Introductiepracticum

WordSmith 4: Concord

Eric Akkerman

Afd. Toegepaste Informatica Letteren

versie 1c – 30 januari 2007

Voorwoord

WordSmith is een verzameling programma-onderdelen die gebruikt kunnen worden bij de analyse van tekstbestanden. De belangrijkste zijn:

Concord: voor het maken van concordanties voor bepaalde woorden of woordgroepen. Door dit programma-onderdeel kunnen ook diverse overzichten worden geproduceerd die inzicht verschaffen in de lexicale patronen die in de tekst voorkomen (zoals collocaties en woordclusters), evenals overzichten van de plaats waar bepaalde woorden voorkomen in een tekst ('word plots'). Concord is het belangrijkste onderdeel van WordSmith.

Wordlist: voor het produceren van woordenlijsten en frequentielijsten op basis van de inhoud van één of meer tekstbestanden. Vanuit dit onderdeel kan ook informatie worden verkregen over de zgn. 'mutual information score' van bepaalde woorden, die aangeeft welke woorden significant vaak in elkaars omgeving worden aangetroffen.

Keywords: voor het identificeren van 'keywords' in een tekstbestand; dit zijn de woorden waarvan de frequentie hoog is vergeleken met een algemene norm. Om dit onderdeel te kunnen gebruiken is een referentiebestand nodig met algemene frequentiegegevens. Vanuit dit onderdeel kan ook informatie worden verkregen over de zgn. 'associates' van keywords. Dit zijn andere keywords die in meerdere teksten samen met het betreffende keyword voorkomen.

Daarnaast bevat WordSmith een aantal handige hulpprogramma's (*utilities*).

Doel van dit practicum is het bieden van een eerste kennismaking met het onderdeel Concord. Na het doorlopen ervan, beschik je over voldoende kennis om met dit onderdeel aan de slag te gaan. De eerste 11 paragrafen geven de meest basale informatie, die in paragraaf 12 nog eens wordt samengevat. In de paragrafen 13 t/m 16 wordt een aantal meer geavanceerde mogelijkheden behandeld. Wie meer wil weten, wordt verwezen naar de uitstekende helpfunctie van het programma zelf, of naar de online manual op het Internet op
`http://www.lexically.net/downloads/version4/html/index.html`
Je kunt de volledige handleiding downloaden op
`http://www.lexically.net/wordsmith/`

Studenten kunnen met WordSmith werken in de zalen 9A-05, 9A-11 en 10A-13, medewerkers kunnen op de pc in hun werkkamer een snelkoppeling (laten) maken naar het programma. Neem hiervoor s.v.p. even contact op met de helpdesk van de afdeling Systeembeheer.

Let op: als je dit practicum digitaal gebruikt, in een Windows-venster naast het WordSmith-venster, loopt WordSmith soms vast door het voortdurend switchen van vensters. Je moet het programma dan weer even opnieuw opstarten.

Ik houd me van harte aanbevolen voor opmerkingen, suggesties en kritiek ten aanzien van de inhoud van dit practicum. Tevens stel ik het op prijs als je eventuele problemen met het programma aan mij door zou willen geven, evenals positieve ervaringen en nuttige toepassingen.

Eric Akkerman
Studiegebied Toegepaste Informatica Letteren
Vrije Universiteit Amsterdam

Inhoudsopgave

1 WordSmith starten en één of meer bestanden selecteren _____ *1*
2 Het maken van een eenvoudige concordantie _____ *2*
3 Het manipuleren van het concordantie-overzicht _____ *3*
4 Het sorteren van de uitvoer _____ *3*
5 Het gebruik van wildcards bij het opgeven van zoekwoorden _____ *4*
6 Het opgeven van meerdere zoekwoorden _____ *6*
7 Het opgeven van een woordgroep ('phrase') als zoekwoord _____ *6*
8 Het bewerken van de concordantie _____ *7*
9 Het opslaan van concordantieregels in een bestand _____ *7*
10 Concordantieregels groeperen (Set) _____ *8*
11 Werken met contextwoorden _____ *9*
12 Samenvatting _____ *9*
13 Zoeken naar lexicale patronen in de tekst _____ *10*
14 Werken met gestructureerde tekst (Tags) _____ *11*
15 Literaire analyse _____ *12*
16 Taalkundige analyse _____ *13*

1 WordSmith starten en één of meer bestanden selecteren

(a)
Zet de computer aan en log in op het facultaire netwerk.
Studenten: start WordSmith 4 via *Start (op de Taakbalk)* → *Programma's* → *Data Analyse* → *WordSmith 4*.
Medewerkers: start WordSmith 4 met behulp van de daarvoor gemaakte snelkoppeling (zie het Voorwoord).

WordSmith werkt via het facultaire netwerk en verwacht een bestand met persoonlijke instellingen, genaamd `Wshell.ini`, op je persoonlijke netwerkstation U:, in een werkgebied (map) met de naam `WSmith`. Als je voor de eerste keer WordSmith gebruikt, en dit bestand nog niet bestaat, krijg je hier bij het starten van het programma een melding over ("Choose Adjust Settings … etc."). Dit is geen probleem, klik op [OK] (dat kan terwijl de zandloper nog op het scherm staat) en ga verder – dit wordt straks verholpen.

Als het programma gestart is, kom je terecht in het venster *Oxford WordSmith Tools*. Van hieruit kun je één of meer bestanden openen voor analyse (menu-optie *File*), de instellingen van WordSmith bekijken en wijzigen (*Settings*), de in het Voorwoord beschreven programma-onderdelen starten en informatie over het programma opvragen (*Help*). De help-optie is zeer uitgebreid en goed leesbaar - maak hier dus gebruik van.

Als je bij het starten van WordSmith een melding kreeg over `Wshell.ini` (zie hierboven), is het goed om hier eerst even iets aan te doen: klik op *Settings* → *Adjust Settings*, klik in het vakje voor "Save" (rechts boven), zodat er een vinkje in komt te staan, en klik vervolgens op [OK]. Sluit ook het volgende venster (waarin wordt gemeld dat de instellingen zullen worden bewaard in `U:\WSmith\Wshell.ini`) af met [OK], en je hebt je eigen instellingenbestand, dat je naar believen kunt wijzigen zonder dat anderen daar last van hebben.[1]

(b)
Hoewel dit ook kan vanuit de verschillende programma-onderdelen, zullen we nu vanuit het hoofdmenu het bestand selecteren waarmee gewerkt gaat worden. Klik op de menu-optie *File* → *Choose texts*. Selecteer (indien nodig) in het linker deel van het venster dat dan verschijnt het station P: (*Public*), daarin de map *Corpora* en daarin weer de map *Til*. Selecteer vervolgens het bestand Alice.txt, dat in het middelste venster rechts verschijnt (door er één maal op te klikken). Dit bestand bevat de tekst van *Alice's Adventures in Wonderland*, van Lewis Carroll. Klik op de blauwe pijl  in de middenblak om het bestand toe te voegen aan de (nu nog lege) bestandenlijst. Bevestig de selectie met een klik op de knop  . Als je weer terug bent in het *WordSmith Tools* venster moet je daar (ergens onderaan) het geselecteerde bestand vermeld zien staan. Zo niet, probeer het dan opnieuw.

[1] Als je later nog eens de instellingen wilt wijzigen, kun je dat ook doen zonder ze te 'saven'. In dat geval gelden de gewijzigde instellingen alleen voor de huidige sessie. Als je meestal met dezelfde bestanden werkt, is het handig om op het tabblad 'Directories' bij 'Texts' de padnaam van de betreffende map in te vullen (voor de MicroConcord corpora van Engels is dit b.v. P:\CORPORA\ENGELS\MCONCORD) en deze instelling te saven. Je komt dan na 'Choose texts' automatisch in de betreffende map terecht. Een andere handige instelling is 'Restore last work' op het tabblad 'General'.

N.B. Je kunt in WordSmith meerdere bestanden tegelijkertijd selecteren voor analyse. Gebruik in het bestandoverzicht (links) <Ctrl> plus muisklik om individuele bestanden aan een selectie toe te voegen of te verwijderen. Gebruik <Del> in het selectievenster om een eerder geselecteerd bestand weer te verwijderen.

2 Het maken van een eenvoudige concordantie

(a) Start het concordantieprogramma via de knop ***C** Concord* en maak het venster van dit programma-onderdeel schermvullend.

(b) Je begint met het maken van een eenvoudige concordantie van het woord *all*. Klik in het menu *File* op de start-knop  (*New …*). Hierna verschijnt het venster 'Getting Started', dat bestaat uit een aantal tabbladen: 'Texts', waarop bestanden kunnen worden geselecteerd, of een eerder gemaakte selectie kan worden gewijzigd, 'Search word', waarop je een zoekwoord kunt opgeven, 'Advanced, waarop je één of meer contextwoorden en een 'zoekhorizon' kunt opgegeven en 'Batch', dat je kunt gebruiken voor meervoudige concordanties. Over contextwoord en zoekhorizon volgt later meer informatie.

(c) Selecteer het tabblad 'Search word', type in het bovenste invulvak het zoekwoord ***all***, en klik op de knop [OK] om het zoekproces te starten. Na enige tijd (de voortgang van het analyseproces wordt op het scherm weergegeven) verschijnt de gewenste concordantie op het scherm.
N.B. Het kan zijn dat Concord een vraag stelt over het sorteren van de concordantieregels. Klik dit scherm weg (beantwoorden met "No" lukt niet…)

Het concordantievenster bevat de volgende kolommen:

(1e kolom): het nummer van de concordantieregel.
Concordance: hierin staan de concordantieregels van het opgegeven zoekwoord.
Set: hierin kun je zelf een letter zetten. Met behulp hiervan kun je bepaalde concordantieregels groeperen. Bij een concordantie van het woord 'school' zou je b.v. de betekenis 'institutie' kunnen markeren met de letter 'a', de betekenis 'gebouw' met de letter 'b', de betekenis 'de gezamenlijke scholieren' met de letter 'c', etc. Met behulp hiervan kun je later dan de concordantie herordenen.
Tag: als met gecodeerde tekstbestanden en een zgn. *tagfile* wordt gewerkt, wordt hier de code (tag) getoond die het dichtst bij het betreffende zoekwoord staat.
Word # etc.: het volgnummer van het zoekwoord in de brontekst, de positie van het woord; idem voor zin, paragraaf, kop en sectie.
File: de naam van het bestand waarin de concordantie regel is gevonden (vooral nuttig als meerdere bestanden zijn geselecteerd voor de analyse).
%: de plaats in de brontekst waar het zoekwoord staat (bij 25% staat het op een kwart van de tekst).

3 Het manipuleren van het concordantie-overzicht

(a) Je staat nu bovenaan het concordantie-overzicht. Je kunt door alle concordantieregels heen scrollen op de manieren die je bij Windows gewend bent. Ga met de toets <PgDn> naar de eerste concordantieregel (N=1) en bekijk ondertussen wat er allemaal langs komt.

(b) De kolommen met gegevens over woorden, zinnen, paragrafen en secties zijn niet altijd nodig of relevant. Je kunt ze verbergen met behulp van de menu-optie *View à Layout*: selecteer op het tabblad "Layout" één voor één alle kolommen tussen 'Tag' en "File' en klik op het keuzerondje voor "Hide".

(c) De ruimte die gebruikt wordt voor het tonen van de concordantieregel is vaak wat klein. Maak deze groter door de muiscursor op de grens tussen de titelbalken 'Concordantie' en 'Set' te plaatsen (hij verandert dan in het teken ß ‖à) en deze grens vervolgens naar rechts te slepen. Zorg er echter wel voor dat de overige kolommen nog te zien zijn.

(d) Zoals je ziet, wordt in eerste instantie elke concordantieregel op één tekstregel getoond. Je kunt de hoeveelheid getoonde tekst wijzigen door in het menu *View* de opties *Grow* en *Shrink* te gebruiken. Probeer uit wat het effect hiervan is, door eerst twee maal te vergroten en daarna weer twee maal te verkleinen.

(e) Tenslotte kun je via de menu-optie *View → Sentence only* instellen dat je in het overzicht zinnen wilt zien. Soms moet je daar overigens even geduld voor hebben. Voor deze mogelijkheid maakt WordSmith gebruik van een standaarddefinitie van een zin (beginnend met een hoofdletter en eindigend met een punt, een vraagteken of een uitroepteken).
N.B. In sommige gevallen (als je al veel hebt gerommeld met de lay-out van de concordantieregels) gaat dit mis. Sluit in dat geval het concordantieoverzicht en laat even een nieuwe concordantie maken.

(f) Schakel de instelling 'Sentence only' weer uit, en zorg dat je op één schermregel weer één concordantieregel ziet (m.b.v. de optie *Shrink*).

Je kunt vanuit het concordantiescherm ook naar de complete brontekst springen. Dit is handig als je de uitgebreide context wilt bestuderen. Probeer dit uit door te dubbelklikken op een concordantieregel. Ga weer terug naar het concordantieoverzicht door op de tab *concordance* te klikken (onderaan het WordSmith-venster).

4 Het sorteren van de uitvoer

Na een zoekopdracht worden de concordantieregels getoond in de volgorde waarin ze in het bronbestand zijn aangetroffen. Het is aan te raden om de regels te ordenen. Daarbij kunnen maximaal drie sorteersleutels gebruikt worden. Stel dat je bijvoorbeeld in eerste instantie sorteert op het woord rechts van het zoekwoord. Je ziet in de concordantie dan dat woordcombinaties als 'all about', 'all over' 'all round' en 'all sorts' bij elkaar staan. Met een tweede sorteersleutel kun je dan bijvoorbeeld alle groepen concordantieregels met hetzelfde rechterwoord onderling weer ordenen op het eerste woord links van 'all'.
Je kunt de sorteersleutels instellen en wijzigen met behulp van de optie *Resort* in het menu *Edit* (of m.b.v. functieknop F6). Je kunt dan de eerste sorteersleutel instellen bij *'Main sort'*, de tweede bij *'Sort 2'* en de derde bij *'Sort 3'*.

(a) Stel de eerste sorteersleutel in op “R1” (eerste woord rechts van het zoekwoord) en klik op [OK]. Bekijk het resultaat. Kijk vooral even goed naar de regels met de combinatie ‘all the’. Stel dan de tweede sorteersleutel in op “L1” (eerste woord links van het zoekwoord). Kijk ook nu weer nar het resultaat, met name bij ‘all the’.

(b) Stel nu de tweede sorteersleutel in op het tweede woord rechts van het zoekwoord ('R2') en klik weer op [OK]. De concordantieregels worden nu direct herordend. Het effect hiervan kun je weer goed zien bij de regels met de combinatie ‘all the’: deze groep is nu intern geordend op het tweede woord rechts van het zoekwoord ‘all’, zodat nu combinaties als ‘all the time’ en ‘all the while’ eruit springen.

(c) Stel dan de eerste sorteersleutel in op het eerste woord links van het zoekwoord ('1L'). Omdat je nu geen tweede sorteersleutel gebruikt, kun je daar klikken op het aankruisvakje voor 'Activated', zodat het betreffende vinkje verdwijnt. Klik dan op [OK] en bekijk het resultaat. Zoals je ziet, komen hierdoor uitdrukkingen als ‘after all’, ‘at all’ en ‘that’s all’ bij elkaar te staan.

(d) Bij het analyseren van teksten speelt het ontdekken van lexicale patronen vaak een belangrijke rol. Het sorteren van de concordantieregels op verschillende manieren is een belangrijke manier om dergelijke patronen op te sporen.[2] Zorg daarom dat je het instellen van de sorteervolgorde begrijpt, eventueel door hier nog wat mee te experimenteren.

(e) Stel tenslotte de eerste sorteersleutel weer in op het eerste woord rechts van het concordantiewoord en de tweede sorteersleutel op het eerste woord links. Vergeet niet daarvoor het betreffende aankruisvakje voor 'Activated' weer aan te vinken.

N.B.

(i) Bij de sorteersleutels staat 'Centre' voor het zoekwoord zelf. Het heeft alleen zin om dit te gebruiken als je met meerdere zoekwoorden werkt (zie hieronder).

(ii) Bij de sorteersleutels heb je ook de mogelijkheid om te sorteren volgens andere sleutels, zoals contextwoord (zie oefening 11), zelfgedefinieerde categorie (zie oefening 10) of voorgedefinieerde code-tag (zie oefening 13).

(iii) Gebruik voor informatie over de overige opties in dit venster de helpfunctie.

5 Het gebruik van *wildcards* bij het opgeven van zoekwoorden

WordSmith kent drie zgn. *wildcards*, waarmee gezocht kan worden naar groepen woorden die met een bepaalde reeks letters beginnen en/of eindigen. Dit zijn de asterisk (*) voor een willekeurig aantal willekeurige karakters, het dakje (^) voor exact één willekeurige letter en het vraagteken (?) voor exact één willekeurig karakter (inclusief cijfers en leestekens).

Sluit het subvenster met de concordantieregels zonder deze te bewaren. Klik weer op de start-knop om een nieuw zoekwoord op te geven. Geef als antwoord op de vragen die Concord je stelt aan dat je de nieuwe opdracht niet in een nieuw venster wilt zien en dat de uitvoer niet bewaard hoeft te worden. Bedenk eerst zelf wat de volgende zoekopdrachten zullen opleveren en probeer ze dan uit. Bekijk de uitvoer steeds goed en noteer welke woorden elke zoekopdracht oplevert. Sluit tussendoor steeds het venster met concordantieregels zonder deze te bewaren.

[2] Andere nuttige technieken op dit gebied worden in oefening 12 behandeld.

N.B. Als je een nieuwe concordantie start zonder eerst een eerder venster te sluiten, vraagt Concord “Start another Concord window?”. Als je deze vraag met “Yes” beantwoord, wordt de concordantie in een apart, nieuw, venster geopend (en blijft het andere behouden). Je kunt dan b.v. goed twee concordanties met elkaar vergelijken. Als je “No” antwoord, wordt vervolgens gevraagd of je de huidige concordantie wilt opslaan. Meestal zal dit niet nodig zijn.

(a) ***man****

.....

Omdat er nu meerdere woorden gevonden worden, is het nuttig om het zoekwoord als sorteersleutel te gebruiken. Probeer dit uit: geef 'Centre' op als eerste sorteersleutel en 'R1' als tweede.

(b) ****man****

.....

(c) ***un^^^^***

.....

Met behulp van de optie “exclude if text contains” op het tabblad “Advanced” kun je één of meer woorden opgeven die moeten worden uitgesloten. Dit kan relevant zijn als je met wildcards werkt.

(d) Herhaal om dit uit te proberen opdracht (a), maar sluit nu het woord ***man*** zelf uit.

.....

N.B.
Als je in de tekst wilt zoeken naar karakters die het programma zelf als *wildcard* gebruikt, moet je daar aanhalingstekens omheen plaatsen (gebruik b.v. "?" om te zoeken naar een vraagteken).

Sluit voor alle volgende oefeningen tussendoor steeds het venster met concordantieregels van de voorgaande oefening, zonder deze te bewaren. Verwijder bij normale zoekopdrachten de tekst uit het invoervak “exclude if text contains”.

6 Het opgeven van meerdere zoekwoorden

In één opdracht kunnen meerdere zoekwoorden worden opgegeven door deze achter elkaar te plaatsen, gescheiden door een *slash* (/). Start een nieuwe concordantie, wis eerst het veld 'but excluding'[3] en geef dan de volgende zoekopdracht op:

man/woman/boy/girl

N.B.

(i) Op deze wijze kunnen maximaal 15 alternatieven worden opgegeven (inclusief *wildcards*), waarbij de totale zoekopdracht echter maximaal 80 karakters kan bevatten. Als dit niet volstaat, is het ook mogelijk om de gewenste woorden op te slaan in een apart bestand. Dit is overigens ook een handige optie als je in verschillende bestanden naar eenzelfde reeks woorden wilt zoeken. Omdat dit niet in het bestek van dit practicum behandeld kan worden, verwijs ik hiervoor naar de help-informatie van het programma (zoek in de index naar "file-based search-words or phrases").

(ii) Je kunt het of-teken (/) ook gebruiken bij het uitsluiten van woorden (zie oefening 5).

7 Het opgeven van een woordgroep ('phrase') als zoekwoord

Je kunt als zoekpatroon tevens een combinatie van meerdere zoekwoorden opgeven. Ook hierbij kun je gebruik maken van *wildcards*. Als de asterisk (*) daarbij wordt omgeven door spaties, wordt deze geïnterpreteerd als 'willekeurig woord'. Ook kan gebruik gemaakt worden van het of-teken /, dat overigens functioneert op het niveau van de zoekopdracht, niet op het niveau van afzonderlijke woorden (zie oefening (b)).

(a) Start een nieuwe concordantie, geef als zoekwoord de woordgroep ***at all*** op en kijk wat het resultaat is. Bij het ordenen wordt 'at' beschouwd als het centrale zoekwoord, en 'all' als het tweede woord. Bedenk wat het effect zal zijn als de eerste sorteersleutel op '2R' gezet wordt, en de tweede op '1L', en probeer dit uit.

(b) Start een nieuwe concordantie en geef als zoekwoord de volgende woordgroep op:
after/at all
Stel de ordening in op Centre + 1R en kijk goed wat het resultaat is. Zoals je ziet, wordt dit niet geïnterpreteerd als enerzijds 'after all' en anderzijds 'at all', maar als enerzijds 'after' en anderzijds 'at all'. Kijk nu wat het resultaat is van de zoekopdracht
after all/at all

(c) Start een nieuwe concordantie, geef als zoekwoord de volgende woordgroep op:
to * to
Bekijk het resultaat en orden de uitvoer op het woord tussen 'to' en 'to' (1R).

[3] Als je dit niet doet, zal dit woord ook bij de volgende opdracht uitgesloten worden, ondanks het feit dat dat daarbij helemaal niet kàn. Dit is een foutje in het programma.

8 Het bewerken van de concordantie

Een concordantie bevat meestal wel regels die zoekwoorden bevatten die eigenlijk niet relevant zijn in het kader van een bepaalde zoekvraag. Je kunt dergelijke regels uit de concordantie verwijderen door de cursor erop te plaatsen en vervolgens op de <Del> toets te drukken. De betreffende regel wordt dan doorgestreept en krijgt dan een lichtgrijze kleur. Je kunt een markering weer verwijderen m.b.v. de <Ins(ert)> toets. Om de gemarkeerde regels daadwerkelijk te verwijderen moet je de optie *Zap* uit het menu *Edit* gebruiken..

Verwijder op de hierboven beschreven wijze de vijf concordantieregels uit het overzicht van oefening 7 (**to * to**) waarbij tussen 'to' en 'to' geen werkwoordsvorm staat (*Alice, her, them, -en you*).

9 Het opslaan van concordantieregels in een bestand

De concordantieregels voor een bepaald zoekwoord kunnen op twee manieren worden opgeslagen: als concordantiebestand (achtervoegsel .CNC) of als tekstbestand (achtervoegsel bij voorkeur .TXT). In het eerste geval kun je het opgeslagen bestand alleen weer met behulp van Concord bekijken. In het tweede geval kun je daarna de betreffende concordantieregels opnemen in je eigen (Word- of WordPerfect-) bestand.

(a)
Gebruik de optie *Layout* uit het menu *View* om alle kolommen behalve die met de concordantieregels te verbergen. Je hebt die in de uitvoer niet nodig.

(b)
Selecteer de menu-optie *File → Save as → Plain Text* (let op dat je hier niet kiest voor *Save*, omdat de uitvoer dan wordt opgeslagen als concordantiebestand). Het venster dat dan verschijnt bevat bovenaan een regel voor het opgeven van station, map en naam. Daarnaast kun je een aantal opties instellen. Zo kun je onder meer je eigen koptekst opgeven, en een specificeren welke rijen en kolommen wel of niet in de uitvoer moeten worden opgenomen.

- Geef als 'header' op: *Constructie "to – werkwoord – to" in Alice.txt*
- Sla dan de concordantie op in de map `WSmith` op station U: als bestand genaamd `toto.txt`.
  N.B. Gebruik de button [Openen] om op te slaan.

N.B. Als je (a) niet had gedaan, had je in het "Save as Plain Text"-venster ook de optie "Columns" in kunnen stellen op "Specify", door het keuzerondje daarachter aan te klikken en in het invulvakje achter "Specify" het cijfer 2 in te vullen. Hiermee geef je aan dat je alleen de tweede kolom (met de concordantieregels) wilt opslaan.

(c)
Selecteer uit de Taakbalk *Start à Programma's à Microsoft Office à Word*, en open in Word het bestand dat je zojuist hebt opgeslagen. Daarvoor moet je in het venster "Openen" eerst het bestandstype "Tekstbestanden" (of eventueel "Alle bestanden") selecteren. Maak het venster van *Word* schermvullend en bekijk hoe het bestand ziet (het kan overzichtelijker zijn

om even de gehele tekst te selecteren middel menu-optie *Bewerken à Alles selecteren*) en de lettergrootte in te stellen op 8 (in plaats van 10). Sluit dan de tekstverwerker weer.

(d)
Standaard worden per concordantieregel 160 karakters weggeschreven, omdat dan bij gebruik van een niet-proportioneel lettertype (zoals Courier) de zoekwoorden overzichtelijk onder elkaar staan. Je hebt dit gezien in het bovenstaande voorbeeld. Stel nu dat je meer context nodig hebt – je moet dan **voordat** je de concordantie opslaat een andere instelling kiezen. Selecteer hiervoor de menu-optie *Settings > Customise*. Ga naar het tabblad 'Concord' en verander het getal 816 bij de optie "Characters to save" in 250. Haal dan weer *Concord* naar de voorgrond, bewaar nogmaals de concordantie (alleen de eerste kolom!) en bekijk wederom het resultaat in *Word*. Zoals je ziet, heb je nu meer context opgeslagen, maar zijn de zoekwoorden niet meer overzichtelijk gecentreerd. Sluit hierna *Word* weer.

N.B.
(i) Per regel kun je met behulp van de instelling die is behandeld onder (d) minimaal 20 en maximaal 80.000 karakters laten wegschrijven.
(ii) Je kunt ook gedeelten van het concordantiescherm selecteren en middels kopiëren en plakken via het klembord overbrengen naar een ander programma.

10 Concordantieregels groeperen (Set)

Vaak is het gewenst om concordantieregels op een bepaalde manier te groeperen. Dit is bijvoorbeeld het geval als je in een (ongecodeerde) tekst zoekt naar het woord ‘huis’, en in de uitvoer de regels waarin ‘huis’ een zelfstandig naamwoord is bij elkaar wilt groeperen, evenals die waarin het een werkwoordsvorm is. Of: als je de betekenissen van ‘deken’ als kleed, als functie van een persoon en als deel van een vissersboot bij elkaar wilt groeperen. Je kunt hiervoor de kolom ‘Set’ gebruiken in het concordantiescherm, door in deze kolom bij elke regel van dezelfde categorie eenzelfde letter te plaatsen. Vervolgens kun je deze letters gebruiken om de concordantie op te sorteren.

(a) Start een nieuwe concordantie, geef als zoekwoord het woord ***too*** op (met twee o’s) en bekijk het resultaat. Zoals je ziet, komt ‘too’ voor in twee betekenissen: “ook” en “te”. Zet achter elke regel waarin de betekenis “ook” is de letter ‘o’ in de Set-kolom en achter elke regel waarin de betekenis “te” is de letter ‘t’. Je kunt de Set-kolom selecteren door erin te klikken.
(b) Klik op de re-sort knop en stel nu de eerste sorteersleutel in op ‘Set’ en de tweede op het eerste woord rechts van het zoekwoord. Bekijk ook nu het resultaat: als het goed is, worden nu de twee betekenissen van ‘too’ gegroepeerd, waarbij de groepen onderling zijn geordend op het woord dat rechts van ‘too’ staat. Zorg er hierbij voor dat de muiscursor niet in de kolom “Set” staat.

N.B. Je kunt een Set-code verwijderen door het cijfer 0 er overheen te typen. Je kunt een hele Set-kolom legen met behulp van de menu-optie *Edit > Clear set column*.

11 Werken met contextwoorden

Je kunt in *WordSmith* ook contextwoorden definiëren. Dat zijn woorden die voorkomen in de context van het zoekwoord. Je kunt daarbij voor zowel de linker- als de rechter context de zgn. *horizon* instellen, dat is het aantal woorden links en rechts van het zoekwoord dat moet worden doorzocht op het contextwoord. De standaard horizon is 5 woorden links en 5 woorden rechts. Het maximum aantal woorden dat je links en rechts kunt instellen is 25.
Stel na afloop de sorteersleutel weer even in op de standaardwaarde.

(a)
Je gaat uitzoeken hoe vaak de woorden ‘up’ en 'down' in elkaars context voorkomen. Start een nieuwe concordantie en selecteer het tabblad “Search”. Geef als zoekwoord ***up***. Selecteer dan het tabblad “Advanced”, vul dan als *context word* het woord ***down*** in en laat de concordantie maken. Zoals je ziet, levert dit zeven concordantieregels op. Omdat de standaard horizon 5 is (zowel links als rechts), vind je nu alleen voorkomens van ‘up’ en ‘down’ die vrij dicht bij elkaar staan.

(b)
Ga terug naar het scherm waar je de zoekwoorden kunt opgeven, laat het zoekwoord staan en verander nu in het optiescherm de horizonwaarden in 10 links en 10 rechts. Zoals je ziet, worden er nu zes extra concordantieregels gevonden, waarbij de woorden ‘up’ en ‘down’ verder van elkaar af staan (gebruik de 'grow'-knop om dit te kunnen zien).
N.B. Bij WordSmith 4 lijkt dit principe niet te werken (*bug*!).

N.B. Bij het opgeven van een contextwoord kun je gebruik maken van dezelfde opties als bij zoekwoorden (*wildcards*, combinaties met de / om alternatieven aan te geven, etc.). Je kunt ook woorden juist uitsluiten als contextwoord. Plaats er in dat geval een tilde (~) voor.

12 Samenvatting

***WordSmith Tools Controller*:**

Actie	**Toets, knop, menu-optie**
WordSmith Tools Controller starten	Vanuit Startmenu Windows XP
Instellingen aanpassen (eventueel) B.v. aantal karakters voor het uitvoerbestand	Menu *Settings à Adjust Settings* Tabblad Concord
Bestand(en) selecteren	Menu *File à Choose Texts* (gebruik knop [All] voor selectie van alle teksten in een map)

Concord:

Actie	**Menu > optie**
Zoekwoord(en) en evt. contextwoorden specificeren (ook: *wildcards*, meerdere zoekwoorden; woordgroepen)	*File > New*
Meer tekst tonen in contextregel	*View > Grow*
Ongewenste concordantieregels verwijderen	toets Del > *Edit > Zap*
Ongewenste kolommen in concordantiescherm verbergen	*View > Layout*
Concordantieregels groeperen	Letters plaatsen in kolom Set, daarna sorteren
Concordantieregels sorteren	*Edit > Resort*
Concordantie opslaan in tekstbestand	*File > Save as > Plain Text*

Je hebt nu voldoende basisvaardigheden geleerd om met het programma Concord te kunnen werken. In de nu volgende oefeningen wordt een aantal meer geavanceerde mogelijkheden van dit programma-onderdeel behandeld.

13 Zoeken naar lexicale patronen in de tekst

In deze oefening wordt aandacht besteed aan een aantal manieren om onderzoek te doen naar lexicale patronen. Vanuit het concordantiescherm staan hiervoor de volgende mogelijkheden tot je beschikking:

Collocaties: woorden die voorkomen in de buurt van het gebruikte zoekwoord, geordend naar frequentie.

Patronen: overzicht van de woorden die voorkomen in de buurt van het gebruikte zoekwoord, georganiseerd op basis van hun frekwentie als eerste buurwoord, tweede buurwoord, etc. links en rechts van het zoekwoord.

Clusters: patronen van frequent voorkomende woordgroepen rondom het zoekwoord.

(a) Start een nieuwe concordantie voor het woord ***time***. Vergeet niet het in oefening 11 opgegeven contextwoord ‘down’ te verwijderen!

(c) Selecteer het tabblad ‘collocates' (onderaan het Concord-venster) om de collocaties van het woord ‘time’ te bekijken. Bestudeer de uitvoer goed en gebruik eventueel de helpfunctie van WordSmith om nadere uitleg op te vragen.

(d) Selecteer het tabblad ‘patterns’ om de patronen van het woord 'time' te bekijken. Bestudeer de uitvoer goed en vraag eventueel met behulp van de knop [?] nadere uitleg op. Bestudeer de uitvoer goed en gebruik eventueel de helpfunctie van WordSmith om nadere uitleg op te vragen.

(c) Selecteer het tabblad ‘clusters’ om de patronen van het woord 'time' te bekijken. Bestudeer de uitvoer goed en vraag eventueel met behulp van de knop [?] nadere uitleg op. Bestudeer de uitvoer goed en gebruik eventueel de helpfunctie van WordSmith om nadere uitleg op te vragen.

14 Werken met gestructureerde tekst (Tags)

Steeds meer tekstbestanden zijn geannoteerd met extra-tekstuele informatie. Meestal gaat het daarbij om codes (*tags*) die informatie verschaffen over de structuur van de tekst. Denk daarbij b.v. aan een indeling in hoofdstukken, paragrafen, aan de markering van hoofdstuktitels en paragraafkopjes. Hoewel dat bij een optimale structurele codering feitelijk niet nodig is, worden in de praktijk ook vaak codes gebruikt om teksgedeelten te markeren die een bepaalde lay-out hebben, zoals cursief en vet. Hoewel dergelijke codes verschillende verschijningsvormen hebben, wordt tegenwoordig steeds vaker gebruik gemaakt van codes in een zgn. XML-formaat. Deze zijn meestal te herkennen aan het feit dat ze tussen hoekige haakjes <> staan, waarbij de eindcode van een tekstelement wordt voorafgegaan door een *slash* (/).

In de voorbeeldtekst `Alice.txt` worden de volgende codes gebruikt:

<title> aan het begin en </title> aan het einde de titel van het boek
<author> aan het begin en </author> aan het einde van de naam van de auteur
<chapter> aan het begin en </chapter> aan het einde van elk hoofdstuk[4]
<htitle> aan het begin en </htitle> aan het einde van elke hoofdstuktitel
 aan het begin en aan het einde van elke paragraaf
<verse> en </verse> aan begin en einde van elk rijmpje, liedje en gedichtje in de tekst
 en om elk stuk tekst dat cursief gedrukt is in de broneditie

WordSmith kent verschillende mogelijkheden voor het werken met dergelijke codes. Voor het merendeel valt dit onderwerp buiten het bestek van deze introductiecursus. Wel wil ik kort aandacht besteden aan de manier waarop van deze codes gebruik gemaakt kan worden door aan te geven dat bij het maken van een concordantie het zoekwoord uitsluitend in een bepaald tekstelement moet worden gezocht.

(a) Start een nieuwe concordantie voor het woord ***said*** en kijk hoe vaak dit woord in de tekst voorkomt. Sluit dan het concordantievenster weer.

(b) Je gaat nu de te doorzoeken tekst beperken tot alle tekst-elementen die zijn gemarkeerd als ‘verse’. Klik daarvoor op de menu-optie *Settings → Customise.* Ga dan naar het tabblad *Tags*. Zorg dat bij ‘Mark-up to ignore’ is ingevuld **<*>** (meestal standaard het geval). Het effect hiervan is dat alle codes tussen < en > genegeerd worden, behalve de tags die zijn opgenomen in een zgn. *tagfile*. Deze is al voor je gemaakt: in het bestand `Alice.tag` (in `P:\Corpora\Tagfiles`) zijn de codes <verse> en </verse> opgenomen als codes die wèl gebruikt moeten worden. Selecteer dit bestand bij ‘Tag file (mark-up to be included)’. Klik dan op de knop [Load] om het bestand daadwerkelijk te activeren en klik op [OK].

[4] De chapter-code heeft ook een zgn. ‘attribuut’ (n) waarmee het nummer van elk hoofdstuk wordt aangegeven.

(c) Klik tenslotte op de knop [Only Part of File]. Ga naar 'Sections to keep' en vul daar in: **<verse>** (inclusief de hoekige haakjes) TO **</verse>**. Klik vervolgens op [OK] om het instellingenvenster te sluiten.

(d) Haal Concord weer naar de voorgrond (door in het betreffende venster te klikken of door op de betreffende knop op de Taakbalk te klikken). Start dan een nieuwe concordantie voor het woord ***said***. Bekijk het resultaat (indien nodig met een vergroot overzicht van de concordantieregel): het aantal treffers is nu veel minder, omdat de zoekactie beperkt is gebleven tot de tekstelementen tussen de codes <verse> en </verse>.

N.B.

(i) In de kolom 'Nearest tag' van het concordantiescherm wordt de start-tag weergegeven die het dichtst bij het zoekwoord van de betreffende concordantieregel staat.

(ii) Vergeet niet een zoekbeperking zoals deze weer ongedaan te maken als je verder gaat (weer via het instellingenvenster).

(iii) WordSmith levert een tagfile voor gebruik bij de MicroConcord-corpora. Dit bestand, met de naam `Mconc.tag`, staat ook in `P:\Corpora\Tagfiles`.

15 Literaire analyse

Bij de analyse van literaire teksten kunnen de WordSmith-onderdelen *Wordlist* en *KeyWords* een nuttige rol spelen. Hiermee kunnen namelijk woorden worden opgespoord die kenmerkend zijn voor het vocabulaire van een bepaald werk. Voor meer informatie over deze onderdelen wordt verwezen naar de help-functie.

In het concordantiescherm kan de plot-functie nuttig zijn (zie tabblad 'plot'). Hiermee wordt een grafisch overzicht gegeven van de plaatsen in het bestand waar het zoekwoord voorkomt. Hiermee wordt in één oogopslag duidelijk waar een bepaald woord of begrip in de tekst voorkomt (bijvoorbeeld vooral in het eerste hoofdstuk en in de laatste twee hoofdstukken).

Een concreet voorbeeld: in *The Tempest* van W. Shakespeare zouden de elementen een belangrijke metaforische betekenis hebben. Je kunt dit onderzoeken door:

- met behulp van *Wordlist* een alfabetische woordenlijst en een frequentielijst te laten produceren op basis van de tekst;
- in deze lijsten te zoeken naar de (frequente) woorden die de elementen aanduiden (*storm, rain, gail, etc*);
- met behulp van *Concord* een concordantie te laten maken van de gevonden woorden (gebruikmakend van de combinatiefunctie /);
- het resultaat te onderzoeken en eventueel te laten plotten;
- als je beschikt over meerdere toneelstukken van Shakespeare in gedigitaliseerde vorm, het gebruik van de betreffende woorden ter vergelijking ook daarin te analyseren.

16 Taalkundige analyse

Bij taalkundig onderzoek wordt vaak gewerkt met geannoteerde tekstbestanden (ook wel *corpora* genoemd). Dergelijke bestanden zijn verrijkt met allerlei inhoudelijke informatie. Een bekend voorbeeld vormen de taalkundige tekstbestanden met zgn. morfo-syntactische codes. Dergelijke codes bevatten informatie over de woordsoort en de flexievorm van elk woord in de tekst. Omdat dergelijke codes er in verschillende corpora heel anders uit kunnen zien (soms zijn het cijfercodes, soms lettercode, soms staan ze na een spatie achter het woord (dus los daarvan), soms zijn ze daaraan verbonden met b.v. een liggend streepje), is het moeilijk om over het werken met dergelijke bestanden iets algemeens te zeggen. Je zult altijd een overzicht van de mogelijke codes en hun betekenis moeten gebruiken.

Wat wel gezegd kan worden als je op zoek bent naar meer ingewikkelde grammaticale structuren, is dat het goed is om het zoekpatroon stap voor stap op een logische wijze op te bouwen. Je kunt je daarbij voorstellen dat het zoekpatroon in een soort venstertje geplaatst wordt dat over het corpus heen geschoven wordt - daar waar het patroon op een deel van de tekst 'past' levert het een *hit* op.

In het Eindhovens corpus, bijvoorbeeld[5], een Nederlandstalig corpus waarin elk woord wordt gevolgd door een cijfercode, hebben bijwoorden een code die begint met het cijfer 5, terwijl bij ‘gewone’ bijwoorden achter die 5 een 0 staat. Als je nu in dit corpus wilt zoeken naar zinnen die een constructie bevatten bestaande uit een gewoon bijwoord, gevolgd door het woord 'niet', gevolgd door weer een bijwoord (waarvan het type er niet toe doet, pak je dat als volgt aan:

- eerst een gewoon bijwoord: dit is een willekeurig woord (**`*`**), gevolgd door een code die met 50 begint (**`50*`**);
- dan het woord 'niet', waarvan de code er niet toe doet (**`niet *`**);
- tenslotte een willekeurig bijwoord: dit is een willekeurig woord (**`*`**), gevolgd door een code die met 5 begint (**`5*`**).

Het totale zoekpatroon wordt dan dus: **`* 50* niet * * 5*`**

In de VU-versie van het LOB corpus, een Brits-Engels corpus waarin elk woord wordt gevolgd door een lettercode (waarin soms ook een cijfer kan voorkomen) die achter het woord is ‘geplakt’ met een liggend streepje, hebben bijvoeglijke naamwoorden de code ‘JJ’ en nevenschikkende voegwoorden de code ‘CC. Als je in dit corpus wilt zoeken naar sequenties van twee gecoördineerde bijvoeglijke naamwoorden (*black and blue* of *red or green*[6]), pak je dat als volgt aan:

- je zoekt eerst een willekeurig woord met een code die met een ‘J’ begint, hetgeen je kunt opgeven als **`*_J*`** (zonder spaties daartussen);
- dan een willekeurige woord met de code CC: **`*`**_cc (eveneens zonder spaties);
- en tenslotte weer een willekeurig woord met een J-code (zie boven);
- tussen de woord_code-combinaties staan natuurlijk wel spaties.

---

[5] In dit voorbeeld wordt ervan uitgegaan dat Concord cijfers beschouwt als onderdeel van een woord, en een getal dus ziet als een woord (dit is de standaardinstelling). Een sequentie als “een 450 mooi 100 huis 000” bestaat voor Concord dus uit zes ‘woorden’.

[6] Die in het corpus zijn gecodeerd als *black_JJ and_CC blue_JJ* en *red_JJ or_CC green_JJ*.

Het totale zoekpatroon wordt dan dus: `*_J* *_CC *_J*`