De verleidingen en gevaren van GrETEL*

Jan Odijk

NT 25 (1): 7–37
DOI: 10.5117/NEDTAA2020.1.002.ODIJ

Abstract
Corpora are a useful and important source of evidence for linguistic research, but they are not the only kind of evidence, do not have any special status as evidence, and have their limitations. Recent very user-friendly applications such as GrETEL make it very easy to search in large and richly annotated corpora on the basis of an example sentence and without knowledge of a query language or the exact nature of the linguistic annotations. It is therefore very tempting to use these applications intensively. That is fine, but also dangerous in ways, because in many cases, in order to interpret the results correctly, the researcher must really be aware of the precise nature of the linguistic annotations and of the way in which the user-friendly interface generates a query on the basis of an example sentence. I will illustrate this with several examples. I also sketch some methods for avoiding or mitigating the dangers and argue that the applications should support these methods also in as user-friendly a manner as possible.

Keywords: corpus analysis, corpus applications, reliability of corpus analysis

1. Inleiding

De beschikbaarheid en de toegankelijkheid van grote corpora is in de laatste jaren enorm toegenomen, zoals ik schets in sectie 2. Corpora bieden

* Ik ben de deelnemers aan de Dag van de Nederlandse Zinsbouw en de beoordelaars van dit artikel erkentelijk voor hun nuttige en constructieve commentaar.

een nuttige en belangrijke bron van bewijsmateriaal voor taalkundig onderzoek. Ik onderbouw dit aan de hand van concrete voorbeelden in sectie 3. Corpora vormen echter niet de enige soort bewijsmateriaal voor taalkundig onderzoek, hebben geen speciale status als zodanig, en hebben hun beperkingen (zie sectie 4). Nieuwe zeer gebruikersvriendelijke applicaties zoals GrETEL maken het heel gemakkelijk grote en rijk geannoteerde corpora te doorzoeken op basis van een voorbeeldzin en zonder kennis van een zoekopdrachttaal of de precieze aard van de taalkundige annotaties, zoals ik illustreer in sectie 5. Het is daarom heel verleidelijk deze applicaties intensief te gebruiken. Dat is goed, maar er schuilen ook grote gevaren in, want in de praktijk moet een onderzoeker wel degelijk weten wat de aard van de taalkundige annotaties is en hoe de gebruikersvriendelijke interface een zoekopdracht op basis van een voorbeeld genereert. Ik illustreer dat met verschillende voorbeelden in sectie 6. Er zijn methodes om deze gevaren te vermijden of te reduceren, waarvan ik er een paar beschrijf in sectie 7. Ik beargumenteer dat de applicaties deze methodes ook zo gebruikersvriendelijk mogelijk moeten ondersteunen. In sectie 8 vat ik mijn conclusies samen.

2. Corpusgebaseerde analyse

Het is nuttig en soms zelfs noodzakelijk om bij een taalkundige analyse gebruik te maken van corpusmateriaal en van software om corpusmateriaal te analyseren. Dit is nuttig omdat het kan leiden tot een betere empirische onderbouwing van hypotheses en theorieën (bijvoorbeeld omdat die dan gebaseerd zijn op meer data). Het kan taalkundig onderzoek versnellen (omdat men sneller en makkelijker relevante data kan vinden), en het gebruik van corpora en geavanceerde analysetools biedt zelfs het potentieel om nieuwe verbanden te vinden in data. Het is soms noodzakelijk wanneer er geen andere typen data beschikbaar zijn.

Het gebruik van corpora en geavanceerde softwaretools om corpora te doorzoeken en te analyseren is in de laatste 10-15 jaar veel toegankelijker geworden door een reeks van projecten die werkten aan de onderzoeksinfrastructuur voor taalkundig onderzoek, zoals het STEVIN-programma (2004-2010, Spyns & Odijk, 2013), CLARIN-NL (2009-2015, Odijk & Van Hessen 2017), CLARIAH-CORE (2015-2019), Nederlab (2014-2018), en vele kleinere projecten.

Hierdoor zijn corpora en andere data beschikbaar gekomen, evenals een reeks van applicaties en tools voor het doorzoeken en analyseren van deze

data. Er zijn veel data beschikbaar gekomen die alleen orthografische representaties bevatten, zoals Dutch Google N-grams,[1] Dutch Twitter N-grams,[2] CHILDES,[3] en TwinNL Dutch Tweets[4] die zoeken via Google Search complementeren. Daarnaast ook data die verrijkt zijn met taalkundige annotaties (zoals woordsoortinformatie, informatie over inflectie, over lemma, etc.) op voorkomens van woordvormen ('tokens'), bijvoorbeeld corpora zoals SoNaR,[5] SoNaR New Media,[6] CGN,[7] en BASILEX,[8] en toepassingen zoals Nederlab,[9] OpenSoNaR,[10] CHN,[11] AutoSearch[12] en CoW.[13] Tot slot zijn er corpora met een syntactische ontleding voor iedere zin in het corpus (treebanks en parsebanks),[14] zoals de CGN-treebank,[15] LASSY-Klein,[16] LASSY-Groot,[17] en CHILDES Dutch,[18] en daarnaast ook een heel reeks van treebank zoek- en analyseapplicaties zoals DACT,[19] PaQu,[20] GrETEL,[21] GrETEL 4,[22] CESAR,[23] en SPOD.[24]

1 https://urd2.let.rug.nl/~gosse/bin/Web1T5_freq.perl
2 http://www.let.rug.nl/gosse/Ngrams
3 http://childes.psy.cmu.edu/data/
4 https://twinl.surfsara.nl/ en http://145.100.59.92/cgi-bin/twitter
5 https://ivdnt.org/taalmaterialen/2026-tstc-sonar-corpus
6 https://ivdnt.org/taalmaterialen/2002-tstc-sonar-nieuwe-media-corpus-1
7 https://ivdnt.org/taalmaterialen/1993-tstc-corpus-gesproken-nederlands
8 https://ivdnt.org/downloads/tstc-basilex-corpus
9 http://www.nederlab.nl/
10 https://portal.clarin.inl.nl/opensonar_frontend/opensonar/search
11 http://chn.inl.nl/
12 http://portal.clarin.nl/node/4222
13 https://corporafromtheweb.org/
14 Het onderscheid tussen treebanks en parsebanks wordt vaak niet gemaakt, maar ik versta onder een treebank een corpus waarin aan iedere zin een *manueel geverifieerde* syntactische analyse is toegewezen, en onder een parsebank een corpus waarin aan iedere zin een *automatisch gegenereerde* syntactische analyse is toegewezen. Kiril Simov wees me op het onderscheid (p.c.).
15 https://paqu.let.rug.nl:8068/?db=cgn&word=&rel=&hword=&postag=&hpostag=&meta=&sn=10
16 https://ivdnt.org/taalmaterialen/2037-tstc-lassy-klein-corpus
17 https://ivdnt.org/taalmaterialen/2056-tstc-lassy-groot-corpus
18 https://paqu.let.rug.nl:8068/?db=childesdutch&word=&rel=&hword=&postag=&hpostag=&meta=&sn=10
19 http://rug-compling.github.io/dact/
20 http://www.let.rug.nl/alfa/paqu
21 https://gretel.ccl.kuleuven.be/gretel3/
22 http://gretel.hum.uu.nl/gretel4/ng/home
23 https://cesar.science.ru.nl/
24 http://www.let.rug.nl/alfa/paqu/spod

Ik heb zelf ook het nodige gedaan om de beschikbaarheid hiervan grotere bekendheid te geven en studenten en onderzoekers op te leiden om deze applicaties te gebruiken in het onderzoek. Zo heb ik in de periode van 2014-2018 meer dan 27 presentaties over corpusgebaseerde analyse gegeven, en is er een lezing van mij over GrETEL op video opgenomen[25] en via het web beschikbaar gemaakt. Ik heb acht relevante publicaties uitgebracht in de periode 2015-2019.[26] Er is een Lingua Special Issue met drie corpusgebaseerde studies uitgebracht onder mijn redacteurschap (Odijk 2016a, 2016b).[27] Er zijn door mij twee LOT-cursussen georganiseerd waarin corpusanalyse een belangrijke rol speelde. Ik geef ook een cursus corpusanalyse in het reguliere universitair curriculum sinds 2014. Ik heb een project geïnitieerd voor de verrijking van het Taalportaal (Van der Wouden et al. 2016a) met corpuslinks (Bouma et al. 2015, Van der Wouden et al. 2016b, 2017). Ik was redacteur van het Open Access boek over CLARIN in Nederland en Vlaanderen, waarin zes hoofdstukken over corpusanalyse gaan (Odijk & Van Hessen 2017). De Lassy Relatiezoekapplicatie[28] is ontwikkeld in Groningen maar dreigde door gebrek aan onderhoud ten onder te gaan. Ik zorgde ervoor dat er geld beschikbaar kwam voor onderhoud en het uitbreiden van de functionaliteit, resulterend in de treebankapplicatie PaQu[29] (Odijk et al. 2017) en later de syntactische profileerapplicatie SPOD.[30] Tot slot werk ik in Utrecht met een team ontwikkelaars aan een upgrade van GrETEL, versie 4.[31]

Ik zal ook in dit artikel allerlei voordelen van corpusanalyse bij taalkundig onderzoek naar voren brengen. Maar mijn activiteiten om corpusanalyse te bevorderen nopen mij ook te wijzen op de nadelen, beperkingen en zelfs de gevaren van corpusanalyse. Dat betreft kwesties die grotendeels bekend zijn onder corpustaalkundigen en andere taalkundigen die ervaring hebben met corpusgebaseerd onderzoek. Maar toepassingen zoals PaQu en GrETEL maken het gebruik van corpora bij het onderzoek zo gemakkelijk en verleidelijk dat de beperkingen en gevaren snel over het hoofd gezien worden. Daar wil ik in dit artikel dan ook met nadruk op wijzen.

25 http://lecturenet.uu.nl/Site1/Catalog/Full/c9f887bc45154af5bd7cdb218216816621
26 Bouma et al. (2015), Odijk (2015, 2016c), Odijk et al. (2017, 2018a, 2018b), Van der Wouden et al. (2016b, 2017).
27 http://dx.doi.org/10.1016/j.lingua.2016.04.003
28 http://portal.clarin.nl/node/14354
29 http://www.let.rug.nl/alfa/paqu
30 http://www.let.rug.nl/alfa/paqu/spod
31 http://gretel.hum.uu.nl/gretel4/ng/home

Hoewel ik zelf vaak corpora bij taalkundige analyse gebruik, en het gebruik ervan door andere onderzoekers stimuleer en faciliteer, zou ik mijzelf nooit een 'corpustaalkundige' noemen, en zou ik die activiteit nooit 'corpustaalkunde' noemen. Het gebruik van die termen suggereert namelijk een speciale status voor corpora als empirisch bewijsmateriaal boven andere soorten. En dat zie ik zeker anders: corpora vormen slechts één mogelijke bron, en een taalkundige moet alle relevante bronnen in beschouwing nemen. Er zijn veel andere vormen: de bekendste zijn resultaten van experimenten, waarin gecontroleerd (zij het vaak in artificiële omstandigheden) data verzameld worden. Dit kunnen allerlei soorten experimenten zijn. Voorbeelden zijn experimenten waarin geprobeerd wordt een relatie te vinden tussen taalkundige processen of eigenschappen en andere processen, bijv. oogbewegingen door middel van eye-tracking, of verwerkingstijd gemeten via responstijden, etc. Ook het afnemen van oordelen van moedertaalsprekers over welgevormdheid, acceptabiliteit en interpretatie van uitingen is een voorbeeld van dergelijke experimenten. Daaronder valt ook het afnemen van de eigen oordelen (introspectie) over welgevormdheid, acceptabiliteit en interpretatie van uitingen, wat soms enigszins denigrerend 'armchair linguistics' wordt genoemd.[32] Dergelijke introspectie levert ook empirische observaties.[33] Vanzelfsprekend geldt voor alle vormen van het verzamelen van bewijsmateriaal dat er de nodige zorg aan besteed moet worden om de experimenten zo te laten lopen zoals ze bedoeld zijn en ongewenste storende factoren zoveel mogelijk uit te sluiten.[34] Voor

32 Eerlijk gezegd zou ik de term 'armchair linguist' als geuzennaam willen omarmen: het is immers een van de meest efficiënte, goedkope en ook meest betrouwbare manieren om taalkundige data te verzamelen. Dat geldt natuurlijk alleen als het met de nodige zorg wordt uitgevoerd, en als men er zich van bewust is dat er (zoals bij ieder experiment) storende factoren kunnen zijn. Zo zijn intuïties onderhevig aan performancefactoren, bijv. het kan gebeuren dat men op een bepaald moment een tweede interpretatie van een dubbelzinnige uiting niet 'ziet', terwijl die op een ander tijdstip wel 'gezien' wordt. Men moet zich er ook van vergewissen dat de oordelen ook door anderen gedeeld worden, want men zou niet graag een theorie uitsluitend baseren op het oordeel van de taalkundige die de theorie voorstelt. Maar dit is niets bijzonders, want men zou ook nooit een theorie uitsluitend baseren op een enkel voorkomen van een verschijnsel in een corpus.

33 Hoewel dat door sommigen ontkend lijkt te worden, bijv. wiktionary definieert armchair linguistics als 'any linguistic enterprise employing introspection rather than empirical methods' (https://en.wiktionary.org/wiki/armchair_linguistics). Zie ook verderop in dit artikel bij de bespreking van werk van Augustinus over IPP en de bespreking van de naam GrETEL.

34 Er is natuurlijk een zeer uitgebreide discussie in de literatuur over de status van intuitieve oordelen als bewijsmateriaal voor taalkundig onderzoek, waar ik hier verder niet op

taalkundige theorieën die claimen iets te zeggen over hoe taal gerepresenteerd is in het brein vormen bijv. hersenscans vanzelfsprekend ook potentieel relevant bewijsmateriaal.

Kortom: al het relevante bewijsmateriaal moet in beschouwing worden genomen, en geen enkele vorm of bron heeft een geprivilegieerde status. De verschillende vormen kunnen elkaar vaak aanvullen. Charles Fillmore heeft dit prachtig verwoord voor 'armchair linguistics' versus corpusgebaseerde taalkunde, en ik citeer dat hier:

> I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way. My conclusion is that the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body. (Fillmore 1992: 35)

3. Voordelen van corpusgebaseerde analyse

Het gebruik van corpora bij taalkundige analyse heeft vele voordelen en is soms noodzakelijk. In bepaalde gevallen moet men gebruik maken van corpora om de eenvoudige reden dat er geen andere data zijn of te krijgen zijn. Voor de studie van een taal uit eerdere tijden zijn er geen moedertaalsprekers maar slechts corpora beschikbaar. Hetzelfde geldt voor spontane spraak van kinderen (transcripties van spontane kindertaal). In andere gevallen zijn er in principe wel andere data te verkrijgen, maar kost dat een grote inspanning en veel geld, bijv. als een taal in een geheel ander deel van de wereld gesproken wordt door een zeer klein aantal mensen.

Corpora vormen authentieke gegevens van daadwerkelijk taalgebruik. Voor de bestudering van spontane taal en voor de bestudering van performancefouten (aarzelingen, herstarts, versprekingen, gevulde pauzes, etc.) zijn corpora met authentiek taalgebruik onontbeerlijk. Omdat in corpora daadwerkelijk taalgebruik geregistreerd is, kan men frequentietellingen

in wil gaan. Enkele relevante werken in deze context zijn Langendoen et al. (1973); Schütze (1996); Edelman & Christiansen (2003); Wasow & Arnold (2005); Gibson & Fedorenko (2010); Gibson, Piantadosie & Fedorenko(2013); Phillips & Lasnik (2003); Featherston (2009); Phillips (2010); Pullum (2017); Linzen & Oseki (2018).

doen van het gebruik van bepaalde woorden, woordcombinaties, en constructies. Corpora met authentiek taalmateriaal kunnen gebruikt worden om de eigen taalintuïties te toetsen. Zo onderzoeken Odijk et al. (2017: 292-293) de complementatie-eigenschappen van de (bijna-synonieme) koppelwerkwoorden *raken* en *worden*. De auteurs nemen op basis van de intuïtieve oordelen in (1) aan dat *worden* NP's en AP's maar geen PP's als predicatief complement kan nemen, terwijl *raken* AP's en PP's maar geen NP's als predicatief complement kan nemen:

(1) a. Zij werd / raakte zwanger.
b. Zij *werd / raakte in verwachting.
c. Zij werd / *raakte dokter.

Maar corpusonderzoek in de Lassy-Klein en Corpus Gesproken Nederlands (CGN) treebanks met behulp van PaQu levert wel degelijk voorbeelden op waarbij *worden* een PP als predicatief complement neemt: bijv. *van* (*groot*) *belang worden, van kracht worden*.

Odijk (2016c) onderzoekt het modificatiepotentieel van de bijna-synonieme woorden *heel*, *erg* en *zeer* in hun betekenis 'in hoge mate'. Het lijkt erop dat *heel* uitsluitend adjectieven kan modificeren, terwijl *erg* en *zeer* ook adposities en werkwoorden kunnen modificeren.[35] Zie (2).

(2) a. Hij is daar heel / erg / zeer blij mee.
b. Hij is daar *heel / erg / zeer in zijn sas mee.
c. … omdat dat mij *heel / erg / zeer verbaast

Maar corpusonderzoek laat zien dat er wel degelijk voorbeelden zijn waarin *heel* een adpositie modificeert, zie (3):

(3) Heel af en toe,[36] heel in de verte, heel in het algemeen, heel in het bijzonder, heel in het kort, heel in het begin, heel op het laatst, heel uit de verte, heel aan het eind

Wanneer corpora bovendien nog verrijkt zijn met taalkundige annotaties, zoals woordsoort, lemma en inflectie-informatie voor woordvoorkomens,

35 In constituent-gebaseerde theorieën zal men waarschijnlijk zeggen dat deze woorden adpositionele en verbale *constituenten* modificeren. Ik wil neutraal blijven in deze kwestie.
36 Het is niet onmiddellijk duidelijk wat de categorie van de uitdrukking *af en toe* is. Nader onderzoek is hiernaar vereist.

of zelfs volledige syntactische structuren voor zinnen of uitingen, worden de mogelijkheden om efficiënt relevante data te vinden in grote databestanden enorm interessant voor taalkundig onderzoek. Met behulp van (naar mijn mening) zeer gebruikersvriendelijke interfaces zoals geboden door applicaties als OpenSoNaR, PaQu en GrETEL wordt het gebruik van die corpora ook erg gemakkelijk.

4. Nadelen en beperkingen van corpusgebaseerde analyse

Naast de voordelen en mogelijkheden die het gebruik van corpora bij taalkundig onderzoek bieden zijn er ook nadelen en beperkingen aan het gebruik van corpora.

Geen minimale paren. Minimale paren spelen een grote rol bij heel veel taalkundig onderzoek, maar kunnen normaal gesproken alleen gecreëerd worden in een gecontroleerd experiment. Minimale paren zijn zo belangrijk omdat ze het mogelijk maken te onderzoeken wat de invloed van een enkele variabele is. Zonder minimale paren zijn er vrijwel altijd storende factoren die het moeilijk of zelfs onmogelijk maken betrouwbare conclusies te trekken. Corpora leveren authentieke data maar dat impliceert dat de kans om minimale paren tegen te komen in corpora erg klein is.

Slechts een steekproef. Ieder corpus, hoe groot ook, is altijd slechts een steekproef van daadwerkelijk taalgebruik. En veel verschijnselen in natuurlijke taal zijn zeldzaam. Zo onderzocht Augustinus (2015) het verschijnsel IPP (Infinitivus Pro Participio). Daartoe bestudeerde ze eerst de literatuur die al eerder verschenen was over dit onderwerp. Na het bestuderen van die literatuur concludeert ze:[37]

> The fact that several authors disagree on the set of verbs that can occur as IPP indicates that an *empirical* investigation is necessary in order to identify the Dutch IPP verbs. (Augustinus 2015: 14; mijn cursivering, JO)

37 Ik heb het woord *empirical* cursief gezet. Augustinus impliceert hiermee dat de auteurs die zij bestudeerd heeft geen empirisch onderzoek hebben gedaan. Dat wens ik te bestrijden. Zelfs als deze auteurs uitsluitend introspectieve observaties hebben gebruikt bij hun onderzoek is dit onderzoek empirisch. Dit geeft mij ook de gelegenheid mijn enige fundamentele bezwaar tegen GrETEL naar voren te brengen, nl. de tweede *E* in het acroniem. De

En ze vangt een zoektocht in treebanks aan met GrETEL. Na die zoektocht vat Augustinus haar bevindingen samen. Ik citeer en benadruk een aantal frases:

> The data **do not contain any IPP verbs that are not listed in the literature on the topic**, and **due to data sparseness**, **not all verbs encountered in the literature occur as IPP in the treebank data**. For those verbs, their IPP status was defined by means of Internet data, in order to construct a typology that is based on empirical data. (Augustinus 2015: 171; mijn markering, JO)

Kortom, corpora leveren niet altijd de data waar men naar op zoek is.

Een tweede voorbeeld komt uit Odijk (2018), die onderzoekt hoe kinderen het modificatiepotentieel van de woorden *heel* en *zeer* leren. In het Nederlands van volwassenen lijkt *heel* geen PP's te kunnen modificeren op een tiental uitzonderingen na (zoals we boven gezien hebben). Het woord *zeer* daarentegen kan productief PP's modificeren. Maar hoe moeten kinderen dat leren als blijkt dat in corpora die typerend zijn voor de input die ze bij eerste-taalverwerving krijgen *heel* dat PP's modificeert frequenter voorkomt dan *zeer* dat PP's modificeert? Zie Tabel 1.

Ook hier blijken de data uit corpora, of althans de frequenties ervan, niet zomaar uitsluitsel te kunnen geven: het uitzonderlijke geval komt vaker voor dan het regelmatige geval, en beide zijn laagfrequent.[38] Niet voor niets heerst er onder taalkundigen en andere onderzoekers die cruciaal corpora gebruiken altijd een zekere 'honger' naar meer data en velen gebruiken niet voor niets het adagium *There is no data like more data*.

Geen negatief bewijsmateriaal. De onwelgevormdheid van zinnen speelt een belangrijke rol als bewijsmateriaal in veel taalkundig onderzoek. Maar corpora kunnen geen informatie over onwelgevormdheid leveren. Er zullen ongetwijfeld onwelgevormde zinnen in corpora voorkomen, maar die zijn niet als zodanig gemarkeerd.

expansie van het acroniem GrETEL is *Greedy Extraction of Trees for Empirical Linguistics* en bevat de frase 'empirical linguistics'. Dat is niet fout maar suggereert dat er vormen van taalkunde zijn die niet empirisch zijn, en daar kan ik mij niet in vinden. Een betere naam voor deze applicatie zou dus de tweede *E* weglaten: GrETL.

38 Alleen in de Lassy-Klein en Wikipedia-data (beide niet echt typerend voor taalverwervingsdata) lijken de verhoudingen andersom te liggen, maar ook daar is de frequentie van *zeer* dat een PP modificeert bijzonder laag.

Tabel 1 Frequentie (per miljoen tokens) van *heel* vs. *zeer* als bepaling van PP's

Corpus (per miljoen tokens)	*heel* mod P	*zeer* mod P
LASSY-Klein	0,0	2,7
CGN	7,9	1,8
VanKampenJAC	3,3	0,0
VanKampen LAUorSAR	6,5	0,0
CHILDES Dutch	5,8	1,6
Basilex	1,7	0,3
Wikipedia	0,3	1,9

Automatisch aangemaakte annotaties. Taalkundige verrijkingen van corpora vergroten de bruikbaarheid ervan enorm. Maar vaak zijn die verrijkingen automatisch aangemaakt, en die zullen zeker fouten bevatten (zie ook Bloem, dit nummer). Ook manueel aangemaakte of manueel geverifieerde en gecorrigeerde annotaties zullen trouwens fouten bevatten. Desondanks blijven dergelijke annotaties zeer nuttig, maar men moet er zich bij de analyse wel goed rekenschap van geven dat er fouten in de annotaties kunnen zitten. Ik zal dat verderop in meer detail illustreren.

Corpusdata en intuïties. Het gebruik van corpusdata kan naar mijn mening niet de intuïties van moedertaalsprekers vervangen.[39] Een taalkundige zal ieder voorkomen van een verschijnsel in een corpus kritisch beschouwen, en voor ieder voorbeeld beoordelen hoe het gewaardeerd moet worden. Er komen immers in corpora allerlei verschijnselen voor, inclusief fouten, versprekingen, valse starts en herhalingen, etc. Iedere beoordeling van dit type is natuurlijk een hypothese, maar intuïties over welgevormdheid, ambiguïteit, etc. zullen bij het opstellen van dergelijke hypothese terecht een belangrijke rol spelen. Zo komt in het Nederlandse CHILDES Van Kampen corpus o.a. de volgende uiting voor, uitgesproken door een volwassene (spreker JAC):

(4) ik kijk heel uit (bestand Laura30.cha)

Bij de analyse hiervan concludeer ik niet dat *heel* voor deze spreker werkwoorden kan modificeren, maar stel ik de hypothese op dat het hier een

39 Vanzelfsprekend zal de ontwikkeling van taaltechnologie op basis van corpusdata, of eerste-taalverwerving door een jong kind, (vrijwel) uitsluitend moeten gebeuren op basis van de inputdata en het verwervingsmechanisme, en zal er geen beroep gedaan kunnen worden op intuïties van moedertaalsprekers. Maar dat betekent niet dat intuïties niet gebruikt moeten of mogen worden bij taalkundig onderzoek.

verspreking van de volwassene is (en dat het proces van eerste-taalverwerving robuust moet zijn tegen dergelijke onwelgevormde input). In dit geval kon ik deze hypothese toetsen: de spreker bevestigde inderdaad dat zij de zin onwelgevormd vindt, en vermoedde dat het een transcriptiefout moest zijn.[40]

Tabel 2 Fragmenten uit SoNaR waarin *heel* een PP modificeert die de auteur onwelgevormd acht

Voorbeeld	Subcorpus	Land	Genre
ze DJFacteuR sjah DJFacteuR is **heel van slag** ocharme drm speelk ni graag online	WR-U-E-A-0000104018	B	chats
-). Als ik **heel per ongeluk** in een andere auto rijd dan	WR-P-E-A-0000506410	B	discussion lists
e midweekje naar oostrozebeke: **heel van de kaart**: lol:	WR-P-E-A-0000188664	U	discussion lists
en Janez Detd trok me **heel over de streep**	WR-P-E-A-0000296105	B	discussion lists
deze topic waar ge zo **heel uit de hoogte** doet omdat die persoon dubbelpostte	WR-P-E-A-0000463665	U	discussion lists
ik bedoel gwn die t-shirts **heel in het zwart** snapte gwn mé het logotje …	WR-P-E-A-0000146523	B	discussion lists
… zo iemand tegenkomen en tis **heel in orde**;) kuskes en succes …	WR-P-E-A-0000055681	B	discussion lists
… ne camera is, ist **heel in orde**:) …	WR-P-E-A-0000045331	B	discussion lists
… of Futures jaimykeuh je ben **heel op het slechte pad** een brug oversteken: -	WR-P-E-A-0000188955	B	discussion lists
… ouders die zich hiervoor interesseren **heel op prijs**. Bedankt, Fred …	WR-P-E-A-0004642784	NL	discussion lists
… niet. Ik zou het **heel op prijs** stellen als de vraagsteller de …	WR-P-E-A-0005355195	NL	discussion lists
… zitten nog niet of slechts **heel ten dele** in deze resultaten. Aan …	WR-P-P-G-0000448580	B	newspaper text
…. En dat was ook **heel ten onrechte**, denk ik. Het …	WR-P-E-G-0000008944	B	subtitles

Bij hetzelfde onderzoek vond ik in SoNaR (Oostdijk et al. 2013) verschillende zinnen waarin *heel* een PP modificeert. Voor de gevallen die boven beschreven zijn in (3) accepteer ik die voorbeelden als voorbeelden die ook in mijn idiolect welgevormd zijn (en ik vermoed dat ze algemeen gangbaar zijn in het hele taalgebied). Maar voor de voorbeelden in Tabel 2 accepteer

40 Ik heb geen toegang tot de oorspronkelijke opnames en kan daarom niet toetsen of het hier een verspreking of een transcriptiefout betreft.

ik dat niet: het gebruik van *heel* als bepaling van een PP in die zinnen is en blijft onwelgevormd voor mij. Natuurlijk kan ik deze data niet negeren, en ik zou kunnen veronderstellen dat anderen *heel* blijkbaar wel zo kunnen gebruiken, mogelijk dialectisch bepaald (de meeste voorbeelden zijn uit een informeel genre en uit België afkomstig), mogelijk omdat zij *heel* kunnen gebruiken zoals ik *geheel* zou gebruiken. Hoe deze data ook geanalyseerd moeten worden, ik behandel ze niet op dezelfde manier als andere voorbeelden (zoals de voorbeelden in (3)) uit het corpus, maar waardeer ze op een bepaalde manier, ingegeven door mijn intuïties, en stel hypotheses op over hoe dit bewijsmateriaal het beste verantwoord kan worden.

5. De verleidingen

Met toepassingen zoals OpenSoNaR, GrETEL en PaQu is de ingang tot corpusgebaseerd onderzoek enorm vergemakkelijkt. Het zijn webapplicaties die altijd toegankelijk zijn als men een internetverbinding heeft. Er hoeven geen data of software gedownload te worden, en geen aanpassingen gedaan te worden in verband met het lokale operatingsysteem. De applicaties bieden een waaier aan gebruikersinterfaces, van zeer eenvoudige interfaces voor beginners en/of relatief eenvoudige zoekopdrachten, tot expertinterfaces die de beheersing van een zoekopdrachttaal en goede kennis van de aard van de data vereisen.

Om te illustreren hoe gemakkelijk het is om met nauwelijks voorkennis toch behoorlijk ingewikkelde zoekopdrachten uit te laten voeren, zal ik hier een voorbeeld in detail uitwerken. Daarmee probeer ik te illustreren hoe verleidelijk het is om met deze gemakkelijke gebruikersinterface zoekopdrachten uit te voeren voor onderzoek. Maar die gemakkelijke toegang brengt ook gevaren met zich mee, want het feitelijke onderzoek begint dan pas, zoals ik in sectie 6 zal beschrijven.

We gebruiken GrETEL 4,[41] en gaan zoeken in het Laura-gedeelte van het Nederlandse CHILDES Van Kampen corpus (Van Kampen, 2009).[42] Dit corpus is door mij opgeladen in GrETEL, het is daar automatisch ontleed door Alpino (en daarmee tot een parsebank (zie voetnoot 14) gemaakt) en beschikbaar gemaakt als *vklaura* om het te doorzoeken. Ik heb het gedeeld met iedereen, dus iedereen kan erin zoeken.[43] Een overzicht van alle

41 http://gretel.hum.uu.nl/gretel4/ng/home
42 https://childes.talkbank.org/access/DutchAfrikaans/VanKampen.html
43 Zie http://gretel.hum.uu.nl/gretel-upload/index.php/treebank/show/vklaura

beschikbare te doorzoeken treebanks en parsebanks is in de applicatie beschikbaar.[44] Ieder gebruiker die inlogt kan ook zijn of haar eigen corpus opladen. Verschillende formaten worden daarbij ondersteund.

We zijn geïnteresseerd in constructies met drie 'kale' werkwoorden, dat wil zeggen werkwoorden zonder *te* en *om* in een specifiek grammaticaal verband. Voorbeeld (5) illustreert deze constructie:

(5) Hij wil dat gaan doen

De werkwoorden in deze constructie zullen we aanduiden met V1, V2 en V3, en het grammaticaal verband tussen deze werkwoorden is dat V1 V2 als verbaal complement neemt, en V2 V3. In dit voorbeeld geldt dat V1=*wil*, V2=*gaan*, en V3=*doen.*

De kracht van GrETEL is nu dat we constructies van dit type kunnen vinden in een corpus op basis van deze voorbeeldzin (Query By Example, QBE). Het gehele proces omvat een aantal stadia die ook duidelijk in de GrETEL interface weergegeven zijn: *Example, Parse, Matrix, Treebanks, Results,* en *Analysis*. De gebruiker kan via die labels naar het betreffende stadium gaan, of stap voor stap door de stadia lopen via de *Next*- en *Previous*-knoppen. We zullen het voorbeeld aan de hand van deze stadia hier beschrijven.

Example. We voeren (5) in in de example-based search interface.

Parse. De zin wordt vervolgens automatisch ontleed door Alpino. Een gebruiker met voldoende kennis van de Alpino-structuren kan die inspecteren, en eventueel besluiten een ander voorbeeld te kiezen.

Matrix. Vervolgens kan de gebruiker aangeven welke aspecten van de voorbeeldzin relevant zijn voor de constructie waar de gebruiker in geïnteresseerd is. In een matrix die eruitziet zoals in Tabel 3 kan de gebruiker voor ieder woord aangeven of het relevant is, en op welke manier er gegeneraliseerd moet worden vanuit het voorbeeld. In het onderwerp *hij* en het lijdend voorwerp *dat* zijn we niet geïnteresseerd: zij mogen er zijn, maar hoeven dat niet. Daarom markeren we die als **Optional.** We zijn wel geinteresseerd in de werkwoorden zoals ze in deze grammaticale configuratie voorkomen. Niet zozeer in de specifieke vormen waarin ze voorkomen (**Word)**, of de lemma's van de specifieke werkwoorden in het voorbeeld (**Lemma**), maar in ieder woord van woordklasse werkwoord dat in deze grammaticale configuratie voor kan komen. Daarom markeren we deze woorden voor **Word Class.**

Tabel 3 GrETEL Selectiematrix

Sentence	*Hij*	*wil*	*dat*	*gaan*	*doen*
Word					
Lemma					
Word Class		X		X	X
Optional	X		X		

Ons voorbeeld is een hoofdzin, maar we zijn ook geïnteresseerd in voorbeelden van dergelijke werkwoorden in bijzinnen. Het onderscheid tussen hoofdzin en bijzin wordt aangeduid door de categorie van de direct dominerende knoop. Daarom markeren we nog de optie 'Ignore properties of the dominating node'.

Door deze keuzes te maken wordt er nu automatisch een zoekopdracht in XPath gevormd en getoond op het scherm, zowel de XPath code (6) als een grafische representatie van de zoekopdracht (Figuur 1).

(6) Xpath-zoekopdracht gevormd op basis van de geselecteerde opties:

```
//node[@cat and
        node[@pt="ww" and @rel="hd"] and
        node[@rel="vc" and @cat="inf" and
                node[@pt="ww" and @rel="hd"] and
                node[@rel="vc" and @cat="inf" and
                        node[@pt="ww" and @rel="hd"]]]]
```

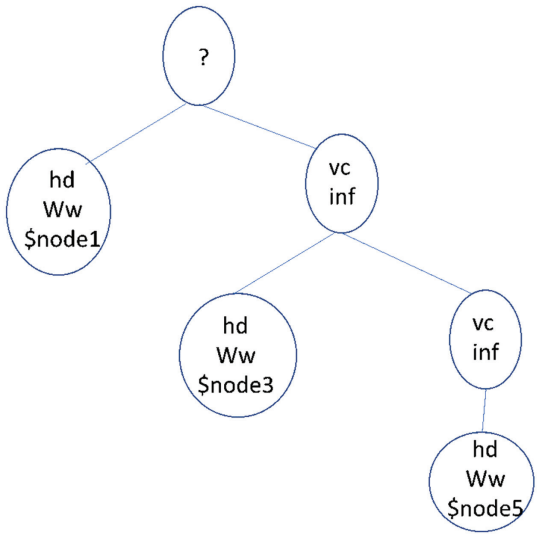
Deze complexe zoekopdracht is nu gecreëerd zonder iets van de zoekopdrachttaal Xpath te hoeven weten, en zonder ook maar iets te hoeven weten over de precieze syntactische structuren die in de parsebank of treebank voorkomen.

Merk op dat met deze zoekopdracht alleen gezocht wordt naar werkwoorden en het aangegeven grammaticaal verband tussen deze werkwoorden. De onderlinge volgorde van de werkwoorden is hierbij niet van belang,[45] en de zoekopdracht laat toe dat er allerlei ander materiaal in een uiting voorkomt dan gespecificeerd in de zoekopdracht (bijv. onderwerpen, complementen van werkwoorden, bepalingen, etc.).[46]

44 http://gretel.hum.uu.nl/gretel-upload/index.php/treebank

45 GrETEL biedt een optie om wel met de onderlinge volgorde rekening te houden, maar die gebruiken we hier niet.

46 De term *greedy* in de expansie van GrETEL slaat op dit aspect van GrETEL: de zoekopdracht levert een resultaat op als er naast de woorden en woordgroepen die gespecificeerd



Figuur 1 *Grafische weergave van de zoekopdracht. $node1, $node3, en $node5 zijn variabelen die later in de analyse gebruikt worden.*

Treebanks. Nu krijgt de gebruiker de mogelijkheid een of meerdere parse- of treebanks te selecteren waarin gezocht gaat worden met deze zoekopdracht. We selecteren *vklaura*. Deze parsebank bestaat uit vijf componenten. Iedere component bevat de syntactische structuren van de transcripties die een periode van één jaar bestrijken. Na deze selectie gedaan te hebben, klikken we op de Next-knop om in het *Results*-stadium te komen.

Results. Op deze pagina wordt de volledige Xpath-zoekopdracht getoond plus een tabel met het aantal hits en zinnen per component, en het totaal aantal hits en zinnen in deze parsebank. Hier vinden we 325 hits in 64.319 zinnen. Zie Tabel 4.

Daaronder worden de hits zelf getoond,[47] waar we hier een klein aantal van weergeven (Tabel 5). De hits worden weergegeven in vier kolommen: een kolom voor een volgnummer van de hit, een voor de naam van

zijn in de zoekopdracht nog meer woorden en woordgroepen optreden in de syntactische structuur.

47 Maximaal 500 per component.

Tabel 4 Aantal hits per component en in totaal

Name	Hits	All Sentences
vklaura	**325**	**64,319**
year1	12	3,746
year2	104	19,881
year3	76	17,071
year4	99	17,685
year5	34	5,936
	325	64,319

de XML-file die de syntactische structuur van de zin bevat, een voor de componentnaam, en een voor de hit zelf. De naam van de XML-file is ook een URL die leidt naar een grafische weergave van de syntactische structuur en de ermee geassocieerde metadata. De hit bestaat uit een weergave van de zin, met daarin vet gemarkeerd de delen van de zin die gedomineerd worden door de knoop die overeenkomt met de topknoop van de Xpath-zoekopdracht.

Tabel 5 Sample van de resultaten van de zoekopdracht

9	year5-59339.xml	year5	dis voor autoos, die mogen niet die een na die andere auto voorbij gaan, **daar moet je netjes achter mekaar blijven rijden**.
10	year5-59641.xml	year5	domme moeder, zegt zijn moeder, **je moet ook niet alleen gaan spelen**.

Inderdaad zien we aan de voorbeelden in Tabel 5 dat de hits constructies van het gewenste type bevatten in totaal andere zinnen dan de voorbeeldzin waar we mee startten, en met (gedeeltelijk) andere werkwoorden (*moet…blijven…rijden, moet…gaan…spelen*) dan in de voorbeeldzin.

GrETEL 4 biedt nu de mogelijkheid de resultaten verder te filteren per component en op basis van metadata, maar we laten dit hier achterwege. In plaats daarvan klikken we op de *Next*-knop om naar het *Analysis*-stadium te gaan.

Analysis. In het *Analysis*-stadium worden opnieuw de XPath-zoekopdracht en een grafische weergave ervan weergegeven. Iedere knoop in de zoekopdracht is geassocieerd met een variabele van de vorm $node*i* waar *i* een optioneel volgnummer is (de hier relevante variabelen waren al aangeduid in Figuur 1). Daaronder staat een draaitabel. De draaitabel bevat een vak met metadata-elementen van de treebank die gebruikt kunnen worden in de draaitabel. Verder kan iedere knoop uit de Xpath-zoekopdracht of

de grafische weergave ervan naar de draaitabel gesleept worden. Als men dat doet, kan men aangeven welke eigenschap van deze knoop men in de draaitabel opgenomen wil zien. Initieel zijn de eigenschappen *lemma* en *pt* (woordsoort) van *$node1* al opgenomen in de tabel. Op basis hiervan krijgen we een overzicht van de werkwoorden die in deze constructie als V1 optreden met hun frequenties, zie Tabel 6.

Tabel 6 Lemma en woordsoort van instantiaties van V1 en hun frequenties

$node1.pt	ww	Totals
$node1.lemma		
doen	1	1
gaan	30	30
hebben	28	28
komen	1	1
kunnen	43	43
laten	5	5
moeten	67	67
mogen	25	25
willen	13	13
zijn	8	8
zullen	104	104
Totals	325	325

De eigenschap *woordsoort* is hier niet zo interessant omdat deze eigenschap hier maar één waarde heeft. We kunnen deze eigenschap uit de tabel slepen en eigenschappen van andere knopen in de tabel zetten, bijv. de eigenschap *lemma* van *$node3* (=V2) en van *$node5* (=V3). Daarmee krijgen we een overzicht van alle drietallen werkwoorden en hun frequenties die in deze constructie in de treebank *vklaura* optreden (hier niet weergegeven). We kunnen ook eigenschappen van metadata opnemen in de tabel, bijv. de eigenschap *speaker*, en daar waardes uit filteren. Tabel 7 geeft een overzicht van alle werkwoordsdrietallen gebruikt door de spreker met code *LAU* in de treebank *vklaura*. Tabel 8 geeft dezelfde informatie maar nu naar de leeftijd van de spreker per maand.

Zo kan een gebruiker heel snel en op intuïtieve manier analyses maken van de resultaten van de treebankzoekopdracht. En voor bepaalde doelstellingen volstaat dit ook, bijv. als men enkele daadwerkelijk voorkomende voorbeelden wil zoeken van bepaalde constructies. Maar als men taalkundig onderzoek wil baseren op deze zoekopdrachten, dan is er meer nodig.

Tabel 7 Werkwoordcombinaties gebruikt door spreker LAU en hun frequenties

		speaker	LAU	Totals
$node1.lemma	**$node3.lemma**	**$node5.lemma**		
doen	doen	maken	1	1
gaan	laten	kopen	1	1
	zitten	op_schuiven	1	1
moeten	gaan	doen	1	1
		stoppen	1	1
	laten	drogen	1	1
mogen	blijven	staan	2	2
	gaan	spelen	1	1
willen	gaan	kleuren	1	1
zijn	blijven	staan	1	1
zullen	leren	fietsen	1	1
Totaal			12	12

Tabel 8 Werkwoordcombinaties gebruikt door spreker LAU en hun frequenties per maand

		speaker	LAU								
		months	**43**	**51**	**55**	**57**	**59**	**60**	**64**	**66**	
$node1. lemma	**$node3. lemma**	**$node5. lemma**									
doen	doen	maken				1					1
gaan	laten	kopen							1		1
	zitten	op_schuiven			1						1
moeten	gaan	doen			1						1
		stoppen						1			1
	laten	drogen		1							1
mogen	Blijven	staan								2	2
	gaan	spelen			1						1
willen	gaan	kleuren				1					1
zijn	blijven	staan	1								1
zullen	leren	fietsen					1				1
Totaal			1	1	3	2	1	1	1	2	12

6. De gevaren

Hoewel het mogelijk is om snel en op een intuïtieve manier zoekopdrachten te maken via de Query-By-Example faciliteiten, moet men zich terdege realiseren dat het feitelijke onderzoek nu pas begint. Zoals ik zal illustreren aan de hand van verschillende voorbeelden moet men wel degelijk de precieze syntactische structuren van de treebank of parsebank kennen, en

moet men ook weten hoe de QBE-methode precies generaliseert vanuit het voorbeeld.

De aard van de syntactische structuren is expliciet gedocumenteerd, bijv. voor CGN in Hoekstra et al. (2003) en voor LASSY in Van Noord et al. (2011). Dit zijn lijvige documenten (respectievelijk 77 en 208 pagina's), die desondanks incompleet zijn. Om de aard van de syntactische structuren te doorgronden is daarom vaak ook eigen exploratie vereist.

Voorbeeld 1. Stel dat we willen weten welke zelfstandige naamwoorden op kunnen treden als (het hoofd van) het lijdend voorwerp van het werkwoord *drinken*. In GrETEL kunnen we dat doen op basis van de voorbeeldzin *Hij drinkt water*. Maar we kunnen het ook doen op basis van de voorbeeldzin *Hij drinkt een biertje*. De resulterende XPath-expressies en links naar de zoekopdrachten in Lassy-Klein staan weergegeven in Tabel 9.

Tabel 9 Xpath-expressie en link naar de zoekopdracht in Lassy-Klein per voorbeeldzin

Voorbeeldzin	Link naar Zoekopdracht	Xpath-expressie
Hij drinkt water	Querylink	//node[@cat and node[@rel="hd" and @pt="ww" and @lemma="drinken"] and node[@rel="obj1" and @pt="n"]]
Hij drinkt een biertje	QueryLink	//node[@cat and node[@rel="hd" and @pt="ww" and @lemma="drinken"] and node[@rel="obj1" and @pt="n"]]

Maar de zoekopdrachten verschillen en ze leveren ook totaal andere resultaten op. Ze leveren ook een ander resultaat dan PaQu oplevert voor een vergelijkbare zoekopdracht (nl. lever de zinnen met woorden van woordsoort *n* die een afhankelijkheidsrelatie met relatielabel *obj1* onderhouden met een vorm van het werkwoord *drinken*).[48] Dit is te zien in Tabel 10.

Hoe komt het dat deze drie zoekopdrachten allemaal een verschillend resultaat opleveren? Dat komt door de aard van de syntactische structuren in de treebanks. Daarin komen drie verschillende gevallen voor: (1) een lijdend voorwerp dat uitsluitend uit een woord bestaat heeft geen NP-knoop boven zich; (2) een lijdend voorwerp dat uit meerdere woorden bestaat heeft wel een NP-knoop boven zich; (3) een lijdend voorwerp dat ook elders in de zin een functie vervult, kan bestaan uit een knoop met uitsluitend een index-attribuut. Dit is expliciet gedocumenteerd (Hoekstra et al. 2003: 5)

48 Met deze link naar de zoekopdracht (woord/postag=n, relatie=obj1, hoofdwoord=+drinken) uitgevoerd op Lassy-Klein: https://paqu.let.rug.nl:8068/?db=lassysmall&word=&rel=obj1&hword=%2Bdrinken&postag=n&hpostag=&meta=&sn=10

Tabel 10 Gevonden woorden als (hoofd van) het lijdend voorwerp bij *drinken* in Lassy-Klein op basis van twee verschillende voorbeeldzinnen en in PaQu

Voorbeeldzin	*Hij drinkt water*	*Hij drinkt een biertje*	PaQu
Rodenbach	1		1
alcohol	11	9	23
bier	1	1	6
water	1	1	4
druppel		2	2
glas		13	15
koffie		1	2
paar		1	1
standaardglas		2	2
Wijn		1	2
Frisdrank			2
Versie			2
Drank			1
Fee			1
Thee			1
Totaal	**14**	**31**	**65**

en uitvoerig beschreven in Odijk et al. (2017: 284-285). De aard van de syntactische structuren vormt natuurlijk een probleem voor alle corpusonderzoek dat men doet in deze treebanks. Maar door GrETELs QBE te gebruiken worden de verschillen in structuur niet of nauwelijks zichtbaar voor de onderzoeker. Het wordt dan ook op iedere cursus over GrETEL expliciet vermeld. Houdt men er geen rekening mee, dan krijgt men niet de resultaten die men verwacht. De conventie om geen constituentknoop aan te nemen boven een enkel woord is bewust gekozen door de treebankbouwers, maar is bij het zoeken bijzonder onpraktisch, omdat het heel veel zoekopdrachten compliceert (zie bijv. ook Van Eynde, Augustinus & Vandeghinste 2016: 107), en veel zoekopdrachten geformuleerd moeten worden als een combinatie van meerdere zoekopdrachten. Daarom zou het wenselijk zijn als zoekopdracht en analyseresultaten opgeslagen kunnen worden en door middel van set-achtige operaties (vereniging, doorsnijding, verschil) gemanipuleerd kunnen worden. Nog beter zou het zijn als de treebanks ook worden aangeboden in een versie met zo'n constituentknoop, en als de indexknopen ook aangeboden zouden worden als volledige knopen,[49] en PaQu biedt al een optie voor dit laatste.

49 Dan moet in GrETEL de ontleedboom die op basis van een voorbeeld door Alpino gegenereerd wordt natuurlijk ook overeenkomstig aangepast worden.

Voorbeeld 2. We nemen de zoekopdracht die we in sectie 5 gebruikt hebben maar passen deze nu toe op de parsebank *vksarah*, voor uitingen gedaan door de spreker met code *SAR*.[50] Dat doen we omdat we voor deze data vergelijkingsmateriaal hebben: Jacqueline van Kampen maakte een lijst beschikbaar van combinaties van drie 'kale' werkwoorden in de relevante grammaticale configuratie die zij handmatig gevonden had in dit bestand.

Het blijkt dat er nogal wat verschillen zijn tussen wat we vinden met GrETEL en wat in Van Kampens lijst staat. GrETEL vindt slechts 17 van de 28 voorbeelden die gevonden hadden moeten worden, en GrETEL vindt één voorbeeld dat niet gevonden had moeten worden.[51] Een deel hiervan is toe te schrijven aan het gebruik van QBE. Het gaat hierbij om de volgende gevallen:

- De meeste infinitieven treden niet alleen op, omdat Alpino een verborgen onderwerp bij infinitieven aanneemt met een woordgroep uit de hogere zin als antecedent. Als er echter geen antecedent is, wordt er ook geen verborgen onderwerp aangenomen, en kan de infinitief wel alleen optreden. Dit heeft tot gevolg dat één voorbeeld niet gevonden wordt (*moeten ook leren boksen*). Dit is eigenlijk hetzelfde probleem als beschreven onder voorbeeld 1 hierboven.
- We hebben in de matrix geselecteerd dat ieder woord van woordsoort 'werkwoord' toegelaten was (wat de inflectie-eigenschappen ook zijn), maar Alpino markeert het inflectioneel karakter van een werkwoord ook in de dominerende knoop: *inf* voor constituenten met een infinitief werkwoord als hoofd, *ppart* voor constituenten met een voltooid deelwoord als hoofd, etc. Hoewel we dus geabstraheerd hebben van inflectionele verschillen op het woord zelf, gebeurt dat niet voor de constituenten die dit woord bevatten, en daarom vinden we alleen resultaten met infinitieven, en niet met deelwoorden. Maar er is er wel degelijk een voorbeeld met een deelwoord, en dat vinden we niet met de QBE-methode: *dan gaat ie op de grond gegooid worden.*[52]

50 Met deze querylink, hier weergegeven als een `short URL': http://shorturl.at/oBM18.

51 Een test op combinaties van twee werkwoorden laat zien dat QBE voor dergelijke gevallen ook veel uitingen niet vindt: 1263/7908=16% in alle uitingen van vklaura, 166/1529=18.6% voor de uitingen door LAU in vklaura; 991/5327=18.6% in alle uitingen van vksarah, 318/1609=19.8% voor de uitingen door SAR in vksarah.

52 Dit is onderdeel van een veel grotere uiting.

- Er zijn goede redenen om de werkwoordscombinatie *laten zien* op twee verschillende manieren te analyseren.[53] Alpino doet dat ook. De ene analyse is zoals we zouden verwachten: *zien* als verbaal complement (*vc*) van *laten*. Maar voor de betekenis 'tonen' analyseert Alpino *zien* als een scheidbaar deel van het werkwoord *laten*. Wat men ook van deze analyse moge vinden, er is geen mogelijkheid in de QBE-methode om naar dergelijke structuren te generaliseren uitgaande van de gekozen voorbeeldzin. En dus mist GrETEL vier voorbeelden: *zal ik het es laten zien wat ik heb?*, *moet je alleen laten zien?, oh, mag je die laten zien?,* en *jij mag ze ook een keer laten zien, he*.

Zelfs als we deze voorbeelden mee zouden nemen, vinden we nog niet alles. Dat komt gedeeltelijk door foute annotatie in het CHAT-bestand. Zo herkent Alpino *za* in *za k het es laten zien?* niet als werkwoord, maar deze uiting had (volgens de CHILDES-conventies) geannoteerd moeten zijn als *za*(*l*) (*i*)*k het e*(*en*)*s laten zien?*, en dan zou Alpino deze uiting wel vinden. Een belangrijker oorzaak zijn uitingen die onwelgevormd zijn volgens Alpino en daarom van Alpino een analyse krijgen waardoor zij niet als deze constructie herkend worden. Dit is natuurlijk geen probleem dat specifiek is voor GrETEL maar algemener voor corpusgebaseerd onderzoek. Voorbeelden zijn:

(7) Onwelgevormde uitingen die niet gevonden worden:
  a. want ik moet moet de jas van van van haar he hebben zoeken.
  b. um, jij moet, jij moet jij moet helpen gaan.
  c. zullen we gewoon gaan zo inschuiven?
  d. <as tie> [//] <as as die> [//] as die wij [?] voor laat lopen gaan.

In (7a) lijkt *hebben* een infinitief als complement te nemen (of misschien is de transcriptie c.q. annotatie incorrect en moet *zoeken* als correctie van *hebben* geïnterpreteerd worden). In (7b) lijkt de meest voor de hand liggende interpretatie die met een voor Standaardnederlands ongebruikelijke volgorde V1 V3 V2 waarin V3 en V2 infinitieven zijn. Alpino analyseert echter *gaan* als complement van *helpen*, en de uiting wordt niet gevonden omdat *gaan* als enige complementwoord optreedt (vgl. voorbeeld 1). In (7c) treedt er materiaal op tussen de twee infinitieven, wat ook onwelgevormd

53 Bijv. in de 'tonen'-betekenis kan er een indirect object met voorzetsel *aan* optreden, wat beide werkwoorden apart niet toelaten.

is in Standaardnederlands (althans, zoals gesproken in Nederland). In (7d) is de zin onwelgevormd en wordt *laat* door Alpino als bijvoeglijk naamwoord geanalyseerd.

Voorbeeld 3. De NP van de vorm *een aantal N* kan wanneer het als onderwerp optreedt met een persoonsvorm in het enkelvoud of het meervoud optreden. Alpino neemt voor de twee gevallen andere structuren aan: *aantal* als hoofd en N als bepaling voor het enkelvoud, en N als hoofd en *een aantal* als determinator voor het meervoud. Maar in de Lassy-Klein en CGN-databanken komt het ook andersom voor! Dit zijn inconsistenties in de annotaties, die helaas bijna altijd in (manueel geannoteerde) data voor zullen komen. Dit is ook geobserveerd door Augustinus et al. (2014: 24, voetnoot 26).[54] Voor de vergelijkbare constructie *een paar* N komt dan echter weer alleen de structuur met N als hoofd voor, zowel met enkelvoudige als met meervoudige persoonsvormen. Als je geen rekening houdt met dergelijke inconsistenties, krijg je vertekende resultaten. In de zoekopdrachten in Tabel 11 wordt geabstraheerd van het getal van het werkwoord (attribuut *pvagr*).

Tabel 11 Zoekopdrachten gebaseerd op de *een aantal* N constructie

Zoekopdracht gebaseerd op *"een aantal mensen is gekomen"*	Zoekopdracht gebaseerd op *"een aantal mensen zijn gekomen"*
//node[@cat and node[@cat="np" and @rel="su" and node[@lemma="een" and @pt="lid" and @rel="det"] and node[@lemma="aantal" and @pt="n" and @rel="hd"] and node[@rel="mod" and @pt="n"]] and node[@pt="ww" and @rel="hd"]]	//node[@cat and node[@cat="np" and @rel="su" and node[@rel="det" and @cat="np" and node[@rel="det" and @pt="lid" and @lemma="een"] and node[@lemma="aantal" and @pt="n" and @rel="hd"]] and node[@pt="n" and @rel="hd"]] and node[@rel="hd" and @pt="ww"]]

Alle mogelijke waardes en frequenties van het attribuut *pvagr* kunnen dan in de analysecomponent van GrETEL 4 zichtbaar gemaakt worden. De

54 De observatie en de behandeling ervan gaat in dit werk nogal impliciet: de zoekopdracht die op basis van een voorbeeld gegenereerd wordt 'is modified in order to become a more general version' (Augustinus et al. 2014: 24). In dit educatieve pakket wordt hier toch een gelegenheid gemist om de beperkingen van de QBE-methode aan de hand van een concreet voorbeeld duidelijk te maken.

resultaten hiervan staan samengevat in Tabel 12. Als er geen rekening gehouden zou worden met de inconsistenties zou de onderzoeker:

- 4 mv-resultaten missen in Lassy-Klein voor de zoekopdracht gebaseerd op voorbeeld *een aantal mensen is gekomen*;
- 10 ev-resultaten missen in CGN voor de zoekopdracht gebaseerd op voorbeeld *een aantal mensen zijn gekomen*;
- 1 ev-resultaat missen in CGN en in Lassy voor de zoekopdracht gebaseerd op voorbeeld *een paar mensen zijn gekomen*.

Daarnaast zijn er nog *none*-resultaten, die optreden als het werkwoord geen persoonsvorm is (in geval van foute part-of-speech-tagging of bij bepaalde knopen die een index-attribuut hebben).[55]

Tabel 12 Resultaten voor de zoekopdrachten gebaseerd op de voorbeeldzinnen *een aantal / paar mensen is / zijn gekomen* in CGN en Lassy-Klein

	voorbeeld	*een aantal mensen*		*een paar mensen*	
voorbeeld	**pvagr**	**CGN**	**Lassy**	**CGN**	**Lassy**
is gekomen	ev	5	27	0	0
	mv	0	4	0	0
	none	0	1	0	0
zijn gekomen	ev	10	0	1	1
	mv	35	25	27	14
	none	1	0	1	0

Voorbeeld 4. Er is bij onderzoek aan de hand van een rijk automatisch of semiautomatisch geannoteerde databank altijd het gevaar van *circulariteit*. Van Eynde (2009) onderzoekt welke werkwoorden een predicatief complement kunnen nemen op basis van de CGN-treebank. Deze treebank is geheel manueel aangemaakt, en daarom is er bij dit onderzoek geen sprake van circulariteit. Maar als men nu hetzelfde zou onderzoeken in een parsebank, dan zou dat zeer waarschijnlijk circulair zijn, omdat de Alpino-grammatica alleen predicatieve complementen toelaat bij werkwoorden die ervoor gemarkeerd zijn om een predicatief complement toe te laten. Men moet dus, om circulariteit te vermijden, weten hoe de Alpino-grammatica met de te onderzoeken verschijnselen omgaat.

55 Dit laatste zou voorkomen kunnen worden als GrETEL ook indexexpansie toelaat, zoals nu al het geval is in PaQu.

7. Het bezweren van de gevaren

Voor de *vksarah* data set hadden we de beschikking over een onafhankelijk geannoteerde dataset (die we als gouden standaard gebruikten) zodat we konden bepalen hoe betrouwbaar de analyse op basis van GrETEL was. Maar in de meeste gevallen is zo'n onafhankelijk geannoteerde dataset natuurlijk niet voorhanden. Hoe kunnen we dan toch bepalen hoe betrouwbaar de resultaten van een zoekopdracht of analyse in GrETEL zijn?

Er zijn allerlei methodes om dit te doen. Ik zal er hier een paar schetsen.[56] Het is van groot belang dat gebruikers van de applicaties deze methodes ook onderwezen krijgen en dat de applicaties deze methodes ook ondersteunen.

Men zou de data manueel kunnen controleren. Dat is echter in veel gevallen niet mogelijk omdat de hoeveelheid data te groot is. Een haalbaarder methode is om een selectie van een superset van de data te maken die relevant is voor het onderzochte verschijnsel. Zo is het voor de analyse van verbale complementatie nuttig te onderzoeken hoe werkwoorden in het algemeen in de treebank of parsebank met elkaar gecombineerd worden. Dat blijkt nog een behoorlijke variatieop te leveren, zie Tabel 13.[57] We hebben dit onderzocht in de CHILDES parsebank in PaQu. De meest voorkomende combinatie is die waarin het ene werkwoord optreedt als verbaal complement (*vc*) van het andere. De tweede meest voorkomende combinatie is die waarin de twee werkwoorden conjuncten zijn in een coördinatiestructuur. Maar we zien ook dat een werkwoord op kan treden als onderwerp, als lijdend voorwerp, als predicaat, als bepaling en als nog veel meer. In Tabel 13 vinden we ook dat een werkwoord soms als *svp* op kan treden bij een ander werkwoord (343 keer in de CHILDES parsebank in de PaQu- applicatie), een grammaticale relatie die vooral voor scheidbare werkwoordspartikels gebruikt wordt.

Een nog grotere superset voor de zoekopdracht met drie kale werkwoorden is de set van alle uitingen met drie werkwoorden waarvan er minimaal twee geen persoonsvorm zijn, los van welk grammaticaal verband ze

56 Zie Bloem (2016) en Bloem (dit nummer) voor (gedeeltelijk andere) algemene strategieën om automatisch gegenereerde ontledingen te evalueren zonder gebruik te maken van een gouden standaard.

57 Wat niet blijkt uit Tabel 13 is dat de combinatiemogelijkheden in hoge mate bepaald worden door de precieze vorm van het werkwoord (d.w.z., is het een infinitief, een deelwoord, of een persoonsvorm). Veel van de gevallen betreffen verder gesubstantiviseerd gebruik van infinitieven.

Tabel 13 Combinatiemogelijkheden van twee werkwoorden in de Childes parsebank in PaQu. Postag = woordsoort van het afhankelijke werkwoord (V2), hpostag = woordsoort van V1, rel= label van de afhankelijkheidsrelatie tussen deze twee woorden

aantal	postag	rel	hpostag
51097	ww	vc	ww
5392	ww	cnj/cnj	ww
917	ww	obj1	ww
865	ww	su	ww
772	ww	predc	ww
610	ww	mod	ww
343	ww	svp	ww
53	ww	predm	ww
34	ww	app	ww
31	ww	obj2	ww
20	ww	ld	ww
3	ww	body/whd	ww
2	ww	obj1/su	ww
2	ww	su/obj1	ww
1	ww	body/rhd	ww

onderling ook hebben. De applicaties GrETEL en PaQU zouden het maken van dergelijke zoekopdrachten moeten ondersteunen. De relevante zoekopdracht voor drie werkwoorden lijkt op het eerste gezicht eenvoudig, maar een naïeve versie leidt tot heel veel overbodige resultaten, namelijk van knopen die een knoop bevatten die al aan de voorwaarde voldoet. We willen daarom de minimale knoop die aan alle voorwaarden voldoet. De techniek hiervoor zoals geschetst in het DACT CookBook[58] werkt hier niet. Een zoekopdracht die deze techniek gebruikt (8) vindt weliswaar alle uitingen maar niet per se alle matches:

(8) //node[count(.//node[@pt="ww"])>=3 and count(.//node[@wvorm!="pv"])>=2 and not(node[count(.//node[@pt="ww"])>=3 and count(.//node[@wvorm!="pv"])>=2])]

58 http://rug-compling.github.io/dact/cookbook/#minimal-dominating-node. Een correcte versie kan hier http://rug-compling.github.io/dact/cookbook/#example-4-minimal-dominating-node-revisited gevonden worden.

De correcte zoekopdracht vereist het gebruik van variabelen in de zoekopdracht, wat momenteel ondersteund wordt door PaQu maar niet door GrETEL:[59]

(9) //node[(some $x in .//node[@pt="ww"], $y in .//node[@pt="ww" and @wvorm!="pv"], $z in .//node[@pt="ww" and @wvorm!="pv"] satisfies ($x/number(@begin) < $y/number(@begin) and $y/number(@begin) < $z/number(@begin) and not(node[.//node = $z and .//node=$x and .//node=$y])))]

Deze zoekopdracht levert, toegepast op de parsebank *vklaura*, 815 resultaten op, waarvan er 83 door de spreker LAU uitgesproken worden. Analyse van deze uitingen van spreker LAU levert het resultaat zoals weergegeven in Tabel 14.

Tabel 14 Analyse van de uitingen van LAU met drie werkwoorden

Utts	LAU	Aantal
relevant		18
	QBE	12
	laten_zien	1
	vc/ww	1
	gemist	4
niet relevant		65

In 65 uitingen is er sprake van een geheel ander grammaticaal verband dan waar we in geïnteresseerd zijn (*niet relevant*). Van de 18 relevante uitingen hadden we er 12 met de QBE-methode gevonden, is er één geval van *laten zien* met een *svp*-analyse, is er één voorbeeld van een verbaal complement dat uitsluitend uit het werkwoord bestaat (*vc/ww*), en hebben we vier additionele uitingen gemist, die weergegeven zijn in (10).

(10) Additionele door de QBE-methode gemiste uitingen in *vklaura*:
a. **kun** je ook **meehelpen** so **doen**?
b. hee, **zouen puzzelen doen**?[60]
c. wat **zou** ze xxx, ze nou weer **gekregen hebben**?
d. **zou** eens **moeten** bij **doen**.

59 Dank aan Gertjan van Noord voor zijn assistentie hierbij.

60 Voor deze zin kan men betwijfelen of die gevonden zou moeten worden. Immers, een analyse met *puzzelen* als gesubstantiveerde infinitief en als lijdend voorwerp bij *doen* lijkt correct hier, en onder deze analyse is de zin niet relevant. Ik heb de zin hier toch genoemd omdat de analyse van Alpino voor deze zin echt volledig verkeerd is.

In totaal worden er dus 6 van de 18 uitingen gemist (33%). Enkele van die uitingen zijn potentieel zeer interessant. Zo lijken er in (10a) en in (10d) elementen tussen de werkwoorden op te treden, wat in de taal van de volwassenen alleen in Vlaanderen voorkomt. De voorbeelden van *vklaura* en *vksarah* laten zien dat de QBE-analyse een redelijk groot aantal relevante uitingen niet vindt, waardoor nadere analyse vereist is. De hier toegepaste techniek (een zo klein mogelijke superset manueel analyseren) reduceert de foutenmarge nu tot de fouten die gemaakt worden in pos-tagging. Dezelfde techniek kan ook toegepast worden om een superset voor de *een aantal N*-constructie te definiëren. In de meeste gevallen is het aantal te controleren uitingen beperkt (een paar honderd) en dus goed manueel te verifiëren. In alle gevallen kan bij grote sets die gecontroleerd zouden moeten worden, de manuele controle op een aselecte steekproef uitgevoerd worden om een indruk te krijgen van de kwaliteit van de analyses. Maar daarvoor dienen de applicaties het nemen van zo'n steekproef wel te ondersteunen. PaQu doet dat (gedeeltelijk), GrETEL in het geheel nog niet. Men zou ook iedere resultaatverzameling of steekproef daaruit manueel willen kunnen annoteren: dat draagt bij aan het repliceerbaar maken van de uitgevoerde test en zou ook gebruikt kunnen worden om de dataset te verrijken.

Circulariteit moet vermeden worden. In mijn eigen werk heb ik dat vermeden door volledige manuele controle van de meeste relevante data en op basis van de controle van een steekproef voor een relevante subset die te groot was om volledig manueel te controleren. Maar soms kan het feit dat de Alpino-ontleder foute analyses maakt hierbij ook helpen. Zo kwam ik voorbeelden tegen waarin het woord *heel*, volgens de ontleding van Alpino, werkwoorden modificeert. Dat waren in de meeste gevallen foute analyses, maar het toont aan dat Alpino niet 'ingebakken' heeft dat *heel* uitsluitend bijvoeglijke naamwoorden kan modificeren, en daarmee wordt circulariteit vermeden. Een soortgelijk geval ontstond met het koppelwerkwoord *raken*: er was een zin in de dataset waarin het koppelwerkwoord *raken* een NP als predicatief complement neemt (zie Odijk et al. 2017: 291). Deze ontleding was fout, maar het toont opnieuw aan dat Alpino niet 'ingebakken' heeft dat het koppelwerkwoord geen NP's als predicatief complement kan nemen. Hier is het adagium van een beroemd Amsterdams filosoof (*elk nadeel heb zijn voordeel*) volledig van kracht.

8. Conclusies

De beschikbaarheid en de toegankelijkheid van grote corpora is in de laatste jaren enorm toegenomen. Corpora vormen een nuttige en belangrijke bron van bewijsmateriaal voor taalkundig onderzoek, maar zij vormen niet het enige bewijsmateriaal voor taalkundig onderzoek, hebben geen speciale status als zodanig, en hebben hun beperkingen. Nieuwe zeer gebruikersvriendelijke applicaties zoals GrETEL maken het heel gemakkelijk grote en rijk geannoteerde corpora te doorzoeken op basis van een voorbeeldzin en zonder kennis van een zoekopdrachttaal of de precieze aard van de taalkundige annotaties. Het is daarom heel verleidelijk die intensief te gebruiken. Dat is goed, maar er schuilen ook grote gevaren in, want in de praktijk moet een onderzoeker wel degelijk weten wat de aard van de taalkundige annotaties is en hoe gebruikersvriendelijke interfaces een zoekopdracht op basis van een voorbeeld genereren. Er zijn methodes om deze gevaren te vermijden of te reduceren, waar ik er hier een aantal van beschreven heb. Die moeten ook aan gebruikers van de applicaties onderwezen worden en de applicaties zouden deze methodes ook zo gebruikersvriendelijk mogelijk moeten ondersteunen. Ik som de gewenste uitbreidingen aan de functionaliteit van deze applicaties hier nog eens op:

- Versies van de treebanks en parsebanks beschikbaar maken met expansie van eenwoordsfrases en indexknopen (sectie 6)
- Opslaan van zoekopdracht en analyseresultaten (sectie 6)
- Set-achtige operaties op (opgeslagen) zoekopdracht en analyseresultaten (sectie 6)
- Selectie van een steekproef van de resultaten (sectie 7)
- Annoteren van (een steekproef van de) zoekopdrachtresultaten (sectie 7)
- Gebruik van nieuwe annotaties in de analysecomponent (sectie 7)
- Ondersteuning van het gebruik van variabelen in Xpath-zoekopdrachten (sectie 7)
- Ondersteuning voor minimale supersetexploratie (sectie 7)

Door dergelijke functionaliteit toe te voegen en het gebruik ervan te onderwijzen aan gebruikers kunnen de gebruikersvriendelijke applicaties om corpora te doorzoeken en te analyseren ook op een 'veilige' manier gebruikt worden, dat wil zeggen, op zo'n manier dat zij betrouwbare resultaten opleveren.

Referenties

Augustinus, Liesbeth, Schuurman, Ineke, Vandeghinste, Vincent & Van Eynde, Frank (2014). *GrETEL. Searching for breadcrumbs in texts* (*CLARIN Educational Module*). Centre for Computational Linguistics KU Leuven. <http://dev.clarin.nl/sites/default/files/EducationalModule-v4b.pdf>

Augustinus, Liesbeth (2015). *Complement raising and cluster formation in Dutch. A treebank-supported investigation*. Doctoraal proefschrift KU Leuven.

Bloem, Jelke (2016). Evaluating automatically annotated treebanks for linguistic research. In: P. Bański, M. Kupietz, H. Lüngen, A. Witt, A. Barbaresi, H. Biber, E. Breiteneder & S. Clematide (red.), *Proceedings of the 4th Workshop on Challenges in the Management of Large Corpora* (*CMLC 4*). Paris: ELRA, 8-14. <http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-CMLC Proceedings.pdf>

Bouma, Gosse, J.M. van Koppen, Frank Landsbergen, Jan Odijk, Ton van der Wouden & Matje van de Camp (2015). Enriching a descriptive grammar with treebank queries. *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories* 14 (*TLT14*), 13-25.

Edelman, Shimon & Morten H. Christiansen (2003) How seriously should we take Minimalist syntax? *Trends in Cognitive Science* 7, 60-61.

Featherston, Sam. (2009). Relax, lean back, and be a linguist. *Zeitschrift für Sprachwissenschaft* 28(1), 127-132.

Gibson, Edward & Evelina Fedorenko. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Science* 14, 233-234.

Gibson, Edward, Steven T. Piantadosi & Evelina Fedorenko (2013). Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2013). *Language and Cognitive Processes* 28(3), 229-240.

Charles J. Fillmore (1992). 'Corpus linguistics' or 'computer-aided armchair linguistics'. In: Jan Svartvik (red.), *Directions in Corpus Linguistics*. Berlin/New York: Mouton de Gruyter, 35-60.

Hoekstra, Heleen, Michael Moortgat, Bram Renmans, Machteld Schouppe, Ineke Schuurman & Ton van der Wouden (2003). *CGN syntactische annotatie*. CGN report Utrecht University.

Kampen, Jacqueline van (2009). The non-biological evolution of grammar: Wh-question formation in Germanic. *Biolinguistics* 3(2-3), 154-185.

Langendoen, D. Terence (1973). The problem of linguistic theory in relation to language behavior: A tribute and reply to Paul Goodman. *Daedalus* 102(3), 195-201.

Linzen, Tal & Yohei Oseki (2018). The reliability of acceptability judgments across languages. *Glossa: a journal of general linguistics* 3(1), 1-25.

Noord, Gertjan van, Ineke Schuurman & Gosse Bouma (2011). *Lassy syntactische annotatie* (revision 19455). Lassy report, RU Groningen. <https://www.let.rug.nl/vannoord/Lassy/sa-man_lassy.pdf>

Odijk, Jan (2015). Linguistic research with PaQU. *Computational Linguistics in The Netherlands Journal* 5, 3-14.

Odijk, Jan (2016a). Linguistic research using CLARIN. *Lingua* 178, 1-4.

Odijk, Jan (red.) (2016b). Linguistic research in the CLARIN infrastructure. *Lingua* 178.

Odijk, Jan (2016c). A use case for linguistic research on Dutch with CLARIN. In: K. De Smedt (red.), *Selected Papers from the CLARIN Annual Conference 2015* (Vol. 123). Linköping: Linköping University Electronic Press, 45-61.

Odijk, Jan (2018). Boosting linguistic research with CLARIN. Lezing op *ESSLLI 2018*, Sofia (Bulgarije), 14 augustus 2018.

Odijk, Jan, Alexis Dimitriadis, Martijn van der Klis, Marjo van Koppen, Meie Otten & Remco van de Veen (2018a). The AnnCor CHILDES Treebank. In: *Proceedings of the Eleventh*

International Conference on Language Resources and Evaluation (*LREC 2018*), *Miyazaki, Japan*. Paris: European Language Resources Association (ELRA), 2275-2283.

Odijk, Jan & A. van Hessen (red.) (2017). CLARIN *in the Low Countries*. London: Ubiquity Press.

Odijk, Jan, M. van der Klis & S. Spoel (2018b). Extensions to the GrETEL Treebank Query Application. In: E. Bejcek (red.), *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories* (*TLT16*). Praag: Charles University, 46-55.

Odijk, Jan, Gertjan van Noord, P. Kleiweg & Erik Tjong Kim Sang (2017). The Parse and Query (PaQu) application. In: J. Odijk, & A. van Hessen (red.), *CLARIN in the Low Countries*. London, UK: Ubiquity Press, 281-297.

Oostdijk, Nelleke, Martin Reynaert, Veronique Hoste & Ineke Schuurman (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In: P. Spyns & J. Odijk (red.), *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*.

Phillips, C. (2010). Should we impeach armchair linguists? *Japanese/Korean Linguistics* 17, 49-64.

Phillips, C. & H. Lasnik (2003). Linguistics and empirical evidence: Reply to Edelman and Christiansen. *Trends in Cognitive Sciences* 7(2), 61-62.

Pullum, Geoffrey (2017). Theory, data, and the epistemology of syntax. In: M. Konopka & A. Wöllstein (red.), *Grammatische Variation: Empirische Zugänge und theoretische Modellierung*. Berlin/New York: Mouton de Gruyter, 283-298.

Schütze, Carson T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*, Chicago: The University of Chicago Press.

Schütze, Carson T. (2016). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Berlin: Language Science Press.

Spyns, Peter & Jan Odijk (red.) (2013). *Essential speech and language technology for Dutch. Results by the STEVIN-programme*. Berlin/Heidelberg: Springer.

Van Eynde, Frank (2009). A treebank-driven investigation of predicative complements in Dutch. An efficient, practical, actually usable approach. *Computational Linguistics in the Netherlands 2009 - Selected Papers from the 19th CLIN Meeting, CLIN 2009*, 131-145.

Van Eynde, Frank, Liesbeth Augustinus & Vincent Vandeghinste (2016). Number agreement in copular constructions: A treebank-based investigation. *Lingua* 178, 104-126.

Wasow, Thomas and Jennifer E. Arnold, (2005). Intuitions in linguistic argumentation. *Lingua* 115, 1481-1496.

Wouden, Ton van der et al. (2016a). Het Taalportaal. Een nieuwe wetenschappelijke grammatica voor het Nederlands en het Fries (en het Afrikaans). *Nederlandse Taalkunde* 21(1), 157-168.

Wouden, Ton van der, Gosse Bouma, Matje van de Camp, Marjo van Koppen, Frank Landsbergen & Jan Odijk (2016b). Enriching a grammatical database with intelligent links to linguistic resources. In: K. De Smedt (red.), *Selected Papers from the CLARIN Annual Conference 2015*. Linköping: Linköping University Electronic Press, 108-117.

Wouden, Ton van der, Gosse Bouma, Matje van de Camp, Marjo van Koppen, Frank Landsbergen & Jan Odijk (2017). Enriching a scientific grammar with links to linguistic resources: The Taalportaal. In: J. Odijk & A. van Hessen (red.), *CLARIN in the Low Countries*. London: Ubiquity Press, pp. 299-310.

Over de auteur

Jan Odijk, Universiteit Utrecht
E-mail: j.odijk@uu.nl