HET PROEFSCHRIFT VAN

Theo Vosse

# Speuren naar spelfouten

*Peter Burger*

Zelden zo'n eigenaardig proefschrift gezien als *The word connection*. Het boek begint als een ouderwetse Amerikaanse misdaadroman met *private eye* Hank, die bezoek krijgt van de knappe, academisch gevormde Deborah. Zij dumpt een stapel boeken over kunstmatige intelligentie, taaltechnologie en automatische zinsontleding op zijn bureau en vraagt hem een computerprogramma te bemachtigen dat automatisch spelfouten corrigeert. Bedrijfsspionage! Om te beginnen verdiept Hank zich in een dissertatie over spellingcorrectie.

En dan volgt het eigenlijke proefschrift van de computertaalkundige Theo Vosse. Vosse (32) werkt als onderzoeker bij de Leidse vakgroep Functieleer en Theoretische Psychologie. Hij onderzocht de mogelijkheden om automatisch spelfouten op te sporen en te verbeteren. En hij ging nog een stap verder: hij schreef het correctieprogramma waar detective Hank naar zoekt – CORR[ie].

Het bedenken van zo'n programma lijkt eenvoudig: voer de computer een woordenboek en laat hem daarin alle ingetypte woorden opzoeken; elk woord dat hij niet kent, is fout. Er bestaat al een programma dat zo werkt: de spellingcontroleur van WordPerfect. Wie zijn typewerk door WP laat nakijken, kent de nadelen van dit systeem: de meeste gesignaleerde 'fouten' zijn correcte woorden die het programma domweg niet herkent. Het aantal valse alarms daalt alleen als de gebruiker zelf woorden aan het woordenboek toevoegt. Dat blijkt ook anders te kunnen.

## ● REISGIDSREDACTEUR

Uitstekende Nederlandse woorden als *speelgoedbrandweerauto's* en *vioolbouwersnederzetting* staan niet in woordenboeken. Dat hoeft ook niet: elke lezer die de delen kent, kan de betekenis van het geheel raden. Maar alleen als die lezer een mens is. Voor de computer is *speelgoedbrandweerauto's* niets anders dan een letterreeks die niet in de woordenlijst voorkomt – dus een mogelijke spelfout.

Vosse rustte CORR[ie] daarom uit met de mogelijkheid om samenstellingen te herkennen die niet in de woordenlijst staan. Zo'n systeem bestond nog niet, onder meer doordat de meeste onderzoekers zich bezighouden met het Engels, dat zelden woorden aan elkaar plakt zoals het Nederlands of het Duits dat doet.

CORR[ie] herkent *reisgidsredacteur* dus als een samenstelling, ook als ze dat woord voor het eerst ziet. Maar hoe weet ze nu dat *kwaliteitesverbetering* een typefout is, en niet de mogelijke samenstelling *kwaliteit+es+verbetering?* Vosse: "CORR[ie] hanteert de vuistregel dat samenstellingen zelden uit drie of vier korte woorden bestaan. Ook een samenstelling met een weinig voorkomend woord van twee of drie letters erin, zoals *es*, is verdacht. Bovendien kent ze een woord dat er maar één letter van verwijderd is: *kwaliteitsverbetering*. Als zo'n correctie bestaat, signaleert ze die."

## ● WEESMEISJE

CORR[ie] is ook getraind om de fout op te sporen in *Hij word nooit ziek*. Omdat zo'n fout resulteert in een bestaand woord, heeft CORR[ie] aan een woordenlijst niet genoeg. Ze moet kunnen ontleden. Vosse: "Dat ontleden kost ongeveer een seconde per woord, dus voor de fouten met d's en t's moet je wel veel geduld hebben. Voor een tekst van tienduizend woorden moet je drie uur wachten."

CORR[ie] ziet ook andere fouten die kennis van de grammatica vereisen. Zoals in de bijzin *...waarbij dus iedere niet uitputtende beschrijving van zon systeem op vele wijzen te interpreteren is. Zon* kan hier geen zelfstandig naamwoord zijn, constateert CORR[ie] terecht: er ontbreekt een apostrof – *zo'n.*

Machinaal ontleden blijft echter een moeilijke zaak. Taalgebruikers van vlees en bloed zien wel dat *De wielrenners zagen de man met de hamer* een andere structuur heeft dan *De goochelaars zagen het weesmeisje met een kettingzaag*. En dat er iets loos is in *De wielrenners zaagden de man met de hamer*. CORR[ie] niet. "Ze heeft geen kaas gegeten van betekenis", zegt Vosse. "Dat blijft een probleem."

Maar ze ziet wel de fout in *faillissement*. Vosse: "Al die verticale streepjes achter elkaar, daar lezen mensen overheen. Bij kunstmatige intelligentie is het een algemene regel dat taken die makkelijk zijn voor mensen, moeilijk zijn voor computers, en vice versa."

"Uiteindelijk is een menselijke corrector veel beter," zegt Vosse, "maar ook duurder. En er zijn steeds minder correctoren: corrigeren is intensief werk dat niet geweldig betaalt. Bovendien zijn er niet zoveel mensen die echt goed spellen."

In de epiloog wordt *The word connection* weer een roman. De laatste scène is klassiek: Hank treft in zijn kantoor een chaos aan van losse papieren, geopende kasten en omgekeerde laden. De concurrentie heeft alle gegevens gestolen. Wishful thinking van Vosse? Hoewel twee uitgeverijen met CORR[ie] werken, zal het programma voorlopig niet op de markt worden gebracht door een softwarebedrijf. Individuele computergebruikers zullen het dus nog enige tijd moeten doen met de dommere spellingcontroleur van WordPerfect. <

T.G. Vosse: *The word connection. Grammar-based spelling error correction in Dutch.* Enschede, Neslia Paniculata, 1994. ISBN 90 75296 01 0. Prijs *f* 35,-.