Practicum Corpus Gesproken Nederlands / Corex

Eric Akkerman, Bureau Informatisering, Faculteit der Letteren, Vrije Universiteit (12-03-2013)[1]

A. Achtergrondinformatie

Het *Corpus Gesproken Nederlands* (*CGN*) is een databank van het hedendaags Nederlands zoals dat wordt gesproken door volwassenen in Nederland en Vlaanderen. Het corpus bevat circa 900 uur spraak en omvat ongeveer 9 miljoen woorden: plm. 3,3 miljoen woorden daarvan zijn afkomstig uit Vlaanderen, ruim 5,6 miljoen woorden werden opgenomen in Nederland. Het *CGN* bestaat uit een groot aantal fragmenten van spraakopnames. Al het materiaal is orthografisch (d.w.z. volgens de geldende spelling) getranscribeerd, waarbij de orthografische transcriptie gekoppeld is aan het spraaksignaal. De orthografische transcriptie vormde het uitgangspunt voor de lemmatisering (aanduiding van stamvormen) en de verrijking van het materiaal met woordsoortinformatie (*parts of speech – POS*). Verder is er voor een selectie van één miljoen woorden een brede fonetische transcriptie vervaardigd en is een deel van het materiaal door middel van een syntactische analyse verrijkt met informatie over zinsopbouw. Ten slotte is een bescheiden deel van het corpus, circa 250.000 woorden, van een prosodische analyse voorzien. Deze analyses zijn allemaal inhoudelijke verrijkingen van het corpus. Men spreekt in dat verband meestal van annotaties. Het CGN heeft verschillende annotatielagen. Hieronder staat ter illustratie een deel van de informatie m.b.t. de uiting "`nou je hebt ze in uh uh rond en vierkant`".

```
5  17267 21281 N01002 fn000248.6
ORT nou je hebt ze in uh uh rond en vierkant.
POS BW() VNW(pers,pron,nomin,red,2v,ev) WW(pv,tgw,met-t)
    VNW(pers,pron,stan,red,3,mv) VZ(init) TSW() TSW()
    ADJ(vrij,basis,zonder) VG(neven) ADJ(vrij,basis,zonder) LET()
LEM nou je hebben ze in uh uh rond en vierkant .
```

Doordat het CGN verschillende annotatielagen heeft, is een eenvoudig programma als WordSmith minder geschikt om erin te zoeken. Bovendien is het met WordSmith zeker niet mogelijk om gegevens uit verschillende annotatielagen met elkaar te combineren, noch om efficiënt gebruik te maken van de metadata van het corpus. Daarom is in het kader van het CGN-project ook een specifiek zoekprogramma ontwikkeld. Dit programma heet *Corex*, hetgeen staat voor CORpus EXploitatie software. De Corex-gebruiker kan luisteren naar spraakbestanden, kan verschillende annotaties bekijken en zoekacties uitvoeren op het CGN. Corex ondersteunt een makkelijke navigatie door de subcorpora, gebaseerd op voorgedefinieerde of door de gebruiker gedefinieerde groeperingen zoals het geslacht van de spreker, de leeftijd en diverse andere beschrijvende gegevens (ook wel metadata genoemd). Het spraaksignaal kan synchroon worden afgespeeld met de annotatiegegevens. Overigens kan dat niet in de facultaire pc-zalen: de geluidsbestanden zijn daar vooralsnog niet beschikbaar.

Meer informatie over het CGN is te vinden via het facultaire corpusoverzicht:
http://www2.let.vu.nl/intranet/resources/corpora/
Meer informatie over Corex is te vinden via de facultaire Werkbank Taalonderzoek:
http://www2.let.vu.nl/intranet/resources/werkbanken/taalonderzoek/programmatuur/corex.php

[1] Met dank aan Theo Janssen en Wilbert Spooren voor hun commentaar op een eerdere versie.

B. Het practicum

In dit practicum leer je een aantal basisvaardigheden met betrekking tot Corex, die je in staat stellen om de betreffende dossieropdrachten van de cursus Taalgebruikstheorie uit te werken. Aan de orde komen het specificeren van deelcorpora, het zoeken naar woorden en woord-combinaties in de orthografische annotatielaag en het bekijken van de sociale dimensies van de sprekers. Dit is slechts een fractie van alle mogelijkheden van het programma.

B.1. COREX starten

In de facultaire pc-zalen kun je Corex starten via het Start-menu:

Start > Programma's > Corex > Corex

Let op: het duurt enige tijd voor het programma is gestart; tot die tijd zie je een venster waarin allerlei instellingen worden geregeld (en dat er een tijdlang nogal statisch kan uitzien, zodat het lijkt alsof er niets gebeurt…). Als het hoofdvenster van Corex is verschenen, kun je aan de slag.
N.B. Als het instellingenvenster op je scherm blijft staan, kun je dit het beste even minimaliseren met behulp van de betreffende systeemknop  rechts boven in de titelbalk.
→ Sluit dit venster niet, want daarmee sluit je tevens Corex!

B.2. Het definiëren van een deelcorpus (eenvoudig, maar met beperkingen)

Binnen Corex kun je op een aantal manieren een deelcorpus definiëren voor je zoekvraag. De meest eenvoudige (maar beperkte – zie par. B.5.) manier om dit te doen, is door in het hoofdvenster een *bookmark* te selecteren en vervolgens in het kader *Metadata Description Tree* een deelcorpus op te bouwen. De volgende *bookmarks* zijn voor dit practicum het meest relevant:

Components	Hiermee kun je bepaalde teksttypen selecteren (conversaties, telefoongesprekken, nieuwsuitzendingen, etc.).
Regions	Hiermee kun je corpussessies selecteren op basis van land (met name Nederland of Vlaanderen) en regio waar de spreker(s) woonden tussen hun 4^{e} en 16^{e} levensjaar. N.B. Dit geeft dus geen informatie over de plaats waar het betreffende geluidsfragment is opgenomen, maar over de herkomst van de spreker(s).
Age/sex	Hiermee kun je sessies selecteren op basis van het geslacht en/of de leeftijd van de sprekers.

Je gaat nu een deelcorpus selecteren dat bestaat uit spontane conversaties.

1. Start Corex, wacht totdat het hoofdvenster is verschenen en maak dit schermvullend.
2. Dubbelklik op de *bookmark* "Components".
3. Klik vervolgens in het kader *Metadata Description Tree* éénmaal op de sleutel voor het woord "Components" of dubbelklik op de "C" voor dat woord om een overzicht te krijgen van alle teksttypes in het corpus.

4. Dubbelklik dan op het teksttype “spontaneous conversation (face-to-face)”. Vervolgens worden alle sessies opgesomd die deel uitmaken van deze categorie; er is hier dus geen verdere indeling.
5. Dubbelklik op de naam van het eerste fragment. Je ziet voor deze sessie dan de verwijzingen naar de metadata (de beschrijvende data, zie par. A), de annotatiebestanden en de geluidsbestanden. Klik éénmaal op “Metadata” om deze te bekijken in het kader “Info/Content”. Zoals je ziet, wordt er heel wat beschreven. Interessant zijn bijvoorbeeld de gegevens over de sprekers; je vindt deze onder het kopje “Participant information”. Kijk hier even naar – hoewel er diverse codes worden gebruikt, zie je ook informatie die direct duidelijk is.
6. Je kunt in een apart venster de orthografische transcriptie van een sessie bekijken door op de naam van de sessie (b.v. *fn000248*) te dubbelklikken. Probeer dit uit. De gekleurde code bij elke annotatie-eenheid is de ID-code van de spreker.
7. Sluit het transcriptievenster weer. Klik éénmaal op de sleutel voor de naam van de sessie om deze weer te sluiten en doe hetzelfde voor de categorie “spontaneous conversation (face-to-face)”.[2]
8. Je gaat nu daadwerkelijk een teksttype selecteren. Klik éénmaal op de naam “spontaneous conversation (face-to-face)”, zodat hij geselecteerd is (de naam wordt dan getoond op een lichtblauwe achtergrond). Klik dan onder aan het kader op de knop [Add] om hem toe te voegen aan je selectielijst. Het woord “List” op de gelijknamige knop wordt dan rood. Dit geeft aan dat er subcorpora zijn geselecteerd. Als je klikt op de knop [List], zie je het resultaat in het venster “Selected nodes”. Sluit dit venster weer en klik op de knop [Save] om je selectie te bewaren.[3] Je kunt nu in het gedefinieerde deelcorpus gaan zoeken.

N.B. Houd altijd goed in de gaten of er bepaalde subcorpora zijn geselecteerd en welke dat zijn, zodat je altijd weet waarop de resultatenlijst van een zoekopdracht is gebaseerd. Je kunt geselecteerde subcorpora met behulp van de knop [Delete] weer (één voor één) deselecteren. Alleen als het woord “List” niet rood wordt weergegeven, wordt een zoekopdracht op het gehele corpus wordt uitgevoerd.

B.3. Zoeken naar woorden en woordcombinaties

Je kunt met behulp van het programma Corex in de orthografische annotaties zoeken naar:
- individuele woorden (zoals *huis*);
- combinaties van woorden (zoals *open huis* en *huis en haard*);
- woorden die een bepaalde letterreeks bevatten (b.v. *huis*, zoals in *huisvesting* en *verhuisd*);
- woorden die binnen een bepaalde afstand van elkaar voorkomen (zoals *naar* en *huis* in *ik ga naar huis* en *ik ga straks naar mijn nieuwe huis*).

Je zult nu zien hoe je m.b.v. Corex kunt zoeken naar een woord. Je doet dit aan de hand van de volgende vraagstelling:

Het Engelse bijvoeglijke naamwoord *cool* heeft enige tijd geleden zijn intrede gedaan in het Nederlands. Nu kunnen bijvoeglijke naamwoorden op drie manieren worden gebruikt:
- als bijvoeglijke bepaling, zoals in *een mooi huis* – dit gebruik wordt ook wel attributief genoemd;

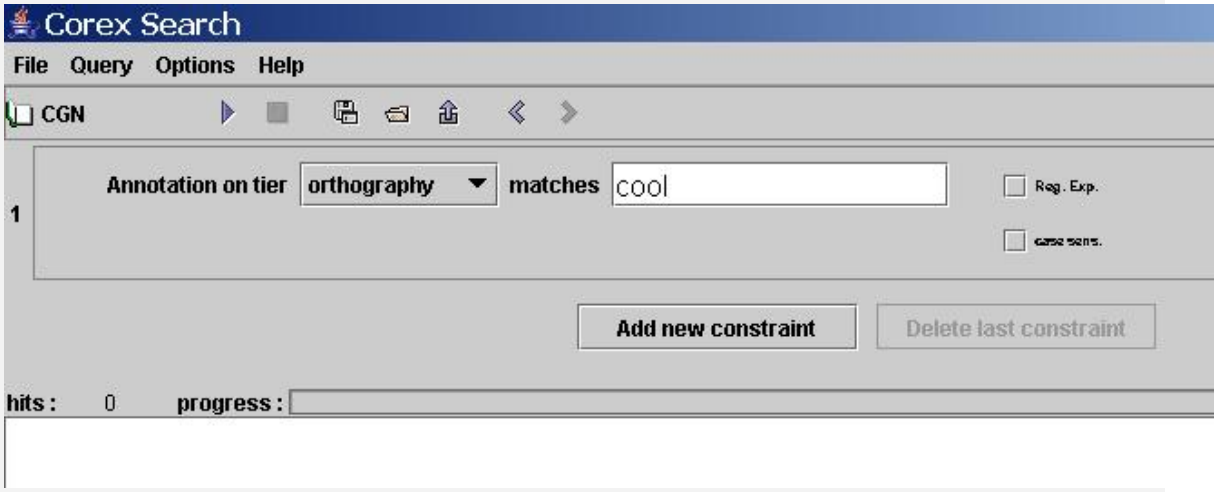
[2] Op dezelfde wijze kun je ook meerdere teksttypen aan je deelcorpus toevoegen: selecteren + knop [Add].
[3] De specificatie van je selectie wordt vooralsnog bewaard in de netwerkfolder met je persoonlijke *settings*.

- als naamwoordelijk deel van het naamwoordelijk gezegde, zoals in *het huis is mooi* – dit gebruik wordt ook wel predicatief genoemd;
- als bijwoordelijke bepaling, zoals in *Jan schrijft mooi* of *Je bent mooi bruin* – dit gebruik wordt ook wel adverbiaal genoemd.

Vragen: (i) Wordt *cool* in het Nederlands attributief, predicatief en adverbiaal gebruikt? (ii) Als dat het geval is, welk gebruik is dan het meest frequent?

Om deze vragen te beantwoorden ga je op zoek naar het woord *cool* in het hierboven gedefinieerde deelcorpus "spontaneous conversation (face-to-face)". Als je dit nog niet hebt gedefinieerd, doe dit dan nu alsnog! Alleen de subcorpora die aan de selectielijst zijn toegevoegd (en daarmee in de *basket* zijn opgenomen) worden doorzocht.

1. Ga indien nodig naar het hoofdvenster van Corex. Selecteer daar in het menu "Search" de optie "Content search". Er verschijnt dan een nieuw venster met de titel "Corex search", waarin je een zoekopdracht kunt formuleren. Maak dit zoekvenster schermvullend. Voor het zoeken naar een woord in de orthografische annotatielaag (de transcriptie) kun je alles laten staan zoals het staat, en in het invulveld achter "matches" het te zoeken woord invullen. Vul hier nu het woord *cool* in:



   N.B. Zorg dat het vakje voor "Reg. exp." <u>niet</u> is aangevinkt.
2. Klik nu op de knop in de CGN-werkbalk (bovenaan) of op Search onderaan het venster om het zoekproces te starten.
   N.B. Als je hierna wordt gevraagd of je met een index search wilt werken, antwoord dan "Nee".
   Een balkje achter progress: geeft aan welk gedeelte van het deelcorpus al is doorzocht; achter hits: wordt het aantal gevonden woorden getoond. Zodra een annotatie-eenheid wordt gevonden die het opgegeven woord bevat, wordt deze ook meteen getoond. Wacht even totdat er ongeveer 10 voorkomens zijn gevonden, en onderbreek dan het zoeken met behulp van de knop (in de CGN-werkbalk) of Stop (onderaan het zoekvenster).
3. Je kunt de resultaten van een zoekopdracht niet afdrukken. Je kunt ze wel kopiëren en vervolgens plakken in Word of een ander programma (Kladblok, Excel, Access, e.d.): selecteer het gedeelte dat je wilt kopiëren (*let op: Corex geeft niet met behulp van de bekende blauw-zwarte achtergrondkleur aan dat je iets hebt geselecteerd, maar dit gebeurt wel!*), klik op de toetscombinatie Ctrl+C om de selectie te kopiëren en plak hem vervolgens in het programma naar keuze (Bewerken > Plakken of Ctrl+V).

N.B. Zorg er altijd voor dat de annotatie-eenheden die je citeert door anderen terug te vinden zijn in het corpus. Neem daarvoor als bronvermelding tenminste het bestandsnummer en het nummer van de annotatie-eenheid over. Je vindt die op de regel boven de annotatie-eenheid, achter het hekje. Als dat relevant is, kan je ook nog een afkorting van het subcorpus geven waarin de annotatie-eenheid is gevonden. Een voorbeeld van een annotatie-eenheid met een correcte bronvermelding is:
{maar} toch die idealen moet je gewoon centraal stellen. (fn000067.15)

4. De uitvoer van een zoekopdracht bevat uitsluitend de annotatie-eenheden die de betreffende *hit* bevatten en de naam van het betreffende sessiebestand. Als je meer context wilt zien, of als je meer informatie over de betreffende sessie of spreker wilt zien, kun je de *Corex viewer* gebruiken. Klik daarvoor met de rechter muisknop op een *hit* en klik vervolgens op "Open transcription" (je kunt ook dubbelklikken op een *hit*). De Corex viewer markeert de annotatie-eenheid met een paarse kleur en toont daarvoor en daarna de betreffende context. Je kunt in dit venster scrollen om meer context te bekijken. De gekleurde codes onder elke eenheid zijn de sprekercodes (ID's). Sluit het Corpus view-venster weer.
5. Je kunt de inhoud van het Corpus view-venster niet op de gebruikelijke manier kopiëren naar een ander programma. Er is wel een trucje om dit te doen. *Tip*: verklein voordat je dit doet het Panel-venster en scroll de inhoud zodanig dat het venster uitsluitend de tekst bevat die je wilt opslaan. Print dan de inhoud van het venster als pdf-bestand, via de menu-optie
   File > Print view > klik op [OK] in het venster 'Pagina-instelling' > selecteer de 'printer' Adobe PDF in het venster 'Afdrukken' > [OK]
   Hierna kan je het betreffende bestand ergens opslaan (doe dit op een logische plaats en onder een logische naam!). Het wordt daarna automatisch geopend in een PDF-reader, waarna je de betreffende tekst kunt kopiëren en vervolgens kunt plakken in je onderzoeksverslag. N.B. Je moet ook nu nog wel even de juiste bronvermelding (zie hierboven) overnemen uit het venster met Search-resultaten.
6. Met de hierboven gebruikte zoekmethode vind je uitsluitend het gehele woord *cool*. Je mist dus eventuele vervoegingen (*coole*) en samenstellingen (*supercool*). Je kunt hier wat aan doen door een zgn. reguliere expressie te gebruiken. Je kunt dan speciale *wildcards* gebruiken voor een willekeurige letter, een willekeurige reeks letters, het begin of het einde van een woord, etc. Omdat bij het werken met reguliere expressies niet langer wordt gezocht naar woorden, maar naar tekenreeksen en omdat de tekenreeks *c-o-o-l* verder niet veel zal voorkomen, kun je het in dit geval eenvoudig aanpakken, door uitsluitend het vakje "Reg. Exp." aan te vinken. Probeer dit en wacht nu tot het gehele subcorpus is doorzocht. Bekijk ondertussen elke hit, tel het aantal keren dat *cool* attributief, predicatief en adverbiaal wordt gebruikt en noteer daarbij de gevallen waarbij *cool* een voorvoegsel heeft (hoe vaak komt dat voor?). Negeer bij het tellen uiteraard de niet-relevante hits (zoals *The New Cool Collective*). Wat is je eindconclusie?

Hieronder volgt een aantal voorbeelden van speciale symbolen die je in reguliere expressies kunt gebruiken.

Symbool	Betekenis	Voorbeeld	Uitleg
\b	woordgrens	**\b**huis	woorden die beginnen met de letters *huis*
.	willekeurig teken	**\b**hui**.\b**	woorden van 4 letters die beginnen met de letters *hui*, gevolgd door één willekeurig teken (zoals *huis*, *huid*, *huig*, etc.)
[]	één van de letters tussen de vierkante haken	**\b**hui**[fg]\b**	woorden van 4 letters die beginnen met de letters *hui*, gevolgd door een "f" of een "g".
?	nul of éénmaal het voorgaande teken	coole**?**	*cool* of *coole*
\	ontdoet het volgende teken van zijn speciale betekenis	niet\?	*niet*, gevolgd door een vraagteken (het vraagteken betekent nu dus niet "nul of éénmaal")

N.B.

i) Om te zoeken naar woordcombinaties, zoals *best wel*, moet je ervoor zorgen dat het vakje voor "Reg. Exp." niet is aangevinkt.

ii) Helaas zijn de *wildcards* van Corex weer anders dan die van een programma als WordSmith, let daar dus op als je ermee wilt gaan werken.

B.4. Sociale kenmerken van de sprekers

Bij een woord als *cool*, dat als anglicisme is geïntroduceerd in het Nederlands, is het interessant om te onderzoeken of het woord door alle sprekers wordt gebruikt, of uitsluitend door één of meer specifieke groepen. Om hierachter te komen moet je kijken naar de sociale kenmerken van de betreffende sprekers, zoals leeftijd, geslacht, land en/of regio van herkomst, opleiding, beroep, etc. Je kunt deze bekijken door een annotatie-eenheid te openen in de Corex viewer en dan de optie *Show metadata* uit het menu *Options* te selecteren. Je kunt dan de gewenste gegevens van het betreffende fragment bekijken in het venster van de *Metadata Description Tree* (klik op de sleutel voor de fragmentnaam, klik dan op "Metadata" en bekijk de gegevens in het venster "Info/content"). Omdat dit nogal omslachtig is en je bovendien zelf de juiste spreker en de gebruikte codes nog moet opzoeken, is een speciaal databasebestandje gemaakt waarin de gegevens van een betreffende spreker op een efficiënte manier kunnen worden opgezocht en bekeken.

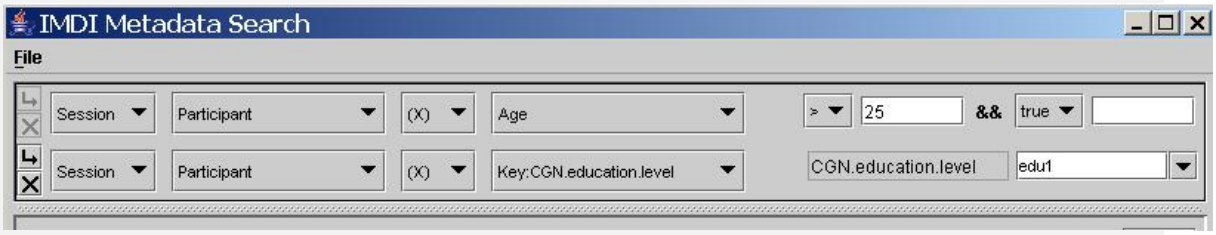
1. Open één van de annotatie-uitingen met *cool* in de Corex viewer.
2. Open de database met sprekergegevens via Windows:
   Start > Programma's > Opleidingen > Nederlands > CGN sprekergegevens
   Maak het venster even op maat voor het betreffende formulier (n door de randen te verslepen).
3. Vul links boven in het venster de ID-code van de spreker in van de annotatie-uiting in de Corex viewer met *cool*, druk op de Enter-toets en bekijk vervolgens zijn of haar gegevens. Doe dit voor nog enkele sprekers.
4. Sluit na afloop het databasevenster weer (als je het vaker wilt gebruiken, kun je het natuurlijk ook minimaliseren). Sluit tevens het venster "Corex Search".

B.5. Een deelcorpus samenstellen op basis van meerdere specifieke kenmerken

De methode die in paragraaf B.4 is beschreven, is handig om snel de gegevens van een aantal sprekers op te zoeken. Je kunt dan in het geval van *cool* een overzichtelijk tabelletje maken waarin je relevante sprekergegevens (zoals leeftijd, geslacht en opleidingsniveau) noteert. Als je echter van te voren al bepaalde intuïties hebt die je wilt toetsen, kun je dat ook op een andere manier doen, namelijk door een specifiek deelcorpus samen te stellen met behulp van een zgn. *metadata search*. Het grote verschil met het definiëren van een deelcorpus zoals beschreven in paragraaf B.2 is dat je bij een *metadata search* meerdere criteria kunt specificeren en dat een zoekopdracht ook echt uitsluitend zoekt in annotatie-eenheden die aan de opgegeven criteria voldoen.[4]

In de laatste oefening toets je de hypothese dat hoger opgeleide personen ouder dan 25 jaar het woord *cool* niet zullen gebruiken. Hiervoor stel je eerst een deelcorpus samen met de annotatie-eenheden van dergelijke personen.

1. Verwijder m.b.v. de knop [Clear] eventuele deelcorpora die eerder zijn geselecteerd.
2. Klik vanuit het hoofdscherm van Corex op *Search > Metadata Search*. Het venster "IMDI Metadata Search" verschijnt. Maak dit venster schermvullend. Met het keuzemenu "Session" hoef je niets te doen.
3. Selecteer in het tweede keuzemenu (waarin "Name" staat) de optie "Participant". Je kunt dan rechts daarvan een kenmerk van de spreker (waaraan wordt gerefereerd met een 'X') opgeven. Selecteer in het keuzemenu waarin "Type" staat de optie "Age". Selecteer in het keuzemenu dat vervolgens verschijnt het teken ">" (groter dan) en vul in het invulvak daarnaast het getal 25 in. Sluit af met de Enter-toets.
4. Geef aan dat je een tweede conditie wilt formuleren door een klik op de knop  . Selecteer ook hier in het tweede keuzemenu (waarin "Name" staat) de optie "Participant". Het gaat hier om dezelfde spreker, dus ook deze conditie geldt voor spreker 'X'. Selecteer in het keuzemenu waarin "Type" staat) de optie "Key.CGN.education.level" en selecteer ten slotte in het laatste keuzevak de code "edu1". Door in deze keuzelijst de cursor even op een code te laten rusten, wordt in een apart kadertje de betekenis getoond. Probeer dat; je ziet dan dat "edu1" staat voor "higher education". Sluit ook deze conditie af met een druk op de Enter-toets. De condities zouden er nu als volgt uit moeten zien:



5. Klik op de knop [Search] onderaan het venster om in het corpus alle relevante annotatie-eenheden op te laten zoeken. Dit kan even duren (met name de eerste keer dat je dit doet). Als in de balk achter progress: de boodschap "Finished in *x* ms" verschijnt, is het corpus doorzocht.

[4] Bij het definiëren van een subcorpus zoals in par. B.2 selecteer je *sessies* waarin bepaalde sprekers voorkomen (bijvoorbeeld mannen jonger dan 24 jaar). Een *content search* zoekt vervolgens echter in de complete sessies waarin die sprekers voorkomen – daarbij kan het gaan om conversaties met sprekers die *niet* aan het betreffende criterium voldoen, in wier uitingen dan dus toch ook wordt gezocht.

6. Nu kun je deze definitie opslaan. Type in het invoervak achter de knop [Save results in:] een duidelijke naam in, bijvoorbeeld “Hoger opgeleiden ouder dan 25 jaar” en klik vervolgens op de knop [Save results in:]. Sluit vervolgens het venster IMDI Metadata Search.
7. Dubbelklik in het hoofdvenster van Corex bij de *bookmarks* op “Search results”. Als het goed is, zie je hier de naam van het zojuist gedefinieerde deelcorpus. Verwijder (als je dat nog niet gedaan hebt) m.b.v. de knop [Clear] eventuele deelcorpora die eerder zijn geselecteerd[5] en voeg dan met de knoppen [Add] en [Save] je eigen deelcorpus aan de selectie toe (zie ook par. B.2). Kijk vervolgens met een *Content Search* hoe vaak het woord *cool* in dit deelcorpus voorkomt. Klopt de hypothese? Hoe oud zijn de sprekers die je toch nog gevonden hebt (ervan uitgaande dat de sessies rond 2000 zijn opgenomen)?

N.B.

i) De relatie tussen de gedefinieerde condities is altijd EN. Ze moeten dus allemaal gelden. (Een OF-relatie is dus niet mogelijk).

ii) Je kunt een conditie verwijderen met de knop  .

Hoewel je in absolute termen nog niet veel van de mogelijkheden van Corex hebt gezien, weet je nu voldoende om te gaan werken aan je dossieropdracht voor dit onderwerp.

Voor die opdracht kan het overigens handig zijn om de resultaten van een *content search* op te slaan, zodat je ze niet in één zitting hoeft te analyseren. In het venster “Corex Search” wordt via de menu-optie *Query > Save result* het materiaal opgeslagen in een extern bestand met het achtervoegsel .RES, waarvan je naam en locatie zelf kunt definiëren. Tijdens een latere zitting kun je het dan weer opvragen via *Search > Read result*. Je krijgt dan echter niet direct iets te zien – daarvoor moet je eerst de betreffende zoekopdracht herhalen. Dit gaat echter wel veel sneller dan wanneer je alle stappen weer opnieuw moet uitvoeren.

B.6. Aandachtspunten

1. Wees alert

Wees nooit te snel tevreden als een zoekopdracht een bepaald resultaat oplevert, maar denk altijd na over mogelijke tekortkomingen daarvan. De orthografische transcriptie van het *CGN* is bijvoorbeeld gemaakt door mensen, die fouten kunnen maken, soms inconsequent zijn in hun beslissingen, of andere beslissingen nemen dan hun mede-transcribeurs. Zo is het in voorbeeld van dit practicum heel goed mogelijk dat bepaalde vormen van *cool* zijn getranscribeerd als *koel*, omdat de uitspraak moeilijk te onderscheiden was of omdat de betekenis onduidelijk is (b.v. in *een koele buurman*). Een ander voorbeeld is dat getranscribeerde vraagzinnen in principe eindigen met een vraagteken, wat echter geen 100% garantie biedt dat alle zinnen die eindigen op een punt, geen vraagzinnen zijn. Ook hierbij heeft een transcribeur bij een twijfelgeval een beslissing moeten nemen.

[5] Voor deze opdracht hoeft het zoeken niet beperkt te worden tot het teksttype *conversation*. Als je dat wel zou willen doen, moet je één of meer condities in de metadata search toevoegen (Session > Content > Key:CGN.textClass.target).

2. Pas op met vergelijken en met generaliseren
Pas op dat je niet te snel algemene conclusies trekt op basis van je corpusresultaten. Als je bijvoorbeeld gevonden hebt dat 35 mannen en 43 vrouwen in het corpus het woord *cool* gebruiken, kun je dit alleen goed vergelijken als je weet hoeveel uitingen / woorden er in totaal door mannen en vrouwen zijn geuit in het gebruikte (deel)corpus. Je mag dus nooit absolute uitkomsten met elkaar vergelijken (behalve als je deelcorpora exact even groot zijn), maar je moet in principe met percentages werken. Daarnaast kunnen je bevindingen heel goed op toeval berusten. Je moet daarom met behulp van statistische procedures (zoals de relatief eenvoudige chi-kwadraattoets) aantonen dat dat niet het geval is, maar dat je bevindingen *significant* zijn. Daarnaast mag je op basis van je bevindingen in een corpus nooit zomaar constateren dat die dus gelden voor *het* Nederlands; ze gelden in principe alleen voor het corpus, behalve als het corpus daar representatief voor is.
N.B. Als je op deze manier met je corpusresultaten omgaat, houd je je bezig met kwantitatieve taalkunde. Hoewel dat op zich vaak een goede ondersteuning is voor inhoudelijk (kwalitatief) taalkundig onderzoek, kun je corpora op zich goed gebruiken voor onderzoek dat uitsluitend kwalitatief van aard is, mits je daarbij maar geen onverantwoorde uitspraken doet.

3. Afdrukken en kopiëren
Het is niet mogelijk om vanuit Corex resultaten op te slaan en/of af te drukken op een voor mensen prettig leesbare manier. Ook in het kopiëren van resultaten is niet op een handige manier voorzien. Dit kan echter wel, zie punt 3 en 5 van paragraaf B3.