MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport S 132

Over het vóórkomen van consonantverbindingen in de Nederlandse spreektaal

door
Constance van Eeden
en
Gerda Klerk-Grobben

1954

1. Inleiding.

Voor het opstellen van teksten ten behoeve van een onderzoek betreffende de verstaanbaarheid van doofstommen werd ons gevraagd een schatting te geven van de frequentie van voorkomen van consonantverbindingen in de Nederlandse spreektaal.

De gegevens hiervoor werden als volgt verzameld. Men liet een groot aantal personen een verhaal vertellen bij vier verschillende plaatjes. Uit de bij ieder van de vier plaatjes vertelde verhalen werd aselect een aantal woorden (voor ieder plaatje $\pm$ 2000) gekozen en in ieder van deze vier steekproeven werden de verschillende consonantverbindingen geteld.

2. Statistisch onderzoek.

Het bleek dat bepaalde woorden bij bepaalde plaatjes zeer veel en bij andere plaatjes niet of zeer weinig gebruikt werden. Om hiervoor een correctie aan te brengen, werd bij ieder van de plaatjes, onafhankelijk van de waarnemingsresultaten, een lijstje samengesteld van woorden, waarvan verwacht mag worden dat ze gemakkelijk met dat plaatje in verband gebracht kunnen worden. Bij elk van deze woorden werd daarna bepaald hoe vaak ze in elk van de steekproeven voorkomen. De correctie op het waarnemingsmateriaal bestaat nu hierin dat uit de steekproeven behorend bij elk plaatje de woorden, voorkomend in dit lijstje, geschrapt worden.

Op het aldus gecorrigeerde materiaal werden twee statistische toetsen toegepast (zie voor de algemene gang van zaken bij het toetsen van een hypothese memorandum S 47 (M 6) [1]):

1. De χ^2 -toets. Deze werd toegepast om te onderzoeken of er systematische verschillen zijn tussen de vier steekproeven wat betreft de frequenties van het voorkomen van de verschillende consonantverbindingen.

Deze toets wordt als volgt uitgevoerd: Is n_{ij} het aantal malen dat de i^{e} consonantverbinding bij de j^{e} steekproef voorkomt, zij verder $n_j = \sum_i n_{ij}$, $t_i = \sum_j n_{ij}$ en zij n het totaal aantal consonantverbindingen bij de vier steekproeven tezamen, dan is de toetsingsgrootheid

$$\chi^2 = \frac{1}{n} \sum_{i,j} \frac{(n_{ij} n - n_j t_i)^2}{n_j t_i} .$$

Deze grootheid bezit, als er geen systematisch verschil tussen

---

1) De memoranda, waarnaar in dit rapport wordt verwezen, zijn achteraan bij gevoegd.

steekproeven is, bij benadering een χ^2-verdeling [2] met $(h-1)(k-1)$ vrijheidsgraden; hierin is h het aantal steekproeven (in ons geval dus 4) en k het aantal consonant-verbindingen.

Het resultaat van deze toets is een overschrijdingskans $< 10^{-4}$, zodat de hypothese van overeenstemming tussen de vier steekproeven met stelligheid verworpen moet worden. Ondanks de door ons aangebrachte correcties (zie boven) zijn er dus toch nog duidelijke verschillen tussen de bij de verschillende plaatjes verkregen resultaten. Dit betekent dus dat er, ook na de correctie, nog verschillen zijn tussen de vier steekproeven wat betreft de frequentie van consonantverbindingen en dus ook van de woorden. Daar echter niet meer bekend is uit welke woorden de steekproeven bestonden is het ons niet mogelijk dit nader te onderzoeken, ten einde nog verdere correcties aan te brengen.

2. Ondanks deze verschillen in frequenties van consonantverbindingen, zou er toch overeenstemming tussen de plaatjes kunnen bestaan in de volgorde naar grootte van deze frequenties. Dit werd onderzocht met de methode van m rangschikkingen (zie memorandum S 47 (M 14)). Bij deze toetsing werden alleen die consonantverbindingen beschouwd welke (na correctie!) meer dan 20 maal bij de vier plaatjes tezamen voorkomen. Dit geschiedde, omdat het, statistisch gezien, niet gunstig zou zijn de zeldzamere consonantverbindingen (waarvoor men dus eigenlijk een uitgebreider onderzoek zou moeten verrichten) in deze toets te betrekken. Het resultaat was een overschrijdingskans $< 10^{-4}$, zodat de nulhypothese, welke geen overeenstemming tussen de vier volgorden veronderstelt, met stelligheid verworpen kan worden ten gunste van de onderstelling dat er een zekere overeenstemming in volgorde bestaat.

Deze misschien wat tegenstrijdig lijkende resultaten van de twee toetsen kunnen met het volgende voorbeeld verklaard worden. Stel we vergelijken drie steekproeven met elk drie consonantverbindingen en we vinden de volgende frequentieverdelingen: zie fig. 1

---

2) Een χ^2-verdeling is een bepaald type van waarschijnlijkheidsverdeling. De verdelingen van dit type worden nader gekarakteriseerd door een parameter welke "het aantal vrijheidsgraden" genoemd wordt. Tabellen van de χ^2-verdeling treft men aan in nagenoeg alle leerboeken van wiskundige statistiek.

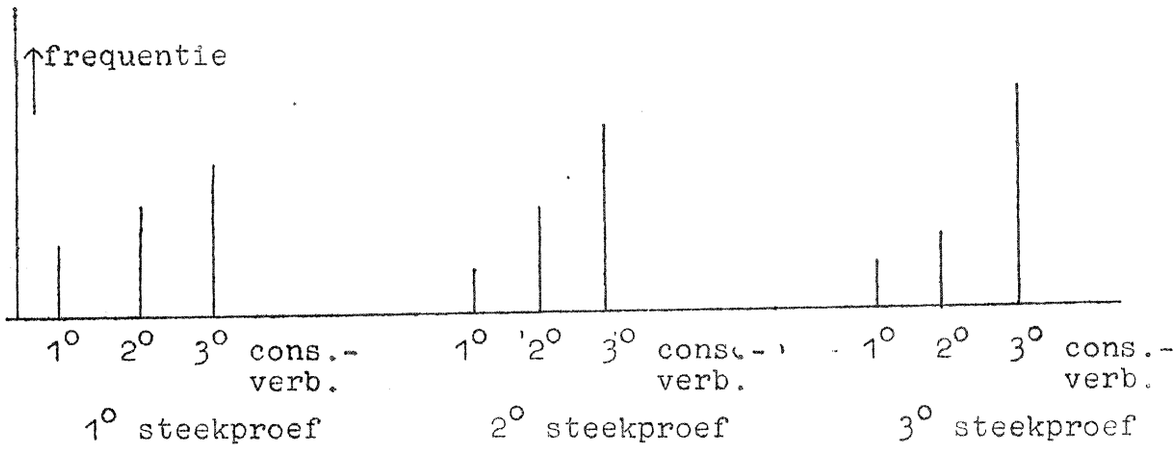

Fig. 1. Voorbeeld ter verduidelijking van de resultaten.

Dan is er dus weinig overeenstemming, dus een aanmerkelijk verschil, in de frequenties, maar toch is de grootte-volgorde der frequenties in elk der drie steekproeven dezelfde.

Berekenen wij uit het gecorrigeerde waarnemingsmateriaal de frequentie, waarmee de verschillende consonantverbindingen voorkomen in de vier steekproeven tezamen genomen, dan vinden wij:

Tabel I

Aantal malen dat een bepaalde consonantverbinding voorkomt op 100 consonantverbindingen

st	15
nt	12
sch	10
cht	7
rt	5
kt	5
ft	4
ts	4
fr	4
gr	4
kl	3
lt	3
bl	3
br	2
dr	2
rk	2
ns	2
ngk	2
tr	2
rg	2
sl	1
lk	1

Tabel I (vervolg)

mt	1
nst	1
kr	1
ks	1
ls	1

De overige consonantverbindingen komen alle minder dan één op de 100 maal voor.

Uit het voorgaande volgt dat aan deze schattingen voor de frequenties weinig betekenis kan worden gehecht, maar dat een schatting van de volgorde van deze frequenties veel meer zin heeft. Met behulp van de methode van m rangschikkingen kan een schatting van de volgorde gegeven worden met behulp van het totaal van de rangnummers in de vier steekproeven voor elk van de consonantverbindingen. Deze schatting voor de volgorde naar dalende frequentie is:

st, nt, sch, cht, rt, kt, ft, gr, fr, ts, lt, kl, br, dr, tr, ns, bl, rk, rg, ngk, sl, mst, lk, mt

MATHEMATISCH CENTRUM
Amsterdam
Statistische Afdeling.

S47(M6)

Algemene gang van zaken bij het toetsen van een hypothese. [1]

De toetsing van een hypothese H_0 berust steeds op een aantal waarnemingen $x_1, x_2, \ldots, x_n$ van één of meer stochastische grootheden[2], of op enige groepen van waarnemingen (bv. twee steekproeven).

Bij een toets behoort een toetsingsgrootheid $\underline{u}$ (soms meer dan één), die een functie is van bovengenoemde stochastische grootheden en die, voor de waargenomen waarden $x_1, x_2, \ldots, x_n$ een waarde aanneemt, die berekend kan worden (bv.: het gemiddelde der waarnemingen, of de spreiding, of het verschil van de gemiddelden van twee waarnemingen).

De toetsingsgrootheid wordt steeds zo gekozen, dat men, op grond van de onderstelling, dat H_0 juist is, de waarschijnlijkheidsverdeling van deze grootheid kan berekenen.

Vervolgens kiest men een verzameling Z van mogelijke uitkomsten van $\underline{u}$, en wel op zodanige wijze, dat de kans, dat $\underline{u}$ een in Z gelegen waarde aanneemt, onder de hypothese H_0, gelijk is aan een gegeven getal α, zodat Z dus van α afhankelijk is[3]. Z heet de kritieke zône van de toets, α de onbetrouwbaarheidsdrempel (Engels: level of significance). Voor α neemt men veelal de waarde 0,05 of 0,01.

Men verwerpt nu H_0 op grond van de waarnemingen $x_1, x_2, \ldots, x_n$, indien de bij deze waarnemingen behorende waarde van $\underline{u}$ in Z ligt. Dit wordt vaak uitgedrukt door te zeggen, dat het resultaat van het experiment "significant" is. De waarde van α moet dan echter worden vermeld. De kans, dat dit zal gebeuren, is, indien H_0 juist is, gelijk aan α. Derhalve is α de kans op ten onrechte verwerping van de juiste hypothese, ook de kans op een fout van de eerste soort genoemd. Indien men deze methode toepast, met $\alpha = 0{,}05$ resp. 0,01, zal men in gemiddeld ongeveer één op 20 resp. op 100 van de gevallen, waarin de hypothese die men toetst juist is, deze toch verwerpen.

1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.

2) Een stochastische grootheid is een grootheid, die een waarschijnlijkheidsverdeling bezit, of, anders gezegd, een grootheid, die voor de elementen van een collectie(universum,populatie) gedefinieerd is en daarop allerlei waarden aanneemt. Stochastische grootheden worden aangegeven door onderstreepte letters.

3) Soms kan men slechts bereiken, dat deze kans $\leq \alpha$ is.

De toetsingstheorie biedt in het algemeen geen mogelijkheid om tot aanvaarding van een hypothese te komen. Indien een bepaalde hypothese H_0 niet verworpen kan worden, is dit gewoonlijk met een hele verzameling van hypothesen tegelijk het geval. Niet-verwerpen staat dus niet gelijk met aanvaarden.

Wel zal men vaak in de loop van een statistische analyse bepaalde onderstellingen, die plausibel schijnen en voor de verdere analyse van nut zijn, toetsen, alvorens ze bij de verdere bewerking van het materiaal te gebruiken. Worden zij dan op grond van de toets niet verworpen, dan houdt dit in zo verre een rechtvaardiging van die onderstellingen in, dat een grote afwijking door de toets veelal wel zou zijn ontdekt. Indien men dan verder de onderstellingen gebruikt, verwaarloost men eventueel aanwezige afwijkingen van onbekende grootte, die echter niet zo groot zijn, dat zij door de toets zijn ontdekt.

Vele toetsen gelden zelf alleen onder bepaalde onderstellingen omtrent de waarschijnlijkheidsverdelingen der stochastische grootheden, waarvan waarnemingen zijn verricht. Deze nevenvoorwaarden dienen steeds uitdrukkelijk te worden vermeld en, zo mogelijk, zelf te worden getoetst.

In plaats van de onbetrouwbaarheidsdrempel α wordt vaak bij de uitslag van een toetsing de overschrijdingskans k opgegeven; dit is de kleinste waarde van α, waarbij in het betrokken geval, nog tot verwerping van H_0 zou zijn overgegaan; anders gezegd: de kleinste α, waarvoor de gevonden waarde der toetsingsgrootheid nog juist in de (bij α behorende) kritieke zône Z ligt. Wordt dus de waarde k opgegeven en werkt men met onbetrouwbaarheidsdrempel α, dan wordt verworpen, indien $k \leq \alpha$ is.

Voor het onderscheid tussen één- en tweezijdige toetsing en de keuze tussen deze twee mogelijkheden vergelijke men bv. de tweede hieronder gegeven litteratuurplaats. Wij moeten hier volstaan met de opmerking, dat éénzijdige toetsing veelal eerder tot verwerping van H_0 leidt, maar dat deze slechts onder bijzondere omstandigheden kan worden toegepast.

Litteratuur:

J.Neyman, First course in probability and statistics, New York, 1950, Chapter 5.

J.Hemelrijk en H.R. van der Vaart, Het gebruik van één- en tweezijdige overschrijdingskansen voor het toetsen van hypothesen, Statistica 4 (1950) p.54-66.

MATHEMATISCH CENTRUM
2de Boerhaavestr.49
A M S T E R D A M O.
Statistische Afdeling
S47 (M14)

Methode der m rangschikkingen [1])

Een duidelijk voorstelling van deze toetsingsmethode verkrijgt men door n elementen te beschouwen, die een bepaald kenmerk, eventueel in verschillende mate, bezitten. dit kenmerk wordt door m waarnemers beoordeeld en ieder van deze waarnemers rangschikt deze n elementen volgens zijn beoordeling naar opklimmende waardering. Op deze wijze ontstaan m rijen van rangschikkingen. We willen nu een maat aangeven voor de overeenstemming tussen deze rangschikkingen, m.a.w. een maat voor de overeenstemming tussen de m beoordelingen. De hypothese H_0, die met deze methode getoetst kan worden, houdt in dat er geen overeenstemming tussen de waarnemers bestaat; precieser gezegd, dat alle rangschikkingen onafhankelijk van elkaar op toevallige wijze zijn ontstaan. Dit is b.v. het geval, als het betrokken kenmerk in werkelijkheid voor alle elementen dezelfde waarde bezit.

We kunnen de afleiding voor de maat van overeenstemming het eenvoudigst geven aan de hand van een voorbeeld.

elementen		A	B	C	D	E	F
rangnummers toegekend door waarnemer	a	5	4	1	6	3	2
	b	2	3	1	5	6	4
	c	4	1	6	3	2	5
	d	4	3	2	5	1	6
		15	11	10	19	12	17

De som van alle rangnummers is $\frac{1}{2}nm(n+1)$. Onder de hypothese H_0 is het theoretische gemiddelde van iedere kolom $\frac{1}{2}m(n+1)$

We beschouwen nu de afwijkingen van dit gemiddelde. In ons voorbeeld is het theoretisch kolomgemiddelde gelijk aan 14. De afwijkingen daarvan zijn

1 −3 −4 5 −2 3

---

1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid

De som der kwadraten van deze afwijkingen noemen wij S.
In ons voorbeeld is S = 64.
Als alle m rangschikkingen gelijk zijn wordt het maximum van S bereikt.
Dit maximum is $\frac{1}{12} m^2(n^3-n)$.
We definiëren nu als <u>coëfficiënt van overeenstemming</u>

$$W = \frac{12\,S}{m^2(n^3-n)}$$

In ons voorbeeld is $W = \frac{12 \times 64}{16 \times 210} = 0{,}229$.

W varieert dus tussen 0 en 1.

De verdeling van <u>S</u> onder de hypothese H_o is exact berekend voor een aantal waarden van n en m [1] , terwijl voor grote m en n benaderingen bekend zijn.

De meeste gebruikelijke benaderingen zijn de volgende.

1°. De χ^2-benadering:

$\underline{\chi}_r^2 = m(n-1)\underline{W} = \frac{12\,\underline{S}}{mn(n+1)}$ heeft voor $m \to \infty$ een χ^2-verdeling met n-1 vrijheidsgraden ([1] pg. 84 [2] pg. 36-37).

2°. De z-benadering:

$\underline{V} = (m-1)\frac{\underline{W}}{1-\underline{W}}$ is bij benadering verdeeld als $\underline{F} = e^{2z}$
(<u>F</u> is de <u>F</u> van Snedecor, <u>z</u> de <u>z</u> van Fisher) met

$\nu_1 = n-1-\frac{2}{m}$

$\nu_2 = (m-1)\nu_1$ vrijheidsgraden ([1] pg. 84 [2] pg.33-36).

Met behulp van de verdelingen van <u>S</u> of <u>W</u> onder de hypothese H_o, kan deze hypothese getoetst worden, waarbij H_o verworpen wordt als <u>W</u> waarden dichtbij 1 (resp. <u>S</u> dichtbij $\frac{1}{12} m^2(n^3-n)$) aanneemt, de kritieke zône is dus van de vorm $W \geqq W_o$ (resp. $S \geqq S_o$).

Het kan voorkomen dat de waarnemers geen onderscheid ontdekken in de mate waarin verschillende elementen het kenmerk bezitten. Ze geven deze elementen dan gelijke rangnummers.

Veronderstel, dat door een waarnemer geen onderscheid wordt gemaakt tussen de elementen, die de rangnummers 3 t/m 6 moeten dragen. Dan wordt als rangnummer van ieder van deze elementen het gemiddelde van de rangnummers $\frac{1}{4}(3 + 4 + 5 + 6) = 4\frac{1}{2}$ gebruikt.

Daar het maximum van <u>S</u> nu verandert, moeten wij een correctie op de formule voor <u>W</u> toepassen. Deze vindt men in [1] (pg.82) en [2] (pg. 28-30). Eveneens veranderen dan de formules voor de χ^2-benadering ([1] pg. 86, [2] pg. 37) en voor de z-benadering ([1] pg. 86 [2] pg. 34), doch deze correcties zijn van weinig betekenis, tenzij het aantal gelijken groot is.

Literatuur: [1] M.G.Kendall, Rank correlation methods, London 1948, Hoofdstuk 6, pag. 80.

Tabel van de verdelingsfunctie van S voor:

n = 3 m = 2 t/m 10
n = 4 m = 2 t/m 6
n = 5 m = 3
op pag. 146-149.

Tabel van de waarden van S, waarvan de overschrijdingskansen onder de hypothese H_o gelijk zijn aan 0,05 of 0,01, berekend met behulp van de z-benadering voor:

n = 3 m = 8,9,10,12,14,15,16,18,20
n = 4 m = 4,5,6,8,10,15,20
n = 5 t/m 7 m = 3,4,5,6,8,10,15,20
op pag. 150.

[2] Ph.van Elteren, Methode der m rangschikkingen, Cursus "Parametervrije Methoden", Hoofdstuk II, Rapport S 59, Mathematisch Centrum (1951), Blz. 18-45.