1 De gegevens

We hebben nu een aantal factoren gezien die van invloed kunnen zijn op de keuze tussen een morfologische en een perifrastische superlatief. In deze paragraaf proberen we deze invloed wat preciezer in kaart te brengen. We stellen ons de vraag hoe sterk de verbanden zijn tussen de fonologische factoren (coda en aantal lettergrepen) en de keuze voor een van beide vormingspatronen.

Laten we beginnen te onderzoeken of het aantal lettergrepen van een adjectief van invloed is. Tabel 2 geeft de typefrequentie van de morfologische en perifrastische superlatieven in het corpus, uitgesplitst naar het aantal lettergrepen van het gebruikte adjectief.

	superlatief	
aantal lettergrepen	morfologisch	perifrastisch
1	74	17
2	44	44
3	30	64
4	6	27
5	0	7

Tabel 2: De typefrequentie van morfologische en perifrastische superlatieven in het corpus, uitgesplitst naar aantal lettergrepen van het adjectief.

[4]Deze factor is niet geheel indiscutabel en vooral moeilijk te onderzoeken. Hoe kunnen we er zeker van zijn dat iemand nadruk legt, en wat weerhoudt de taalgebruiker ervan het hoofdaccent van het adjectief extra te beklemtonen om zo meer nadruk te leggen op het overtreffende karakter van de toegeschreven kwaliteit (vgl. *de compléétste krant van Nederland*)? We zullen deze factor dan ook niet verder onderzoeken.

De gegevens laten een duidelijke tendens zien waarbij het aantal perifrastische superlatieven stijgt naarmate het aantal lettergrepen toeneemt. De groep met de zwakste tendens, namelijk geen, is die van tweelettergrepige adjectieven. Beide vormingsprocedés hebben hier een typefrequentie van 44. Hier komen we later nog op terug. Verder valt nog de vrij grote groep perifrastische superlatieven met één lettergreep op. Dergelijke gevallen ontkennen het absolute karakter van de invloed van het aantal lettergrepen op de keuze tussen de twee vormingspatronen van de superlatief.

Laten we nu de tweede genoemde factor, de coda van de laatste lettergreep, bestuderen. In de E-ANS wordt slechts een aantal eindklanken genoemd dat kan aansturen op een perifrastische superlatief. Voor een volledig beeld moeten we echter ook alle andere eindklanken in beschouwing nemen om te bepalen hoe sterk de invloed van de coda is als geheel. Tabel 3 geeft een overzicht van de gegevens:

superlatief	coda														
	d	f	g	j	k	l	∅	m	n	ng	p	r	s	t	w
morfologisch	15	6	35	3	24	8	0	10	8	4	5	13	1	19	3
perifrastisch	24	4	17	1	28	21	1	5	8	1	2	8	21	18	1

Tabel 3: Typefrequentie morfologische en perifrastische superlatieven in het corpus, uitgesplitst naar de coda van de laatste lettergreep van het adjectief (in deze tabel orthografisch weergegeven, waarbij ∅ voor het ontbreken van een coda, dus een uitgang op een vocaal, staat).

We zien dat adjectieven met eindklank /s/ een sterke voorkeur voor de perifrastische superlatief vertonen.[5] Voor de andere eindklanken geeft de tabel geen eenduidig beeld maar zijn er wel tendensen.

Concluderend kunnen we stellen dat er duidelijke tendensen zijn, maar dat ze verre van absoluut zijn. Een categoriaal regelmodel ligt daarom op basis van deze factoren niet voor de hand.

2.2 De herkomst van het adjectief?

Een derde waarschijnlijke determinant is de herkomst van het adjectief. [6] Er is vaak op gewezen dat uitheemse woorden zich in verschillende opzichten anders

[5]Merk op dat ook hier een tegenvoorbeeld bestaat, namelijk mals in 'de *malste filet-pure*' (CGN:fv400623.39). Een interessant voorbeeld, omdat er naast de eindklank /s/ ook van semantische ambiguïteit sprake is.

[6] Ariane van Santen wees ons op deze mogelijke factor, waarvoor onze dank. Naast deze variabele hebben we nog een reeks andere factoren onderzocht die echter geen van alle effect leken te hebben. Zo hebben we bijvoorbeeld naar het klemtoonpatroon van de adjectieven gekeken. Deze

gedragen dan inheemse woorden. Zo hechten inheemse achtervoegels zich doorgaans beter aan inheemse woorden en voelen uitheemse affixen zich beter thuis bij uitheemse woorden (zie bijvoorbeeld Booij (2007, 71-72). Het zou daarom interessant zijn te onderzoeken of in het geval van uitheemse adjectieven als *bizar*, *neutraal* of *extreem*, in plaats van het inheemse achtervoegsel *-st(e)* het omschrijvende *meest* wordt gekozen.

De uitheemse herkomst van een woord kan echter niet altijd eenduidig worden vastgesteld. Sommige van oorsprong uitheemse woorden hebben door de eeuwen heen hun burgerrecht verworven en kunnen niet langer beschouwd worden als uitheems. Hierom hebben wij de herkomst van de adjectieven op een indirecte manier proberen te benaderen, namelijk door naar het klinkerpatroon van de adjectieven te kijken. Inheemse, van oorsprong Germaanse woorden kenmerken zich (vaak) door een ander klinkerpatroon dan uitheemse woorden. Zo vinden we bij inheemse woorden vaak een gereduceerde klinker in de laatste lettergreep, terwijl veel uitheemse woorden daar juist een volle klinker hebben, vergelijk:

(9) a *ernstig*, *hevig*, *spannend*, *aardig* (vol – gereduceerd)

b *bizar*, *actief*, *brutaal*, *intiem*, *primair* (vol vol)

De correlatie tussen het klinkerpatroon en de herkomst is zeker niet honderd procent (denk aan morfologisch complexe inheemse adjectieven als *ondiep*, *grijpgraag* of *drinkbaar*), maar we kunnen het klinkerpatroon van een adjectief opvatten als een symptomatische indicator van de herkomst van een woord.

Speelt het klinkerpatroon een rol bij de keuze voor een perifrastische of morfologische superlatief? Met name voor bisyllabische adjectieven lijkt dat zo te zijn. Bekijk tabel 4:

We zien dat woorden met twee volle klinkers (typisch uitheems) een duidelijke voorkeur hebben voor perifrastische superlatieven. Woorden met een gereduceerde klinker in de laatste lettergreep (typisch inheems) worden juist aangetrokken door de morfologische overtreffende trap. Het is interessant om te zien dat waar het aantal lettergrepen van een adjectief geen voorkeur laat zien voor een perifrastische of morfologische superlatief, namelijk voor bisyllabische woorden, het klinkerpatroon (en dus indirect de herkomst) van het adjectief een sterk bepalende factor is. De factoren vullen elkaar aan en zijn vermoedelijk samen beter in staat de keuze tussen een van de twee superlatieven te voorspellen. Dergelijke elkaar aanvullende factoren duiden erop dat een multifactoriële, probabilistische

factor heeft weliswaar enig effect, maar dat wordt tenietgedaan in combinatie met het klinkerpatroon van de adjectieven. Ook hebben we onderzocht of priming een mogelijke factor is. De resultaten lieten zien dat de keuze voor een van de twee superlatieven niet bepaald wordt door een eventuele direct voorafgaande superlatief. Verder hebben een reeks sociolinguïstische variabelen onderzocht, waaronder regio, opleidingsniveau en register. Ook deze factoren hadden geen effect.

klinkerpatroon	superlatief	
	morfologisch	perifrastisch
$++$	9	26
$+-$	26	10
$-+$	9	8

Tabel 4: Typefrequentie morfologische en perifrastische superlatieven op basis van bisyllabische adjectieven in het corpus, uitgesplitst naar klinkerpatroon (waarbij '$+$' voor een volle vocaal staat en '$-$' voor een gereduceerde).

aanpak geschikter is dan een categoriale om de geobserveerde variatie in het gebruik van de superlatief te beschrijven. In de volgende paragraaf zullen we het specifieke probabilistische model bespreken dat we voor ons onderzoek hebben gebruikt: het model van *Memory-Based Learning*, zoals ontwikkeld door Daelemans & Van den Bosch (2005).

3 Het model

Als taalgebruikers door middel van analogisch redeneren bepalen welke nieuwvorming ze produceren, moeten er beperkingen zijn op de verzameling van vormen die model kunnen staan voor de nieuwvorming. Wanneer er beperkingen zijn die bepaalde items uitsluiten en andere gevallen juist selecteren, vervalt immers het tegenargument dat alles als modelwoord voor de nieuwvorming kan dienen. Wanneer deze beperkingen daarnaast geformaliseerd zijn, ontstaat de mogelijkheid tot rigoreuze, kwantitatieve toetsing van het idee dat analogie op basis van eerdere taalervaringen het onderliggende principe van productiviteit kan zijn.

Een model dat deze beperkingen in een geformaliseerde vorm biedt, is het *Memory-Based Learning model* (MBL) van Daelemans & Van den Bosch (2005). De rationale van dit model is als volgt. De vorming van een woord is een categorisatieprobleem waarbij de verschillende alternerende vormingspatronen de categorielabels zijn waaruit de taalgebruiker moet kiezen. De categorisatie van een item in een van de categorieën gebeurt op basis van de eerder verwerkte taalervaringen van de taalgebruiker die het meest lijken op het te categoriseren geval. Wanneer op basis van een grondwoord al een keer een afleiding is gemaakt, is de categorisatie eenvoudig, omdat de afleiding als geheel uit het geheugen kan worden opgehaald. Bij daadwerkelijke nieuwvormingen kan dit evident niet. Hier kiest de taalgebruiker voor het categorielabel van de woorden die het meest op het

nieuwe woord lijken. Naar analogie van die woorden wordt het vormingsprocedé bepaald.

De formele procedure die hieraan ten grondslag ligt, is het k-*Nearest Neighbors* algoritme (k-NN). Dit algoritme selecteert binnen een verzameling datapunten de deelverzameling van datapunten die het meest lijkt op het te categoriseren item. Door alleen die meest gelijkende deelverzameling te gebruiken (of de k meest gelijkende deelverzamelingen) legt het model een strenge beperking op aan de mogelijkheid van items om als modelvorm te dienen voor het te categoriseren item.

Omdat MBL uitgaat van losse items en niet van abstracties die over items gemaakt worden, is het een instantiatie van het idee dat (taal)kennis gebaseerd is op unieke verwerkte (taal)ervaringen en niet op – al dan niet uit de ervaring onttrokken – abstracte regels (Posner & Keele 1968, Pierrehumbert 2001, Bybee 2006). Zulke unieke ervaringen worden in de overwegend Engelstalige literatuur exemplars genoemd. In het vervolg zullen we deze term hanteren om een verwerkte taalervaring aan te duiden die een datapunt in de trainingset vormt.

Verder is MBL een model dat niet zozeer stelt wat de inhoud van taalkennis zou moeten zijn, maar dat slechts een cognitieve procedure van analogisch redeneren voorstelt. Deze procedure kan op elk categorisatieprobleem met elke verzameling determinanten worden toegepast. Het model is dus eerder procedureel en exemplar-gebaseerd dan declaratief en abstract.

Om de mate van overeenkomst van een exemplar en een nieuw item te bepalen, kunnen we deze operationaliseren als een afstandsmaat die de conceptuele afstand weergeeft tussen de twee. Het idee hierachter is dat hoe kleiner de afstand is, hoe meer de items op elkaar lijken. In deze studie gebruiken we de Manhattanmetriek als afstandsmaat. De Manhattan-afstand tussen een nieuw item en een exemplar berekenen we door de afstanden tussen de waarden op elk kenmerk van de twee items te berekenen en die bij elkaar op te tellen. Formeler gesteld: de afstand Δ tussen twee items X en Y, elk met n kenmerken, is de som van de afstanden δ tussen de waardes van die gevallen, x en y, op elk van de kenmerken (zie vergelijking 10).

$$\Delta(X, Y) = \sum_{i=1}^{n} \delta(x_i, y_i) \tag{10}$$

Een kunstmatig voorbeeld uit een niet-talig domein helpt een en ander te verduidelijken. Stel dat een artificiële leerder zes verwerkte ervaringen met appels heeft. De leerder heeft die appels f als lekker f niet lekker geëvalueerd (evaluatie $= +, -$), en heeft verder een aantal eigenschappen van de appels in de herinnering verwerkt. Tabel 5 geeft een overzicht.

Als de leerder nu op een nieuwe appel stuit en wil bepalen of deze de moeite

appel	evaluatie	kleur	hardheid	diameter
1	+	rood	hard	8 cm
2	+	rood	hard	11 cm
3	+	geel	hard	7 cm
4	−	rood	zacht	9 cm
5	−	geel	zacht	11 cm
6	−	geel	hard	12 cm

Tabel 5: Een kunstmatig voorbeeld van een verzameling exemplars.

waard is om te eten, kan zij deze appel met alle exemplars vergelijken en de meest overeenkomstige appel nemen als 'modelappel' voor de voorspelling van de smaak van het nieuwe geval.

Omdat de herinnering aan de appel verschillende types informatie bevat (categoriale en numerieke), moeten de afstanden tussen waarden genormaliseerd worden om ze vergelijkbaar te maken. Voor categoriale variabelen, zoals 'kleur', kunnen we stellen dat de afstand tussen de waarden van twee items op een kenmerk 0 is als ze identiek zijn, en 1 als ze niet overeenkomen. Voor numerieke variabelen, zoals 'diameter', geldt dat de afstand tussen de waarden van twee variabelen het absolute numerieke verschil is, gedeeld door het totale bereik van de variabelen. Formeler: de afstand δ voor een variabele i tussen de waarde x_i van item X en de waarde y_i van item Y is het absolute verschil tussen x_i en y_i, gedeeld door het bereik van de waarden van i, oftewel het verschil tussen de hoogste en laagste waarde van i in de dataset.

$$\delta(x_i, y_i) = \begin{cases} \frac{(x_i - y_i)}{max_i, min_i} & \text{indien numeriek, anders} \\ 0 & \text{als } x_i = y_i \\ 1 & \text{als } x_i \neq y_i \end{cases} \tag{11}$$

We kunnen nu de afstand van een te categoriseren appel tot de opgeslagen herinneringen van appels berekenen. Stel dat onze nieuwe appel geel en zacht is en een diameter van 8 centimeter heeft. De afstand tot exemplar 3 uit tabel 5 is dan bijvoorbeeld $\Delta = 1.2$. Op 'kleur' verschillen de twee items niet, op 'hardheid' verschillen ze van waarde en is de afstand $\delta = 1$, en op 'diameter', een kenmerk met een bereik van 5 centimeter, verschillen ze 1 centimeter, en dus is voor dit kenmerk de afstand $\delta = 0.2$. Als we op dezelfde manier voor alle exemplars uit het geheugen de afstand tot de nieuwe appel proberen te berekenen, zien we dat de kleinste afstand tussen de nieuwe appel en een exemplar $\Delta = 0.6$ is, te weten

bij exemplar 5. De set van Nearest Neighbors bestaat dus slechts uit exemplar 5, en de voorspelling is unaniem dat de appel niet lekker is.

In dit voorbeeld zijn we ervan uitgegaan dat alle kenmerken even zwaar meewegen in het bepalen van de afstand tussen de items. Sommige kenmerken sturen echter veel sterker aan op een bepaalde uitkomst dan andere. In tabel 5 zien we bijvoorbeeld dat de waarde 'zacht' op het kenmerk 'hardheid altijd samengaat met de uitkomst 'niet lekker'. Hierom verwachten we dat de waarden van dit kenmerk er meer toe doen in het bepalen van de uitkomst dan de waarden van het kenmerk 'kleur', die gelijkmatiger verdeeld zijn over de twee uitkomsten. Om dit te verdisconteren kennen we elk van de kenmerken een gewicht toe dat weergeeft hoe sterk bepaald waardes correleren met de categorielabels. Dit gewicht kan op verschillende manieren worden berekend. In deze studie gaan we uit van de Gain Ratio (Quinlan 1993), een informatie-theoretische maat gebaseerd op de entropievermindering in de uitkomsten die het te wegen kenmerk bewerkstelligt, genormaliseerd over het aantal waarden van dat kenmerk. Deze maat blijkt goed te werken met kenmerken met veel nominale waarden, en is voor veel taalkundige doelen dus een bruikbare standaard. Verdere motivatie voor het gebruik van de Gain Ratio, de formele beschrijving en een overzicht van andere gewichtsmaten is te vinden in Daelemans & Van den Bosch (2005, 29-32). De globale afstand kan nu worden berekend door elk van de afstanden tussen de waarden van de twee items op de kenmerken te vermenigvuldigen met hun gewicht. Dit resulteert in de verbeterde vergelijking (12), waarin w_i het gewicht van kenmerk i is.

$$\Delta(X, Y) = \sum_{i=1}^{n} w_i \delta(x_i, y_i) \tag{12}$$

Voor het kunstmatige voorbeeld vinden we dat de Gain Ratio voor de drie kenmerken 'kleur', 'hardheid' en 'diameter' respectievelijk 0.082, 0.500 en 0.296 is. De kleinste afstand tussen het nieuwe item en een exemplar is nu niet die tussen het nieuwe item en exemplar 5 (vergelijking 13) , maar die tussen het nieuwe item en exemplar 4 (vergelijking 14). Appel 4 wordt dus gekozen als de modelappel voor de voorspelling van de smaak van de appel.

$$\begin{aligned} &\Delta(\text{exemplar 5, nieuw item}) = \\ &\delta_{\text{kleur}} \cdot w_{\text{kleur}} + \delta_{\text{hardheid}} \cdot w_{\text{hardheid}} + \delta_{\text{diameter}} \cdot w_{\text{diameter}} = \\ &0 \cdot 0.082 + 0 \cdot 0.500 + 0.6 \cdot 0.296 = 0.178 \end{aligned} \tag{13}$$

$$\begin{aligned} &\Delta(\text{exemplar 4, nieuw item}) = \\ &\delta_{\text{kleur}} \cdot w_{\text{kleur}} + \delta_{\text{hardheid}} \cdot w_{\text{hardheid}} + \delta_{\text{diameter}} \cdot w_{\text{diameter}} = \\ &1 \cdot 0.082 + 0 \cdot 0.500 + 0.2 \cdot 0.296 = 0.141 \end{aligned} \tag{14}$$

Naast de wegingsmaat van de kenmerken is er nog een aantal andere parameters in te stellen. Het valt echter buiten het bereik van deze studie om deze mee te nemen, en wij verwijzen de lezer naar Daelemans & Van den Bosch (2005). Één parameter is echter nog van belang om te noemen, te weten het aantal Nearest Neighbor verzamelingen dat we meenemen. Het is mogelijk om alleen die exemplars mee te nemen die de kleinste afstand tot het te categoriseren geval hebben. Soms verbetert de nauwkeurigheid van het model door niet alleen die verzameling te nemen, maar ook de verzameling van exemplars die de op-één-na kleinste afstand tot het nieuwe item hebben. Het aantal verzamelingen k van exemplars met een steeds groter wordende afstand tot het te classificeren exemplar, houden we voor deze studie op $k = 1$[7]

4 Resultaten

4.1 Opzet

Zoals gezegd bestaat het gebruikte corpus uit de componenten *a* tot en met *n* van het Corpus Gesproken Nederlands. De reden dat we component *o* hebben weggelaten, is dat deze voorgelezen geschreven taal bevat, in plaats van (min of meer) spontane gesproken taal. Binnen het gebruikte corpus liet vooronderzoek zien dat er geen significante verschillen zijn tussen de regio's, noch tussen de verschillende componenten, waardoor de homogeniteit van de dataset niet in het geding is. In deze studie hebben we ons verder beperkt tot prenominaal gebruikte adjectieven. Dit levert 2277 tokens op, zoals in paragraaf 2 al is uiteengezet. 156 types adjectieven gingen samen met de morfologische markering *-ste*, terwijl 157 types een perifrastische vorming vertoonden. Voor experiment 1 gaan we uit van de types.[8] De data werden gedeeltelijk handmatig en gedeeltelijk automatisch gecodeerd.

Voor de evaluatie van het model gebruiken we het trainingsregime van leaving-one-out cross-validation (Manning & Schtze 2003: 211). Hierin wordt het model getraind op $N-1$ datapunten en getest op het ene datapunt waarop niet getraind is. In het geval van het k-NN model betekent dit dat we voor één van de datapunten binnen de overige datapunten de verzameling van Nearest Neighbors bepalen en dat het model vervolgens de uitkomst van het item op basis van deze verzameling

[7] Maar zie Keuleers & Sandra (n.d.) voor onderzoek naar de mate waarin andere waardes op deze parameter de nauwkeurigheid van het model beïnvloeden.

[8] We hebben gekozen voor types omdat we zijn geïnteresseerd in hoe goed het model met nieuwe vormen kan omgaan en niet zozeer in hoe goed het model vormen uit het geheugen kan ophalen. Zie Albright (2009, 205-207) voor een argumentatie voor het gebruik van types in plaats van tokens.

voorspelt. Door dit te herhalen voor ieder datapunt ontstaat een nauwkeurige evaluatie van het potentieel van het model om met ongeziene gevallen om te gaan. Het trainen en testen werd integraal gedaan met het hiertoe ontwikkelde programma TiMBL.[9]

In deze studie gebruiken we vier maten om het model te evalueren: de *accuracy*, de *precision*, de *recall* en de F*-score* (Manning & Schütze 1999, 267-271). De algemene accuracy berekenen we door het aantal goed voorspelde gevallen te delen door het aantal datapunten. Omdat deze maat relatief weinig inzicht verschaft, noemen we hem ter volledigheid bij de experimenten, maar richten we onze aandacht meer op de overige drie. De precision, recall en F-score worden per uitkomstcategorie u berekend. De precision is de proportie van de als u gecategoriseerde items die het model correct voorspelt, en wordt berekend door het aantal goed voorspelde gevallen van u (ware positieven) te delen door alle gevallen die het model correct of abusievelijk als u voorspelt. De recall daarentegen geeft de proportie van de eigenlijke gevallen van u die door het model correct voorspeld worden. Dit berekenen we door de ware positieven door de som van de ware positieven en de gevallen die ten onrechte niet als u gecategoriseerd zijn (foute negatieven), te delen. In de vergelijkingen (15) en (16) is de definitie van de precision en de recall te vinden. De F-score, ten slotte, geeft het gewogen harmonische gemiddelde van de precision en de recall. Op deze manier weerspiegelt deze maat zowel de nauwkeurigheid als de mate waarin de precision en recall overeenkomen. Hoe verder precision en recall uit elkaar liggen, hoe lager de F-score immers zal uitvallen, ook al kan de accuracy gelijk blijven. Als we de precision en recall even zwaar wegen, zoals wij in deze studie zullen aanhouden, kan de formule gesimplificeerd worden tot vergelijking (17).

$$\text{precision} = \frac{\text{ware positieven}}{\text{ware positieven} + \text{ foute positieven}} \tag{15}$$

$$\text{recall} = \frac{\text{ware positieven}}{\text{ware positieven} + \text{ foute negatieven}} \tag{16}$$

$$F\text{-score} = \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{17}$$

4.2 Experiment 1

In het eerste experiment hebben we onderzocht hoe goed het model presteert op basis van de in paragraaf 2 genoemde variabelen (het aantal lettergrepen, de coda van de laatste lettergreep en het klinkerpatroon). De volledige matrix (waarvan tabel 6 een gedeelte geeft) is als input opgegeven aan het programma `TiMBL`.

[9]Zie voor een beschrijving en handleiding: `http://ilk.uvt.nl/timbl/`.

adjectief	aantal lettergrepen	coda	klinkerpatroon	superlatief
droog	1	g	$+$	morfologisch
goedkoop	2	p	$++$	morfologisch
ongelegen	4	n	$+-+-$	perifrastisch
recent	2	t	$-+$	perifrastisch
gevaarlijk	3	k	$-+-$	morfologisch

Tabel 6: Gedeelte van de matrix die is opgegeven aan het programma TiMBL. In de kolom 'klinkerpatroon' staat een $+$ voor een volle klinker en een $-$ voor een gereduceerde. De coda's zijn in hun orthografische representatie opgenomen.

superlatief	precision	recall	F-score
morfologisch	0.658	0.684	0.671
perifrastisch	0.680	0.654	0.667

Tabel 7: Resultaten experiment 1. De tabel geeft voor de morfologische en de perifrastische superlatief de precision, recall en F-score.

Hoe goed kan het analogiemodel omgaan met de geobserveerde superlatiefvariatie? Tabel 7 geeft een overzicht van de resultaten:

Het model kan redelijk goed omgaan met de superlatiefalternantie. Zowel de morfologische superlatief als de perifrastische superlatief bereiken een F-score van rond de 0.67. De accuracy van het model ligt rond de 67 procent: 210 van de 314 superlatieven zijn 'correct' geclassificeerd. [10] Deze 67 procent laat zich het beste vergelijken met een basisclassificatie van de data waarin, zonder de onderzochte factoren mee te rekenen, de waarschijnlijkste realisatie voor alle observaties wordt voorspeld oftewel de pure kans op een perifrastische of morfologische superlatief. Tabel 8 geeft de waardes van dit basismodel:

De vergelijking met de basisclassificatie laat een duidelijke verbetering zien. De onderzochte factoren dragen ertoe bij dat het model 17% meer correct classificeert.

Hoewel de drie variabelen een verbeterd model vormen, kan het model nog niet wat het zou moeten kunnen. 67 procent is een redelijke uitkomst, maar 33 procent van de geobserveerde variatie kan niet worden voorspeld. In de volgende paragraaf zullen we daarom een verbeterd model presenteren waarin we de super-

[10] Het gaat hier uiteraard niet over normatieve correctheid, maar over overeenstemming met de gegevens uit het corpus.

superlatief	precision	recall	F-score
morfologisch	0.492	0.492	0.492
perifrastisch	0.508	0.508	0.508

Tabel 8: Basisclassificatie voor het model in experiment 1. De tabel geeft voor de morfologische en de perifrastische superlatief de precision, recall en F-score.

latieven onderzoeken in hun syntactische context.

4.3 Experiment 2

Zoals we al in voetnoot 2.1 vermeldden, leken geen van de aanvankelijk onderzochte variabelen goede voorspellers te zijn van de keuze tussen de twee patronen van superlatiefvorming. Een van de factoren die in een vervolgstudie het onderzoeken waard bleek, was de aantrekkingskracht tussen het adjectief en het nomen waar het adjectief bij hoort.

De motivatie hierbij heeft betrekking op de rol van opslag bij de productie van taal. Als twee woorden vaak samen voorkomen, is het waarschijnlijk dat die opeenvolging geroutiniseerd raakt (Bybee 2006, 715). Hoe sterker die routine wordt, hoe groter de kans dat een taalgebruiker de reeks niet langer als een opeenvolging van twee zelfstandige onderdelen vormt, maar als een patroon dat als geheel uit het geheugen wordt opgehaald.

Wanneer een ontleedbare opeenvolging als geheel geproduceerd en geïnterpreteerd wordt, kunnen we ons ook voorstellen dat de toepassing van elders productieve grammaticale regels niet opgaat voor dat geheel. De ontleding van het patroon als opeenvolging van twee zelfstandige onderdelen wordt immers een steeds zwakkere associatie wanneer het patroon steeds meer een routine wordt. Tussen een adjectief en een nomen (bijvoorbeeld: *leuke dag*) kunnen we nog een adjectief plaatsen (*leuke, zonnige dag*), of het adjectief dat er staat modificeren met een comparatief- of superlatiefmorfeem (*leukere/leukste dag*). Wanneer de combinatie van een adjectief en een nomen (voortaan: *A+N-combinatie*) als geheel wordt opgehaald uit het geheugen, heeft een dergelijke 'inbraak' op het patroon niet de voorkeur of is deze wellicht zelfs onmogelijk. De reden hiervoor is dat het patroon dan eerst opnieuw ontleed zou moeten worden als een adjectief gevolgd door een nomen, alvorens er nieuwe morfemen tussen de twee geplaatst kunnen worden. Voor de keuze van de superlatiefmarkering betekent dit dat bij A+N-combinaties die een sterke aantrekkingskracht tussen het adjectief en het nomen vertonen, perifrastische vorming te verwachten is.

Het begrip 'aantrekkingskracht' is tot nu toe op een informele manier gebruikt. Willen we deze factor echter als variabele in het model gebruiken, dan moet hij geformaliseerd worden. Een maat uit de informatietheorie die hiervoor bruikbaar is, is de Pointwise Mutual Information (PMI, Fano (1961)). Deze maat, weergegeven in vergelijking (18), drukt de logaritmische transformatie van de verhouding uit tussen de waargenomen en de verwachte kans dat twee woorden (w_1 en w_2) opeenvolgend voorkomen. De verwachte kans van opeenvolgend voorkomen van w_1 en w_2 is daarbij gebaseerd op de situatie waarin w_1 en w_2 onafhankelijk van elkaar voorkomen, en is dus het product van de relatieve frequenties van de twee woorden in het corpus. De verhouding tussen de waargenomen en de verwachte kans dat de twee woorden opeenvolgend voorkomen kunnen we interpreteren als de aantrekkingskracht tussen die twee woorden (vgl. Jurafsky & Martin 2009: 696). Bij een $PMI > 0$ spreken we van aantrekkingskracht en als $PMI < 0$, stoten w_1 en w_2 elkaar af. Daarnaast geldt dat hoe hoger de PMI is, hoe sterker de aantrekkingskracht tussen w_1 en w_2.

$$PMI = log_2 \frac{p(w_1, w_2)}{p(w_1) \times p(w_2)} \tag{18}$$

Deze factor veroorzaakt echter niet alleen een verbetering van de nauwkeurigheid van het model, zoals we verderop zullen zien, maar roept ook vragen over exemplarmodellen op. Om een verbetering te bereiken moeten we namelijk op een andere manier de granulariteit van de gebruikte exemplars definiëren.

Tot nu toe hebben we aangenomen dat taalgebruikers de onvervoegde vorm van het adjectief gebruiken om nieuwe superlatieven te vormen. Nieuwe gevallen van hetzelfde adjectief-type gelden dan ook als identiek en worden niet apart opgeslagen (Bybee 2006, 716). In morfologische studies heeft dit idee ertoe geleid dat woordtypes als het relevante niveau van opslag worden gezien (Bybee 1995, Albright 2009). De keuze voor dit niveau is echter niet triviaal, want in syntactische exemplar-modellen is het exemplar juist een analyse van de waargenomen zin (Bod 1998). Wat de eenheid van opslag is, is dus op meerdere niveaus van abstractie en granulariteit te definiëren (zie ook Verbeemen, Vanpaemel & Pattyn (2007)). Dit tweede onderzoek begint bij een hoger niveau van ontleding, waarin we veronderstellen dat taalgebruikers volledige naamwoordsgroepen opslaan bij het verwerken van taal.

Omdat we niet langer de unieke adjectieven als datapunten nemen, verandert de verhouding tussen de frequentie van de perifrastische en de morfologische vormingen drastisch. Waar in experiment 1 de twee uitkomsten evenveel datapunten omvatten, is de verhouding tussen perifrastische en morfologische vormingen in dit experiment ongeveer 1 op 6. Er zijn 204 unieke perifrastische A+N-combinaties te vinden, tegenover 1297 morfologische.

De frequentiegegevens van de adjectieven en nomina zijn verkregen door de

aantrekking A+N	superlatief	
	morfologisch	perifrastisch
LAAG	467	33
MIDDEN	444	56
HOOG	376	125

Tabel 9: Frequentie morfologische en perifrastische superlatieven (A+N-combinaties) in het corpus uitgesplitst naar de aantrekkingskracht tussen het adjectief en het nomen.

superlatief	precision	recall	F-score
morfologisch	0.948	0.971	0.959
perifrastisch	0.797	0.678	0.732

Tabel 10: Resultaten experiment 2. De tabel geeft voor de morfologische en de perifrastische superlatief de precision, recall en F-score.

frequentie van de lemma's te verzamelen in de gebruikte componenten van het *Corpus Gesproken Nederlands*. Op basis van deze gegevens kon automatisch de PMI berekend worden voor elke A+N-combinatie. Vervolgens zijn de data in drie even grote klasses van aantrekkingskracht geplaatst, zoals in tabel 9 te zien is. Het bereik van de waardes was voor de drie klasses als volgt: $LAAG = -2.73 \leq PMI \leq 4.91$, $MIDDEN = 4.91 \leq PMI \leq 7.70$, $HOOG = 7.70 \leq PMI \leq 21.94$. Een eerste blik leert hier dat A+N-combinaties met een hoge aantrekkingskracht inderdaad relatief meer perifrastische vormingen hebben, terwijl combinaties met een lage PMI vrijwel geen perifrastische vormingen vertonen. Wel dient hierbij opgemerkt te worden dat in geen van de klasses de perifrastische vormingen frequenter zijn dan de morfologische, wat bij de eerdere drie factoren wel het geval was. Het effect van de aantrekkingskracht zal dan ook niet erg groot zijn.

Het model bereikte een accuracy van 0.929 wanneer het getraind was op de drie eerder besproken variabelen en het kenmerk van aantrekkingskracht met de drie niveaus als scalaire waardes. Dit betekent dat het model 1395 van de 1501 unieke A+N-combinaties goed voorspeld heeft. De F-scores lopen wat meer uiteen: waar de morfologische categorie een F-score van 0.959 heeft, is deze 0.732 voor de perifrastische vormingen. Tabel 10 geeft een overzicht.

Deze gegevens zijn niet erg informatief als we niet weten hoe het model ongeïnformeerd presteert. Op basis van de relatieve frequentie van de uitkom-

superlatief	precision	recall	F-score
morfologisch	0.857	0.857	0.857
perifrastisch	0.143	0.143	0.143

Tabel 11: Basisclassificatie voor het model in experiment 2. De tabel geeft voor de morfologische en de perifrastische superlatief de precision, recall en F-score.

superlatief	precision	recall	F-score
morfologisch	0.937	0.976	0.959
perifrastisch	0.806	0.603	0.690

Tabel 12: Resultaten experiment 2 met de aantrekkingskracht buiten beschouwing gelaten. De tabel geeft voor de morfologische en de perifrastische superlatief de precision, recall en F-score.

sten behaalt het model een accuracy van 0.756, met een F-score van 0.856 voor de morfologische vormingen en 0.143 voor de perifrastische vormingen (zie tabel 12. Daarnaast is het van belang te weten hoe het model presteert zonder de nieuwe variabele. Tabel 11 geeft de resultaten weer van het A+N-model met slechts de drie eerder bestudeerde factoren. De accuracy in dat geval was 0.923.

De accuracy van 0.929 in het A+N-model met alle vier variabelen lijkt hoog, maar dat valt mee als we het vergelijken met het basisclassificatiemodel, dat een accuracy van 0.756 heeft. Wanneer de categorielabels zo sterk in frequentie verschillen, wordt het behalen van een hoge accuracy vergemakkelijkt. Het correct voorspellen van de minderheidscategorie wordt daarentegen moeilijker. De kans op een morfologische superlatief is immers bij voorbaat groter omdat de relatieve frequentie hoger is. Zonder verdere informatie is de beste 'gok' dan ook een morfologische superlatief. De F-score van het basisclassificatiemodel voor de perifrastische vormingen is daarom ook zo laag: er zijn meer dan zes keer zoveel morfologische datapunten.

Wanneer het model getraind is met de vier variabelen, zien we echter een aanzienlijke stijging in de F-score: van 0.143 naar 0.732. Hieruit kunnen we opmaken dat, ondanks de druk van de meerderheidscategorie, het model een bepaalde niche voor de perifrastische vormingen heeft gevonden op basis waarvan gegeneraliseerd kan worden naar nieuwe items. Hierin schuilt wellicht ook de notie van productiviteit: een minderheidscategorie kan uitbreidbaar zijn, maar dan moet ze een voldoende coherent patroon in de psychologische ruimte van varia-

belen vormen. Zoals Aronoff (1976, 45) stelt: 'productivity goes hand in hand with semantic coherence'. Dat kunnen we als volgt veralgemeniseren: hoe sterker en coherenter een patroon is (in welk domein dan ook), hoe beter het model op basis van dat patroon zal generaliseren naar nieuwe items. Het A+N-model met vier variabelen vormt dus zeker wat betreft de perifrastische vormingen een verbetering over het basisclassificatiemodel, maar ligt dit nu aan het feit dat het exemplarniveau anders gedefinieerd is, of aan de toevoeging van de vierde variabele? De zo goed als identieke accuracy van het A+N-model met drie variabelen suggereert dat het aan de aard van de datapunten ligt en niet aan de factor van aantrekkingskracht. Ook zonder deze factor voorspelt het model nog 92% van de gevallen goed.

Opnieuw is deze maat niet alleszeggend. Als we kijken naar de scores voor de perifrastische superlatief, dan zien we een stijging van 0.043 voor de F-score en zelfs van 0.075 voor de recall, terwijl de precision maar weinig (0.006) zakt. Deze stijging in de recall betekent dat een groter percentage van de perifrastische superlatieven in het corpus ook als zodanig gecategoriseerd is. Hoewel de algemene accuracy dus maar minimaal stijgt, is de beduidende vooruitgang in de voorspelling van de perifrastische vormingen voor ons een reden om te stellen dat deze factor een kleine rol speelt bij de keuze van superlatiefvorming. Ten slotte kunnen we ons afvragen in hoeverre de PMI weergeeft wat we willen dat de maat weergeeft. Omdat we op lemmafrequenties hebben gezocht, hebben ook elementen die al in de collocatie een morfologische superlatiefmarkering hebben, een hoge PMI, bijvoorbeeld *zwak* + *schakel* ($PMI = 12.91$), dat alleen als *zwakste schakel* en niet als *zwakke/zwakkere schakel* voorkomt. Daarnaast is de maat erg gevoelig voor laagfrequente woorden, waardoor een duidelijk collocationeel patroon als *gewoon* + *zaak* uit het idioom *de gewoonste zaak van de wereld*, een lage PMI heeft (0.17) door de hoge frequentie van de twee lemma's.

4.4 Vergelijking experiment 1 en 2

We hebben experiment 2 uitgevoerd om te onderzoeken of de empirische dekking van het model verbeterd zou kunnen worden. De vraag naar verbetering impliceert een vergelijking tussen de twee modellen en dat is in dit geval niet onproblematisch. Doordat we in het tweede experiment een andere definitie van onze datapunten hebben aangenomen (A+N-combinaties in plaats van adjectief-types), is het basisclassificatiemodel veranderd. Hierdoor is vervolgens een vergelijking van de verschillende evaluatiematen niet langer mogelijk. Binnen experiment 2 zien we echter wel dat de toevoeging van de factor van aantrekkingskracht een positief effect heeft op de empirische dekking, voornamelijk op de voorspelling van de perifrastische vorming. Volgens ons is dit voldoende reden om aan te nemen dat aantrekkingskracht een determinant is van de superlatiefalternantie, hoewel na-

der onderzoek naar een methode om de gewenste collocaties te isoleren hiervoor noodzakelijk is.

5 Besluit

Het Nederlands beschikt over twee manieren om de superlatief uit te drukken: morfologisch *-ste* en perifrastisch *meest*. Deze alternantie vormt een categorisatieprobleem voor taalgebruikers: Wanneer gebruik je -ste en wanneer meest? Met name in de generatieve taalkunde gaat men ervan uit dat een analogiemodel geen uitkomst kan bieden voor een dergelijk categorisatieprobleem, omdat het mechanisme van analogie niet restrictief genoeg zou zijn. In tegenstelling tot regels is met analogie elke vorm een mogelijke en daarom heeft het geen voorspellende kracht. In dit artikel hebben wij laten zien dat de superlatiefalternantie wel degelijk te modelleren is met behulp van een analogisch leermechanisme. Op basis van computationele modellering met het *Memory-Based Learning* model van Daelemans & Van den Bosch (2005), hebben we laten zien dat het afwisselend voorkomen van *-ste* en *meest* bij de formatie van een superlatief met een redelijke mate van precisie kan worden voorspeld. Het is uiteraard geen nieuw idee dat analogie een cruciale rol speelt in taalproductie en taalverwerking. De traditionele toepassing van analogie (in de vorm van proportionele analogie) is echter dermate algemeen dat er geen zinvolle voorspellingen mee gedaan kunnen worden. Het model van analogie dat wij in deze studie onder de aandacht hebben willen brengen, doet dit wel en biedt ons de mogelijkheid op een kwantitatieve manier te onderzoeken of analogie het onderliggende principe van productiviteit kan zijn. Dit onderzoek vormt in dat kader slechts een ontkenning van de onmogelijkheid van analogie als het onderliggende cognitieve principe. Een zinvolle bijdrage aan het categorisatiedebat binnen de taalwetenschap zou de verdere ontwikkeling zijn van modellen die verschillende visies op categorisatie onder bepaalde *ceteris paribus* condities vergelijken, zoals het *Varying Abstraction Model* (Verbeemen e.a. 2006).

Referenties

Albright, A. (2009), Modeling analogy as probabilistic grammar, *in* J. Blevins & J. Blevins, eds, 'Analogy in grammar: Form and acquisition', Oxford University Press: Oxford.

Albright, A. & Hayes, B. (2003), 'Rules vs. analogy in english past tenses: A computational/experimental study', *Cognition* **90**(2), 119–161.

Aronoff, M. (1976), *Word Formation in Generative Grammar*, Cambridge: MIT Press.

Baayen, R. H. (1991), 'Quantitative aspects of morphological productivity', *Yearbook of morphology* **149**.

Becker, T. (1990), *Analogie und morphologische Theorie*, München: Wilhelm Fink Verlag.

Bod, R. (1998), *Beyond grammar: an experience-based theory of language*, Center for the Study of Language and Information,[Stanford University].

Booij, G. (2007), *The Grammar of Words. An Introduction to Linguistic Morphology*, Oxford: Oxford University Press. Oxford Textbooks in Linguistics.

Bybee, J. (1995), 'Regular morphology and the lexicon', *Language and Cognitive Processes* **10**, 425–455.

Bybee, J. (2006), 'From usage to grammar: The minds response to repeti- tion', *Language* **82**, 529–551.

Chapman, D. & Skousen, R. (2005), 'Analogical modeling and morphological change: the case of adjectival negative prefix in english', *English Language and Linguistics* **9**(2), 333–357.

Chomsky, N. (1986), *Knowledge of language: its nature, origin, and use*, Greenwood Publishing Group.

Daelemans, W. & Van den Bosch, A. (2005), *Memory-based language processing*, Studies in Natural Language Processing, Cambridge University Press, Cambridge, UK.

Fano, R. M. (1961), *Transmission of Information: A Statistical Theory of Communications.*, New York, MIT Press.

Hüning, M. (1998), Woordensmederij. De geschiedenis van het suffix -erij, PhD thesis, Leiden University.

Keuleers, E. & Sandra, D. (n.d.), 'Similarity and productivity in the english past tense', Unpublished manuscript. Submitted for publication.

Keuleers, E., Sandra, D., Daelemans, W., Gillis, S., Durieux, G. & Martens, E. (2007), 'Dutch plural inflection: the exception that proves the analogy', *Cognitive Psychology* **54**, 283–315.

Krott, A., Baayen, R. H. & Schreuder, R. (2001), 'Analogy in morphology: modeling the choice of linking morphemes in dutch', *Linguistics* **39**(1), 51–93.

Manning, C. D. & Schütze, H. (1999), *Foundations of statistical natural language processing*, MIT Press.

Paul, H. (1920), *Prinzipien der Sprachgeschichte*, Vol. 9, Niemeyer, 1975 (Studienausgabe).

Pierrehumbert, J. (2001), Exemplar dynamics: Word frequency, lenition, and contrast, *in* J. Bybee & P. Hopper, eds, 'Frequency effects and the emergence of lexical structure', John Benjamins: Amsterdam, pp. 137–157.

Pinker, S. (1999), *Words and Rules. The ingredients of language*, London: Phoenix.

Posner, M. & Keele, S. (1968), 'On the genesis of abstract ideas', *Journal of Experimental Psychology* **77**(3/1), 353–363.

Quinlan, J. R. (1993), *C4.5: programs for machine learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Skousen, R. (1989), *Analogical modeling of language*, Kluwer, Dordrecht.

Verbeemen, T., Vanpaemel, W. & Pattyn, S. (2007), 'Beyond exemplars and prototypes as memory representations of natural concepts: A clustering approach', *Journal of Memory and Language* **56**, 537–554.