Kwantitatieve Analyse
Oeuvre Hugo Claus

Guy De Pauw
CLiPS - Universiteit Antwerpen

17 juni 2010

1 Methodologie

We onderzoeken de onderstaande werken, ingedeeld in vier fasen: begin, midden, einde en controle.

Jaar	Titel	Bron
Begin		
1950	De Metsiers	PDF
1952	De hondsdagen	PDF
1956	De koele minnaar	PDF
1962	De verwondering	PDF
1963	Omtrent Deedee	PDF
Midden		
1972	Het jaar van de kreeft	PDF
1972	Schaamte	PDF
1977	Jessica!	PDF
1978	Het verlangen	PDF
1983	Het verdriet van België	PDF
Einde		
1988	Een zachte vernieling	PDF
1989	De zwaardvis	PDF
1994	Belladonna	PDF
1996	De geruchten	PDF
1998	Onvoltooid verleden	PDF
Controle		
2000	Een slaapwandeling	OCR

Elke tekst werd linguïstisch geanalyseerd met behulp van het Tadpole[1] pakket. Deze software doet aan *tokenization* (oa het bepalen van zinsgrenzen, leestekens afsplitsen) en linguïstische analyse. Een voorbeeld van de uitvoer:

ID	woord	lemma	morf	woordsoort	hoofd	label
1	’	’	[’]	LET()	0	ROOT
2	Agnes	Agnes	[Agnes]	SPEC(...)	5	su
3	,	,	[,]	LET()	2	punct
4	je	je	[je]	VNW(...)	5	su
5	mag	mogen	[mag]	WW(...)	0	ROOT
6	drie	drie	[drie]	TW(...)	7	det
7	keer	keer	[keer]	N(...)	8	obj1
8	raden	raden	[raad][en]	WW(...)	5	vc
9	,	,	[,]	LET()	8	punct
10	’	’	[’]	LET()	8	punct
11	zeg	zeggen	[zeg]	WW(...)	0	ROOT

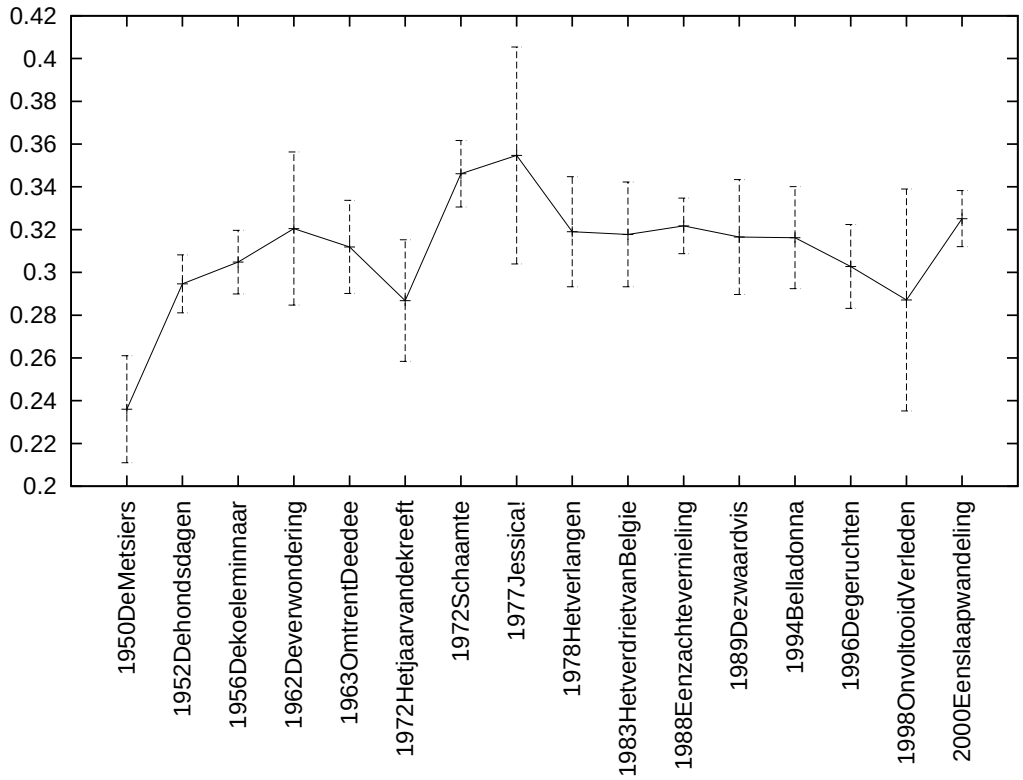
De kolom **woord** bevat het woord zoals het wordt aangetroffen in de brontekst. Vervolgens vinden we het **lemma** van het woord, een rudimentaire **morf**ologische analyse en de **woordsoort**. De kolom **hoofd** geeft aan welk **ID** in de zin het syntactische hoofd van het woord is, terwijl **label** de aard van de syntactische relatie uitdrukt. Op basis van de kolommen **hoofd** en **label** is het mogelijk een dependentie-analyse van de zin te construeren (zie ook p. 51).

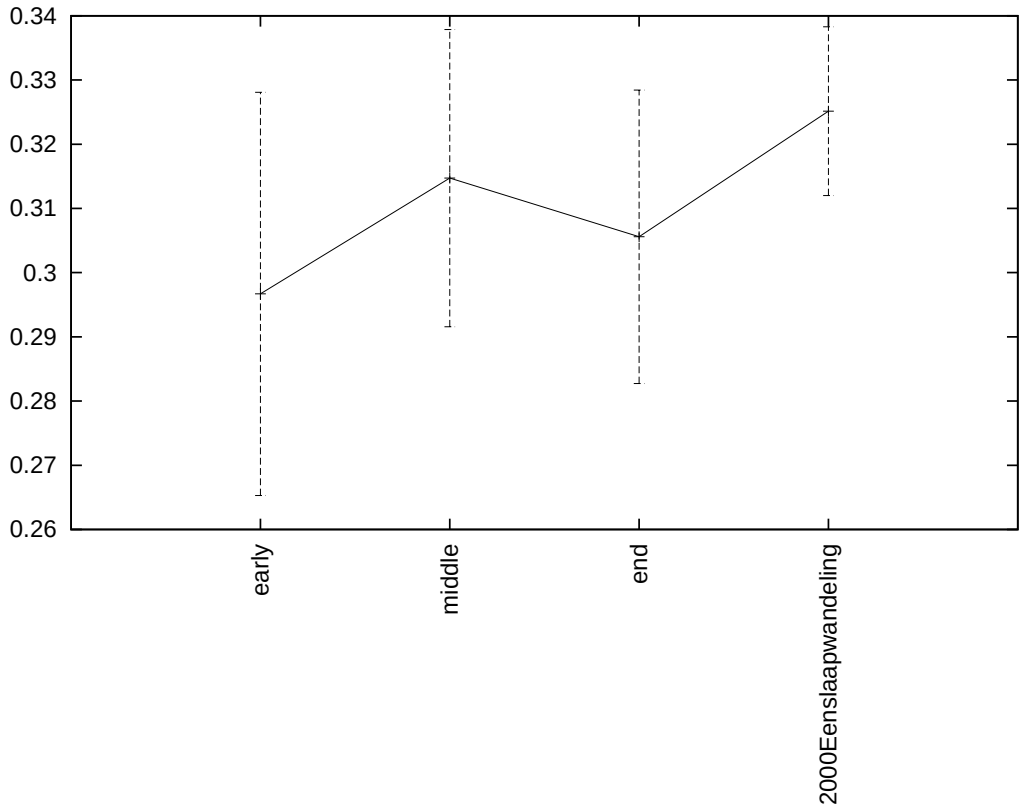
Voor de meeste berekeningen wordt elke tekst onderverdeeld in stukjes van 4000 woorden. Zo worden er voor *Het Verdriet van België* bijvoorbeeld 75 stukjes gemaakt en wordt de berekening voor elk stukje apart gemaakt. Vervolgens wordt het gemiddelde (en standaarddeviatie) van die 75 berekeningen beschouwd als het uiteindelijke resultaat. Deze methode probeert ervoor te zorgen dat de absolute lengte van een werk wordt uitgeschakeld als factor bij de berekeningen.

Op de volgende pagina’s vind je voor elke berekening twee grafieken: een gedetailleerde grafiek per werk en een grafiek die de werken onderverdeeld in drie grotere periodes.

[1] http://ilk.uvt.nl/tadpole

2 Type/Token Ratio: Grafieken





Type/Token Ratio: Uitleg

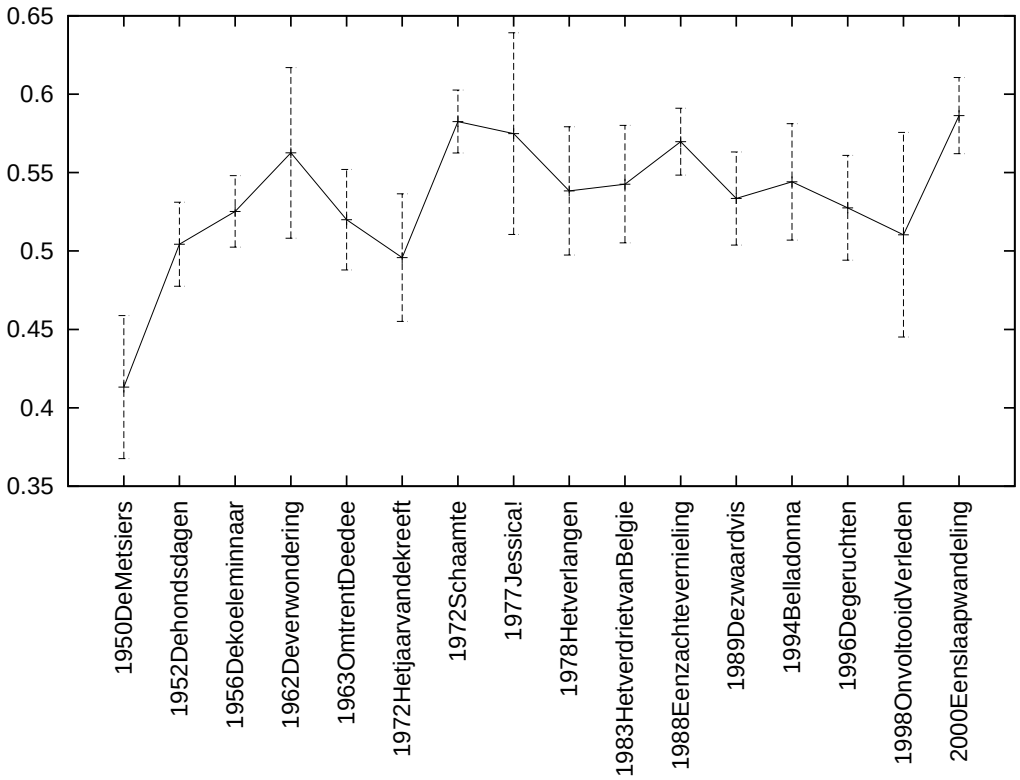
De *type/token ratio* meet de lexicale variatie binnen een document. In deze metingen is een *token* een woordvorm, een *type* het lemma voor die woordvorm. Een voorbeeld:

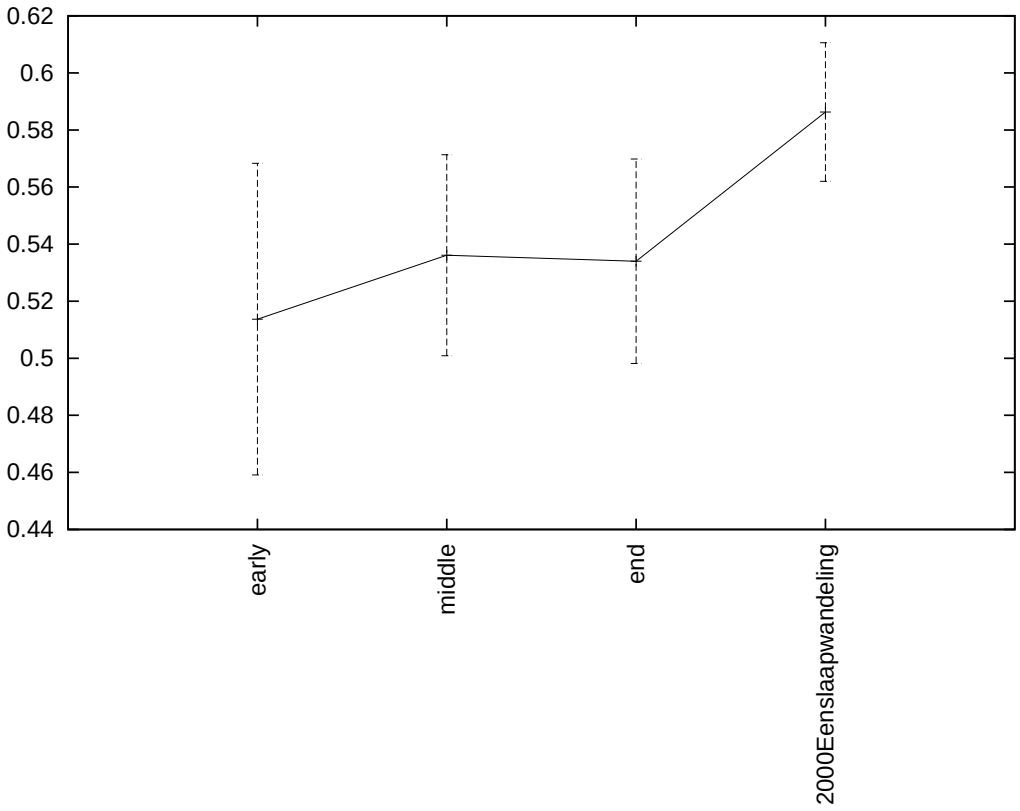
Hij heeft het gehad

Er zijn vier tokens in deze zin, maar er zijn slechts drie verschillende types (*hij hebben* (2x) *het*). De type/token ratio (TTR) voor deze zin is dan 3/4, ofwel 0.75.

Hoe hoger de TTR, hoe groter de lexicale variëteit van de tekst, of - kort door de bocht gezegd - hoe rijker de woordenschat van de auteur. De effecten van Alzheimer zouden moeten leiden tot een verarmde woordenschat, ttz een lagere type/token ratio. De grafieken tonen aan dat dit effect niet merkbaar is voor *Een Slaapwandeling* en dat eerder een tegenovergestelde tendens kan worden vastgesteld.

3 Type/Token Ratio (inhoudswoorden): Grafieken

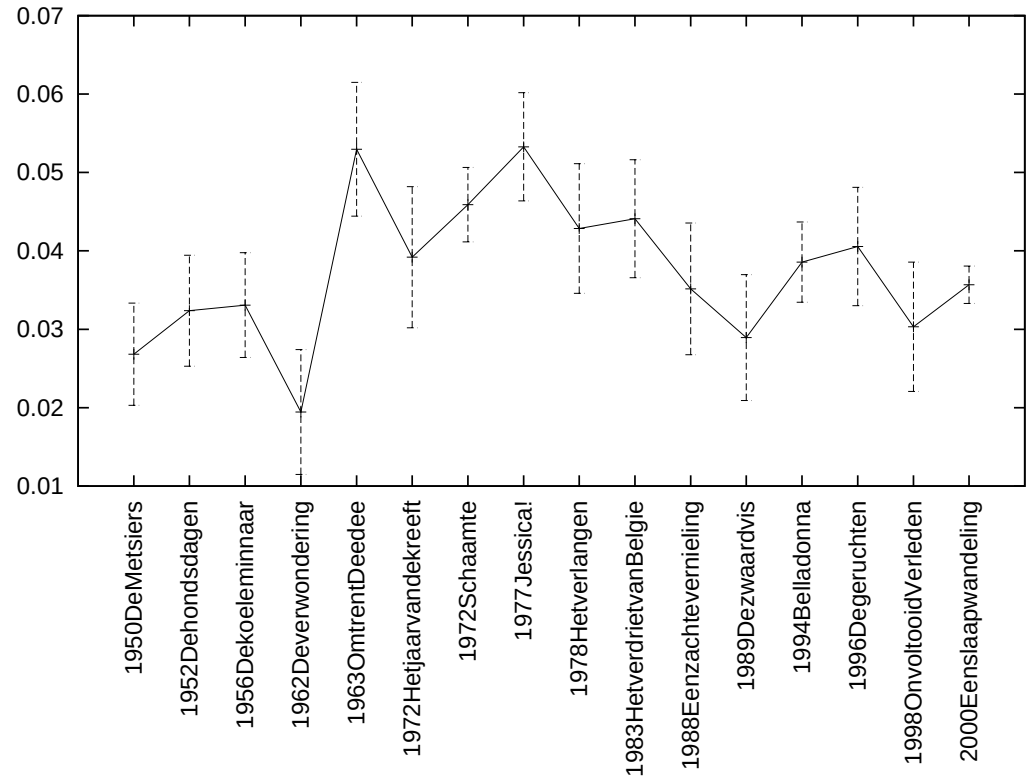


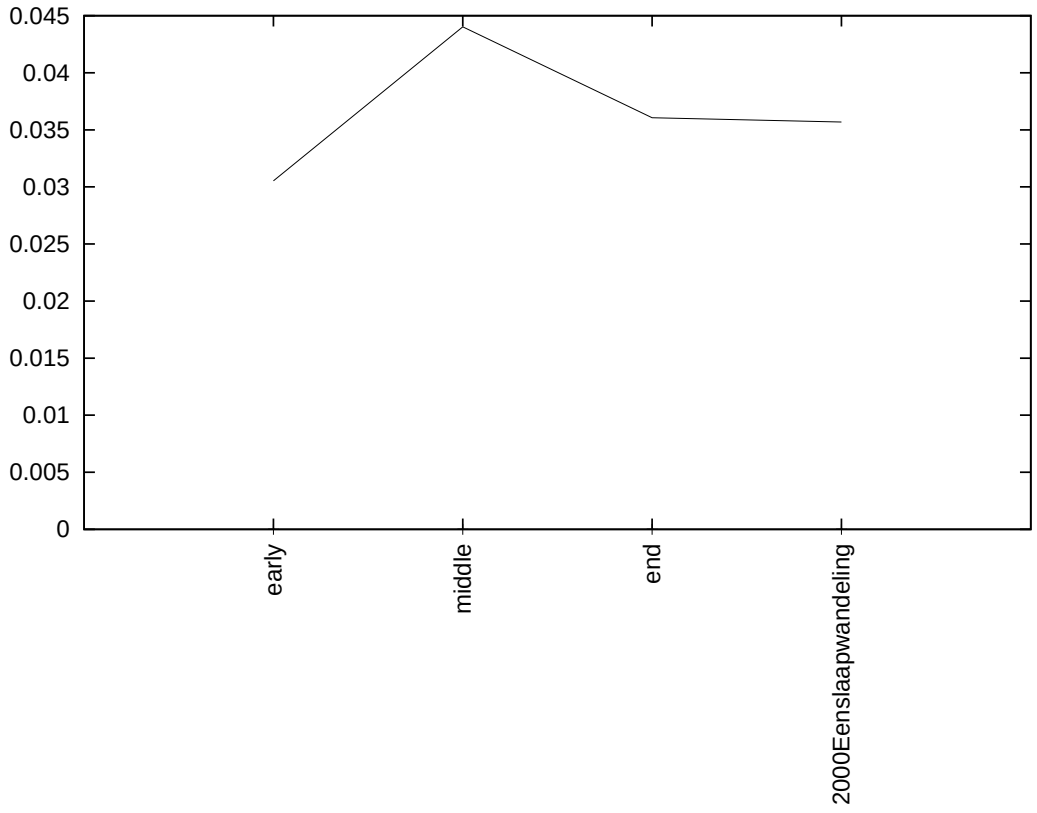


Type/Token Ratio (inhoudswoorden): Uitleg

Hier wordt TTR berekend uitsluitend op basis van inhoudswoorden. Lidwoorden, voegwoorden, voornaamwoorden en voorzetsels worden met andere woorden niet meegerekend. Dit levert globaal gezien hogere TTR waarden op, maar de tendens blijft dezelfde.

4 Percentage eigennamen: Grafieken

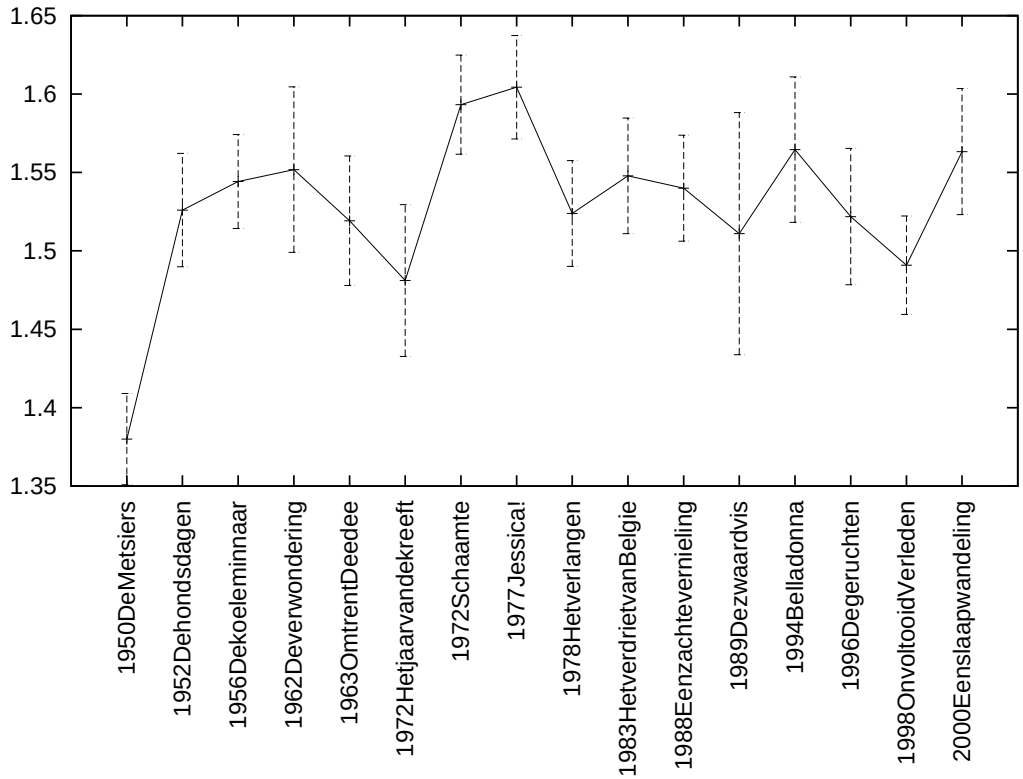


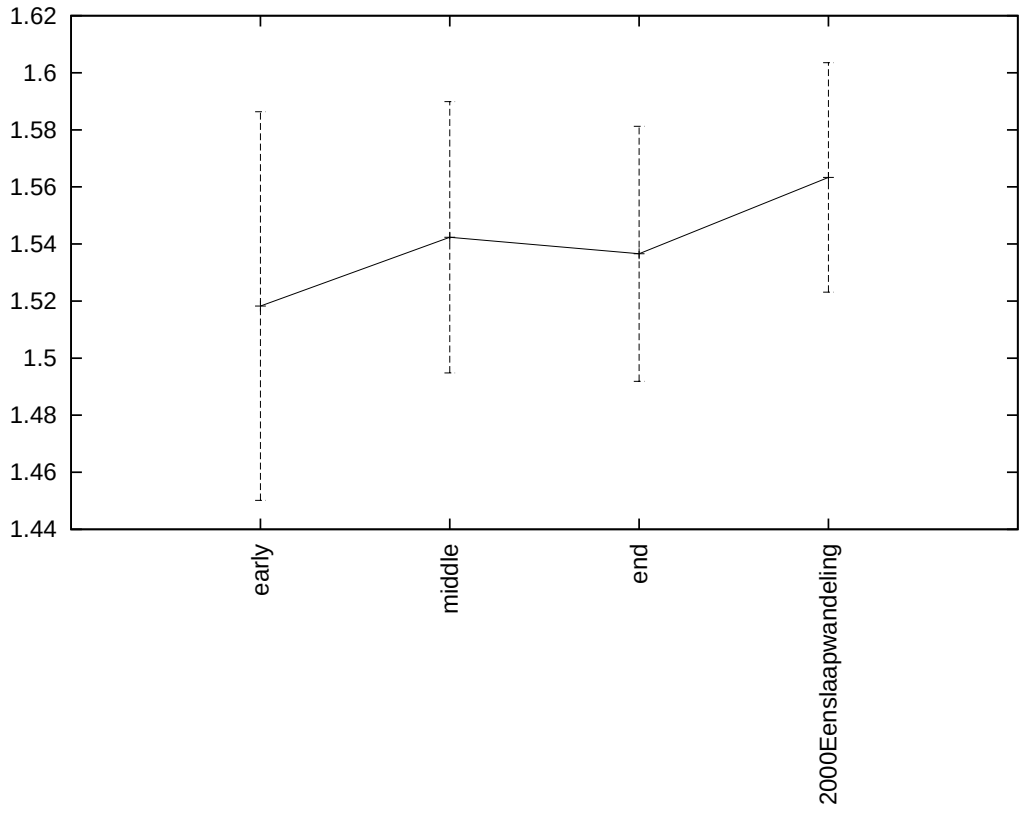


Percentage eigennamen: Uitleg

Hier wordt het percentage van eigennamen berekend. Verwacht wordt dat Alzheimer patiënten meer moeite hebben met het herinneren en (re)produceren van eigennamen en het gebruik hiervan zullen vermijden. Ook voor deze berekening vinden we geen significante effecten terug in *Een Slaapwandeling*.

5 Gemiddelde woordlengte: Grafieken

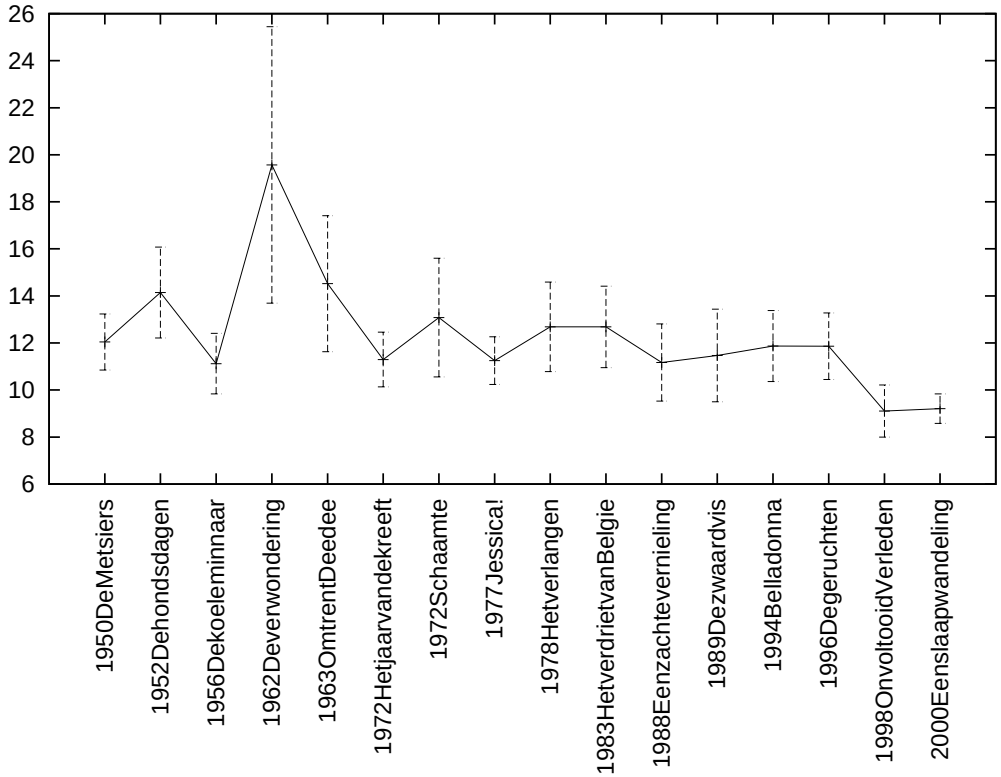


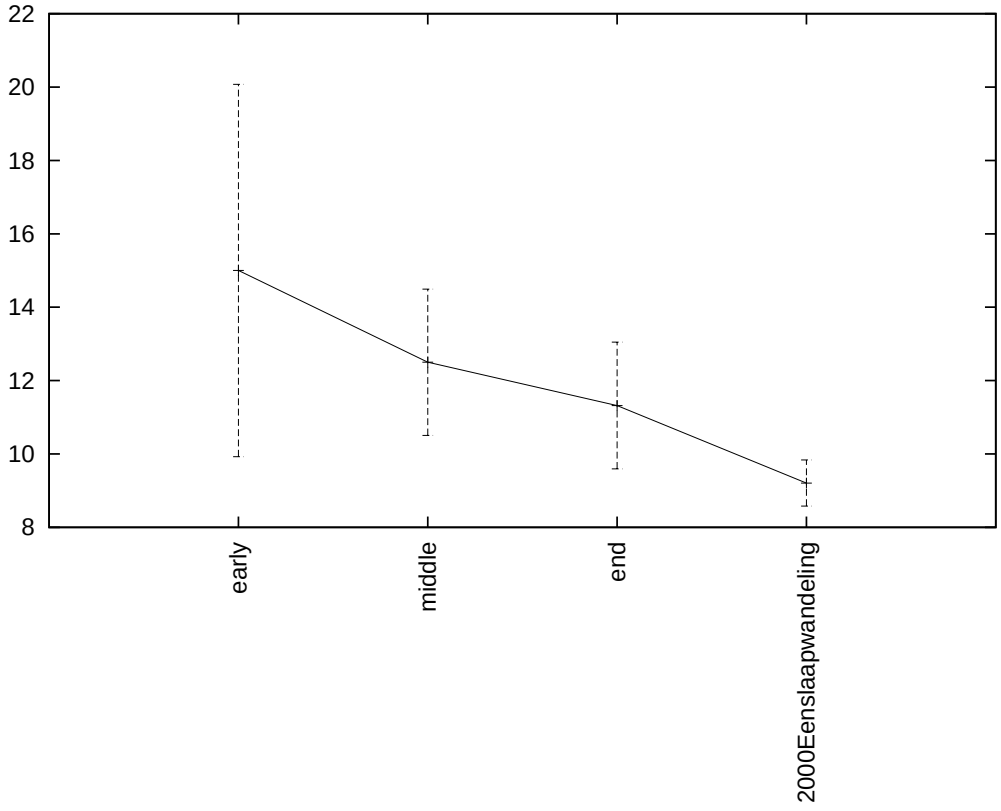


Gemiddelde woordlengte: Uitleg

De gemiddelde woordlengte wordt berekend aan de hand van het aantal syllaben per woord. Een vereenvoudigd taalgebruik zou moeten leiden tot het gebruik van over het algemeen minder lange woorden. Ook in deze berekening geldt eerder het tegendeel voor *Een Slaapwandeling*. Een uitschieter hier is wel *De Metsiers*.

6 Gemiddelde zinslengte: Grafieken

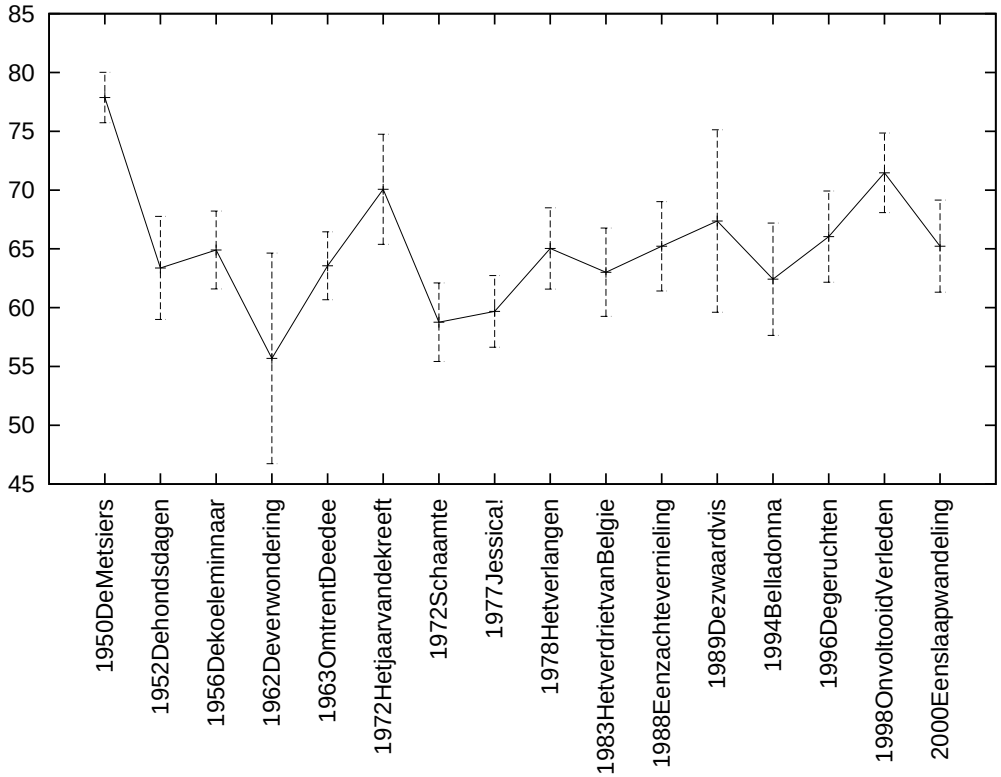


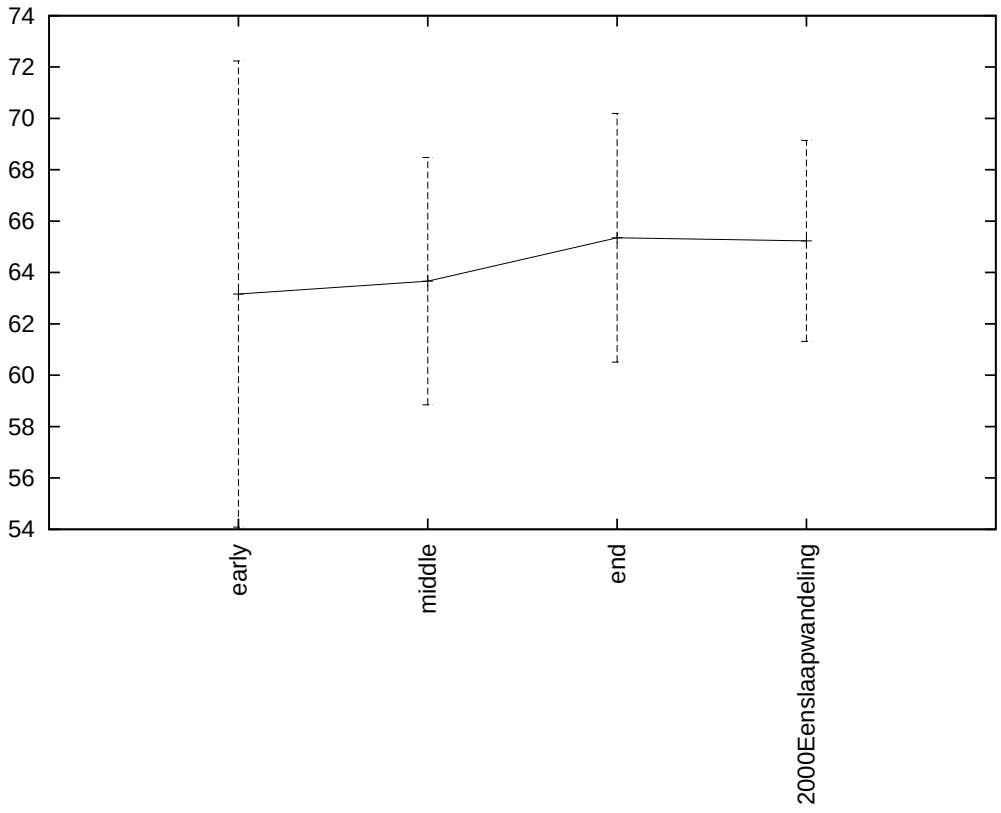


Gemiddelde zinslengte: Uitleg

Zinslengte wordt eenvoudigweg berekend aan de hand van het aantal woorden per zin. In de onderste grafiek zien we een duidelijke tendens naar kortere zinnen doorheen het oeuvre. Hoewel deze tendens het verwachte dieptepunt bereikt in *Een Slaapwandeling*, toont de bovenste grafiek aan dat *Een Slaapwandeling* een gelijkaardige gemiddelde zinslengte heeft als *Onvoltooid Verleden* uit 1998. Een opvallende uitschieter qua zinslengte is *De Verwondering*.

7 Flesch Reading Ease: Grafieken





Flesch Reading Ease: Uitleg

De Flesch Reading Ease score drukt de *leesbaarheid* van een tekst uit. De score wordt berekend op basis van de gemiddelde zins- en woordlengte met behulp van de volgende formule:

206.835 - (1.015 * gemiddelde zinslengte) - (84.6 * gemiddelde woordlengte)

De resulterende score kan je op de volgende manier interpreteren[2]:

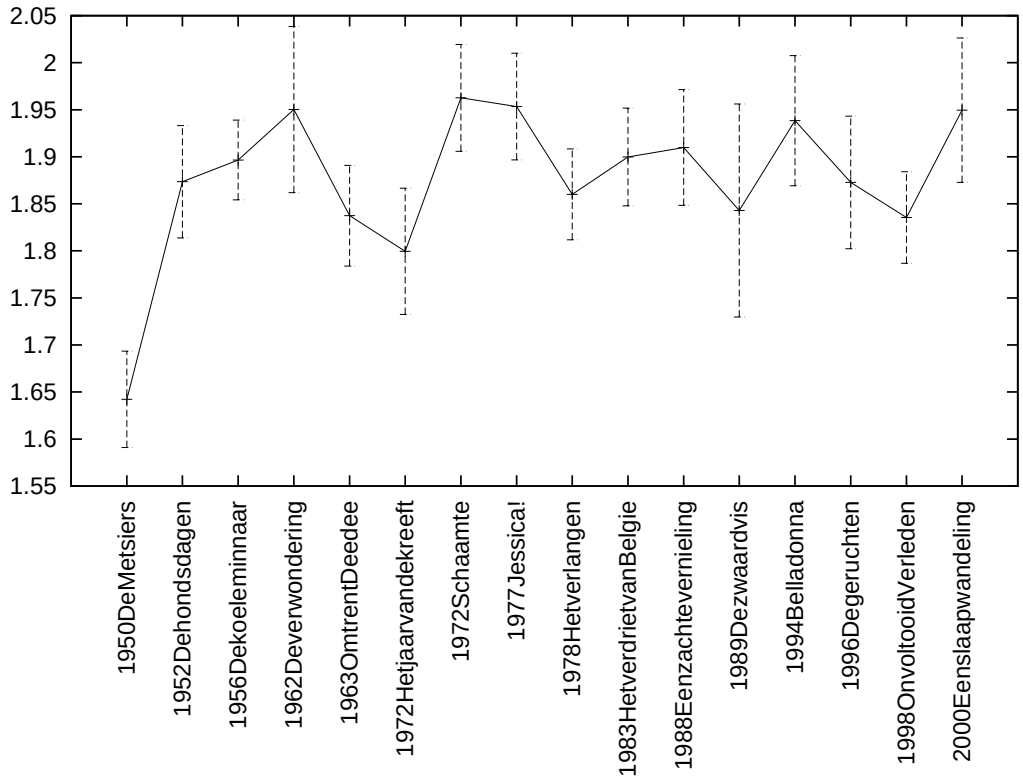
90.0	-	**100.0**	easily understandable by an average 11-year-old student
60.0	-	**70.0**	easily understandable by 13- to 15-year-old students
0.0	-	**30.0**	best understood by university graduates

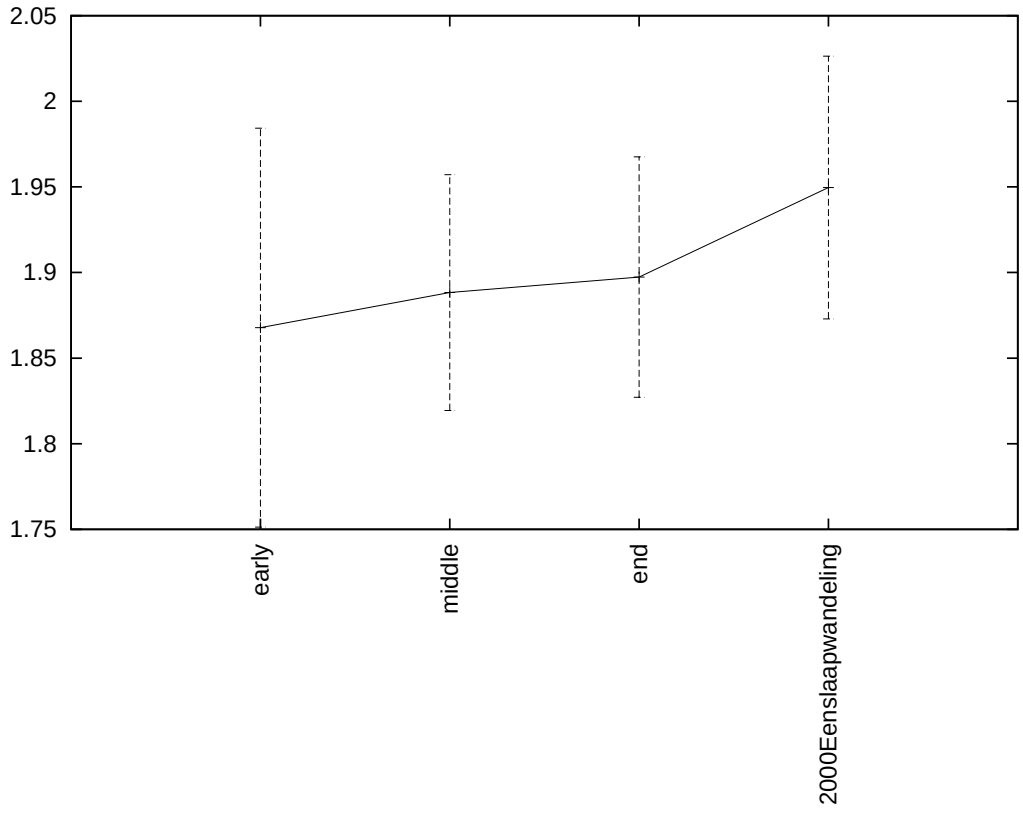
Met andere woorden: hoe hoger de score, hoe *leesbaarder*, *gemakkelijker* de tekst. Let wel dat de formule en de interpretatie van het resultaat ontwikkeld zijn op basis van het Engels en in absolute termen niet noodzakelijk toepasselijk zijn op het Nederlands. Met deze score kunnen de werken onderling wel met elkaar vergeleken worden in termen van leesbaarheid.

Over het algemeen blijft de leesbaarheid constant doorheen het oeuvre met *De Metsiers* en *De Verwondering* als uitschieters.

[2] http://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_test

8 Gemiddelde inhoudswoordlengte: Grafieken

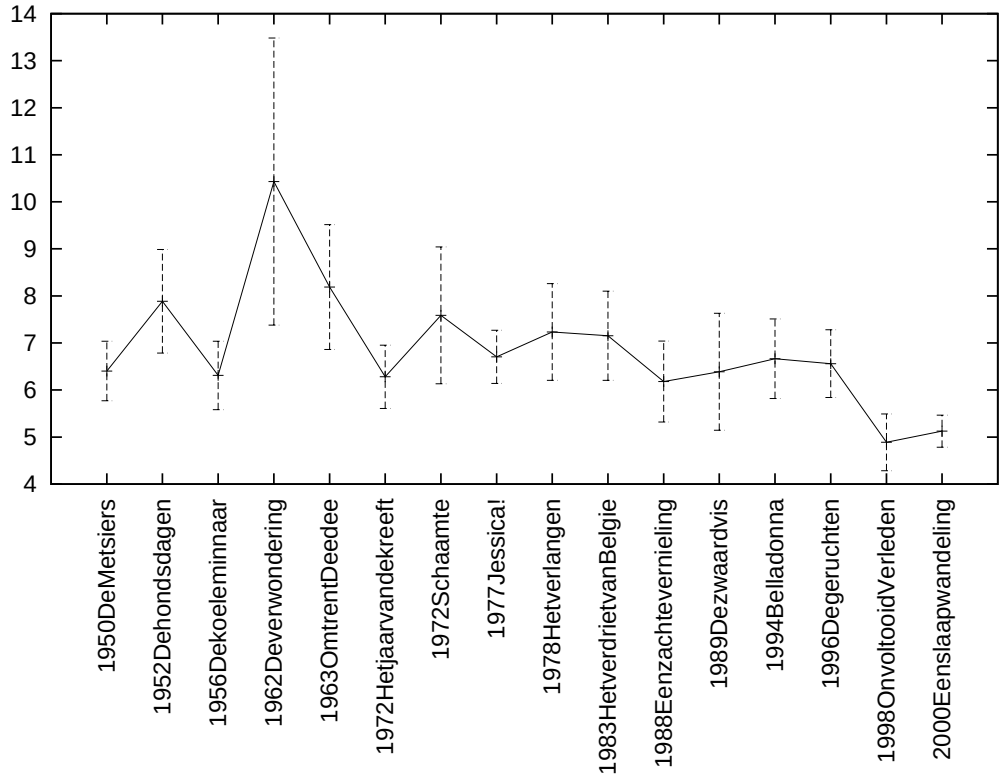


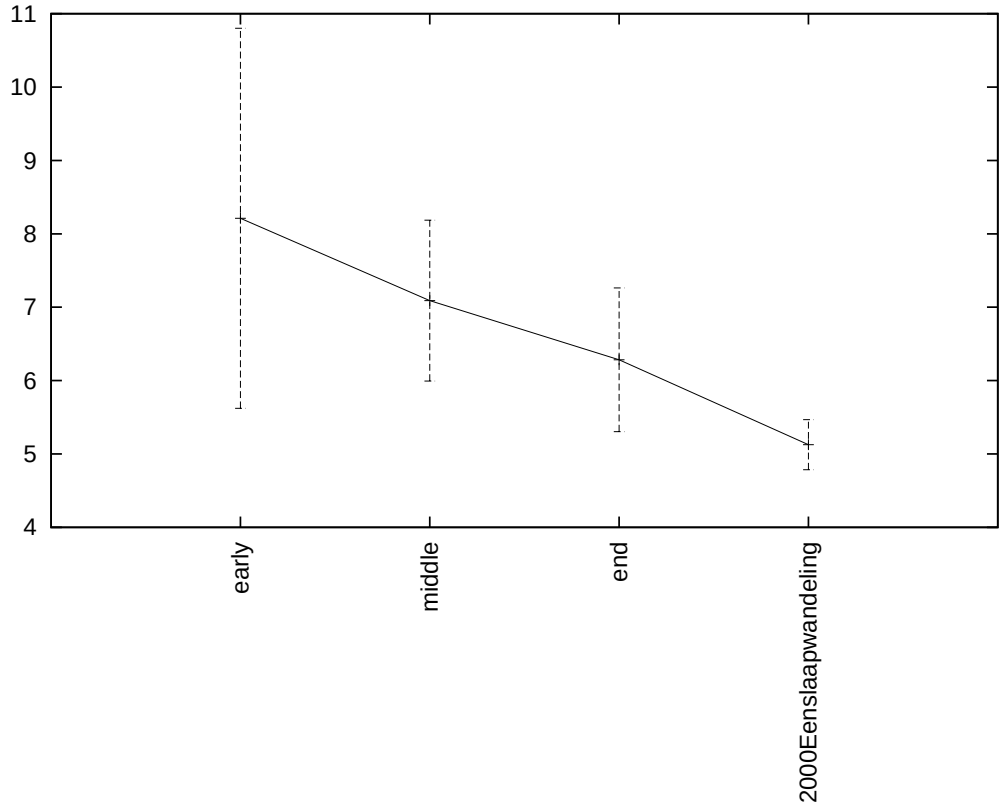


Gemiddelde inhoudswoordlengte: Uitleg

De lengte van functiewoorden heeft men per definitie niet onder controle. Daarom kan het handig zijn om de gemiddelde lengte te berekenen uitsluitend op basis van inhoudswoorden. Deze berekening toont echter dezelfde tendens.

9 Gemiddeld aantal inhoudswoorden per zin: Grafieken

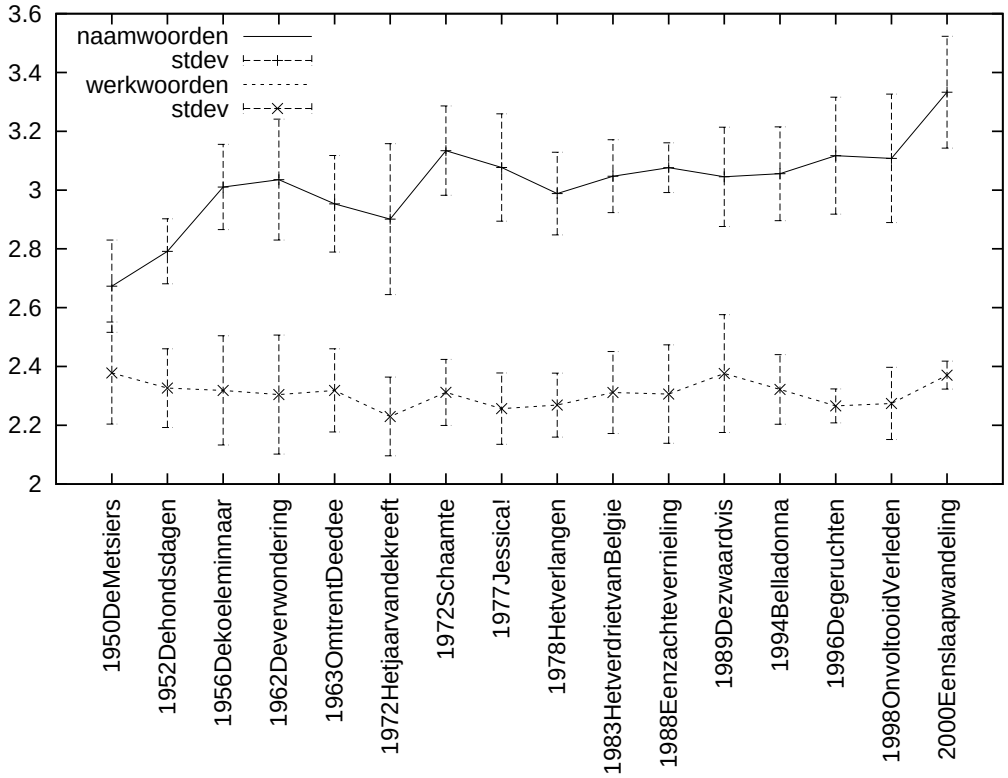


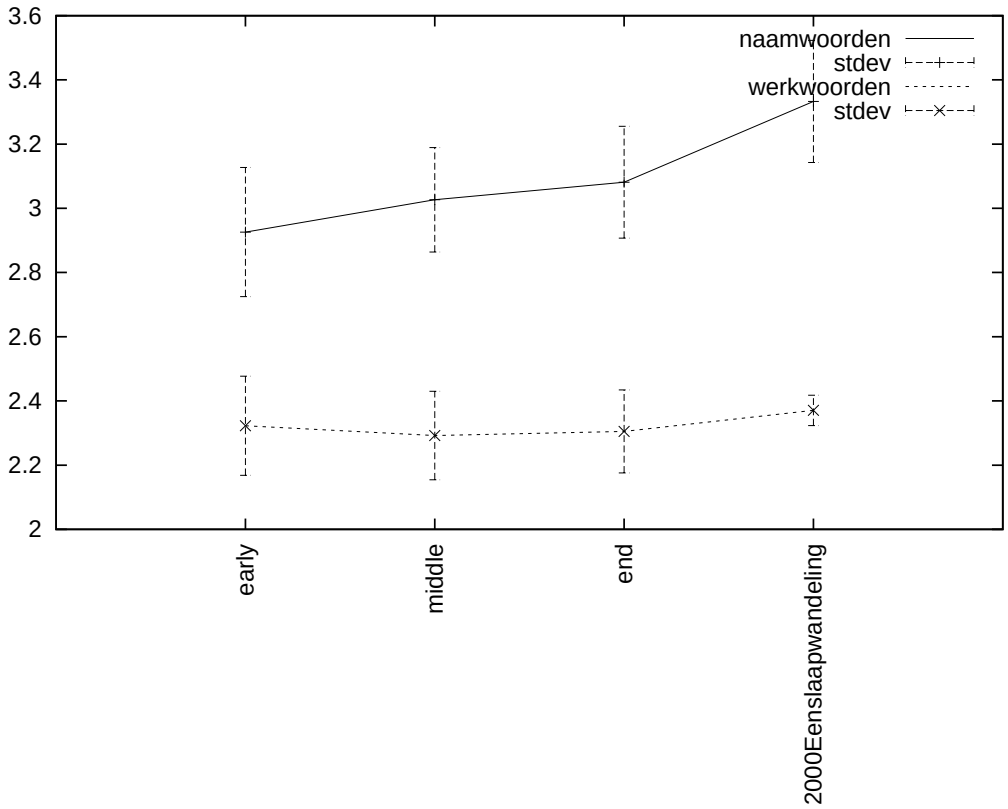


Gemiddeld aantal inhoudswoorden per zin: Uitleg

Zinslengte berekend uitsluitend op basis van inhoudswoorden. Gelijkaardige tendens als in de berekening op alle types van woorden.

10 Specificiteit van woorden: Grafieken





Specificiteit van woorden: Uitleg

Alzheimer patiënten zouden geneigd zijn meer algemene woorden te gebruiken en dus specifiek woordgebruik te vermijden. Om de specificiteit van de woordkeuze te berekenen maken we gebruik van *Wordnet*, een hiërarchisch georganiseerd lexicon. Wordnet neemt de vorm aan van een netwerk, waar woorden met elkaar in verbinding staan. Zo worden onder andere synoniemen en antoniemen aangeduid binnen Wordnet. Voor deze berekening zijn we enkel geïnteresseerd in de concepten *hyponiem*[3] en *hyperoniem*[4].

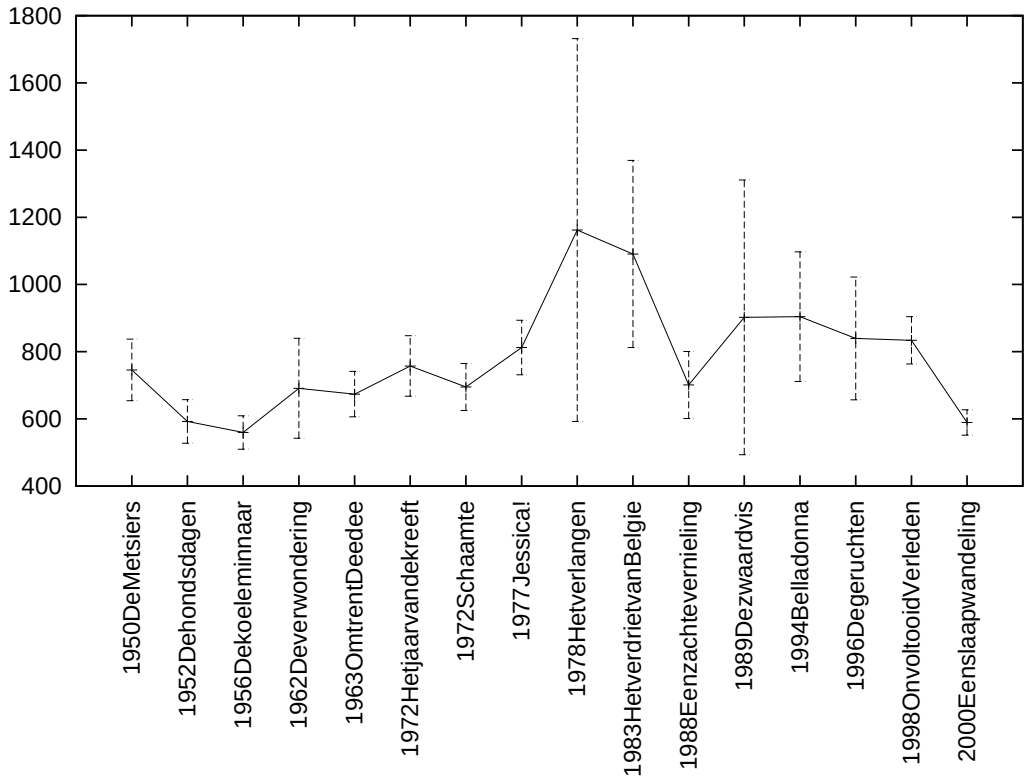
Het woord *takenpakket* is bijvoorbeeld een hyponiem van *pakket*. *pakket* is een hyponiem van *groep*; *groep* is een hyponiem van *iets*. *iets* heeft geen hyperoniem. Voor elk lemma in de tekst dat in Wordnet voorkomt, berekenen we het kortste pad naar een woord dat geen hyperoniem heeft. Hoe langer dat pad, hoe specifieker het woord in kwestie.

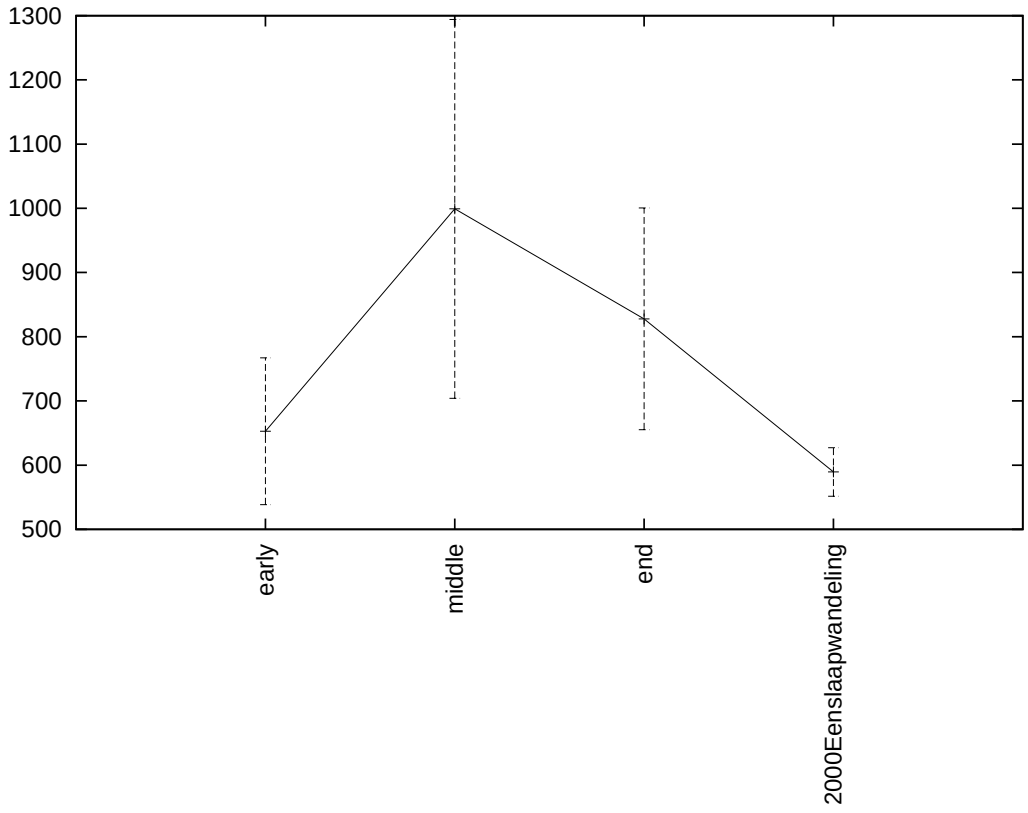
Dit kunnen we niet vaststellen voor *Een Slaapwandeling*, noch voor naamwoorden, noch voor werkwoorden. Alweer lijkt de tendens eerder tegengesteld.

[3] *is een hyponiem van* kan geparafraseerd worden als *is een specifiek soort van*

[4] *is een hyperoniem van* kan geparafraseerd worden als *is een meer algemene vorm van*

11 Frequentie in corpus: Grafieken





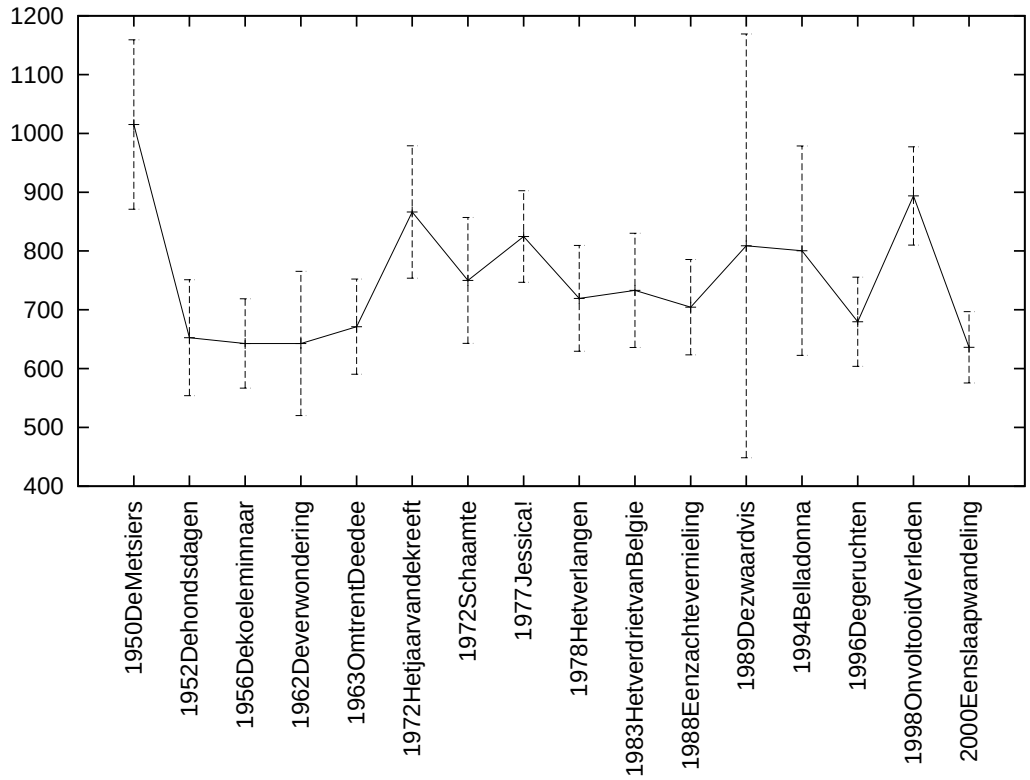
Frequentie in corpus: Uitleg

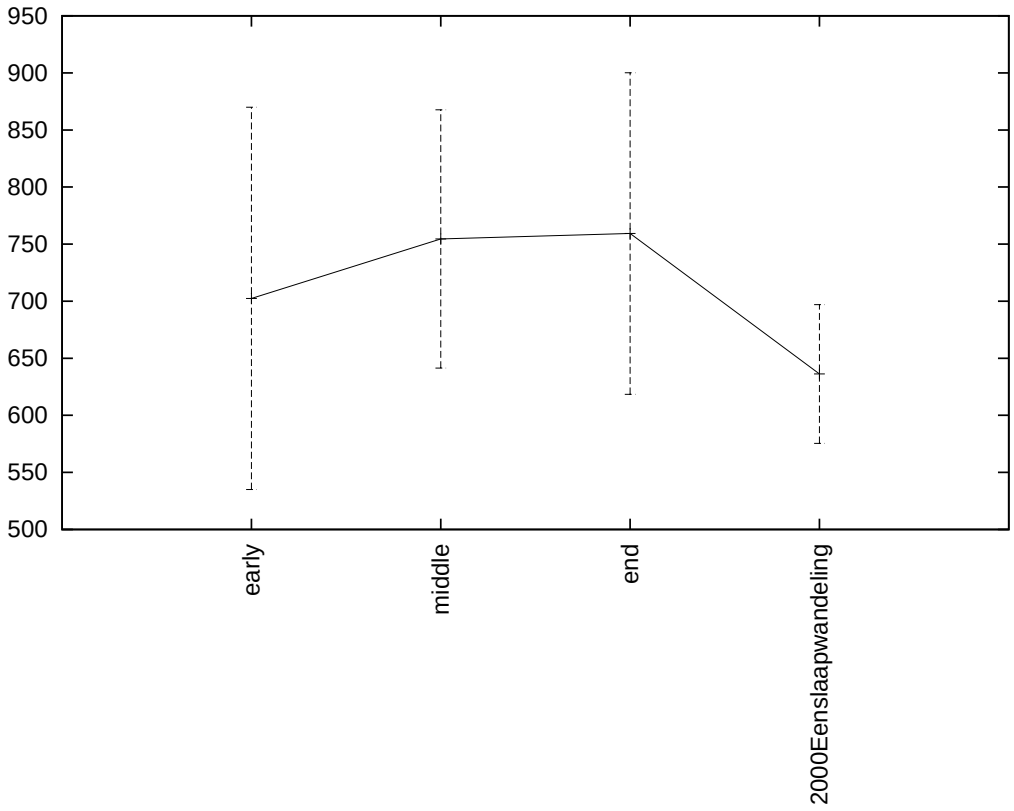
Deze berekening zoekt voor elk woord de frequentie op in de frequentielijst samengesteld op basis van het CGN[5]. Hoe hoger de waarde, hoe meer frequent het woord. Of met andere woorden: hoe lager de waarde, hoe *ongebruikelijker* het woord. De waarden zelf zijn niet informatief in absolute termen. We beschouwen enkel de relatieve verhouding tussen de werken/perioden onderling.

Van Alzheimer patiënten wordt verwacht dat ze geneigd zijn om meer voor de hand liggende woorden te gebruiken. Ook hier zien we een omgekeerde trend. Een mogelijke verklaring hiervoor zou een verhoogd percentage aan out-of-vocabulary woorden kunnen zijn, maar de metingen hieromtrent tonen aan dat *Een Slaapwandeling* hier niet significant afwijkt van de rest van de teksten (zie p. 27).

[5] Corpus Gesproken Nederlands

12 Frequentie in corpus (inhoudswoorden): Grafieken

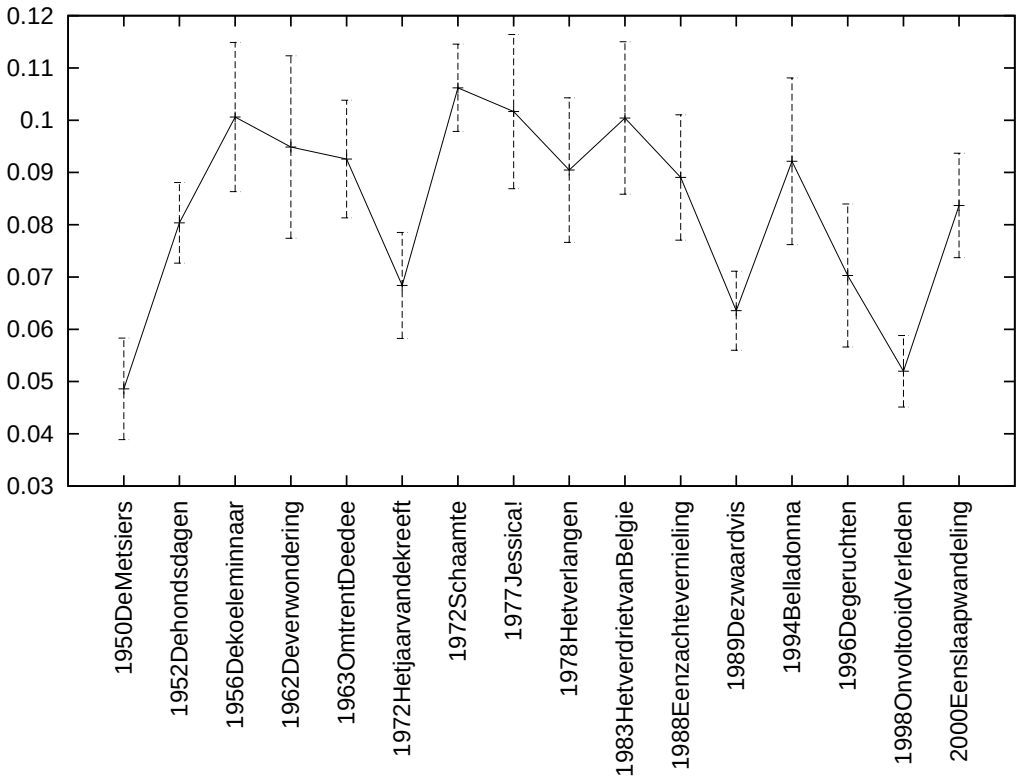


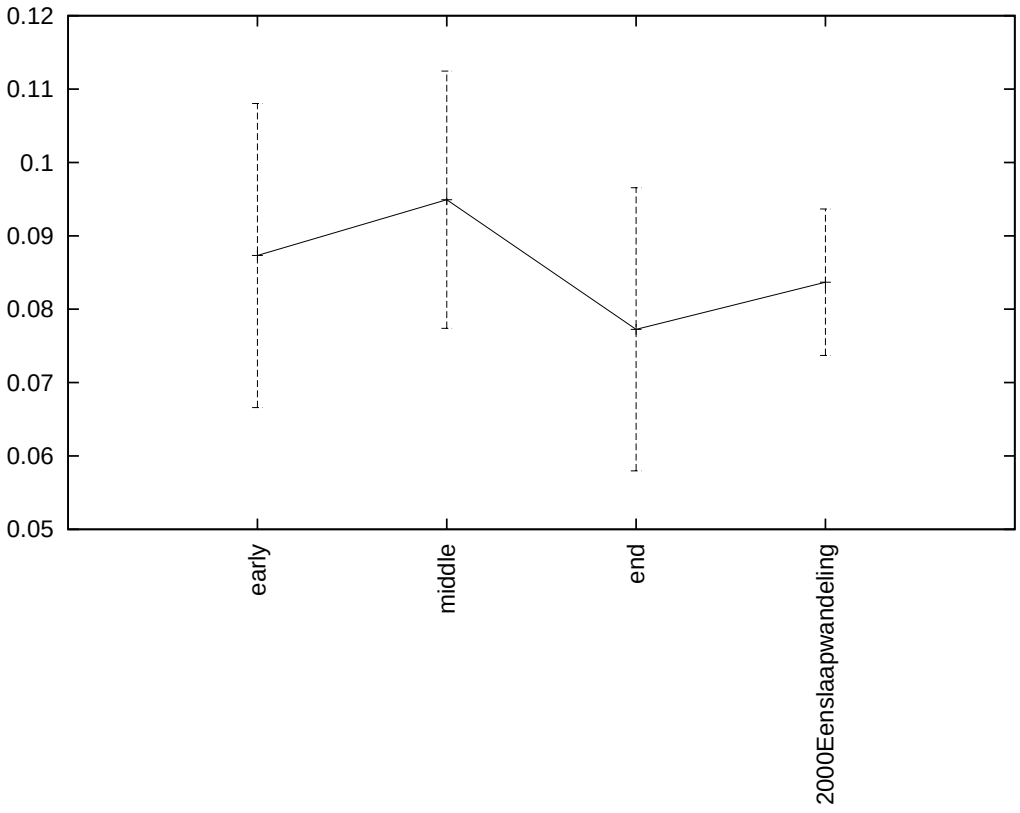


Frequentie in corpus (inhoudswoorden): Uitleg

De gemiddelde frequentie van inhoudswoorden. Grosso modo dezelfde tendenzen zijn merkbaar.

13 Percentage van Out-of-Vocabulary woorden: Grafieken

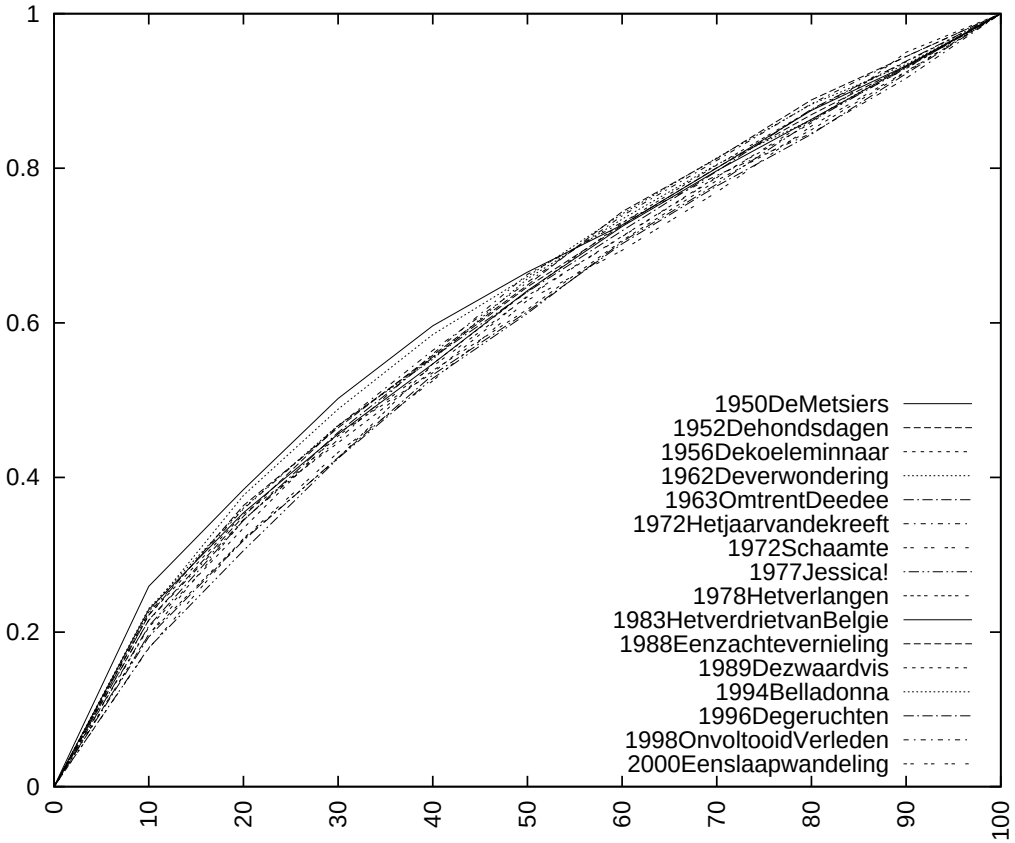


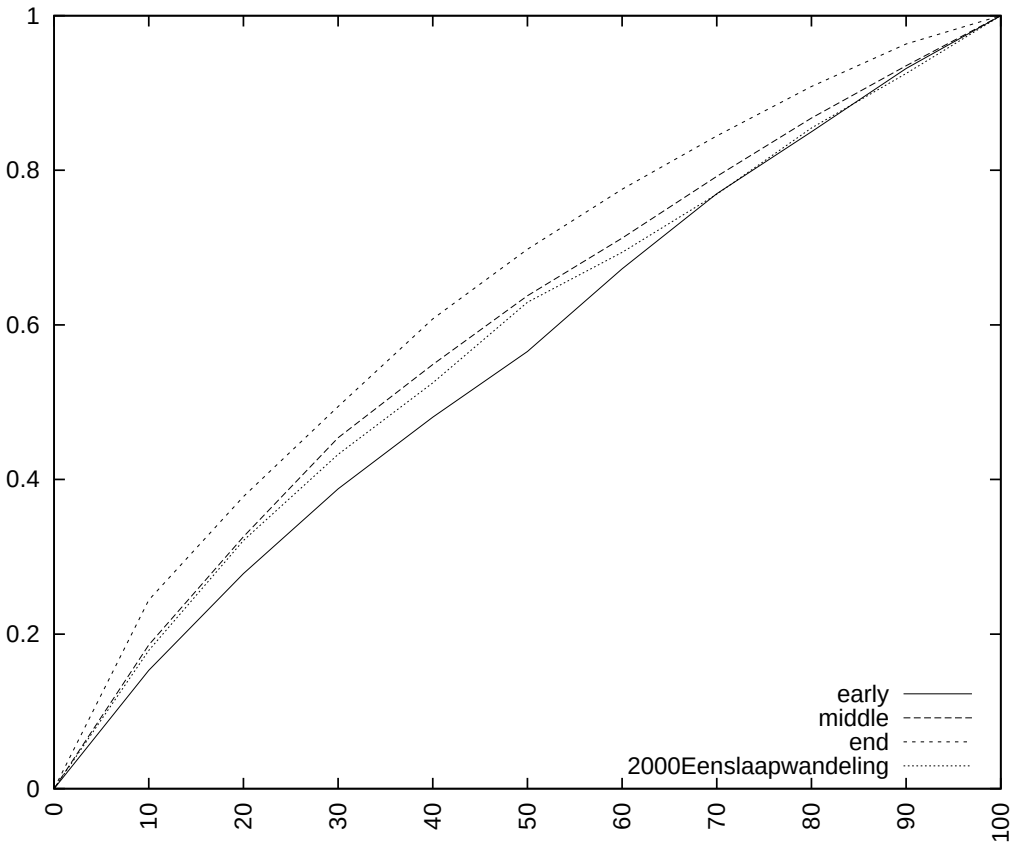


Percentage van Out-of-Vocabulary woorden: Uitleg

Het percentage van woorden in de tekst dat niet voorkomt in de frequentielijst (en met andere woorden zeer laag frequente woorden of potentiële nieuwvormen zijn). Dit percentage werd berekend uitsluitend op basis van inhoudswoorden. Hoe hoger dit percentage, hoe *ongebruikelijker* het woordgebruik. Geen duidelijke tendenzen zijn echter merkbaar.

14 Percentage van gebruikte lemmata: Grafieken





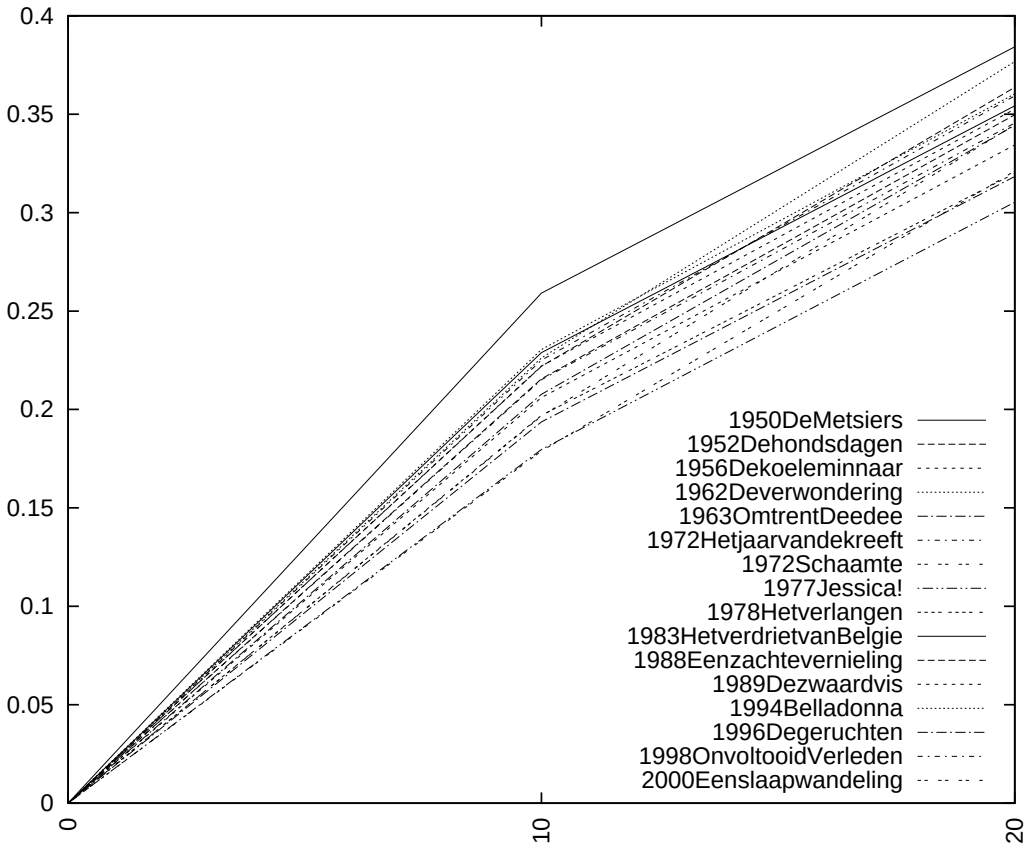
Percentage van gebruikte lemmata: Uitleg

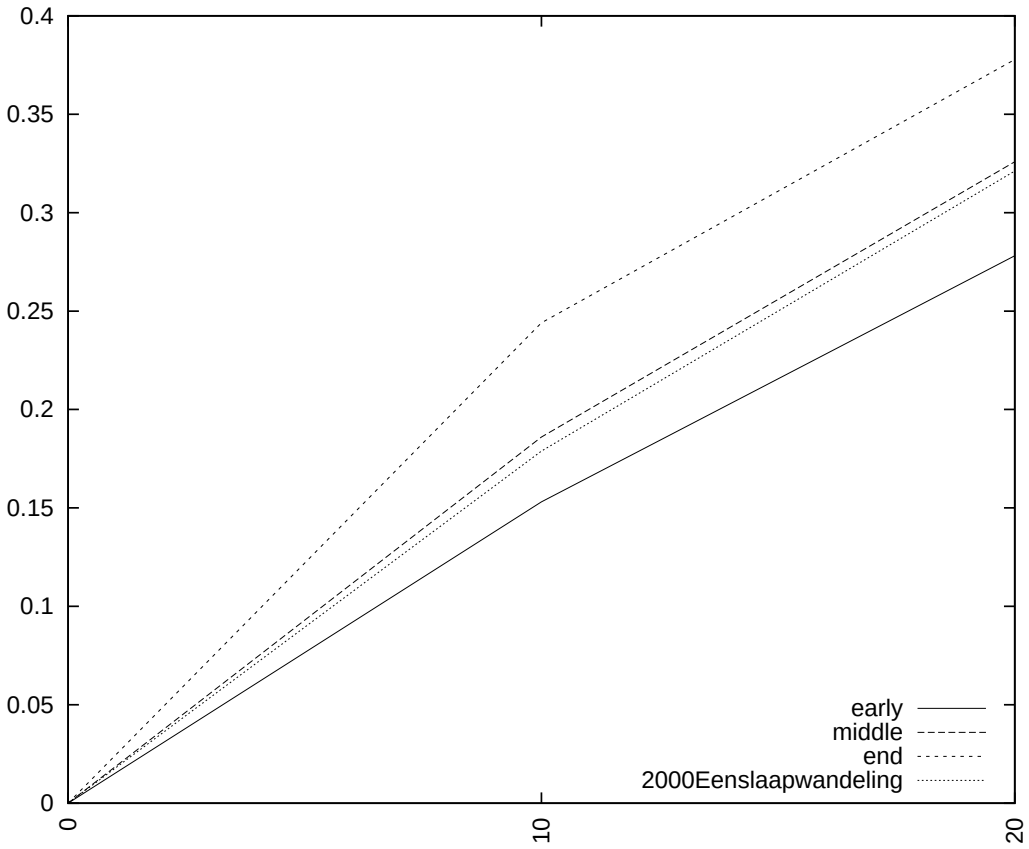
In deze berekening tellen we eerst het totaal aantal lemmata dat in het werk wordt gebruikt. Vervolgens gaan we na - in tranches van 10% - hoeveel procent van deze lemmata er op dat moment reeds werden aangetroffen.

Een voorbeeld ter verduidelijking: de plot van *De Metsiers* in de bovenste grafiek. Van alle lemmata die in het werk worden gebruikt, zijn er na 10% van het boek reeds 25.9% gebruikt, in tegenstelling tot de meeste werken, waar er op dat moment slechts een 20-tal procent van de lemmata werden geïntroduceerd. Men zou deze berekening kunnen beschouwen als een meting van hoe snel de auteur zijn lexicale repertoire opgebruikt. Of nog anders uitgedrukt (onder zwaar voorbehoud): een steile start, gevolgd door een afvlakkende curve, wijst erop dat de auteur doorheen het werk minder nieuwe concepten introduceert en het werk dus thematisch relatief constant blijft.

Een Slaapwandeling bewandelt hier het gemiddelde parcours.

15 Percentage van gebruikte lemmata (eerst 20%): Grafieken



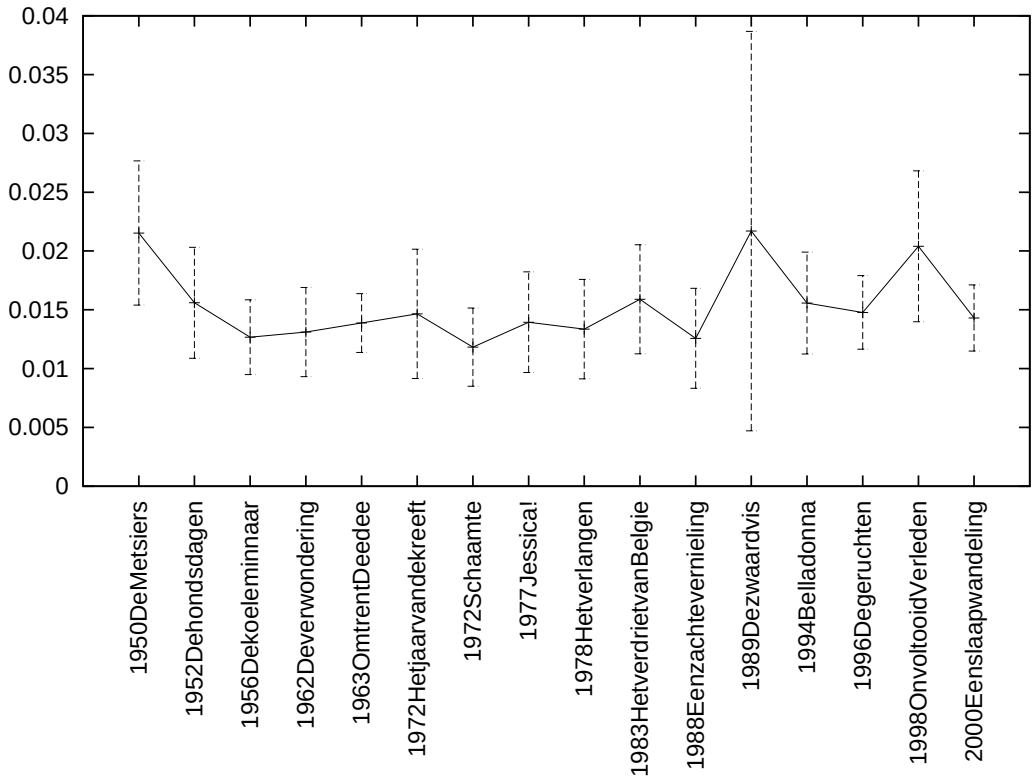


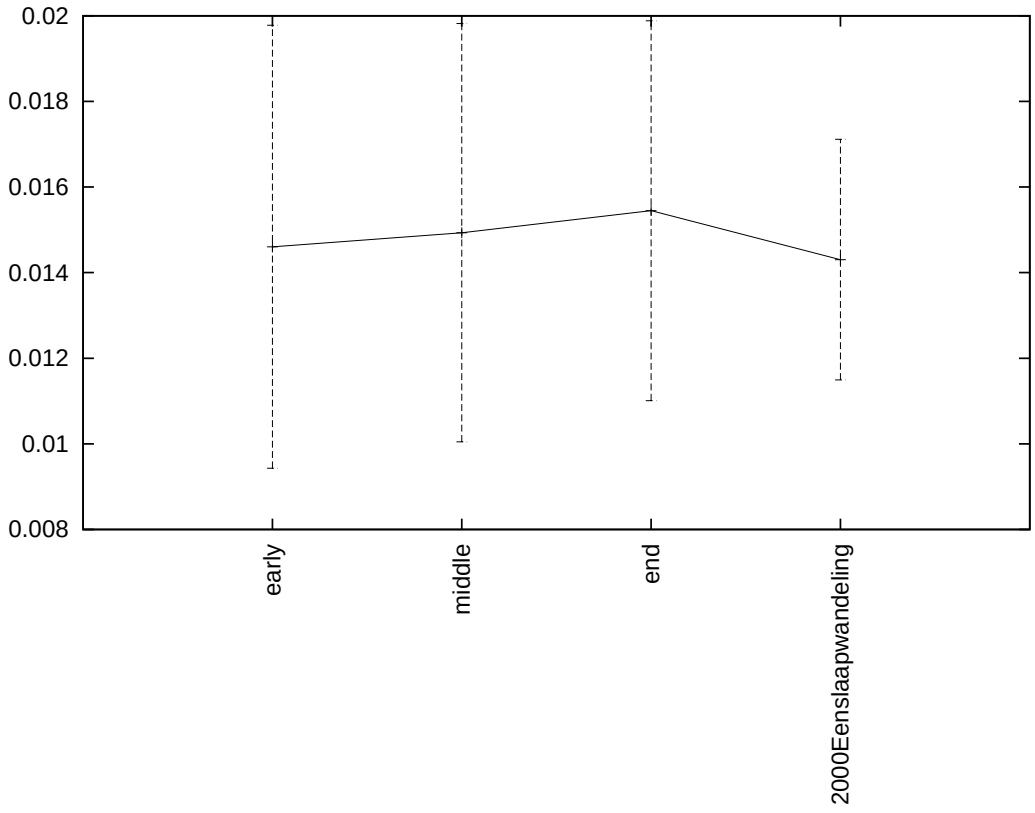
Percentage van gebruikte lemmata (eerst 20%): Uitleg

In deze grafieken zien we een uitvergroting van de eerste 20%. Hier zien we wel dat *Een Slaapwandeling* (als we *Jessica!*[6] buiten beschouwing laten) een relatief trage groeicurve heeft, wat er (weerom onder voorbehoud) op wijst dat het werk thematisch gezien relatief homogeen is.

[6] Kan de trage groeicurve van *Jessica!* worden verklaard door het feit dat het oorspronkelijk een toneelstuk is?

16 Lexicale herhaling (context=2): Grafieken





Lexicale herhaling (context=2): Uitleg

Hier wordt berekend hoe vaak er lexicale herhaling is binnen een bepaalde tijdspanne. Een voorbeeld:

Ik heb het gehad

In deze zin hebben we lexicale herhaling, aangezien we twee woordvormen van hetzelfde lemma *hebben* vinden binnen een context van twee woorden.

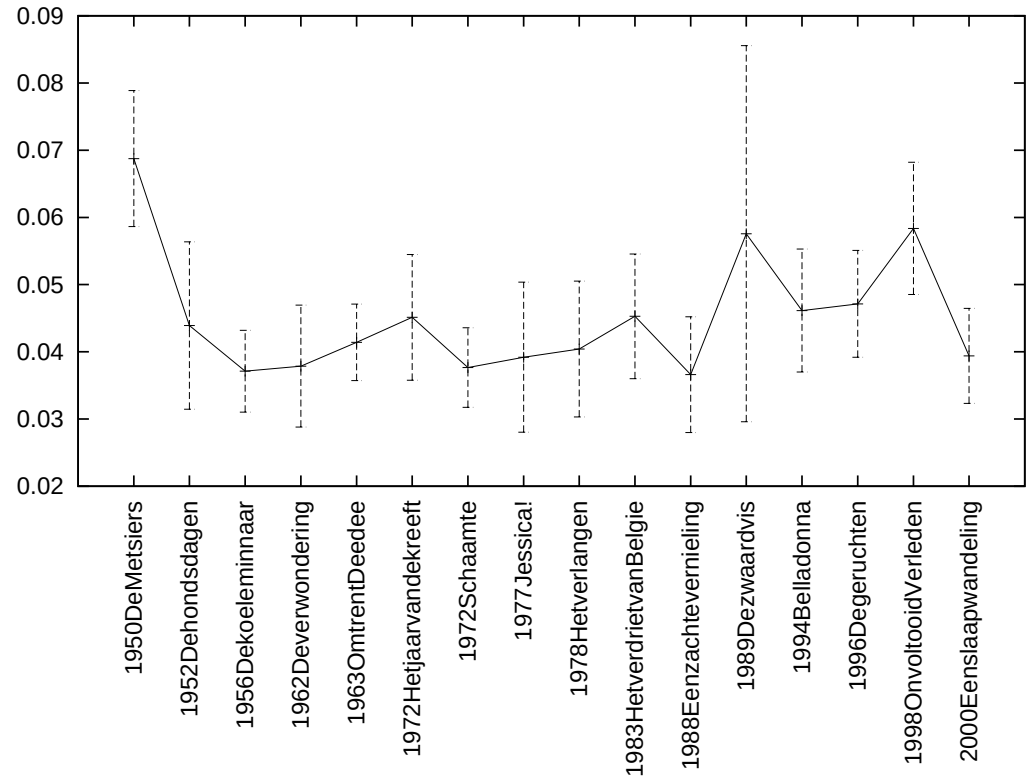
Een ander voorbeeld:

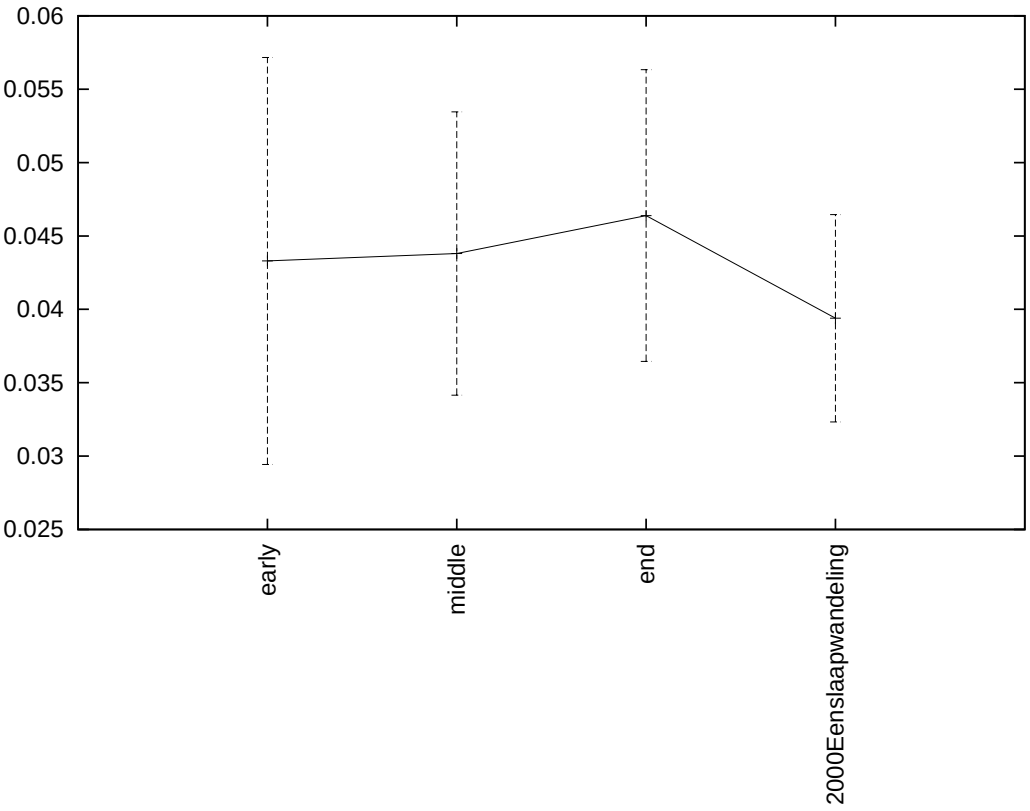
Ik heb het helemaal gehad

Voor een context van **2** woorden vinden we in deze zin geen lexicale herhaling. Voor een context van **3** dan weer wel.

Alzheimer patiënten zouden neigen tot een hogere graad van lexicale herhaling. Voor een context van 2 woorden is dit uiteraard minimaal. Op de volgende bladzijden kan je dezelfde berekening vinden voor een steeds groter wordende context.

17 Lexicale herhaling (context=5): Grafieken

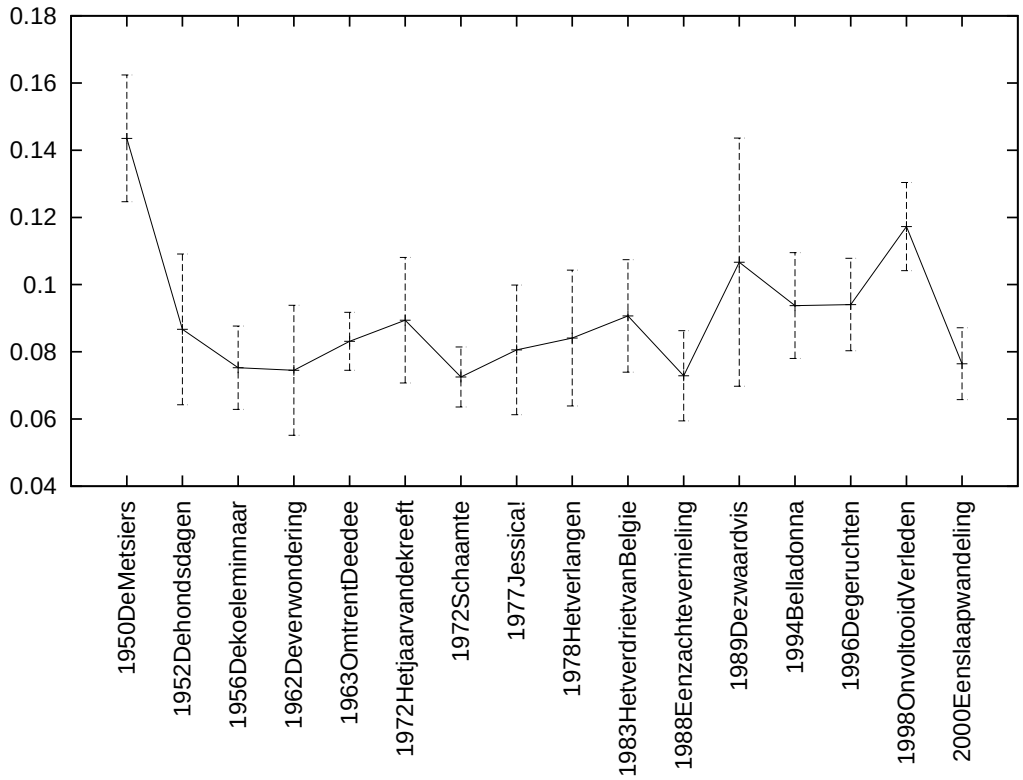


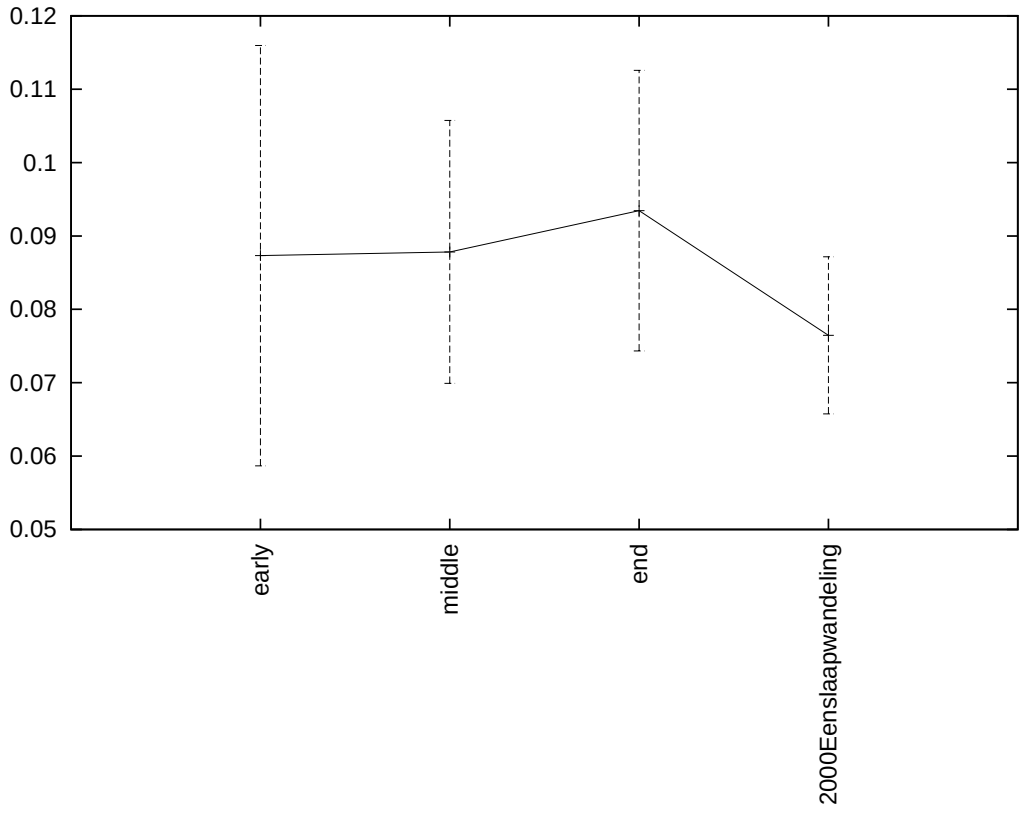


Lexicale herhaling (context=5): Uitleg

Licht merkbare tegenovergestelde tendens.

18 Lexicale herhaling (context=10: Grafieken)

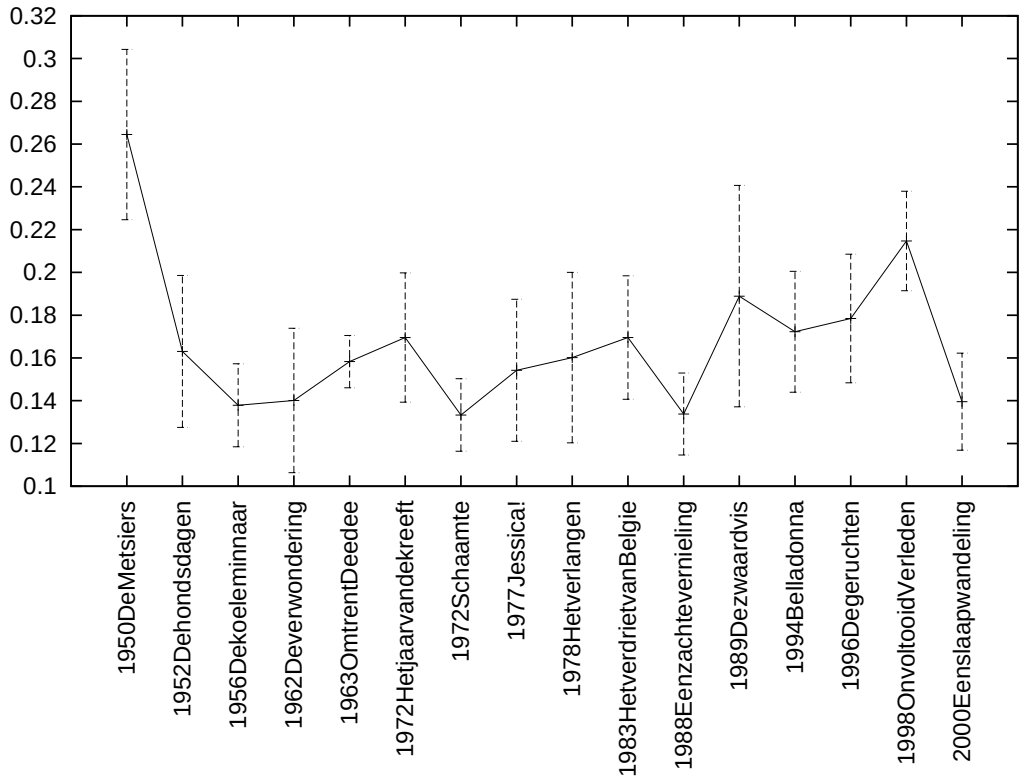


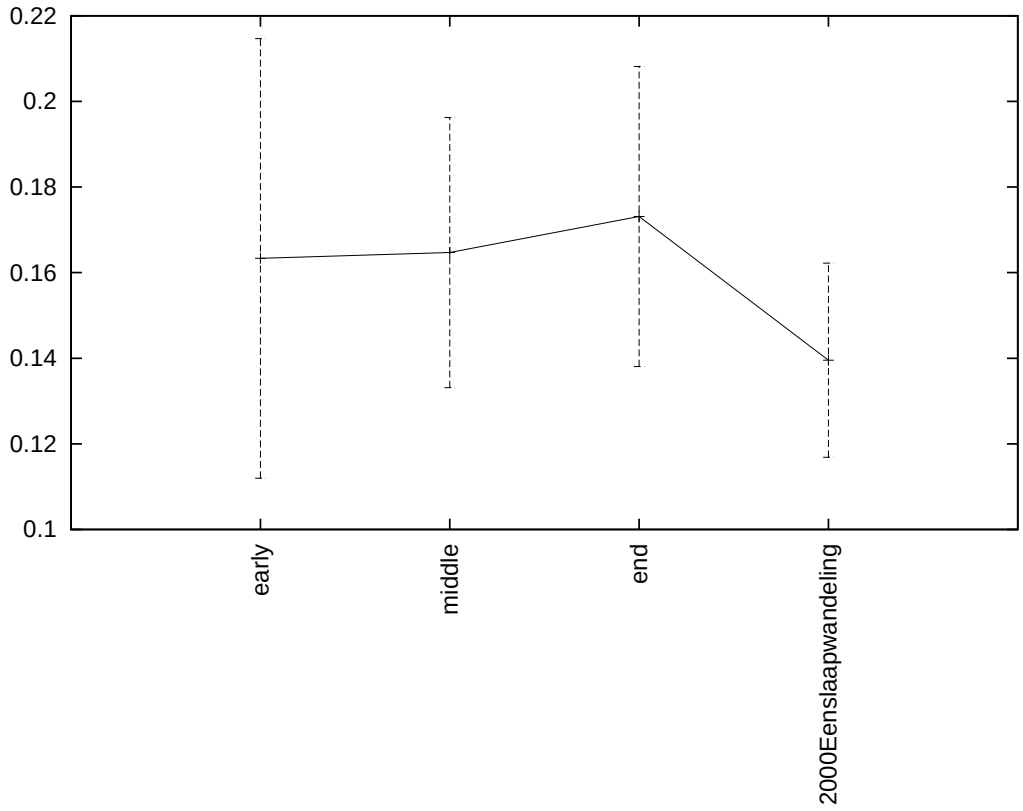


Lexicale herhaling (context=10): Uitleg

Licht merkbare tegenovergestelde tendens.

19 Lexicale herhaling (context=20): Grafieken

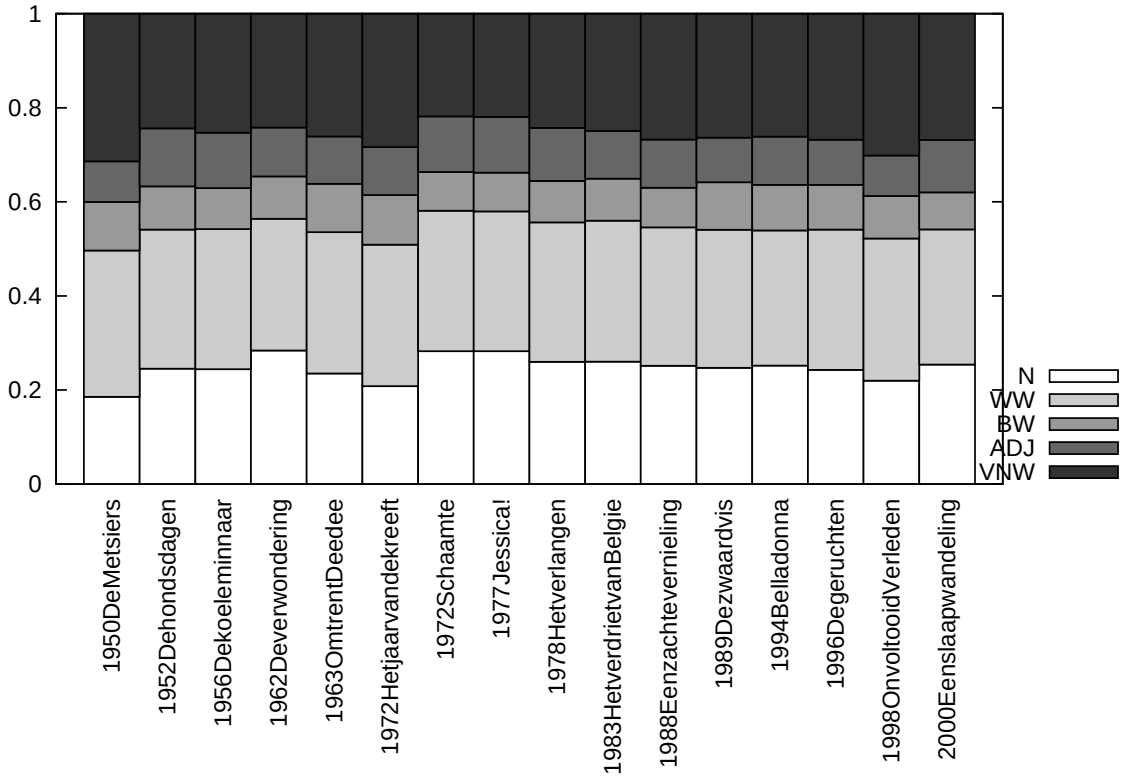


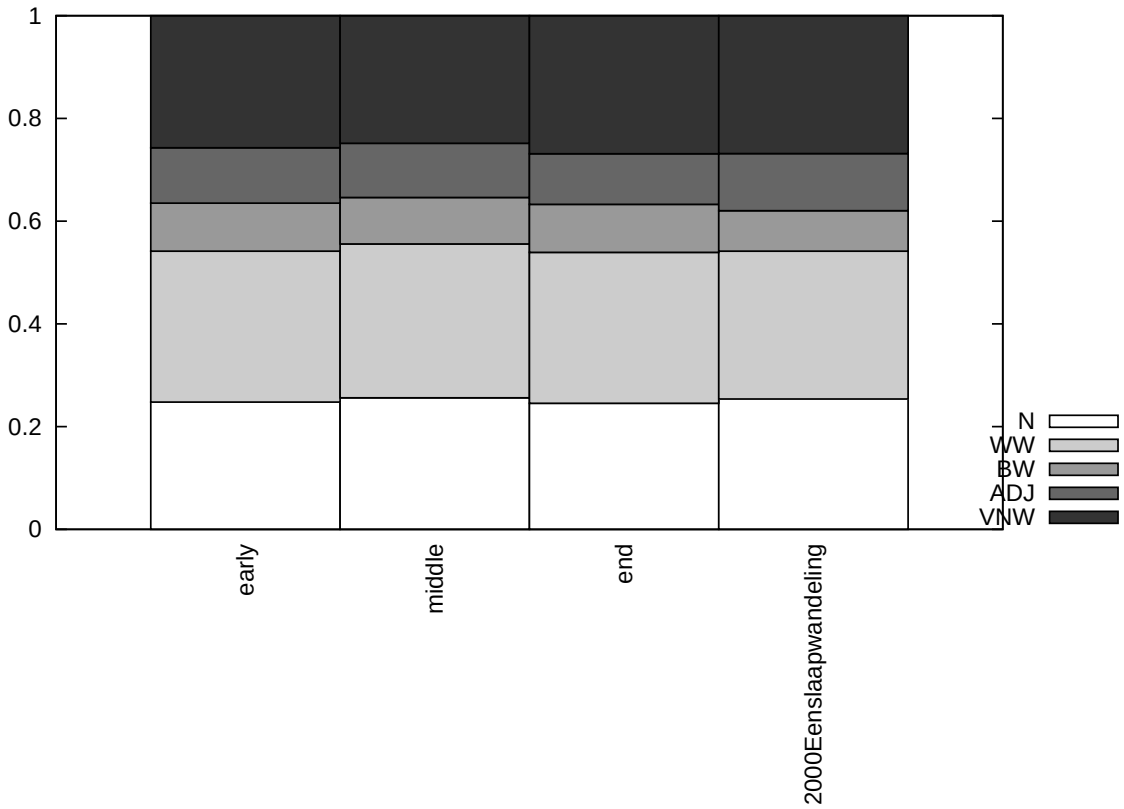


Lexicale herhaling (context=20): Uitleg

Licht merkbare tegenovergestelde tendens.

20 Distributie van woordsoorten: Grafieken

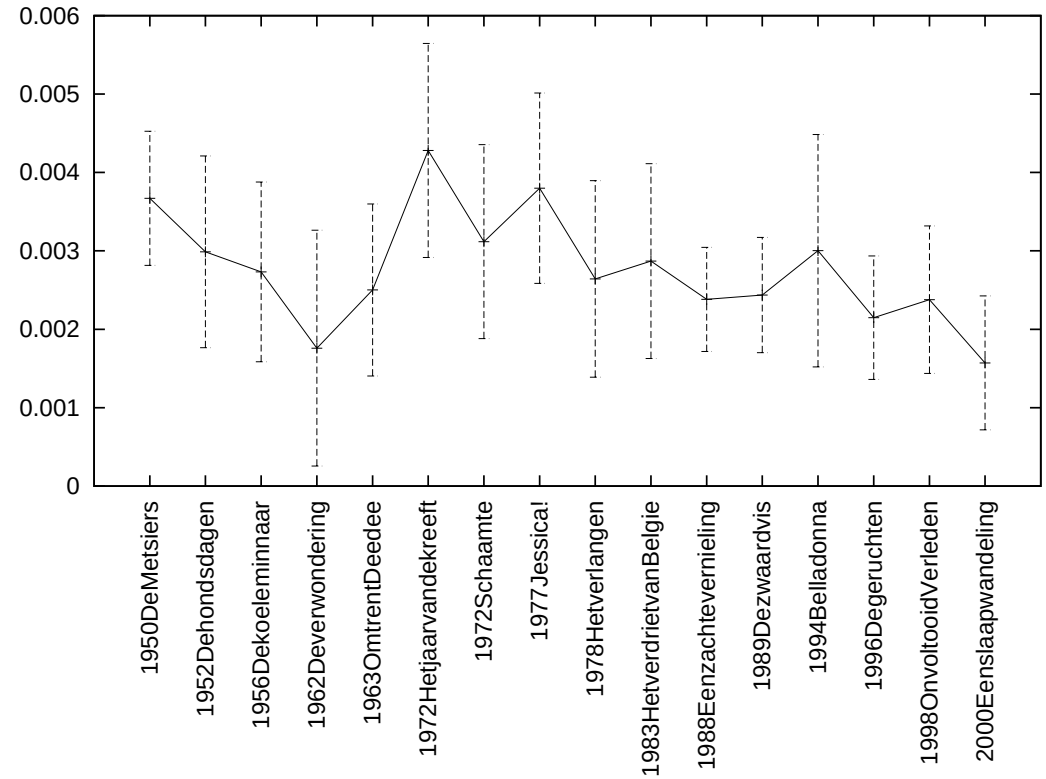


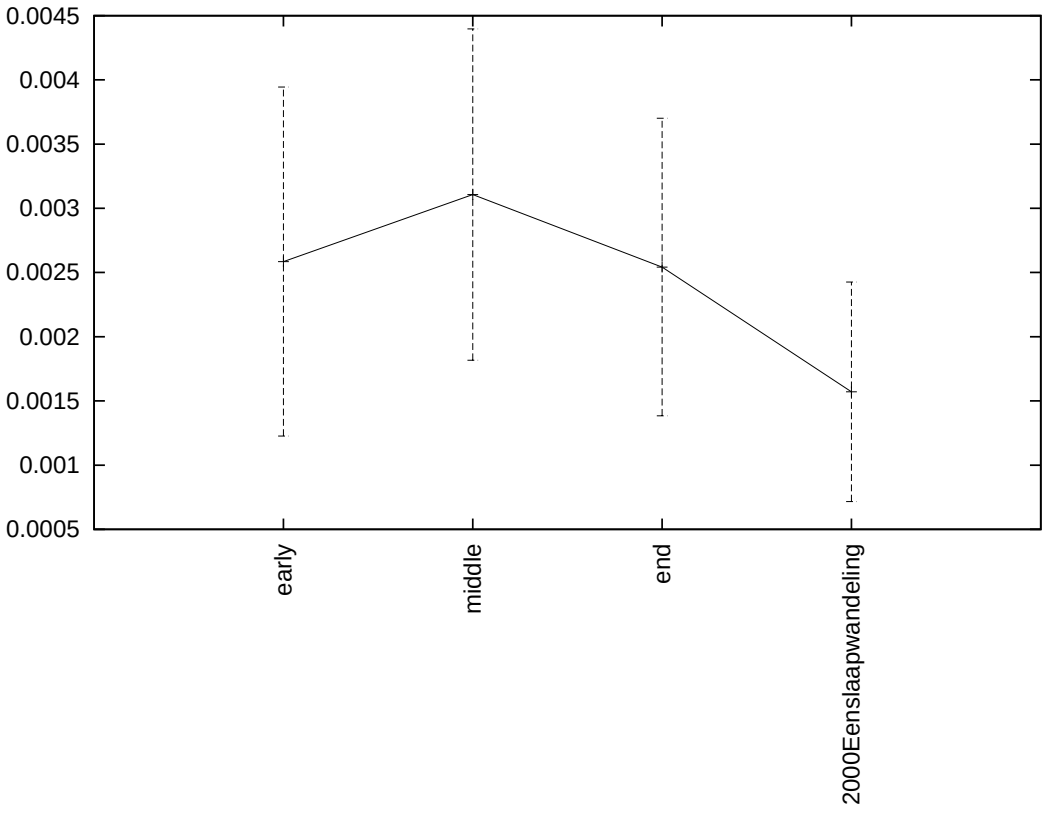


Distributie van woordsoorten: Uitleg

Hier wordt de relatieve distributie van een aantal woordsoorten weergegeven. Alzheimer patiënten zouden minder naamwoorden gebruiken en dit compenseren met meer werkwoorden en voornaamwoorden. Deze tendens is niet merkbaar bij Claus.

21 Percentage Interjecties: Grafieken

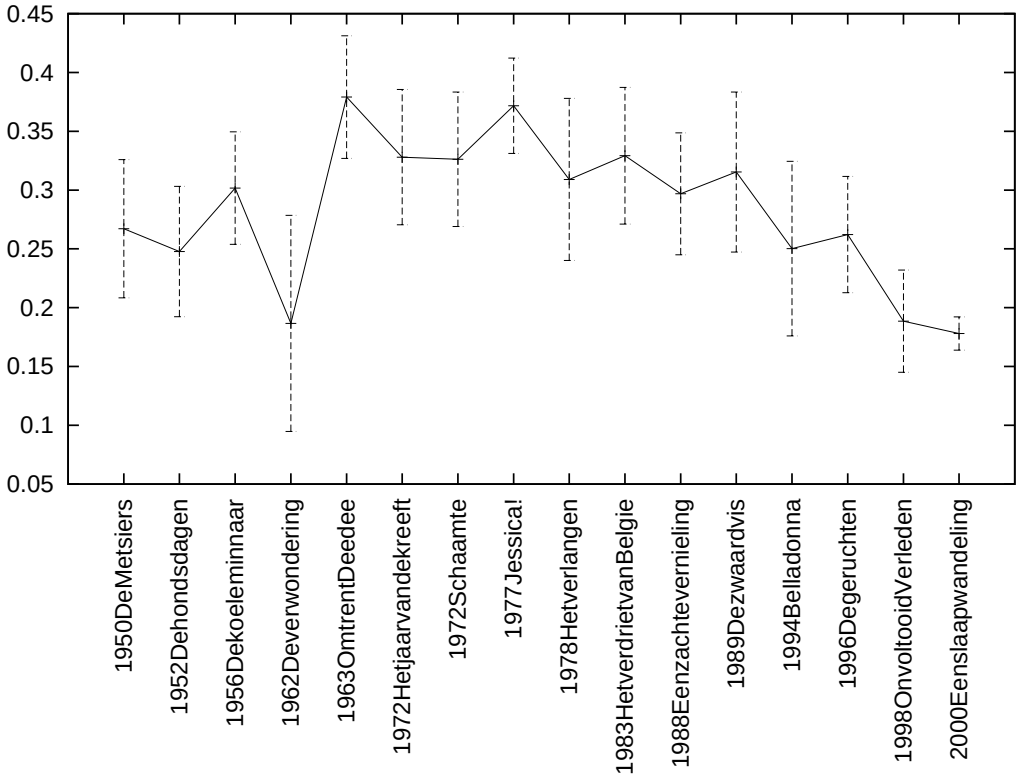


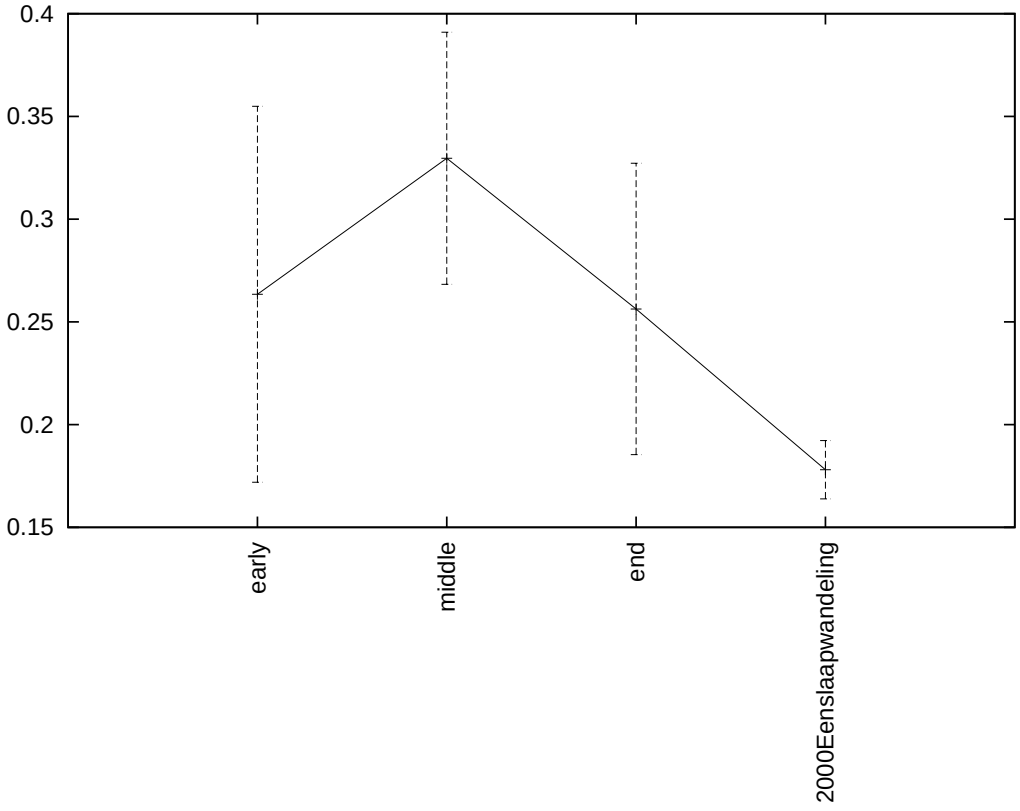


Percentage Interjecties: Uitleg

Men zou kunnen verwachten dat schrijftaal de spreektaal nabootst. Dit zou moeten leiden tot een verhoogd percentage aan interjecties. De omgekeerde trend kan worden vastgesteld.

22 Percentage dialoog: Grafieken

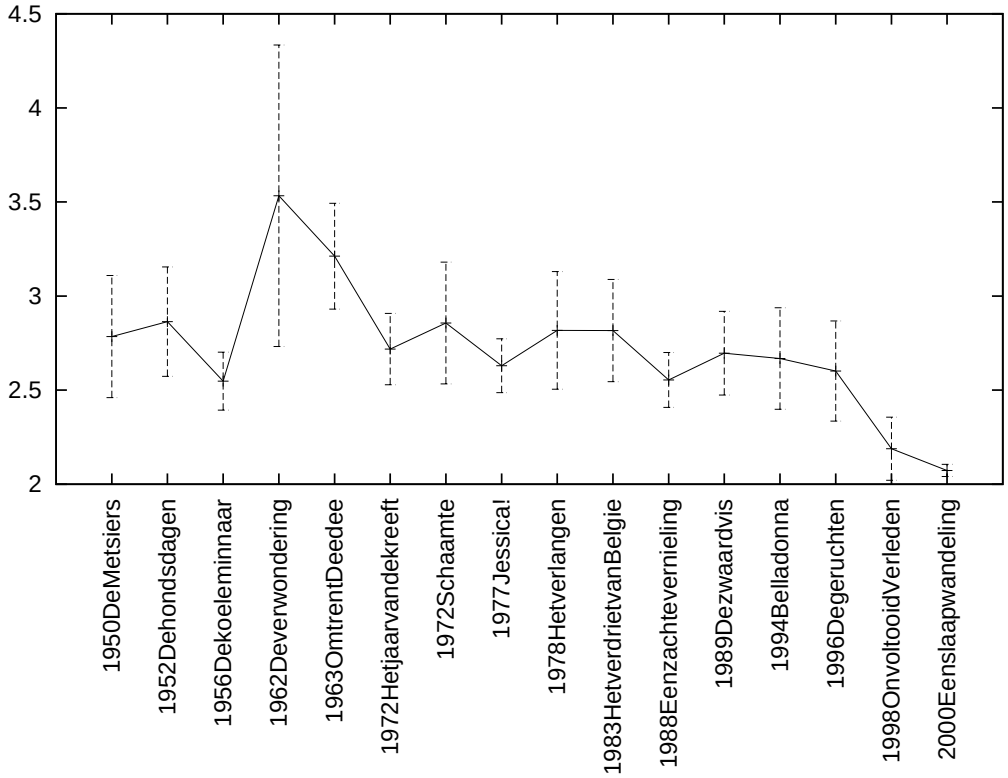


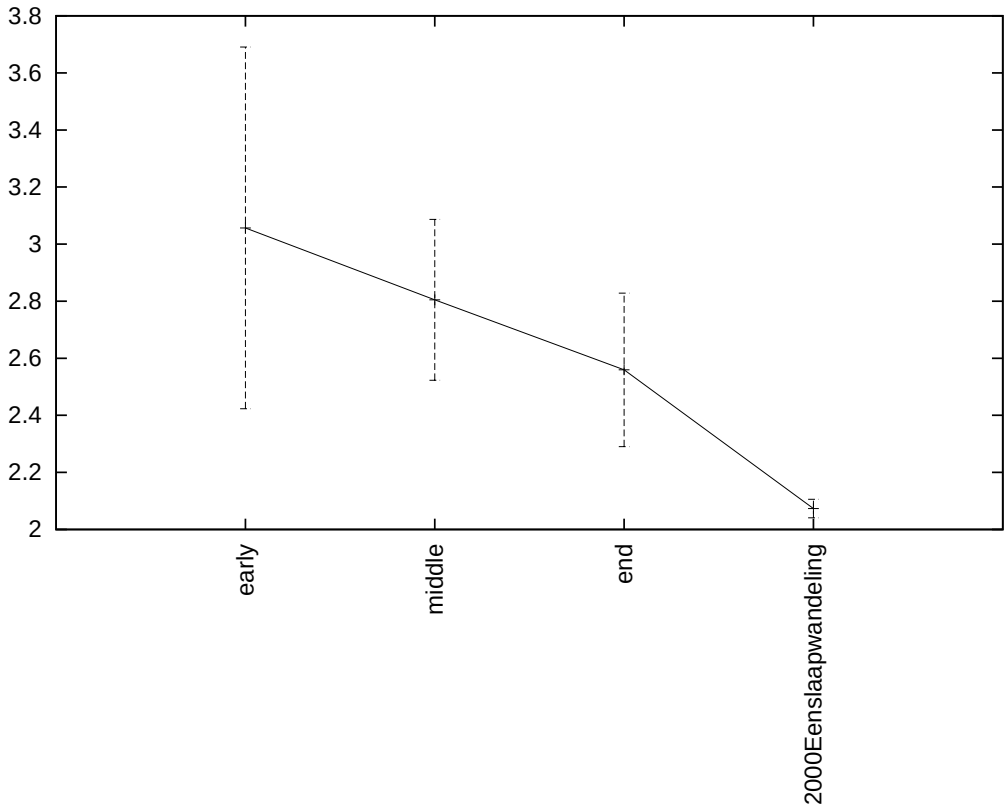


Percentage dialoog: Uitleg

In het onderzoek naar de werken van Iris Murdoch werd een verhoogd percentage aan dialogen vastgesteld. De metingen op het werk van Claus laten een omgekeerde tendens zien.

23 Gemiddeld aantal (bij)zinnen per zin: Grafieken

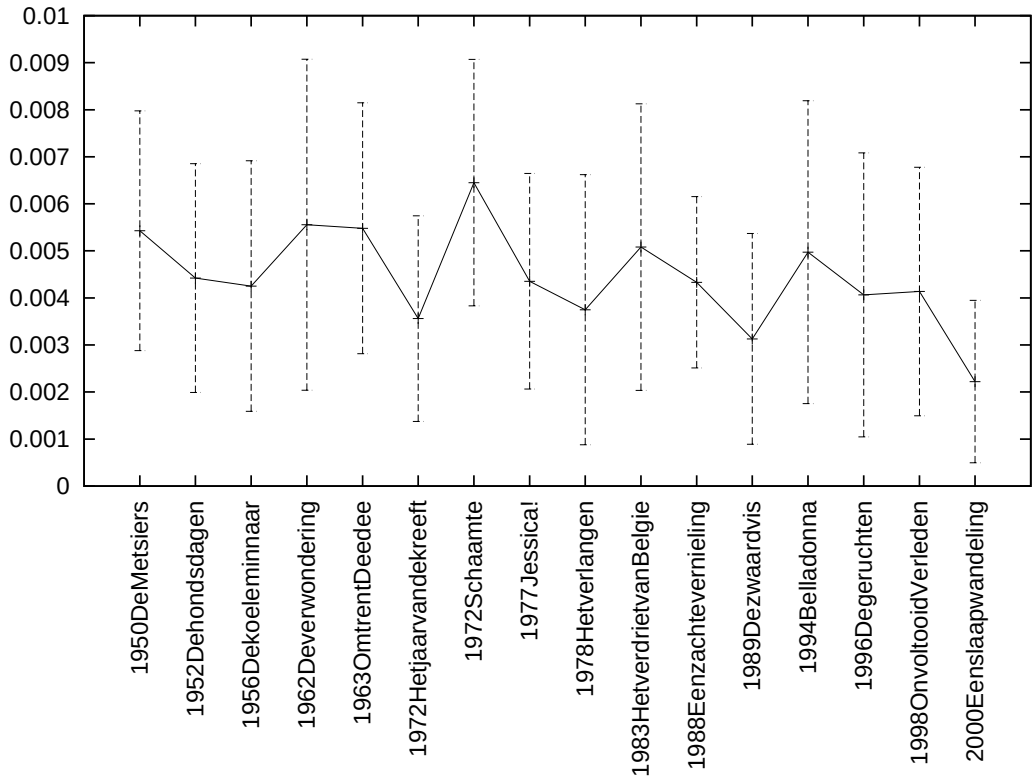


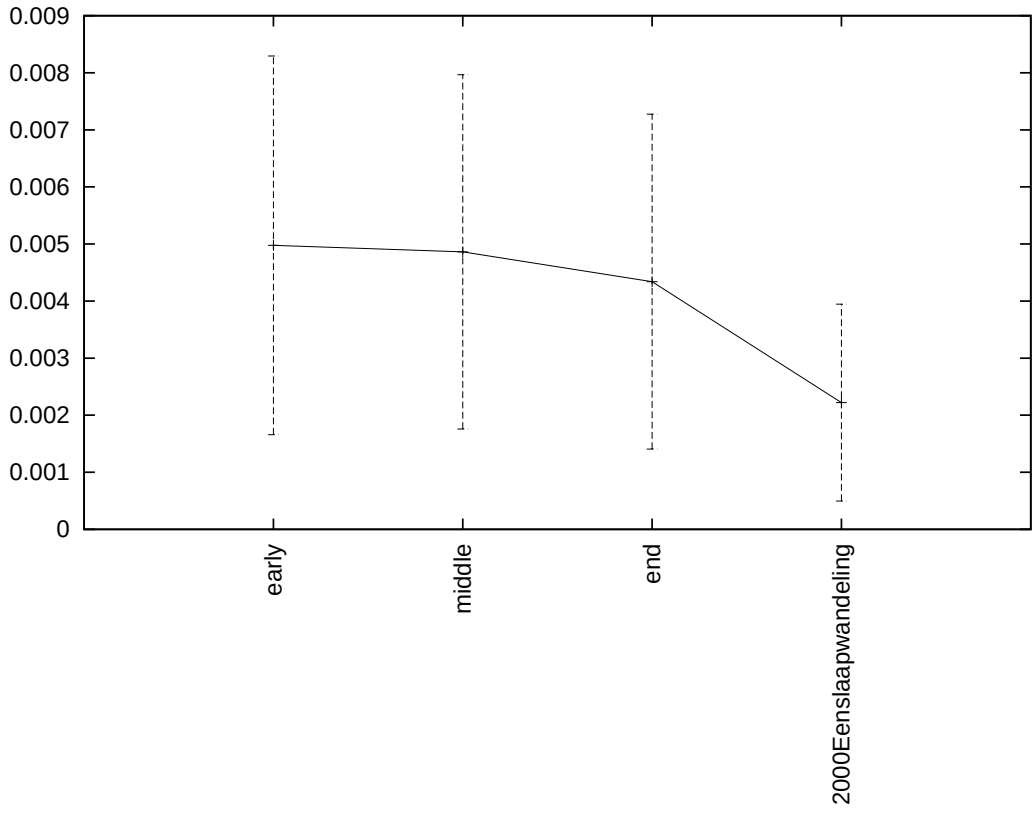


Gemiddeld aantal (bij)zinnen per zin: Uitleg

Alzheimer patiënten zouden over het algemeen syntactisch eenvoudigere constructies moeten maken. Dit kan onder meer gemeten worden door het gemiddeld aantal (bij)zinnen per zin te berekenen. De grafieken laten in deze wel een duidelijke tendens zien. In *Een Slaapwandeling* worden gemiddeld minder bijzinnen gebruikt, maar het is duidelijk dat dit een tendens is die in de natuurlijke evolutie van het oeuvre merkbaar is.

24 Percentage Passiefconstructies: Grafieken

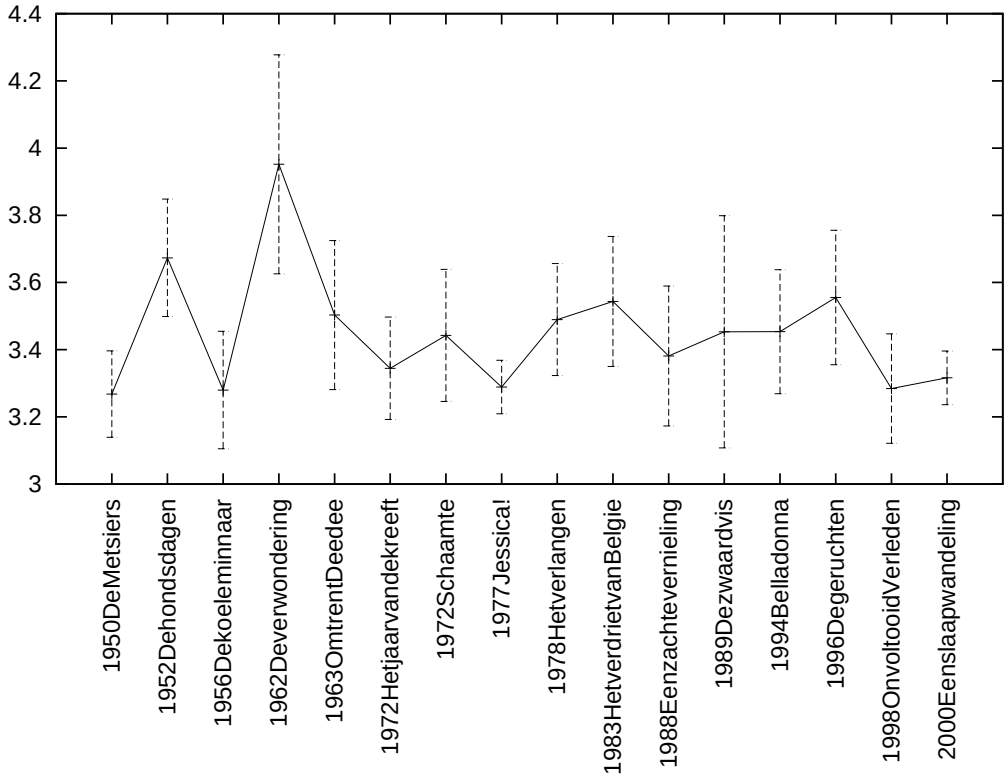


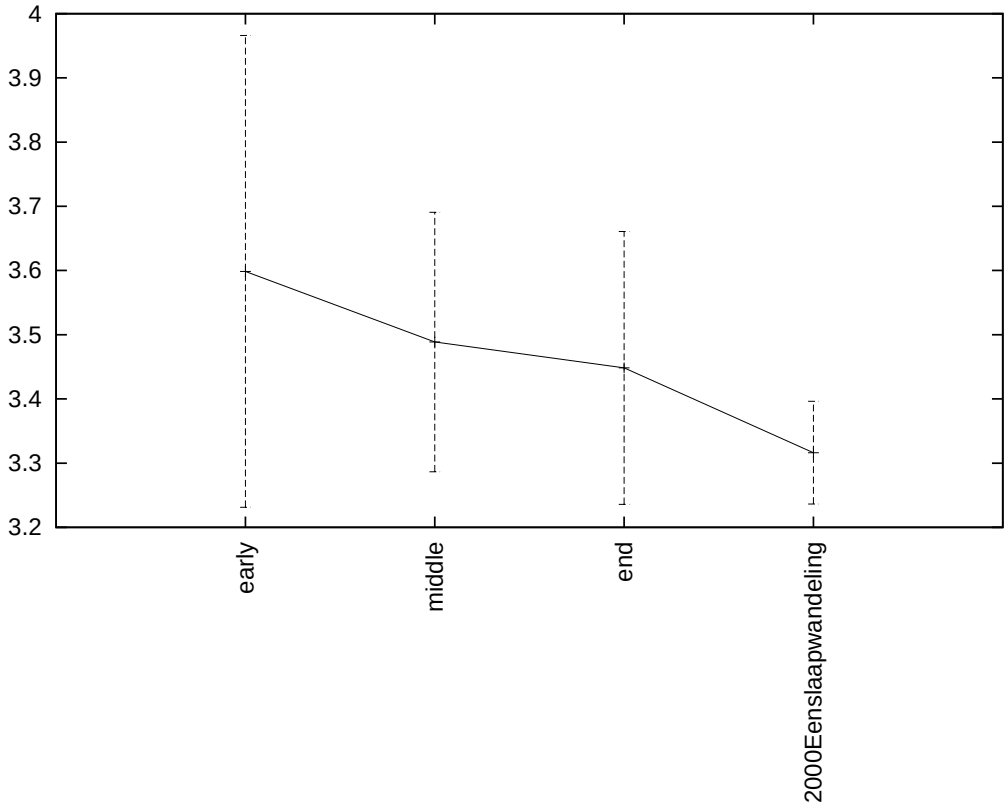


Percentage Passiefconstructies: Uitleg

Passiefconstructies vereisen een grotere cognitieve inspanning. Het percentage van passiefconstructies (tov het totaal aantal (bij)zinnen) zou achteruit moeten gaan bij Alzheimer patiënten. Deze trend wordt bevestigd.

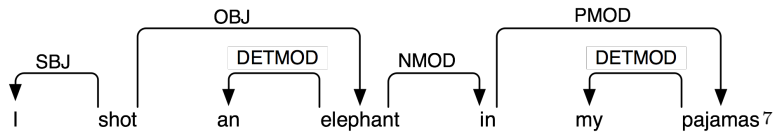
25 Gemiddelde inbedding in dependentie-analyse: Grafieken





Gemiddelde inbedding in dependentie-analyse: Uitleg

Deze berekening meet de complexiteit van de gebruikte syntactische structuren. In de onderstaande syntactische structuur krijgt het woord *my* een inbeddingsscore van **4**, aangezien er vier stappen moeten worden gezet om van het hoofd van de zin (*shot*) tot aan *my* te geraken. Hoe complexer de zin, hoe dieper de woorden zijn ingebed met betrekking tot het hoofd.



Complexe syntactische structuren eisen veel werkgeheugen voor een taalgebruiker. Alzheimer patiënten zouden daarom de voorkeur geven aan minder complexe syntactische structuren. Deze tendens kan in zekere mate worden vastgesteld.

[7] http://nltk.googlecode.com/svn/trunk/doc/images/depgraph0.png