Redactionele en computationele voorbeeldselectie

Dirk Kinable (Instituut voor Nederlandse Lexicologie)

1. Computergestuurde voorbeeldselectie: een intrigerende functionaliteit

Voorbeeldzinnen in woordenboeken citeert een lexicograaf uit bronnen of worden door hemzelf bedacht. Deze tweedeling 'quoted' en 'invented' geldt ook als basisonderscheid in de vakliteratuur over lexicografische documentatie van betekenisonderscheidingen.[1] Veelal formuleert dergelijk onderzoek een voorkeur voor aangehaalde vindplaatsen of citaten in eigenlijke zin, zowel om hun grotere authenticiteit en autoriteit als om het daarmee samenhangend vermogen tot illustratie en onderbouwing.[2] In dit opzicht beschikken corpusgebaseerde woordenboeken, zoals ook het *Algemeen Nederlands Woordenboek*, blijkbaar over goede perspectieven. De materiaalverzameling van het ANW illustreert lemma's met meer dan 70 miljoen concordanties, afkomstig uit teksten die een breed veld bestrijken van ruim 3500 domeinen, tientallen kranten en meer dan 600 literaire teksten. Dat aan voordelen eveneens nadelen kunnen kleven, geldt echter ook hier. Met de brede, authentieke materiaalbasis dient zich voor de woordenboekschrijver onder meer meteen het probleem aan van de selectie van voorbeeldzinnen.

Dergelijke problemen in verband met de beheersing van grote materiaalverzamelingen vergen noodzakelijk de hulp van computerprogrammatuur. In het geval van het ANW is daarbij gekozen voor Sketch Engine, een tool voor corpora die talloze zoek- en sorteermogelijkheden biedt met ordeningen van concordanties naar lemma, woordomgeving, syntactische verbanden, geografische herkomst van de bronnen enz. Naast de al beschikbare mogelijkheden wordt ook aan de uitbreiding van functionaliteiten gewerkt.[3] Deze ontwikkeling van geavanceerde versies opent perspectieven voor het hierboven gesignaleerde probleem van de voorbeeldselectie. Intrigerend in dit verband is de toegevoegde functionaliteit 'Sort good dictionary examples'. Deze sorteert de vindplaatsen van een bepaald lemma op geschiktheid in een afdalende volgorde. De vraag rijst in hoeverre deze functionaliteit al operationeel is voor de bewerking van ANW-artikelen. Een steekproefsgewijze vergelijking kan aan deze discussie alvast een bijdrage leveren.

2. Een proeftraject van honderd lemma's

Als sample dient in dit vergelijkend onderzoek een traject van 100 artikelen die ik voor het ANW heb bewerkt zonder hulp van voornoemde functionaliteit. Dit traject behelst zowel monoseme als polyseme woorden (54, respectievelijk 46), bovendien gespreid over verschillende woordsoorten. Bij de monoseme woorden zijn er 44 substantieven, 9 werkwoorden en 1 adjectief; bij de polyseme lemma's zijn de aantallen respectievelijk 31, 9 en 4, waarnaast 2 woorden zowel als substantief en als adjectief voorkomen (*consonant* en *melomaan*). Deze 100 artikelen betreffen woorden die het ANW-corpus elk met circa 50 aanhalingen documenteert. Bij de bewerking zijn er daarvan in totaal 517 weerhouden. Hoe de sortering door Sketch Engine zich tot deze redactionele selectie verhoudt, belichten we hierna.

Als toetsingscriterium zijn we uitgegaan van de vraag in hoeverre de functionaliteit voor automatische voorbeeldselectie de redactioneel gekozen citaten kan sorteren naar een positie binnen de eerste 15 aanhalingen op het genoemde aantal van circa 50. Voor de lexicograaf zou het namelijk al behoorlijke tijdwinst opleveren, als hij zich voor zijn voorbeeldkeuze vooral kon concentreren op grosso modo het naar voor gesorteerde eerste

[1] Zgusta 1971: 266-267; Landau 2001: 208; Cowie 2002: 74.
[2] Zgusta 1971: 265; Landau 2001: 166, 210; Harras 2002: 612.
[3] Zie: <www.sketchengine.co.uk> en <beta.sketchengine.co.uk>.

derde van het corpusmateriaal. Een dergelijke reductie zou bovendien steeds meer renderen, naarmate de aantallen vindplaatsen voor omvangrijkere woorden oplopen. Omgekeerd mag ook duidelijk zijn dat bij woorden met duizenden vindplaatsen een derde van het materiaal nog steeds een behoorlijke werklast betekent voor een redacteur, ook al kan deze ondersteuning zoeken door andere functionaliteiten van Sketch Engine zoals Word Sketch en het opvragen van materiaalsamples. De gekozen marge gaat daarmee al bij al uit van een redelijke verwachting ten aanzien van de efficiëntie van de sorteerfunctionaliteit.

Overschouwen we de resultaten van onze steekproef, dan levert dat het volgende totaalbeeld op. Van de 517 redactioneel gekozen voorbeelden kregen er door de sorteerfunctie 187 of 36,17% een plaats toegewezen bij de eerste 15 vindplaatsen van de circa 50 per lemma, terwijl 330 of 63,82% terechtkwamen bij de daaropvolgende circa 35 vindplaatsen. De steekproef laat gelijkaardige verhoudingen zien, wanneer men vervolgens de 517 voorbeelden opsplitst naar monoseme en polyseme woorden.Van de 173 citaten bij monoseme woorden, blijkt 39,3% (68) binnen de zone van de eerste 15 voorbeelden geplaatst te worden en 60,69 % (105) daarbuiten. Bij de polyseme lemma's met een groter aantal citaten (344) als gevolg van hun verschillende betekenissen, ligt deze verhouding in een vergelijkbare orde van grootte op 34,59 % (119) en 65,4 % (225).

Aantal redactioneel geselecteerde voorbeelden op 100 artikelen	517	
Automatische sortering	Bij de eerste 15 citaten	Als citaat 16-49/51
Monoseem	68 (39,30 %)	105 (60,69 %)
Polyseem	119 (34,59 %)	225 (65,40 %)
Totaal	187 (36,17 %)	330 (63,82 %)

Hoewel de bovenstaande tabel duidelijke verschillen laat zien tussen redactionele en computationele voorbeeldkeuze, is het perspectief dat zich in de cijfers aftekent, beslist positief te noemen. Dat geldt des te meer wanneer men rekening houdt met een acceptabele opwaardering van de bovenstaande scores. Neemt men bijvoorbeeld die monoseme woorden of betekenissen van polyseme woorden onder de loep, waar de computer geen enkel van de redactioneel gekozen voorbeelden opneemt bij de eerste 15 citaten, dan stelt men vast dat in deze computersortering minstens 54 voorbeelden voorkomen die men als gelijkwaardig met de redactionele keuze kan beschouwen. Dat zou het totaalresultaat van 36,17% al met minimaal 10,4% verbeteren. Het zou dus een misvatting zijn om de menselijke keuze steevast als het *nec plus ultra* te beschouwen. Niettemin is duidelijk dat in deze alliantie tussen mens en machine de computer wordt aangestuurd door criteria die de taalonderzoeker bepaalt.Voor een beter begrip van de wijze waarop bovenstaande resultaten tot stand kwamen, is het dan ook nodig even stil te staan bij de gebruikte programmatuur. Deze staat te boek onder de naam 'GDEX' als initiaalwoord voor 'Good Dictionary Examples' (Kilgarriff e.a. 2008: 425). Voor selectie uit omvangrijke corpora toetst deze software voorbeeldzinnen aan de volgende maatstaven:

1. Een goede informatieve lengte van 10 tot 25 woorden.
2. De frequentie van gangbare woorden.
3. De afwezigheid van vaak extra context vergende aanwijzende pronomina en terugverwijzende woorden.
4. De aanwezigheid van de collocatie in de hoofdzin.

5. De vorm van een volzin met een hoofdletter aan het begin en een afsluitende punt, uitroepteken of vraagteken.
6. De aanwezigheid van woorden met een opvallende aanwezigheid of 'saillantheid' bij combinaties.
7. De positie van het trefwoord aan het zinseinde, na een inleidende, verduidelijkende context.

Naarmate aanhalingen aan deze criteria voldoen, krijgen ze een hogere score; in het omgekeerde geval volgt een lagere waardering. Niet alle criteria tellen daarbij in dezelfde mate mee. Vooral de eerste twee kenmerken, zinslengte en woordfrequentie, krijgen het meeste gewicht (Kilgarriff e.a. 2008, 426-427).

Realiseren deze criteria, zoals de tabel hierboven illustreerde, al een positief, zelfs opwaardeerbaar resultaat, dan suggereert de frequente sortering van informatieve, redactioneel gekozen citaten naar de plaatsen 16-49/51 ook het belang van verdere optimalisering van de automatische voorbeeldselectie. In eerste instantie blijkt het mogelijk meer waardevolle citaten naar voren te sorteren dan nu het geval is, door in het eerste derde van het corpusmateriaal redundante voorbeelden te weren. Aanleiding voor het meermaals voorkomen van een voorbeeldzin is vaak dat websites informatiedelen van elkaar overnemen, herhaling van het trefwoord in een voorbeeldzin of een onvolkomenheid in het corpusmateriaal zelf. Waar de zinnen met het trefwoord geheel of vrijwel volledig overeenstemmen (zoals in onze steekproef onder meer bij *klepel*, *moeien* en *toucheren*), volstaat echter een enkel voorbeeld. Hetzelfde geldt voor collocaties. Illustratief in dit verband zijn onder meer de lemma's *generisch*, *mediaal*, *moesson* en *plengen*, waar er van de eerste 15 gesorteerde voorbeelden respectievelijk 11 worden ingenomen door de verbinding *generisch (geneesmiddel*, *middel*, *product)*, 9 door de combinatie *mediale (band*, *knieband)*, 6 door de verbinding *droge*, resp. *natte moesson* en 14 door de verbinding *een traan*, *tranen*, *krokodillentranen plengen.* Door het aantal 'beste' voorbeelden per voorbeeldzin of per collocatie in het eerste derde deel van het lemmamateriaal te limiteren en GDEX de redundante treffers door toekenning van een lage score naar achteren te laten schuiven, komen vooraan posities vrij voor nieuw informatief voorbeeldmateriaal.

Niet alleen redundantie binnen de eerste 15 voorbeeldzinnen beïnvloedt de GDEX-scores negatief. Vooral het hoge cijfer van de 330 redactioneel geselecteerde voorbeelden in de posities 16-49/51 haalt de score van de sorteerfunctie sterk naar beneden.

Zoomen we in op dit totaal van 330, dan laten zich een aantal groepen onderscheiden. Een wezenlijk aandeel daarin hebben in totaal 136 citaten (41,21%) die afkomstig zijn van: (a) betekenissen van monoseme woorden waarvoor GDEX alle redactioneel gekozen citaten naar de posities 16-49/51 doorschuift (32 voorbeeldzinnen); (b) betekenissen waarvoor op 1 redactionele aanhaling na alle redactioneel gekozen concordanties in dit segment terechtkomen (104 voorbeeldzinnen).

De twee voorgaande subgroepen hebben gemeen dat GDEX de betreffende semantische onderscheidingen wel documenteert binnen het eerste derde van het corpusmateriaal met al dan niet redactioneel geselecteerde citaten. In tegenstelling hiermee laat zich binnen het totaal van 330 vindplaatsen nog een categorie onderscheiden die afkomstig is uit lemmadelen die *GDEX* geheel ongedocumenteerd laat binnen de eerste 15 voorbeeldzinnen en dus ook niet met gelijkwaardige concordanties illustreert. Deze kleinere maar wel bijzondere groep bevat 66 voorbeeldzinnen of 20 % op het totaal van 330. Deze situatie met 20% vindplaatsen voor betekenissen, combinaties en verbindingen die ontbreken in het eerste derde deel van het gesorteerde corpusmateriaal noopt de lexicograaf ertoe ook te zoeken in de volgende tweederde om zijn artikel met voorbeelden te documenteren. Bij lemmamateriaal met duizenden concordanties kan zo'n aantasting van het rendement van de

sorteerfunctie heel andere implicaties inhouden dan bij de overzichtelijke lemma's van onze steekproef. We belichten deze categorie voorbeeldzinnen dan ook even extra.

Het aantal citaten betreffende betekenissen en belangrijke subbetekenissen die GDEX binnen de eerste 15 gepresenteerde voorbeeldzinnen niet documenteert, bedraagt 23. Zij zijn afkomstig van de polyseme woorden; de monoseme zijn hier niet aan de orde, aangezien hun enige betekenis logischerwijze wel wordt geïllustreerd binnen de vindplaatsen die GDEX vooraan sorteert. Genoemde 23 voorbeelden illustreren 14 betekenissen van de 120 die bij de polyseme lemma's werden onderscheiden. Voor 11,66 % van de betekenissen in deze groep ontbreken dus voorbeelden in de eerste 15 posities en brengt GDEX alle redactioneel gehonoreerde citaten onder in de laatste twee derden van het corpusmateriaal.

Een tweede categorie binnen de groep van 66 betreft 35 aanhalingen voor een dertigtal combinatietypes, zoals bijvoorbeeld een substantief in een voorzetselgroep of in een objectpositie bij een werkwoord (*naar anijs smaken*, *een pantomime opvoeren*). Onder combinaties verstaat het ANW woordgroepen die vooral de syntactische combinatiemogelijkheden demonstreren van een woord in de vorm van vertrouwde vrije verbindingen en van lexicale en grammaticale collocaties.

Ten slotte laten zich nog 8 voorbeeldzinnen uit de 66 onderscheiden voor 7 verbindingen, waaronder *generische naam* en *mediale tennisellebоog*. In het ANW vormen verbindingen een ruime categorie van onder meer vaste, idiomatische woordgroepen, combinaties met een gevestigd karakter, zegswijzen, uitdrukkingen, vergelijkingen e.d.

3. Redactionele afwegingen bij voorbeeldselectie

Blijkt uit de voorgaande cijfers dat toch een behoorlijk aantal redactioneel geselecteerde voorbeelden naar posities verderop in het corpusmateriaal worden verwezen, dan dringt zich ook de noodzaak op van reflectie over mogelijke oorzaken.

De voormelde criteria waarop GDEX zich baseert, gaven al aan dat vooral formele aspecten of computationeel goed berekenbare normen aan de basis liggen van de voorbeeldsortering. De globale zinvolheid van deze GDEX-criteria behoeft verder geen betoog. Beantwoorden zij echter ook voldoende aan de normen waardoor een redacteur zich bij zijn lexicografisch werk laat leiden? De vakliteratuur over lexicografische voorbeelden althans adviseert in ieder geval de nodige inhoudelijke afwegingen. Zij bepaalden mee de selectie van de hierboven geïnventariseerde voorbeeldzinnen en hebben blijkbaar vaak toch tot andere keuzes geleid dan deze van GDEX. Zonder exhaustief te willen zijn ten aanzien van de criteria uit de vakliteratuur of de ANW-lemma's uit de steekproef, licht ik dit toe aan de hand van de groep van 330 redactioneel geselecteerde voorbeeldzinnen die GDEX lager waardeerde

Primair wijst de vakliteratuur doorgaans op de belangrijke relatie tussen de definitie en de voorbeeldzinnen en hun onderling aanvullende waarde voor de woordenboekgebruiker. Het spreekt voor zich dat deze wederzijdse betrokkenheid slechts mogelijk is bij voorbeeldzinnen met een informatieve lexicografische context die de woordbetekenis helder laat vaststellen. Dergelijke informatie kan echter verschillende vormen aannemen. In *Dictionaries* formuleert Harras als eerste basisregel voor informatieve voorbeeldselectie dat goede citaten prototypische eigenschappen weergeven van wat het trefwoord aanduidt (Harras 1989: 611-613). Een voorbeeldzin van het ANW-lemma *moesson* vermeldt een dergelijke elementaire eigenschap: 'de uit het oosten waaiende moesson'. Het predicaat *waaiend* duidt een prototypisch kenmerk aan[4] en desambigueert het polyseme woord *moesson* in de zin van 'het periodiek voorkomende natuurelement wind'. Voorgaande redactioneel geselecteerde

[4] Zie voor een vergelijkbaar voorbeeld met prototypische kenmerken in een syntactische combinatie ook Harras (1989: 611), met een verwijzing naar een subject-werkwoordcombinatie voor 'vogel' en typisch daarmee verbindbare werkwoorden (*tjilpen*, *uitvliegen*, *nestelen*, enz.).

vindplaats illustreert het belang van prototypische eigenschappen voor voorbeeldselectie des te meer, daar in dit geval een hoofdbetekenis aan de orde is die geheel buiten beeld blijft in de eerste 15 GDEX-voorbeelden.

Niet alleen voor betekenissen maar ook voor combinatietypes laten zich gemakkelijk voorbeelden aanwijzen waar prototypische karakteristieken aanleiding waren voor redactionele voorbeeldkeuzes die GDEX naar de posities 16-49/51 doorschuift; zo bijvoorbeeld bij het lemma *bubbel* waar het trefwoord als subject vaak in typische combinatie voorkomt met werkwoorden als [*opborrelen*, *uiteenspatten*], en bij de dansstijl *jive,* waar zich voor het combinatietype adjectief-substantief kenmerken laten optekenen als [*pittig*, *vlug*] die de dansstijl inderdaad typeren, zoals Google met beeldmateriaal illustreert.

Naast de relevantie van prototypische kenmerken bij voorbeeldkeuzes onderstreept Harras het belang van voorbeeldzinnen die collocaties bevatten om de betekenisonderscheidende werking die dergelijke typische verbindingen vaak hebben. Illustratief is in dit verband onder meer het werkwoord *plengen*. Het idiomatische verband met typische objecten als [*wijn*, *een offer*] leidt bij betreffend lemma tot de onderscheiding van de hoofdbetekenis ‘ritueel uitgieten’, die GDEX in de posities 1-15 niet illustreert.

Ten slotte wijst Harras erop dat betekenisverhelderende informatie in de context soms ook voorkomt in de vorm van semantisch verwante woorden zoals (bijna-)synoniemen en tegengestelde begrippen. Zo’n voorbeeld van een ANW-citaat met een synoniem, dat GDEX verderop positioneert, biedt het lemma *zoel*. Het synoniem verduidelijkt dat deze vindplaats niet in de eveneens voorkomende positieve betekenis ‘mild van temperatuur’ is te lezen, maar een attestatie vormt voor de negatieve betekenisomschrijving ‘onaangenaam vochtig en warm, drukkend’:

Het is een **zoele**, heiige dag, die de zon krachtig laat vermoeden maar niet echt toont.

De voorgaande voorbeelden hebben betrekking op een drietal types van interpretatiesturende informatie in de zinscontext, die concordanties geschikt maakt om definities te documenteren. Vaak liggen nog andere inhoudelijke aspecten van de relatie tussen definitie en voorbeeldzin ten grondslag aan redactionele voorbeeldkeuzes. Zoals de vakliteratuur adviseert en de 330 voorbeeldzinnen laten zien, worden voorbeeldzinnen vaak aangewend om definities die noodzakelijk generaliserend of meer abstract van aard zijn, te concretiseren en tevens met nieuwe gegevens te complementeren (Zgusta 1971: 264-265, Nikula 1985: 188-189). Een informatief voorbeeld biedt het ANW-lemma *limerick*. De ANW-definitie stelt als kern dat het om een ‘vijfregelig puntdicht’ gaat, en voegt daaraan kenmerken toe betreffende het rijmschema, metrum en de typische inhoud zoals de introductie van een persoon en plaatsnaam in de eerste regel en de grappige wending in de laatste. De voorbeeldzin hierna concretiseert verder de verhouding van de regels, de aard en inhoud van een dergelijke tekst en geeft bovendien een concreet voorbeeld dat ook de werking van rijm en metrum toont:

In de **limericks** van Lear, die bestaan uit vijf regels, waarvan de laatste regel de eerste ten dele herhaalt, wordt in de tweede regel verteld wat er met iemand aan de hand is, in de derde en vierde wat er aan gedaan wordt en in de laatste regel wordt de oplossing van het probleem gegeven, of alleen maar geconstateerd dat het nu eenmaal zo is en dat er niets aan te doen is. […] Maar meestal gaat het drastischer toe: there was an old person of Rimini, who said, "gracious! goodness! o gimini!” When they said, "please be still!” she ran down a hill, and was never more heard of at Rimini.

Redactioneel wel gehonoreerd op inhoudelijke gronden, krijgt dit voorbeeld echter een lager waarderingscijfer op basis van de GDEX-criteria, waarbij mogelijk onder meer de zinslengte in het nadeel speelde. Een lagere waardering kreeg ook het volgende voorbeeld bij het lemma *larynx*. Ook hier zien we concretisering en complementering, meer bepaald bij de

synoniemdefinitie ‘strottenhoofd’, door de benoeming en situering van orgaandelen. Samen met de mogelijkheden van digitaal beeldmateriaal kan dit soort voorbeelden een definitie beslist ondersteunen:

De keel is aan de bovenkant begrensd door de neusholte; de scheiding tussen beiden wordt gevormd door de huig. Aan de voor-onderkant wordt de keelholte begrensd door het strottenhoofd (de **larynx**); deze kan worden afgesloten door middel van het strotklepje. De achter-onderkant wordt begrensd door de slokdarm.

Op een belangrijk criterium in verband met voorbeelden die de definitie aanvullen, wijst ten slotte nog Cowie. Toegevoegde waarde leveren namelijk ook concordanties die de culturele context toelichten (Cowie 2002: 77), zoals in het lemma *secretaire* uit ons proeftraject. Waar de ANW-definitie abstraheert op grond van functie, vorm en gebruik van betreffend meubel, informeert het volgende citaat ook over de chronologie, de herintroductie en de status:

Een meubelstuk dat de jongste jaren werd opgediept van onder het stof der eeuwen, is de **secretaire**. De 17de-eeuwse glorie van dit stuk huisraad herleeft in een moderne uitvoering. In veel gevallen is de secretaire een siermeubel, dat door de geraffineerde vormgeving een portie humor of verfijning in het interieur brengt.

In het voorgaande werd de keuze van concordanties vooral belicht vanuit de samenhang die bestaat tussen voorbeeldzin en definitie. In het navolgende gaan we kort in op nog twee andere criteria, waarvan de lexicografische vakliteratuur het belang onderstreept. Meer specifiek is een tweede rol voor voorbeeldzinnen weggelegd op het vlak van de illustratie van contextuele kenmerken zoals syntaxis, collocatie, register enz. (Atkins en Rundell 2008: 454). Overeenkomstig zal ook de ANW-redacteur vaak voorbeelden kiezen in functie van bijvoorbeeld combinatietypes. Ik beperk me tot enkele elementaire syntactische verbanden. Bij de combinaties en verbindingen van een substantief en een voorafgaand adjectief laten zich bijvoorbeeld optekenen: [*diepe*] *kerf* en, in de medische sfeer, [*peritoneale*] *dialyse* dat voorbij de vaktalige grenzen ook voorkomt in krantenrubrieken en op websites voor nierpatiënten. Beide voorbeelden zijn daarbij representatief voor de situatie dat Sketch Engine wel de hoogste score toekent aan een bepaalde combinatie (*diep* 5x, *peritoneaal* 6x) maar dat GDEX het betreffende syntagma toch niet sorteert naar het eerste derde van de concordanties. Ook selecteert GDEX voor dit eerste segment niet altijd een voorbeeld wanneer een combinatietype door een ruimere klasse van eenmalige concordanties wordt vertegenwoordigd. Illustratief zijn bijvoorbeeld de redactioneel wel gehonoreerde combinatietypes [*bronchiaal*, *coronair*, *invaliderend*, *pijnlijk*] *spasme* en [*duidelijke*, *verrassende*] *clou*.

Een vergelijkbare situatie doet zich geregeld voor bij de combinaties van werkwoorden met een object (bv. *toucheren* ‘licht of even raken’ met het object [*elkaar*]), of met een bijwoord ([*behoorlijk*, *flink*, *grif*] *dokken*). In dit opzicht vormen ook de scheidbare werkwoorden een traditioneel lastige categorie voor software. Zo behoren alle voorbeelden voor de combinatie *een dronk uitbrengen* tot de groep van 330 lager gewaardeerde concordanties.

Een derde en laatste redactioneel criterium dat we hier nog voor het voetlicht brengen, houdt verband met het toepassingsbereik van woorden. Zo komen bij sommige woorden naast het algemene gebruik in eigenlijke toepassing ook gespecialiseerde betekenissen voor. Andere woorden worden dan weer toegepast in concreet en abstract, of in een eigenlijk en metaforisch of metonymisch verband. Ook om deze semantische, contextgerelateerde ‘range of application’ te demonstreren zijn voorbeelden van groot nut.[5] Zo’n breder toepassingsbereik wordt in het traject van onze steekproef geïllustreerd bij het lemma *dialyse*. In zijn algemene toepassing kan men dit woord definiëren als ‘het onttrekken van bepaalde stoffen aan een

[5] Zgusta 1971: 263-264 en Landau 2001: 210-211.

oplossing door stoffen met grotere en kleinere deeltjes te scheiden door middel van een halfdoorlaatbare, filterende wand'. De voorbeelden die het ruimere gebruik van dit scheikundige procedé illustreren, behoren tot de klasse van 330 die GDEX in de posities 16-49/51 onderbrengt. Posities 1-15 daarentegen, worden geheel besteed aan de tweede, gespecialiseerde betekenis die betrekking heeft op de kunstmatige zuivering van bloed bij onvoldoende nierfunctie. Aan deze betekenisverhouding ligt hier mogelijk betekenisspecialisering of verkorting uit *nierdialyse* ten grondslag. Een andere semantische verhouding doet zich voor bij het werkwoord *mangelen*. Om het toepassingsbereik in beeld te brengen, moeten we zowel de eigenlijke toepassing met betrekking tot wasgoed documenteren, als de overdrachtelijke, metaforische betekenis 'iemand of iets zwaar in de verdrukking brengen'. Ook hier sorteert GDEX de redactioneel geselecteerde voorbeelden voor de eigenlijke betekenis naar de posities 16-49/51. Dat is eveneens zo bij *kalebas*, waar de voorbeelden ditmaal een toepassingsbereik laten zien met een metonymische betekenisverhouding tussen de eigenlijke plantnaam en de voortgebrachte vrucht.

4. Opties voor verbetering

Het is begrijpelijk dat de complexe vertaalslag van dergelijke redactionele afwegingen naar programmatuur misschien niet meteen euforie opwekt bij ontwikkelaars van lexicografische software. Het is echter de vraag of uitzicht op oplossingen hier geheel ontbreekt. Naar zich laat aanzien, dienen zich onder meer twee benaderingswijzen aan om de hierboven onderscheiden groep van 330 te verkleinen of om, met andere woorden, een lexicografische sorteerfunctie meer te doen aansluiten bij de redactionele voorbeeldkeuze. Een eerste benaderingswijze betreft het criterium 'collocaties' dat in de reeks waarop GDEX zich baseert al voorkomt, maar blijkbaar zwaarder zou kunnen wegen in verhouding tot de kenmerken zinslengte en woordfrequenties. Grotere nadruk op de syntactische combineerbaarheid en de aanwezigheid van typische trefwoorden in deze combinaties zou voor tal van informatieve zinnen uit de groep van 330 wel een prominente plaats opleveren in de sortering van het lemmamateriaal. Vaak gaat het om elementaire grammaticale verbanden, zoals deze van werkwoord en subject. Ik wijs in dit verband slechts op een concordantie van het lemma *cartograaf.* Aandacht voor de relatie subject-werkwoord leidt er tot opname van een voorbeeld dat de kenmerkende bezigheid van deze beroepsgroep aanduidt: tekenen. De elementaire grammaticale relatie subject-werkwoord illustreert hiermee bovendien haar nut om, overeenkomstig een hiervoor al besproken lexicografisch advies, prototypische kenmerken te achterhalen:

Cartografen tekenen de plattegrond van deze fictieve stad, architecten ontwerpen de gebouwen.

Ook door andere elementaire syntactische verbanden en typische trefwoorden daarbinnen zou een sorteerfunctie voor concordanties redactioneel wél geselecteerde combinaties naar voren kunnen halen. In onze steekproef geldt dat onder meer voor voorbeeldzinnen die relaties documenteren zoals deze van werkwoord en object (bijvoorbeeld *een crucifix* [*ophangen, kussen*], [*een ambt*] *postuleren*), van een werkwoord met een voorzetselbepaling (*lurken aan* [*een rietje*]) of van een substantief en een adjectief ([*gouden*, *voorbije*] *epoque, melomaan* [*publiek*]). Wil men een sorteerfunctionaliteit in dit opzicht verder optimaliseren, dan zal dat omgekeerd ook eisen stellen aan het aangeleverde corpusmateriaal zelf. Meer bepaald zal de tagging ervan op voldoende niveau moeten staan. Ook parsering van het corpus zal waarschijnlijk tot betere resultaten leiden.

Een tweede weg voor verbetering van de voorbeeldsortering is van meer inhoudelijk-semantische aard en heeft betrekking op de zogenaamde contextanten. Dit zijn woorden die buiten de echte collocaties in een ruimere context inhoudelijk verbindbaar zijn met het

trefwoord, zoals bijvoorbeeld ‘passaatwind’ en ‘cycloon’ uit de contextantenreeks in het lemma *moesson*. Gebruik van dergelijke woorden als een van de zoekcriteria bij voorbeeldsortering zou juist de gemiste hoofdbetekenis met de betekeniskern ‘halfjaarlijkse wind’ binnen de eerste 15 posities hebben geattesteerd. Zo bijvoorbeeld in de volgende concordantie die nu na automatische GDEX-sortering geheel achteraan in het lemmamateriaal belandde:

Daarboven waaien ze, sirocco’s, **moessons**, passaatwinden, zelfs cyclonen, al die winden die ik ken uit Aardrijkskunde.

Dat contextanten nuttig zijn voor voorbeeldsortering wekt niet echt verbazing. Zoals Atkins en Rundell (2008: 294) kernachtig formuleren: ‘context disambiguates’ en contextanten hebben in de lexicale, betekenissturende omgeving nu eenmaal een relevant aandeel.[6]

Het voorgaande voorbeeld is beslist geen alleenstaand geval. Waar *GDEX* vindplaatsen voor betekenissen en subbetekenissen mist in het eerste derde van het respectievelijke lemmamateriaal, zouden contextanten juist aansluiting bij de redactionele voorbeeldkeuze bevorderen. Dat is onder meer zo voor *baldakijn*, *gouvernement*, *mangelen*, *tartaar*, *titelen* enz.[7] Ook in betekenissen waar GDEX wel redactioneel geselecteerde of gelijkwaardige voorbeelden opneemt, vormen contextanten een middel om meer redactioneel geselecteerde concordanties vooraan te plaatsen. Enkele voorbeelden hiervan zijn de al vermelde trefwoorden *larynx* (in de hiervoor geciteerde aanhaling op grond van de contextanten ‘keelholte, strottenhoofd, strotklepje’) en *limerick*, alsook het lemma *halo* in de betekenis 2.0 ‘ongewenst, wazig lichtschijnsel op beeldopnames dat ontstaat rond heldere beeldgedeelten of langs sterk contrasterende omtreklijnen’. In de volgende redactioneel geselecteerde aanhaling komen woorden uit de contextantenrubriek voor als ‘belichting, emulsie, film’:

Tenslotte zit dan nog helemaal onderaan een anti-halolaag die voorkomt dat sterk licht terug in de emulsie wordt gereflekteerd en daar op de verkeerde plaats voor een tweede belichting, een **halo**, zorgt. […] Een zwart-wit film heeft naast de beschermingslagen één emulsielaag.

Ten slotte zijn contextanten eveneens nuttig om concordanties op te sporen die de culturele context verduidelijken. In feite dragen zij daardoor bij aan de toepassing van een criterium waarvan Cowie het belang aangaf. Het volgende ANW-citaat dat GDEX naar de posities 16-49/51 verschoof, informeert over een bepaalde reiscultuur in verband met een ruimer verspreid leenwoord in het Nederlands: *hostel*. Uit de lemmarubriek met de contextanten treffen we hier *budgetreiziger*, *backpacker* en *jeugdherberg* aan:

In Australië kent men twee soorten hostels of herbergen, speciaal voor budgetreizigers: de privé ‘backpacker’hostels en de YHA hostels (jeugdherbergen). Beide bieden accommodatie met kookfaciliteiten in ‘een gastvrije, ontspannen sfeer. […] In totaal zijn er meer dan 130 hostels, toegankelijk ongeacht de leeftijd, met slaapzalen, keukens en zitkamers waar u andere reizigers uit de hele wereld ontmoet.

5. Slotbeschouwing

[6] Daarbij is het van belang aan te stippen dat, waar GDEX zich op de syntactische eenheid van de zin richt, betekenisvolle contextanten vaak ook buiten de zinsgrenzen voorkomen (Kilgarriff e.a. 2008: 427).

[7] Contextanten die concordanties voor de ontbrekende betekenissen zouden prioriteren zijn bijvoorbeeld: *beeld*, *gotisch*, *monument*, *voorgevel* bij *baldakijn* 1.1 ‘overkapping uit beeldhouwwerk of houtsnijwerk’; *detachement*, *innemen*, *militair* bij *gouvernement* 4.0 ‘militair bestuur over een stad of zone’; *droger*, *wasmachine* bij *mangelen* 1.0 ‘wasgoed kreukvrij maken’; *mayonaise*, *pickels*, *saus* bij *tartaar* 2.0 ‘tartaarsaus’; *boek*, *literair*, *publiceren* bij *titelen* 2.0 ‘een boek een titel geven’.

De voorgaande gegevens laten zien dat achter het aanvinken van een functionaliteit als automatische voorbeeldsortering een veelzijdige problematiek schuilt. Dat de software in eerste instantie vooral op meer formele criteria werd gebaseerd, is begrijpelijk. Daarnaast blijkt ook de noodzaak van een oplossing voor inhoudelijke aspecten van lexicografische voorbeeldselectie. Deze laatste heeft in verschillende opzichten betrekking op de relatie tussen definitie en voorbeeldzin, alsook op contextuele kenmerken en het semantisch toepassingsbereik van woorden. In sommige gevallen kan dat implicaties hebben voor de GDEX-criteria. Ik vermeld slechts de tweede maatstaf die concordanties lager waardeert naarmate woorden voorkomen die minder gebruikelijk of zeldzaam zijn. Alleen al in het traject van onze steekproef laten zich tal van voorbeelden aanwijzen die stammen uit een breed en divers grensgebied tussen algemene en meer gespecialiseerde taal of zelfs vaktaal die via media, internet en onderwijs frequent doordringt tot brede kringen van taalgebruikers. Voorbeeldzinnen met dergelijke trefwoorden (en hun contextanten) die juist voor raadpleging in een woordenboek in aanmerking komen, zouden bij een evenwichtige balans voldoende geprioriteerd moeten worden voor eventuele opname in een woordenboekartikel.

Hoewel er beslist de nodige ruimte voor verbetering is, moet men zich echter ook realiseren dat er altijd wel discrepantie zal bestaan tussen menselijke keuzes en wat computationeel haalbaar is, zowel door de complexiteit van taal als door de veelheid aan lexicografische afwegingen. Dat besef is niet alleen goed voor het gevoel van zelfwaarde van de redacteur, maar het behoedt hem ook voor de naïeve gedachte dat met de spreekwoordelijke ‘druk op de knop’ een computationele ‘deus ex machina’ hem kan komen redden.

Bibliografie

Atkins, B.T.S. en M. Rundell (2008), *The Oxford guide to practical lexicography*. Oxford University Press.

Cowie, A.P. (2002), ‘Examples and collocations in the French “Dictionnaire de langue”’, in: Corréard, M.-H. (ed.), *Lexicography and Natural Language Processing*. *A Festschrift in honour of B.T.S. Atkins*, Euralex, 73-90.

Harras, G. (1989), ‘Zu einer Theorie des lexikographischen Beispiels’, in: *Wörterbücher. Dictionaries. Dictionnaires*. *Ein internationales Handbuch zur Lexikographie*. Band I. Berlin-New York, Walter de Gruyter, 607-614.

Kilgarriff, A., M. Husák, K. McAdam, M. Rundell, P. Rychlý (2008), ‘*GDEX*: Automatically Finding Good Dictionary Examples in a corpus’, in: *Proceedings of the XIII EURALEX International Congress*. 1. Barcelona, 425-432.

Landau, I. (2001), *Dictionaries*. *The art and craft of lexicography*. Cambridge University Press.

Nikula, H. (1986), ‘Wörterbuch und Kontext. Ein Beitrag zur Theorie des lexikalischen Beispiels’, in: A. Schöne (Hrsg.), *Kontroversen, alte und neue*. Bd. 3. Tübingen, Max Niemeyer Verlag, 187-192.

Zgusta, L. (1971), *Manual of lexicography*. Prague / The Hague / Paris, Academia / Mouton.