Schaalvergroting in het syntactische alternantie-onderzoek
Een nieuwe analyse van het presentatieve er *met automatisch gegenereerde predictoren*

Dirk Speelman, Stefan Grondelaers, Benedikt Szmrecsanyi & Kris Heylen

Abstract

In this paper, we revisit earlier analyses of the distribution of *er* 'there' in adjunct-initial sentences to demonstrate the merits of computational upscaling in syntactic variation research. Contrary to previous studies, in which major semantic and pragmatic predictors (viz. adjunct type, adjunct concreteness, and verb specificity) had to be coded manually, the present study operationalizes these predictors on the basis of distributional analysis: instead of hand-coding for specific semantic classes, we determine the semantic class of the adjunct, verb, and subject automatically by clustering the lexemes in those slots on the basis of their 'semantic passport' (as established on the basis of their distributional behaviour in a reference corpus). These clusters are subsequently interpreted as proxies for semantic classes. In addition, the pragmatic factor 'subject predictability' is operationalized automatically on the basis of collocational attraction measures, as well as distributional similarity between the other slots and the subject. We demonstrate that the distribution of *er* can be modelled equally successfully with the automated approach as in manual annotation-based studies. Crucially, the new method replicates our earlier findings that the Netherlandic data are easier to model than the Belgian data, and that lexical collocations play a bigger role in the Netherlandic than in the Belgian data. On a methodological level, the proposed automatization opens up a window of opportunities. Most important is its scalability: it allows for a larger gamut of alternations that can be investigated in one study, and for much larger datasets to represent each alternation.

Keywords: syntactic variation, Belgian vs. Netherlandic Dutch, existential sentences, computational methods, distributional analysis

1. Inleiding

De taalkunde kent een lange en rijke traditie in het bestuderen van syntactische alternantiepatronen, vooral, maar zeker niet uitsluitend, in het kader van

taalgebruiksgebaseerd ('usage-based') onderzoek. Als we inzoomen op kwantitatieve corpuslinguïstische benaderingen, zien we dat regressie als analysetechniek een dominante positie heeft ingenomen. De techniek wordt gebruikt binnen uiteenlopende onderzoekstradities en voor uiteenlopende doelen. Soms staat de studie van puur formele variatie, die niet met betekenisverschillen correleert, centraal; dat is vooral het geval in sociolinguïstische benaderingen die fonetische, lexicale, maar ook morfosyntactische variabelen gebruiken om de linguïstische afstand tussen demografische groepen en gesprekscontexten te meten. In andere gevallen wordt de techniek net ingezet bij het in kaart brengen van de semantische/functionele verschillen tussen de syntactische varianten, die de keuze voor de een of andere variant meebepalen. Dit is onder meer binnen de constructiegrammatica gebruikelijk. Regressie is een krachtige techniek die ons toelaat om voor een specifiek alternatiepatroon – zoals de keuze tussen constructies als *De koningin gaf de appel aan Sneeuwwitje* en *De koningin gaf Sneeuwwitje de Appel* – het samenspel in kaart te brengen van predictoren van uiteenlopende aard, gaande van taalinterne factoren zoals accent- en intonatiepatronen, lexicon, syntaxis, semantiek en pragmatiek, over patronen die aan cognitieve verwerking gerelateerd zijn, bv. verwerkingsgemak of belasting van het geheugen, tot taalexterne factoren zoals genre, modus, (formaliteit van) situatie, sprekers-/schrijverskenmerken, enz. (zie onder meer Gries 2001; Grondelaers, Speelman & Geeraerts 2002; Bresnan et al. 2007). [1]

Regressieanalyse is zeker niet de enige methode voor de studie van alternantiepatronen. Met name het laatste decennium duiken er verschillende krachtige en beloftevolle nieuwe analysemethodes op in het veld van de syntactische alternanties, zoals conditional inference trees en random forests (Tagliamonte & Baayen, 2012), naive discriminative learning (Baayen, 2011), en exemplar-based learning (Vandenbosch 2012; De Troij et al. geaccepteerd). Afgaand op het aantal studies waarin de technieken worden gebruikt, is het niettemin veilig om te stellen dat regressieanalyse nog altijd stevig op de troon zit.

In dit artikel richten we onze aandacht op de remediëring van enkele inherente beperkingen van regressieanalyse, en dan met name op de beperkingen die volgen uit het feit dat de codering van predictoren, in het bijzonder semantische (Levin 1993; Gries 2005) en pragmatische predictoren (Arnold et al. 2005), tot nu toe gewoonlijk manueel geschiedt. Dat zorgt ervoor dat de methode niet onbeperkt schaalbaar is: enkele duizenden attestaties van

[1] Regressie bestaat in verschillende vormen. Bij syntactisch alternantie-onderzoek denken we in eerste instantie

een alternantiepatroon manueel coderen voor een aantal predictoren is de bovengrens van wat in één studie haalbaar is. Bovendien brengt manuele codering van met name semantische en pragmatische predictoren steeds een zeker gevaar van subjectiviteit met zich mee, niet alleen voor wat het bepalen van de mogelijke categorieën (d.i. de *levels* van de variabele) betreft, maar ook voor wat betreft het toekennen van die categorieën aan de individuele attestaties. De bekende strategieën om met dat laatste om te springen (lees: werken met verschillende codeerders en nagaan in hoeverre hun coderingen overeenstemmen), maken het eerst genoemde probleem, het tijdrovende karakter van de codering, alleen maar erger.

Om deze beperkingen te ondervangen, verkennen we de mogelijkheid om semantische en pragmatische predictoren *automatisch* te coderen, met behulp van computationele technieken zoals associatiematen voor co-occurrentiepatronen en distributionele modellen. We treden daarbij in de voetsporen van Levshina & Heylen (2014), die de alternantie tussen de causatieven *doen* en *laten* (zie (1) en (2)) op basis van drie factoren modelleerden, namelijk de semantische categorie van het lexicale hoofd in elk van de drie slot fillers, namelijk de *causer* (hier 'vrouw'), de *causee* (hier 'kinderen') en het *effected predicate* (hier 'tennissen'). Vernieuwend daarbij was dat ze de relevante semantische categorieën voor de slot fillers niet manueel toekenden, en evenmin a priori bepaalden wat de mogelijke categorieën zouden zijn. Elk van die categorieën werd bottom-up opgebouwd door de lexemen in elk slot te clusteren op basis van hun distributionele similariteit, i.e. de mate waarin hun co-occurrentiepatronen in een groot referentiecorpus op elkaar gelijken. De onderliggende gedachte daarbij is dat de semantische equivalentie tussen woorden onder meer blijkt uit hoe vergelijkbaar hun 'buurwoorden' zijn: naarmate woorden qua betekenis equivalenter zijn, vertonen ze ook meer overlappende co-occurrentiepatronen in de vorm van dezelfde woorden die zich in hun omgeving ophouden (vergelijk het beroemde adagium van J. R. Firth, de grondlegger van de distributionele methode: "you shall know a word by the company it keeps" (1957: 11)). Om die reden kan distributionele similariteit, als die berekend wordt op basis van een groot referentiecorpus, gebruikt worden als een benaderende maat voor semantische gerelateerdheid.

(1) De vrouw *deed* haar kinderen tennissen tegen hun zin.
(2) De vrouw *liet* haar kinderen tennissen tegen hun zin.

---

aan logistische regressie, en dan in het bijzonder aan *mixed-effects logistic regression*.

In de huidige studie passen we diezelfde methode toe op de alternantie tussen de aan- en afwezigheid van *er* in bepalingsinitiële zinnen zoals (3) en (4). In ons geval zijn de slots waarvoor we (telkens inzoomend op het belangrijkste inhoudswoord) automatisch semantische klassen genereren de *bepaling* (hier 'asbak'), het *werkwoord* (hier 'liggen') en het *onderwerp* (hier 'sigarettenpeuk').

(3) In de asbak ligt een sigarettenpeuk.
(4) In de asbak ligt *er* een sigarettenpeuk.

We breiden de methode hier echter uit omdat we genoodzaakt zijn niet alleen semantische predictoren, maar ook een *pragmatische* predictor automatisch te modelleren. Omdat uit alle voorafgaande corpusgebaseerde studies (cf. infra) blijkt dat de aan- of afwezigheid van *er* in de eerste plaats bepaald wordt door de voorspelbaarheid van het onderwerp ('sigarettenpeuk' in (3)-(4)), modelleren we die factor met *bottom-up* uit een referentiecorpus gedistilleerde maten die kwantificeren hoe voorspelbaar het onderwerp van de zin is op basis van de voorafgaande bepaling en het voorafgaande werkwoord. De betreffende maten zijn zowel gebaseerd op co-occurrentiepatronen (vanuit het idee dat het frequente samen optreden van een specifieke onderwerpsreferent met een specifieke bepalingsreferent of een specifiek werkwoord dat onderwerp meer 'verwacht' of 'voorspelbaar' kan maken) als distributionele similariteit (vanuit het idee dat ook de semantische relatie tussen die entiteiten en het onderwerp het onderwerp meer 'verwacht' of 'voorspelbaar' kan maken).

Hoewel de in Levshina & Heylen (2014) en in dit artikel beschreven geautomatiseerde manier om syntactische alternanties te bestuderen geenszins bedoeld is om studies op basis van manuele codering te vervangen (daarvoor is de benadering beslist te grofkorrelig), creëert deze nieuwe benadering o.i. nieuwe mogelijkheden. Ten eerste laat de techniek toe om studies van individuele alternantiepatronen op te schalen van enkele duizenden naar (minstens) enkele tienduizenden attestaties. Ten tweede elimineert de methode in hoge mate het gevaar voor subjectiviteit bij de codering. Ten derde maakt de techniek een tweede vorm van schaalvergroting mogelijk: het wordt mogelijk om (al dan niet in één enkele studie) een hele reeks alternantiepatronen samen onder de loep te nemen. Zo kan in kaart worden gebracht in welke mate het relatieve belang van verschillende types predictoren (semantisch,

pragmatisch, lexicaal, …) in regressiemodellen al dan niet gelijk loopt over verschillende alternantiepatronen, en over verschillende regionale of anders gedefinieerde variëteiten.

Omdat de *er*-alternantie reeds uitvoerig werd bestudeerd in eerdere studies (Grondelaers et al. 2002, 2009; Grondelaers, Speelman & Geeraerts 2002, 2008; Grondelaers & Speelman 2007) die als belangrijkste aandrijvers manueel gecodeerde semantisch/pragmatische categorieën identificeerden, kunnen we die oudere studies als negenproef gebruiken voor de validiteit van de automatische methode. Daarbij stellen we ons enerzijds de vraag of de nieuwe methode überhaupt in staat is om een kwaliteitsvol model van de *er*-variatie te bouwen. Anderzijds zoomen we in op enkele belangrijke specifieke vaststellingen uit de oude studies, en stellen we de vraag in welke mate deze vaststellingen in de nieuwe studie al dan niet bevestigd worden.

De structuur van dit artikel is als volgt. In Sectie 2 gaan we nader in op de pionierstudie van Levshina & Heylen (2014). In Sectie 3 bespreken we kort de resultaten van eerdere studies over de *er*-alternantie en motiveren we op basis daarvan specifiekere onderzoeksvragen voor het huidige onderzoek. In Sectie 4 beschrijven we de dataverzameling en de analysemethode van onze casestudy. In Sectie 5 vatten we de resultaten samen, en in Sectie 6 volgt dan een korte nabespreking van de resultaten. Sectie 7 overloopt de conclusies.

2. Automatisch gegenereerde semantische predictor in Levshina & Heylen (2014)

In Levshina & Heylen (2014) werd voor het eerst gebruik gemaakt van geautomatiseerde distributionele semantiek om de alternantie tussen *doen*- en *laten*-causatieven te modelleren (maar zie ook Lapata 1999; Theijssen et al., 2010 voor eerdere pogingen om predictoren in regressies (deels) te automatiseren). Levshina & Heylen (2014) bepalen de voorkeur voor specifieke causatiefconstructies zoals gezegd op basis van de semantische klasse van de belangrijkste inhoudswoorden in de *causer*-, de *causee*- en de *effected predicate*-slots. In (1) en (2) zijn dat respectievelijk *vrouw*, *kinderen* (of liever het basislemma *kind*), en *tennissen*. Met behulp van distributionele semantiek wordt voor elk van die woorden een vector gebouwd, een ‘gebruiksprofiel’ dat belichaamt welke andere woorden in hun buurt voorkomen, en hoe vaak dit gebeurt. Een vector is in essentie een reeks cijfers die de frequentie reflecteren van een groot aantal referentiewoorden die in de buurt van het onderzochte woord voorkomen; in dat opzicht vormen vectoren, een beetje kort door de

bocht, het in cijfers uitgedrukte ‘semantische paspoort’ van een woord. Vervolgens worden de aangetroffen lexemen in de drie slots op basis van hun respectievelijke vectoren in semantische klassen geclusterd: het spreekt daarbij vanzelf dat woorden waarvan de distributionele similariteit groter is omdat hun gebruikscontexten meer op elkaar gelijken in dezelfde klasse terechtkomen (Lin 1998; Turney & Pantel 2010).

Levshina & Heylen (2014) vergelijken een hele reeks manieren om vectoren te berekenen. Bovendien variëren ze het aantal clusters dat gebouwd wordt voor elk van de drie slots, gaande van een beperkt aantal ruime categorieën tot een groter aantal fijnmazigere. Verschillende combinatiemogelijkheden worden uitgeprobeerd in de zoektocht naar de combinatie die er het beste in slaagt om de *doen/laten*-alternantie te modelleren. Het model dat in hun studie als ‘beste’ wordt bestempeld bevat acht semantische categorieën voor de *causer*, acht semantische categorieën voor de *causee*, en maar liefst 23 voor het *effected predicate*. Een eerste conclusie is dus dat het werkwoord-*slot* een veel fijnmazigere classificatie vereist dan de twee nominale *slots*. Een tweede vaststelling is dat de informatie in de verschillende slots, zoals ze dat zelf noemen, ‘niet-additief’ is. Uit de vergelijking van modellen met verschillende predictoren blijkt weliswaar dat elk van de *slots* op zich informatie bevat die de modellen beter maakt (waarbij *effected predicate* informatiever is dan *causer*, en *causer* informatiever dan *causee*), maar het blijkt ook zo te zijn dat modellen doorgaans beter worden als informatie uit twee of drie s*lots* samen in het model zit. De meerwaarde die resulteert uit het combineren van meerdere *slots* in één model is evenwel niet-additief, in de zin dat de mate waarin het model stelselmatig beter wordt bij het toevoegen van extra predictoren minder uitgesproken is dan men zou kunnen verwachten: het gecombineerde effect van de verschillende *slots* op de causatiefalternantie is kleiner dan de som van de effecten van de individuele *slots* op de causatiefalternantie. De auteurs verklaren dit patroon tentatief als het gevolg van (semantische) coherentie tussen de verschillende *slots*, die tot gevolg heeft dat de informatie in de verschillende slots niet helemaal onafhankelijk is (Stefanowitsch & Gries 2005: 11).

Het beste model uit Levshina & Heylen (2014) vertoont een uitstekende *goodness of fit* (C > 0,9), waaruit de onderzoekers concluderen dat de semantiek van de verschillende *slots* een goede voorspeller is voor het alternantiepatroon, en dat distributionele semantiek dus een uitstekend instrument is om de relevante semantische dimensies bloot te leggen.

3. Eerdere studies over de *er*-alternantie in bepalingsinitiële zinnen

In dit artikel willen we de methode uit Levshina & Heylen (2014) toepassen op de alternantie tussen de aan- en afwezigheid van het existentiële of presentatieve *er* in bepalingsinitiële zinnen zoals (3) en (4), die hier voor het leesgemak nog eens worden herhaald:

(3) In de asbak ligt een sigarettenpeuk.
(4) In de asbak ligt *er* een sigarettenpeuk.

Ook hier kunnen we verschillende constructie-*slots* onderscheiden. Ten eerste is er de *bepaling*: in voorbeelden (3) en (4) is dat telkens 'in de asbak', met als belangrijkste inhoudswoord 'asbak'. Ten tweede is er het *werkwoord*; in de voorbeelden is dat telkens 'ligt', met als lemma 'liggen'. Ten derde is er het onderwerp 'een sigarettenpeuk', met als belangrijkste woord 'sigarettenpeuk'.[2] We zullen onderzoeken in welke mate de 'semantische klasse' van de woorden in deze *slots*, berekend zoals in Levshina & Heylen (2014), een goede predictor is voor de alternantie.

Er valt echter te verwachten dat semantische predictoren, anders dan bij de *doen/laten*-alternantie, niet volstaan om tot een goede modellering van de alternantie te komen. Eerder onderzoek heeft immers aangetoond dat pragmatische predictoren minstens even belangrijk zijn voor het modelleren van deze *er*-alternantie. Op basis van eerder corpuslinguïstisch onderzoek met manueel gecodeerde datasets (Grondelaers, Speelman & Geeraerts 2002, 2008), maar ook op basis van experimenteel psycholinguïstisch onderzoek (Grondelaers et al. 2002, 2009), weten we dat met name de voorspelbaarheid van het onderwerp de *belangrijkste* predictor is voor de aan- of afwezigheid van *er*. Daarom zullen we in de huidige studie ook werken met automatisch gegenereerde corpusgebaseerde maten voor deze voorspelbaarheid (zie ook Jaeger 2005; Levy en Jaeger 2007). De pragmatische aandrijver 'onderwerpsvoorspelbaarheid' is natuurlijk niet volledig onafhankelijk van de semantische predictoren: een zeer concrete, sterk afgebakende plaatsbepaling als 'in zijn broodtrommeltje'

[2] We beperken ons in deze studie tot tokens van de *er*-alternantie waarvan de drie *slots* lexicaal gevuld zijn; alle andere instanties werden door ons retrieval-script geweerd. Meer in het bijzonder wordt gezocht naar het nomen dat bepaling en onderwerp realiseert, en het werkwoord dat in deze constructie altijd net voor *er* of net voor het onderwerp verschijnt. De reden hiervoor is dat de distributionele vergelijking van vectoren van woorden van verschillende woordsoorten (bv. de vergelijking van de vector voor een voorzetsel met de vector voor een nomen) niet zo gemakkelijk te interpreteren is. In principe is het nochtans mogelijk om binnen één *slot* met woorden van verschillende woordsoorten rekening te houden bij het bouwen van een vector voor dat *slot*. Via

bepaalt de voorspelbaarheid van een onderwerp als ‘een pindakaasboterham’ natuurlijk in veel hogere mate dan de tijdsbepaling in de zin *Zondag kregen we soep bij de lunch maar gisterenmiddag was er een pindakaasboterham* of de abstractere plaatsbepaling in de zin *In Antwerpen stond er soep op het menu maar in Brussel was er een pindakaasboterham*. Het is in dit verband logisch dat de twee laatste zinnen met *er* geconstrueerd moeten worden terwijl *In zijn broodtrommeltje was een pindakaasboterham* veel makkelijker zonder kan. Bepalingstype (locatief vs. temporeel) en bepalingsconcreetheid zijn dan ook variabelen die onveranderlijk bevestigd werden als predictoren van *er*, net zoals ‘werkwoordelijke specificiteit’: als we *was* in de laatste zin door *zat* vervangen is *er* nog minder nodig. Bemerk dat de predictoren bepalingstype en bepalingsconcreetheid een semantisch en pragmatische dimensie bundelen. Hetzelfde geldt voor werkwoordelijke specificiteit. Je kan die drie predictoren immers bekijken als semantische classificaties die zo opgebouwd zijn dat de verschillende semantische klassen samenvallen met een verschillende mate van voorspelbaarheid van het onderwerp. In de huidige studie zullen we trachten, in de mate van het mogelijke, die twee dimensies uit elkaar te halen door de aparte introductie van enerzijds semantische klassen (die niet noodzakelijk allemaal rechtlijnig langs de as van voorspelbaarheid lopen) en anderzijds maten voor de voorspelbaarheid van het onderwerp (die niet noodzakelijk helemaal parallel lopen met semantische klassen).

Een ander resultaat uit de net genoemde corpusgebaseerde studies over de *er*-alternantie is het bestaan van kwantitatieve en kwalitatieve nationale verschillen tussen Nederland en België (zie vooral Grondelaers, Speelman & Geeraerts 2008). Ten eerste is *er* relatief frequenter in België. Verder werd gevonden dat (i) het effect van bepalingstype en bepalingsconcreetheid opvallend groter was in Nederland dan in België, en dat (ii) Nederlandse modellen veel vlotter een hoge *goodness of fit* behaalden: opgenomen predictoren verklaren en voorspellen het gedrag van *er* in Nederlandse bepalingszinnen veel beter dan in Belgische. In Nederland (iii) zijn over het algemeen ook minder predictoren nodig om tot goede modellen te komen: de net genoemde aandrijvers bepalingstype, bepalingsconcreetheid en werkwoordelijke specificiteit volstaan er om een uitstekende fit te bereiken.

---

technieken voor anafoorresolutie is het bijvoorbeeld mogelijk om ook pronominale *slot*-fillers van een zinvolle vector te voorzien. Van deze meer geavanceerde toepassingen wordt in de huidige studie geen gebruik gemaakt.

Aansluitend bij deze regionale verschillen verwijzen we ook naar Vandenbosch (2012) en De Troij et al. (geaccepteerd), waar werd getoond dat het modelleren van de *er*-alternantie met behulp van memory-based learning beduidend succesvoller was voor data uit Nederland dan voor data uit België. Memory-based learning (Daelemans & Van den Bosch 2005) is een leeralgoritme dat nieuwe constructionele keuzes voorspelt op basis van louter lexicale gelijkenis met een trainingset van 'oude' keuzes. Omdat deze modellen uitsluitend gebruik maken van ongeclassificeerde lexicale informatie (met name de vijf woorden links en de vijf woorden rechts van het slot waarin *er* al dan niet voorkomt), zonder verder rekening te houden met andere factoren zoals syntactische status, lijkt het erop dat de *er*-alternantie in Nederland veel meer dan in België (mee)bepaald is door de aanwezigheid in de zin van specifieke lexemen. Vandaar de onderzoeksvraag of lexicale informatie ook in deze studie nuttiger zal blijken voor de modellering van de Nederlandse dan van de Belgische data.

Hieronder lijsten we de onderzoeksvragen nog even op:

1. Is het *er*-gebruik in de data even goed modelleerbaar met de nieuwe, automatisch gegenereerde predictoren als met de oude, manueel gecodeerde aandrijvers? Met andere woorden, is het met het nieuwe model even goed mogelijk om contexten met een hogere kans op *er* te onderscheiden van contexten met een lagere kans op *er*? We zullen voor de vergelijking gebruik maken van de zogenaamde C-index, een maat voor de statistische *goodness of fit*.
2. Hoe gelijklopend zijn de nieuwe modellen met de modellen uit eerder onderzoek? Meer in het bijzonder:[3]
    a. Wordt bevestigd dat informatie over de bepaling voor de Nederlandse data belangrijker is dan voor de Belgische?
    b. Wordt bevestigd dat de Nederlandse data gemakkelijker te modelleren zijn dan de Belgische?

[3] De specifieke deelvragen bij onderzoeksvraag 2 vormen een operationele verenging van de algemene vraag hoe goed de nieuwe en de oude modellen overeenkomen. Die verenging is nodig, omdat de nieuwe modellen bijzonder veel informatie opleveren (ze leggen zowel globale als heel wat meer lokale patronen bloot), waardoor een exhaustieve vergelijking van de nieuwe en de oude modellen buiten het bestek van dit artikel valt. Hoewel het beslist interessant zou zijn om nader in te gaan op de interpretatie van het effect van de individuele predictoren (en alle deelcategorieën binnen die predictoren) die de automatische analyse oplevert, beperken we ons hier dus bewust tot een vergelijking van de automatische en de manuele modellen op basis van de globale 'modelleerbaarheid' van de data, de 'modelleerbaarheid' van de Belgische en de Nederlandse data, en het relatieve belang in beide types modellen van de semantische, pragmatische en lexicale predictoren in zowel de Belgische data als de Nederlandse dataset.

c. Wordt bevestigd dat lexicale informatie voor de Nederlandse data belangrijker is dan voor de Belgische?

4. Casestudy: Dataverzameling en analysemethode

De Nederlandse dataset voor de casestudy, waarnaar we voortaan verwijzen als NL-TwNC, bestaat uit 15000 attestaties van de *er*-alternantie die met een script uit het Twente News Corpus (TwNC; ca. 400 miljoen tokens) werden gehaald (Ordelman et al. 2007). Het script zocht naar zinnen die begonnen met een bepaling, gevolgd door een werkwoord/persoonsvorm, optioneel gevolgd door *er* en vervolgens gevolgd door een onbepaald onderwerp. Vergeleken met de datasets waarop de in Sectie 3 vermelde corpusstudies gebaseerd zijn, is de selectie nu op twee manieren verschillend. Enerzijds is de nieuwe selectie ruimer: nu worden ook andere bepalingstypes dan bepalingen van plaats en tijd toegelaten. Anderzijds is de nieuwe selectie ook enger: enkel bepalingen waarin het belangrijkste woord een nomen en meer bepaald een soortnaam is (geen eigennaam dus), worden toegelaten (dezelfde restrictie geldt trouwens ook voor het onderwerp).

De Belgische dataset voor de casestudy, voortaan BE-LeNC, bestaat eveneens uit 15000 attestaties, op volledig analoge manier geselecteerd uit het naar het TwNC gemodelleerde Leuven News Corpus (LeNC; ca 1,2 miljard tokens).

De datasets NL-TwNC en BE-LeNC zijn willekeurig geselecteerd uit een grotere collectie van treffers (in totaal vond het script 191002 treffers in het TwNC en 404663 treffers in het LeNC). In de rest van deze sectie bespreken we de variabelen waarvoor de items in NL-TwNC en BE-LeNC werden gecodeerd. We beschrijven hierbij niet elk technisch detail van de berekeningen. Voor deze details verwijzen we de geïnteresseerde lezer naar Speelman, Heylen & Grondelaers (te verschijnen).

De responsvariabele `er` codeert de aan- of afwezigheid van *er* binair. Mogelijke waarden zijn `aanwezig` en `afwezig`.

De predictor `semcat_werkw` codeert de semantische klasse van het werkwoord. Vier verschillende types van distributionele modellen werden gebouwd (zie ook Padó & Lapata 2010). *Bag-of-words* vectoren bevatten de frequentie van een aantal referentiewoorden in een

'raam' rond het onderzochte woord; wij hebben *bag-of-words* vectoren met spanwijdte 4:4 (vier woorden links, vier woorden rechts) en 7:7 gebouwd. Vectoren op basis van dependentierelaties inventariseren informatie over woorden die specifieke afhankelijkheidsrelaties met het doelwoord onderhouden, zoals bijvoorbeeld 'is het onderwerp van' of 'is het indirect object van'. Vectoren op basis van subcategorisatie-informatie zijn vooral geschikt voor het peilen naar de semantische essentie van werkwoorden omdat ze als contextinformatie het volledige subcategorisatieschema van een werkwoord bevatten.

De vectoren die elk van de vier benaderingen opleverden werden vervolgens met behulp van *k means clustering* (Everitt, Landau & Leese 2001) gegroepeerd in *k* groepen (deze oefening werd herhaald voor *k* lopend van 5 tot 30: het algoritme werd gevraagd vectoren te clusteren in 5 tot 30 groepen). Ten slotte werd, op basis van hun potentieel om de *er*-alternantie te modelleren, uit al deze benaderingen één benadering geselecteerd: net zoals in Levshina & Heylen (2014) bleek het distributionele model op basis van subcategorisatieschema's succesvoller dan de andere modellen, en werd het daarom gekozen.[4] Voor de clusteroplossing werd geopteerd voor *k*=22, omdat een verdere verhoging van *k* geen substantiële verbetering van de *goodness of fit* van de regressiemodellen opleverde. In wat volgt zullen we m.a.w. werken met 22 semantische klassen van werkwoorden.

De predictor `semcat_ondrw` codeert de 'semantische klasse' van het hoofd van de onderwerps-NP. Hier werd met één type distributioneel model gewerkt (*bag-of-words* met *span* 4:4). Voor het clusteren werd gewerkt met *k* lopend van 5 tot 10. Uiteindelijk werd geopteerd voor *k*=8 (8 semantische klassen), omdat een verdere verhoging van *k* geen substantiële verbetering van de *goodness of fit* van de regressiemodellen opleverde. Voor de predictor `semcat_bepal`, die de semantische klasse van het belangrijkste woord van de bepalings-NP codeert, werd met een identiek distributioneel model gewerkt, zij het dan met *k*=9.

De negen predictoren `voorspelb_1`, `voorspelb_2`, …, `voorspelb_9` coderen verschillende aspecten van de voorspelbaarheid van het onderwerp op basis van de bepaling en/of het werkwoord. Ten eerste hebben we gewerkt met collocatiesterkte (Evert 2009; Gries

[4] Het gebruikte criterium voor de vergelijking van het succes van de modellen is ook hier de C-index.

2013) tussen bepaling of werkwoord enerzijds, en onderwerp anderzijds (en ook tussen bepaling en werkwoord). Ten tweede hebben we gewerkt met (op distributionele gegevens gebaseerde) semantische gelijkenis tussen bepaling of werkwoord enerzijds en onderwerp anderzijds (en ook tussen bepaling en werkwoord). De onderliggende redenering daarbij is dat zowel een hoge mate van co-occurrentie als een hoge mate van semantische gerelateerdheid de voorspelbaarheid van het éne woord op basis van het andere woord doorgaans verhogen. Om collocatiepatronen te meten gebruiken we voor elk mogelijk paar van de drie *slots* drie types maten: de absolute frequentie van de co-occurrentie, een aantal op *effect size* gebaseerde associatiematen (*pointwise mutual information*, *Delta P*, *DICE*, *relative risk*), en een op een significantietoets gebaseerde associatiemaat (*signed log-likelihood ratio test statistic*). Om semantische verwantschap te meten maken we gebruik van verschillende types distributionele modellen. Op deze manier hebben we 62 verschillende maten voor voorspelbaarheid, die geregeld sterk correleerden. Om die 62 maten tot een overzichtelijk aantal voorspelbaarheidspredictoren te herschalen, hebben we ze met behulp van principaalcomponentanalyse herleid tot 9 niet correlerende principaalcomponenten (in NL-TwNC verklaart die analyse 82% van de variantie, in BE-LeNC 86%).

De factoren `werkw`, `onderw` en `bepal` ten slotte bevatten (het lemma van) het belangrijkste inhoudswoord in de werkwoords-, onderwerps- en bepalingsslots, die als *random factors* opgenomen worden in het *mixed-effects model*. De automatische identificatie van dit belangrijkste inhoudswoord (hoofd van de NP in de bepalings-PP; hoofd van de onderwerps-NP; werkwoordsvorm net voor *er* of net voor het onderwerp) kon automatisch gebeuren, omdat het script waarmee de attestaties van de *er*-alternantie uit de corpora werden gehaald, de patronen weerde waarin deze identificatie problematisch is (zoals in het geval van bv. nevenschikking binnen de bepaling of binnen het onderwerp).

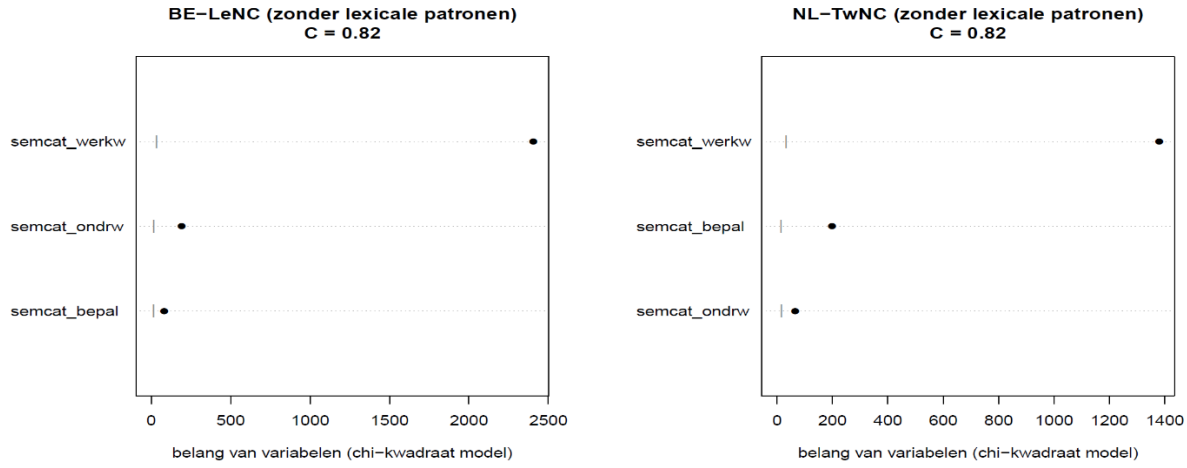
Belangrijk om op te merken, voor we overgaan tot de bespreking van de resultaten, is dat zowel de semantische variabelen als de pragmatische variabelen apart berekend zijn voor NL-TwNC en BE-LeNC. Concreet betekent dit dat noch de semantische klassen noch de negen voorspelbaarheidsdimensies dezelfde betekenis hebben in NL-TwNC als in BE-LeNC. We hebben er bewust voor gekozen om evenveel voorspelbaarheidsdimensies te hebben in beide datasets, en om bovendien telkens evenveel ‘semantische klassen’ (voor elk van de drie *slots*) te gebruiken, omdat de data aangeven dat de optimale keuzes (aantal principaalcomponenten;

aantal clusters) telkens ruwweg overeenstemden voor beide datasets. Die opgelegde equivalentie heeft het voordeel dat de regressie-analyses voor beide datasets dezelfde complexiteit hebben. We moeten bij de interpretatie echter onthouden dat de betreffende dimensies en klassen in de beide datasets slechts overeenstemmen qua aantal, niet (volledig) qua inhoud.

5. Casestudy: Resultaten

Met behulp van de in Sectie 4 beschreven predictoren worden in deze sectie aparte regressiemodellen gebouwd voor BE-LeNC en NL-TwNC. We beschrijven, telkens zowel voor BE-LeNC als voor NL-TwNC, achtereenvolgens vier modellen, waarin we telkens andere predictoren toevoegen aan het model. We beschouwen het vierde model als het finale model waar we naartoe werken, maar we tonen ook de tussenstappen omdat die extra inzicht verschaffen in het relatieve belang van, en de relatieve verhoudingen tussen zowel de drie categorieën van predictoren (semantisch, pragmatisch, lexicaal) als de individuele aandrijvers binnen elke categorie. Wat met name informatief kan zijn, is hoe het relatieve belang van een predictor soms afneemt (of liever 'gecorrigeerd wordt') na het toevoegen aan het model van een andere predictor.

In het eerste model, een *fixed effects only model*, introduceren we de drie semantische predictoren. In Figuur 1 wordt, links voor BE-LeNC en rechts voor NL-TwNC, het relatieve belang van de predictoren in de modellen getoond. Het belang van een predictor, afleesbaar op de x-as, is geoperationaliseerd als de teststatistiek (*model chi-squared*) voor een *log likelihood ratio test* voor het belang van die predictor in het model (waarbij een model met die predictor wordt vergeleken met een model zonder die predictor). Hoe hoger de teststatistiek (positie van het bolletje), hoe hoger het belang van de predictor. Het kleine verticale streepje geeft aan waar de grens voor significantie ligt in de *log likelihood ratio test* (waarbij posities rechts van het streepje wijzen op significantie). Op de y-as zijn de namen van de predictoren leesbaar, gesorteerd volgens dalende belangrijkheid.



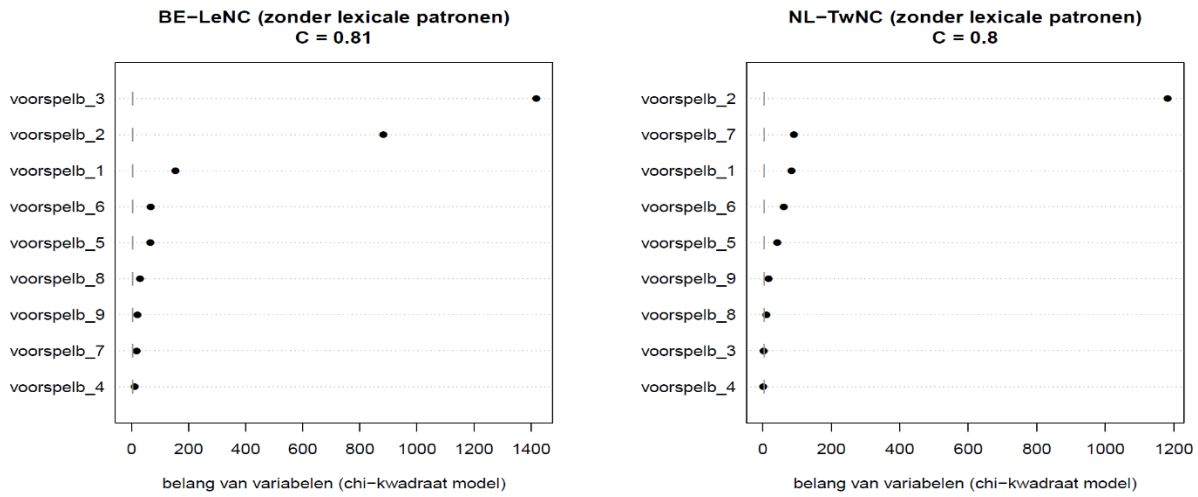
Figuur 1. Belang van de predictoren in model 1 (enkel semantische predictoren)

De *goodness of fit* van de modellen (C-index, opgenomen in de titel van beide grafieken) lijkt te suggereren dat de modellen een behoorlijke kwaliteit en voorspellende waarde hebben, hetgeen aantoont dat de semantische factoren op zichzelf al een belangrijke rol spelen bij de modellering van *er*. We zien verder dat zowel in BE-LeNC als in NL-TwNC de semantische klasse van het werkwoord veruit de belangrijkste predictor is. Vervolgens is er een verschil tussen beide modellen: in NL-TwNC is de semantische klasse van de bepaling relatief belangrijker dan in BE-LeNC. Bemerk dat die observatie onderzoeksvraag (2.a) lijkt te bevestigen, al is het nog te vroeg om verstrekkende conclusies te trekken. Verderop zullen we namelijk vaststellen dat de verklarende kracht van de semantische klassen, vooral in het geval van het werkwoord, in dit model enigszins 'geflatteerd' is omdat specifieke lexicale effecten van individuele verbale lexemen onderliggend werkzaam zijn: één hoogfrequent woord of een klein aantal hoogfrequente woorden kunnen daarbij een hele semantische klasse domineren.[5]

In het tweede model (Figuur 2), ook een *fixed effects only model*, laten we de semantische klassen achterwege, om uitsluitend in te zoomen op de voorspelbaarheidspredictoren. We zien dat de *goodness of fit* (C-index) iets lager ligt dan in model 1, maar nog steeds wijst op een goede voorspellende waarde van de modellen: ook de voorspelbaarheidspredictoren hebben zoals verwacht een duidelijke impact op het gedrag van *er*. Louter statistisch gezien is

[5] Model 1 bevat (net als de verderop volgende modellen) heel wat informatie die in dit artikel niet besproken wordt. Zo is het heel leerzaam om de woorden in de verschillende 'semantische klassen' te inspecteren en het effect van elke individuele ´semantische klasse' te inspecteren. Als exploratieve techniek is een dergelijke detailinspectie van het model zeer nuttig. In dit artikel beperken we ons evenwel bewust tot de 'grote lijnen' die de modellen uittekenen.

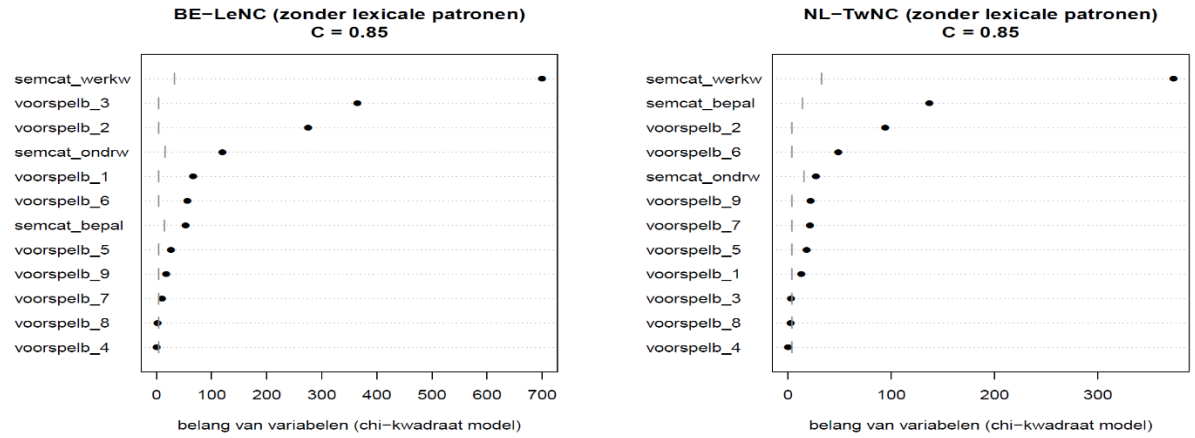
het niet volledig koosjer om het relatieve belang van de semantische versus de pragmatische predictoren vast te stellen op basis van de vergelijking van de C-indices in model 1 versus die in model 2. Ten eerste is het zo dat in beide families van predictoren (zowel de semantische als de pragmatische) de operationaliseringen grofkorrelig zijn. Ten tweede wordt in model 1 systematisch met categoriale predictoren gewerkt en in model 2 met numerische predictoren, wat de vergelijking bemoeilijkt. Het vergelijken binnen één type model van BE-LeNC en NL-TwNC is echter een stuk gefundeerder. Het belangrijkste verschil is dat er in NL-TwNC maar één uitgesproken voorspelbaarheidspredictor is (voorspelb_2 vervat informatie over zowel bepalings-werkwoord- als werkwoord-onderwerp co-occurrentiepatronen, samengebundeld in één predictor), terwijl er in BE-LeNC twee uitgesproken predictoren zijn: aan de Belgische kant reflecteert voorspelb_3 een hoge co-occurrentiefrequentie van werkwoord en onderwerp, en voorspelb_2 een hoge co-occurrentiefrequentie van bepaling en werkwoord. Opvallend is verder dat alle negen voorspelbaarheidsaandrijvers aan Belgische kant significant blijven, terwijl voorspelb_3 en voorspelb_4 aan Nederlandse kant geen significantie halen.



Figuur 2. Belang van de predictoren in model 2 (enkel pragmatische predictoren)

In het derde model (Figuur 3), nog steeds een *fixed effects only model*, combineren we de semantische klassen en de voorspelbaarheidsdimensies. We stellen dan vast dat beide types predictoren een belangrijke rol spelen: ze tillen de *goodness of fit* van de beide modellen naar een hoger niveau (zie de respectieve C-indexen), wat aangeeft dat ze tot op zekere hoogte complementair zijn. Anderzijds zien we tegelijkertijd (vooral in NL-TwNC, maar ook in BE-LeNC) dat sommige voorspelbaarheidsdimensies aan belang inboeten (ook in hun relatieve positie tegenover andere voorspelbaarheidsdimensies) en zelfs geen significantie meer halen

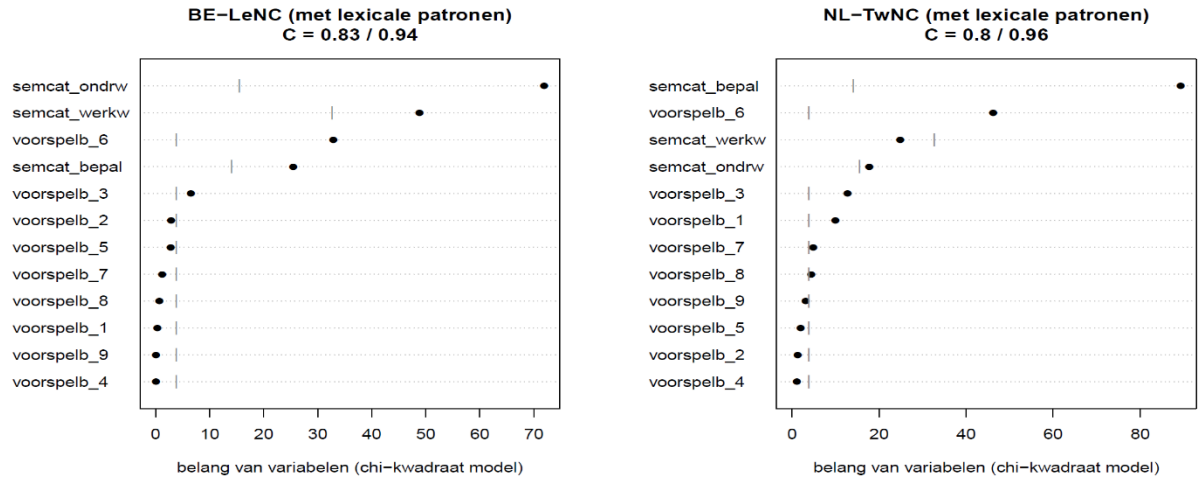
in de *log likelihood ratio test*. Dit toont dat de informatie in sommige van de voorspelbaarheidsmaten en sommige semantische klassen deels overlapt. Dat gebeurt bijvoorbeeld wanneer een bepaalde voorspelbaarheidsdimensie vooral informatie bevat *op basis van het werkwoord*, en bepaalde semantische categorieën van het werkwoord 'brede', taxonomisch onspecifieke werkwoorden bevatten die zich minder lenen tot het voorspellen van het onderwerp; de introductie van dat laatste type categorie in het model leidt dan tot een afname van het belang van de betreffende voorspelbaarheidsmaat. De onderlinge relatieve positie van de drie semantische variabelen blijft, vergeleken met model 1, wel stabiel, zowel in BE-LeNC als in NL-TwNC. Met name het relatief grotere belang van de semantische klasse van de bepaling in de Nederlandse data blijft overeind (cf. onderzoeksvraag 2.a). Anderzijds hebben we nog geen duidelijke indicaties gevonden dat het in de Nederlandse data gemakkelijker is om een hogere *goodness of fit* te bereiken dan in de Vlaamse: in modellen 1, 2 en 3 waren de C-indexen in BE-LeNC en NL-TwNC telkens opvallend vergelijkbaar.



Figuur 3. Belang van de predictoren in model 3 (semantische en pragmatische predictoren)

In Figuur 4 tonen we model 4, een *mixed-effects model* met, naast alle fixed-effect predictoren uit model 3, ook random intercepts voor de factoren `werkw`, `onderw` en `bepal`. Met deze random effects kunnen we de impact van individuele lexemen op de alternantie ondervangen. Als, bijvoorbeeld, het schijnbare effect van een bepaalde semantische werkwoordscategorie niet zozeer te wijten is aan de globale semantische categorie, maar aan één hoogfrequent lexicaal item binnen die categorie (bijvoorbeeld, het werkwoord *zijn*), dan zal het model zonder random intercepts voor `werkw` hier geen rekening mee kunnen houden. Een model dat dergelijke random intercepts wel bevat zal detecteren dat

een werkwoord zich apart gedraagt en zal daarom een betere schatting kunnen maken van het globale effect van de semantische categorie.



Figuur 4. Belang van de predictoren in model 4 (semantische, pragmatische en lexicale predictoren)

We tonen in deze modellen een dubbele C-index. Het eerste getal (0,83 voor BE-LeNC; 0,8 voor NL-TwNC) geeft telkens de C-index waarbij het model predicties maakt voor attestaties op basis van de fixed-effect predictoren, maar niet op basis van de random factoren. Het tweede getal (0,94 voor BE-LeNC; 0,96 voor NL-TwNC) geeft telkens de C-index waarbij het model predicties maakt voor zowel de fixed-effect predictoren als de waarden voor de random factors.

We willen er nogmaals op wijzen dat de vergelijking van het relatieve belang van de drie hoofdcategorieën van predictoren (semantisch, pragmatisch, lexicaal) niet zo evident is, omdat ze statistisch gezien met erg verschillende grootheden corresponderen (categoriale *fixed effects*, numerische *fixed effects*, *random effects*). Toch kunnen we heel wat leren uit de confrontatie van de drie types factoren.

Ten eerste valt op dat de introductie van lexicale factoren door toevoeging van random effects het belang van de semantische klasse van het werkwoord opvallend reduceert. De voor de hand liggende interpretatie is dat het grote belang van die semantische werkwoordsklassen in de vorige modellen voor een belangrijk deel gebaseerd was op onderliggende lexicale effecten. Omdat er 23 werkwoordcategorieën zijn, en het aantal items per categorie zo beduidend kleiner is dan bij de nomencategorieën, bestaat de kans dat

bepaalde categorieën gedomineerd worden door slechts één of enkele specifieke werkwoorden met bovengemiddelde frequentie.

Ten tweede valt op dat verschillende predictoren (vooral voorspelbaarheidspredictoren, maar in NL-TwNC ook de semantische klasse van het werkwoord) geen significantie meer behalen. Ook dat is een indicatie dat die predictoren in de vorige modellen te veel teerden op de impact van één of een aantal individuele lexemen. Met name hoogfrequente individuele werkwoorden kunnen voorspelbaarheidsmaten vertekenen: in een lexicaal ongeïnformeerd model kunnen ze suggereren dat een bepaald patroon een algemeen voorspelbaarheidseffect is, terwijl het eigenlijk deels of misschien vooral het effect is van een individueel werkwoord. Een lexicaal geïnformeerd model kan het aparte effect van dat éne werkwoord isoleren en zo een realistischer beeld geven van het bredere, algemene effect van de voorspelbaarheidsmaat.

Ten derde zien we nu wel een indicatie van een hogere *goodness of fit* voor NL-TwNC, met name voor de tweede C-index. Op zich is dit verschil (0,96 > 0,94) niet erg groot, maar de toename qua fit bij gebruik van de random factoren is veel groter in NL-TwNC (0,8 naar 0,96) dan in BE-LeNC (van 0,83 naar 0,94). Deze observatie kan geïnterpreteerd worden als een indicatie van het grotere belang van de lexicale patronen in NL-TwNC, en met name als een bevestiging van onderzoeksvraag (2.c). In de volgende sectie gaan we dieper in op die laatste bevinding.

6. Bespreking

Het relatief grotere gewicht van lexicale factoren in Nederlandse modellen van syntactische alternanties blijkt niet alleen uit de *er*-alternantie. Zo laat Pijpops (2019) zien dat ook de keuze tussen transitieve constructies (bv. *iemand huwen*, *iemand spreken*) en constructies met een prepositioneel object (bv. *met iemand huwen*, *met iemand spreken*) in Nederlandse modellen meer gestuurd wordt door lexicale factoren (bv. het specifieke werkwoord) dan in Belgische. Een mogelijk gerelateerd patroon is de observatie in verschillende (op manuele codering gebaseerde) studies dat alternanties in de Nederlandse data gemakkelijker te modelleren blijken dan in de Belgische data. Dat komt omdat er minder predictoren nodig zijn om tot een goed model te komen, en omdat het effect van die predictoren bijzonder robuust is. Dat is niet alleen het geval bij onze *er*-alternantie, maar ook voor woordvolgordevariatie in werkwoordclusters (rode en groene woordvolgorde; De Sutter,

Speelman & Geeraerts 2005) en voor causatieve constructies (*doen*/*laten*-alternantie; Speelman & Geeraerts 2009).

De robuustere, eenvoudiger te beregelen alternantiepatronen in Nederland kunnen tentatief gekoppeld worden aan de vroegere standaardisering van het Nederlands in Nederland. Het is niet waarschijnlijk dat de top-down planning die met die standaardisering gepaard ging impact gehad heeft op constructionele keuzes: anders dan fonetische en lexicale keuzes zitten die te diep in onze syntactische motor om actief stuurbaar te zijn. Veel plausibeler is het idee dat in een taal die langer in standaardvorm bestaat een graduele uitkristallisering van de taakverdeling tussen varianten in een alternantiepatroon plaatsvindt: de grotere uniformiteit in taalgebruik die het gevolg is van het standaardiseringsproces vertaalt zich dan in een grotere uniformiteit in de manier waarop varianten in een alternantiepatroon zich tot elkaar verhouden, en dus in een duidelijkere taakverdeling tussen die varianten. In Grondelaers, Speelman & Geeraerts (2008) hebben we beargumenteerd dat plaatsbepalingszinnen in het Nederlandse Nederlands functioneel gespecialiseerd zijn om voorspelbare informatie te introduceren, terwijl *er*-initiële zinnen de functie van onvoorspelbare-informatiedrager in zich bergen. Die specialisering correleert met een 'conventionalisering' van de constructionele ingrediënten tot constituenten die de voorspelbaarheid van het onderwerp bevorderen (in bepalingsinitiële zinnen), of constituenten die dat niet doen (in *er*-initiële zinnen). Het gevolg daarvan is natuurlijk een makkelijkere modelleerbaarheid. Maar die uitgekristalliseerde taakverdeling kan ook leiden tot een 'verstening' in specifieke lexicale voorkeuren: de gereduceerde combinatoriek die het gevolg is van functionele specialisatie zorgt ervoor dat een kleiner aantal lexemen vaker in elkaars buurt voorkomen, en in die hoedanigheid in hogere mate kunnen colloceren met elkaar, en/of met *er*.

Momenteel is deze gedachtegang nog niet meer dan een werkhypothese om een aantal opvallende constructionele Noord/Zuid-verschillen te verklaren. Het moet met name nog blijken of het beschreven fenomeen (makkelijkere voorspelbaarheid van, en grotere lexicale invloed op Nederlandse alternanties) zich ook in andere alternantiepatronen voordoet. Het probleem daarbij is dat we geen goed zicht hebben op welke syntactische alternantiepatronen in het Nederlands gevoelig zijn voor nationale variatie. Het artikel *Vissen naar variatie* (Grondelaers et al., dit nummer) stelt een computationele methode voor om in Nederlandse en Belgische ondertitelcorpora met vertaalsoftware op zoek te gaan naar een ruimer aantal constructionele patronen die nationale verschillen indiceren. De in dit artikel besproken

regressieopschaling is met name bedoeld om de analyse van een groot aantal nationaal stratificerende syntactische variabelen mogelijk te maken en naar een hoger niveau te tillen. Alleen op die manier kunnen eventuele diepe verschillen tussen de grammatica van het Belgische en Nederlandse Nederlands ontdekt en in kaart worden gebracht.

7. Conclusies

Als we onze bevindingen bespreken in termen van de onderzoeksvragen die we hierboven formuleerden, dan kunnen we de volgende conclusies trekken.

1. De automatische modellen voor de *er*-alternantie behalen hoge waarden voor *goodness of fit*, die minstens vergelijkbaar zijn met de waarden uit de modellen op basis van manuele codering. Dat stemt uiteraard optimistisch. We moeten evenwel in rekening brengen dat de vergelijking niet helemaal eerlijk is. Zoals eerder aangegeven gooien de automatisch gegenereerde predictoren 'een breed net uit', en bevatten ze wellicht meer informatie dan we er strikt genomen mee willen operationaliseren, waardoor het niet volledig correct is om hun prestatie rechtstreeks te vergelijken met die van de manueel gecodeerde predictoren, die vaak meer 'gerichte' informatie bevatten. De manueel gecodeerde variabele 'bepalingsconcreetheid', bijvoorbeeld, rangschikt bepalingsreferenten van (i) driedimensionaal concreet ('broodtrommeltje'), over (ii) tweedimensionaal concreet ('weide'), (iii) zowel concrete als abstracte interpretatie mogelijk ('school') naar (iv) volledig abstract ('wiskunde'). De automatische analyse deelt de attestaties in negen categorieën onderling gelijkende nomina op, maar die negen categorieën zijn niet zonder meer in termen van concreetheid te hiërarchiseren. Niettemin interpreteren we de hoge C-indices van de modellen als een indicatie dat de benadering werkbaar en beloftevol is.
2. De nieuwe modellen bevestigen grotendeels, maar niet volledig de oudere modellen:
    a. Als we kijken naar het relatieve belang van de variabelen in de modellen beschreven in Sectie 5, en met name ook in het finale model 4, vinden we dan de bevestiging dat 'semantische klassen' van de bepaling voor de Nederlandse data een belangrijkere predictor is dan voor de Belgische? Naar effectgrootte hebben we in dit onderzoek niet gekeken (dat zou een detailstudie van de individuele semantische klassen impliceren), maar qua verklarende waarde is de variabele alvast relatief belangrijker in de Nederlandse data.

b. We vinden niet echt duidelijke indicaties in de nieuwe modellen dat de Nederlandse data gemakkelijker te modelleren zijn dan de Belgische.
c. Het grotere verschil tussen de eerste en de tweede C-index voor de Nederlandse data in het finale model 4 (in Sectie 5) bevestigt dat specifieke lexicale patronen in de Nederlandse data een grotere impact op de *er*-alternantie hebben dan in de Belgische data.

Daarnaast lijkt het ons belangrijk om, enigszins los van de onderzoeksvragen uit deze studie, onze ervaring met de in dit artikel beschreven en toegepaste methode samen te vatten. Op methodologisch gebied is onze vaststelling dat deze nieuwe methode een bruikbare maar complexe techniek is waarvan de resultaten met enige omzichtigheid moeten worden geïnterpreteerd. Om te beginnen moet en kan de techniek nog verder worden verfijnd. De voorgestelde operationaliseringen voor semantische en pragmatische predictoren gooien weliswaar een breed net uit, maar vangen op die manier ook wel wat ruis. Daarom is de techniek vooral nuttig om de grote krachtlijnen van variatiepatronen in kaart te brengen. We hopen in de casestudy aangetoond te hebben dat de techniek ons belangrijke dingen kan leren over het relatieve belang van verschillende types van predictoren. Daarnaast spreekt het vanzelf dat het arsenaal aan automatisch berekenbare predictoren nog moet worden uitgebreid om de voorziene bredere waaier aan alternantiepatronen aan te kunnen. Wij geloven niet dat dit arsenaal ooit volledig de rijkdom zal kunnen benaderen van wat manueel gecodeerd kan worden als predictor, maar we zijn ervan overtuigd dat onze aanpak kan uitgebreid worden tot een niveau waarop grootschalige vergelijkingen van grotere collecties alternantiepatronen haalbaar en zinvol worden.

Referenties

Arnold, Jennifer E., Thomas Wasow, Ryan Ginstrom & Anthony Losongco (2000). Heaviness vs. Newness: the effects of structural complexity and discourse status on constituent ordering. *Language* 76, 28-55.

Baayen, R. Harald (2011). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics* 11, 295-328.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen (2007). Predicting the dative alternation. In: Gerlof Bouma, Irene Krämer & Joost Zwarts (Eds.), *Cognitive Foundations of Interpretation*. Amsterdam: Royal Netherlands Academy of Arts and Sciences, 69-94.

Daelemans, Walter & Antal Van den Bosch. (2005). *Memory-based language processing*. Cambridge: Cambridge University Press.

De Sutter, Gert, Dirk Speelman & Dirk Geeraerts. (2005). Regionale en stilistische effecten op de woordvolgorde in werkwoordelijke eindgroepen. *Nederlandse Taalkunde* 10, 97-128.

De Troij, Robbert, Stefan Grondelaers, Dirk Speelman & Antal Van den Bosch. (geaccepteerd). Lexicon or grammar? Using Memory-Based Learning to investigate the syntactic relationship between Belgian and Netherlandic Dutch. Te verschijnen in *Natural Language Engineering*.

Everitt, Brian S., Sabine Landau, and Morven Leese (2001). *Cluster analysis*. London: Arnold.

Evert, Stefan. (2009). Corpora and collocations. In: Anke Lüdeling & Merja Kytö (Eds.), *Corpus linguistics: an international handbook*. Vol. 2. Berlin and New York: Mouton De Gruyter, 1212–1248.

Firth, John R. (1957). *Papers in Linguistics 1934-1951*. Londen: Oxford University Press.

Gries, Stefan Th. (2001). A multifactorial analysis of syntactic variation: Particle movement revisited. *Journal of Quantitative Linguistics* 8(1), 33-50.

Gries, Stefan Th. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 43, 365-399.

Gries, Stefan Th. (2013). 50-something years of work on collocations: what is or should be next . . . *International Journal of Corpus Linguistics* 18(1), 137-165.

Grondelaers, Stefan, Dirk Speelman & Dirk Geeraerts (2002). Regressing on er. Statistical analysis of texts and language variation. In: Annie Morin & Pascale Sébillot (Eds.), *6ièmes journées internationales d'analyze statistique des données textuelles* (6th international conference on textual data statistical analysis). Rennes: Institut National de Recherche en Informatique et en Automatique, 335-346.

Grondelaers, Stefan, Marc Brysbaert, Dirk Speelman & Dirk Geeraerts (2002). *Er* als accessibility marker: on- en offline evidentie voor een procedurele interpretatie van presentatieve zinnen. *Gramma/TTT: Tijdschrift voor Taalwetenschap* 9 (1), 1-22.

Grondelaers, Stefan & Dirk Speelman (2007). A variationist account of constituent ordering in presentative sentences in Belgian Dutch. *Corpus Linguistics and Linguistic Theory* 3, 161-193.

Grondelaers, Stefan, Dirk Speelman & Dirk Geeraerts (2008). National variation in the use of *er* "there": regional and diachronic constraints on cognitive explanations. In: Gitte Kristiansen & René Dirven (Eds.), *Cognitive sociolinguistics: language variation, cultural models, social systems*. Berlin & New York: Mouton de Gruyter, 153-204.

Grondelaers, Stefan, Dirk Speelman, Denis Drieghe, Marc Brysbaert & Dirk Geeraerts (2009). Introducing a new entity into discourse: comprehension and production evidence for the status of Dutch *er* "there" as a higher-level expectancy monitor. *Acta Psychologica* 130 (2), 153-160.

Jaeger, T. Florian. (2005). Optional *that* indicates production difficulty: Evidence from disfluencies. In: *Proceedings of DiSS05, Disfluency in Spontaneous Speech Workshop*, 103—109. Aix-en-Provence, France.

Lapata, Maria. (1999). Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College Park, Maryland, USA: Association for Computational Linguistics, 397-–404.

Levin, Beth (1993). *English Verb Classes and Alternations: A preliminary investigation*. Chicago, IL: University of Chicago Press.

Lin, Dekang (1998). Automatic retrieval and clustering of similar words. In: *Proceedings of the 17th international conference on Computational linguistics*. Montreal, Canada, 768-774.

Levshina, Natalia & Kris Heylen (2014). A radically data-driven Construction Grammar: Experiments with Dutch causative constructions. In: Ronny Boogaart, Timothy Colleman & Gijsbert Rutten (Eds.), *Constructions in Germanic – Extending the scope*. Berlin & New York: Mouton de Gruyter, , 17-–46. doi:10.1515/9783110366273.17

Levy, Roger& T. Florian Jaeger (2007). Speakers optimize information density through syntactic reduction. In: Bernhar Schölkopf, John C. Platt & Thomas Hoffman (Eds.),

Advances in neural information processing systems Vol. 19,. Cambridge, MA: MIT Press, 849—856.

Ordelman, Roeland, Francisca De Jong, Arjan Van Hessen & Henri Hondorp (2007). TwNC: a Multifaceted Dutch News Corpus. *ELRA Newsletter* 12 (3–4). Available at http://doc.utwente.nl/68090/ (last accessed 23 December 2014)

Padó, Sebastian & Mirella Lapata (2010). Dependency-based construction of semantic space models. *Computational Linguistics* 33, 161-199.

Pijpops, Dirk (2019). *How, why and where does argument structure vary? A usage-based investigation into the Dutch transitive-prepositional alternation*. Doctorale dissertatie KU Leuven.

Speelman, Dirk & Dirk Geeraerts (2009). Causes for causatives: the case of Dutch *doen* and *laten*. In: Eve Sweetser & Ted Sanders (Eds.), *Causal categories in discourse and cognition*. Berlin & New York: Mouton de Gruyter, 173-204.

Speelman Dirk, Kris Heylen & Stefan Grondelaers (te verschijnen). A bottom-up, data-driven operationalization of semantic classes and predictability in syntactic alternation research. In: Tanja Karoli Christensen & Jensen Torben Juel (Ed.) *Explanations in Sociosyntax*. Cambridge University Press.

Stefanowitsch Anatol & Stefan Th. Gries (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1(1), 1-43.

Tagliamonte, Sali & R. Harald Baayen (2012). Models, forests and trees of York English: *Was/were* variation as a case study of statistical practice. *Language Variation and Change* 24, 135-178.

Theijssen, Daphne, Lou Boves, Hans van Halteren & Nelleke Oostdijk (2010). Evaluating automatic annotation: automatically detecting and enriching instances of the dative alternation. *Language Resources and Evaluation* 46, 565-600.

Turney, Peter D. & Patrick Pantel (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37, 141-188.

Vandenbosch, Antal. 2012. Example-based modeling of syntactic alternations. Plenary talk at New Ways of Analyzing Syntactic Variation, November 16th 2012, Radboud University Nijmegen.

Over de auteur(s)

Dirk Speelman, KU Leuven
E-mail: dirk.speelman@kuleuven.be

Stefan Grondelaers, Radboud Universiteit Nijmegen
E-mail: s.grondelaers@let.ru.nl

Benedikt Szmrecsanyi, KU Leuven
E-mail: benedict.szmrecsanyi@kuleuven.be

Kris Heylen, KU Leuven
E-mail: kris.heylen@kuleuven.be