Verwerking van achttiende-eeuws Nederlands met Frog

Erik Tjong Kim Sang
Meertens Instituut
erikt(at)xs4all.nl

13 februari 2014

1 Inleiding

Frog [1] is een verzameling programma's voor de taalkundige analyse van hedendaagse Nederlandstalige teksten. In het Nederlabproject[1] willen we ook teksten in ouder Nederlands taalkundig analyseren. In dit document gaan we na of Frog achttiende-eeuwse teksten correct kan verwerken. Bij deze test passen we Frog toe op twee oude teksten en een recente tekst.

2 Teksten

We hebben de volgende drie teksten geselecteerd voor onze test:

- bladzijde 2 uit *Poetische en Stigtelyke Mengeldigten*, van Adrianus Hardy, uit 1750. `http://resolver.kb.nl/resolve?urn=dpo:2002:mpeg21:0012`
- bladzijde 2 uit *Romeinsche Geschiedenissen*, deel 16, van M. Stuart uit het jaar 1800. `http://resolver.kb.nl/resolve?urn=dpo:10362:mpeg21:0008`

[1] `www.nederlab.nl`

- bladzijde 2 uit *Benoît, of hoe integreert men het gecrêpepapierde kaasdoosje-voorvaderdag in de loft-met-starck...* van Noortje Wiesbauer uit 1993 (als vergelijkingsmateriaal).
  `http://dbnl.nl/tekst/_bra004199301_01/_bra004199301_01_0064.php` [p. 161]

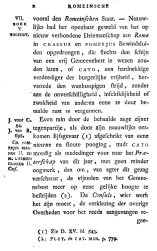
De eerste twee teksten verkregen we via de website Early Dutch Books Online[2]. Van deze twee teksten waren de gescande versies beschikbaar (tekst met scanfouten). De derde tekst verkregen we via de website Digitale Bibliotheek voor de Nederlandse Letteren[3]. Van deze tekst was een PDF-versie beschikbaar, waaruit met knippen en plakken een tekstdocument komt worden afgeleid.

3 Voorbewerking

Voordat de teksten konden worden aangeboden aan het programma Frog, moesten ze eerst worden omgezet in het FoLiA [2], het XML-formaat wordt gebruikt voor de in- en uitvoerbestanden van Frog. Voor de recente tekst volstond het om een extra programma te schrijven dat paragrafen tussen paragraaftags plaatste en een XML-inleiding en -uitleiding toevoegde.

De twee oudere teksten konden niet op dezelfde manier worden omgezet omdat op de bladzijden meer informatie aanwezig was dan de tekst:

- bladzijdenummers
- hoofdstuktitels (herhaald op elke pagina)
- noten (zowel onder als naast de tekst)
- de eerste lettergreep van de volgende pagina



Al deze informatie stond in de tekst tussen de bladzijden. Om het voor Frog mogelijk te maken deze informatie correct te verwerken, zou alles moeten worden gemarkeerd als informatie die door het programma mag worden overgeslagen. Het was echter teveel werk om hiervoor een programma te schrijven. Daarom hebben we deze informatie manueel verwijderd en hebben beide teksten ingekort tot de eerste 20 bladzijden (plus eventuele inleiding). Hierna

[2] `www.earlydutchbooksonline.nl`, thans `www.delpher.nl`

[3] `www.dbnl.org`

konden ze ook worden omgezet in FoLiA met het programma dat was gebruikt voor de recente tekst[4].

4 Verwerking met Frog

Nadat de drie teksten waren omgezet in het formaat FoLiA, konden ze zonder probleem binnen enkele minuten worden verwerkt door Frog. Het programma genereerde drie nieuwe FoLiA-bestanden met daarin de taalkundige analyse van de tekst in elk document. Voor de vergelijking gebruiken we onze eigen FoLiA-browser, een verzameling van Javascriptprogramma's voor visualisatie van de inhoud van FoLiA-bestanden.

5 Scankwaliteit

We begonnen onze analyse met de controle van de kwaliteit van de teksten. De twee oude teksten zijn gedigitaliseerd met behulp van optical character recognition (ocr) en hierdoor zijn sommige letters verkeerd herkend. Van de tweede bladzijde van elk document telden we de woorden en getallen (niet de leestekens), en controleerden we met behulp van de PDF-bestanden op de websites hoeveel daarvan correct waren herkend:

- Hardy (1750): 139 woorden; 2 fouten, 98% correct
- Stuart (1800): 147 woorden; 4 fouten, 97% correct
- Wiesbaden (1993): 289 woorden: 0 fouten, 100% correct

Bij deze controle hebben we herkenning van de lange s (f) als de letter f goedgerekend. Deze verwisseling kwam diverse keren voor in de twee oude documenten en zou met behulp van een woordenlijst gemakkelijk te corrigeren moeten zijn. Voor de gecontroleerde bladzijden vallen de foutaantallen erg mee.

[4]Ko van der Sloot van de Universiteit Tilburg heeft ook een programma geschreven voor het omzetten van EDBO-documenten naar FoLiA-formaat. Het extra materiaal dat in deze sectie wordt gemeld komt bij die documenten ook in de FoLiA-versie terecht.

6 Zinsgrenzen

Zinsgrenzen zijn in de oudere documenten soms lastig te herkennen omdat niet consequent gebruik wordt gemaakt van leestekens:

- Hardy (1750): 9 zinnen; 9 fouten, 0% correct
- Stuart (1800): 4 zinnen; 0 fouten, 100% correct
- Wiesbaden (1993): 14 zinnen: 0 fouten, 100% correct

Het programma heeft alleen bij de tekst van Hary moeite om de zinsgrenzen te vinden. Dit komt doordat deze tekst in dichtvorm is opgezet met gebruik van komma's en inspringen om zingrenzen aan te geven. Frog herkent hier dan ook maar een zinsgrens, op een verkeerde plek door een vraagteken die is ontstaan door een ocr-fout.

7 Lemma's

Frog bepaalt voor elk woord in een tekst het lemma, wat kan worden gezien als de corresponderende woordenboekingang. Voor veel woordvormen is maar een mogelijke woordenboekingang. Voor andere woorden moet een keuze gemaakt worden. Zo zou de woordvorm *vis* zowel kunnen passen bij het gelijkvormige zelfstandige naamwoord als bij het werkwoord *vissen*.

- Hardy (1750): 139 woorden; 38 fouten, 72% correct
- Stuart (1800): 147 woorden; 23 fouten, 84% correct
- Wiesbaden (1993): 289 woorden: 6 fouten, 98% correct

In het evaluatieproces hebben we ocrfouten op basis van de verwisselingen lange s↔f en ij↔y geaccepteerd. Verder hebben we van alle woorden vereist dat zij werden gelinkt aan een correct hedendaags equivalent. Het lemmatisatieproces genereert meer fouten bij het verwerken van oudere teksten. De fouten worden voornamelijk veroorzaakt door vocabulaireverschillen en ocrfouten.

8 Syntactische woordklassen (part-of-speech)

Frog kent ook syntactische klassen als werkwoord en zelfstandig naamwoord toe aan woorden. Bij de evaluatie hebben we alleen naar de hoofdklassen gekeken. Eigenschappen als enkelvoud vs meervoud en tijdsbepalingen hebben we buiten beschouwing gelaten.

- Hardy (1750): 139 woorden; 38 fouten, 72% correct
- Stuart (1800): 147 woorden; 24 fouten, 84% correct
- Wiesbaden (1993): 289 woorden: 16 fouten, 94% correct

De scores zijn vergelijkbaar met die van de lemma-analyse. Part-of-speech tags worden in het hedendaagse werk slechter herkend dan lemma's. De meest voorkomende fout is het benoemen van bijwoorden als bijvoeglijke naamwoorden.

9 Namen (named entities)

Frog bevat ook een module voor naamherkenning. Naast het onderscheid tussen namen en andere woorden, wordt ook de klasse van een herkende naam aangegeven, bijvoorbeeld: persoon, organisatie of locatie. We hebben alleen het onderscheid wel-geen getest:

- Hardy (1750): 4 namen; 2 correct, 14 fouten, 13% correct
- Stuart (1800): 8 woorden; 3 correct, 11 fouten, 21% correct
- Wiesbaden (1993): 4 namen: 4 correct, 0 fouten, 100% correct

De namen in de hedendaagse tekst worden correct herkend. Voor de oudere teksten maakt het systeem meer fouten dan dat het namen correct herkent. Een probleem hierbij is capitalisatie: in de oude teksten is de eerste letter van diverse zelfstandige woorden een hoofdletter. Daarnaast bevatte een tekst (Stuart) namen in smallcaps, wat door de optical character recognition werd omgezet kleine letters en daarna lastig was te herkennen als naam.

10 Conclusies

We hebben twee oude teksten verwerkt met het taalanalyseprogramma Frog en hebben vervolgens de resultaten vergeleken met die van een recente tekst. Zoals verwacht, werden de oude teksten slechter verwerkt dan de recente tekst. De prestaties van het systeem waren het slechtst voor de oudste tekst en voor complexere analyses. Het herkennen van zinsgrenzen ging, afhankelijk van de aangeboden tekst, goed. Herkenning van lemma's, syntactische woordklassen en namen ging beduidend slechter bij de twee oude teksten.

Bij de resultaten moet worden aangetekend worden dat we voor de test niet de meest ingewikkelde bladzijden hebben uitgekozen. Daarnaast hebben we extra materiaal, zoals paginanummers, kopjes en voetnoten, handmatig van de bladzijden verwijderd. Als we deze informatie in de bestanden hadden laten staan dan waren de testresultaten ongetwijfeld slechter geweest. Het opschonen van de teksten is wenselijk maar het is ondoenlijk om dit handmatig te doen voor al ons materiaal en het nog maar de vraag of dit proces automatiseerbaar is.

We kunnen op dit moment de EDBO-documenten verwerken met Frog maar de kwaliteit van de resultaten zal niet erg hoog zijn. Voor een verbetering van de analyseresultaten kunnen de volgende processen proberen te verbeteren:

1. **Tekenherkenning**: verbeterde letterherkenning (optical character recognition) zal waarschijnlijk leiden tot een verbetering van alle analyses van Frog. Het ocr-proces kunnen we niet overdoen maar mogelijk kan postprocessing (bijvoorbeeld met TICCL of een lijst van veelgemaakte fouten) de kwaliteit van de teksten verbeteren.

2. **Markering van buitentekstelijk materiaal**: paginanummers, kopjes en noten moeten apart worden verwerkt door het programma. Als we deze al gemarkeerd zijn of als zij automatisch kunnen worden herkend dan kunnen daardoor de analyses van Frog worden verbeterd.

3. **Toevoeging lexicon ouder Nederlands**: dit zou helpen om van meer woorden het lemma en de woordklasse goed te herkennen. Het INL kan mogelijk geschikte lexica leveren.

4. **Recapitalisatie**: omdat de herkenning van namen sterkt leunt op de aanwezigheid en afwezigheid van hoofdletters, heeft het alternatieve hoofdlettergebruik in oude teksten een negatieve invloed op het herkennen van namen. Vooraf het hoofdlettergebruik standaardiseren zou de naamherkenning verbeteren.

Niet voor alle taken is kant-en-klare software beschikbaar. Verbetering van de resultaten van Frog op oude Nederlandse teksten zal extra werk kosten maar hier is bij de planning van Nederlab rekening mee gehouden.

A Geteste teksten (zoals gescand)

Hardy (1750)

Uw vlugge ftyl , hoewel * vvat kreupel afgefchreyen , Gaf ftraks een ftaaltje van den hersfenryken bol , Ik zag geleertheid en taalkunde daar in leven , Van Godtvrugt en verftandt en leesvrugt even vol : My dagt , ik zag U naar den groten Tempel flappen , Vol geest en vuur , verzelt van ? t agtbaar Priesterdom * t Geen U geleidde naar dehooge kansfel trappen , Den aandagt ftelen van den Godtgewyden drom : My dagt , ik hoorde daar de blyde Maagden reij en In 't feestkleed uitgedost , voor uwe voeten neer Gezeten , vol van vreugd haar heilig danklied fpreijen , Ten prys van U , maar ook vooral van Uwen Heer , 'k Meen ' Jezus , die Uw hart door Zynen gloed deet branden , Uw tong ontftekende met heilig autervuur Niet meer van ftierenvleesch ; maar beetere offerhanden , In vollen vlam gezet , niet binnen Zalems muur , Maar in de vrugtbaare en genaderyke ftreeken Van Neetlands Zion , daar de vrede Koning woont ,

Stuart (1800)

vooral den Romeinfchen Staat . Naauwlijks had het openbaar geweld van het op nieuw verbondene Driemanfchap aan Rome in crassus en pompejus Bewindsluiden opgedrongen , die Hechts den fchijrt van een vrij Gemeenebest in wezen zouden laten , of cato , een hardnekkige verdeediger der burgerlijke vrijheid , hervattede den wanhoopigen ftrijd , zonder aan de onverfchilligheid , infchiklijkheid of zwakheid van anderen eene te duur betaalde rust te vergunnen . , Even min door de behaalde zege zijner tegenpartije , als door zijn naauwlijks ontkomen lijfsgevaar (1) afgefchrikt van eene "nieuwe en ftoute pooging , tradt cato) moedig als mededinger voor naar het Pree [torfchap van dit jaar , met geen minder oogmerk , dan om , van agter dit gezag verfchanst , de vijanden van het Gemeenebest meer op eene gelijke hoogte te beftrijden (2) . De Confuls , wier werk het zijn moest , de verkiezing der overige Overheden voor het reeds aangevangen re -

Wiesbaden (1993)

X-en had weer zo'n origineel thema ! Samengevat zou men kunnen stellen dat ze a.h.w. scrabbelen met hetzelfde alfabet en dezelfde woord - en letterwaarden ....

SPELEN betekent oorspronkelijk 'zich continu bewegen , zich vlot kunnen bewegen . Het spel wordt daartoe beschermd door spelregels . Speelt men , dan moet men dus ernstig spelen . Overtredingen gelden enkel binnen het spel . Maar wie weigert het spel mee te spelen , bekritiseert niet zozeer de manier waarop het spel gespeeld wordt , maar het spel zelf , en de overtuiging , het geloof en de wil dat er de basis van vormt . En dat is voor de spelers die het paradoxaliter ernstig menen , onvergeeflijk , en vervult de sfeer met ongemak , wrevel tot agressie . Om een 'verbroken spel te vermijden , geldt dus de eis van de algemene instemming , zeg maar ' conformisme .

Als men dus zo opgaat in het spel , dat het werkelijkheid wordt , dan ' speelt men niet meer in de eigenlijke betekenis van het woord . Misschien bestaat de inwijding in l'art de vivre dan ook in het vlotten van de evenwichtsoefeningen tussen spelen en leven , die zich met al onze menselijke aktiviteiten vermengd hebben : het spel van de lektuur , van het gasten-ontvangen , van het reizen , van het zakendoen , ...

Enerzijds ligt het belang van het spel(en) in de mogelijkheid te ontsnappen aan de vernauwende wereld van het zakelijke , van de orde van de noodzakelijkheden , de mens te bevrijden van het determinisme , om zo meer en meer zichzelf te vinden .

Anderzijds , zoals verder nog ter sprake zal komen , ontstaan via het spelen nieuwe determinanten , wordt de ' onvrijheid in de hand gewerkt , en verliest de mens uiteindelijk juist zijn kans op persoonlijkheid .

Zo kan men zich inbeelden dat Benoîts personages sculpturen als buren hebben , Van Beirendoncks creaties appreciëren maar liever Armani dragen , Wittamer de ontdekking

Referenties

[1] A. Van den Bosch, G.J. Busser, W. Daelemans, and S. Canisius. An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pages 99–114. Leuven, Belgium, 2007.

[2] M. van Gompel and M. Reynaert. FoLiA: A practical XML format for linguistic annotation - a descriptive and comparative study. *CLIN Journal*, 3:63–81, 2013.