TNTL 123 (2007) 83-96

Roeland van Hout en Ans van Kemenade

De kwantitatieve diepgang van taalvariatie-onderzoek

Abstract — The study of linguistic variation has shifted towards quantitative analysis over the last decades, as can be seen in both sociolinguistics and historical linguistics. What is the added value of a quantitative analysis? Its value is shown on the basis of some concrete research examples where a so-called linguistic variable is analysed by including internal (linguistic) and external (non-linguistic) variables or factors. The basic statistical technique is chi square analysis. The more advanced applications involve logistic regression analysis. The increasing availability of larger language corpora enhances the quantitative development and implies a further extension by investigating complex syntactic variation patterns.

1 Inleiding

Onderzoek naar taalvariatie heeft in de loop van de laatste veertig jaar ingrijpende veranderingen ondergaan, zoals te zien is aan de ontwikkelingen in de sociolinguïstiek en de historische taalkunde. Er is een intensievere uitwisseling en samenwerking ontstaan met theoretisch gerichte taalkundige analyses. Daarnaast vond de kwantitatieve analyse steeds meer ingang in het onderzoek. Dat ging aanvankelijk schoorvoetend, zeker in de historische taalkunde, maar in het hedendaagse onderzoek zijn kwantitatieve analyses bijna gemeengoed geworden. Wat is nu de meerwaarde van de kwantitatieve benadering? Dat is het beste te illustreren aan de hand van enkele voorbeelden waarin duidelijk wordt gemaakt wat voor soort van vragen vanuit een kwantitatief perspectief beantwoord kunnen worden.

Aanvankelijk richtte de kwantitatieve analyse zich op de sociale en fonologische variatiepatronen, maar meer recent, met het beschikbaar komen van grote(re) taalcorpora, staan ook, of misschien beter gezegd juist ook syntactische en lexicale variatiepatronen in de schijnwerpers. Meer dan ooit is er een wisselwerking mogelijk tussen data en theorie, omdat er meer mogelijkheden zijn om taalkundige hypotheses te toetsen aan empirische gegevens. Dat geldt overigens ook voor dialectgeografisch onderzoek, maar daar zullen we in deze bijdrage verder geen aandacht aan besteden.

Waar begint het kwantitatieve onderzoek naar sociale en talige variatie? Dat blijft moeilijk vast te stellen, maar een bekend onderzoek is Fischer (1958), waarin een aantal sociale factoren onderzocht werd wat betreft de keuze tussen de /ng/ variant en /n/ variant in het Engelse tegenwoordige deelwoord, bijvoorbeeld *walking* vs. *walkin* of *running* vs. *runnin*. Het was evenwel het werk van Labov dat de kwantitatieve omslag in het taalvariatie-onderzoek definitief heeft bewerkstelligd. De moderne kwantitatieve benadering vindt zijn oorsprong in het onderzoek van Labov op Martha's Vineyard, een toeristisch eilandje voor de kust van Massachusetts, en New York City. Labov slaagde erin om met getallen om te gaan, op een systematische en doortastende manier. Hij wist taalvariatie te verbinden met ver-

schillende mogelijke verklaringsbronnen, zowel van sociale als linguïstische aard, of, anders geformuleerd, van interne en externe aard. De theoretische onderbouwing voor de analyse van kwantitatieve variatiepatronen is reeds te vinden in Labov, Weinreich & Herzog (1964), een uiteenzetting over fundamentele kwesties in taalvariatie- en taalveranderingsonderzoek die nog steeds niets van haar actualiteitswaarde heeft verloren.

Eén van de bekendste voorbeelden van de kwantitatieve relatie tussen een variabel taalverschijnsel en een sociale factor is het warenhuisonderzoek van Labov in New York City, waarbij de verschillen tussen de warenhuizen gelijk staan met een onderverdeling naar de sociale klasse van het winkelend publiek. Een hogere sociale status 'betekende' het vaker uitspreken van de postvocale (r) in de beide woorden *fourth floor*. Elk warenhuis had een vierde verdieping en Labov vroeg waar hij een bepaald artikel kon vinden dat juist op die vierde verdieping verkocht werd. De resultaten zijn door hem overigens nooit diepgaander statistisch geanalyseerd. Dat wordt wel gedaan in Paolillo (2002) en we willen in deze bijdrage illustreren hoe statistische rekenmodellen de bronnen van taalvariatie kunnen onderscheiden en wegen. Vervolgens komt de historische taalkunde aan bod, waar we vooral ook aandacht besteden aan de analyses van Kroch (1989), omdat hij de tijdsfactor op een specifieke wijze wil modelleren. Daarna komt een voorbeeld van eigen onderzoek aan de orde. De voorbeelden maken duidelijk hoe dicht de sociolinguïstiek en de historische taalkunde tegen elkaar aanzitten en zelfs met gemak onder de paraplu van variatielinguïstiek geplaatst kunnen worden. In paragraaf 4 gaan we in op recente ontwikkelingen en wensen en mogelijkheden voor toekomstig onderzoek.

2 De warenhuizen: sociolinguïstisch onderzoek

Chambers (2002) geeft een inleidend en breed overzicht over de stand van zaken in het sociolinguïstisch variatie-onderzoek. Zijn boek, maar dat geldt ook voor andere inleidende boeken in de variationistische ('variational') sociolinguïstiek, maakt duidelijk hoe frequent kwantitatieve gegevens en grafieken deel uitmaken van de inhoudelijke argumentatie. Praktisch alle verbanden en oorzakelijke verklaringen berusten op graduele taalvariatiepatronen. Daarentegen is er eigenlijk helemaal geen statistiek in Chambers' boek te vinden. Dat is alleen te begrijpen tegen de achtergrond van taalkundig geschoolde sociolinguïsten die nog steeds niet verzot zijn op het uitwerken van complexere statistische technieken.

In het onderzoek zelf speelt de statistiek wel degelijk een rol van betekenis. De sociolinguïsten hebben zelfs een eigen statistisch analyseprogramma, VARBRUL/GOLDVARB, dat met name ontwikkeld is door David Sankoff (zie Sankoff & Labov 1979). Wat kan deze statistische techniek?

De techniek veronderstelt een afhankelijke variabele, dat wil zeggen een taalvariabele, met twee mogelijke verschijningsvormen. Labov onderzocht de realisatie van de postvocale (r) in New York City. De (r) kent twee mogelijke verschijningsvormen of uitkomsten: realisatie of geen realisatie. Deze tweewaardige ofwel binomiale variabele kan gekoppeld worden aan mogelijke verklarende variabelen. Labov keek in zijn warenhuisonderzoek naar het verschil tussen de drie onderzochte

warenhuizen, Saks, Macy's en Klein, het verschil tussen de twee onderzochte woorden (*fourth* en *floor*) en ten slotte naar het verschil tussen een nadrukkelijke en niet-nadrukkelijke context. De gegevens in de nadrukkelijke context verkreeg hij door na het antwoord van de informant nog eens te vragen: *wat zei u*? Dat leverde dan de nadrukkelijke realisatie op. Deze drie variabelen met de bijbehorende frequenties van de uitspraak van de postvocalische (r) zijn te vinden in tabel 1.

Tabel 1 De realisatie van de postvocale (r) uitgesplitst naar de variabelen warenhuis, woord en nadruk; de gegevens betreffen de gevonden frequenties voor niet-realisatie (= 0) en realisatie (= 1) van de (r).

		Saks		Macy's		Klein's		totaal		
		0	1	0	1	0	1	0	1	
fourth	– nadruk	39	17	81	33	63	3	183	53	77.5%
	+ nadruk	24	16	48	13	40	6	112	35	76.2%
floor	– nadruk	18	31	62	48	59	5	139	84	62.3%
	+ nadruk	12	21	20	31	33	7	65	59	52.4%
totaal		93	85	211	125	195	21	499	231	
		52.2%		62.8%		90.3%		68.4%		

Tabel 1 laat aan de hand van de totalen zien dat de verschillen tussen de warenhuizen groot zijn. Waar in Saks het aandeel van de niet-realisaties 52.2% bedraagt heeft Klein's een percentage van 90.3%. Saks heeft de hoogste sociale status en dat houdt in dat de postvocalische (r) veel vaker gerealiseerd wordt. De percentages aan de rechterkant van de tabel laten minder grote verschillen zien voor de variabelen woord en nadruk. Woord levert een verschil op, waarbij het slotwoord *floor* meer realisaties oplevert. Nadruk lijkt, vooral in het woord *fourth*, geen tot nauwelijks een effect te hebben.

Uiteraard kan voor elk van de drie verklarende variabelen apart worden nagegaan of er een significante samenhang is met de afhankelijke variabele. Dat kan met behulp van de zogeheten chi-kwadraattoets, waarvoor dan telkens per verklarende variabele een contingentietabel ofwel kruistabel wordt gemaakt. In een dergelijke kruistabel wordt een taalvariabele gekruist met een verklarende variabele. Tabel 1 is een kruistabel waarin de taalvariabele (r) gekruist wordt met drie verklarende variabelen tegelijkertijd. De chi-kwadraattoets per verklarende variabele levert dan op dat er een significant verband is voor warenhuis en woord en niet voor nadruk. Een echte stap voorwaarts is echter om de verzameling van variabelen gezamenlijk te analyseren, vooral ook omdat de verklarende variabelen wellicht onderling weer samenhangen en elkaar beïnvloeden. Een ander zeer wezenlijk argument voor een overall-analyse is dat daarmee de interne variabelen (woord, nadruk) en de externe variabelen (in dit geval het warenhuis) samen worden gewogen. Hoe zit het complete beeld eruit?

Die complete analyse kan geleverd door VARBRUL (de naamgeving is afgeleid van 'variable rule' oftewel variabele regel) of z'n latere versie GOLDVARB (zoals de naam aangeeft werden de mogelijkheden nog aantrekkelijker). Dit statistisch pakket is uitermate populair geworden in de analyse van taalvariatie en het is zelfs ook

toegepast in (tweede-)taalverwervingsonderzoek en gebarentaalonderzoek. Technisch gezien gaat het om een binaire logistische regressie-analyse: er is een tweewaardige afhankelijke variabele en er is een verzameling van mogelijke verklarende variabelen. Wat gemakkelijker gezegd: de afhankelijke variabele kan worden uitgedrukt in een percentage en we willen verklaren welke onafhankelijke variabelen oftewel factoren een verklaring kunnen bieden voor de hoogte van het percentage. Om wiskundige redenen wordt niet het percentage als afhankelijke variabele gebruikt, maar het logaritme van de zogeheten odds: de kans dat iets voorkomt of gebeurt gedeeld door de kans dat het niet voorkomt of gebeurt. Als gedacht wordt dat de kans dat Nederland wereldkampioen voetval wordt 5% is, ofwel 1 op de 20, dan is dat in termen van odds 1 tegen 19 (p tegen 1-p = .05/.95 = 1/19; hierbij staat 'p'voor probabiliteit, dat wil zeggen waarschijnlijkheid van voorkomen, in dit geval de waarschijnlijkheid van het kampioenschap).

Dit type analyse is ook beschikbaar in algemene statistische softwarepakketten die meer aangepast zijn aan de standaardmanier van rapporteren van de uitkomsten. In Tabel 2 is een deel van de uitkomsten vervat van een dergelijke analyse met het pakket SPSS.

Tabel 2 Classificatietabel die aangeeft welk deel van de resultaten (non-realisatie vs. realisatie van de (r)) correct voorspeld wordt door de drie verklarende variabelen die in de analyses zijn opgenomen: warenhuis, nadruk en woord.

Classification Table[a]

	Observed		Predicted R		
			−r	+r	Percentage Correct
Step 1	r	−r	449	50	90,0
		+r	148	83	35,9
		Overall Percentage			72,9

a The cut value is ,500

Tabel 2 laat zien welke voorkomens van de (r) correct voorspeld werden en welke niet. Het totaal percentage correct voorspeld is 72.9%. In 148 gevallen luidt de voorspelling een non-realisatie, terwijl het feitelijk een realisatie was. Voor 50 gevallen geldt het omgekeerde: realisatie voorspeld, maar een non-realisatie gevonden ofwel geobserveerd. Stelt dit resultaat tevreden? Bij een analyse van variatiepatronen zouden we pas tevreden kunnen zijn als elk individueel voorkomen precies voorspeld zou kunnen worden. Daar staat tegenover dat het kwantitatieve onderzoek naar taalvariatie pas succesvol kon zijn door aan te nemen dat er inherente variatie en waarschijnlijkheidsprocessen aan het werk zijn. Een waarschijnlijkheidsproces impliceert dat op het laagste niveau, dat wil zeggen het niveau van de afzonderlijke gebeurtenissen, de voorspellingen niet perfect zijn.

Hoe wordt nu de inbreng van de afzonderlijke variabelen geëvalueerd? Het overzicht is te vinden in Tabel 3, waar een reeks maten te vinden is. Er is een aantal complicaties waarop hier amper kan worden ingegaan. Eén ervan is de status van

de variabele warenhuis. Omdat deze variabele drie waarden heeft moet bepaald worden of het om een nominale variabele gaat met drie categorieën (de warenhuizen zijn verschillend) of om een schaalvariabele (de warenhuizen zijn geordend op een schaal, bijvoorbeeld op een schaal lopend van 1 tot 3). In de analyse in Tabel 2 is gekozen voor een nominale interpretatie en vandaar dat er sprake is van warenhuis(1) en warenhuis (2). Warenhuis (3) fungeert als ijkpunt of vergelijkingspunt (in dit geval ook als ook een nulpunt), wat inhoudt dat in warenhuis (1) en (2) meer (r) realisaties voorkomen. Daarom is er zowel een positieve B voor warenhuis (1) als warenhuis (2). De B van warenhuis (1) is groter dan die van warenhuis (2) omdat het verschil tussen Saks en Klein's groter is dan het verschil tussen Macey's en Klein's.

Tabel 3 Overzicht van de statistische analyse van de verklarende variabelen in een logistische regressie van de (non-)realisatie van de (r).

Variables in the Equation

		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 1[a]	warenhuis			66,044	2	,000	
	warenhuis(1)	2,256	,282	64,182	1	,000	9,549
	warenhuis(2)	1,800	,262	47,392	1	,000	6,052
	nadruk	,320	,179	3,200	1	,074	1,377
	woord	,991	,175	32,104	1	,000	2,694
	Constant	-4,249	,465	83,582	1	,000	,014

a Variable(s) entered on step 1: warenhuis, nadruk, woord.

De B in tabel 3 is de coëfficiënt die aangeeft wat de inbreng van een effect (het gewicht) is in de logistische functie. Van belang is de significantie, die bepaald wordt met behulp van de Wald-coëfficiënt (B gedeeld door het kwadraat van de standaardfout (= S.E.)); het is algemeen gebruik een significantieniveau van 5% oftewel $p < .05$ te hanteren. Alle effecten zijn significant behalve dat van woord, waar een waarschijnlijkheidswaarde van .074 gegeven wordt. Dat is een waarde hoger dan .05. De variabel woord is niet significant.

De meeste gebruikelijke manier om de sterkte van een effect vast te stellen, is de zogeheten Exp(B), die helemaal rechts staat in tabel 3. Het gaat om de odds-ratio die het verschil weegt tussen twee odds-waarden. De odds-waarde voor Saks is 85/93, wat grofweg genomen 0.9 is. De odds voor Klein's is 21/195, hetgeen grofweg .1 is. De twee odds worden onderling gewogen door ze te delen en de uitkomst voor de odds-ratio is dan 9. Dat komt dicht in de buurt van de waarde die in tabel 3 staat. De waarde in tabel 3 is preciezer, omdat in die berekening ook de invloeden van de overige variabelen zijn meegenomen en gewogen. Warenhuis (2) geeft aan dat het aantal realisaties van de (r) bij Macy's in termen van odds in vergelijking met Klein's de factor 6 heeft. Dat is lager dan voor Saks, overeenkomstig de getallen in tabel 1. De variabele nadruk is niet significant en de odds wijken daarom niet sterk van de waarde 1 af. Er blijkt weer wel een duidelijk verschil tussen de beide betrokken woorden. De odds-ratio is bijna 3, hetgeen weer lager is dan de odds-ratio's voor de warenhuizen. Dat is echter ook wat we verwachten, want de cijfers laten

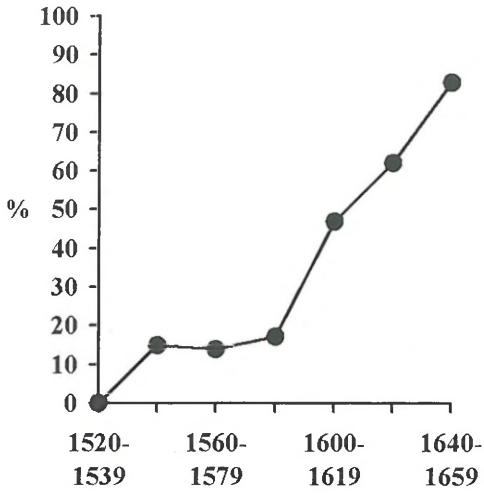
duidelijk veel grotere verschillen zien voor de warenhuizen. Met de odds-ratio kunnen we evenwel heel precies inschatten hoe sterk een effect is.

De hier gehanteerde techniek van de logistische regressie blijkt voor de analyse van taalvariatie een reeks van mogelijkheden te bieden: (1) getoetst kan worden of bepaalde variabelen een significant verband hebben met de onderzochte taalvariabele; (2) de sterkte van het verband kan worden bepaald (indeling naar zwakkere versus sterkere effecten op basis van de odds-ratio); (3) het effect van een variabele kan worden onderzocht in de context van andere variabelen (competitie, modellering); (4) de techniek kan zowel in toetsende als exploratieve zin gebruikt worden. Daarbij mag niet vergeten worden dat de complexiteit van de techniek snel oploopt met het aantal variabelen dat in de analyse betrokken wordt.

Tot slot kan vermeld worden dat deze techniek ook uitstekend toegepast kan worden op de hedendaagse veranderingen in de Nederlands (r). Vooral postvocalisch wordt deze klank zwak gerealiseerd, terwijl de Gooise (r) zeer snel in opkomst is. De (r) blijkt overigens opmerkelijk vaak betrokken te kunnen zijn in variatie- en veranderingsprocessen (zie voor een overzicht Van de Velde & Van Hout 2001).

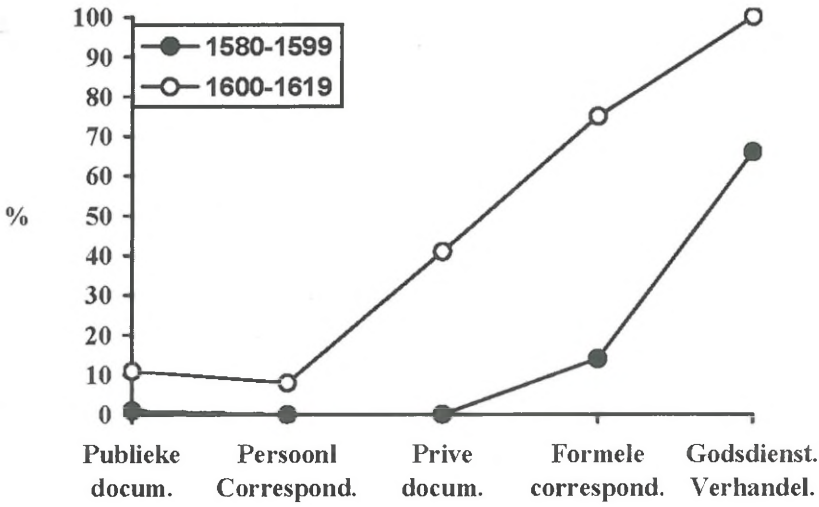
3 De tijdsfactor: taalhistorisch onderzoek

De kwantitatieve benadering van historisch taalonderzoek heeft zich de afgelopen twee decennia voor het Engels en vooral in sociolinguïstisch opzicht snel ontwikkeld. De factor tijd wordt daarbij aan een variabel taalverschijnsel gekoppeld. Een voorbeeld vormt het aandeel van Engelse betrekkelijke voornaamwoorden in het (Engels) Schots die in de loop van de tijd de oorspronkelijke Schotse varianten vervangen. Devitt (1989) bestudeerde de verengelsing van het Middelschots. Figuur 1 laat de verspreiding zien van de zuidelijke Engelse *wh*-woorden in betrekkelijke bijzinnen (*which*, *whilk*, *who*, *where-*), die de oorspronkelijke *quh*-vormen vervingen (*quhilk*, *quha*, *quh-*; Devitt 1989: 19).



Figuur 1 Verengelsing van Schotse betrekkelijke voornaamwoorden in de tijd, gemeten in percentage wh-*vormen ten opzichte van totaal aantal vormen* (wh- *plus* quh-*vormen), naar Devitt (1989: 18, 87).*

Figuur 1 laat een duidelijke toename over zeven periodes van telkens twintig jaar zien, beginnend bij nul (alleen Schotse vormen) tot meer dan 80%, hetgeen een bijna complete verengelsing inhoudt. Dit percentage geeft het gemiddelde over een groot aantal teksten en tekstsoorten en het is vervolgens van belang dat aanvullende analyses gedaan kunnen worden, die de onderliggende factoren onthullen die een rol spelen. Devitt doet daarvoor een beroep op genreverschillen die variëren op de dimensie van formaliteit. Zij laat zien dat er grote genreverschillen zijn in de diffusie van de zuidelijke Engelse vormen. Dat onderscheid blijkt duidelijk uit Figuur 2, waar voor twee tijdsperiodes het aandeel van de Engelse vormen wordt weergegeven.

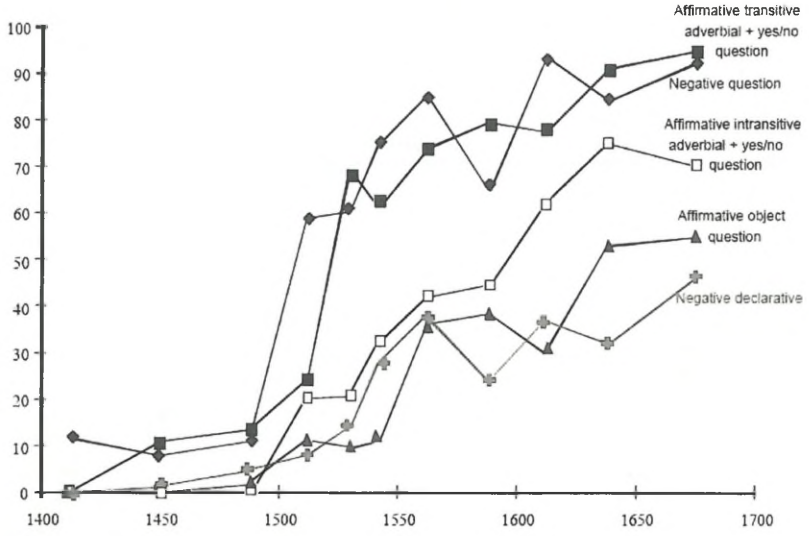


*Figuur 2 Verengelsing (*wh*-vormen) van Schotse betrekkelijke voornaamwoorden in diverse genres in twee tijdsperiodes. Percentages* wh*-vormen (naar Devitt 1989: 61, 97).*

Op de horizontale as van figuur 2 staan de genres en de lijnen representeren twee tijdsperiodes. Het genre van de publieke documenten moet even apart toegelicht worden omdat het gaat om nationaal-Schotse documenten hetgeen de vermijding van Engelse elementen verklaart. De vier overige genres liggen op een formaliteitscontinuüm en figuur 2 laat zien dat de verengelsing toeneemt in de formelere genres. De formaliteitsdimensie neemt in het sociolinguïstisch variatie-onderzoek een prominente plaats in en wat we daar synchroon vonden vinden we nu in een diachroon perspectief. De hogere scores voor de formele genres zijn niet verrassend, omdat het om de invoering gaat van een variant die door de heersende klasse gebruikt wordt. Hoe kunnen we de tijdsdimensie in het variatiemodel verwerken? Zoals Figuur 1 en 2 laten zien verlopen de veranderingsprocessen in de tijd niet volgens een rechte lijn. Het verloop is niet lineair, maar curvilineair en meestal in een S-achtige vorm. Dat is bekend van groei- en veranderingsprocessen in het algemeen. Het standaardpatroon is er één van een voorzichtig begin, dan een versnelling, weer gevolgd door een vertraging, met een afvlakkend verloop op het eind. Technisch gezien valt dat uit te leggen als een wisselwerking tussen een exponen-

tiële groei en restrictieve omstandigheden die in gewicht toenemen naarmate de groei vordert (in taalveranderingsperspectief: het aantal elementen dat kan veranderen neemt af met de benadering van het populatiemaximum oftewel verzadigingspunt). Het gevolg is een logistische curve, zo genoemd door de Belgische wiskundige Pierre-François Verhulst.

Kroch (1989) heeft de logistische curve gebruikt om taalverandering in tijd te modelleren. Logistische curves kunnen parallel lopen. De twee lijnen in Figuur 2 bijvoorbeeld vertonen een soort van parallel verloop. Parallelle logistische curves in de tijd houdt in dat ze hetzelfde verloop hebben, maar dat ze in de tijd gezien eerder of later starten. Een parallel verloop houdt in dat ze met dezelfde snelheidscoëfficiënt verlopen. Het voorbeeld dat Kroch geinspireerd heeft zijn de data van Ellegård (1953). De data van Ellegård (1953) betreffen het voorkomen van *do*-support in het Engels van de vijftiende tot en met de zeventiende eeuw, waarbij een onderscheid gemaakt wordt naar vijf verschillende zinsconstructies waarin *do* geïntroduceerd werd. Het gaat om dezelfde taalverandering, maar om verschillende constructies dan wel contexten waarin deze verandering zichtbaar wordt. Het kwantitatieve verloop van de verandering is weergegeven in figuur 3.



Figuur 3 Verloop over de tijd van do*-insertie in vijf syntactische constructies op basis van de data van Ellegård (1953).*

Figuur 3 laat schommelingen zien die mede van doen kunnen hebben met het corpusmateriaal, een verschijnsel dat normaal is bij het trekken van steekproeven van en uit teksten. De figuur laat zeker in de periode van eind 1400 tot ergens tot aan het einde van 1600 een scherpe stijging zien, zoals die voorspeld wordt door een logistische functie. Twee contexten lijken daarin voorop te lopen, de negatieve vraagconstructie en de bevestigende transitieve constructie. Krochs stelling is dat zolang een verandering zich vrijelijk kan ontwikkelen de verschillende con-

texten paralelle S-curves opleveren. In sommige contexten treedt *do*-insertie eerder op, in andere weer later, maar het verloop is vervolgens hetzelfde.

Dat houdt in dat de volgende functie geldt: de verandering in een bepaalde context wordt bepaald door een tijdstip waarop die verandering inzet plus een hellingshoek. Die hellingshoek is een constante in logaritmische zin maar verandert met de tijd in exponentiële zin. Dat betekent dat we met een enkel getal het verloop van een S-curve kunnen omvatten.

Kroch (1989) komt vervolgens tot een opmerkelijk resultaat als hij de statistische analyses beperkt van meetmoment 1 tot meetmoment 7, dat wil zeggen tot 1575. De vijf geschatte functies hebben dan allemaal dezelfde hellingshoek (dezelfde snelheidscoëfficiënt). De vijf contexten hebben hun eigen positie in de verandering, zij lopen voorop of wellicht wat achter, maar het verloop is vervolgens identiek. Kroch (1989) beargumenteert vervolgens dat 1575 een keerpunt was omdat toen een ingrijpende herschikking van de Engelse grammatica plaatsvond. Zijn VARBRUL-analyses leveren identieke waardes op voor de hellingscoëfficiënt voor de gegevens van voor 1575. Dat resultaat is overigens eenvoudig na te rekenen in SPSS. De conclusie is dan dat een taalverandering zich vrij voorspelbaar zo niet mechanisch kan voltrekken nadat deze is ingezet. De verschillende contexten vormen een soort mal waarlangs zich de verandering voltrekt en hun inbreng wijzigt zich niet tijdens de verandering. Er zijn andere omstandigheden die een inbreuk kunnen doen op het verloop, zoals de syntactische heranalyse van het Engels die door Kroch rond 1575 gepostuleerd wordt (zie ook Kroch & Taylor 2000, Pintzuk 2005, Pintzuk Taylor 2006). De resultaten van Kroch doen in elk geval de vraag rijzen hoe we nu de factor tijd in de veranderingsmodellen moeten implementeren. Een belangrijke literatuurverwijzing op dit punt is Nevalainen & Raumolin-Brunberg (2003). In hun historisch-sociolinguïstische benadering staat het concept van de variabele regel (en daarmee de logistische functie) centraal, waarbij een onderscheid gemaakt wordt in een tijdsfactor, een sociale dimensie en een geografische dimensie.

Uiteraard kan verder logistische regressie-analyse worden gebruikt om vast te stellen welke variabelen of factoren van invloed zijn op een taalvariabele met twee waarden. Voor een bepaalde periode kan nagegaan worden welke verklarende variabelen van invloed zijn en wat hun sterkte is. Dit illustreren we weer aan de hand van een Engels voorbeeld. Het betreft een geval van woordschikkingvariatie, dit keer van nominale subjecten ten opzichte van discoursepartikels als *þa* en *þonne* (beiden letterlijk 'toen'). In het volgende voorbeeld verwijst *sio*, dat in dit verband eerder 'dit' dan 'de' betekent, concreet terug naar 'met geduld':

> Forðæm bið se sige micle mara ðe man mid geðylde gewinð, forðæm sio **gesceadwisnes ðonne** hæfð ofercumen ðæt mod & ... (cocuraC,CP_[Cotton]:33.218.19.42)
> Daarom is de zege veel groter die men met geduld behaalt, omdat dit onderscheidingsvermogen de geest overkomen is....

De achterliggende aanname (Van Kemenade & Los 2006) is dat partikels als *þa* en *þonne* als functie hebben om een discoursedomein af te bakenen. In dat licht is het opmerkelijk dat nominale subjecten een duidelijke variatie in positie laten zien,

wat de vraag doet rijzen waardoor deze variatie wordt bepaald. Het algemene totaal is in tabel 4 te vinden.

Tabel 4 De relatieve positie van het nominale subject t.o.v. discoursepartikels als *þa/þonne* in de bijzin.

	Bijzin
Nominale subject volgt op *þa/þonne*	129
Nominale subject gaat vooraf aan *þa/þonne*	221

De hypothese is als volgt: nominale subjecten die voorafgaan aan *þa/þonne* verwijzen concreet en specifiek terug naar een antecedent in de discourse. Dat kunnen ze mede zo doen omdat het definiete aanwijzende voornaamwoord dat deel uit maakt van de zelfstandige naamwoordsgroep, niet alleen een aanwijzende (deictisch) maar ook een verwijzende (anafore) functie kan hebben. Nominale subjecten die volgen op *þa/þonne* kunnen wel definiet (maar niet specifiek) zijn.

De mogelijk verklarende factoren zijn ondergebracht in een logistische regressie-analyse met als mogelijke verklarende variabelen: positie, specificiteit en definietheid (zie Van Kemenade, Milicev en Baayen in voorbereiding). De analyse is in dit geval gedaan met de logistische regressie-analyses uit het statistische softwarepakket R. De uitkomsten staan in tabel 5.

Tabel 5 De uitkomsten in R van een logistische regressie-analyse van drie factoren die van invloed zijn op de positie van het subject ten opzichte van discoursepartikels.

	Coef	S.E.	Wald Z	P
Intercept	1.987	0.2409	8.25	0.0000
NPPosition=mid	-1.403	0.3253	-4.31	0.0000
NPType=indefinite	-3.926	0.4461	-8.80	0.0000
NPType=properName	2.394	1.0353	2.31	0.0208

Tabel 5 laat uitkomsten zien die vertaalbaar zijn in de uitkomsten die we voor de SPSS-analyse hadden. Zo staat 'intercept' voor wat in SPSS 'constant' genoemd wordt, in een vergelijking met twee variabelen te vergelijken met de waarde van y waarbij x=0. Alle drie de gedefinieerde variabelen blijken een eigen significante inbreng te hebben. De variabele NPosition heeft betrekking op de positie van het subject in de zin als geheel (begindeel vs. middendeel). We geven hier geen verdere uitleg, maar stellen wel vast dat het interessante is dat we er met deze techniek in slagen om een systeem dat betrekkelijk diffuus is (er zijn meerdere factoren die een rol spelen, en de uitkomsten lijken nogal variabel) toch nauwkeurig te kwantificeren en te doorgronden.

4 Nieuwe ontwikkelingen en discussiepunten

Hoe staat het met de toekomstige ontwikkelingen in het taalvariatie-onderzoek? Het meest wezenlijke punt is dat dit onderzoek een centrale plaats claimt in het

taalkundige onderzoek. Verschijnselen als gradualiteit en waarschijnlijkheid lijken hun plaats op te eisen in de modellering van taal. Dit kan ook andersom geformuleerd worden. Het zou verbazingwekkend zijn als het menselijk taalsysteem zich zou ontrekken aan processen die gebaseerd zijn op waarschijnlijkheidsdistributies en processen die gradueel van aard zijn. Meer concreet zullen in de toekomst de volgende ontwikkelingen een rol spelen:

Het onderzoek naar taalvariatie lijkt volwassen te gaan worden wat betreft het gebruik van de statistiek. Er is in het verleden bijvoorbeeld te weinig aandacht geweest voor mogelijke valkuilen, die onder meer te maken hebben met afhankelijkheidsrelaties in de data (zie Rietveld, Van Hout & Ernestus 2004). Er is een sequentie-effect, hetgeen betekent dat in teksten varianten soms in clusters voorkomen. Er is het probleem van de eenheid van analyse, want vaak worden in de sociolinguïstiek de gegevens onafhankelijk van de specifieke spreker geanalyseerd. En iets vergelijkbaars in de historische taalkunde waar gegevens geanalyseerd worden los van de specifieke tekst waar ze vandaan komen. De gegevens worden gekenmerkt door de periode waaruit ze komen en niet door de specifieke tekst waaruit ze kwamen.

Verder zullen algemenere statistische pakketten gebruikt gaan worden, zoals we nu al zien in corpuslinguistisch onderzoek (Bresnan et al., te verschijnen; Grondelaers & Speelman 2007). Dat geeft een betere controleerbaarheid en een aansluiting bij modernere technieken.

Die aansluiting moet ook duidelijk maken welke keuzes er zijn in de statistische modellering. De logistische functie is een uitstekende keuze, maar er zijn alternatieven voor het leggen van een verbindende functie tussen de taalvariabele en de verklarende variabelen of factoren, zoals bijvoorbeeld de Poisson-distributie (die bedoeld is voor zeldzame gebeurtenissen). Verder doen zich nieuwe mogelijkheden voor met zogeheten multi-level modellen. Dat wordt geïllustreerd in Bresnan et al. (te verschijnen) waar het lexicale materiaal op een apart niveau kan worden geïmplementeerd, los van de overige factoren.

Multi-level-modellen geven ook weer bijzondere mogelijkheden om voor de factor tijd een speciale positie in te ruimen. Dat zal moeten omdat Krochs hypothese over de constantheid van het veranderingsproces in meerdere omstandigheden niet zal opgaan omdat er herschikkingen in het taalsysteem plaatsvinden. Een duidelijk voorbeeld waarin de gewichten van de constraints (de verklarende variabelen) over de tijd veranderen is Poplack & Malvar (2006). Zij analyseren vier varianten die de voorbije eeuwen in het Braziliaans Portugees in gebruik waren om de toekomstige tijd uit te drukken. Zij kiezen niet voor een multinominale (d.i. meerwaardige) logistische regressie-analyse, maar ze analyseren telkens één van de varianten ten opzichte van de drie andere varianten. Voor elk van de drie onderzochte tijdsperiodes worden dan telkens vier binaire logistische regressies uitgevoerd. De winst of het verlies van een constructie blijkt in een uitbreiding of beperking van het semantisch-pragmatisch toepassingsdomein te liggen, hetgeen in de loop van de tijd een verandering van de gewichten impliceert. De *ir*-constructie (= gaan) breidt zijn domein uit van het uitdrukken van de onmiddellijke nabije toekomst (negentiende eeuw) tot de meest algemene constructie in het moderne

Braziliaanse Portugees. Kunnen deze veranderingen in een totale multi-level analyse voor de drie periodes samen omvat worden? Hierbij zou apart niveau voor de factor tijd ingeruimd moeten worden.

Verder ligt er nog het verzwegen probleem van mogelijke interacties tussen de verklarende variabelen. We gaan daar verder niet op in, maar hier ligt ook een belangrijke theoretische kwestie, die al aangesneden werd in Sankoff & Labov (1979). Zij stellen dat een interactie-effect tussen linguïstische constraints of condities inhoudt dat de constraints of condities niet correct zijn geformuleerd. Een dergelijke beperking geldt zeker niet voor niet-talige variabelen.

Tot slot moet grondig nagedacht worden over de toegestane complexiteit van variatiemodellen en over de toegestane talige onafhankelijke (verklarende) variabelen. De statistische analyses moeten op inhoudelijke en theoretische gronden worden ingeperkt, ook al om te voorkomen dat telkens weer wisselende verzamelingen van constraints worden aangedragen als verklaringen voor overeenkomstige linguïstische verschijnselen. Als we de opkomst van *gaan* als hulpwerkoord van de toekomstige tijd in het Nederlands willen verklaren dan zou dat bij voorkeur gaan op grond van dezelfde constraints als in Poplack & Malvar (2006).

5 Afsluiting

We willen graag afsluiten met een taalkundige vraag. In deze bijdrage hebben we steeds gesproken over twee varianten en de competitie daartussen. Daarmee sluiten we onderzoek naar afhankelijke variabelen van een gradueel niveau niet uit, maar daarvoor zijn andere technieken beschikbaar. Metingen op klinkers oftewel vocalen wordt vaak gedaan op basis van de zogeheten eerste en tweede formant, harde metingen op basis van het akoestische signaal. Een multivariate ANOVA-analyse voor vocaalformanten wordt nader toegelicht in Rietveld & Van Hout (2005: 174-177). Maar taalvariabelen in een linguïstisch kader kunnen vaak in onderscheiden varianten uiteengelegd worden. Varianten zijn alternatieve vormen voor een specifiek taalverschijnsel. Dat lijkt de beste definitie van een taalvariabele.

Vooral vanuit de hoek van de generatieve grammatica bestaat er echter een voorkeur voor competitie op een hoger niveau dan varianten, namelijk tussen grammatica's. Daarmee blijft de grammatica zelf vrij van waarschijnlijkheidsmechanismen en onzekerheidsprincipes. Dit standpunt maakt duidelijk dat we steeds de vraag moeten stellen op welk niveau van het taalsysteem zich een verandering voltrekt en hoe de grammatica van een taalgemeenschap van toestand A in toestand B terecht komt.

Daarbij komt ook nog de vraag waarom bepaalde variabele verschijnselen niet veranderen. Een voorbeeld hiervan vormt het volgende paar Nederlandse zinnen:

(1) a. Hij zei dat Jan gisteren het boek gelezen had
b. Hij zei dat Jan het boek gisteren gelezen had.

Deze variatie is een stabiele variabele in het hedendaagse Nederlands. Het is ook 'echte' variatie, in die zin dat sprekers beide varianten gebruiken en dat er geen betekenisverschil is tussen de twee varianten. Stabiele variatie is een uitzondering; in

historisch syntactisch onderzoek blijkt steeds weer dat, diachroon gezien, woordschikkingsvariatie slechts zelden stabiel is, en meestal een signaal dat er sprake is van taalverandering. Zo kende het vroegere Nederlands een woordschikkingsvariant waarbij een object, ook in de bijzin, aan het subject vooraf kon gaan, zie bijvoorbeeld Shannon (2003). Het volgende voorbeeld uit een Drentse gerechtelijke tekst komt uit het DiT corpus (Dutch in Transition, momenteel in de maak aan de RU Nijmegen).

(2) dat den acker Reyner gebruket hevet (anno 1420)

Mogelijk waren de varianten relatief vrij, maar hier moet nog nader onderzoek volgen. Ribbert (2006) laat in elk geval zien dat deze variant in de loop van de vijftiende eeuw in Drenthe steeds verder in onbruik raakte. Om een en ander te onderzoeken hebben we vooral grote gedigitaliseerde corpora nodig. Dat geldt voor de historische kant en dat geldt voor de sociolinguïstische kant.

Hoe zit het dan met de diepgang? De kwantitatieve analyses maken het mogelijk om onderliggende verklarende variabelen te wikken en te wegen in expliciete modellen. Het wordt mogelijk om dieper te graven in de verklaringsbronnen van taalvariatie en taalverandering. We kunnen ons intussen al helemaal niet meer voorstellen hoe we de grote hoeveelheden gegevens zouden kunnen doorgronden zonder kwantitatieve analyses. De tweede vorm van diepgang schuilt in de mogelijkheid om taalveranderingsmodellen, bij voorkeur ontwikkeld op theoretische gronden, te toetsen aan de gevonden gegevens.

Bibliografie

Bresnan, Cueni, Nikitina & Baayen (te verschijnen) – J. Bresnan, A. Cueni, T. Nikitina & R. H. Baayen: 'Predicting the Dative Alternation'. Te verschijnen in: G. Bourne, I. Kraemer & J. Zwarts (ed.): *Cognitive foundations of interpretation*. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Chambers 2003 – J. Chambers: *Sociolinguistic theory*. Oxford: Blackwell, 2003 (tweede druk).

Devitt 1989 – A. Devitt: *Standardizing written English: Diffusion in the case of Scotland 1520-1659*. Cambridge: Cambridge University Press, 1989.

Ellegård 1953 – A. Ellegård: *The Auxiliary do: The Establishment and Regulation of its Use in English*. Stockholm: Almqvist en Wiksell, 1953.

Fischer 1959 – J. Fischer: 'Social influences on the choice of a linguistic variant'. In: *Word* 14 (1959), p. 47-56.

Grondelaers & Speelman 2007 – S. Grondelaers & D. Speelman: 'A variationist account of constituent ordering in presentative sentences in Belgian Dutch'. In: *Corpus Linguistics and Linguistic Theory* (2007).

Van Kemenade & Los 2006 – A. van Kemenade en B. Los: 'Discourse adverbs and clausal syntax in Old and Middle English'. In: A. van Kemenade & B. Los (eds.): *The Handbook of the History of English*. Malden, Mass. [etc.]: Blackwell, 2006, p. 224-248.

Van Kemenade, Milicev en Baayen (in voorbereiding) – A. van Kemenade, T. Milicev en R.H. Baayen: 'Discourse reference by topics'.

Kroch 1989 – A. Kroch: 'Reflexes of grammar in patterns of language use'. In: *Language Variation and Change* 1 (1989), p. 199-244.

Kroch 2001 – A. Kroch: 'Syntactic change'. In: M. Baltin & C. Collins (eds.): *The handbook of contemporary syntactic theory*. Oxford: Blackwell, 2001.

Kroch & Taylor 2000 – A. Kroch en A. Taylor: 'Verb object order in early Middle English'. In: S. Pintzuk, G. Tsouls en A.Warner (red.): *Diachronic Syntax: Models and Mechanisms*. Oxford: Oxford University Press, 2000, p.132-163.

Labov 1966 – W. Labov: *The stratification of English in New York City*. Washington: Center for Applied Linguistics, 1966.

Nevalainen & Raumolin-Brunberg 2003 – T. Nevalainen en H. Raumolin-Brunberg: *Historical sociolinguistics: Language change in Tudor and Stuart England*. London: Longman, 2003.

Paolillo 2002 – J. Paolillo: *Analyzing linguistic variation. Statistical models and methods*. Stanford: Stanford University Press, 2002.

Pintzuk 2005 – S. Pintzuk: 'Arguments against a universal base: evidence from Old English'. In: *ELL* 9 (2005), p. 1139-156.

Pintzuk & Taylor 2006 – S. Pintzuk & A. Taylor: 'The loss of OV order in the history of English'. In: A. van Kemenade & B. Los (eds.): *The Handbook of the History of English*. Malden, Mass. [etc.]: Blackwell, 2006, p. 249-278.

Poplack & Malvar 2006 – S. Poplack en E. Malvar: 'Modelling linguistic change. The past and the present of the future in Brazilian Portuguese'. In: F. Hinskens (ed.): *Language Variation – European Perspectives*. Amsterdam/Philadelphia: Benjamins, 2006, p. 169-199.

Ribbert 2006 – A. Ribbert: 'Object preposing in 15th century Drenthe'. In: B. Los en J. van deWeijer (ed): *Linguistics in the Netherlands 2006*. Amsterdam: John Benjamins, 2006.

Rietveld & Van Hout 2005 – T. Rietveld en R. van Hout: *Statistics in language research: Analysis of Variance*. Berlin: Mouto de Gruyter, 2005.

Rietveld, Van Hout & Ernestus 2004 – T. Rietveld, R. van Hout en M. Ernestus: 'Pitfalls in corpus research'. In: *Computers and the Humanities* 38 (2004), p. 343-362.

Sankoff & Labov 1979 – D. Sankoff en W. Labov: 'On the uses of variable rules'. In: *Language in Society* 8 (1979), p. 189-222.

Shannon 2003 – T.F. Shannon: 'Drift in Dutch: Fleshing out the factors of change'. In: A.Verhagen en J. M. van de Weijer (ed.): *Usage-Based Approaches to Dutch*. Utrecht: LOT, 2003.

Van de Velde & Van Hout 2001 – H. van de Velde en R. van Hout (red.): *r-atics. Sociolinguistic, phonetic and phonological characteristics of /r/*. Brussel: Université Libre de Bruxelles, 2001 (Etudes & Travaux).

Adressen van de auteurs

Ans van Kemenade, Engelse taal en cultuur, Radboud Universiteit Nijmegen, Erasmusplein 1, NL-6525 HT Nijmegen

Roeland van Hout, Taalwetenschap, Radboud Universiteit Nijmegen, Erasmusplein 1, NL-6525 HT Nijmegen