Foto: Anda van Riet

De toepassing van digitale corpora binnen (en buiten) het talenlokaal

Digitale corpora bieden steeds meer mogelijkheden om op systematische wijze te kijken naar authentiek taalgebruik, geproduceerd door 'echte' mensen en ontsloten via grote, gratis toegankelijke online tekstbanken. Ook de didactische toepasbaarheid van corpora, searchtools en lesmaterialen groeit gestaag. Tijd voor een kijkje in de keuken van de corpora.

David Geneste

Het gebruik van corpora in de (taal)les staat nog in de kinderschoenen, enkele bescheiden initiatieven uitgezonderd. Een reden hiervoor is de terughoudendheid van leraren; zij moeten immers affiniteit met taalkunde en programmatuur hebben, en hun leerlingen een onderzoekende en systematische werkhouding aanleren. Geen makkelijke taak dus. Toch wordt in onder meer Groot-Brittannië en Duitsland al geruime tijd geëxperimenteerd met het didactisch potentieel van corpora. Bijvoorbeeld om te kijken hoe komma's echt worden gebruikt, of welke signaalwoorden wel of niet worden ingezet in gesprekken.

De (toegepaste) corpuslinguïstiek is een relatief recente tak van de taalkunde die zich bezighoudt met het samenstellen en analyseren van grote, digitale corpora die aan de hand van speciale exploitatiesoftware doorzocht kunnen worden. Ter illustratie: het internet is een soort mondiaal corpus, met Google als zoekmachine. Probleem hierbij is dat we (nog!) niet echt gericht naar specifieke taalvormen kunnen zoeken, en dat het niet altijd duidelijk is wat de bron of herkomst van een tekst is. Professionele corpora bestaan wel uit authentiek en geverifieerd taalgebruik gebaseerd op grote hoeveelheden geschreven of gesproken en geannoteerde bronnen, bijvoorbeeld tijdschriftartikelen of nieuwsuitzendingen. Taalkundigen gebruiken deze corpora bijvoorbeeld om empirisch te kijken naar hoe het Engelse verbindingswoord *however* zich in het 'echt' gedraagt.

In Nederland wordt op relatief kleine schaal met corpora gewerkt, maar in Engelstalige gebieden bestaan er verschillende van dergelijke tekstbanken. Een groot (450 miljoen woorden) en veelgebruikt Engelstalig corpus is *The Corpus of Contemporary American English* (COCA, <http://corpus.byu.edu/coca>), dat via Brigham University in de VS beschikbaar wordt gesteld en ook in deze bijdrage als voorbeeld dient. Hoewel Engelse corpora (Britse en Amerikaanse) zeer goed vertegenwoordigd worden op het internet, zijn er ook voor andere talen corpora te vinden. De instructietaal is dan vaak wel Engels.

Corpora werden in eerste instantie ontwikkeld en gebruikt door taalwetenschappers. Zo zijn alle voorbeeldzinnen in de Engelstalige Cobuild-woordenboeken ontleend aan de *Bank of English*, een groot wetenschappelijk corpus opgezet door de universiteit van Birmingham. Maar na enige training kunnen ook leraren en leerlingen corpora inzetten als les- en leermiddel, temeer omdat sommige corpora, bijvoorbeeld COCA, uitgerust zijn met een laagdrempelige zoekmachine en online tutorials. Didactisch uitgangspunt bij het gebruik van corpora in de taalles is het *data-driven learning*, een taalverwervingsmodel waarin leerlingen met grote hoeveelheden taalgegevens (in corpora) worden geconfronteerd. Corpora bieden dus nooit kant-en-klare (woordenboek)oplossingen, maar wel systematische input die de onderzoekende leerder gebruikt om patronen te zien en tot conclusies te komen (Bennet, 2010).

Corpora bieden nooit kant-en-klare (woordenboek)oplossingen, maar wel systematische input die de onderzoekende leerder gebruikt om patronen te zien

Gebruiksmogelijkheden van corpora voor leerling en leraar

Het kweken van een actieve werkhouding bij leerlingen is geen sinecure, zoals we weten, en het bijbrengen van de nodige technische handigheden en termen die nodig zijn, ook niet. Waarom zouden we corpora dan überhaupt gebruiken? Ten eerste zijn online corpora veelal gratis toegankelijk, en voorzien van software en tools om mee te werken. Met het oog op steeds omvangrijker wordende digitale leeromgevingen, kan het onderwijs hier z'n voordeel mee doen. Een meer inhoudelijke reden is dat corpora authentiek taalgebruik bieden; een corpus is immers een verzameling van echte en geverifieerde taaluitingen in een specifieke context. Een academisch corpus bijvoorbeeld bestaande uit essays van eerstejaarsstudenten wordt altijd gecontroleerd en zo nodig verbeterd zodat grove fouten worden vermeden. Vervolgens kan een corpus ondervraagd worden met een zoekopdracht. Hiervoor zijn verschillende searchtools beschikbaar, die het mogelijk maken gericht te zoeken naar een woord of woordcombinatie. Stel dat een leerling meer wilt weten over het verbindingswoord *however*. Door te zoeken op een *KWIC* (Key Word in Context) krijgt hij dit sleutelwoord in tal van zinsverbanden te zien. Het sleutelwoord wordt steeds in een afwijkende vorm of kleur weergegeven en onder dezelfde sleutelwoorden in andere zinnen geplaatst, zodat de omgeving snel gescand, gelezen en geïnterpreteerd kan worden. Ook is er informatie over de bron of het genre te vinden. (Zie figuur 1 voor een voorbeeld uit het British National Corpus.)

Het is natuurlijk juist de interpretatie van al die gegevens die leerlingen als lastig ervaren. Want wat moeten ze nou eigenlijk met die informatie? Essentieel voor een succesvolle verwerking is dat leerlingen een gerichte zoekopdracht meekrijgen en uiteindelijk eigen relevante vragen weten te formuleren. Bovendien dient de vraag gekoppeld te zijn aan een voor de leerlingen relevante taak, bijvoorbeeld het analyseren van een gesprek of het reviseren van een werkstuk.

Leraren kunnen ook hun voordeel doen met corpora door ze zelf samen te stellen op basis van de input van hun leerlingen. Een dergelijk *learners' corpus* kan aangeven wat specifieke probleemgebieden zijn. Op deze wijze kan een veel preciezere *needs analysis* opgesteld worden. Er wordt ook steeds meer relatief eenvoudig te bedienen software ontwikkeld, waarmee individuen zelf corpora kunnen maken en analyseren. *TextStat*, gemaakt door de Freie Universität in Berlijn, is zo'n handig programma dat eenvoudig digitale tekstbestanden leest (<http://neon.niederlandistik.fu-berlin.de/textstat>). Voorwaarde is wel dat leraren zelf een corpus aanmaken, bijvoorbeeld door geschreven producten van leerlingen in een tekstverwerkingsbestand te knippen en te plakken. Bovendien moet het corpus duidelijk omschreven worden in termen van onder meer omvang, niveau en aanleiding, zodat later altijd achterhaald kan worden wat de status van het *learners' corpus* is.

Vanzelfsprekend vergt het gebruik van corpora oefening en geduld. Zeker wanneer men met meer functies in een corpus wilt werken (zoeken naar concordanties of collocaties, bijvoorbeeld) is een systematische werkhouding geboden. Uit een enquête gehouden onder vwo 5-leerlingen blijkt dat zij het nut van een corpus wel zien, zeker als hulpmiddel bij revisie of als alternatief voor een woordenboek, maar dat ze het selecteren van relevante gegevens uit grote hoeveelheden informatie intimiderend vinden. Maar deze vrees voor grote hoeveelheden informatie leeft natuurlijk niet alleen bij de moderne vreemde talen.

Lesopdracht voor het vak Engels in de bovenbouw

Om voorgaande concreter te maken is een op een corpus gebaseerde opdracht voor het vak Engels bijgevoegd, gebruikmakende van de basisfuncties van COCA (zie kader 1). Deze opzet kan met de nodige ingrepen ook aan anderstalige corpora aangepast worden (zie kader 2). Het voorbeeld is ontleend aan een analyseonderzoek naar het gebruik van Engelse verbindingswoorden. Daarbij is ook getracht aan aantal knelpunten te remediëren door leerlingen met corpora te laten werken.

```
s enshrined in the Constitution of the Republic of Namibia.  HOWEVER,  in exercising their press freedom, media practition
se monuments forty centuries look down upon you». Napoleon,  HOWEVER,  was given little time to enjoy his victory. A Briti
poverty spread throughout almost every part of the nation".  HOWEVER,  most workers stress that rural deprivation isn't a
e into account the biological and genetic factors of nature  HOWEVER,  in believing that everything is learnt and nothing
ional way, he decided it might make an attractive necklace.  HOWEVER,  when his mother, Adele Britton, tried to remove it,
urvey organizations can undertake the large surveys needed.  HOWEVER,  a good number of geographers have conducted recreat
y Historia Brittonum , formerly attributed to Nennius(where  HOWEVER   Rowena is not named). It was filled-out and popular
ound at court functions, she was unnoticed by the nobility.  HOWEVER,  she had no intention of revealing her gratification
sible, yet ROS betrays no surprise at all -- he feels none.  HOWEVER,  he is nice enough to feel a little embarrassed at t
l Omar Torrijos, the support to defeat a coup. Gen Noriega,  HOWEVER,  soon started freelancing for Cuban, Israeli, and Ta
```

Figuur 1. Output (deels) van zoekopdracht 'however' in het British National Corpus

Voor deze opdracht gaan jij en een partner het gebruik van het Engelse verbindingswoord *however* analyseren. Zoals je weet zijn verbindingswoorden erg belangrijk om je schrijf- en spreekvaardigheid samenhangend en vloeiender te maken, maar ze kunnen soms lastig zijn in het gebruik. Je doet deze opdracht samen met behulp van COCA, het grootste Engelstalige corpus ter wereld.

STAP 1
Ga naar <http://corpus.byu.edu/coca/> en type *however* in de werkbalk. Klik vervolgens op *however* in het nieuwe veld en je krijgt duizenden KWIC's (Key Words in Context).

STAP 2
Zoals je weet kan *however* verschillende betekenissen hebben. Zoek voor elk van de drie verschillende betekenissen van *however* hieronder *vijf* voorbeeldzinnen in COCA. Bespreek met je partner of er inderdaad sprake is van een van de drie volgende betekenissen; bij twijfel neem je een ander voorbeeld.
- *however* voor tegenstelling
- *however* voor versterking
- *however* voor voortzetting

STAP 3
Knip en plak alle voorbeeldzinnen in een tekstbestand en verdeel ze over de drie categorieën. Geef ook aan wat de bron is en of het een gesproken (SPOK) zin is.

STAP 4
Bepaal voor elk van de drie categorieën regelmatige patronen. Deze moeten betrekking hebben op:
- de plaats van *however* in de zin;
- het gebruik van komma's;
- gesproken of geschreven tekst.

STAP 5
Schrijf je conclusies zo helder mogelijk op en deel ze met je groepsgenoten.

Kader 1. Op een corpus gebaseerde lesopdracht voor het vak Engels

TAAL	WEBSITE CORPUS	KENMERK
Duits	<http://corpora.ids-mannheim.de/ccdb/>	geheel Duitstalig
Frans	<www.lextutor.ca/concordancers/concord_f.html>	Frans–Engels
Italiaans	<http://badip.uni-graz.at/>	gesproken taal
Nederlands	<http://lands.let.kun.nl/cgn/home.htm>	geheel Nederlandstalig
Spaans	<www.lextutor.ca/concordancers/concord_s.html>	Spaans–Engels
Turks	<http://tscorpus.com/eng/>	Turks–Engels

Kader 2. Overzicht van gratis, online corpora voor verschillende talen

Als casus nemen we het verbindingswoord *however*. Aangezien hier de ruimte ontbreekt om een complete beschrijving van dit woord te geven, volstaan we met de aanname dat leerlingen drie problemen ondervinden met het gebruik van *however*: semantisch (de onderliggende relatiebetekenis wordt verkeerd begrepen, onder meer door verwarring met andere betekenissen van het woord), syntactisch (de positie in de zin en de aanduiding daarvan aan de hand van komma's is onnatuurlijk) en stilistisch (er bestaat verwarring tussen geschreven en gesproken, formele en informele taal). De opdracht richt zich dus specifiek op deze drie kenmerken en dwingt leerlingen na te denken over de betekenis van het woord in verschillende contexten, zelf naar patronen te zoeken en deze te internaliseren door ze onder woorden te brengen. Wat betreft de uitgangspositie: het gaat om havo/vwo-bovenbouwleerlingen die bekend zijn met connectiviteit, corpora en de basisfuncties van COCA. Ze hebben toegang tot een computer en krijgen vijftig minuten om aan de opdracht te werken. Bij een jongere of minder ervaren doelgroep kan er ook voor gekozen worden een beperktere, voorgeselecteerde en eventueel geprinte lijst met concordanties te gebruiken om te voorkomen dat leerlingen verdwalen in de grote hoeveelheid treffers en de searchtools.

Tot slot

In Nederland is nog niet onderzocht wat het rendement van een corpuslinguïstische aanpak in taal- en vaklessen is, maar er is geen reden om aan te nemen dat het effect hier minder zou zijn dan in andere landen. Kritiek heeft zich tot op heden (deels terecht) gericht op de relatief grote technische kennis die nodig is om aan de slag te gaan. Corpustools moeten daarom onderwijsvriendelijker gemaakt worden en voorzien van uitgewerkte en direct toepasbare lesapplicaties. Visuele aantrekkelijke woordwolken (zie bijvoorbeeld <www.wordle.net>) zijn daar een mooi voorbeeld van.

Lastiger is het gegeven dat veel leerlingen (en leraren!) gewend zijn om in termen van correct of incorrect taalgebruik te denken, in plaats van zelf conclusies te trekken over wat wel of geen acceptabel taalgebruik is. Voor de Nederlandse context bieden corpora vooral interessante perspectieven voor taal- en vakdocenten binnen het tweetalig onderwijs (ook bekend als CLIL). Een interessante vraag is bijvoorbeeld hoe corpora kunnen bijdragen aan de ontwikkeling van meer academisch taalgebruik onder leerlingen.

Leraren kunnen kennismaken met corpora op tal van zomercursussen, bijvoorbeeld aan de universiteiten van Birmingham en Lancaster. Ook is er inmiddels de jaarlijkse Teaching and Language Corpus Conference (zie <http://talc10.ils.uw.edu.pl>). Het voornaamste is echter gewoon eens aan de slag te gaan met COCA en te bekijken wat deze applicatie voor les en curriculum te bieden heeft. ■

Literatuur

Bennett, G. R. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. Ann Arbor, MI: University of Michigan Press.

Meer informatie

Anderson, W., & Corbett, W. (2009). *Exploring English with online corpora: An introduction*. Basingstoke, UK: Palgrave MacMillan.

Lackman, K. (2010). *Classroom games from corpora: Using corpora to teach vocabulary*. Te raadplegen via http://www.kenlackman.com/files/CorporaGamesBook103.pdf

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom*. Cambridge, UK: Cambridge University Press.

De Nederlandse uitgeverijen presenteerden de afgelopen weken weer stapels literatuur. Het is vrijwel ondoenlijk om de hele waslijst in kaart te brengen. Om die reden volgt hieronder een selectie van de meest bruikbare en in het oog springende titels.

Aan het begin van de roman *Margot* (Querido, 248 blz.) van **Sophie Zijlstra**, voelen de joodse zusjes Margot en Anne zich thuis in Amsterdam, waar hun ouders zich na de vlucht uit Duitsland gevestigd hebben. Maar in de zomer van 1941 verandert alles. Margot, de oudste van de twee, wordt de toegang ontzegd tot het meisjeslyceum, het tennispark, haar roeiteam en de ijsbaan in de Apollohal. Haar wereld, tot dan toe die van een zorgeloze tiener, wordt razendsnel steeds kleiner. Sophie Zijlstra deed onderzoek naar het korte leven van Margot Frank en baseerde deze roman op historische gegevens. Daardoor kan Margot nu voor het eerst uit de schaduw van haar beroemde zusje Anne treden.

De roman *Het meisje dat uit de lucht kwam vallen* (Anthos, 302 blz.) van de Engelse auteur **Simon Mawer** speelt zich af in het Engeland van de jaren veertig. Marian Sutro heeft zojuist de middelbare school afgerond en wil zich nuttig maken tijdens de oorlog. Als enige van haar medeleerlingen spreekt ze Frans, waardoor ze de aandacht trekt van Mr. Potter, een rekruteringsofficier voor de Britse geheime dienst. Marian wordt opgeleid tot spion en heeft een missie waarvan de uitkomst de afloop van de oorlog weleens zal kunnen beïnvloeden.

Van striptekenares **Barbara Stok** verscheen het album *Vincent* (Nijgh & Van Ditmar, 144 blz.), een *graphic novel* over het verblijf van de schilder Vincent van Gogh in het Zuid-Franse Arles. Vincent droomt ervan om daar een kunstenaarshuis te stichten voor zichzelf en zijn artistieke vrienden. Maar door aanvallen waarin hij volledig in de war is, uitmondend in het beruchte 'oorincident', valt die droom in duigen. Zijn broer Theo blijft hem onvoorwaardelijk steunen. Een fraaie en subtiele 'verstripping' van een even kleurrijk als dramatisch leven.

De roman *Terug naar het bloed* (Prometheus, 544 blz.) van de Amerikaan **Tom Wolfe** begint met het beeld van een politieboot die over de golven van Biscayne Baai voor de kust van Miami dendert. Aan boord bevindt zich de jonge agent Nestor Camacho. Uitgerekend deze jongeman, zoon van uit Cuba gevluchte ouders, jaagt geheel ongewild de Cubaanse gemeenschap tegen zich in het harnas. De roman geeft een dynamisch en gedetailleerd beeld van het geraas van deze tijd, vol rake typeringen en hilarische inzichten.

Werner Plöts, hoofdpersoon in de roman *De dag dat we Andy zijn arm afzaagden* (De Bezige Bij, 300 blz.) van de Vlaamse schrijver **Marnix Peeters**, trekt als tiener de wijde wereld in. Met aan zijn zijde Orzas, een vadsige beer die een poot mist, sukkelt hij van de regen in de drup. Wie hij ook ontmoet, er wordt voortdurend van deze te vriendelijke jongen geprofiteerd, hetzij voor geld, hetzij voor seks. Een eigentijds sprookje en een soms keiharde roman, waarin gruwelijke taferelen en hilarische passages elkaar afwisselen.

De biografie *Marten Toonder* (De Bezige Bij, 640 blz.) is het indringende levensverhaal van de geestelijk vader van Tom Poes en daarmee de meest bekende stripmaker die Nederland ooit gekend heeft. Alles aan en van Marten Toonder heeft minstens twee gezichten of kanten. Hij is schrijver en tekenaar, kunstenaar en zakenman, Nederlander en Ier, kluizenaar en acteur, realist en magiër, de argeloze en de slimme, een open en een gesloten boek. Over zijn leven en werk schreef **Wim Hazeu** (die eerder biografieën schreef over onder anderen Jan Jacob Slauerhoff en Gerrit Achterberg) een schitterend en indrukwekkend boek. ■

Jacob Moerman