Een Verkenning van COREX

Introductie van het Exploitatieprogramma bij het

Corpus Gesproken Nederlands

Auteur: Erik Weijers

1 maart 2004

Voorwoord

Deze handleiding vervult de rol van de inspirerende docent. Hopelijk wordt bij de lezer de belangstelling gewekt en de nieuwsgierigheid bevredigd. Als deze handleiding de docent is, dan is de Engelstalige handleiding, die ook op deze DVD staat, het leerboek. Het document dat u nu in handen heeft, geeft enkele voorbeelden van toepassingen van Corex en is niet allesomvattend. Zo komen syntax, prosodie en het lexicon niet aan de orde. Voor een gedetailleerd overzicht van alle onderdelen van Corex verwijzen we naar de handleiding *corexmanual*.

Inhoudsopgave

1. Algemeen

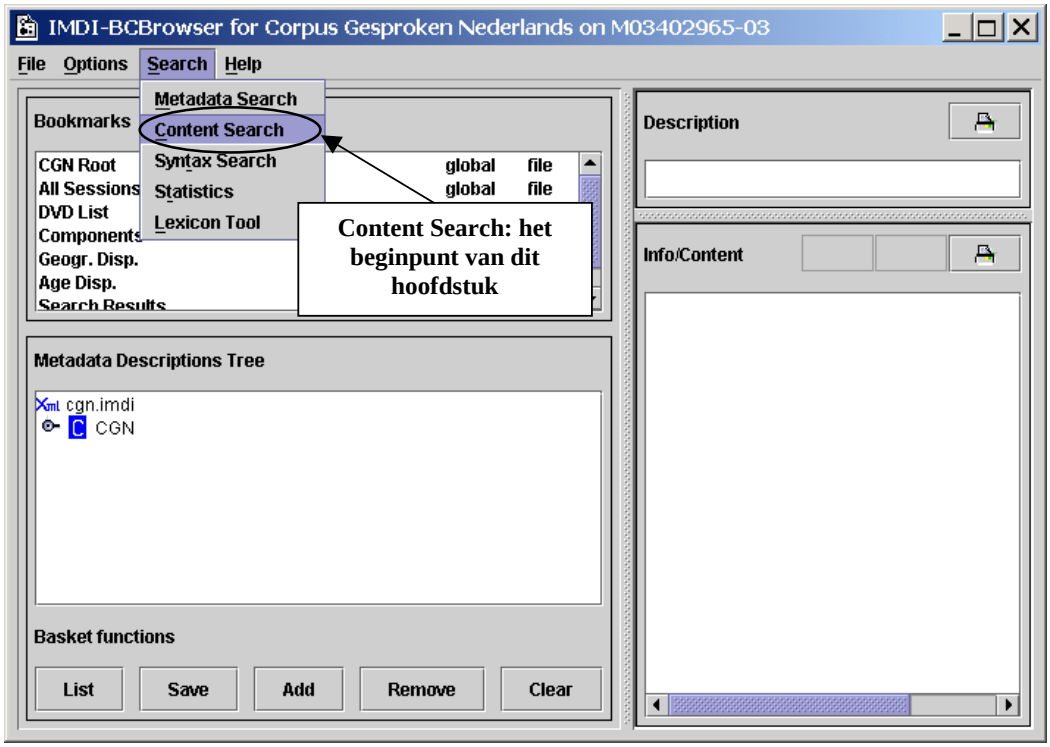
In dit eerste hoofdstuk leggen we aan de hand van een eenvoudig voorbeeld uit hoe u kunt werken met Corex. We demonstreren hoe er in het Corpus Gesproken Nederlands gezocht kan worden naar woorden. Er wordt gedemonstreerd hoe de gevonden geluidsfragmenten beluisterd kunnen worden en hoe het opslaan en weer openen van de zoekresultaten werkt.

De Installatie van COREX

Om Corex te installeren, moet u eerst de Corex-DVD in de DVD-drive doen en openen. U opent de DVD door in de Verkenner te dubbelklikken op de drive waar de DVD in zit. In de directory COREX/scripts staan de installatie-files en de readme-file. De installatie-file voor Windows-systemen is install.bat. De installatie-file voor Unix-systemen is install.sh.

Het hoofdvenster

U start Corex op door in de Corex-directory (de precieze naam van de directory is afhankelijk van de versie, bijvoorbeeld COREX6) de file *corex.bat* te dubbelklikken. Op Unix en MacOS-systemen heet de file *corex*. Voor overige informatie over de installatie verwijzen we naar de *readme*-file in bovengenoemde Corex-directory. Nadat Corex is opgestart, verschijnt het volgende venster in beeld. We besparen onszelf een opsomming van alle vensters en knoppen. In de loop van dit hoofdstuk zal de functie van de meeste onderdelen vanzelf duidelijk worden.

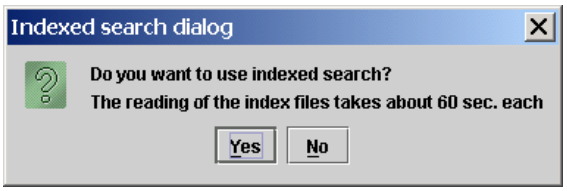


Zoeken naar woorden

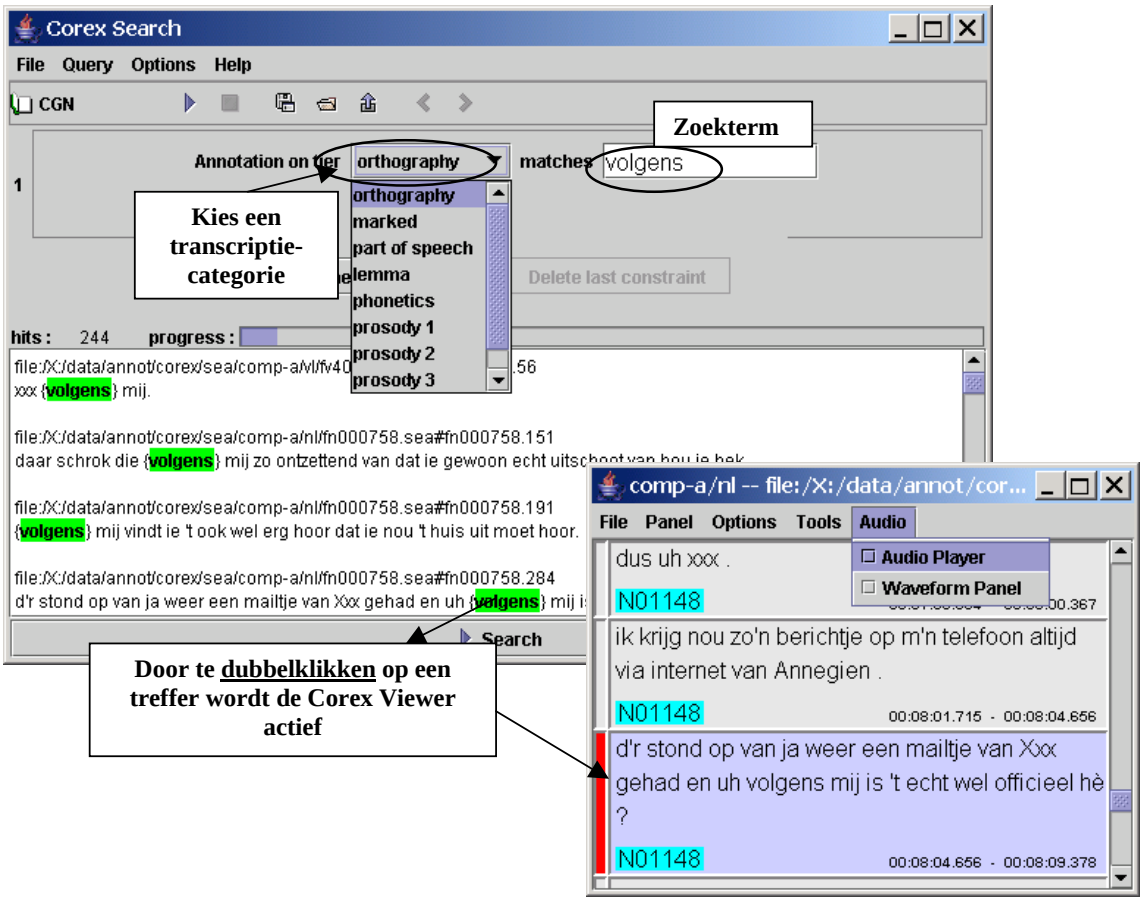
We kunnen het CGN beschouwen als een enorme verzameling geluidsfragmenten waarvoor meerdere transcripties en annotaties beschikbaar zijn. De geluidsfragmenten met de bijbehorende annotaties noemen we een sessie Voor alle sessies zijn de orthografische (spelling) transcripties beschikbaar. Van deze orthografische transcriptie is ook het bijbehorende lemma en de corresponderende *Part of Speech tag* vastgelegd. Deze drie typen annotaties vormen het hart van het geannoteerde CGN.

Een veel gebruikt beginpunt bij de zoektocht in het CGN is de orthografische annotatie. Die maakt het mogelijk om simpelweg een woord in te voeren en Corex te laten zoeken naar alle exemplaren in het corpus. Deze zoekoptie valt onder c*ontent search*. (content search is één manier van zoeken, metadata search een andere. De laatste zal worden behandeld in het volgende hoofdstuk.) Een content search wordt uitgevoerd door in het menu Search te kiezen voor Content Search.

Vervolgens vraagt Corex of u gebruik wilt maken van een *indexed search*. Door gebruik te maken van deze optie, heeft de computer veel minder tijd nodig om een zoekopdracht uit te voeren. De keerzijde is dat de computer veel werkgeheugen gebruikt.



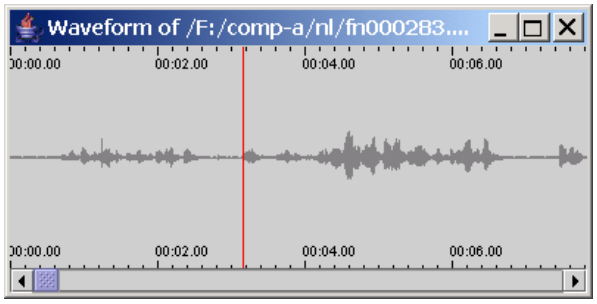
Nadat u een keuze gemaakt heeft over de indexed search, verschijnt het venster waarin de zoekopdracht kan worden ingevuld (volgende pagina). We zijn bijvoorbeeld geïnteresseerd in de manieren waarop het woord “volgens” wordt uitgesproken. We verwachten dat de uitspraak in veel gevallen zal neerkomen op iets als “vogges” (zie hoofdstuk 5 voor meer voorbeelden over fonetiek). We kiezen de optie orthography en voeren “volgens” in. We klikken eenmaal op Search (de lange balk onderaan). Corex begint nu met het zoeken naar alle voorkomens van dit woord. Dit zijn er vele honderden. De mogelijkheid bestaat om het zoeken te onderbreken, door op de knop Stop te klikken (de tekst van de knop Start verandert tijdens het zoekproces in Stop). De zoekresultaten verschijnen één voor één in beeld. Als u een zeldzamer woord zoekt dan “volgens”, kan het even duren voordat de eerste zoekresultaten verschijnen.



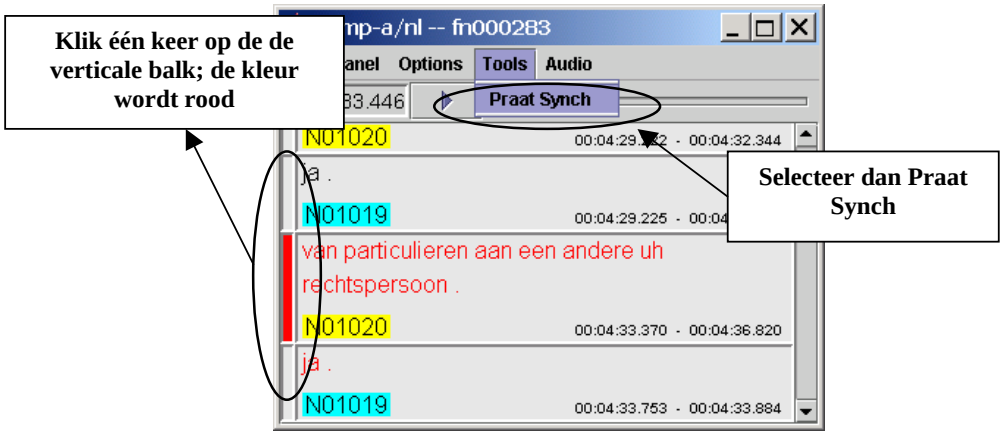
De beluistering van geluidsfragmenten

Door te dubbelklikken op een van de gevonden voorkomens wordt de Corex Viewer geactiveerd. De Corex Viewer markeert de zogenaamde annotatie-eenheid waarin het gezochte woord voorkomt met een paarse kleur. Door in het menu Audio te kiezen voor Audio Player (zie boven) kan het geluidsfragment beluisterd worden.

Door onder het menu Audio te kiezen voor Waveform Panel, kan het geluidssignaal gevisualiseerd worden. Tijdens het afspelen van het fragment loopt een rode lijn synchroon met het geluid door de Waveform panel.



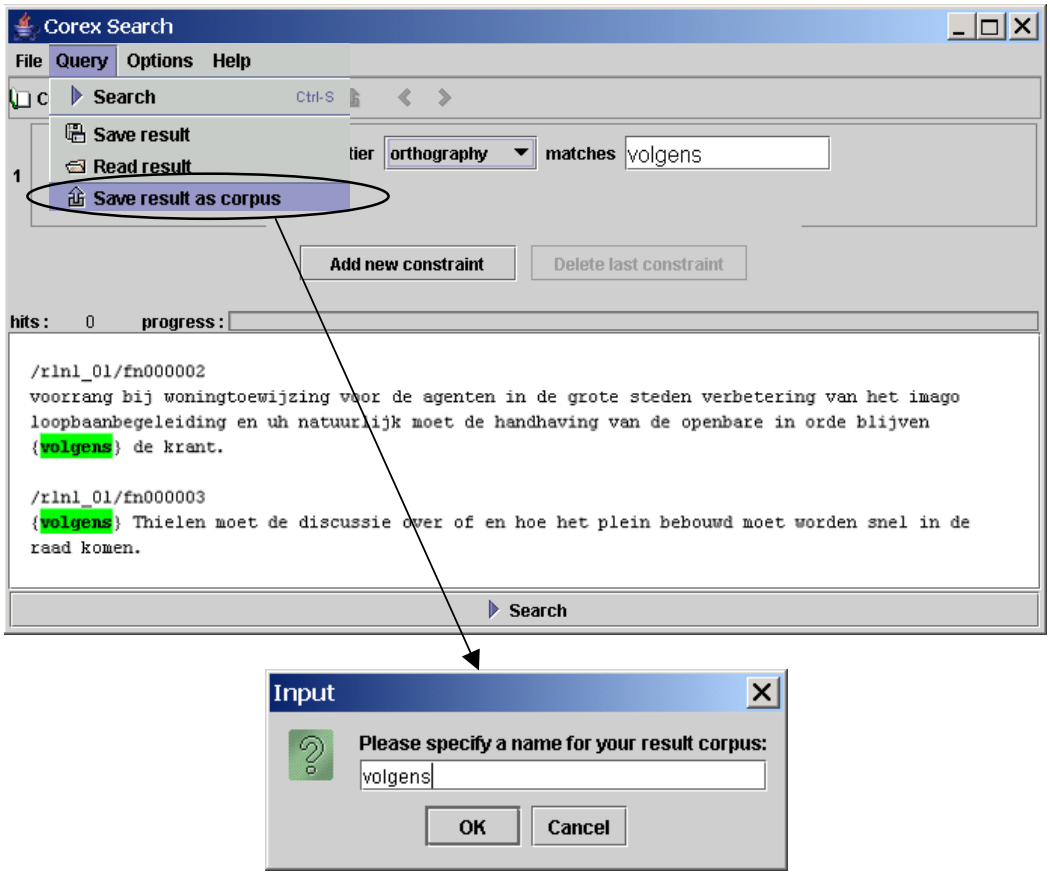
Een fragment openen met Praat



Door te klikken op Tools, Praat Synch kan het gekozen fragment geopend worden met het programma Praat (dit programma moet dan natuurlijk wel geïnstalleerd zijn op de computer). Let er op dat het fragment in kwestie geselecteerd is, dat wil zeggen dat de balk links van de tekst met rood is gemarkeerd. Is dit niet het geval, klik er dan een keer op.

Het opslaan van een zelfgemaakt corpus

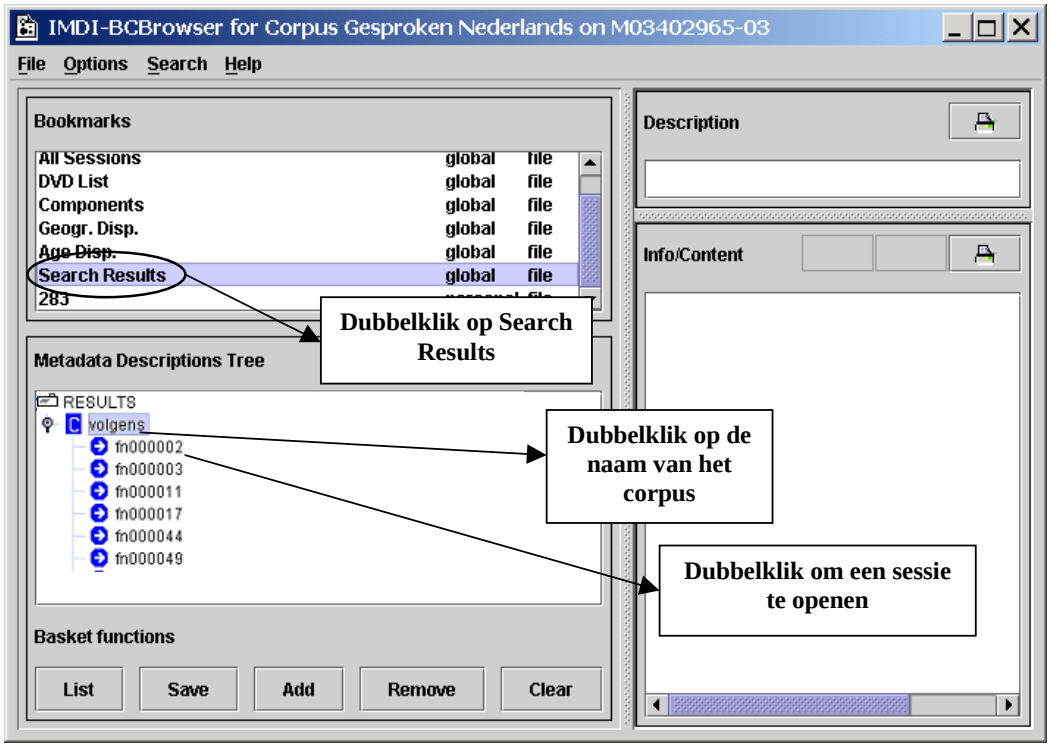
Als u de gevonden geluidsfragmenten wilt bewaren en daarnaast de bijbehorende metadata, waaronder informatie over de sprekers, dan is de optie Save result as corpus… aan te bevelen. De optie Save result is ook een mogelijkheid. Dit is een goede optie als u bij een andere gelegenheid nog meer content searches wil uitvoeren binnen de nu gevonden resultaten. Het voordeel is de hogere snelheid waarmee de zoekopdrachten zullen verlopen. Het nadeel is het verlies van de links naar de metadata, wat in ons voorbeeld niet wenselijk is. Daarom kiezen wij hier voor Save result as corpus.



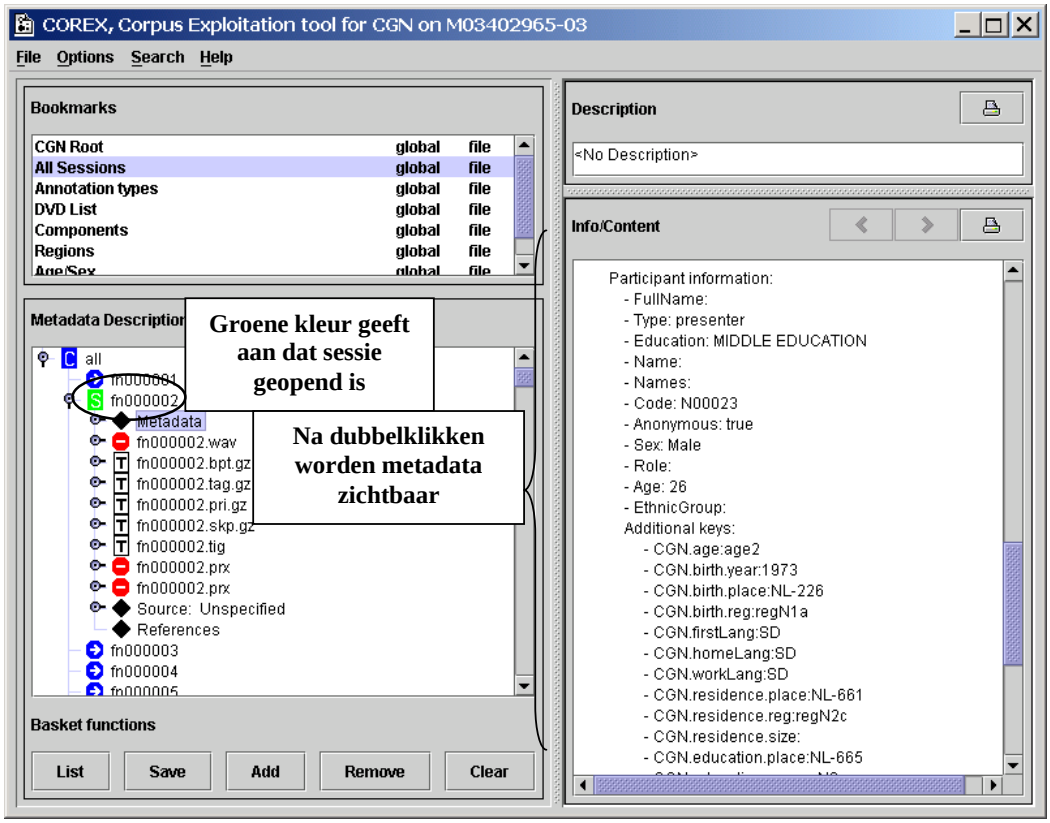
U wordt gevraagd om uw zelfgemaakte corpus een naam te geven en op OK te drukken. Gebruik geen vreemde tekens in uw filenaam; werk met cijfers en letters. U hoeft geen bestandsextensie of locatie op te geven; dit doet Corex zelf. In de volgende sectie demonstreren we hoe u uw zelfgemaakte corpus weer opent.

Het openen van een zelfgemaakt corpus

De fragmenten van het eerder opgeslagen corpus, genaamd “volgens”, zijn opnieuw te beluisteren. De geluidsfragmenten met hun annotatie blijven dan gekoppeld aan de zogenaamde metadata, waaronder informatie over de sprekers. Deze gekoppelde informatie noemen we een *sessie*. Het CGN bestaat uit duizenden sessies, geordend in subcorpora. Het is mogelijk om zelf een subcorpus toe te voegen aan deze lijst.



Onder het kopje Bookmarks, in het hoofdvenster van Corex, kunnen de opgeslagen resultaten weer geopend worden. Eerst dubbelklikt u op Search Results. Als u vervolgens dubbelklikt op de naam van het corpus, verschijnen de verschillende sessies in een rij in beeld. In dit voorbeeld zijn dat dus alle sessies waarin het woord “volgens” een of meerdere keren is uitgesproken.



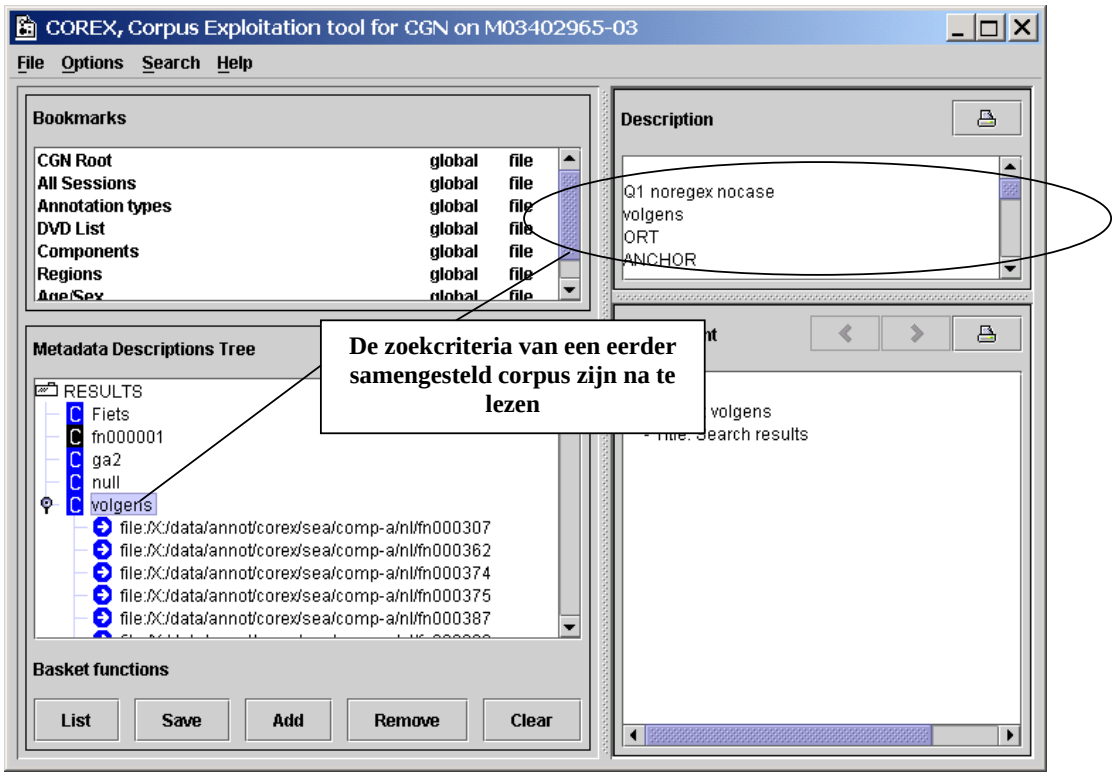
In bovenstaand voorbeeld is er eerst gedubbelklikt op het fragment fn000002, waarna de kleur is veranderd van blauw naar groen. Door vervolgens te dubbelklikken op (bijvoorbeeld) Metadata, worden gegevens over de opnamesessie, waaronder gegevens over de sprekers, zichtbaar in het rechter venster. In bovenstaand voorbeeld zien we onder meer dat de spreker die gecodeerd is als N00023, een man is van 26 jaar oud (voor een overzicht en verklaring van de metadata, zie de appendix). Door nogmaals te dubbelklikken op een geopende (groene) sessie, kan het fragment weer bekeken en beluisterd worden met de Corex Viewer.

Een zelf opgeslagen corpus kan weer verwijderd worden door er met de rechter muisknop op te drukken en te kiezen voor de optie Remove. Dit kan ook buiten Corex om gedaan worden. De locatie van de Search Results is in de Corex-directory (naam hangt af van de versie: bijvoorbeeld COREX6) in het volgende pad: \IMDI-Tools\SEARCH\RESULTS.

Het Terugzien van de Zoekopdracht

Stel u heeft een complexe zoekopdracht uitgevoerd en het resulterende subcorpus opgeslagen. Enige tijd later wilt u nagaan wat ook alweer de zoekcriteria waren die u had toegepast. Dit kan door het betreffende subcorpus te openen en in het venster Description te kijken. In het onderstaande voorbeeld is het subcorpus *volgens* geopend. In het Description-venster is na te lezen wat de opgelegde beperkingen aan

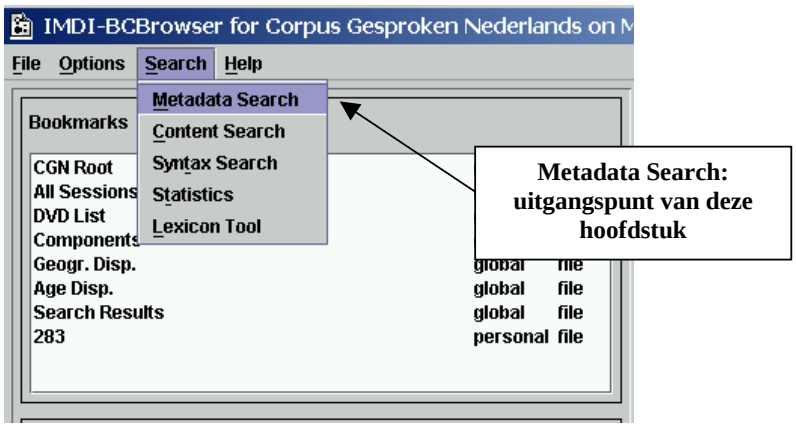
de orthografie zijn: “volgens”. “nogregex” en “nocase” staat ook nog vermeld, om aan te geven dat er geen gebruik is gemaakt van reguliere expressies of *case sensitive search*.



Tot zover dit hoofdstuk. In het volgende hoofdstuk komt aan de orde hoe zoekopdrachten verfijnd en gecombineerd kunnen worden, met gebruikmaking van onder andere metadata search.

2. Metadata Search

In dit hoofdstuk geven we een voorbeeld van een metadata search in Corex. Hierbij komen zowel sprekereigenschappen aan de orde, zoals leeftijd en geboorteregio, alsook sessie-eigenschappen, zoals het type spraak. We demonstreren hoe een zelf samengesteld subcorpus steeds verder op maat gemaakt kan worden door meerdere beperkingen op te leggen aan de sprekereigenschappen. Het zelfgemaakte subcorpus kan worden opgeslagen, waarna er nieuwe metadata- of content searches in uitgevoerd kunnen worden.

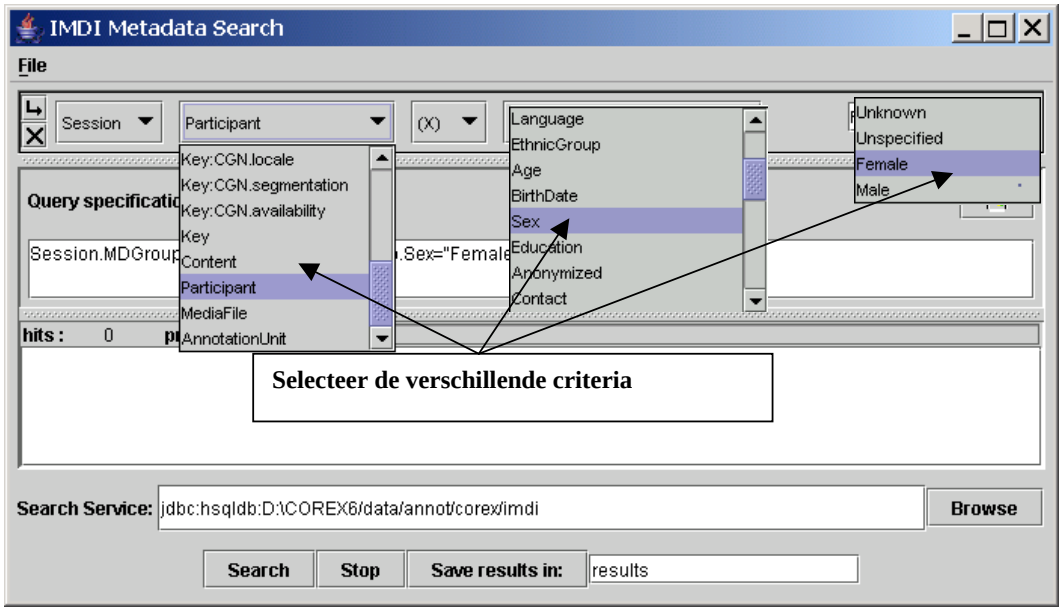


Allereerst bekijken we het openingsscherm. We klikken op Search → Metadata Search, waarna het volgende venster verschijnt (zie volgende pagina). We gaan nu in verschillende stappen de gewenste referentiegroep selecteren.

Wat wordt onze referentiegroep? In dit voorbeeld laten we ons inspireren door het socio-linguïstische verschijnsel van het *poldernederlands*. Dit is de naam die Dr. Jan Stroop geeft aan een vorm van het ABN "[…]die gesproken wordt door een succesvolle categorie Nederlanders: jonge vrouwen met een hogere opleiding. […] Het verschil tussen de *ei* van het ABN en deze *aai* [die van het poldernederlands] is dat bij de laatste de mond verder geopend wordt; het is een geval van 'verlaging' (van de onderkaak), in fonologische termen gesproken. Zo'n verlaging is meestal systematisch en beperkt zich niet tot één bepaalde klinker." (citaat uit *Stroop, Jan., 'Wordt het Poldernederlands model?', in* ***Noordzee, taal en letteren****. 1e jaargang , nummer 1-2, maart - april 1998, blz. 11-13.* Citaat aangehaald in website over poldernederlands: http://cf.hum.uva.nl/poldernederlands/).

Zoeken naar sprekereigenschappen

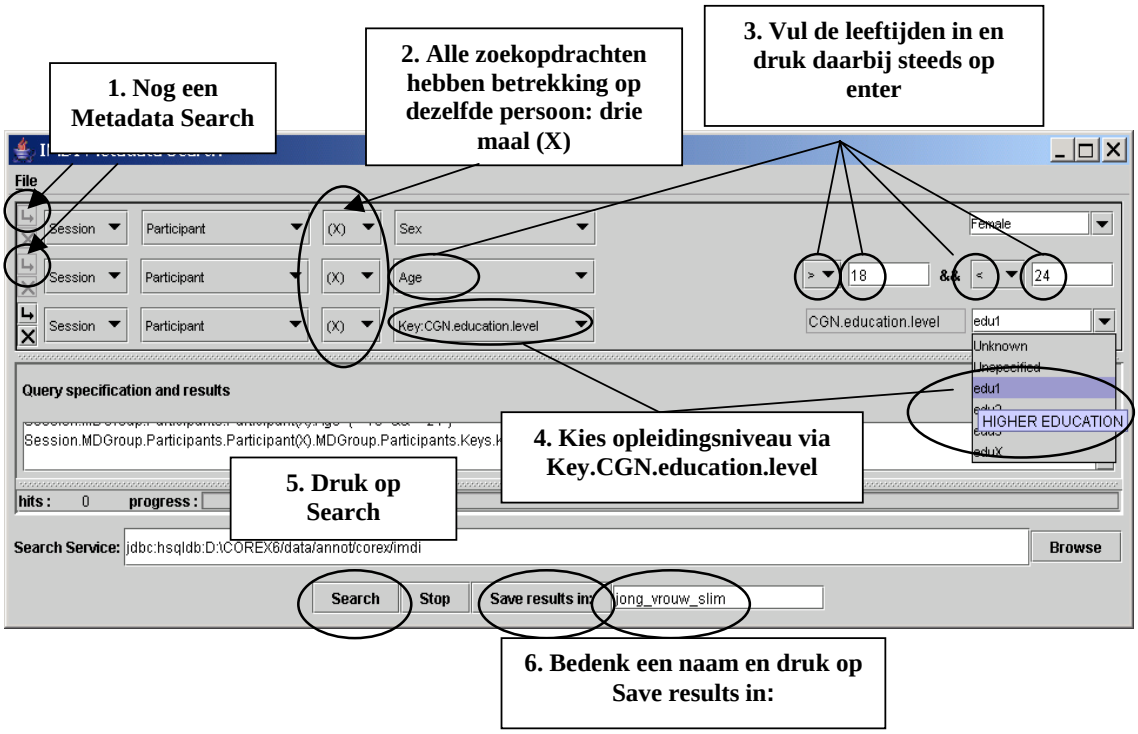
We willen met behulp van Corex uitzoeken of het poldernederlands gesproken wordt door vrouwen van 18-24 jaar met een hoog opleidingsniveau.



Bij elk pull down menu zijn er meerdere mogelijkheden. Wij kiezen voor Participant, Sex, Female. Daarmee hebben we een sexe-criterium gespecificeerd. Vervolgens willen we de geselecteerde groep vrouwelijke sprekers verder inperken naar leeftijd en opleidingsniveau.

Meerdere beperkingen opleggen

Druk op ↳ (zie onderstaand venster), het symbool dat uiterst linksboven in het venster staat. We kunnen nu, op dezelfde wijze als boven, een nieuwe beperking opleggen aan de referentiegroep. We selecteren Participant en Age. Om de cijfers 18 en 24 te kunnen invullen is het wellicht nodig dat u het venster vergroot, door met de linker muisknop te drukken op de rechter begrenzing en de rand naar rechts te slepen. We kiezen > 18 en < 24. *Let er op dat u op Enter drukt na het invullen van de cijfers.* De knop rechts van Participant staat standaard op (X). Dit laten we zo, ook voor de volgende zoekopdracht. Hiermee geven we aan dat alle zoekopdrachten betrekking hebben op dezelfde persoon. Zouden we geïnteresseerd zijn in dialogen, bijvoorbeeld interviews met een mannelijke en een vrouwelijke spreker, dan zouden we kiezen voor (X) female en (Y) male (of andersom).



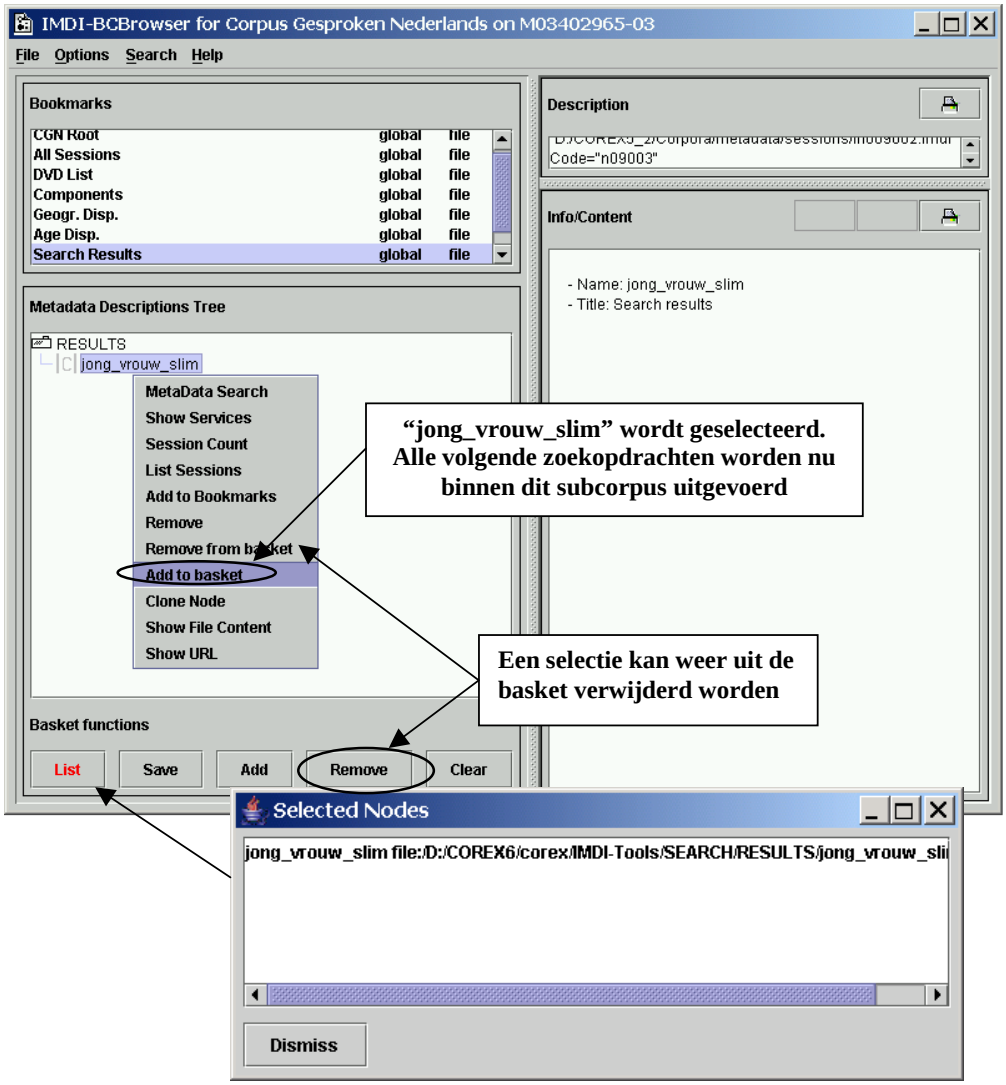
Tenslotte willen we het opleidingsniveau van de geselecteerde groep bepalen. We kiezen voor Participant, Key.CGN.education.level. In het pull down menu verschijnen een aantal gecodeerde opties, zoals edu1. Door met de muis even te blijven zweven boven deze code, wordt de betekenis zichtbaar in een blauw balkje (zie ook de appendix voor een overzicht). Edu1 is het gewenste opleidingsniveau, namelijk een hogere opleiding. Let op: als u veel zoekcriteria toepast, kan het zijn dat ze niet meer allemaal in het venster passen. Vergroot dan het venster door op de bekende knop rechtsboven te drukken: 

Als alle condities zijn geselecteerd, druk dan op Search. Als het zoekproces voltooid is, vullen we een naam in en drukken op Save Results in:. Het subcorpus wordt opgeslagen en is nu bereikbaar via Search Results, zichtbaar in het hoofdvenster van Corex. We hebben nu een eigen subcorpus samengesteld dat alle functionaliteit heeft van het oorspronkelijke CGN. Dat wil zeggen: de fragmenten kunnen beluisterd worden, de Metadata kunnen bekeken worden en er kunnen opnieuw een metadata search en een content search op uitgevoerd worden. Dit laatste is wat we nu gaan doen.

Zoeken binnen het zelfgemaakte subcorpus

We hebben nu een eigen corpus samengesteld van hoog opgeleide vrouwen van tussen de 18 en 24 jaar. Dit corpus is opgeslagen onder de naam “jong_vrouw_slim” (de liggende streepjes zijn niet verplicht; ze zijn er alleen voor de leesbaarheid). Binnen dit zelfgemaakte corpus kunnen weer nieuwe zoekopdrachten worden

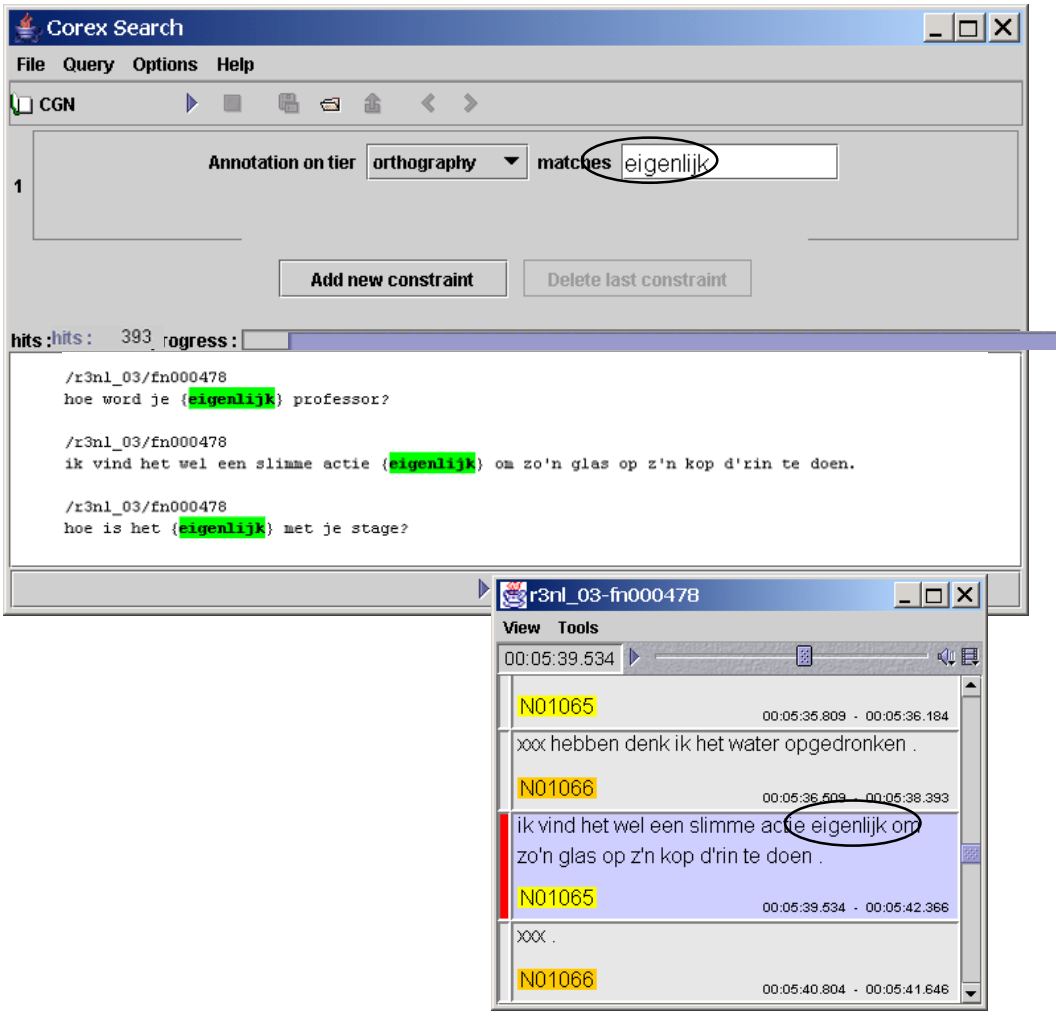
uitgevoerd, zowel content searches als metadata searches. Wij zijn geïnteresseerd in de vraag of onze vrouwen de *ei* in het woord *eigenlijk* uitspreken als een *aai*. Het selecteren van alle voorkomens van (bijvoorbeeld) *eigenlijk* kan met content search, zoals we ook zagen in het vorige hoofdstuk.



Het is van belang voor onze zoekopdracht dat de content search alleen binnen het subcorpus jong_vrouw_slim gaat plaatsvinden en niet in het hele CGN . Dit kan met de optie Add to basket. Eerst dubbelklikken we op Search Results in het hoofdvenster (hier worden zelfgemaakte subcorpora opgeslagen). Dan klikken we met de linker muisknop op ons subcorpus jong_vrouw_slim. De kleur verandert in een soort paars. Na een druk op de rechter muisknop verschijnt het pull down menu, waar de optie Add to Basket te vinden is. Als we hierop klikken wordt het geselecteerde

corpus aangegeven met een [C] . Het woord List wordt rood van kleur. Door op List te klikken, kan bekeken worden welke subcorpora er geselecteerd zijn. Er kunnen namelijk meerdere subcorpora worden toegevoegd aan de selectie. Het verwijderen van een subcorpus uit de selectie gebeurt door te kiezen voor de optie Remove from basket, die te vinden is in hetzelfde pull down menu als Add to Basket.

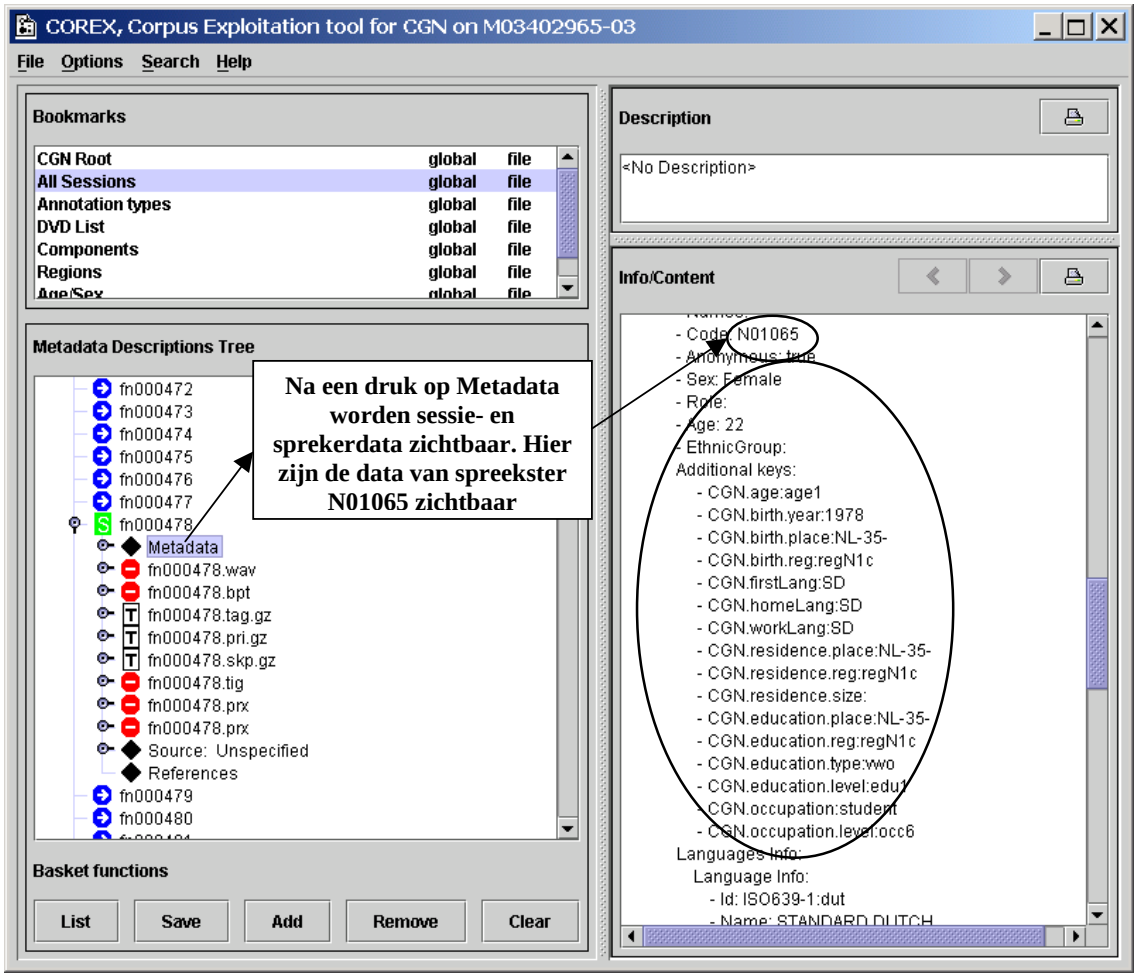
Nu het gewenste subcorpus is geselecteerd, kan een content search worden uitgevoerd. Dit doen we door de gelijknamige knop in het hoofdvenster een keer aan te klikken. De procedure die dan volgt is identiek aan het zoeken met content search in het hele CGN, zoals die in Hoofdstuk 1 beschreven is. Hieronder is het scherm weergegeven waarin de zoekresultaten staan: de voorkomens van het woord *eigenlijk*, uitgesproken door hoogopgeleide vrouwen van tussen de 18 en 24 jaar.



Na beluistering van enkele fragmenten hadden we beet. We vonden een spreekster die de *ei* als een *aai* uitspreekt. (we zochten handmatig, maar we hadden ook kunnen zoeken via de *tier* phonology binnen content search. Een deel van het corpus is immers handmatig fonetisch geannoteerd, zie Hoofdstuk 5). Wat zijn de eigenschappen van deze spreekster? We vervolgen de weg in de metadata.

De weg vinden in de Metadata

Een goede manier om kennis te maken met de metadata is door een fragment te openen en te zien hoe de metadata geordend zijn. In onderstaand voorbeeld is fragment fn000478 (waarin bovenstaande spreekster aan het woord is) geopend. Het fragment kan gevonden worden door in de Metadata Descriptions Tree te dubbelklikken op All Sessions. Dit is de numerieke ordening van alle sessies van het CGN. Overigens kan het fragment ook gevonden worden door met metadata search op Session/Name te zoeken. De naam die ingevuld wordt is dan de naam van het fragment, in dit geval fn000478.



We dubbelklikken op het blauwe pijltje behorend bij fragment fn000478. Dan drukken we eenmaal op Metadata. Een verklaring van alle soorten metadata en de manier waarop ze gecodeerd zijn, is te vinden in de appendix van deze handleiding.

In bovenstaand voorbeeld, in het venster Info/Content, is te zien dat niet alle Metadata gespecificeerd zijn. Dit verschilt per fragment en per spreker. Het

geboortejaar van de spreekster uit het voorbeeld is wel gespecificeerd, maar haar role niet.

Verder zien we dat er meerdere typen files bekeken kunnen worden, die alle betrekking hebben op dit fragment. Sommige zijn echter niet beschikbaar. De files die niet beschikbaar zijn, zijn weergegeven met een rood verbodsbord: ⛔. Beschikbare files zijn weergegeven met een 🅃. Een .wav-file die beschikbaar is, wordt aangeduid met een teken dat op een M lijkt. Voor het gekozen fragment zijn de .tag.gz-file, de .pri-file en de .skp-file beschikbaar. De .pri-file is de primaire orthografische transcriptie, de .tag.gz-file is de file waar *part-of-speech* (POS) informatie in staat, alsook lemma-informatie. De .skp-file bevat de tijdcodes van het begin en het einde van elk woord uit de audio file. Het is nuttig om deze en andere annotatiefiles eens te openen en te kijken hoe de transcripties er uit zien.

De files die voor dit fragment niet beschikbaar zijn, zijn de .wav-file, oftewel het geluidsfragment, en de .bpt-file. Dit laatste betekent dat er voor deze file geen brede fonetische transcriptie beschikbaar is. Overigens zijn de .wav files voor alle sessies beschikbaar (alleen niet ten tijde van het schrijven van deze handleiding).

Zoeken met de CGN-keys

De metadata-categorieën die speciaal voor Corex zijn gemaakt noemen we de *CGN-Keys*. Al deze datacategorieën hebben namen die beginnen met *CGN.*, bijvoorbeeld CGN.age. De keys zijn terug te vinden achter Session, Session/Content en achter Session/Participant. Achter Session en Session/Content zitten de keys met gegevens over de *opnamesessie,* zoals type spraak, opnameduur en de wijze van oplijning van het signaal. Achter Session/Participant zitten de keys over *sprekereigenschappen* die speciaal voor het CGN gecodeerd zijn, zoals geboorteregio en opleidingsniveau..

Meerdere wegen naar de data

Het is belangrijk om te weten dat bepaalde data op meerdere manieren gevonden kunnen worden en, anderzijds, dat bepaalde wegen niet naar data leiden. Dit komt doordat Corex conform is met de IMDI metadataset, de standaard voor verschillende linguïstische *resources*. Bepaalde datavelden van IMDI zijn niet van toepassing op het CGN en zullen dus geen zoekresultaten opleveren (bijvoorbeeld: zoeken via Location/county/American Samoa levert niets op). Anderzijds zijn data die wel relevant zijn voor het CGN soms via meerdere zoekpaden te vinden. Bij de constructie van de metadata search-architectuur van Corex is namelijk zoveel mogelijk CGN-specifieke informatie (de CGN keys) in de IMDI-datavelden gestopt. Bijvoorbeeld: leeftijdinformatie over de sprekers is bereikbaar via Session/Participant/Age (Een IMDI-veld) en ook via Session/Participant /Key.CGN.age. Overigens is er een klein verschil in functionaliteit tussen deze twee manieren van zoeken naar leeftijd. Met Participant/Age kunnen de leeftijdsgrenzen exact worden gekozen, terwijl de key.CGN.age werkt met vooraf bepaalde leeftijdsintervallen (zie appendix).

De CGN Participant Keys betreffende locatie: *place* en *region*

De meeste participant keys spreken voor zich (zo niet, zie de appendix voor een verklaring). De belangrijke CGN keys die nadere uitleg behoeven zijn *place* en *region*. Het CGN heeft twee eigen codeerwijzes voor het specificeren van deze variabelen. Wat betreft de regio's: Nederland en Vlaanderen zijn elk in vier hoofdregio's verdeeld, namenlijk een kernregio, een overgangsregio en twee perifere regio's. Voor Nederland is voor elk van de hoofdregio's een verder onderscheid gemaakt in subregio's. Daarnaast zijn er codes die gebruikt worden om aan te duiden dat de regio onbekend is. De locatie kan nog fijnmaziger gespecificeerd worden met de CGN key *place*. De eerste drie cijfers van de postcode moeten worden ingevuld, voorafgegaan door de landcode en een streepje, bijvoorbeeld NL-664. Het laatste cijfer van de postcode kan ook worden weggelaten en in plaats daarvan vervangen worden door een punt. Het postcodegebied wordt daarmee vergroot tot alle codes die beginnen met (in bovenstaand voorbeeld) 66. De punt in de zoekopdracht fungeert dan als variabele. *Region* en *Place* worden gebruikt voor de geboorteplaats van de spreker, de woonplaats van de spreker en de plaats waar de opleiding genoten is. Al deze keys vallen onder Participant/Key. In het onderstaande venster is een voorbeeld getoond van een metadata search die toegespitst is op een participant key, namelijk CGN.birth.reg.

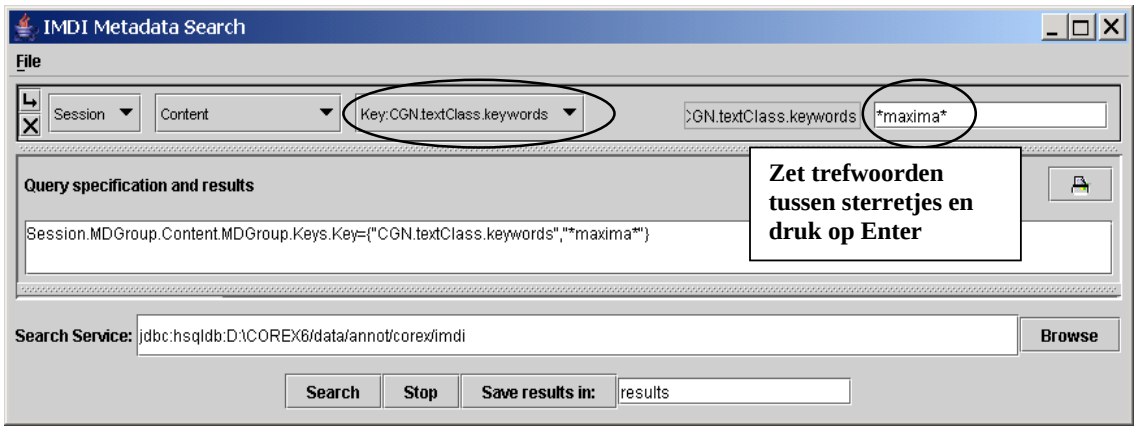
De CGN Session keys

Een andere belangrijke route binnen metadata search zijn de keys die betrekking hebben op eigenschappen van de opnamesessie. Een voorbeeld hiervan is CGN.wordCount, te weten het aantal woorden in een sessie. In de appendix staan alle keys opgesomd en verklaard.

De CGN Content Keys

Tenslotte zijn er de Content Keys van CGN (zoeken via Session/Content), die betrekking hebben op het type spraak en de inhoud van het gesprek. De key die specificeert wat voor een soort interactie het betreft, is CGN.textclass.target. Deze is uitgesplitst in vier categorieën. De eerste daarvan, text type, heeft betrekking op het type spraak. Gaat het om een telefoongesprek of om een *face-to-face* conversatie, of om een ander type spraak. Ook de andere drie categorieën van CGN.textclass.target verschaffen informatie over de context van de spraak (zie appendix).

De content key CGN.textclass.keywords biedt de gelegenheid tot het zoeken op inhoudelijke trefwoorden van de tekst. Bijvoorbeeld: zoeken naar fragmenten die als thema "Maxima" hebben. Elke sessie heeft een of meerdere inhoudelijke trefwoorden meegekregen. Dit betekent dus dat met het invullen van een keyword niet alle orthografische transcripties worden doorzocht, alleen de lijst van trefwoorden.



3. Constraints in Content Search

We gaan dieper in op het zoeken met meerdere *constraints* binnen content search. De POS-tags komen aan bod als voorbeeld.

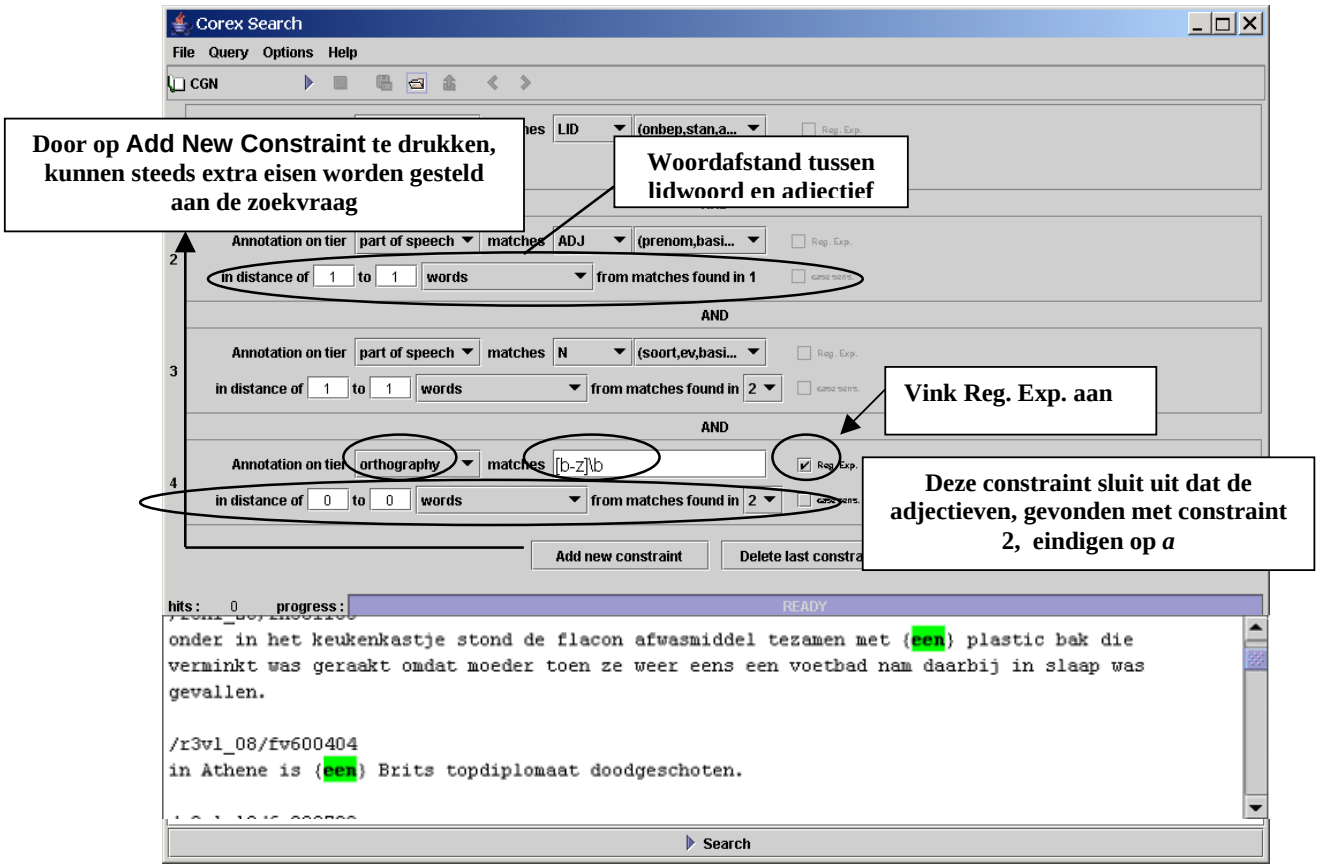
Het Koppelen van Constraints in Content Search

Het volgende voorbeeld illustreert hoe er gewerkt kan worden met Part of Speech tags (POS tags). Ook maakt het duidelijk hoe meerdere constraints samenwerken om een zoekopdracht samen te stellen.

De volgende zoekopdracht is geïnspireerd door een probleemstelling die mij is aangereikt door dr. Frank van Eynde. "In nominale groepen met een zijdig substantief (*de*-woord) als hoofd krijgen de prenominale bepalingen een buigings-e, zoals in *elke blauwe knikker* en *iedere man*. Het gebruik van onverbogen vormen in zulke NPs is echter niet uitgesloten. Zo kunnen, als het substantief een persoonsnaam is, de prenominale bepalingen – onder bepaalde omstandigheden – onverbogen blijven, zoals in *een groot staatsman* en *menig Belgisch zakenman*. […] Een interessante vraag is welke substantieven dit patroon toelaten. […]"

Om dit te onderzoeken gaan we een zoekopdracht samenstellen die drie POS-tags combineert. Het eerste lid van de prenominale bepaling, het lidwoord, moet onbepaald zijn (*een*). Om precies te zijn zoeken we naar lidwoorden met de tag LID (onbep,stan,agr). Herinneren we ons dat een verklaring van deze en andere afkortingen zichtbaar wordt door met de muis boven de afkorting in kwestie te blijven zweven. We voegen een constraint toe door op Add new constraint te drukken. We willen namelijk de onverbogen prenominale adjectieven zoals *groot* selecteren die volgen op het al eerder geselecteerde onbepaald lidwoord. We selecteren ADJ (prenom,basis,zonder). De relatie tussen deze twee constraints moet nu nog worden gespecificeerd. In dit geval moet het adjectief direct volgen op het lidwoord. We kiezen dus voor in distance of 1 to 1 words from annotation found in 1. Dan volgt het enkelvoudige zijdige substantief N (soort,ev,basis,zijd,stan), dat direct moet volgen op het adjectief. Vandaar: "… 1 tot 1 words from annotation found in 2". Let op dat u nu "annotation found in 2"selecteert in plaats van "…1" (of natuurlijk "2 to 2 words from annotation found in 1").

De zoekresultaten die met de bovenstaande drie constraints worden behaald zijn enerzijds niet volledig (hier komen we op terug) en anderzijds zitten er valse treffers tussen. Althans, valse treffers voor onze doeleinden. Er bestaan namelijk prenominale adjectieven die geen verbogen vorm hebben, zoals *extra*. Deze adjectieven zullen ook tussen de zoekresultaten zitten. Wie geen zin heeft om deze treffers allemaal achteraf handmatig uit het resultaatcorpus te verwijderen, kan dit vaak ook bereiken door één of meerdere extra constraints toe te voegen (zie hieronder).



Ter illustratie van de mogelijkheid om valse treffers bij voorbaat uit te sluiten, wordt er nog een orthografische constraint opgelegd. We vullen de reguliere expressie *[b-z]\b* in (denk eraan om Reg. Exp. aan te vinken) Dit betekent dat het adjectief op alle letters mag eindigen behalve een *a* (zie hoofdstuk 4 voor een uitgebreid overzicht van het gebruik van reguliere expressies). Hiermee wordt ook *extra* uitgesloten. (deze aanpak is natuurlijk wel gevaarlijk. Weten we wel zeker dat er geen adjectieven bestaan die eindigen op *a* en die toch een buigings-e kunnen krijgen? Het gaat hier echter alleen om het demonstreren van de mogelijkheden van Corex, niet om de verantwoording van een werkwijze) Hoe leggen we vast dat deze constraint betrekking heeft op het adjectief? Heel eenvoudig: door te specificeren "in distance of 0 to 0 words from annotation found in 2". Een afstand van 0 tot 0 wil zeggen: een afstand van precies nul, ofwel, betrekking hebbend op hetzelfde woord.

Zoals gezegd is deze zoekopdracht niet volledig voor onze doeleinden. Zo zou het eerste lid van de NP ook een vragende of onbepaalde determiner (bijvoorbeeld *welk*, of *ieder*) mogen zijn. Helaas is het niet mogelijk om deze constraint met een OF-operator toe te voegen. Het is dus niet mogelijk om te zeggen: het eerste woord dat ik zoek is OF een onbepaald lidwoord OF een vragende determiner, enz. Dit betekent dat we alleen een volledig overzicht van alle gewenste combinaties kunnen bereiken door steeds een nieuwe zoekopdracht te maken en het resulterende subcorpus op te slaan. Later kunnen de verschillende subcorpora dan eventueel worden samengevoegd met behulp van een tekstverwerkingsprogramma.

Daarnaast moet er nog steeds veel handmatig worden verwijderd. Zie bijvoorbeeld de treffer “een plastic bak”. Ook dit adjectief kan geen verbogen vorm hebben, zodat het ontbreken van een buigings-e hier niets zegt over het erop volgende substantief.

Het selecteren van de juiste POS-tag

In bovenstaand voorbeeld wordt het lidwoord met de tag (onbep,stan,agr) gekozen als eerste constraint. Het is ook mogelijk om te abstraheren over de drie varianten van onbepaalde lidwoorden die het CGN onderscheidt. Dat wil zeggen, het is mogelijk om in één zoekopdracht alle onbepaalde lidwoorden te selecteren. Dit kan door met de rechter muisknop te klikken op het vak waarin een reeds geselecteerde subcategorie lidwoorden, bijvoorbeeld (onbep,stan,agr) te lezen is.

Na de klik wordt het vak wit en kan er met de linkermuisknop een positie worden gekozen. Als we gaan staan bij agr en vervolgens klikken op de Delete-knop op het toetsenbord, wordt agr vervangen door een sterretje: *. Als we boven stan gaan staan en nogmaals op Delete drukken, verandert ook stan in een sterretje. Nu worden alle onbepaalde lidwoorden gezocht, ongeacht of ze de tag stan, gen of dial hebben.

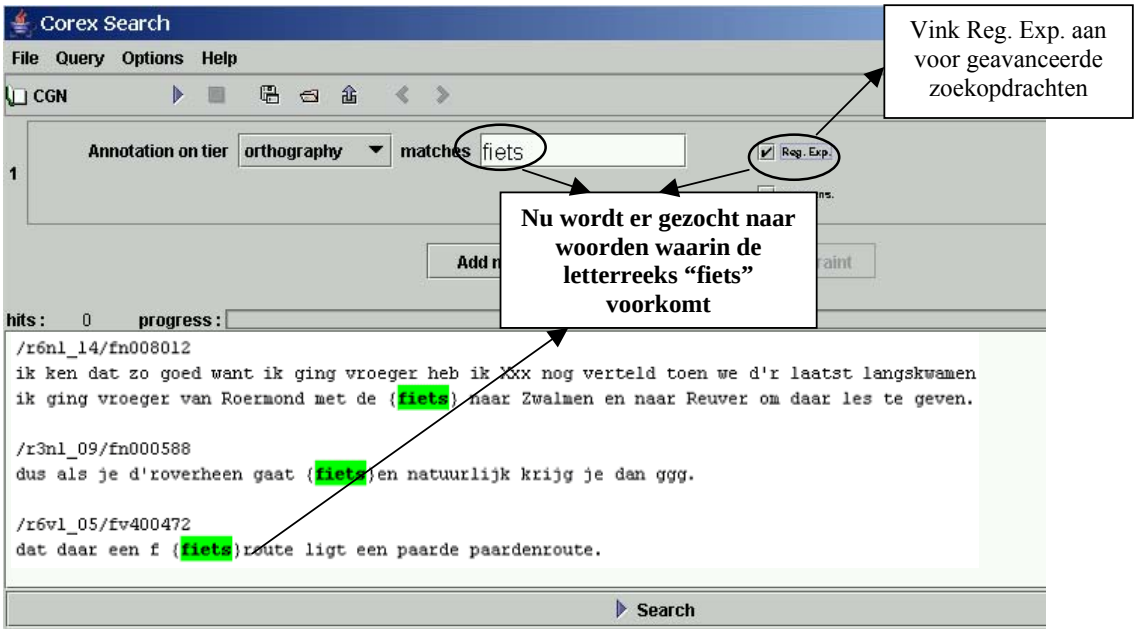
Nota bene: het is niet mogelijk om bijvoorbeeld de volgende keuze te maken: LID (onbep, *, agr). Alleen de laatste codering kan steeds verwijderd worden.

4. Zoeken met Reguliere Expressies

Zoeken met reguliere expressies geeft meer mogelijkheden. Zo kunnen we zoeken naar woorden die eindigen of beginnen met een bepaalde letter. De volgende zoekopdrachten worden, net als de bovenstaande voorbeelden, uitgevoerd binnen Content Search. Nu vinken we het vakje naast Reg. Exp. (Regular Expression) aan. Ook op andere plaatsen binnen Corex kan er gebruikt gemaakt worden van reguliere expressies, bijvoorbeeld bij het lexicon.

LETTERREEKSEN BINNEN WOORDEN

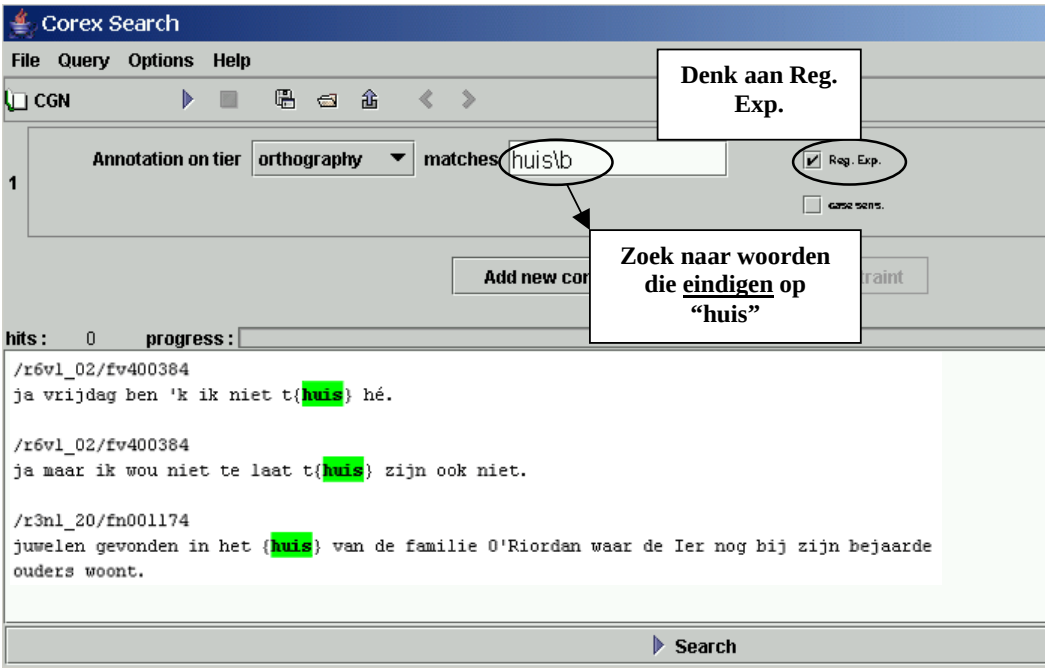
Als we Reg. Exp. aanvinken en vervolgens *fiets* invoeren dan krijgen we alle voorkomens van de letterreeks *fiets* te zien, ook als deze letterreeks deel uitmaakt van een woord, zoals in het geval van "fietsen" en "fietsroute".



WOORDEN DIE EINDIGEN OF BEGINNEN MET …

De tekencombinatie *\b* staat voor een woordbegrenzing. Door *huis\b* in te voeren, zoekt Corex naar de lettercombinatie *huis* aan het einde van een woord. Dit levert resultaten op als "thuis" (en ook "huis" zelf), en laat woorden als "huiselijk" buiten beschouwing. *\b* kan ook aan het begin van de zoekterm worden geplaatst. *\bhui* zoekt naar alle woorden die beginnen met "hui". Let op: *\b* en *hui* schrijven we aan elkaar. Zouden we een spatie invullen, dan zou Corex ook gaan zoeken naar een spatie tussen de woordbegrenzing en de *h*. (in dit geval is dat overigens niet erg, want we zoeken naar woordgrenzen, die meestal samenvallen met een spatie).

Willen we in deze zoekopdracht het woord "huis" uitsluiten, dan geven we de zoekopdracht *\whuis\b*. Backslash *w* staat voor elk karakter dat deel kan uitmaken van een woord. Hiertoe behoren alle hoofd- en kleine letters en alle cijfers. Spaties vallen hier niet onder en dus valt " huis" af en is "thuis" geldig.



EEN VAN DE VOLGENDE KARAKTERS

Stel, u zoekt naar woorden die de lettercombinaties "bde" of "gde" of "fde" bevatten. Hier gebruiken we haken: *[bgf]de* is de juiste zoekopdracht. Hiermee geven we aan dat de eerste letter een b of een g of een f, maar geen andere letter mag zijn en dat de volgende letters "de" moeten zijn.

NIET EEN VAN DE VOLGENDE KARAKTERS

Stel we zoeken naar woorden van vier letters die beginnen met *hui*. We willen echter de overbekende woorden 'huis' en 'huid' uitsluiten. Dit kan door op de plaats van het laatste karakter het volgende te plaatsen: [^sd]. Dit betekent: niet de karakters s of d, verder alles toegestaan. (Het dakje staat voor *niet*) De reguliere expressie wordt aangevuld met de woordgrenzen en het cluster 'hui': *\bhui[^sd]\b*. 'huil' is een voorbeeld van een treffer.

WELK KARAKTER DAN OOK

Met de punt "." geven we de opdracht om naar een willekeurig karakter te zoeken. Dus de zoekopdracht *\bro.d\b* zoekt naar woorden van vier letters die beginnen met *ro* en eindigen op *d.* Het derde karakter is een "joker". Woorden als rood en rond voldoen dus aan de zoekopdracht. We kunnen ook invullen *\bro.*r*. Dan worden alle woorden gevonden die beginnen met *bro*, vervolgens een onbepaalde reeks (van 1 tot oneindig) willekeurig karakters bevatten, en eindigen op *r*. Een woord als *broodbakker* voldoet dus aan dit criterium.

WOORDEN VAN EEN BEPAALDE LENGTE

De lengte van een zoekterm duiden we aan met behulp van accolades. {2,5} wil zeggen een lengte van twee tot en met vijf karakters. Stel u bent op zoek naar woorden van zeven letters precies. Een goede zoekopdracht is dan *\b\w{7}\b*. De code *\w* betekent: elk karakter dat deel kan uitmaken van een woord. De *7* tussen accolades

wil zeggen: een lengte van zeven karakters precies. Weer sluiten we af met *\b*. Zouden we dit weglaten, dan levert het zoekproces ook woorden op die langer zijn dan zeven letters. Voor dat laatste doel bestaat ook een reguliere expressie. Als we na de komma geen cijfer invullen, (bijvoorbeeld *\b\w{7,}*) dan zoeken we naar woorden die minimaal de lengte van het eerste cijfer hebben, tot een onbeperkte lengte.

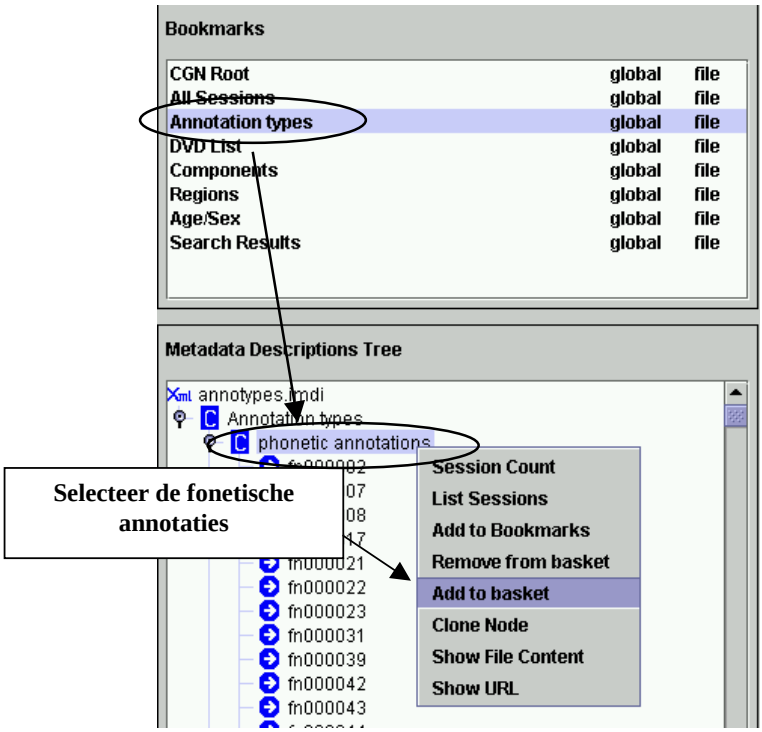
Een volledig overzicht van de syntax van reguliere expressies kunt u vinden in elke handleiding van de programmeertaal *Perl*. Ook op het world wide web is hierover veel informatie te vinden. Een goede website is:
http://www.english.uga.edu/humcomp/perl/regex2a.html#2.2

WOORDEN DIE EEN HOOFDLETTER BEVATTEN

Het zoeken naar woorden met hoofdletters werkt alleen als het vakje Case Sens. (moeilijk leesbaar) is aangevinkt. Dit is het vakje naast dat van Reg. Exp. Het vakje van Reg. Exp. zelf hoeft niet aangevinkt te zijn.

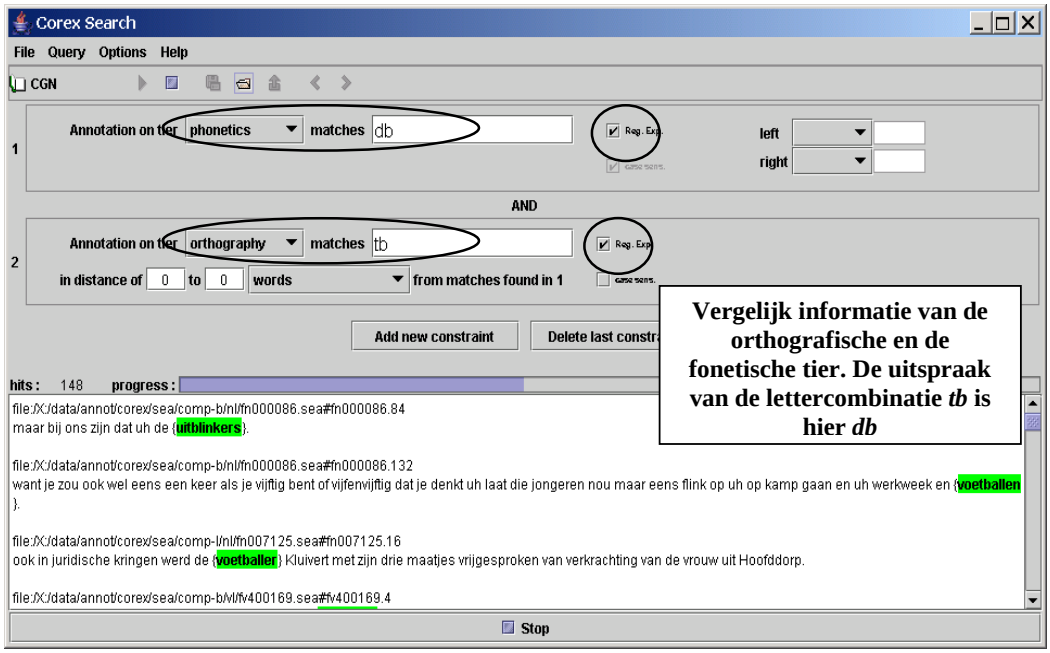
5. De Fonetische Transcriptie

Een deel van het CGN is fonetisch getranscribeerd. Deleties, inserties en substituties van fonemen zijn verwerkt in deze transcriptie. Het transcriptieprotocol is gedocumenteerd en staat ook op deze DVD.



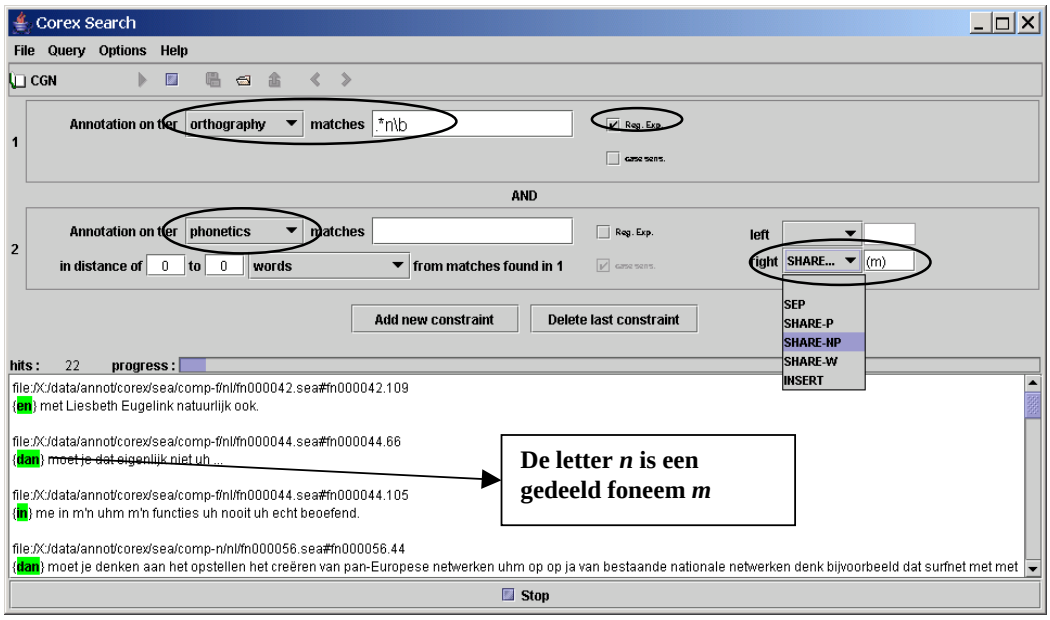
Alle sessies waarvan deze brede fonetische transcriptie (de **.bpt** file) voorhanden is, zijn gebundeld en te vinden onder Bookmarks → Annotation types → phonetic annotations. Ga hier naar toe en dubbelklik er op met de linker muisknop. Klik vervolgens op met de rechter muisknop erop en selecteer Add to basket.

In het onderstaande voorbeeld wordt een fonetische constraint gekoppeld aan een orthografische constraint. Er worden woorden gevonden die de lettercombinatie *td* bevatten die uitgesproken wordt als een *db*. Let er op dat Reg. Exp. aangevinkt is. Het gaat hier immers niet om hele woorden, maar om letters die deel uitmaken van een woord. Omdat de orthografische en de fonetische constraint betrekking hebben op hetzelfde woord, selecteren we *in distance of 0 to 0 words [...]*



Dan is er nog de toegevoegde mogelijkheid om te zoeken naar assimilaties, degeminatie en inserties over woordgrenzen heen. Dit kan met de knoppen left en right, die betrekking hebben op de linker- en rechter woordbegrenzing. Hier zit een pull down menu aan vast met de volgende opties:

- SEP geschieden
- SHARE-P een gedeeld plosief (plofklank)
- SHARE-NP een gedeeld non-plosief
- SHARE-W uitzondering: het woord/de woorden ‘da’s’
- INSERT een geïnserteerd foneem



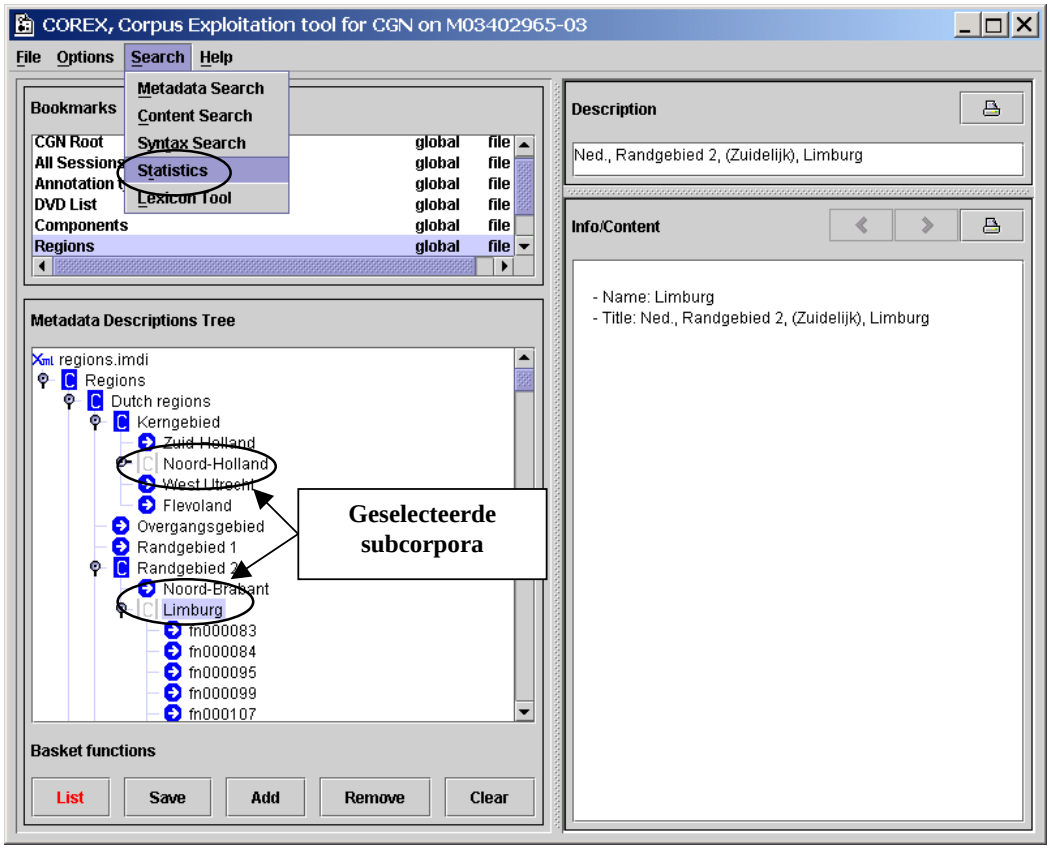
In het bovenstaande voorbeeld is er gezocht naar woorden die eindigen op de letter *n* – de eerste constraint, maar waarbij deze orthografische *n* het gedeelde foneem *m* is – de tweede constraint.

- De eerste constraint: het is een reguliere expressie voor een zoekvraag binnen de orthografie. Er wordt gezocht naar woorden die eindigen op een *n*. Vandaar de *\b* (boundary). Zie hoofdstuk 4 over reguliere expressies.
- De tweede constraint heeft betrekking op hetzelfde woord als de eerste constraint. Vandaar *in distance of 0 to 0 words [...]*. We kiezen voor Right en SHARE-NP. Het gaat immers om een gedeeld non-plosief aan de rechter begrenzing van het woord. Om precies te zijn zoeken we naar het foneem *m*. Dit vullen we in in de tekstbox.

De inhoud van het tekstvak moet tussen haakjes gezet worden. In het tekstvak staat default (.+). Dit kan gewoon blijven staan. Het is een reguliere expressie die betekent dat er niets wordt uitgesloten (syntactisch staat er: welk karakter dan ook, een of meerdere keren). Alleen als we op zoek gaan naar een bepaald foneem, vullen we zelf iets in tussen de haakjes. We kunnen ook hier gebruik maken van reguliere expressies, bijvoorbeeld als we zoeken naar een klasse van fonemen (zie hoofdstuk 4).

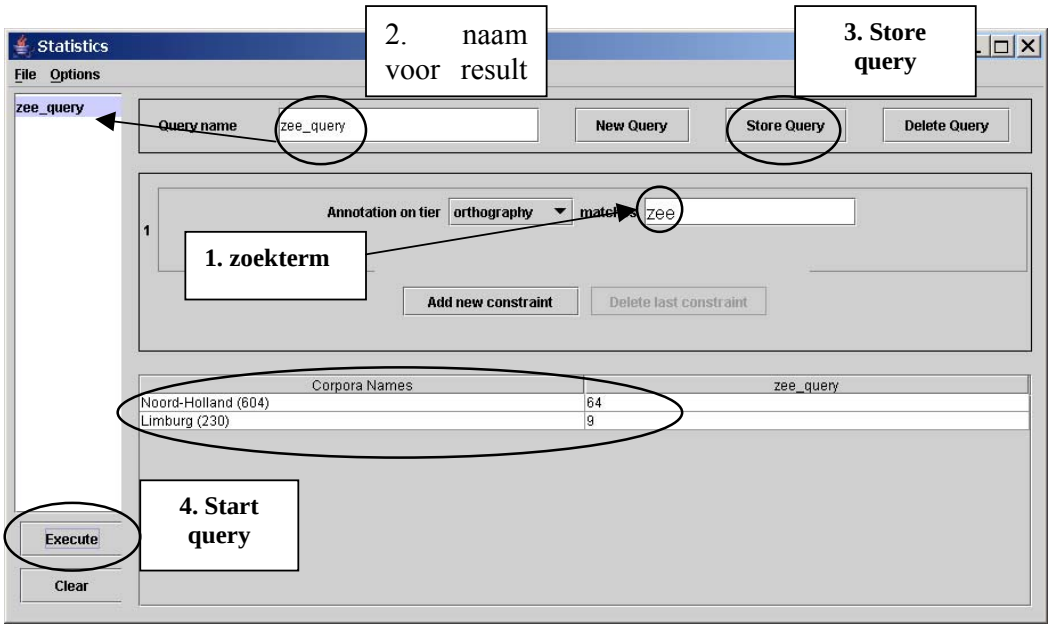
6. Statistiek

Het Statistics panel biedt de mogelijkheid om het aantal voorkomens van een woord of zoekterm in een van de andere *tiers* te tellen (part of speech, fonetiek, etc.). De telling kan uitgesplitst worden naar verschillende subcorpora. Op dezelfde wijze als bij content search kunnen er constraints worden toegevoegd aan een zoekopdracht. Dit maakt het mogelijk om bijvoorbeeld het aantal malen dat woord x en woord y in elkaars nabijheid zijn, te tellen.



Als we een Statistics search gaan uitvoeren, is het belangrijk dat er iets in de basket zit. Bij andere zoekfuncties wordt er *default* gezocht in het hele CGN. Dit is niet zo bij Statistics search. Open het (sub)corpus van uw keuze door te dubbelklikken. Klik nogmaals aan en kies dan voor de knop Add (of rechter muisknop en Add to basket). Ga dan naar het Search menu en kies Statistics.

We vragen ons bijvoorbeeld af: komt het woord 'zee' vaker voor in sessies met sprekers uit Noord-Holland dan in sessies met sprekers uit Limburg? We selecteren de twee subcorpora die sprekers bevatten uit de regio's in kwestie (zie boven) en kiezen voor Search → Statistics.

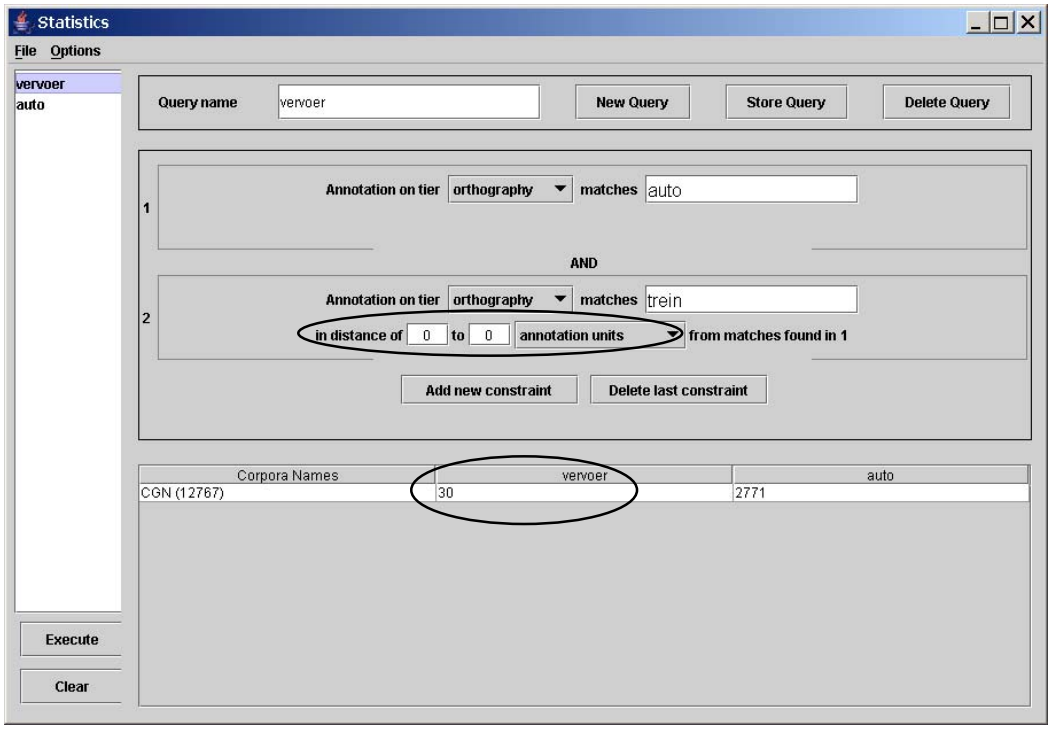


Het is belangrijk om de stappen in bovenstaand venster te volgen. Stap 1 en 2 mogen ook worden omgedraaid. Stap 3 – het opslaan van de query is verplicht. Doe dit pas als de zoekterm en de query name zijn ingevuld! Als deze drie stappen gedaan zijn, kan op Execute worden gedrukt (linksonder).

We zien aan de resultaat-frequenties dat het woord zee vaker wordt gebruikt in Noord-Holland dan in Limburg. Het aantal sessies met Noordhollandse sprekers is weliswaar groter dan het aantal sessies met Limburgse sprekers (zie de getallen tussen haakjes), maar niet genoeg om het verschil van een factor zeven goed te maken. Hierbij moet worden aangetekend dat niet elke sessie evenveel woorden bevat. Het aantal sessies is dus alleen een ruwe maat.

Het volgende voorbeeld is een zogenaamde *mutual information statistic*. Dat is een maat voor de sterkte waarmee woorden geassocieerd zijn. In het onderstaande voorbeeld is gekozen voor *auto* en *trein*. De verwachting is dat deze woorden vaak in dezelfde context voorkomen, dat wil zeggen vaker dan verwacht mag woorden op basis van hun onafhankelijke frequenties.

Allereerst maken we een query voor de woorden auto en trein. Met Add new constraint bepalen we het aantal keer dat *auto* en *trein* binnen dezelfde annotatie-eenheid voorkomen. Let op de selectie van de conditie in distance of 0 to 0 **annotation units** from matches found in 1. De query krijgt de toepasselijke naam vervoer – vergeet niet deze naam in te vullen en op Store Query te drukken.



Het aantal treffers is 30 (voor het hele CGN). Dit getal op zich zegt natuurlijk niets. Het moet vergeleken worden met de onafhankelijke frequenties van de woorden trein en auto. In bovenstaand voorbeeld is dit gedaan voor de frequentie van het woord auto (2771). Voor *trein* zou hetzelfde moeten gebeuren, waarna een maat voor de *joint probability* berekend zou moeten worden. Echter, omdat dit hoofdstuk over statistiek in Corex gaat en niet over statistiek in het algemeen, laat ik het hierbij.

Appendix : CGN Keys

N.B. Niet alle keys zijn voor elk fragment gespecificeerd

Session Keys (via Session)

CGN.wordCount	number of words in the sample: number
CGN.recCount	duration of the sample expressed in the total number of seconds: number
CGN.byteCount	indication of the size of the .wav file (expressed in terms of a number of units): number
CGN.tempoAV	Average number of words per hour: number
CGN.reDVDate	recording date: date or year
CGN.locName	place where the recording was made represented in terms of the (reduced) postal code or description of place in which the recording was made; possibly unknown or unspecified.
CGN.locale	description of the type of space in which the recording was made: loc1 = room of average size; loc2 = open air; loc3 =public place; loc4 = large room; unspecified
CGN.segmentation:	De wijze waarop het spraaksignaal is gesynchroniseerd met de annotatie: manueel of automatisch
CGN.availability	label of the dvd on which the sound file can be found; eg CGN_WAV_01
CGN.fon.available	{true,false} whether there is a manual transcribed phonetic annotation available or not
CGN.syn.available	{true,false} whether there is a syntax annotation available or not
CGN.pro.available	{true,false} whether there is a prosodic annotation available or not

Participant Keys (via Session/Participant)

CGN.age	age class to which speaker belonged at the time the sample was recorded; age0 = under 18 years of age; age1 = 18-24 years of age; age2 = 25-34 years of age; age3 = 35 -44 years of age; age4 = 45-55 years of age; age5 = over 55 years of age; ageX = age unknown
CGN.birth.year	year of birth; in case the information is not available 19nn is given as birthYear
CGN.birth.place	place of birth, represented in terms of the first three digits of the postal code preceded by the country code (B for Belgium, NL for The Netherlands; eg B-994, NL-832). For larger cities in the Netherlands several postal codes may apply. When this is the case, the postal code has been represented in terms of the first two digits that these codes have in common, while the variable third digit has been replaced by a hyphen (eg NL-25-). Where information concerning the place of birth is not available, xxx has been used. For speakers not born in Belgium or The

	Netherlands the place of birth is represented by the country code only.
CGN.birth.region	(geographical) region where the speaker was born. For a list of regions distinguished, see below.
CGN.firstLang	language (variety) speaker was raised in: SD = Standard Dutch; regiolect (eg regiolect: Antwerpen); dialect (eg dialect:Bree); unknown
CGN.homeLang	language (variety) speaker uses at home: SD = Standard Dutch; regiolect (eg regiolect: Antwerpen); dialect (eg dialect:Bree); unknown
CGN.workLang	language (variety) speaker uses at work: SD = Standard Dutch; regiolect (eg regiolect: Antwerpen); dialect (eg dialect:Bree); unknown
CGN.residence.place	speaker's place of residence, represented in terms of the first three digits of the postal code preceded by the country code (B for Belgium, NL for The Netherlands; eg B-994, NL-832). For larger cities in the Netherlands several postal codes may apply. When this is the case, the postal code has been represented in terms of the first two digits that these codes have in common, while the variable third digit has been replaced by a hyphen (eg NL-25-).
CGN.residence.reg	(geographical) region where the speaker resides. For a list of regions distinguished, see below
CGN.residence.size	indication of the (present) size of the place where the speaker resided while he was between 4 and 16 years of age; size1 = over 100,000 inhabitants; size2 = between 50,000 and 100,000 inhabitants; size3 = between 25,000 and 50,000 inhabitants; size4 = between 10,000 and 25,000 inhabitants; size5: between 5,000 and 10,000 inhabitants; size6 = fewer than 5,000 inhabitants; sizeX = size unknown
CGN.education.place	place where speaker attended secondary eductation place where speaker lived for the most part between ages 4 and 16) represented in terms of the first three digits of the postal code preceded by the country code (B for Belgium, NL for The Netherlands; eg B-994, NL-832). For larger cities in the Netherlands several postal codes may apply. When this is the case, the postal code has been represented in terms of the first two digits that these codes have in common, while the variable third digit has been replaced by a hyphen (eg NL-25-).
CGN.education.opleiding	type education; eg lager onderwijs, mbo, universiteit
CGN.education.reg	(geographical) region where speaker lived while he/she attended secondary eductation (region where speaker lived for the most part between ages 4 and 16). For a list of regions distinguished, see below.
CGN.education.level	level of education: edu1 = high, edu2 = middle, edu3 = low, eduX = unknown
CGN.occupation.level	occupational level. For the Netherlands occupational levels occ1 up to and including occ9 are distinguished. For Flanders, the levels occa up to and including occj are distinguished. In case someone has been trained for one occupation but presently

holds some different job, this has been indicated by combining two or more occLevel descriptions, as for example in occC+G where a professor is also a politician. occX is used whenever the occupational level is unknown.

CGN.occupation speaker's occupation

Content Keys (via Session/Content/Keys)

CGN.textclass.target gives information about four aspects: text type, degree of preparedness, mode, and domain;
text type specifies the component to which a sample belongs; 15 text types are distinguished; tta-tto (see list below)
degree of preparedness: prep1 = scripted, prep2 = unscripted, prep3 = more-or-less scripted;
mode: mod1 = broadcast, radio; mod2 = broadcast, tv; mod3 = non-broadcast
domain: dom1 = private; dom2= public

CGN.textclass.keywords one or more keywords that characterize the subject matter in the sample

CGN.activity short description of activity speaker(s) was (were) involved in at the time of recording

Text types:

tta	spontaneous conversations (face-to-face)
ttb	interviews with teachers of Dutch
ttc	spontaneous telephone dialogues (recorded via a switchboard)
ttd	spontaneous telephone dialogues (recorded on MD with local interface)
tte	simulated business negotiations
ttf	interviews/discussions/debates (broadcast)
ttg	(political) discussions/debates/meetings (non-broadcast)
tth	lessons recorded in a classroom
tti	live (eg sport) commentaries (broadcast)
ttj	newsreports/reportages (broadcast)
ttk	news (broadcast)
ttl	commentaries/columns/reviews (broadcast)
ttm	ceremonious speeches/sermons
ttn	lectures/seminars
tto	read speech

Geographical regions:

regN1aThe Netherlands, central region, Zuid-Holland, excl. Goeree Overflakee
regN1bThe Netherlands, central region, Noord-Holland, excl. West Friesland
regN1cThe Netherlands, central region, West Utrecht, incl. the city of Utrecht

regN2aThe Netherlands, transitional region, Zeeland, incl. Goeree Overflakee and Zeeuws-Vlaanderen
regN2bThe Netherlands, transitional region, Oost Utrecht, excl. stad Utrecht
regN2cThe Netherlands, transitional region, Gelders rivierengebied, incl. Arnhem and Nijmegen
regN2dThe Netherlands, transitional region, Veluwe up to the river IJssel
regN2eThe Netherlands, transitional region, West Friesland
regN2fThe Netherlands, transitional region, Polders
regN3aThe Netherlands, peripheral region 1 (north east), "Achterhoek"
regN3bThe Netherlands, peripheral region 1 (north east), Overijssel
regN3cThe Netherlands, peripheral region 1 (north east), Drenthe
regN3dThe Netherlands, peripheral region 1 (north east), Groningen
regN3eThe Netherlands, peripheral region 1 (north east), Friesland
regN4aThe Netherlands, peripheral region 2 (south), Noord-Brabant
regN4bThe Netherlands, peripheral region 2 (south), Limburg
regNx The Netherlands, unknown
regV1 Flanders, central region (Antwerpen and Vlaams-Brabant)
regV2 Flanders, transitional region (Oost-Vlaanderen)
regV3 Flanders, peripheral region 1 (West-Vlaanderen)
regV4 Flanders, peripheral region 2 (Limburg)
regVx Flanders, unknown
regW Wallonia
regZ region known to be outside of The Netherlands and Flanders
regX region unknown

Occupational level:

Nederlandse codes

occ1 occupation requiring higher level of education (doctor, lawyer, etc.)
occ2 occupation requiring middle level of education (teacher, journalist, etc.)
occ3 occupation requiring lower level of education (mechanic, teacher nursery school, bank employee, etc.)
occ4 occupation not requiring any level of education (garbage collector, cleaning lady, taxi driver, etc.)
occ5 holding no job, unemployed
occ6 holding no job, attending school
occ7 holding no job; housewife
occ8 holding no job, declared unfit
occ9 holding no job; other

Vlaamse codes

occA occupation in higher management or government
occB occupation requiring higher education
occC employed on the teaching or research staff in a university or a college
occD employed in an administrative office or a service organisation
occE occupation not requiring any level of specification
occF self-employed
occG politicians
occH employed with the media (journalist, reporter) or artist
occI student, trainee

occJ holding no job
occX unknown