Klankfrequenties in het Nederlands

Kim Luyckx, Hanne Kloots, Evie Coussé en Steven Gillis[1]

"[H]et middel om hier tot exacte precisie te komen ligt voor de hand: tèl ze!" (Van Ginneken, 1915, p. 68)

1. Inleiding

In nagenoeg alle taalgerelateerde disciplines is er behoefte aan frequentiegegevens. Kennis van woordfrequenties kan o.a. goede diensten bewijzen bij de selectie van woorden voor een curriculum Nederlands als tweede taal. Woorden die frequent gebruikt worden, zullen cursisten doorgaans eerder en vaker nodig hebben dan minder frequente items (bv. Appel & Vermeer, 1994, p. 202). Ook taaltechnologen hebben nood aan informatie over woordfrequentie. De prestaties van een spraakherkenner gaan er namelijk flink op vooruit als het taalmodel frequentiegegevens bevat (bv. Strik, 2001, pp. 284-285). Dat we de factor woordfrequentie ernstig moeten nemen, bleek bijvoorbeeld ook uit een psycholinguïstische studie die de jubilaris uitvoerde samen met enkele Antwerpse collega's. In een onderzoek naar de spelling van homofone werkwoordsvormen (bv. *treed vs. treedt*) was het foutrisico namelijk het hoogst voor de vorm met de laagste frequentie in geschreven taal (Daems, 2000; Sandra, Frisson, & Daems, 1999; Sandra, Daems, & Frisson, 2001).

Wie geïnteresseerd is in de woordfrequenties van het geschreven Nederlands, kan terecht bij Uit den Boogaart (1975) en Baayen, Piepenbrock en Gulikers (1995). Voor het gesproken Nederlands beschikken we over de gegevens van De Jong (1979) en over de woordfrequentielijsten bij het recente *Corpus Gesproken Nederlands*.[2]

Wie echter behoefte heeft aan klankfrequenties, wordt geconfronteerd met een opvallende leemte in de literatuur. Voor het Nederlands zijn namelijk nauwelijks *klank*frequentietellingen uitgevoerd. Dat het erg zinvol en noodzakelijk is om zicht te krijgen op de Nederlandse klankfrequenties, werd nochtans al aangegeven door Van Ginneken (1915), al was die misschien toch net iets te optimistisch wat het tijdschema van de onderneming betreft: "Het compleet uittellen eener taal is werk voor een enkele week, maar het zal misschien eeuwen lang vruchten dragen" (Van Ginneken, 1932, pp. 3-4). Aan Van Ginnekens oproep om – zowel voor het Nederlands als voor andere talen – systematisch klankfrequenties te berekenen, werd helaas niet massaal gevolg gegeven. Bovendien zijn de resultaten van de weinige bestaande tellingen – vanuit hedendaags perspectief – gebaseerd

op relatief kleine corpora en/of op corpora van geschreven taal. In deze bijdrage willen we (ten minste een deel van) deze leemte invullen. We hebben klankfrequenties berekend voor een verzameling van 855.892 woorden uit het *Corpus Gesproken Nederlands*. Voor we onze eigen tellingen bespreken, geven we echter een overzicht van de bestaande studies naar Nederlandse klankfrequenties (paragraaf 2) en laten we zien welke taalgerelateerde wetenschappelijke disciplines nood hebben aan klankfrequenties (paragraaf 3).

2. Bestaande klankfrequentietellingen

De oudste ons bekende publicatie waarin op een systematische manier klankfrequenties werden berekend voor het Nederlands is Huizing & Moolenaar-Bijl (1944). Om de gehoorscherpte van patiënten met een hoorapparaat te controleren, werden zogenaamde 'hoorplankjes' ontwikkeld, woordenlijsten waarin klanken met dezelfde frequentie voorkwamen "als in beschaafd Nederlandsch" (p. 437). Om zo'n 'hoorplankje' te kunnen samenstellen, verzamelden Huizing & Moolenaar-Bijl (1944, p. 435) 10.000 woorden "beschaafd Nederlandsch" en telden hoe vaak elke klank erin voorkwam. Het ging om geschreven taal, "hoewel men zoveel mogelijk gesprekken had genomen en eenvoudig Nederlands" (Eggermont, 1956, p. 221). Om wat voor teksten het precies ging, hoe en door wie ze getranscribeerd werden, komen we helaas niet te weten. Wel werd even gewezen op (de mogelijke impact van) regionale verschillen in de uitspraak. De onderzoekers hebben namelijk de uitspraak in de "Noordelijke provincies" van Nederland als richtsnoer genomen. Een telling met het "Hollandsch" als uitgangspunt zou dus andere cijfers kunnen opleveren (Huizing & Moolenaar-Bijl, 1944, p. 436).
Eggermont (1956) wilde aantonen dat klankfrequenties in gesproken taal niet noodzakelijk overeenkomen met die in schrijftaal. Uit Engelstalig onderzoek was namelijk al langer bekend dat de klankfrequenties in gesproken en geschreven taal kunnen verschillen (bv. French, Carter, & Koenig, 1930). Daarom nam Eggermont in de zomer van 1955 zo'n 10.000 woorden "hedendaagse conversatie" op. De informanten spraken "praktisch allen Algemeen Beschaafd met hoogstens een lichte regionale inslag" en behoorden sociaal gezien tot de zogenaamde 'middengroepen', zoals onderwijzers, ambtenaren, leraren (Eggermont, 1956, p. 222). De geluidsopnamen werden door Eggermont fonetisch getranscribeerd. In totaal bevatten de transcripties 33.900 klanken. Hoe de transcripties tot stand gekomen zijn, is helaas niet (meer) te achterhalen. Het verschil tussen de schrijftaalfrequenties van Huizing en Moolenaar-Bijl (1944) en de spreektaalfrequenties van Eggermont (1956) bleek uiteindelijk trouwens kleiner te zijn dan verwacht.
De volgende telling werd uitgevoerd door Van den Broecke (1976, pp. 68-69). Die gebruikte het krantentaalcorpus van Van Berckel, Brandt Corstius, Mokken, en van Wijngaarden (1965) als basis. Dat corpus bevatte artikelen uit een tiental Nederlandse nationale kranten, verschenen op 19 juni 1956 (ca. 50.000 woorden). Uit het corpus van Van Berckel et al. (1965) selecteerde Van den Broecke de 1000

frequentste woorden. Hij transcribeerde ze fonemisch, met de foneeminventaris uit Cohen, Ebeling, van Holk, en Fokkema (1962) en zijn eigen Standaardnederlands ("my own ABN") als uitgangspunt (Van den Broecke, 1976, p. 69). Hij somt ook nog enkele specifieke regels op die hij gehanteerd heeft tijdens het transcriberen. Daarmee is Van den Broecke (1976) meteen de enige van wie een soort transcriptieprotocol is overgeleverd.

Ook de volgende jaren bleef Van den Broecke gefascineerd door frequentiegegevens en krantencorpora. In de jaren 80 kwam een nieuw corpus tot stand, samengesteld uit redactionele tekst van *De Haarlemse Courant*. Toen Van den Broecke (1988) zijn foneemfrequenties publiceerde, telde dat corpus zo'n 2,3 miljoen woorden. Hoe het corpus precies samengesteld is en hoe de bijbehorende fonematische transcriptie tot stand gekomen is, is helaas niet gedocumenteerd. In de telling van Van den Broecke (1988) wordt wel voor het eerst expliciet een onderscheid gemaakt tussen *token*frequenties (berekend op basis van *alle* woorden uit het corpus) en *type*frequenties (berekend op basis van het aantal *verschillende* woorden uit het corpus). Dat onderscheid is belangrijk: klankfrequentie is immers onlosmakelijk verbonden met woordfrequentie. Een hoge tokenfrequentie kan er bijvoorbeeld op wijzen dat een klank in veel verschillende woorden voorkomt, maar kan even goed betekenen dat de klank slechts voorkomt in een handvol hoogfrequente woorden (zie bv. ook Dewey, 1923; French et al., 1930; Denes, 1963; Mines, Hanson, & Shoup, 1978).

Bovenstaand literatuuroverzicht laat zien dat het onderzoek naar Nederlandse klankfrequenties enkele belangrijke hiaten vertoont. Vaak komen we bijvoorbeeld maar weinig te weten over het gebruikte corpus en de transcriptieprocedure. De meeste tellingen waren bovendien gebaseerd op geschreven taal, en dat terwijl klanken toch per definitie akoestische eenheden zijn. Gerber en Vertin (1969) toonden zelfs aan dat de klankfrequenties voor gesproken Brits- en Amerikaans-Engels (= twee variëteiten van dezelfde taal) beter met elkaar correleren dan de frequenties voor gesproken en geschreven Amerikaans-Engels (= een enkele taalvariëteit). De enige klankfrequentietelling met gesproken Nederlands als basis (Eggermont, 1956) is intussen al 50 jaar oud.[3] Bovendien is ze gebaseerd op een – vanuit hedendaags perspectief – relatief beperkt corpus. Ten slotte valt ook op dat het Nederlandstalige onderzoek zich tot nu toe beperkt heeft tot de studie van de klankfrequenties in Nederland. Het zou interessant zijn om ook Vlaams materiaal te onderzoeken. Bij onze eigen klankfrequentietelling proberen we deze hiaten zo veel mogelijk in te vullen.

3. Bruikbaarheid

Ons onderzoek komt tegemoet aan een dringende behoefte in diverse klankgerelateerde wetenschappelijke disciplines. Zo zou informatie over klankfrequenties in het Nederlands van volwassen sprekers bijvoorbeeld kunnen helpen bij de studie van de verwervingsvolgorde van klanken bij jonge kinderen (Gillis, 2000,

p. 148 e.v.). Een hypothese zou kunnen zijn dat kinderen het eerst die klanken verwerven die ze het vaakst horen. Ook spraakherkenners zouden zeker profiteren van kennis over klankfrequenties: bij twijfel tussen verschillende varianten zou de computer dan de klank kunnen suggereren die het vaakst voorkomt in alledaags taalgebruik. Verder zijn de frequenties bijvoorbeeld ook interessant voor audiologen. Als die onderzoek doen naar verstaanbaarheid bij (o.a.) slechthorenden, maken ze vaak gebruik van fonetisch gebalanceerde ("phonetically balanced") woorden- of zinnenlijsten. De frequentie waarmee een klank voorkomt in zo'n lijst, zou idealiter een afspiegeling moeten zijn van de frequentie in reëel taalgebruik (Bosman, 1988, p. 310-311).
Het zou ook interessant zijn om de klankfrequenties in beklemtoonde en onbeklemtoonde syllaben te vergelijken. Zo'n vergelijking zou (o.a.) empirische evidentie kunnen opleveren voor de taalhistorische intuïtie dat veel onbeklemtoonde volle vocalen in de loop der eeuwen hun kleur hebben verloren en als het ware zijn "afgesleten" tot sjwa (bv. Van Loey, 1970). Ook in de fonologische literatuur vinden we de intuïtie dat sjwa vooral voorkomt in onbeklemtoonde syllaben (bv. Booij, 1995). Een klankfrequentietelling kan laten zien hoe volle vocalen en sjwa zich verhouden in het hedendaagse gesproken Nederlands.
Bij een klankfrequentietelling kunnen we ons natuurlijk ook toespitsen op de syllabestructuur. Informatie over de frequentie van de respectieve syllabetypes zou o.a. goede diensten kunnen bewijzen bij de studie van zowel taaltypologie (bv. is er een verband tussen de grootte van de klankinventaris en de grootte van de syllabe-inventaris in een taal?), eerstetaalverwerving (bv. verwerven kinderen eerst de syllabestructuren die ze het vaakst horen?, cf. Levelt, Schiller, & Levelt, 1999) en woordherkenning (bv. draagt een hoge syllabefrequentie ertoe bij dat een woord sneller herkend wordt?; zie ook Schiller, Meyer, Baayen, & Levelt, 1996, pp. 9-10). Informatie over de frequentie van syllabestructuren is zeker ook bruikbaar bij psycholinguïstisch onderzoek naar de zogenaamde *Mental Syllabary* (Cholin, Levelt, & Schiller, 2006).
In deze bijdrage kunnen we uiteraard niet al deze onderzoeksvragen en -domeinen behandelen. De vragen illustreren echter wel dat een klankfrequentietelling, ook al is er dan misschien meer dan "een enkele week" (Van Ginneken, 1932, p. 4) voor nodig, een bijzonder zinvolle en welkome onderneming is.

4. Methode

Onze klankfrequentietelling is gebaseerd op het *Corpus Gesproken Nederlands* (CGN), een verzameling van circa 9 miljoen woorden gesproken Standaardnederlands die tot stand kwam in de periode 1998-2004. Twee derde van de spraak is afkomstig uit Nederland, een derde uit Vlaanderen. Het CGN bevat verschillende types van spraak, bv. telefoondialogen, discussies, lessen en spontane conversaties. Voor meer informatie over de totstandkoming en de samenstelling het CGN verwijzen we naar de URL's in voetnoot 2.

Bij ons onderzoek concentreren we ons op het deel van het CGN waarvoor een geverifieerde brede fonetische transcriptie bestaat. Zo'n transcriptie is beschikbaar voor ongeveer een tiende van het totale corpus. Onze tellingen zijn gebaseerd op een (sub)corpus van 855.892 woorden. We hebben namelijk één type spraak buiten beschouwing gelaten: de voorgelezen teksten uit de zogenaamde 'blindenbibliotheek' (= CGN-component 'o'). Deze teksten zijn immers oorspronkelijk bedoeld om (in stilte) *gelezen* te worden, niet om te worden uitgesproken.

De brede fonetische transcripties van het CGN kwamen als volgt tot stand. Op basis van de orthografische transcripties (d.w.z. een grafemische transcriptie, gedocumenteerd in Goedertier & Goddijn, 2000) werd eerst een fonetische transcriptie gegenereerd via automatische grafeem-naar-foneemomzetting. Aan de basis daarvan lagen bestaande uitspraaklexica: in Nederland CELEX (Baayen et al., 1995), in Vlaanderen FONILEX (Mertens & Vercammen, 1998). Vervolgens werd de automatische transcriptie geverifieerd door menselijke transcribenten. Hoe deze verificatie precies in zijn werk ging, is gedocumenteerd in Gillis (2001) en Gillis, Depoorter, en Goddijn (2003).

Heel wat afspraken uit het CGN-transcriptieprotocol hebben betrekking op klankverschijnselen die typisch zijn voor gebonden spraak. Verbindingsklanken en andere klankinserties (bv. *duet* > [dywɛt], *melk* > [mɛləkˈ]) werden weergegeven, net als assimilatie (bv. *onbepaald* > [ombəpalt]). Geminaten op een woordgrens werden weergegeven en gemarkeerd als twee afzonderlijke segmenten (bv. *om meer* > [ɔm mer]). Graduele processen zoals bijvoorbeeld de verstemlozing van fricatieven en de monoftongering van diftongen werden daarentegen niet getranscribeerd. Ook woorden die in de orthografische transcripties gemerkt waren met een asterisk – o.a. dialectwoorden (*d), vreemde woorden (*v) en afgebroken woorden (*a), cf. Goedertier & Goddijn (2000) – werden zo goed mogelijk fonetisch getranscribeerd. Meer details zijn te vinden in Gillis (2001) en Gillis et al. (2003).

In deze bijdrage presenteren we de resultaten van vier tellingen. Voor een eerste, algemeen overzicht delen we de consonanten in volgens articulatiewijze en de vocalen volgens vocaalkwaliteit (telling 1). Vervolgens presenteren we de frequenties van alle klanken afzonderlijk (telling 2). Omdat enkele Engelstalige studies (bv. Fry, 1947; Denes, 1963) wijzen op (potentiële) verschillen tussen variëteiten van dezelfde taal, i.c. Brits- en Amerikaans-Engels, leek het ons ook interessant om meteen ook de frequenties in het Belgisch- en het Nederlands-Nederlands te vergelijken. Vervolgens focussen we op de syllabestructuur van monosyllabische woorden (telling 3). Ten slotte vergelijken we de verhouding tussen volle vocalen en sjwa in beklemtoonde en onbeklemtoonde syllaben (telling 4). Telkens wordt een onderscheid gemaakt tussen type- en tokenfrequentie. Technische noot: bij woorden met diverse uitspraakvarianten is de typefrequentie gebaseerd op de variant die het frequentst voorkwam. Zo werd *melk* soms uitgesproken als [mɛlk], soms als [mɛləkˈ]. De tweede variant kwam het vaakst voor, dus die werd als basis genomen bij de berekening van de typefrequenties.

Uiteraard waren we bij onze tellingen gebonden aan de symboolset van de CGN-transcripties. Het transcriptieprotocol bevat bijvoorbeeld geen apart symbool voor tongpunt-r en huig-r. Wie wil weten welke r-variant het frequentst voorkomt in het CGN, zal alsnog zelf de geluidsopnamen moeten raadplegen. Verder zijn de brede fonetische transcripties bij het CGN niet gesyllabificeerd noch is er klemtoon aangeduid. Voor sommige tellingen kan dat problemen opleveren, bijvoorbeeld bij een studie van de syllabestructuur of als we de klankfrequenties in beklemtoonde en onbeklemtoonde syllaben willen vergelijken.
Dit probleem werd als volgt aangepakt. Voor telling 3 en 4 werden de brede fonetische transcripties bij het CGN opgelijnd met (d.w.z. gekoppeld aan) de meest formele uitspraakvariant uit FONILEX (Mertens & Vercammen, 1998). Hoe de oplijning precies gebeurde, is toegelicht in Coussé, Gillis en Kloots (2007). De syllabificering gebeurde met behulp van een computerscript, gebaseerd op het *Core Syllabification Principle* (Clements, 1990). Dit principe maakt gebruik van kennis over de maximale onset en sonoriteit (zie ook Booij, 1995). Ambisyllabische consonanten werden bij de tweede syllabe gerekend (bv. *appel* > [ɑ-pəl]).[4] Wat de factor klemtoon betreft, deden we eveneens een beroep op FONILEX. 'Klemtoon' verwijst hier dus naar de lexicale klemtoon. Met zinsaccent of intonatie werd geen rekening gehouden. Voor alle duidelijkheid: we namen alleen de klemtoon over uit FONILEX, niet de uitspraak zelf. Als FONILEX twee klemtoonvarianten bevatte (bv. *bi'kini vs. 'bikini*), werd automatisch en consequent de eerste variant geselecteerd.[5]
Niet alle woorden uit het fonetisch getranscribeerde deel van het *Corpus Gesproken Nederlands* komen ook voor in FONILEX. Het CGN bevat flink wat zeldzame woorden, maar bijvoorbeeld ook eigennamen, dialectwoorden en vreemde woorden. Daarnaast bevat het corpus heel wat uitingen die typisch zijn voor gesproken taal, bijvoorbeeld versprekingen, afgebroken woorden en sprekersgeluiden (bv. kuchen, lachen). Vormen die niet voorkomen in FONILEX zijn bij telling 3 en 4 buiten beschouwing gelaten. Telling 3 en 4 zijn dus gebaseerd op een iets kleiner (deel van het) corpus (772.108 woorden) dan telling 1 en 2 (855.892 woorden).

5. Resultaten

Voor een eerste, ruwe verkenning van de data hebben we de klanken gegroepeerd in een aantal traditionele categorieën (Tabel 1). De consonanten zijn ingedeeld volgens articulatiewijze. Daarbij onderscheiden we plosieven ([p], [b], [t], [d], [k], [g]), fricatieven ([f], [v], [s], [z], [x], [ɣ], [ʃ], [ʒ], [h]), nasalen ([m], [n], [ŋ], [ɲ]), liquida ([l], [r]) en glides ([w], [j]). De vocalen werden gecategoriseerd volgens vocaalkwaliteit: (fonologisch) lange[6] vocalen ([a], [e], [i], [o], [u], [y], [ø]), (fonologisch) korte vocalen ([ɑ], [ɛ], [ɪ], [ɔ], [ʏ]), sjwa, diftongen ([ɛi], [ɔu], [œy]), de zogenaamde 'leenvocalen' ([ɛː], [ɔː], [œː]) en nasale vocalen ([ɑ̃], [ɛ̃], [ɔ̃], [œ̃]).
Bij de consonanten zijn de plosieven en de fricatieven het frequentst. Ook de liquida en de nasalen zijn nog erg goed vertegenwoordigd, zeker als we rekening houden met het feit dat deze categorieën slechts uit twee (liquida) respectievelijk

vier (nasalen) klanken bestaan. De glides zijn het minst goed vertegenwoordigd. Bij de vocalen scoren drie categorieën nagenoeg even hoog: lange vocalen, korte vocalen en sjwa. De diftongen zijn veel minder frequent. Leenvocalen en nasale vocalen maken slechts een klein deel uit van het corpus. Die laatste observatie sluit aan bij onze ervaring dat leenvocalen en nasale vocalen doorgaans weinig aandacht krijgen in beschrijvingen van het Nederlandse klanksysteem. Wellicht gaat de aandacht van taalbeschrijvers vooral uit naar klanken en klankverschijnselen die frequent voorkomen.

Tabel 1

Type- en tokenfrequentie van consonanten (ingedeeld volgens articulatiewijze) en vocalen (ingedeeld volgens vocaalkwaliteit)

	Typefrequentie		Tokenfrequentie	
	n	**%**	**n**	**%**
plosieven	57.035	20,43%	585.999	19,74%
fricatieven	46.129	16,53%	426.097	14,36%
liquida	34.375	12,31%	249.275	8,40%
nasalen	26.752	9,58%	361.742	12,19%
glides	8.209	2,94%	146.344	4,93%
lange vocalen	35.380	12,67%	365.887	12,33%
korte vocalen	33.090	11,85%	428.832	14,45%
sjwa	31.829	11,40%	341.260	11,50%
diftongen	5.955	2,13%	61.638	2,08%
leenvocalen	248	0,09%	615	0,02%
nasale vocalen	136	0,05%	531	0,02%
TOTAAL	**279.138**		**2.968.220**	

In Tabel 2 bekijken we de klankfrequenties meer in detail (telling 2). Links staan de cijfers voor het volledige onderzochte corpus ('Totaal'), rechts daarvan de cijfers voor Nederland en Vlaanderen afzonderlijk. Opvallend is dat sjwa in alle tellingen het hoogst scoort. Een kleine kanttekening bij de frequentie voor [n]: de automatische grafeem-foneemomzetter paste geen n-deletie (bv. *kopen* > [kopə]) toe aan het woordeinde. Concreter geformuleerd: woorden die eindigen op <-n> bevatten in de automatische transcriptie consequent een eind-n. Bij de verificatie door menselijke transcribenten werd een aantal van die eind-n'en geschrapt, maar het valt niet uit te sluiten dat de geverifieerde transcriptie uiteindelijk toch nog een aantal eind-n'en bevat die in werkelijkheid niet zijn uitgesproken, zeker omdat de transcribenten expliciet de opdracht kregen om bij twijfel de automatisch gegenereerde klank te laten staan (Gillis, 2001). De frequenties voor [n] moeten dus met de nodige voorzichtigheid geïnterpreteerd worden (zie ook Van Son & Pols, 2001).

Tabel 2

Type- en tokenfrequentie voor Nederland en Vlaanderen tezamen ('Totaal') en voor Nederland en Vlaanderen afzonderlijk

	Totaal				Nederland				Vlaanderen			
	Typefrequentie		Tokenfrequentie		Typefrequentie		Tokenfrequentie		Typefrequentie		Tokenfrequentie	
Klank	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
ə	31.829	11,40%	341.260	11,50%	24.796	11,76%	239.224	11,74%	13.962	11,35%	102.036	10,97%
t	22.760	8,15%	230.181	7,75%	17.397	8,25%	158.544	7,78%	9.955	8,09%	71.637	7,70%
r	21.116	7,56%	145.178	4,89%	15.417	7,31%	93.923	4,61%	9.834	7,99%	51.255	5,51%
s	17.550	6,29%	127.590	4,30%	13.355	6,34%	89.347	4,38%	7.195	5,85%	38.243	4,11%
n	15.253	5,46%	224.331	7,56%	10.981	5,21%	148.964	7,31%	7.650	6,22%	75.367	8,10%
l	13.259	4,75%	104.097	3,51%	10.035	4,76%	70.593	3,46%	5.818	4,73%	33.504	3,60%
k	11.359	4,07%	99.065	3,34%	8.557	4,06%	67.692	3,32%	4.899	3,98%	31.373	3,37%
ɑ	10.167	3,64%	147.076	4,96%	7.197	3,41%	95.118	4,67%	4.923	4,00%	51.958	5,58%
d	9.536	3,42%	162.758	5,48%	7.289	3,46%	112.061	5,50%	4.154	3,38%	50.697	5,45%
i	8.142	2,92%	68.329	2,30%	5.753	2,73%	45.366	2,23%	3.803	3,09%	22.963	2,47%
e	7.887	2,83%	80.069	2,70%	5.803	2,75%	54.948	2,70%	3.652	2,97%	25.121	2,70%
a	7.776	2,79%	105.170	3,54%	6.106	2,90%	73.417	3,60%	3.138	2,55%	31.753	3,41%
χ	7.723	2,77%	73.415	2,47%	6.806	3,23%	56.079	2,75%	2.559	2,08%	17.336	1,86%
m	7.625	2,73%	99.312	3,35%	5.756	2,73%	68.918	3,38%	3.197	2,60%	30.394	3,27%
ɛ	7.476	2,68%	110.294	3,72%	5.657	2,68%	75.970	3,73%	3.197	2,60%	34.324	3,69%
p	6.878	2,46%	39.241	1,32%	5.219	2,48%	27.361	1,34%	2.894	2,35%	11.880	1,28%
ɔ	6.846	2,45%	63.924	2,15%	5.165	2,45%	42.906	2,11%	3.014	2,45%	21.018	2,26%
o	6.642	2,38%	65.397	2,20%	5.179	2,46%	46.592	2,29%	2.738	2,22%	18.805	2,02%
ɪ	6.469	2,32%	88.699	2,99%	5.157	2,45%	62.932	3,09%	2.496	2,03%	25.767	2,77%
b	5.952	2,13%	43.816	1,48%	4.484	2,13%	30.063	1,48%	2.556	2,08%	13.753	1,48%
f	5.831	2,09%	53.192	1,79%	4.998	2,37%	41.270	2,03%	1.921	1,56%	11.922	1,28%
j	4.277	1,53%	76.133	2,56%	3.393	1,61%	54.310	2,67%	1.552	1,26%	21.823	2,35%
v	4.031	1,44%	37.373	1,26%	2.397	1,14%	20.567	1,01%	2.674	2,17%	16.806	1,81%

	Totaal				Nederland				Vlaanderen			
	Typefrequentie		Tokenfrequentie		Typefrequentie		Tokenfrequentie		Typefrequentie		Tokenfrequentie	
Klank	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*
w	3.932	1,41%	70.211	2,37%	3.054	1,45%	50.681	2,49%	1.659	1,35%	19.530	2,10%
ŋ	3.742	1,34%	34.652	1,17%	2.901	1,38%	24.718	1,21%	1.540	1,25%	9.934	1,07%
ɣ	3.418	1,22%	29.289	0,99%	1.849	0,88%	14.322	0,70%	2.518	2,05%	14.967	1,61%
ɛi	3.417	1,22%	40.684	1,37%	2.592	1,23%	26.971	1,32%	1.490	1,21%	13.713	1,47%
z	3.312	1,19%	45.919	1,55%	2.240	1,06%	26.980	1,32%	1.766	1,44%	18.939	2,04%
h	2.765	0,99%	39.153	1,32%	2.225	1,06%	30.678	1,51%	977	0,79%	8.475	0,91%
u	2.590	0,93%	27.823	0,94%	1.943	0,92%	19.316	0,95%	1.144	0,93%	8.507	0,91%
ʏ	2.132	0,76%	18.839	0,63%	1.753	0,83%	13.542	0,66%	703	0,57%	5.297	0,57%
œy	1.651	0,59%	9.662	0,33%	1.309	0,62%	6.753	0,33%	706	0,57%	2.909	0,31%
y	1.628	0,58%	14.976	0,50%	1.059	0,50%	8.738	0,43%	880	0,72%	6.238	0,67%
ʃ	1.236	0,44%	19.027	0,64%	882	0,42%	13.933	0,68%	574	0,47%	5.094	0,55%
ɔu	887	0,32%	11.292	0,38%	665	0,32%	9.072	0,45%	387	0,31%	2.220	0,24%
ø	715	0,26%	4.123	0,14%	527	0,25%	3.051	0,15%	293	0,24%	1.072	0,12%
ɡ	550	0,20%	10.938	0,37%	455	0,22%	8.426	0,41%	184	0,15%	2.512	0,27%
ʒ	263	0,09%	1.139	0,04%	152	0,07%	769	0,04%	153	0,12%	370	0,04%
ɛ:	174	0,06%	434	0,01%	101	0,05%	222	0,01%	99	0,08%	212	0,02%
ɲ	132	0,05%	3.447	0,12%	95	0,05%	3.079	0,15%	59	0,05%	368	0,04%
ɑ̃	68	0,02%	265	0,01%	27	0,01%	72	0,00%	59	0,05%	193	0,02%
œ:	37	0,01%	73	0,00%	19	0,01%	29	0,00%	24	0,02%	44	0,00%
ɔ:	37	0,01%	108	0,00%	31	0,01%	89	0,00%	15	0,01%	19	0,00%
ɔ̃	33	0,01%	108	0,00%	7	0,00%	21	0,00%	28	0,02%	87	0,01%
ɛ̃	31	0,01%	151	0,01%	13	0,01%	29	0,00%	20	0,02%	122	0,01%
œ̃	4	0,00%	7	0,00%	1	0,00%	4	0,00%	3	0,00%	3	0,00%
TOTAAL	**279.138**		**2.968.220**		**210.797**		**2.037.660**		**123.062**		**930.560**	

Vervolgens focussen we op de syllabeopbouw van de monosyllabische woorden (telling 3). Tabel 3 bevat een top 5 van de meest frequente syllabestructuren in het fonetisch getranscribeerde deel van het CGN.

Tabel 3
Top 5 van de meest frequente syllabestructuren bij monosyllabische woorden (V = vocaal, C = consonant)

Typefrequentie			Tokenfrequentie		
	n	**%**		**n**	**%**
1. CVC	728	32,75%	1. CVC	179.400	32,00%
2. CVCC	456	20,51%	2. CV	160.445	28,62%
3. CCVC	398	17,90%	3. VC	133.347	23,78%
4. CCVCC	245	11,02%	4. CVCC	24.768	4,42%
5. CV	89	4,00%	5. V	17.259	3,08%
overige	307	13,82%	overige	45.430	8,10%
TOTAAL	**2.223**		**TOTAAL**	**560.649**	

De structuren uit onze top 5 scoorden ook hoog in de (schrijftaal)tellingen van Schiller et al. (1996). Veel ervan worden ook vrij vroeg verworven door kinderen (Levelt et al., 1999). Verder valt nog op dat de structuren uit de top 5 in totaal 86,18% (types) resp. 91,90% (tokens) van het onderzochte corpus vertegenwoordigen. Daarnaast zijn er natuurlijk nog heel wat andere structuren mogelijk (bv. V, VCC, CCCVC), maar uiteindelijk vormen die (tezamen) slechts een kleine minderheid van het onderzochte corpus. Opmerkelijk is ten slotte ook nog dat de top 5 voor types en tokens niet overeenkomt, wat illustreert dat het zinvol is om afzonderlijke type- en tokenfrequenties te berekenen.
Ten slotte laten we nog even zien waarom het interessant kan zijn om bij de berekening van klankfrequenties ook een onderscheid te maken tussen beklemtoonde en onbeklemtoonde syllaben (telling 4). We focussen hier op de percentages voor lange en korte vocalen, diftongen en sjwa (Tabel 4). Met één oogopslag is duidelijk dat sjwa, de meest frequente klank uit Tabel 2, vooral voorkomt in onbeklemtoonde syllaben. Als we kijken naar de tokenfrequenties, blijkt zelfs bijna 30% van de onbeklemtoonde klanken een sjwa te zijn. Tabel 4 levert dus duidelijk empirische evidentie op voor de taalhistorische en de fonologische intuïtie dat sjwa's typisch zijn voor onbeklemtoonde syllaben (bv. Van Loey, 1970; Booij, 1995).

Tabel 4
Type- en tokenfrequentie in beklemtoonde syllaben ('+ klem') en onbeklemtoonde syllaben ('- klem') bij lange en korte vocalen, diftongen en sjwa

	Typefrequentie				Tokenfrequentie			
	+ klem		- klem		+ klem		- klem	
	n	%	n	%	n	%	n	%
lange vocalen	10.673	16,93%	9.362	9,07%	279.830	16,03%	46.090	5,62%
korte vocalen	9.346	14,82%	9.352	9,06%	314.098	17,99%	51.108	6,23%
diftongen	2.443	3,87%	1.326	1,28%	51.622	2,96%	5.174	0,63%
sjwa	171	0,27%	20.699	20,05%	42.146	2,41%	240.516	29,30%
overige vocalen	125	0,20%	39	0,04%	509	0,03%	153	0,02%
consonanten	40.292	63,90%	62.474	60,51%	1.057.332	60,57%	477.735	58,21%
TOTAAL	**63.050**		**103.252**		**1.745.537**		**820.776**	

6. Besluit

Doelstelling van deze bijdrage was de berekening van klankfrequenties in het *Corpus Gesproken Nederlands* (CGN). Daarmee vullen we een opvallende leemte in de literatuur: de recentste klankfrequentietelling, gebaseerd op gesproken Nederlands, dateert namelijk uit de jaren 50 van de vorige eeuw. De cijfers in deze bijdrage zijn interessant voor onderzoekers uit diverse taalkundige disciplines. Ook in de psycholinguïstiek, een domein dat de jubilaris na aan het hart ligt, is kennis van klankfrequenties meer dan welkom. Zo kunnen de cijfers bijvoorbeeld gebruikt worden in onderzoek naar de verwervingsvolgorde van klanken en naar processen van woordherkenning. Ook leveren de tellingen informatie op over het klanksysteem van het Nederlands. Ten slotte kan kennis over de numerieke verhoudingen tussen klanken ook licht werpen op de (opbouw van de) Nederlandse woordvorm, al is en blijft een woord natuurlijk altijd veel meer dan louter een combinatie van klanken:

> *Er zal wel geen zoöloog zijn die meent dat de fuut is opgebouwd uit twee zwemvoeten plus nog een aantal onderdelen; de opvatting van sommige fonologen dat woorden zijn of worden gevormd 'by combining the phonemes' is niet de onze. (Bakker, 1971, p. 14)*

Noten

1 De eerste drie auteurs zijn respectievelijk wetenschappelijk medewerker, Postdoctoraal Onderzoeker en Aspirant van het Fonds voor Wetenschappelijk Onderzoek – Vlaanderen.

2 Informatie over het *Corpus Gesproken Nederlands i*s te vinden via de CGN-website <http://lands.let.kun.nl/cgn/home.htm> en via de TST-centrale <http://www.tst.inl.nl> (onder "Producten").

3 Dat er de laatste jaren geen nieuwe klankfrequentietellingen gepubliceerd zijn, betekent uiteraard niet dat de interesse voor dit soort gegevens is afgenomen. Zo berekende V. Kuperman (Max Planck Instituut voor Psycholinguïstiek - Nijmegen) klankfrequenties voor de spontane spraak uit het IFA-corpus (pers. comm. 03/07/2006; zie ook Van Son & Pols, 2001).

4 Het script hield dus geen rekening met de taalspecifieke regel dat Nederlandse syllaben niet op een fonologisch korte vocaal zouden kunnen eindigen. Voor een uitvoerige bespreking van deze problematiek en empirische (tegen)evidentie, zie Gillis en De Schutter (1996), Schiller, Meyer en Levelt (1997) en Kloots (2005).

5 Het ware natuurlijk mooier geweest als we de klemtoonvariant hadden kunnen selecteren die de sprekers effectief gebruikt hebben. De grootte van het corpus maakte het echter onmogelijk om de geluidsopnamen een voor een te beluisteren en desgevallend de plaats van de klemtoon manueel aan te passen.

6 De termen 'lang' en 'kort' fungeren hier louter als label voor (abstracte) fonologische categorieën (zie bv. Booij, 1995). Ze zeggen niet noodzakelijk iets over de duur van de betreffende klinkers.

Referenties

Appel, R., & Vermeer, A. (1994). *Tweede-taalverwerving en tweede-taalonderwijs*. Bussum: Dick Coutinho.

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database Consortium (CD-ROM). Linguistic data.* (2de release) Philadelphia: University of Pennsylvania.

Bakker, J. (1971). *Constant en variabel. De fonematische structuur van de Nederlandse woordvorm*. Proefschrift Universiteit van Amsterdam. Asten: Schriks' drukkerij.

Booij, G. (1995). *The phonology of Dutch*. Oxford: Clarendon Press.

Bosman, A. (1988). Spraakverstaan meten. In M. van den Broecke (Ed.), *Ter Sprake. Spraak als betekenisvol geluid in 36 thematische hoofdstukken* (pp. 307-313). Dordrecht/Providence RI: Foris Publications.

Cholin, J., Levelt, W., & Schiller, N. (2006). Effects of syllable frequency in speech production. *Cognition*, 99, 205-235.

Clements, G. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. Beckman (Eds.), *Papers in laboratory phonology I. Between the grammar and physics of speech* (pp. 283-333). Cambridge: Cambridge University Press.

Cohen, A., Ebeling, C., van Holk, A., & Fokkema, K. (1962). *Fonologie van het Nederlands en het Fries*. Den Haag: Nijhoff.

Coussé, E., Gillis, S., & Kloots, H. (2007). Verkort, verdoft, verdwenen. Vocaalreductie in het Corpus Gesproken Nederlands. *Nederlandse Taalkunde,* 7, 109-138.

Daems, F. (2000). Walgelijk en ongerijmd: Over de leerbaarheid van de werkwoordspelling. In S. Gillis, J. Nuyts & J. Taeldeman (Eds.), *Met taal om de tuin geleid. Een bundel opstellen voor Georges De Schutter ter gelegenheid van zijn pre-emeritaat*

(pp. 95-113). Antwerpen: Universitaire Instelling Antwerpen.

de Jong, E. (Ed.) (1979). *Spreektaal. Woordfrequenties in gesproken Nederlands*. Utrecht: Bohn, Scheltema & Holkema.

Denes, P. (1963). On the statistics of spoken English. *The journal of the Acoustical Society of America,* 35, 892-904.

Dewey, G. (1923). R*elativ [sic] frequency of English speech sounds*. (licht aangepaste herdruk, 1950) Cambridge: Harvard University Press.

Eggermont, J. (1956). De klankfrequentie in het hedendaagse gesproken Nederlands. *De nieuwe taalgids,* 49, 221-223.

French, N., Carter, C., & Koenig, W. (1930). The words and sounds of telephone conversations. *Bell System technical journal,* 9, 290-324.

Fry, D. (1947). The frequency of occurrence of speech sounds in Southern English. *Archives néerlandaises de phonétique expérimentale,* 20, 103-106.

Gerber, S., & Vertin, S. (1969). Comparative frequency counts of English phonemes. *Phonetica*, 19, 133-141.

Gillis, S., & De Schutter, G. (1996). Intuitive syllabification: universals and language specific constraints. *Journal of child language,* 23, 487-514.

Gillis, S. (2000). Fonologische ontwikkeling. In S. Gillis & A. Schaerlaekens (Eds.), *Kindertaalverwerving. Een handboek voor het Nederlands* (pp. 131-184). Groningen: Martinus Nijhoff.

Gillis, S. (2001). P*rotocol voor brede fonetische transcriptie*. Intern CGN-document, beschikbaar via de Centrale voor Taal- en Spraaktechnologie <http://www.tst.inl.nl> en via de oorspronkelijke CGN-website <http://lands.let.kun.nl/cgn/doc_Dutch/topics/version_1.0/annot/phonetics/info.htm>.

Gillis, S., Depoorter, G., & Goddijn, S. (2003). *Phonetic transcription.* Manuscript.

Goedertier, W., & Goddijn, S. (2000). *Protocol voor orthografische transcriptie*. Intern CGN-document, beschikbaar via de TST-centrale <http://www.tst.inl.nl> en via de oorspronkelijke CGN-website <http://lands.let.kun.nl/cgn/doc_Dutch/topics/version_1.0/annot/orthography/info.htm#protocol>

Huizing, H., & Moolenaar-Bijl, A. (1944). De beteekenis der klankfrequentie in het Nederlandsch voor de oorheelkunde. *Nederlandsch tijdschrift voor geneeskunde,* 88, 435-437.

Kloots, H. (2005). *Vocaalreductie in het Standaardnederlands in Vlaanderen en Nederland*. Proefschrift Universiteit Antwerpen.

Levelt, C., Schiller, N., & Levelt, W. (1999). The acquisition of syllable types. *Language acquisition,* 8, 237-264.

Mertens, P., & Vercammen, F. (1998). *Fonilex manual. Fonilex: a pronunciation*

*database of Dutch in Flanders. V*ersie 1.0b, <http://bach.arts.kuleuven.ac.be/fonilex>.

Mines, M., Hanson, B., & Shoup, J. (1978). Frequency of occurrence of phonemes in conversational English. *Language and speech,* 21, 221-241.

Sandra, D., Daems, F., & Frisson, S. (2001). Zo helder en toch zoveel fouten! Wat leren we uit psycholinguïstisch onderzoek naar werkwoordfouten bij ervaren spellers? *Vonk*, 30, 3-20.

Sandra, D., Frisson, S., & Daems, F. (1999). Why simple verb forms can be so difficult to spell: The influence of homophone frequency and distance in Dutch. *Brain and language,* 68, 277-283.

Schiller, N., Meyer, A., Baayen, H., & Levelt, W. (1996). A comparison of lexeme and speech syllables in Dutch. *Journal of quantitative linguistics,* 3, 8-28.

Schiller, N., Meyer, A., & Levelt, W. (1997). The syllabic structure of spoken words: evidence from the syllabification of intervocalic consonants. *Language and speech,* 40, 103-140.

Strik, H. (2001). 'Dat heb ik helemaal niet gezegd!'. De prestaties van de spraakherkenner. *Onze taal,* 11, 284-286.

Uit den Boogaart, P. (Ed.; 1975). *Woordfrequenties in geschreven en gesproken Nederlands*. Utrecht: Oosthoek, Scheltema & Holkema.

van Berckel, J., Brandt Corstius, H., Mokken, R., & van Wijngaarden, A. (1965). *Formal properties of newspaper Dutch*. Amsterdam: Mathematisch Centrum Amsterdam.

van den Broecke, M. (1976). *Hierarchies and rank orders in distinctive features*. Assen/Amsterdam: Van Gorcum.

van den Broecke, M. (1988). Frequenties van letters, lettergrepen, woorden en fonemen in het Nederlands. In M. van den Broecke (Ed.), *Ter Sprake. Spraak als betekenisvol geluid in 36 thematische hoofdstukken* (pp. 400-407). Dordrecht/Providence RI: Foris Publications.

van Ginneken, J. (1915). De statistiek en de taalwetenschap. *De nieuwe taalgids,* 9, 65-95.

van Ginneken, J. (1932). *De ontwikkelingsgeschiedenis van de systemen der menschelijke taalklanken.* Amsterdam: Noord-Hollandsche Uitgevers-Maatschappij.

Van Loey, A. (1970). *Schönfelds historische grammatica van het Nederlands. Klankleer, vormleer, woordvorming*. (8ste druk) Zutphen: W.J. Thieme & Cie.

van Son, R., & Pols, L. (2001). Structure and access of the open source IFA-corpus. *In Proceedings of the IRCS workshop on linguistic databases* (pp. 245-253). Philadelphia: University of Pennsylvania.