Handleiding Eindhoven-corpus

‘Vu-versie’

Dave van Grootheest

Vrije Universiteit Amsterdam
1992

1. Inleiding

Het Eindhoven-corpus probeert een soort dwarsdoorsnede te geven van het Nederlandse taalgebruik. Het telt rond de 720.000 woorden, en bestaat uit 6 subcorpora, elk in een apart bestand:

1.	Cdbl -	Corpus dagbladen	(regel 1- 6632)
2.	Cobl -	Corpus opiniebladen	(regel 6633-13445)
3.	Cgbl -	Corpus gezinsbladen	(regel 13446-21216)
4.	Crno -	Corpus romans en novellen	(regel 21217-31479)
5.	Cpwe -	Corpus populair-wetenschappelijke teksten	(regel 31480-37234)
6.	Cgtl -	Corpus gesproken taal	(regel 37235-51077)

Tevens is een bestand toegevoegd dat het Renkema-corpus bevat, met daarin "taal uit Den Haag":

7.	Camb -	Corpus ambtenarentaal	(regel 51078-53035)

Het Renkema-corpus maakt geen deel uit van het eigenlijke Eindhoven-corpus, maar is daar qua vorm en codering wel aan aangepast.
We zullen voor het gemak spreken over 7 sub-corpora, hoewel het dus eigenlijk 6 subcorpora en een apart corpus betreft.

Voor meer informatie over de achtergronden van het Eindhoven-corpus en het Renkema-corpus verwijs ik naar de volgende publikaties.
- Voor het Eindhoven-corpus:
  P.C. Uit den Boogaard (red.), 1975, Woordfrequenties in geschreven en gesproken Nederlands, Utrecht: Oosthoek, Scheltema en Holkema
- Voor het Renkema-corpus:
  J. Renkema, 1981, De taal van "Den Haag": een kwantitatief-stilistisch onderzoek naar aanleiding van oordelen over taalgebruik, 's-Gravenhage: Staatsuitgeverij

N.B. Men dient zich te bedenken, dat de beschrijvingen van de vorm en opbouw van beide corpora, die in deze boeken worden gegeven, niet zonder meer van toepassing zijn op de corpora in hun huidige -aangepaste- vorm. Voor up-to-date informatie bent u dus aangewezen op deze handleiding.

2. Opbouw van het corpus

Elk van de 7 subcorpora bestaat uit een aantal tekstfragmenten.
De lengte van deze fragmenten verschilt per subcorpus. Het boek van Uit den Boogaard geeft de volgende cijfers:

Cdbl: 75 woorden	Cpwe: 250 woorden
Cgbl: 125 woorden	Crno: 300 woorden
Cobl: 115 woorden	

(De fragmentlengte in Camb en Cgtl verschilt per fragment.)

Zo'n tekstfragment ziet er bv. als volgt uit:

```
[ <voorbeeld-1-cvrb>
1     < dit_360 is_240 de_370 eerste_103 zin_000 . >
2     < en_700 dit_360 is_240 zin_000 2_470 ! >
```

3 < met_600 deze_370 zin_000 sluiten_2510 we_301 het_370 tekstfragment_000 tenslotte_500 af_6205 . >]

- Elk fragment begint met "[" plus spatie.

- Dan volgt de bronvermelding. Dit is een verwijzing naar de herkomst van het fragment, die tussen "<" en ">" staat.

- Vervolgens komen de zinnen, 1 per regel (behalve bij te lange zinnen: hierover later meer).

- Elke zin begint met een zinsnummer, gevolgd door TAB. Dan volgt de eigenlijke zin, voorafgegaan door "<" plus spatie en afgesloten door spatie plus ">".

- Na de laatste zin wordt het fragment afgesloten door spatie plus "]".

- Elk woord is voorzien van een code, die woordsoort-informatie bevat. Deze code, die we de morfocode zullen noemen, is d.m.v. een underscore ("_") met het woord verbonden.

- Een morfocode bestaat normaliter uit 3 cijfers. In zin 3 van het bovenstaande voorbeeld staan 2 woorden met een 4-cijferige code. Dat vierde cijfer is toegevoegd in het geval van woorden die in hun basisvorm 1 woord vormen (in het voorbeeld: "afsluiten"), of anderszins bij elkaar horen, zoals bij woord-ellips (bv. "land- en tuinbouw"). Als het vierde cijfer van de morfocode een 0 is, dan verwijst dat naar het eerstvolgende woord met 4-cijferige code, waarvan het vierde cijfer 5 is (zie voorbeeld). Komt het verschijnsel meerdere malen in 1 zin voor, dan wordt niet alleen de combinatie 0 en 5, maar ook de combinatie 1 en 6 (en desnoods 2 en 7) als vierde cijfer gebruikt, om verwarring te voorkomen.

- Voor de betekenis van de morfocodes: zie Hfdst. 5.

Tenslotte een drietal DEFINITIES:

a) Onder "woord" wordt in deze handleiding verstaan:
   elk teken en elke tekenreeks die van een morfocode voorzien is.
   Het gaat dan om 1 of meerdere letters, cijfers en/of symbolen.
   Dus niet alleen een woord ("groot"), maar ook bv. een getal ("12"),
   een bedrag ("17.000,-") of een symbool ("&").
   Kortom: elke tekenreeks die wordt voorafgegaan door een spatie, en
   gevolgd door een underscore ("_").

b) Met "zin" wordt het volgende bedoeld:
   een teksteenheid die wordt begrensd door "<" en ">", en wordt
   voorafgegaan door een zinsnummer. (N.B. Bronvermeldingen worden
   ook begrensd door "<" en ">", maar hebben geen zinsnummer.)

c) De term "regel" tenslotte staat NIET voor een schermregel, maar
   voor een zgn. logische regel. Dat is een reeks van maximaal plm. 1000
   karakters, afgesloten met een LineFeed (="Return"), die over meerdere
   schermregels verspreid kan staan. In een tekstverwerker (zoals bv.
   "vi") heeft elke logische regel een eigen regelnummer.
   Normaal gesproken bevat 1 regel 1 zin, of een bronvermelding.
   Lange zinnen (meer dan 1000 karakters, d.w.z. meer dan 12 scherm-
   regels) zijn echter gesplitst (komt zo'n 65 keer voor), zodat ze op
   twee of zelfs drie regels staan. Ze zijn dan gesplitst op een plaats

waar normaal een spatie staat; die spatie is dan vervangen door een <RETURN>. Bij wijze van voorbeeld:

```
1      < als_730 dit_360 regel_000 `e`en_470 is_273 ,
dan_510 is_273 dit_360 regel_000 twee_470 ,
waarna_730 deze_370 zin_000 eindigt_243 op_600 regel_000 3_470 . >
```

Dit is dus 1 zin, verspreid over 3 regels - in de editor "vi" bv. zou elke regel een eigen regelnummer hebben, terwijl er maar 1 zinsnummer is. (Voor de goede orde: dit is maar een voorbeeld. Omdat deze zin bij lange na geen 1000 karakters telt, zou hij in het corpus gewoon op 1 regel staan!)
Programma's als "grep", die regel-gericht werken, zullen in het geval van gesplitste zinnen dus slechts de gevonden regel tonen, en niet de hele zin.

3. De in het corpus voorkomende tekens

In de zinnen van het corpus komen voor:

A.) Letters

Alleen kleine letters komen in het corpus voor. Hoofdletters zijn er niet, en ook letters met accenten komen niet voor. Een accent op een letter wordt nl. aangegeven d.m.v. een zgn. inverted comma (` ) voorafgaand aan de letter: het getal "1", voluit geschreven, wordt dus "`e`en".
Om welk accent het gaat is helaas niet aangegeven.

B.) Cijfers

Cijfers komen voor in getallen, bedragen, woorden als "a4-papier", en in morfocodes.
Wat dat laatste betreft: het is goed om te bedenken dat bv. het GETAL 990 altijd wordt voorafgegaan door een spatie (en gevolgd door een underscore), terwijl de MORFOCODE 990 wordt voorafgegaan door een underscore (en gevold door een spatie). Zo zijn ze van elkaar te onderscheiden.

C.) Bijzondere tekens

Een verzamelnaam voor alle tekens die geen letter en geen cijfer zijn. Het gaat hier om de volgende tekens:

```
 !  "  %  &  '  (  )  ,  -  .  /  :  ;  =  ?  @  `  {  }
```

Verder zijn er de spatie en de underscore, die alleen als scheidingsteken fungeren, en dus niet binnen een woord voorkomen.

De volgende tekens komen dus NIET voor in de zinnen van het corpus:

```
 #  $  *  +  <  >  [  \  ]  ^  |  ~
```

(N.B. < > [ ] komen dus wel voor als begrenzers van zinnen, bronvermeldingen en fragmenten!)

Bij de bijzondere tekens die wel voorkomen kunnen we verschillende categorieen onderscheiden:

- Leestekens:
  ! " ( ) , - . : ; ?

Dit zijn de interpunctie-tekens, die op zinsniveau dienen om woorden, woordgroepen of een hele zin te markeren/scheiden/verbinden etc.
Een leesteken staat vrijwel altijd alleen; de enige uitzondering hierop is "..." (drie puntjes).
Leestekens zijn omgeven door spaties. Ze hebben GEEN morfocode.

- Geisoleerde symbolen:
  % & - / = @

Deze lijken qua distributie (plaats en manier van voorkomen) veel op leestekens. Ze hebben echter een andere functie: je zou kunnen zeggen dat ze voor een woord staan ("procent", "en", "tot", "min", etc.).
Ze zijn dan ook voorzien van een morfocode, nl. 992 (rekenkundig gebruikte symbolen) of 993 (overige symbolen).

- Tekens binnen woorden:
  ! " % & ' ( ) , - . / : @ { } `

Deze tekens kunnen binnen een tekenreeks voorkomen. Soms hebben ze op woordniveau een zelfde soort functie als leestekens. Ze maken echter deel uit van het woord, en worden dan ook niet omgeven door spaties.

Uit het bovenstaande blijkt dat een teken kan behoren tot meerdere categorieen. Maar in een concreet geval valt uit de directe omgeving altijd op te maken, om welke categorie het gaat.
De symbolen @ { } ` zijn meta-symbolen met een speciale functie in de tekst, waar ik bij de bespreking van de afzonderlijke tekens (Hfdst. 4) op terug zal komen.

4. De bijzondere tekens afzonderlijk bekeken

We zullen nu de bijzondere tekens een voor een de revue laten passeren, zodat ook individuele eigenaardigheden aan de orde kunnen komen.

* ! (Uitroepteken)
-- Leesteken.
-- Binnen woord. Komt slechts 1 keer voor (in Cobl):
   - tora!-tora!-tora!_010 (naam van een film)

* " (Dubbele quote)
Dubbele quotes komen in principe paarsgewijs voor.
-- Leesteken: telkens twee dubbele quotes die een zin (of een grotere eenheid), een zinsdeel of een woord markeren.
-- Binnen woord: telkens twee dubbele quotes die een deel van het woord markeren. Bv.:
   - "d"-niveau_000
   - "libelle"-onderzoek_000

```
*  % (Procentteken)
-- Geisoleerd symbool (morfocode 993).
-- Binnen woord (komt alleen voor in Camb):
      -  %-regeling_000
      -  1%-operatie_000


*  & (Ampersand)
-- Geisoleerd symbool (morfocode 993): alleen Cgbl.
-- Binnen woord:
      -  snoeshaan-&-co_010
      -  fodor&-zoon_010

*  ' (Apostrof-teken)
-- Binnen woord: alleen als apostrof-teken (dus niet als quote). Bv.:
      -  ncrv's_012
      -  's_372 avonds_002

*  ( en ) (Haakjes)
-- Leestekens: omgeven een zin (of een grotere eenheid), een zinsdeel
   of een woord.
-- Binnen woord: omgeven een deel van het woord. Bv.:
      -  db(a)_001
      -  (ether-)zendtijd_000

*  , (Komma)
-- Leesteken.
-- Binnen woord: vooral in bedragen. Bv.:
      -  2,95_470
      -  jawereld,neewereld_010

*  - (Streepje)
-- Leesteken: gedachtenstreepje.
-- Geisoleerd symbool: minteken (morfocode 992) of "tot"-symbool
   (morfocode 993).
-- Binnen woord: o.a. koppelteken, "tot"-symbool. Veel gebruikt als
   koppelteken in namen (die namelijk meestal als 1 woord geschreven
   zijn). Bv.:
      -  west-duitsland_010
      -  1963-1964_470
      -  europese-commissie_010

*  . (Punt)
-- Leesteken.
-- Binnen woord: vooral in afkortingen en getallen. Bv.:
      -  a.s._103
      -  14.000_470

*  / (Slash)
-- Geisoleerd symbool (morfocode 993).
-- Binnen woord: o.a. deelstreep in getallen. Bv.:
      -  31/2_470   (dit staat in het corpus meestal voor
                      "drie-en-een-half"...!)
      -  en/of_700  (maar dit kan ook geschreven zijn in "losse"
                     woorden!)

*  : (Dubbele punt)
-- Leesteken.
-- Binnen woord: komt twee keer voor, in een verwijzing naar een tekst
   uit de Bijbel.
      -  prediker_010 3:4_470
```

- lucas_010 2:16_470

* ; (Puntkomma)
-- Leesteken.
* = (is-gelijk-teken)
-- Geisoleerd symbool: met morfocode 992 (betekenis "is gelijk aan...") of 993 (betekenis "betekent...").

* ? (Vraagteken)
-- Leesteken.
-- Binnen tekenreeks: meestal in tweevoud ("??"), al dan niet in combinatie met andere tekens. Bv.:
- zo'n_370 ... (??_998) ding_000 of_700 'n_450 trui_000 of_700 ...
(in Cgtl: duidt een onduidelijke passage aan)
- tabel_000 ??_470
(in Camb: staat voor een onbekend getal)
- artikel_000 ??.2.17_470
(idem)
- 19??_470
(idem, een onbekend jaartal)
- fout(?)_998
(in Cgtl)

* @ ("At"-symbool)
Meta-symbool: staat voor allerlei weinig voorkomende symbolen, en is dus een soort symbool-symbool.
-- Geisoleerd symbool (morfocode 993): staat meestal voor "plus-minus" of het graden-symbool, maar kan ook staan voor bv. een Grieks woord.
-- Binnen woord: vaak in een wiskundige term o.i.d. (vooral Cpwe).
- @@@_992 (Cpwe: een wiskundige term)
- @2_460 (Cpwe: "wortel twee")
- 8@5_470 (Camb: betekenis onbekend)

* ` (Inverted comma)
Meta-symbool: geeft aan dat de volgende letter een accent heeft.
-- Binnen woord.
- h`och_010 (een naam)
- `e`en_470

* { en } (Accolades)
Meta-symbolen: markeren een woord of woorddeel dat elders in de zin is weggelaten.
-- Binnen woord:
- land-_0000 en_700 tuin{bouw}_0005
- genees-_0000 en_700 andere_103 {middelen}_0005
- runder-_0000 , kalfs-_0000 , varkens-_0000 of_700 gemengd_216 {gehakt}_0005
- frequentie-_0000 (of_700 : waarschijnlijkheids-_0000) {verdeling}_0005
Deze markering zou bv. van belang kunnen zijn bij het produceren van een woordenlijst, om zowel "landbouw" als "tuinbouw" en zowel "geneesmiddelen" als "middelen" in zo'n lijst te krijgen.
Bovenstaande gevallen zijn tevens goede voorbeelden van het gebruik van 4-cijferige morfocodes. Als de twee bovenste voorbeelden in 1 zin gebruikt zouden zijn, zouden ze bv. als volgt zijn gecodeerd:
- land-_0000 en_700 tuin{bouw}_0005 , genees-_0001 en_700 andere_103 {middelen}_0006

5. Overzicht van de morfocodes en hun betekenis

```
--- DE EERSTE TWEE CIJFERS -------------------------------------

0..  als 1e cijfer : SUBSTANTIEVEN ..............................
 0.  als 2e cijfer : "gewone" substantieven
 1.   "  "    "    : eigennamen
 2.   "  "    "    : adjectivisch gebruikte substantieven
 8.   "  "    "    : interjectivisch gebruikte substantieven
 9.   "  "    "    : woorden in zelfnoemfunctie


1..  als 1e cijfer : ADJECTIEVEN ................................
 0.  als 2e cijfer : "gewone" adjectieven
 2.   "  "    "    : substantivisch gebruikte adjectieven
 5.   "  "    "    : adverbiaal gebruikte adjectieven
 8.   "  "    "    : interjectivisch gebruikte adjectieven



2..  als 1e cijfer : WERKWOORDSVORMEN ............................

(0-3  als 2e cijfer: deelwoorden en infinitieven)
 0.  als 2e cijfer : deelw./inf. van intransitief gebruikte ww.
 1.   "  "    "    :   "   / "   van transitief gebruikte ww.
 2.   "  "    "    :   "   / "   van reflexief gebruikte ww.
 3.   "  "    "    :   "   / "   van hulp- of koppelww.


(4-8  als 2e cijfer: persoonsvormen)
 4.  als 2e cijfer : pers.vorm van intransitief gebruikte ww.
 5.   "  "    "    :     "     van transitief gebruikte ww.
 6.   "  "    "    :     "     van reflexief gebruikte ww.
 7.   "  "    "    :     "     van hulp- of koppelww.
 8.   "  "    "    :     "     van interjectivisch gebruikte ww.



3..  als 1e cijfer : PRONOMINA a ................................
 0.  als 2e cijfer : personalia
 2.   "  "    "    : possessiva    (zelfstandig)
 3.   "  "    "    :     "         (bijvoeglijk)
 4.   "  "    "    : reflexiva     (zelfstandig)
 5.   "  "    "    :     "         (bijvoeglijk)
 6.   "  "    "    : demonstrativa (zelfstandig)
 7.   "  "    "    :     "         (bijvoeglijk)


4..  als 1e cijfer : PRONOMINA b ................................
 0.  als 2e cijfer : interrogativa (zelfstandig)
 1.   "  "    "    :     "         (bijvoeglijk)
 2.   "  "    "    : relativa      (zelfstandig)
 3.   "  "    "    :     "         (bijvoeglijk)

 4.   "  "    "    : indefinita    (zelfstandig)
 5.   "  "    "    :     "         (bijvoeglijk)
 6.   "  "    "    : cardinalia    (zelfstandig)
```

```
 7.   "  "    "    :     "         (bijvoeglijk)
 8.   "  "    "    : ordinalia     (zelfstandig)
 9.   "  "    "    :     "         (bijvoeglijk)


5..  als 1e cijfer : BIJWOORDEN ..............................
 0.  als 2e cijfer : "gewone" bijwoorden
 1.   "  "    "    : aanwijzende en onbepaalde bijwoorden
 3.   "  "    "    : betrekkelijke bijwoorden
 4.   "  "    "    : voornaamwoordelijke bijwoorden
 5.   "  "    "    : vragende voornaamwoordelijke bijwoorden
 6.   "  "    "    : betrekkelijke voornaamwoordelijke bijwoorden
 8.   "  "    "    : interjectivisch gebruikte bijwoorden


6..  als 1e cijfer : PRE- en POSTPOSITIES .....................
 0.  als 2e cijfer : voorzetsels
 1.   "  "    "    : 2e deel gesplitste voornaamwoordelijke bijwoorden
 2.   "  "    "    : niet-verbaal deel van samengestelde werkwoorden
 3.   "  "    "    : 2e deel gescheiden voorzetselgroepen
 4.   "  "    "    : achterzetsels
 5.   "  "    "    : "te" bij infinitief / voorzetsel met een
                     "te"-plus-infinitief-constructie als rest
 6.   "  "    "    : voorzetsels met een door onderschikkend voegwoord
                     ingeluide bijzin als rest


7..  als 1e cijfer : CONJUNCTIES .............................
 0.  als 2e cijfer : nevenschikkende voegwoorden
 1.   "  "    "    : onderschikkende voegwoorden
 2.   "  "    "    : voegwoorden van vergelijking
 3.   "  "    "    : onderschikkende voegwoorden die afwijkende
                     hoofdzinsvolgorde regeren
 4.   "  "    "    : inleidend deel van nevenschikkende voegwoord-groepen


8..  als 1e cijfer : INTERJECTIES ............................
 0.   "  "    "    : "echte" interjecties
 1.   "  "    "    : substantivisch gebruikte onomatopeeen


9..  als 1e cijfer : RESTGROEPEN .............................
 9.  als 2e cijfer
  0  als 3e cijfer : anderstalig
  1   "  "    "    : niet-lexicale verbindingselementen
  2   "  "    "    : rekenkundige termen/symbolen
  3   "  "    "    : overige symbolen
  8   "  "    "    : niet tot de steekproef behorende elementen
{992 en 993 zijn nieuwe codes: kwamen niet voor in originele corpus}
```

--- HET DERDE CIJFER ----------

werkwoordsvormen: DEELWOORDEN en INFINITIEVEN .....
(2.. als 1e cijfer, 0-3 als 2e cijfer)

..0 infinitief (verbaal)
..1 infinitief (substantivisch)
..2 onvoltooid deelwoord (onverbogen)
..3 " " (verbogen)
..4 " " (meervoud)
..5 " " (bijwoordelijk)
..6 voltooid deelwoord (onverbogen)
..7 " " (verbogen)
..8 " " (meervoud)
..9 " " (bijwoordelijk)

werkwoordsvormen: PERSOONSVORMEN .....
(2.. als 1e cijfer, 4-8 als 2e cijfer)

..1 1e persoon singularis presens
..2 2e " " "
..3 3e " " "
..4 pluralis presens
..5 singularis preteritum
..6 pluralis preteritum
..7 imperatief zonder pronomen personale
..8 imperatief met pronomen personale
..9 overige vormen

ANDERE WOORDSOORTEN (1e cijfer niet 2 en niet 9) .....
..0 basisvorm
..1 meervoudsvorm
..2 genitief
..3 overige verbogen vormen
..4 comparatief (onverbogen)
..5 " (genitief)
..6 " (andere verbogen vormen)
..7 superlatief (onverbogen)
..8 " (genitief)
..9 " (andere verbogen vormen)

--- HET VIERDE CIJFER ----------

woord met code ...0 hoort bij woord met code ...5
" " " ...1 " " " " " " ...6
" " " ...2 " " " " " " ...7

(N.B. "hoort bij..." wil zeggen:
"vormt in basisvorm 1 woord met, of staat in een bijzondere
verhouding tot..." -- zie ook Hfdst. 2)

Bijlage 1: Achtergrondinformatie over deze versie van het corpus

```
Verschillen met vorige versies
------------------------------
Het uitgangspunt van deze versie van het corpus was de PC-versie van
A. Verhagen. Die PC-versie week op een aantal punten af van het
originele corpus. Exacte documentatie hierover ontbreekt echter.
Vast staat wel dat A. Verhagen onder andere allerlei vormfouten
verbeterd heeft en zinnen gesplitst heeft die te lang waren om door de
gangbare programmatuur (bv. grep, vi, ex) op de juiste wijze te worden
verwerkt.
Hieronder zal ik een opsomming geven van de belangrijkste veranderingen
die ik zelf in het corpus heb aangebracht.

[1]
Ik heb allerhande vormfouten verbeterd, die nog steeds in ruime mate
aanwezig waren. Enkele veel voorkomende fouten:
-- fouten m.b.t. spaties (dubbele spatie, geen spatie, spatie verkeerd
   geplaatst), vooral in de buurt van getallen
-- verkeerd geplaatste apostrof-tekens (na de morfocode i.p.v. aan het
   einde van het woord)

[2]
Naast regelrechte fouten kwamen er ook allerlei andere eigenaardigheden
in het corpus voor. Zo werd i.p.v. het uitroepteken de rechter accolade
gebruikt, terwijl het uitroepteken zelf gebruikt werd om een accent op
de volgende letter aan te geven (bv. "!e!en").
Ik heb het uitroepteken in ere hersteld, en heb als accentteken de
inverted comma (`) gekozen (bv. "`e`en").

[3]
De codering van leestekens was op z'n zachtst gezegd eigenaardig:
het uitroepteken (weergegeven als "}") en de puntjes ("..." etc.)
hadden wel een morfocode, de overige leestekens niet. Deze morfocodes
heb ik laten vervallen, zodat in het algemeen geldt dat leestekens geen
morfocode hebben.

[4]
Het gebruik van quotes (aanhalingstekens) was nogal inconsistent:
waar de ene keer enkele quotes (') gebruikt werden, werden de andere
keer dubbele quotes (") gebruikt. Het kwam zelfs voor dat een citaat
met een enkele quote begon en met een dubbele eindigde.
Al met al viel er vrijwel geen lijn in te ontdekken, en daarom heb ik
het gebruik van de quotes gestandaardiseerd: overal waar een quote als
aanhalingsteken wordt gebruikt, staat nu een dubbele quote, zodat de
enkele quote alleen nog als apostrof-teken wordt gebruikt.

[5]
Een aantal symbolen was niet van een morfocode voorzien (vooral "@",
het "symbool-symbool", maar ook bv. "%", "&", "/"). Speciaal voor deze
gevallen heb ik de codes 992 (rekenkundige termen/symbolen) en 993
(overige symbolen) ingevoerd, zodat nu alle niet-leestekens in de zin
een morfocode hebben.

[6]
Anderstalige woordgroepen en zinnen waren vaak aan elkaar geschreven,
voorzien van de morfocode 990. Omdat dit het lezen bemoeilijkt en soms
heel lange reeksen oplevert, waar bepaalde programma's weer over kunnen
struikelen, heb ik deze reeksen zo goed en zo kwaad als dat ging
```

uiteengerafeld, en alle losse woorden voorzien van de morfocode 990.
(Voor wie het ook eens wil proberen: enige kennis van Nederlandse dialecten, Middelnederlands, Duits, Frans, Spaans, Latijn en vooral Engels, plus enig puzzel-oplossend vermogen, wordt aanbevolen.)

[7]
Ik heb het systeem van woordellips-markering wat uitgewerkt. Oorspronkelijk stond er alleen een "[" op de plaats waar het gezamenlijke deel van de twee woorden in kwestie begon of ophield (bv. "lees- en weet[honger". Ik heb dit systeem zo aangepast, dat het gezamenlijke woorddeel nu omgeven is door accolades ("lees- en weet{honger}") .
N.B. Bij woordellips hebben de twee woorden in kwestie allebei een 4-cijferige morfocode.

[8]
De bronvermeldingen heb ik zodanig gestandaardiseerd, dat elke bronvermelding nu eindigt op de naam van het subcorpus (in kleine letters; bv. <tel-2-10-1-cdbl>).

[9]
De nogal complexe situatie, waarbij een fragment (na de bronvermelding) begon met "<<>" en eindigde met ">>", heb ik opgeheven door respectievelijk "[" en "]" als fragmentbegin en -einde te gebruiken.

[10]
Desambiguering van een aantal dubbelzinnige tekens is een van de belangrijkste effecten van de door mij doorgevoerde wijzigingen. Maar de meest grootschalige ambiguiteit tot nu toe was wel die van de spatie, die zowel woord van morfocode scheidde als woorden (+ morfocode) van elkaar. Deze ambiguiteit heb ik opgelost door de spatie tussen woord en morfocode te vervangen door een underscore. Een grootschalige operatie, die echter alleszins de moeite waard is: de leesbaarheid van het corpus wordt verbeterd, het is in een oogopslag duidelijk wat waar bij hoort, getallen en codes zijn veel beter uit elkaar te houden, en allerlei ambiguiteiten zijn de wereld uit.

Enkele aantekeningen

(1)

De belangrijkste ambiguiteiten die nu tot het verleden behoren:
-- Getallen en morfocodes:
     zoeken naar het getal   100 is zoeken naar " 100_",
     zoeken naar de morfocode 100 is zoeken naar "_100 ".
-- Het streepje: kon losstaand een gedachtenstreepje voorstellen, een minteken, of een "tot"-symbool . Nu niet meer:
     gedachtenstreepje: " - "
     minteken: " -_992"
     "tot"-symbool: " -_993"

(2)
In Cgtl werden puntjes-reeksen ("..." etc.) op twee manieren gebruikt:
  - om een spreekpauze of aarzeling aan te duiden (morfocode 991)
  - om een onduidelijke passage aan te duiden (morfocode 998)
In het eerste geval heb ik net zo gehandeld als bij de puntjes in de andere subcorpora: ik heb de morfocode verwijderd, en ze dus als leestekens beschouwd.
(Het zijn dus "gewone" puntjes geworden, maar met dien verstande dat ze

in Cgtl een speciale functie hebben, nl. dat ze spreekpauzes/aarzelingen aangeven.)
In het tweede geval heb ik de code wel laten staan: het gaat hier niet om een leesteken, maar om inhoudelijke (zij het onduidelijke) informatie.
Zo is tevens het onderscheid behouden tussen de beide functies van de puntjes in Cgtl.

(3)
Oorspronkelijk waren in Cgtl alle uitingen van interviewers gecodeerd als "998" ("niet tot de steekproef behorend"). A. Verhagen heeft deze zinnen van relevante morfocodes voorzien, en deze zinnen gemarkeerd als "vermoedelijke uitingen van de interviewer".
In de huidige versie is deze markering verwerkt als "int_998" aan het begin van zo'n zin.

Bijlage 2: Technische informatie over deze versie van het corpus

```
 [1]  Voor elk subcorpus geldt de volgende grammatica:

 1.  subcorpus  ::= (block)+
 2.  block      ::= "[ " HEAD SNTS " ]" <LineFeed>
 3.  HEAD       ::= "<" HEADER ">"
 4.  HEADER     ::= ([a-z] | [0-9] | [-./])+
 5.  SNTS       ::= (SNT)+
 6.  SNT        ::= <LineFeed> SNTNR <TAB> "< " SENTENCE ">"
 7.  SNTNR      ::= ([0-9] | [ab])+
 8.  SENTENCE   ::= (EXP)+
 9.  EXP        ::= ((WORD "_" CODE) | INTERP) SEP
10.  WORD       ::= ([a-z] | [0-9] | [!"%&'(),-./:=@`{}])+
11.  CODE       ::= [0-9][0-9][0-9] | [0-9][0-9][0-9][0-9]
12.  INTERP     ::= [!"(),-.:;?] | "..."
13.  SEP        ::= " " | <LineFeed>
14.  <LineFeed> ::= CHR(10)
(N.B. "(...)+" wil zeggen: "1 of meer maal...")


[2]  Het aantal regels waarin een bepaald bijzonder teken voorkomt:


      camb  cdbl  cgbl  cgtl  cobl  cpwe  crno
      ----  ----  ----  ----  ----  ----  ----
!        0    28   185   181    92    93   423
"       97  1034  1474   339  1225   819  2614
%       36     0    10     3    29    22     0
&        0     2     4     0     7     2     1
'       47   318   458  3809   429   184   378
(      149   406   323   142   453   426    76
)      150   405   321   143   456   437    78
,      934  3152  3623  4650  3370  3170  4145
-      546  2340  1206   379  1906  1039   861
.     1811  6515  6922  8821  6220  5411  8880
/       29    38    27     0    31    14     1
:       28   396   632   397   700   551   487
;       43    79   183    47   181   308   217
=        1     0     0     0     2    15     0
?      210   110   464  1276   321   193   980
@        2     1     7     0     4    48     0
`      214   652   734   328   805   741   595
{       38    70    36     8    64    89    15
}       38    70    36     8    64    89    15


N.B.:
In deze tabel zijn de bronvermeldingen niet meegerekend.
Voor de bronvermeldingen is de onderstaande tabel van toepassing.

      camb  cdbl  cgbl  cgtl  cobl  cpwe  crno
      ----  ----  ----  ----  ----  ----  ----
-      489  1619   800   334  1095   492   415
.        0     0    96     4     0    24    16
/        0     0     0    62     0     0     0
```

[3] Aanvullende aantekeningen:

a. Langste tekenreeks

De langste tekenreeks is 89 karakters lang, en staat in Camb, regel 52119.

b. Regelnummering

Over het algemeen lopen de regelnummers regelmatig op.
Er zijn echter vier lege regels verwijderd (Cobl:8432, Cgbl:18735, Crno:27500 en Cpwe:37140), en in een aantal gevallen zijn twee zinnen die op 1 regel stonden gesplitst, zodat sommige regelnummers eindigen op "a" of "b".

c. Spatie en LineFeed

Een spatie en een LineFeed zijn in de huidige vorm van het corpus feitelijk equivalent.
Bijvoorbeeld:
- Als een te lange zin gesplitst is in twee regels, dan eindigt de eerste regel niet op een spatie, en de tweede regel begint er ook niet mee. In dit geval is de LineFeed dus gebruikt als woordscheider.
- In het hele corpus komen geen dubbele spaties voor, maar ook geen opeenvolgingen van spatie-LineFeed of LineFeed-spatie, evenmin als twee LineFeeds achter elkaar.