Dat gebeurd mei niet:
Computationele modellen voor verwarbare homofonen

Walter Daelemans en Antal van den Bosch

1. Inleiding

Van de lagere school tot de universiteit, van e-mail tot krant worden lezers geconfronteerd met verwarring in teksten tussen gelijk uitgesproken maar verschillend geschreven woorden (homofonen). Een deel van deze chaos komt voort uit toevalligheden; de maand *mei* en het bezittelijk voornaamwoord *mij* hebben een verschillende afkomst maar klinken toevallig hetzelfde. Deze gevallen moeten uit het hoofd geleerd worden. Een ander deel van de verwarring is een gevolg van conventionele spellingregels, zoals de regels voor flectie, waarvan de meest beruchte de *dt*-regels zijn. De woordvormenparen *gebeurt* en *gebeurd* en *wordt* en *word* klinken hetzelfde, maar afhankelijk van de syntactische context waarin ze voorkomen verschilt de spelling. Hier is niet alleen geheugenwerk vereist maar ook een bewuste toepassing van spellingregels. Frans Daems heeft niet alleen een belangrijke rol gespeeld in de bewaking en aanpassing van de spellingregels voor het Nederlands vanuit zijn werk in de Werkgroep Spelling van de Raad voor de Nederlandse Taal en Letteren van de Nederlandse Taalunie, maar ook in de studie van de oorzaken van de veelvuldige fouten in het schrijfproces van zowel beginnende als ervaren schrijvers. Zo werd door Daems en collega's met behulp van experimenteel psycholinguïstisch onderzoek aangetoond dat frequentie van voorkomen een belangrijke rol speelt in de keuze voor de syntactisch correcte varianten van homofone werkwoordsvormen (Sandra, Daems & Frisson, 2001). Bij fouten is het meestal de meest frequente vorm die ten onrechte gekozen werd.

Experimenten op het gebied van het schrijfproces bij notoire twijfelgevallen zoals homofone woordvormen kunnen ons meer inzicht geven in de kennis en processen die een rol spelen bij de verwerking van geschreven taal. Op hun beurt kunnen dergelijke inzichten bijdragen aan discussies over keuzes bij het aanpassen van spellingconventies. Een recente aanvulling op de mogelijkheden van psycholinguïstisch onderzoek is het gebruik van computationele modellen. Zo simuleert Van den Bosch (2006) met behulp van een computermodel wat het effect is van verschillende spellingwijzigingen op de leerbaarheid van de Nederlandse spelling. Hoewel het hier om leerbaarheid door een zelflerend systeem gaat, is het niet

vergezocht om een correlatie te veronderstellen met leerbaarheid door mensen. Het ligt dan voor de hand om voorstellen tot wijziging van de spellingregels ook op basis van dergelijke modellen te evalueren vooraleer ze los te laten in de taalgemeenschap.

We bouwen voort op deze methode en bestuderen in dit artikel de leerbaarheid van verwarbare homofonen met een computermodel dat tot stand komt via een zelflerende techniek. Het model implementeert een specifieke hypothese over hoe mensen taal verwerken, namelijk door het impliciet onthouden van de woordvormen in de context waarin ze voorkomen.

2. De spelling van homofonen in het schrijfproces

De spelling van het Nederlands wordt voor een deel bepaald, naast een complexe etymologische geschiedenis, door een delicate balans tussen een morfologisch en een fonologisch principe. De twee principes staan in conflict met elkaar. Waar het morfologische principe ernaar streeft om de morfologische oorsprong (vooral de spelling van de stam) van woordvormen te bewaren, probeert het fonologisch principe de spelling zo dicht mogelijk bij de uitspraak te houden. Een transparante morfologische structuur verbetert de begrijpelijkheid van een spellingsysteem, maar bemoeilijkt het schrijven. Wanneer het morfologische principe zou domineren zouden we *het huiz* schrijven, *wij loopen* en *wij ligen*. Wanneer het fonologische principe steeds de overhand zou halen, was er geen -*dt*-probleem meer; we schreven dan *hij wort* en *het is gebeurt*. Verwarbare homofonen ontstaan soms uit een verschillende etymologie die toevallig dezelfde uitspraak oplevert (zoals *noch* en *nog*, en *mei* en *mij*), maar daarnaast vaak uit het conflict tussen de morfologische en fonologische principes. De eerder genoemde -*dt*-vormen zijn daar een voorbeeld van. We schrijven *hij wordt* omdat op die manier de identiteit van de stam bewaard blijft, maar de keerzijde is dat een werkwoordvorm met dezelfde uitspraak (*ik word*) verschillend gespeld wordt, waardoor verwarring kan ontstaan met het homofone woord. In deze studie bekijken we beide types homofonen en onderzoeken de leerbaarheid ervan met cognitief geïnspireerde modellen.

3. Context- en geheugengebaseerde desambiguering van homofonen

Lezers worden blootgesteld aan woordvormen in context. Het is een aannemelijk uitgangspunt dat niet alleen het aantal voorkomens van een woordvorm (de frequentie) een

effect heeft op het geheugen van de lezer, en dus ook op de productie ervan bij het schrijven, maar ook het aantal voorkomens van een woordvorm in specifieke contexten. Zo komt de woordvorm *houdt* ongeveer 107.000 keer voor in een groot corpus van 578 miljoen woorden aan hedendaagse Nederlandse tekst, terwijl de vorm *houd* ongeveer 12.000 keer voorkomt. Ervan uitgaand dat dezelfde proportie ook wel zal voorkomen in tekst die lezers van het Nederlands tegenkomen, zou volgens het eerder genoemde frequentie-effect van Daems en collega's de fout *ik houdt* dus relatief meer moeten voorkomen dan *hij houd*. Maar ongeacht de fouten die we te lezen krijgen, zullen we toch het fragment *ik houd* veel meer te zien krijgen dan *ik houdt*, en *hij houdt* veel meer dan *ik houdt*. Hetzelfde geldt voor andere contexten die mogelijk nuttige informatie geven voor de selectie tussen homofone vormen. Onze hypothese is dat woordvormen in context geheugensporen vormen die onbewust gebruikt worden in het schrijfproces. Dat het zo vaak goed gaat bij spelling is niet alleen het gevolg van het correct toepassen van spellingregels, maar vooral van het effect van die geheugensporen, en waar het fout gaat, is dat het gevolg van het ontbreken van voldoende representatieve contexten in het geheugen.

3.1. Definitie als classificatietaak

Om deze hypothese te testen, construeren we een computationeel model dat het leerproces (door lezen) en het relevante deel van het schrijfproces (namelijk de beslissing tussen homofone varianten) simuleert. We vertrekken van een corpus van ongeveer 578 miljoen woorden hedendaags geschreven Nederlands, waarvan 90% gebruikt wordt als leermateriaal. De resterende 10% van het corpus wordt opzij gezet als testmateriaal om het gedrag van het uiteindelijke model te evalueren. Deze 90%-10%-splitsing wordt tien maal herhaald in een tienvoudig kruisvalidatie-design. Uit het leermateriaal worden per paar van verwarbare homografen (bv., *mei* en *mij*) de contexten uit het leermateriaal gehaald waarin de verschillende varianten voorkomen. Deze voorbeelden representeren de blootstelling van de lezer aan deze vormen en hun contexten tijdens het lezen. Het model construeert een geheugenstructuur uit deze voorbeelden waarbij de context gerepresenteerd wordt in lexicale vorm (de woorden uit de context zelf) en in syntactische vorm (de woordsoorten van de woorden in de context van de homofone woordvorm). De taak die het model moet oplossen is het produceren van de correcte vorm uit het groepje homofone vormen, gegeven een context; de taak is als het ware te reconstrueren welke van de alternatieve mogelijkheden de schrijver van de originele tekst zal hebben bedoeld, gezien de context. Op deze manier wordt de taak

gerepresenteerd als een classificatietaak. De *input*, het gegeven materiaal dat geclassificeerd wordt, is een context van een homofoon in termen van naburige woorden en de woordsoorten van die woorden, en de *output* is de contextueel correcte variant van de homofoon. Door de aard van de context (alleen woorden of zowel woorden als woordsoorten) te variëren, kunnen we te weten komen welke informatie een systeem nodig heeft voor het leren van de taak. Door het gedrag van het systeem te evalueren op de testdata en leercurves te analyseren kunnen we uitspraken doen over eventuele cognitieve relevantie van het model.

3.2. Geheugengebaseerd leren

Voor taken die geformuleerd worden als classificatieproblemen zijn een groot aantal algoritmen beschikbaar die uit voorbeelden leren. Omwille van de centrale rol die geheugen volgens ons speelt bij lees- en schrijfprocessen, ligt het voor de hand om als leermethode geheugengebaseerd leren te gebruiken (Daelemans & Van den Bosch, 2005). Deze leertechniek gaat uit van het principe dat leren niets meer is dan het (compact) opslaan van voorbeelden in het geheugen, en dat nieuwe problemen opgelost worden naar analogie met de meest gelijkende voorbeelden in het geheugen. Dergelijke methodes zijn in het verleden met goede resultaten toegepast, vooral in de taaltechnologie, maar ook in de computationele psycholinguïstiek (Keuleers et al., 2007). Het specifieke algoritme dat werd gebruikt voor de experimenten is IGTree (Daelemans et al., 1997), een approximatie van het klassieke '*k*-nearest neighbor'-algoritme (Cover en Hart, 1967). IGTree bewaart alle voorbeelden in het geheugen (hier contexten en de bijhorende woordvorm) en voorspelt welke van de homofonen de juiste is in de gegeven context, op basis van extrapolatie uit het homofoon van het meest gelijkende voorbeeld in het geheugen. IGTree hanteert een strikte volgorde van naburige woorden in het bepalen van gelijkenis, die automatisch bepaald is op grond van de voorspellende kracht van de positie van het woord, geschat via de informatietheoretische metriek *information gain*. Deze metriek zorgt ervoor dat twee contexten een grotere gelijkenis krijgen toegekend wanneer ze dezelfde woorden bevatten die direct naast het homofone woord staan, dan wanneer hun gelijkenis slechts bestaat uit woorden die op enige afstand van het homofone woord staan. De onmiddelijke buurwoorden van een homofoon woord zijn doorgaans verreweg het meest voorspellend voor de keuze van de juiste spelling.

3.3. Dataverzameling

Uit CELEX (Baayen et al., 1993) extraheerden we een lijst van 15.896 verzamelingen met ieder twee of meer homofonen op grond van de aanwezige orthografische en fonemische informatie. De meeste verzamelingen (93%) zijn paren, maar de lijst bevat ook aanmerkelijk grotere verzamelingen (met als grootste verzameling *weidde - weidden - weide - weiden - wijdde - wijdden - wijde – wijden,* met acht homofonen). Vervolgens zochten we naar alle voorkomens van iedere homofonenverzameling en legden deelcorpora aan met de voorkomens van ieder woord in de verzameling in een directe context van vijf linker- en rechterbuurwoorden. Uit deze verzameling van deelcorpora selecteerden we tien hoogfrequente paren met meer dan 160.000 voorkomens per homofoonvariant. In Tabel 1 worden de tien paren opgesomd en verdeeld in vier groepen: de verwarring tussen *ij* en *ei* in de drie voornaamwoorden *zij*, *mij* en *wij* met hun *–ei* varianten, de verwarring van de grafemen *ch* en *g* in twee gevallen, en gerelateerde verwarringen tussen *d* en *t*, en tussen *d* en *dt* in werkwoordsvormen.

Verwarbaarheids-categorie	*Confusie-set*	*Aantal contexten*	*Frequentste woord*	*Minder frequente woord*
ij/ei in voornaam-woorden	zij, zei	622.647	zij, 370.932	zei, 251.715
	mij, mei	278.954	mij, 193.437	mei, 85.517
	wij, wei	161.350	wij, 159.408	wei, 1.942
ch/g	noch, nog	1.590.379	nog, 1.569.788	noch, 20.591
	licht, ligt	221.247	ligt, 149.287	licht, 71.960
d/t	moed, moet	818.965	moet, 807.802	moed, 11.163
	bekend, bekent	317.388	bekend, 313.014	bekent, 4.374
	wand, want	207.182	want, 201.952	wand, 5.230
d/dt in werkwoorden	word, wordt	1.149.078	wordt, 1.127.134	word, 21.944
	vind, vindt	270.210	vindt, 196.019	Vind, 74.191

Tabel 1. De tien geselecteerde homofoonparen met hun totale en uitgesplitste frequentie van voorkomen in een corpus van 578 miljoen woorden.

De uitgesplitste tellingen van voorkomen van de individuele woorden in Tabel 1 laten zien dat er soms grote verschillen bestaan (*bekend* komt bijvoorbeeld 71 maal vaker voor dan *bekent*), maar dat er ook paren zijn die elkaar niet ver ontlopen in frequentie, bijvoorbeeld *zij* en *zei*.

3.4. Experimentele opzet

Met behulp van een automatische 'part of speech'-tagger (MBT, Daelemans et al., 1996) werd het gehele achtergrondcorpus van 578 miljoen woorden voorzien van woordsoortinformatie. De tagger kent naar schatting 96.5% correcte woordsoortinformatie toe aan woorden in willekeurige tekst. Deze informatie wordt meegenomen in ieder van de deelcorpora per homofoonpaar en wordt gebruikt in de tweede van de twee experimentele condities. In de eerste conditie wordt de locale context rondom een homofoon woord gerepresenteerd door de vijf linker- en rechterbuurwoorden; in de tweede conditie worden van deze tien woorden ook de door de tagger toegekende woordsoorten meegenomen.

Ieder deelcorpus van de tien geselecteerde paren wordt systematisch opgedeeld volgens de eerdergenoemde opzet van een tienvoudig kruisvalidatie-experiment, zodat een gemiddelde onderscheidend vermogen gemeten kan worden. Om recht te doen aan de soms scheve verdeling van de alternatieven in een homofonenverzameling is het belangrijk om het onderscheidend vermogen van het leeralgoritme niet te meten in termen van accuraatheid, d.w.z. het percentage correct gemaakte beslissingen. Als de meest voorkomende van een paar van homofonen honderd maal vaker voorkomt dan de minder voorkomende van de twee, dan kan een leeralgoritme 99% accuraatheid halen door altijd de meest frequente vorm te voorspellen, maar dit hoge resultaat verbloemt het feit dat het algoritme volledig faalt op de minder frequente vorm. Betere metrieken zijn de *precision* en *recall* van ieder van de twee mogelijke uitkomsten (bijvoorbeeld *mij* of *mei*). De *precision* van een uitkomst is de ratio tussen het aantal correcte voorspellingen en het totale aantal voorspellingen van die uitkomst. De *recall* van een uitkomst is de ratio tussen het aantal correcte voorspellingen van die uitkomst en het totaal aantal keer dat het die uitkomst had moeten zijn. Bij twee mogelijke uitkomsten die uiteenlopen in frequentie komt het vaak voor dat met name de *recall* van de minder frequente uitkomst onder druk komt te staan; veel leeralgoritmen zijn geneigd te vaak de hoger frequente uitkomst te voorspellen en daarmee veel gevallen van de lager frequente uitkomst te missen.

4. Resultaten

De *precision*- en *recall*scores van IGTree op de tien homofoonparen, wanneer de contexten alleen gerepresenteerd worden door de naburige woorden (zie Tabel 2), laten zien dat de meer frequente alternatieven in ieder homofoonpaar ("Woord 1") beter worden

voorspeld, meestal zowel in termen van *precision* als van *recall*, dan de minder frequente ("Woord 2"). Als het verschil in frequentie groot is, zoals bij *wij – wei* en *bekend – bekent*, wordt de frequentere vorm vrijwel perfect voorspeld en wordt de minder frequente vorm relatief slecht voorspeld, met name in termen van *recall*. De relatief laagfrequente woorden *noch* en *bekent* worden bijna in de helft van hun voorkomens verward voor hun frequentere homofoon. Omdat in beide gevallen de minder frequente vorm veruit de minderheid vormt van voorkomens van het homofoonpaar is de accuraatheid van IGTree (het totale percentage correct gemaakte beslissingen) in beide gevallen wel hoger dan 99%.

Woord 1	**precision**	**recall**	**Woord 2**	**precision**	**recall**
zij	95.7	98.9	zei	98.2	93.4
mij	98.9	99.2	mei	98.3	97.5
wij	99.9	99.9	wei	98.3	90.0
nog	99.4	99.8	noch	78.3	51.3
ligt	93.7	98.0	licht	95.4	86.3
moet	99.8	99.9	moed	94.6	82.4
bekend	99.4	99.9	bekent	89.3	53.4
want	99.8	99.9	wand	99.1	94.0
wordt	99.6	99.6	word	90.5	81.1
vindt	97.1	98.9	vind	97.1	92.5

Tabel 2. Voorspellingsscores in termen van precision en recall van IGTree op de tien homofoonparen, op basis van contexten van alleen woorden. "Woord 1" is het frequentere woord; "Woord 2" het minder frequente.

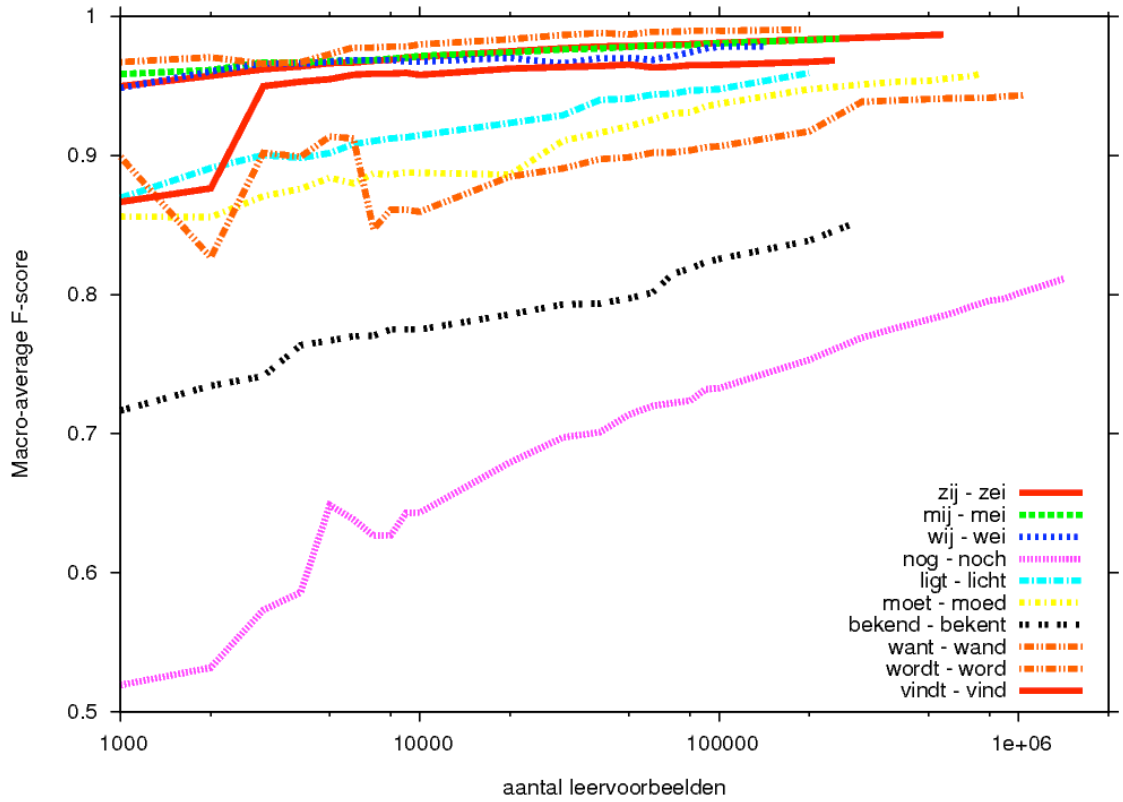
Wanneer woordsoortinformatie ook beschikbaar is voor IGTree gaat de voorspelbaarheid van met name de minder frequente homofonen in een aantal gevallen sterk vooruit, zoals te zien is in de vetgedrukte grijze cellen in Tabel 3. In zeven van de tien gevallen stijgt de *recall* aanzienlijk, soms met een aantal procenten, zoals bij *zei* (van 93.4 naar 98.9) en *wand* (van 94.0 naar 98.0). In drie gevallen brengt de toevoeging van woordsoortinformatie echter geen positief effect teweeg: bij *mij – mei*, *nog – noch*, en *bekend – bekent*. Bij de laatste twee gevallen lijkt de zeer scheve frequentieverdeling tussen de hoogfrequente en de laagfrequente variant niet te compenseren met meer generieke syntactische contextinformatie; in het eerste geval (*mij – mei*) is er al een zekere balans tussen de hoge *precision* en *recall*, en lijkt de minder frequente vorm *mei* onveranderlijk verward te worden met *mij* in ongeveer één op de vijftig gevallen.

Woord 1	precision	recall	Woord 2	precision	Recall
zij	**98.3**	98.4	zei	98.9	**98.9**
mij	99.0	99.1	mei	97.9	97.8
wij	99.9	99.9	wei	98.5	**93.8**
nog	99.4	99.8	noch	78.6	51.6
ligt	**96.5**	97.9	licht	95.6	**92.6**
moet	99.8	99.9	moed	94.0	**88.9**
bekend	99.4	99.9	bekent	89.3	53.3
want	99.8	99.9	wand	*97.6*	**98.0**
wordt	99.7	99.9	word	**93.3**	**84.3**
vindt	97.7	99.0	vind	97.2	**93.9**

Tabel 3. Voorspellingsscores in termen van precision en recall van IGTree op de tien homofoonparen, op basis van contexten van woorden en hun woordsoorten. Vetgedrukte scores in een lichtgrijze cel duiden op een toename van meer dan 1.0 vergeleken met de voorspellingsscore zonder woordsoortinformatie (zie Tabel 2).

4.1. Leercurves

Een belangrijk aspect van de verschillen in onderscheidend vermogen, naast de scheve verdeling van de alternatieven binnen een homofoonverzameling, is de absolute hoeveelheid voorbeelden die in een corpus gevonden kunnen worden. Tot nu toe rapporteerden we over alle voorbeelden die we van een homofonenpaar konden vinden in een corpus van 578 miljoen woorden, en dat leverde variabele hoeveelheden voorbeelden op. Het is ook mogelijk te vergelijken hoe goed verschillende homofoonparen onderscheiden kunnen worden bij gelijke hoeveelheden leervoorbeelden per homofoonpaar. Figuur 1 biedt hiertoe zogenaamde leercurves voor de tien onderzochte homofonenparen. De horizontale as, die een logaritmische schaal heeft, markeert het aantal leervoorbeelden (locale contexten); de curves zijn gemeten op de experimentele conditie waarbij woorden en woordsoortinformatie in de context zijn opgenomen, met pseudo-exponentieel toenemende aantallen leervoorbeelden. Het onderscheidend vermogen van de tien modellen, uitgedrukt in een enkel getal, is gemeten in termen van *macro-average F-score*, een ongewogen gemiddelde van de *F-score* van beide alternatieven, waarbij de *F-score* op zijn beurt het harmonisch gemiddelde is van de *precision* en de *recall* van iedere uitkomst, berekend met de formule (2 * *precision* * *recall*) / (*precision* + *recall*). Deze maat weegt de voorspellingen op de minderheidsuitkomst even zwaar als de voorspellingen op de meerderheidsuitkomst.



Figuur 1. Leercurves van de tien homofoonparen. De x-as heeft een logaritmische schaal. Voorspellend vermogen is gemeten in termen van macro-gemiddelde F-score.

Wat allereerst opgemerkt kan worden over de leercurves in Figuur 1 is dat alle lijnen een opwaartse trend vertonen, maar dat een aantal enigszins afvlakt. Dat geldt met name voor de homofoonparen die het beste onderscheiden worden, met een *macro-average F-score* van 95% of hoger; hier lijken 10 à 20.000 voorbeelden al het best haalbare onderscheidende vermogen op te leveren. Minder goed onderscheiden homofoonparen als *nog - noch* en *bekend – bekent* lijken nog een orde of twee maal zoveel leermateriaal nodig te hebben voordat ze een *macro-average F-score* van meer dan 90% zullen bereiken.

Een tweede opmerking die gemaakt kan worden naar aanleiding van de leercurves is dat het voorspellend vermogen dat haalbaar is met de gebruikte methode, tot op zekere hoogte relatief is ten opzichte van het aantal beschikbare leervoorbeelden. Hoe meer voorbeelden, hoe beter het onderscheidend vermogen op ongezien materiaal, hoewel voor homofonen als *zij – zei*, *mij – mei* en *wij – wei* het plafond bereikt lijkt te zijn rond een *macro-average F-score* van 98%.

4.2. Confusie-analyse

Op grond van de resultaten opgesomd en afgebeeld in Tabellen 2 en 3 en Figuur 1 kunnen we vaststellen dat de frequentieverdeling van de alternatieven en de hoeveelheid beschikbaar leermateriaal van invloed zijn op het onderscheidend vermogen van een geheugengebaseerd leeralgoritme. We zien dat relatief veel fouten de minder frequente woorden treffen. In absolute zin is dit echter niet altijd het geval. In Tabel 4 worden de absolute aantallen confusies opgesomd tussen "Woord1" (het frequentere woord) en "Woord2" (het minder frequente woord), in beide experimentele condities.

Homofoonpaar	Alleen woorden in context		Woorden en woordsoorten in context	
	Woord 1 als woord 2	Woord 2 als woord 1	Woord 1 als woord 2	Woord 2 als woord 1
zij - zei	404	1611	**408**	**381**
mij - mei	152	215	**190**	**189**
wij - wei	4	16	3	13
nog - noch	309	1026	289	1018
ligt – licht	303	953	278	504
moet - moed	57	190	57	118
bekend - bekent	21	191	19	191
want - wand	3	30	**11**	**8**
wordt - word	189	441	137	367
vindt - vind	226	575	197	466

Tabel 4. Absolute aantallen verwarringen tussen het frequentere woord (Woord 1) en het minder frequente woord (Woord 2) van de tien bestudeerde homofoonparen, in beide experimentele condities.

De cijfers in Tabel 4 laten zien dat meestal het minder frequente woord meer verward wordt voor het frequentere woord dan andersom, maar in de eerste plaats ontlopen de absolute getallen elkaar doorgaans niet meer dan een orde van grootte. In een aantal gevallen waarbij ook informatie over woordsoorten beschikbaar is (vetgedrukt in de rechterkolommen van Tabel 4) wordt het frequentere woord zelfs vaker verward voor het minder frequente woord dan andersom: bij *zij – zei*, *mij – mei*, en *want – wand*. Op basis van deze verwarringsstatistieken, in combinatie met de tellingen van Tabel 1, kan in grote lijnen geconcludeerd worden dat bij relatief minder scheve verdelingen de verwarbaarheid tussen twee alternatieven niet altijd uitvalt in het voordeel van de meest frequente vorm; contextinformatie, met name woordsoortinformatie, is hiervoor verantwoordelijk.

5. Discussie

De resultaten van onze experimenten suggereren dat frequentie inderdaad een belangrijke rol speelt bij de selectie van homofonen in het schrijfproces. Ze laten echter ook zien dat de keuze mogelijk niet alleen afhangt van frequentie, maar ook van de onmiddellijke context van het homofoon. Een computationeel lerend model dat informatie over de naburige woorden van een homofoon en hun syntactische klasse bewaart en gebruikt, is in staat om het aantal verwarringen van minder frequente woorden redelijk terug te brengen, in het geval van syntactische informatie soms zelfs in het voordeel van het minderheidswoord. De leercurves tonen aan dat er ook lastige gevallen overblijven met erg scheve verdelingen en beslissingen waar lokale context niet voldoende informatie geeft (zoals *nog* - *noch* en *bekend* - *bekent*). Uit het model blijkt ook geen verschil tussen types van verwarbare paren (op basis van etymologie of op basis van conflicten tussen morfologisch en fonologisch principe). Een meer informatieve indeling is tussen gevallen waarbij de lokale context voldoende desambigueert en die waar dat niet het geval is, en frequentie dus een belangrijkere rol speelt.

Vanuit ons model kunnen we concrete predicties maken over het soort verwarringsfouten dat in geschreven tekst waarschijnlijker is dan andere. In psycholinguïstisch vervolgonderzoek kan dit verder onderzocht worden.

Referenties

Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1993). The CELEX lexical database on CD-ROM. Philadelphia, PA: Linguisic Data Consortium.

Cover, T., & Hart, P.E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*, 21-27.

Daelemans, W., Zavrel, J., Berck P. & Gillis, S. (1996). Memory-Based Part of Speech Tagging. In: Ejerhed, E. and Dagan, I. (eds.) *Proceedings of the Fourth Workshop on Very Large Corpora, Copenhagen*, Denmark, 14-27.

Daelemans, W. & Van den Bosch, A. (2005). *Memory-Based Language Processing*. Cambridge: Cambridge University Press.

Daelemans, W., Van den Bosch, A., & Weijters, A. (1997). IGTree: Using trees for compression and generalization in lazy learning algorithms. *Artificial Intelligence Review, 11*, 407-423.

Keuleers, E., Sandra, D., Daelemans, W., Gillis, S., Durieux, G., & Martens, E. (2007). Dutch plural inflection: The exception that proves the analogy. *Cognitive Psychology, 54:4*, 283-318.

Sandra, D., Daems, F., & Frisson, S. (2001). Zo helder en toch zoveel fouten! Wat leren we uit psycholinguïstisch onderzoek naar werkwoordfouten bij ervaren spellers? In *VONK: Tijdschrift van de Vereniging voor het Onderwijs in het Nederlands*, *30:3*, 3-20.

Van den Bosch, A. (2006). Spelling space: A computational test bed for phonological and morphological changes in Dutch spelling. *Written Language and Literacy, 9:1*, 25-44.