



# Literary translations and tools for stylometric research

Karina van Dalen-Oskam

18 September 2017



UNIVERSITEIT VAN AMSTERDAM

# My background

- **Literary studies** (Middle Dutch, Modern Dutch and English)
- **Literary onomastics** (Literary Name studies)
- **Lexicography** (Early Middle Dutch Dictionary, 1200-1300)
- And a bit of **Translation studies**

# Stylometry

# THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate

SPEAKEASY

## The Science That Uncovered J.K. Rowling's Literary Hocus-Pocus

By BEN ZIMMER

Jul 16, 2013 8:33 am ET



DAVID CHESKIN / PRESS POOL J.K. Rowling

The literary world is still abuzz over [the revelation](#) by London's Sunday Times that J.K. Rowling of "Harry Potter" fame secretly wrote the well-received crime novel "The Cuckoo's Calling" under the pen name Robert Galbraith. In chasing the scoop, the reporters called upon two experts in the field of authorship attribution to determine if "Galbraith" was really Rowling. The experts ran the texts through software programs

THE CHRONICLE OF HIGHER EDUCATION

SECTIONS

FEATURED

Why Bechtel Is Dying

Can Design Thinking Redesign Higher Ed?

Your Daily Briefing

Get the Teaching News

TECHNOLOGY

### The Professor Who Declared, It's J.K. Rowling



Duquesne U.

"Nothing that we do is magic," says Patrick Juola, a computer scientist at Duquesne U., whose computer program analyzed linguistic style to identify a detective novel as having been written, pseudonymously, by Harry Potter's creator.

By Steve Kolowich | JULY 29, 2013

Patrick Juola has practiced stylometry, the science of linguistic style, for decades. But he was never famous for it until this month, when he helped unmask the world's best-known living author.

Mr. Juola, an associate professor of computer science at Duquesne University, was one of two academics enlisted by London's Sunday Times to confirm a tip that J.K. Rowling, creator of the Harry Potter series, had written a new detective novel under a nom de plume.

After Ms. Rowling acknowledged the ruse, Mr. Juola found himself caught up in a tale spanning

# Patrick Juola

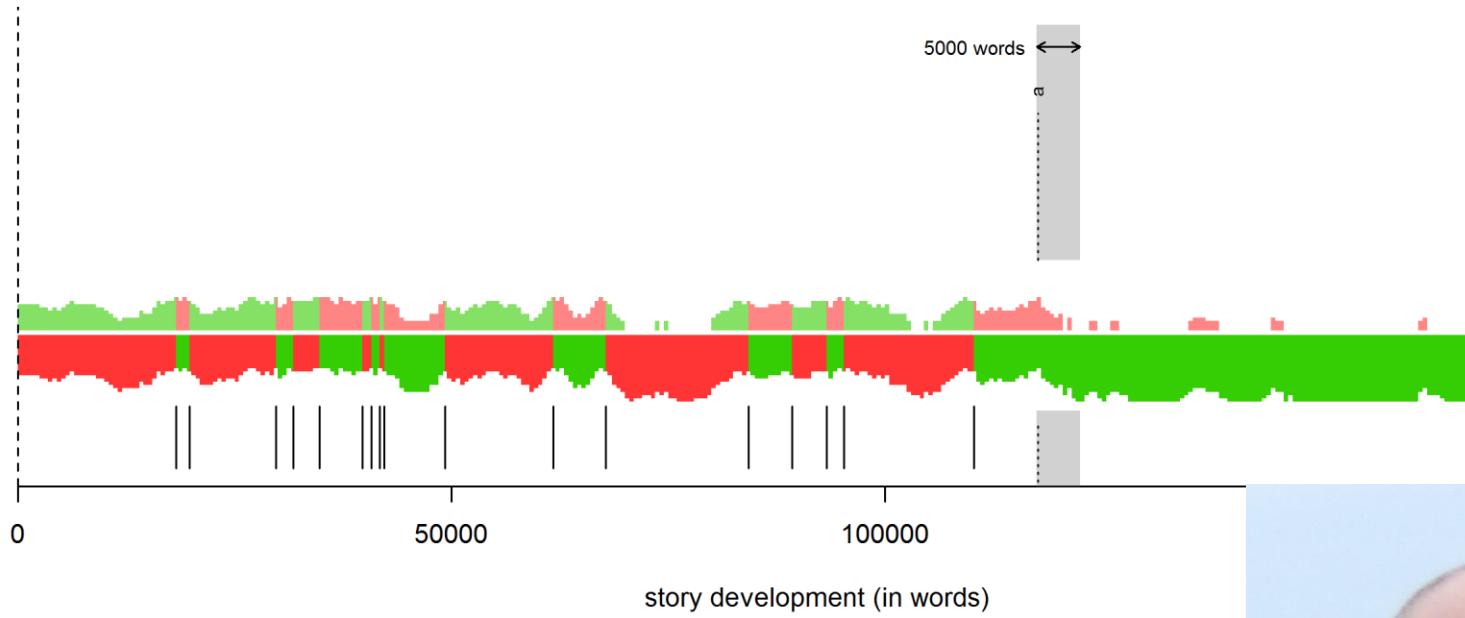
# Onderzoekers werpen met de computer nieuw licht op auteurschap Wilhelmus

Het Wilhelmus wordt ook wel het oudste volkslied ter wereld genoemd, maar we weten niet wie het schreef. Met nieuwe computertechnieken zijn onderzoekers uit Antwerpen, Amsterdam en Utrecht een mogelijke auteur op het spoor gekomen. Het



Datum 10 mei 2016

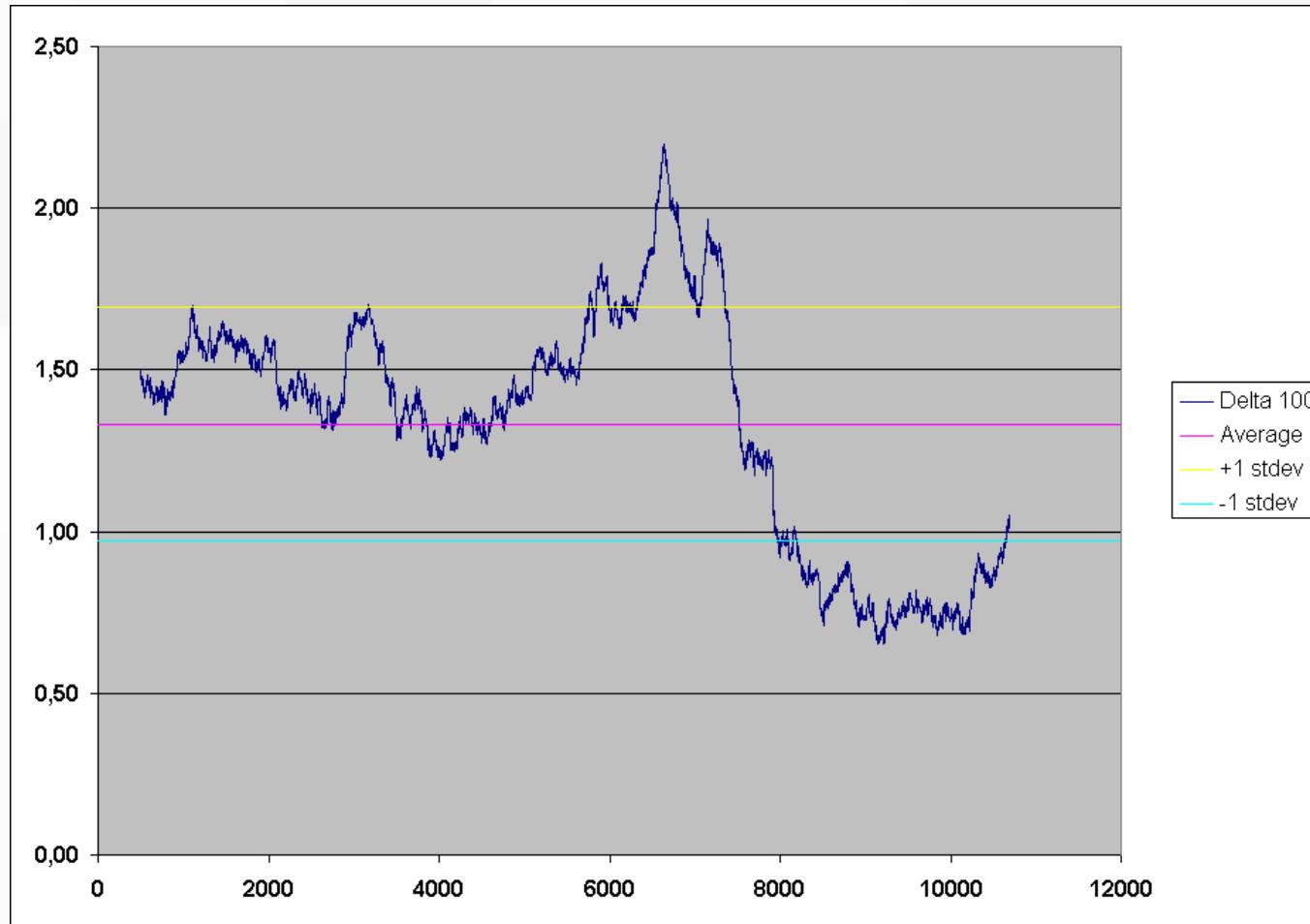
Mike Kestemont

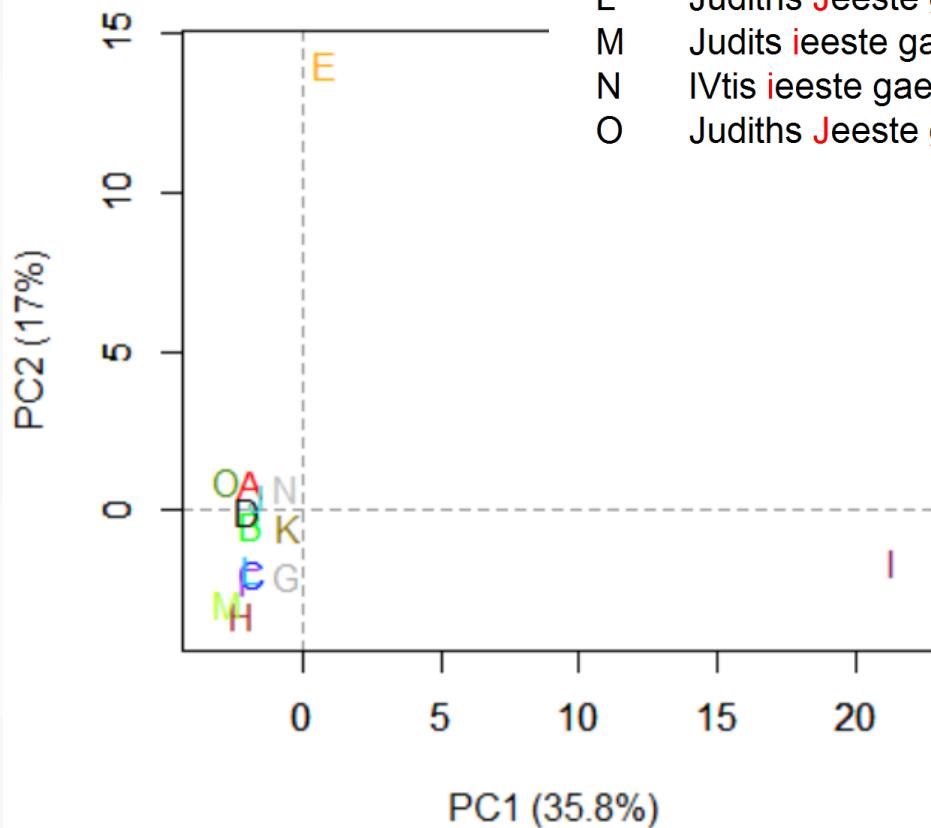


# Jan Rybicki



# Authors





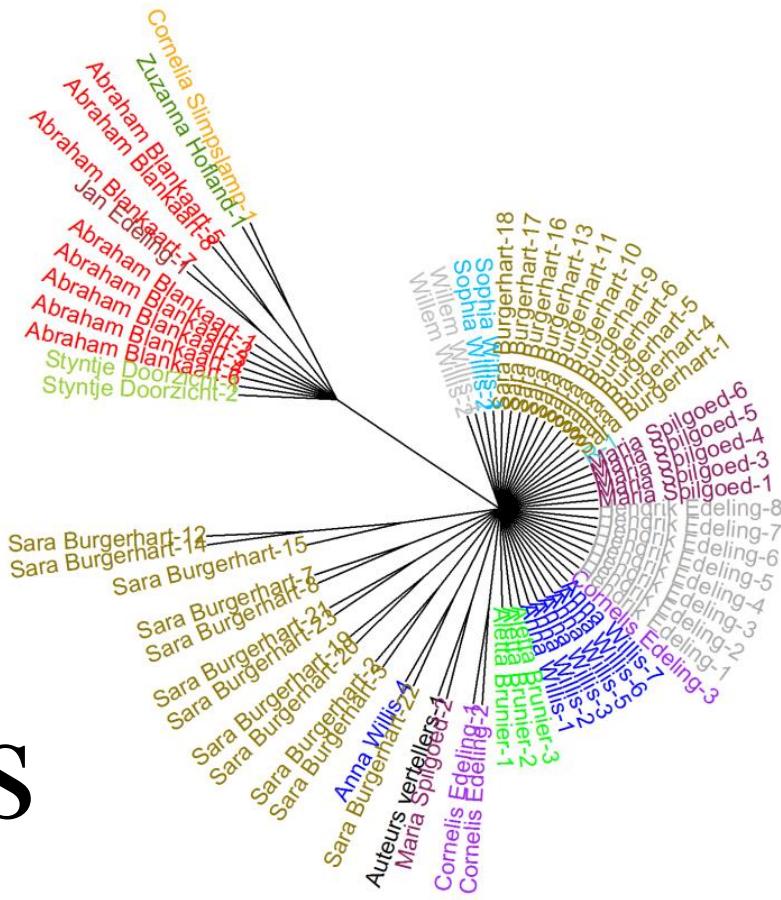
- |   |                                      |   |  |
|---|--------------------------------------|---|--|
| A | Judits <b>i</b> eeste gaet hier an   | A | wi bintse <b>o</b> f het ware .i. lam                    |
| B | Judits <b>i</b> eeste gaet hier an   | B | [w]i bintse <b>o</b> f <sup>t</sup> ware een lam         |
| C | Judiths <b>g</b> eeste gaet hier an. | C | Wi bindse alle <b>a</b> lst ware een lam.                |
| D | Judits <b>i</b> eeste gaet hier an   | D | Wi bindse <b>o</b> cht het waer .i. lam                  |
| E | vdiths <b>i</b> eeste gaet hier an   | E | So <b>moghedise</b> <b>verwinnen</b> <b>a</b> ls .i. lam |
| F | Judiths <b>J</b> eeste gaet hier an  | F | Wi <b>sellense</b> binden <b>a</b> ls een lam            |
| G | Judits <b>g</b> eeste gaet hier an   | G | wi bindenze <b>o</b> f <sup>t</sup> waer een lam         |
| H | Judits <b>g</b> eeste gaet hier an   | H | Wij bijnden se al <b>o</b> f <sup>t</sup> waer een lam   |
| I | Vdits <b>g</b> eeste gaet hier an    | I | Wi bindense <b>d</b> an <b>wel</b> <b>a</b> ls een lam   |
| J | Iudith <b>J</b> eeste gaet hier an   | J | Wij binden se <b>o</b> f <sup>t</sup> ware lam           |
| K | IVdits <b>i</b> eeste gaet hier an   | K | Wi bindense <b>o</b> f het ware .i. lam                  |
| L | Judiths <b>J</b> eeste gaet hier an  | L | Wi <b>sellense</b> binden <b>a</b> ls .J. lam            |
| M | Judits <b>i</b> eeste gaet hier an   | M | Wi <b>sellense</b> binden <b>a</b> ls .i. lam            |
| N | IVtis <b>i</b> eeste gaet hier an    | N | Wi bindense <b>o</b> f <sup>t</sup> ware een lam         |
| O | Judiths <b>J</b> eeste gaet hier an  | O | wi bindense <b>o</b> f <sup>t</sup> ware .J. lam         |

# Scribes



# Characters

## Senders from Sara Burgerhart Bootstrap Consensus Tree



1-1000 MFW Culled @ 0%  
Classic Delta distance Consensus 0.5



# Literary quality Translation quality

*just counting words...*



# A new definition of 'style'

- Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively.

J. Berenike Herrmann, Karina van Dalen-Oskam,  
Christof Schöch, Revisiting Style, a Key Concept in  
Literary Studies. *Journal of Literary Theory* 2015; 9(1):  
25–52

# The Riddle of Literary Quality

[Home](#)[News](#)[Project team](#)[Proposal](#)[Publications](#)[Riddle in media](#)[Think Tank](#)

## The Riddle of Literary Quality

*The Riddle of Literary Quality* is a research project of the [Huygens Institute for the History of the Netherlands](#) in collaboration with the [Fryske Akademy](#) and the [Institute for Logic, Language and Computation](#) (University of Amsterdam). *The Riddle* officially started January 15th 2012 and will run for four years. The project is funded by the [Computational Humanities Programme](#) of the [Royal Netherlands Academy of Arts and Sciences](#).

## Main research question:

- Do modern novels which are considered to be 'literary' show a different usage of stylistic features than novels which are considered to be 'not literary'?
- In other words: What are the current conventions of literariness?

WAT VIND JIJ  
VAN MIJ?



[Home](#)   [Enquête](#)   [Meer informatie](#)

# HET NATIONALE LEZERSONDERZOEK

VUL DE  
ENQUÊTE IN



Wat is literaire kwaliteit? Wat maakt een roman goed of slecht?  
Wanneer is een roman literair, of juist niet? En zijn deze dingen  
objectief vast te stellen?

Misschien wel. Wellicht zijn er elementen in een boek die bepalen  
dat het als goed of minder goed wordt gezien. Wetenschappers aan  
het Huygens Instituut voor Nederlandse Geschiedenis en de  
Universiteit van Amsterdam gaan dit onderzoeken.

Hiervoor is de mening van de lezer nodig. U dus. Welke boeken  
vindt u goed (of juist niet goed) en welke vindt u literair (of juist  
niet) en waarom?

Doe daarom mee aan het Nationale Lezersonderzoek en vul de  
enquête in. Dat kost ongeveer 5 minuten.

Veel plezier!

- [Delen op Facebook](#)
- [Delen op Twitter](#)



# The National Reader Survey, 2013

Questions to the respondents:

- Age, sex
  - Level of education
  - Zipcode
- 
- How many books do you read per year?
  - Fiction, non-fiction, or both?
  - Opinion on sixteen statements



# Which of these 401 books did you read

- Labeled as fiction by publisher
- Most sold and most lent in 2010-2012
- Published in Dutch for the first time in or after 2007
- Originally Dutch or translated into Dutch
  - 152 Dutch
  - 249 translated
    - 180 from English
    - 69 from nine other languages



# What is your opinion

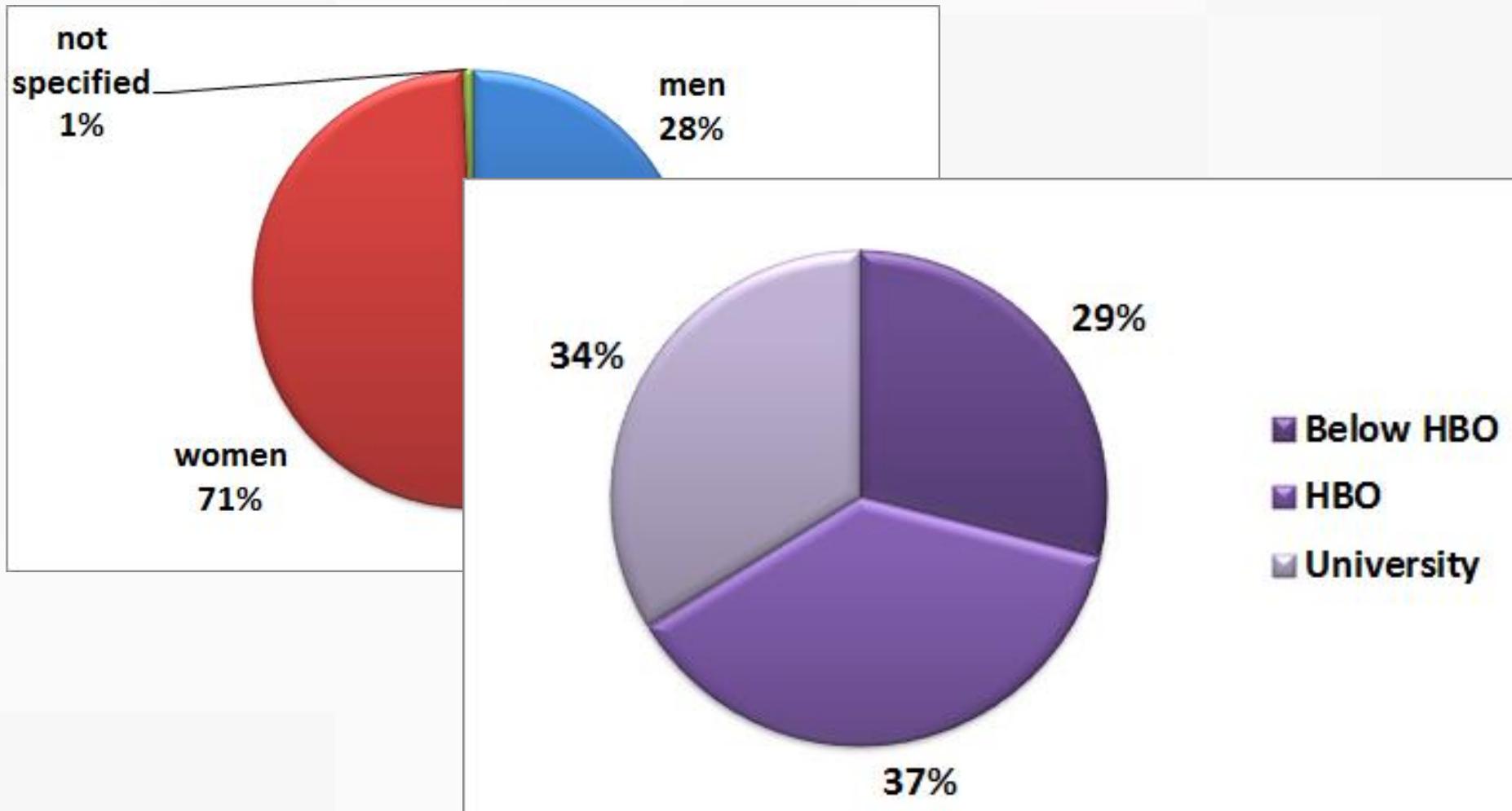
- About seven books you have read
  - **Question: How do you rate this book on a scale of literariness?**  
1 = definitely not literary, 7 = highly literary;  
8 = don't know
  - **Question: How do you rate this book on a scale of general quality?**  
1 = very bad, 7 = very good; 8 = don't know
- About seven books you did not read
  - **rate these on the same two scales**



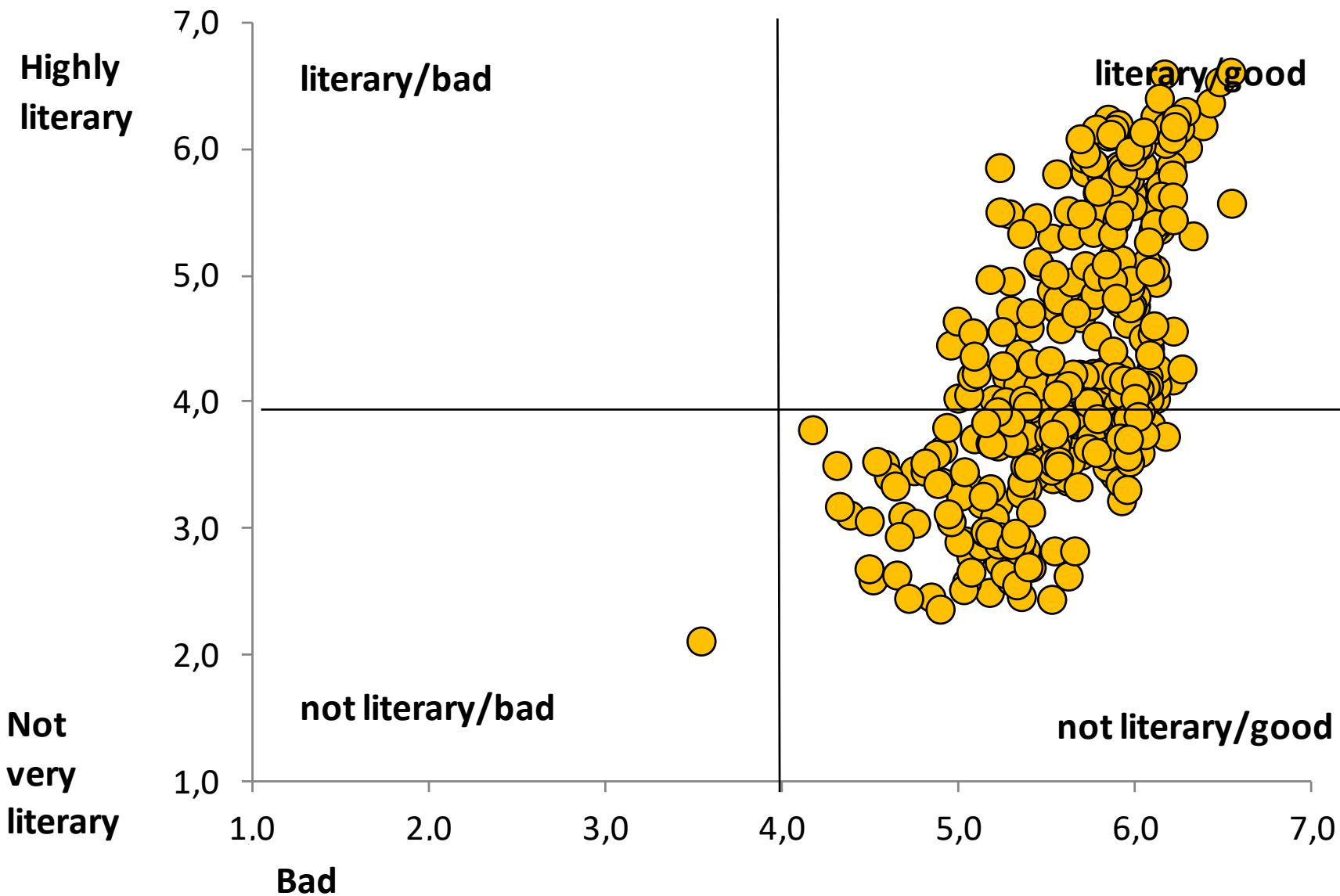
# Motivate your opinion

...for one of the books you read and scored on literariness

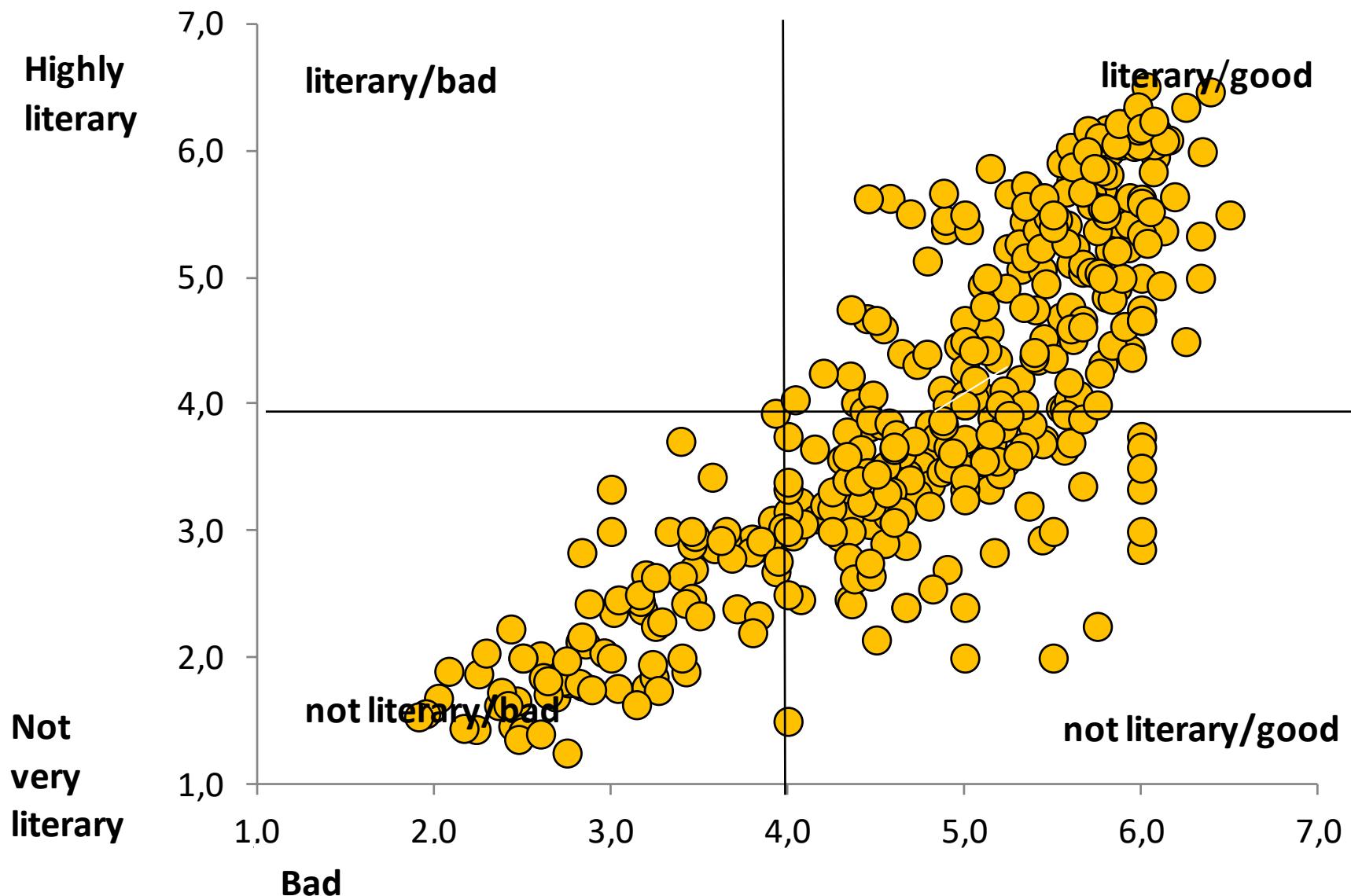
13,782 respondents



# Scores for books read



# Scores for books not read





## Top-10 **least** literary

		N	Mean
1	<b>E.L. James</b> Vijftig tinten grijs	320	2.1
2	<b>Sophie Kinsella</b> Shopaholic & Baby	313	2.4
3	<b>Sophie Kinsella</b> Mag ik je nummer even?	218	2.4
4	<b>Lauren Weisberger</b> Chanel Chic	212	2.5
5	<b>Sophie Kinsella</b> Mini Shopaholic	190	2.5
6	<b>Sophie Kinsella</b> Ken je me nog?	319	2.5
7	<b>Jill Mansell</b> Versier me dan	192	2.5
8	<b>Lauren Weisberger</b> Champagne in Ch Marmont	115	2.5
9	<b>Katie Fforde</b> Trouwplannen	150	2.6
10	<b>Jill Mansell</b> Drie is te veel	259	2.6



## Top-10 **most** literary

		N	Mean
1	<b>Julian Barnes</b> Alsof het voorbij is	816	6.6
2	<b>Erwin Mortier</b> Godenslaap	810	6.6
3	<b>Erwin Mortier</b> Gestameld liedboek	575	6.5
4	<b>Michel Houellebecq</b> De kaart en het gebied	708	6.4
5	<b>Tom Lanoye</b> Sprakeloos	727	6.4
6	<b>Haruki Murakami</b> Norwegian Wood	848	6.3
7	<b>A.F.Th. van der Heijden</b> Tonio	768	6.3
8	<b>Stephan Enter</b> Grip	752	6.3
9	<b>J. Bernlef</b> Geleende levens	509	6.2
10	<b>Umberto Eco</b> Begraafplaats van Praag	573	6.2



## Top-10 worst

		N	Mean
1	<b>E.L. James</b> Vijftig tinten grijs	325	3.5
2	<b>James Worthy</b> James Worthy	194	4.2
3	<b>Kluun</b> Haantjes	221	4.3
4	<b>Heleen van Royen</b> De mannenester	223	4.3
5	<b>Suzanne Vermeer</b> Cruise	207	4.4
6	<b>Emile Proper / Sabine van den Eynden</b> Gooische vrouwen	102	4.5
7	<b>Suzanne Vermeer</b> Après-ski	238	4.5
8	<b>E.L. James</b> Vijftig tinten donkerder	201	4.5
9	<b>Naima El Bezaz</b> Vinexvrouwen	459	4.5
10	<b>Elizabeth Gilbert</b> Eten, bidden, beminnen	631	4.6



## Top-10 best

		N	Mean
1	<b>Jan Brokken</b> Baltische zielen	546	6.5
2	<b>Julian Barnes</b> Alsof het voorbij is	819	6.5
3	<b>Erwin Mortier</b> Gestameld liedboek	576	6.5
4	<b>Tom Lanoye</b> Sprakeloos	727	6.4
5	<b>David Mitchell</b> De niet verhoorde gebeden van Jacob de Zoet	738	6.4
6	<b>Lawrence Hill</b> Het negerboek	549	6.3
7	<b>A.F.Th. van der Heijden</b> Tonio	769	6.3
8	<b>Laurent Binet</b> HhhH	672	6.3
9	<b>Haruki Murakami</b> Norwegian Wood	848	6.3
10	<b>Jo Nesbø</b> Het pantserhart	323	6.3



## Some motivations

Mooi, spannend en verrassend boek. Schrijfstijl minder verrassend.

***Beautiful, full of suspense, surprising book. Writing style not so surprising.***

Een boek zonder goed verhaal dat geen plezier geeft in het lezen. Dat moet dus wel literatuur zijn.

***A book without a good story, without any reading pleasure. That must be literature.***

## Respondents about genre

Het leest deels als een damesroman dat vind ik nu niet bepaald literair.

***It partly reads as a ladies' novel and I don't consider that literary at all.***

het verhaal was matig, taalkundig gezien was het niet goed. Absoluut niet boeiend. Het stijgt niet uit boven een bouquetreeks roman. Onbegrijpelijk dat dit boek een prijs gewonnen heeft.

***The story was average, linguistically it wasn't very good. It doesn't rise above a Harlequin novel. I don't understand why this book won a prize.***

# 10 Computational 01 Stylistics 11 Group

## 0101000 011010110

[main page](#)

### Navigation

- ▼ [main page](#)
  - [more photos](#)
- [corpora](#)
- ▼ [papers and articles](#)
  - [preprints](#)
- ▼ [projects](#)
  - Computer Methods in Textual Studies 2014
  - Go Set A Watchman while we Kill the Mockingbird In Cold Blood
  - testing big dendograms
  - Testing consensus networks
  - testing rolling delta
  - Testing rolling stylometry
  - translationese
- ▼ [stylo R package](#)
  - [installation hints](#)
  - [scripts](#)
  - [scripts: obsolete](#)
  - [Stylometry@Kraków](#)
  - [workshops](#)

# Stylo Package for R

Jan Rybicki,  
Maciej Eder,  
Mike Kestemont

**latest news:** the version **0.6.3** of the R package "stylo" released! Click [here](#) for further details.

Apart from the package as mentioned above, we still provide some scripts. Click [here](#) if you're interested.

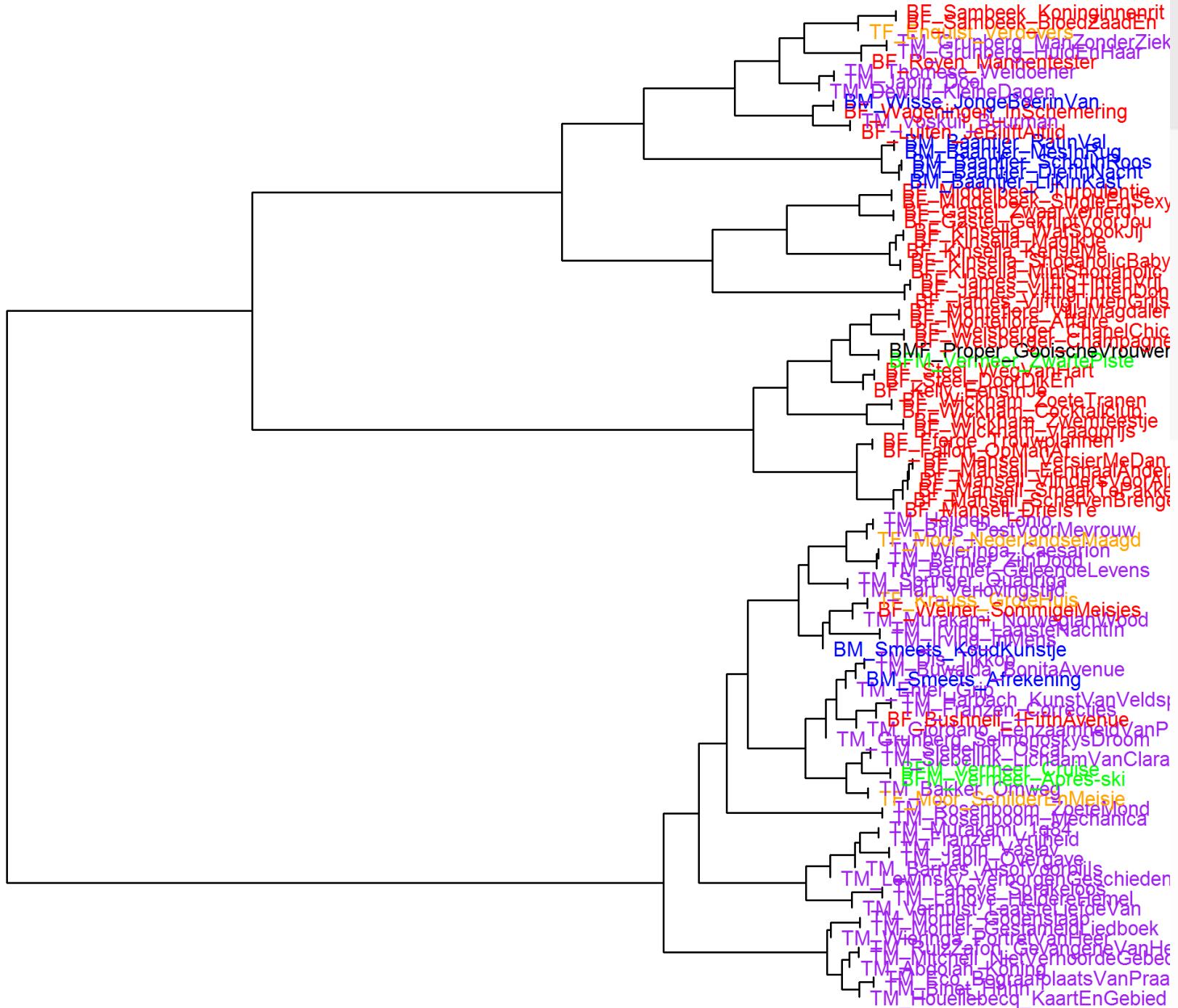
This [HOWTO](#) of the package "stylo" lets you make your first stylometric analysis in no time.

Also, it might be worthwhile to visit this [discussion group](#).

**Maciej Eder** (center) is Director of the Institute of Polish Language at the Polish Academy of Sciences, and Associate Professor at the Institute of Polish Studies at the Pedagogical University of Kraków, Poland. He is interested in European literature of the Renaissance and the Baroque, classical heritage in early modern literature, and scholarly editing (his most recent book is a critical edition of 16th-century Polish translations of *Dialogue of Salomon and Marcolf*). A couple of years ago while doing research on anonymous ancient texts, Eder discovered the fascinating world of computer-based stylometry and non-traditional authorship attribution. His work is now focused on a thorough re-examination of current attribution methods and applying them to non-English languages, e.g. Latin and Ancient Greek.

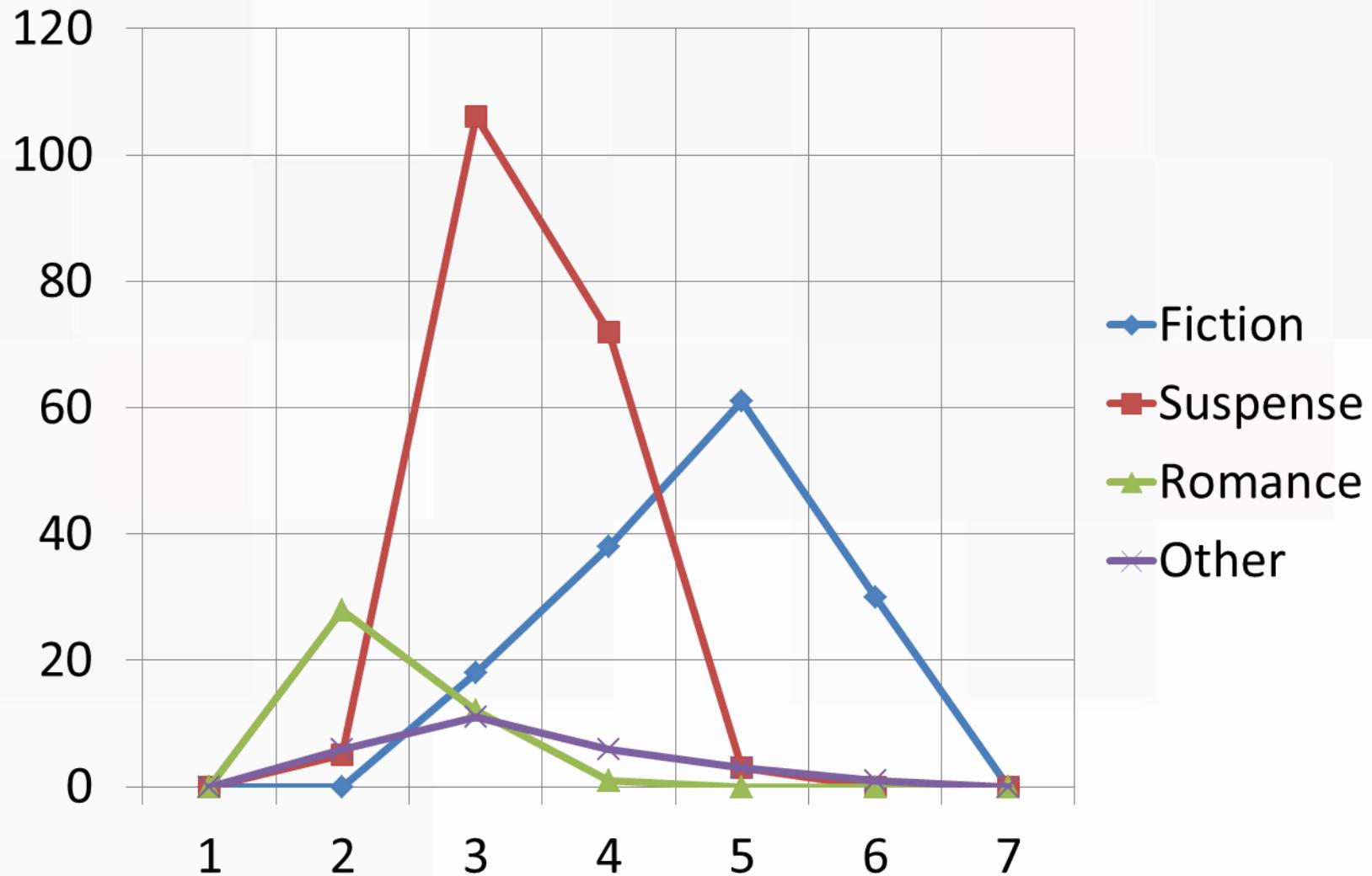
**Jan Rybicki** (left) is Assistant Professor at the Institute of English Studies, Jagiellonian University, Kraków, Poland; he also taught at Rice University, Houston, TX and Kraków's Pedagogical University. His interests include translation, comparative literature and humanities computing (especially stylometry and authorship attribution). He has worked extensively (both traditionally and digitally) on Henryk Sienkiewicz and the reception of the Polish novelist's works into English, and on the reception of English literature in Poland. Rybicki is also an active literary translator, with more than twenty translated novels by authors such as Coupland, Fitzgerald, Golding, Gordimer, Ie Carré or Winterson.





## Mean score literariness: **4.3**

- Literary fiction      **5.21**      147 novels
- Suspense                **3.91**      186 novels
- Romance                **2.95**      41 novels
- *Other\**                **3.81**
- 27 titles, non-fiction, collections of stories, young adult general, young adult fantasy, regional novels, Fifty shades....



## Readers of many languages!

Heb het in het Frans gelezen. Prachtig taalgebruik en actuele problematiek zeer origineel en tussen de regels weergegeven.

***Read it in French. Beautiful language use, current issues, very original and ‘between the lines’.***

Personen worden nogal vlak neergezet. Stijl is matig. (In het Engels gelezen) Veel herhalingen.

***Characters are rather flat. Style not very good (read it in English). A lot of repetition.***

## Some opinions

Te oppervlakkig. Gaat niet diep genoeg op de materie in. Kan ook aan de vertaling liggen.

***Too shallow. Does not delve deep enough into the topic. Could also be the translation.***

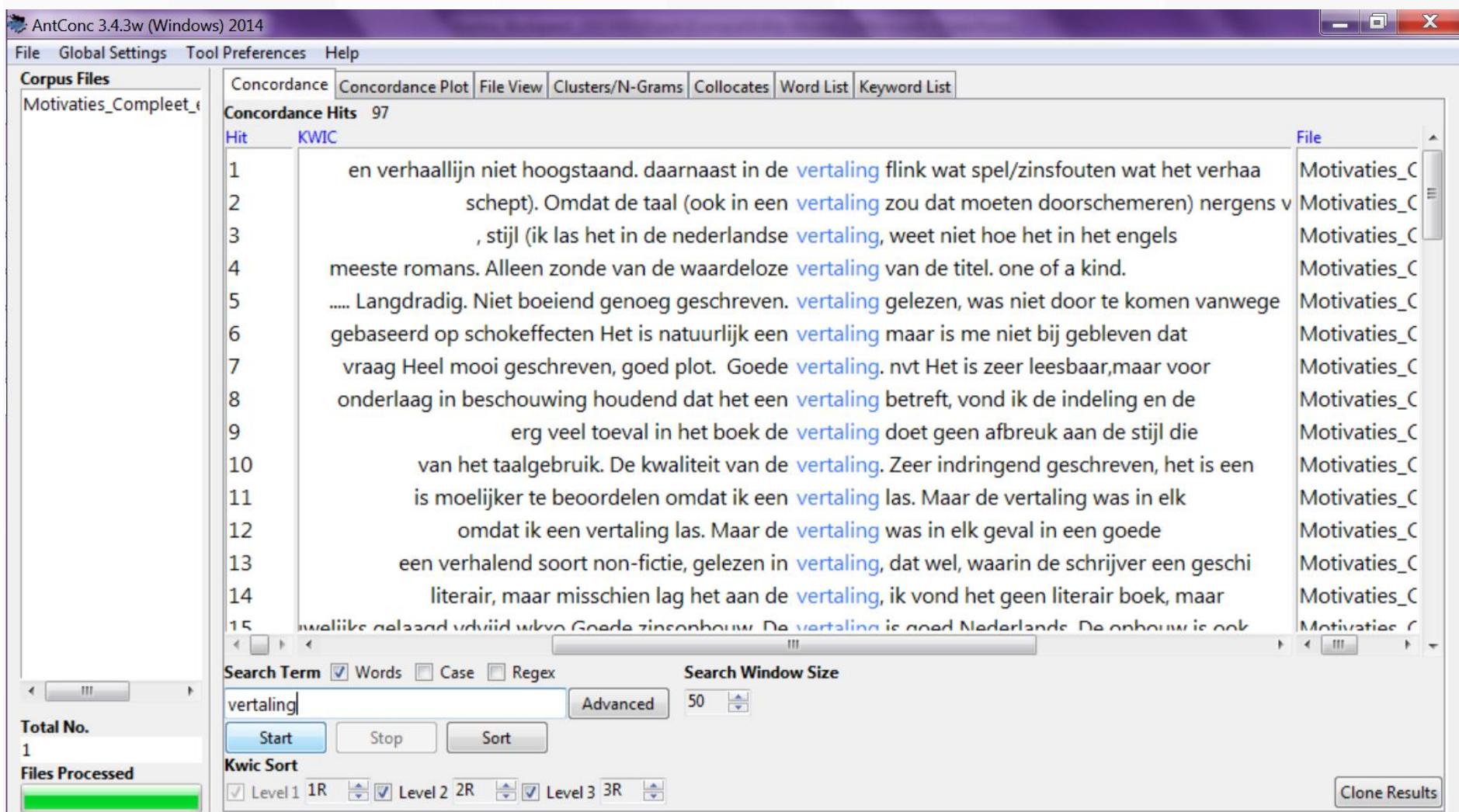
Verhaal een beetje over de top. Niet altijd mooi van taal. (Kan ook aan de vertaling liggen natuurlijk!)

***Story a bit over the top. Language not always agreeable. (Could also be the translation, of course!).***

# Original *versus* translation

	Dutch	Translated	Total
Literary fiction	73	74	147
Suspense	58	128	186
Romance	5	36	41
<i>Other</i>	16	11	27

# AntConc, by Laurence Anthony



The screenshot shows the AntConc 3.4.3w software interface. The menu bar includes File, Global Settings, Tool, Preferences, and Help. The main window displays a concordance search for the word "vertaling". The search results are shown in a table with columns for Hit number, KWIC (Context), and File. The KWIC column shows the context of the word in the text, and the File column shows the source file for each hit. The search term "vertaling" is highlighted in blue in the text. The search window size is set to 50. The total number of hits is 97. The search term "vertaling" is entered in the search term field. The Kwick Sort button is visible at the bottom left.

AntConc 3.4.3w (Windows) 2014

File Global Settings Tool Preferences Help

Corpus Files

Motivaties\_Compleet\_1

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Concordance Hits 97

Hit KWIC File

1 en verhaallijn niet hoogstaand. daarnaast in de **vertaling** flink wat spel/zinsfouten wat het verhaa Motivaties\_C

2 schept). Omdat de taal (ook in een **vertaling** zou dat moeten doorschemeren) nergens v Motivaties\_C

3 , stijl (ik las het in de nederlandse **vertaling**, weet niet hoe het in het engels Motivaties\_C

4 meeste romans. Alleen zonde van de waardeloze **vertaling** van de titel. one of a kind. Motivaties\_C

5 .... Langdradig. Niet boeiend genoeg geschreven. **vertaling** gelezen, was niet door te komen vanwege Motivaties\_C

6 gebaseerd op schokeffecten Het is natuurlijk een **vertaling** maar is me niet bij gebleven dat Motivaties\_C

7 vraag Heel mooi geschreven, goed plot. Goede **vertaling**. nvt Het is zeer leesbaar, maar voor Motivaties\_C

8 onderlaag in beschouwing houdend dat het een **vertaling** betreft, vond ik de indeling en de Motivaties\_C

9 erg veel toeval in het boek de **vertaling** doet geen afbreuk aan de stijl die Motivaties\_C

10 van het taalgebruik. De kwaliteit van de **vertaling**. Zeer indringend geschreven, het is een Motivaties\_C

11 is moeilijker te beoordelen omdat ik een **vertaling** las. Maar de vertaling was in elk Motivaties\_C

12 omdat ik een vertaling las. Maar de **vertaling** was in elk geval in een goede Motivaties\_C

13 een verhalend soort non-fictie, gelezen in **vertaling**, dat wel, waarin de schrijver een geschi Motivaties\_C

14 literair, maar misschien lag het aan de **vertaling**, ik vond het geen literair boek, maar Motivaties\_C

15 ...welk... gelezen vdyldt wkyo Goede zinsopbouw. De **vertaling** is goed Nederlands. De opbouw is ook Motivaties\_C

Search Term  Words  Case  Regex

vertaling Advanced

Search Window Size 50

Total No. 1

Files Processed

Kwick Sort

Level 1 1R  Level 2 2R  Level 3 3R

Clone Results



	Keywords in motivations of Dutch literary titles		Keywords in motivations of translated literary titles
Taalgebruik	<i>Language use</i>	Vertaling	<i>Translation</i>
Moeder	<i>Mother</i>	Verhaallijn	<i>Plot</i>
Taal	<i>Language</i>	Herhalingen	<i>Repetitions</i>
Journalistiek	<i>Journalism</i>	Verhaal	<i>Story</i>
Proza	<i>Prose</i>	Deel	<i>Volume</i>
Jeugd	<i>Youth</i>	Boek	<i>Book</i>
Stijl	<i>Style</i>	Hoofdpersonen	<i>Main characters</i>
Boekje	<i>Booklet</i>	Spel	<i>Play</i>
Uitdieping	<i>In-depth study</i>	Kaart	<i>Map</i>
Verdriet	<i>Sorrow</i>	Verhouding	<i>Relation</i>
Kinderen	<i>Children</i>	Drijfveren	<i>Motivations</i>
Verhaaltjes	<i>Stories</i>	Huwelijk	<i>Marriage</i>
Zinnen	<i>Sentences</i>	Richting	<i>Direction</i>
Afstand	<i>Distance</i>	Slavernij	<i>Slavery</i>
Poes	<i>Cat</i>	Woordkeus	<i>Choice of word</i>
volzinnen	<i>Stylish sentences</i>	Kind	<i>Child</i>

## Mean score literariness: 4.3

Dutch :	<u>4.5</u>	Translated:	<u>4.1</u>
Lit. fiction NL:	<u>5.3</u>	Lit. fiction translated:	<u>5.1</u>
Suspense NL:	<u>3.7</u>	Suspense translated :	<u>4.0</u>
Romance NL:	<u>2.8</u>	Romance translated :	<u>3.0</u>
Other NL:	<u>3.8</u>	<i>Other translated:</i>	<u>3.8</u>

# A new experiment

- Are literary translations considered to be less literary than originals?
- Is there a hierarchy in quality related to original language / country of origin?
- If so, is this hierarchy the same in different countries?

# The experiment

- Great-Britain, France, the Netherlands, and Germany
- Representative sample of citizens between age 18-65, 150 participants for each country
- Opinion about literariness of 24 novels, six per language - three male, three female authors



# The books we selected



Amanda Hodgkinson  
David Mitchell  
Emma Donoghue  
Hilary Mantel  
John Boyne  
Julian Barnes

22 Brittania Road  
The Thousand Autumns of Jacob de Zoet  
Room  
Bring up the bodies  
The house of special purpose  
The sense of an ending

22 Brittania Road  
Les mille automnes de Jacob de Zoet  
Room  
Le pouvoir  
La maison des intentions particulières  
Une fille, qui danse

22 Brittania Road  
De niet verhoorde gebeden van  
Kamer  
Het boek Henry  
Het winterpaleis  
Alsof het voorbij is

22 Brittania Road  
Die tausend Herbste  
Raum  
Falken  
Das Haus zur beson  
Vom Ende einer Ges

Delphine de Vigan  
Jean-Marie Gustave Le Clézio  
Laurent Binet  
Marie Ndiaye  
Maylis de Kerangal  
Michel Houellebecq

Nothing Holds Back the Night  
The African  
HhhH  
Ladivine  
The Heart / Mend the living  
The map and the territory

Rien ne s'oppose à la nuit  
L'Africain  
HhhH  
Ladivine  
Réparer les vivants  
La carte et le territoire

Niets weerstaat de nacht  
De Afrikaan  
HhhH - Himmlers hersens hete  
Ladivine  
De levenden herstellen  
De kaart en het gebied

Das Lächeln meiner  
Der Afrikaner  
HHhH Himmlers Hirn  
Ladivine  
Die Lebenden reparieren  
Karte und Gebiet

Anna Enquist  
Arnon Grunberg  
Gerbrand Bakker  
Herman Koch  
Margriet de Moor  
Renate Dorresteijn

Counterpoint  
Tirza  
The detour  
The dinner  
The storm  
A Heart of Stone

Contrepont  
Tirza  
Le détour  
Le diner  
Une catastrophe naturelle  
Un coeur de pierre

Contrapunt  
Tirza  
De omweg  
Het diner  
De verdronkene  
Een hart van steen

Kontrapunkt  
Tirza  
Der Umweg  
Angerichtet  
Sturmflut  
Hertz aus Stein

Charles Lewinsky  
Charlotte Roche  
Herta Müller  
Ingo Schulze  
Julia Franck  
Uwe Tellkamp

Melnitz  
Wetlands  
The hunger angel  
Adam and Evelyn  
The Blind Side of the Heart / The Blindness  
The Tower: tales from a lost country

Melnitz  
Zones Humides  
La bascule du souffle  
Adam et Èvelyne  
La femme de midi  
La tour: histoire en provenance d'une

Het lot van de familie Meijer  
Vochtige streken  
Ademschommel  
Adam en Evelyn  
De middagvrouw  
De toren: verhaal uit een verzo

Melnitz  
Feuchtgebiete  
Atemschaukel  
Adam und Evelyn  
Die Mittagsfrau  
Der Turm: Geschich



<b>Titel</b>	<b>auteur</b>	<b>UK (A)</b>	<b>FR (B)</b>	<b>NL (C )</b>	<b>DE (D)</b>
22 Britannia Road	Amanda Hodgkinson	4,65 D	4,37	4,30	4,25
De niet verhoorde gebeden van Jacob de Zoet	David Mitchell	4,85	4,90	4,71	5,04
Kamer	Emma Donoghue	4,72	4,44	4,46	4,74
Het boek Henry	Hilary Mantel	4,43	4,49	4,52	4,55
Het winterpaleis	John Boyne	4,75 B	4,37	4,46	4,67
Alsof het voorbij is	Julian Barnes	4,65	4,60	4,51	4,95 BC
Niets weerstaat de nacht	Delphine de Vigan	4,53	4,74	4,52	4,81
De Afrikaan	Jean-Marie Gustave Le Guin	4,64	4,95 C	4,43	4,76
HhhH - Himmlers hersens heten Heydrich	Laurent Binet	4,40	4,33	4,66	4,35
Ladivine	Marie Ndiaye	4,55	4,54	4,33	4,40
De levenden herstellen	Maylis de Kerangal	4,25	4,51	4,53	4,68 A
De kaart en het gebied	Michel Houellebecq	4,28	4,71 A	4,54	4,54
Contrapunt	Anna Enquist	4,58	4,55	4,85	4,66
Tirza	Arnon Grunberg	4,48	4,62	5,00 AD	4,43
De omweg	Gerbrand Bakker	4,43	4,53	4,61	4,63
Het diner	Herman Koch	4,54	4,35	4,82 BD	4,20
De verdronkene	Margriet de Moor	4,51	4,19	4,49	4,76 B
Een hart van steen	Renate Dorrestein	4,60	4,65	4,71	4,45
Het lot van de familie Meijer	Charles Lewinsky	4,44	4,78	4,59	4,52
Vochtige streken	Charlotte Roche	4,62 BD	4,21 D	4,20 D	3,03
Ademschommel	Herta Müller	4,45	4,74	4,55	4,65
Adam en Evelyn	Ingo Schulze	4,48	4,42	4,33	4,48
De middagvrouw	Julia Franck	4,61	4,69	4,52	4,59
De toren: verhaal uit een verzonken land	Uwe Tellkamp	4,57	4,38	4,37	4,88 BC

Overall

4,54

4,54

4,54

4,54

## Three questions:

If you compare **English** literature to **German** literature, which of the statements below do you find most adequate:

1. There is no difference
2. English literature usually has a higher literary quality
3. German literature usually has a higher literary quality
4. I do not know

Next, the same question for the other two languages.

# Results for each country

- **English:** English has higher quality than Dutch and German, comparable with French
- **French:** French has higher quality than Dutch and German, comparable with English
- **Dutch:** Dutch has a lower quality than English, but comparable with German and French
- **German:** German has a lower quality than English, but comparable with French and higher than Dutch

# Question: Rank all four literatures



UK

UK



FR

F



NL



DE

D

1. English
2. French
3. German
4. Dutch

1. French
2. English
3. German
4. Dutch

1. **English**
2. Dutch
3. French
4. German

1. German
2. **English**
3. French
4. Dutch

# Hierarchy total



UK

1. English



FR

2. French



DE

3. German



4. Dutch

# Some motivations for English

aus dem gefühl heraus

***Just a feeling***

based on my own nationality

beaucoup d'auteurs littéraires

***Many literary authors***

Beter woordgebruik en andere thema's

***Better language use and different themes***

harry potter is the best thing ever

Hebben bij mijn weten de meeste grote auteurs voortgebracht

***To my knowledge have the largest number of great authors***

het is gewoon een feit, net als britse films of series

***It's just a fact, same as British films or series***

we are the home of it :)

we have dickens and eliot etc

we have the best authors

# Some motivations for German

allein durch die Namensangabe

***Just by looking at the names***

anything is better than english

ben de taal magtig [sic: spelling error for 'machtig']

***I master the language***

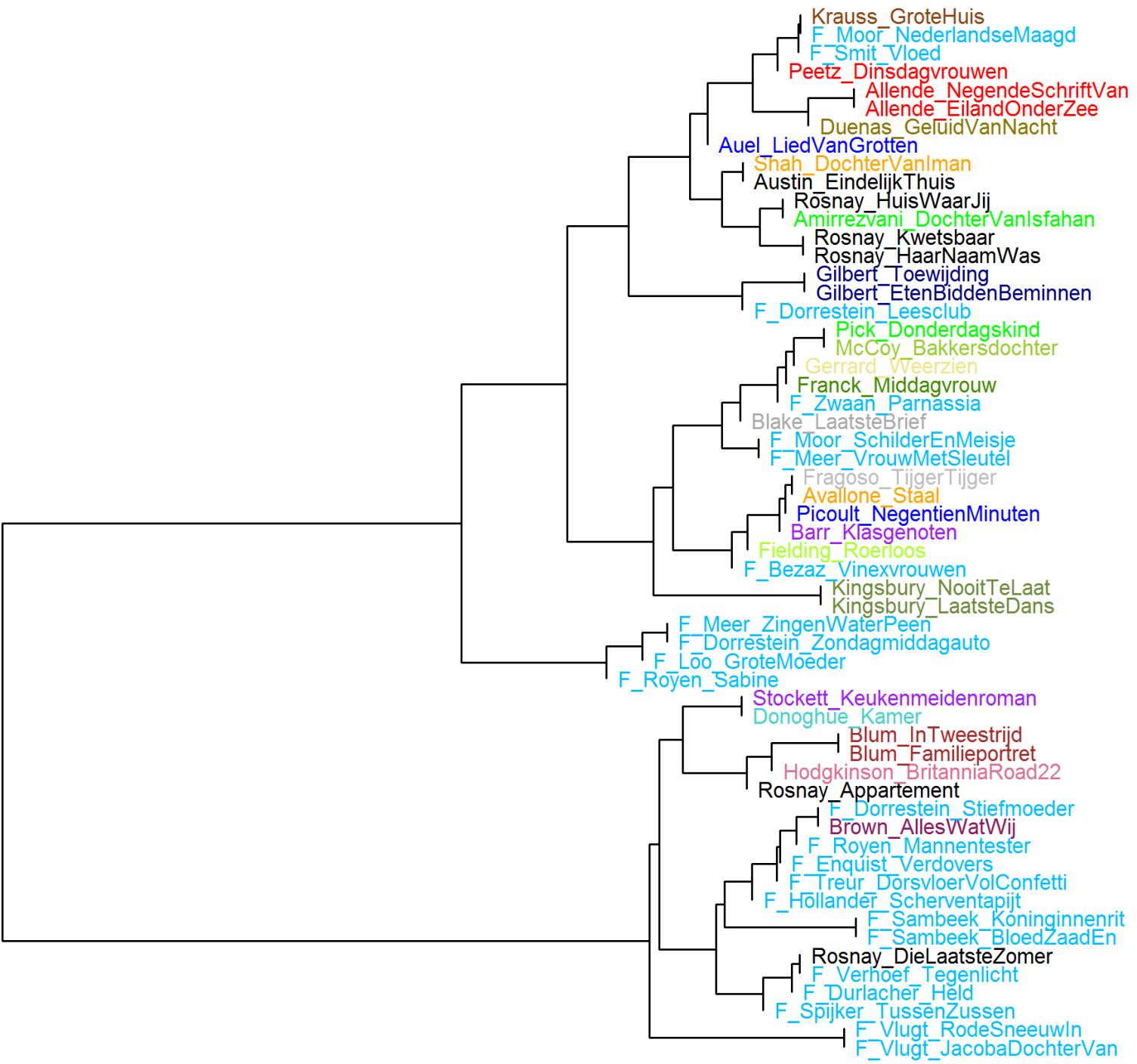
Plus detaille

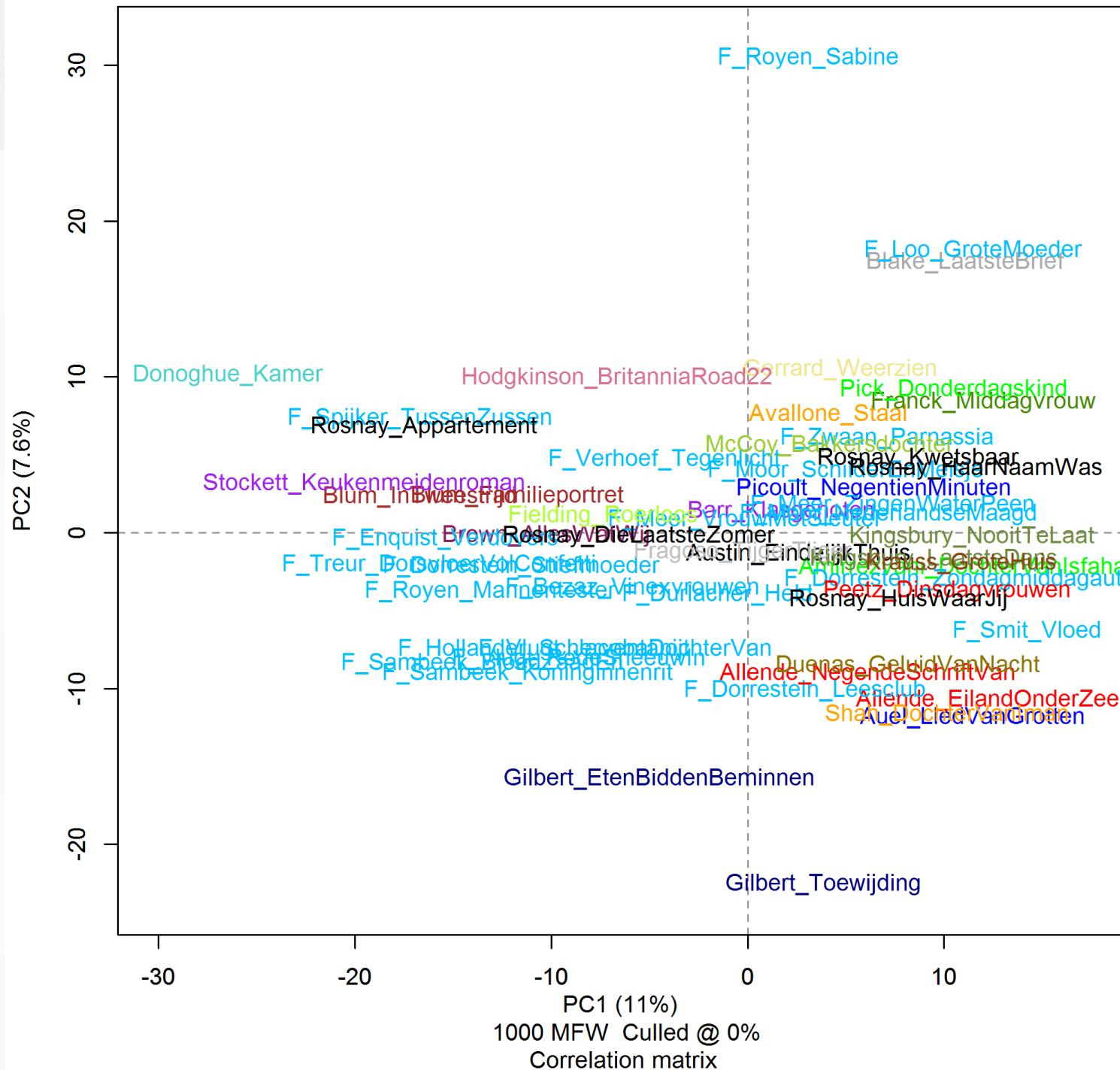
***More detailed***

## Back to the novels

- Comparing vocabulary and word frequencies in Dutch originals and translations
- Limited to literary fiction
- Limited to female authors

	<b>Female authors</b>	Male authors
Novels orig. NL	<b>23</b>	50
Transl. from English	<b>25</b>	19
Transl. From other lang.	9	21
<i>Total</i>	<b>57</b>	<b>90</b>





A	B	C	D	E	F	G	H	I	J	K	L	M
1	13329	1579.63	mijn	34	360	387.359	dennis	67	107	238.628	axel	
2	661	1474.14	katelijne	35	178	385.396	hendrik	68	106	236.398	spanjaarden	
3	590	1301.84	isabella	36	170	379.129	elsje	69	218	234.977	prins	
4	488	1088.32	suzan	37	168	374.669	nico	70	102	227.478	margaretha	
5	485	1081.63	annelies	38	164	365.748	ewoud	71	102	227.478	sophia	
6	478	1066.02	willem	39	163	363.518	egon	72	100	223.017	eugenie	
7	477	1012.13	jacob	40	39017	356.172	ik	73	100	223.017	stijn	
8	421	938.903	drik	41	159	354.597	marit	74	98	218.557	breda	
9	376	838.545	mich	42	154	343.447	renate	75	157	215.121	hans	
10	360	802.862	andries	43	153	341.216	tessa	76	145	214.006	jean	
11	355	791.711	lideweij	44	1742	337.806	wordt	77	95	211.866	puck	
12	347	741.724	lucien	45	149	332.296	joop	78	95	211.866	rosalie	
13	16466	721.667	is	46	1753	330.774	komt	79	691	209.119	ligt	
14	318	709.195	tess	47	145	323.375	bötticher	80	93	207.406	folkert	
15	350	684.737	an	48	141	314.454	frits	81	173	206.824	roos	
16	276	615.528	aron	49	136	303.303	christiaan	82	92	205.176	agnes	
17	341	593.379	claire	50	554	301.398	oma	83	100	204.919	ok	
18	261	504.405	joost	51	9412	295.654	me	84	91	202.946	guido	
19	224	499.559	allard	52	42509	294.948	het	85	116	201.156	vera	
20	218	486.178	marieke	53	149	292.158	nederland	86	94	199.334	johan	
21	292	483.563	jan	54	130	289.922	gideon	87	89	198.485	betje	
22	223	476.044	sandra	55	130	289.922	sacha	88	168	195.47	koning	
23	5605	454.047	heeft	56	2449	288.515	toch	89	2388	191.884	mij	
24	64592	451.961	de	57	125	278.772	leni	90	822	191.805	mama	
25	231	439.893	amsterdam	58	261	278.357	patiënt	91	85	189.565	georgette	
26	269	436.648	von	59	122	272.081	heinz	92	84	187.335	cathérine	
27	194	432.654	josefien	60	119	265.391	renée	93	5545	186.325	ook	
28	193	430.423	rogier	61	121	259.046	herman	94	2938	182.903	daar	
29	197	427.567	humphrey	62	138	252.101	marie	95	82	182.874	anneke	
30	2325	402.997	vader	63	112	249.779	reinier	96	82	182.874	edelen	
31	177	394.741	fintan	64	229	249.521	collega	97	82	182.874	evelien	
32	176	392.51	filips	65	109	243.089	jeroen	98	141	182.181	leida	
33	209	387.501	schilder	66	159	241.602	kasteel	99	180	178.414	sonia	
								100			adriana	

A	B	C	D	E	F	G	H	I	J	K	L	M
1	2951	2345.03	ayla	34	1535	547.667	mam	67	389	309.122	felipe	
2	11998	2289.15	zei	35	677	537.983	elsie	68	652	305.12	tim	
3	74369	2270.52	ze	36	1747	511.46	anna	69	383	304.354	pepik	
4	49780	1734.3	haar	37	4018	487.044	keek	70	371	294.818	gostaham	
5	2083	1655.27	karena	38	600	476.794	aurek	71	365	290.05	patsy	
6	33487	1234.92	was	39	597	474.41	frankie	72	349	277.335	obersturmführer	
7	1392	1091.92	grot	40	612	473.725	kevin	73	766	275.559	glimlachte	
8	1396	1063.41	miss	41	2037	470.641	t	74	1218	270.37	baby	
9	1336	1061.66	jondalar	42	3892	470.543	vroeg	75	869	267.747	r	
10	1316	1045.77	kirsten	43	58292	467.641	dat	77	22043	256.786	om	
11	1273	1011.6	sofia	44	601	465.02	warren	78	916	252.645	d	
12	1246	990.143	kari	45	715	439.325	wolf	79	315	250.317	janine	
13	84611	986.378	en	46	525	417.195	matt	80	311	247.138	gordiyeh	
14	1236	982.197	elin	47	618	416.913	patrick	81	1521	245.355	voordat	
15	10315	971.159	toen	48	3649	413.693	wist	82	1196	244.425	probeerde	
16	1200	953.589	casey	49	581	408.139	emma	83	2645	243.919	hen	
17	1126	894.784	abby	50	499	396.534	pavel	84	305	242.371	iris	
18	21842	893.494	had	51	488	387.793	reba	85	300	238.397	skeeter	
19	1175	880.256	trudy	52	592	385.083	amanda	86	299	237.603	zelandonia	
20	1135	877.028	margaret	53	519	382.181	ralph	87	3706	231.314	ging	
21	2145	799.776	peter	54	463	367.926	aibileen	88	287	228.067	naheed	
22	1225	778.557	charles	55	6347	365.919	waren	89	301	228.005	negende	
23	967	768.434	josie	56	478	358.392	ryan	90	282	224.093	fereydoon	
24	948	753.335	zelandoni	57	442	351.239	jonayla	91	280	222.504	mélanie	
25	833	661.95	marnie	58	442	351.239	tamsin	92	3627	215.414	terwijl	
26	826	656.387	janice	59	6340	350.422	kon	93	1640	211.076	god	
27	808	628.923	lizzie	60	439	348.855	izzy	94	1042	209.487	hoewel	
28	787	625.395	marta	61	451	346.396	nicole	95	707	204.823	oké	
29	774	615.065	silvana	62	416	330.578	minny	96	623	204.733	alex	
30	1022	611.391	john	63	398	316.274	lacy	97	316	199.231	dollar	
31	735	584.073	janusz	64	396	314.684	celia	98	2521	197.879	deed	
32	722	573.743	drew	65	391	310.711	anneliese	99	248	197.075	riki	
33	716	568.975	hilly	66	390	309.916	jordan	100	271	196.162	tony	

# Possible follow-up steps

- Concordance analysis selected words
- Compare on title level with survey scores
- Alignment originals and translations
- *Dynamic* contemporary corpus....
- ...with detailed genre metadata
- 'Keyness alert'

# Other ways of counting words

- Vocabulary richness
- Amount of dialogue
- Sentence length
- Sentence complexity
- Distribution of topics

# Topic Modeling Literary Quality

*Kim Jautze,<sup>1</sup> Andreas van Cranenburgh,<sup>1,2</sup> and Corina Koolen<sup>2</sup>*

{kim.jautze, andreas.van.cranenburgh}@huygens.knaw.nl, c.w.koolen@uva.nl

<sup>1</sup> Huygens ING, Royal Netherlands Academy of Arts and Sciences

<sup>2</sup> Institute for Logic, Language and Computation, University of Amsterdam

## Introduction

To what extent can topic models explain variation in perceptions of literary quality? We try to find correlations between topics and judgments of literary quality using a topic model of 401 recent bestselling Dutch novels. Instead of examining topics on a macro-scale in a geographical or historical interpretation (e.g., Jockers 2013; Riddell 2014), we take a new perspective: whether novels have a dominant topic in their topic distributions (mono-topicality), and whether certain topics may express an explicit or implicit genre in the corpus. We hypothesize that there is a relationship between these aspects of the topic distributions and perceptions of literary

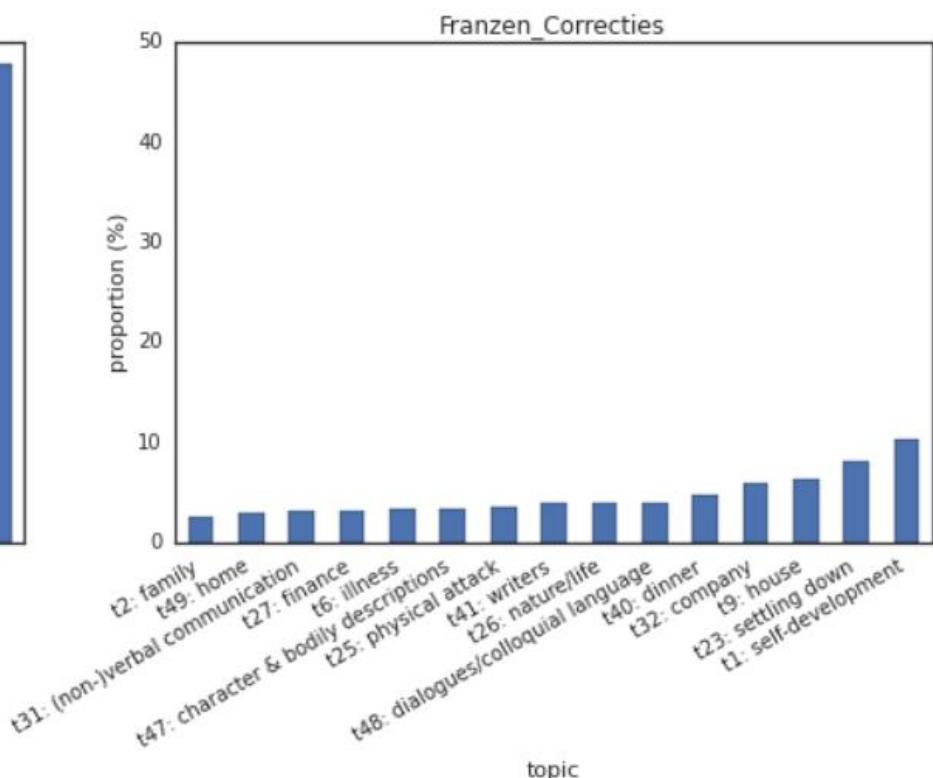
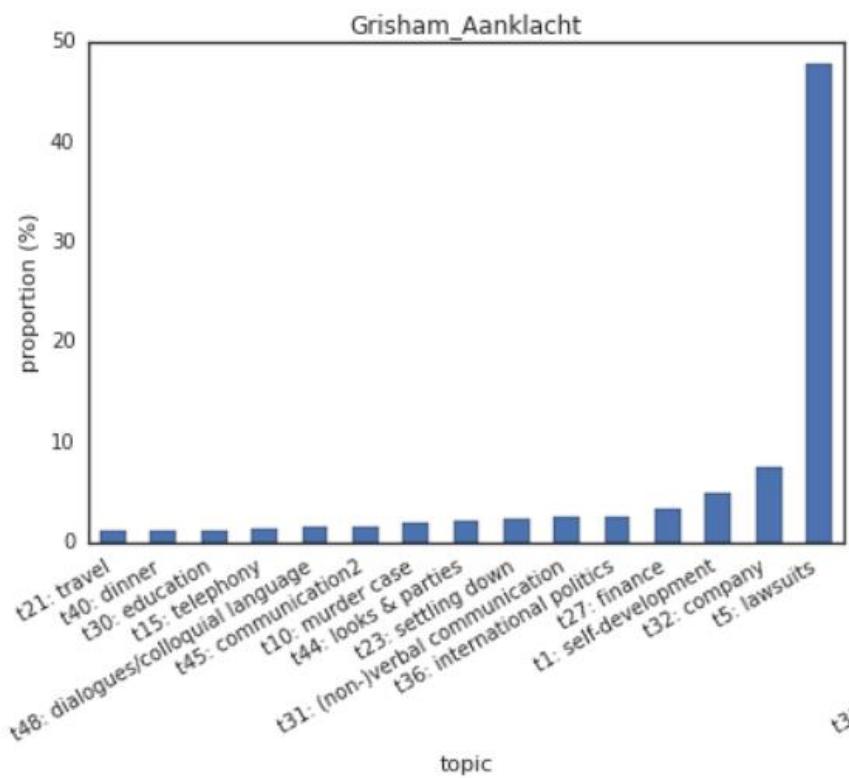


Fig. 2: Distribution of the top 15 topics in novels with high (left) and low (right) mono-topicality

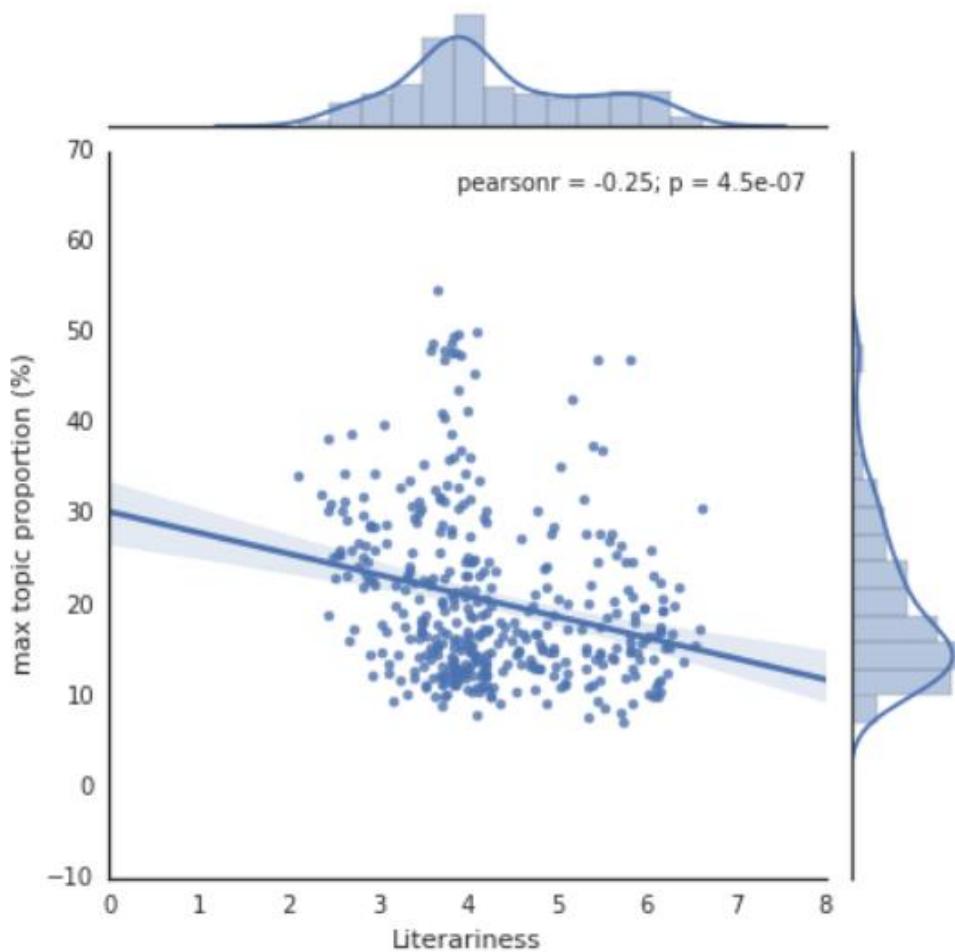


Fig. 3: Correlation between share of the most prominent topic per book and mean literariness ratings

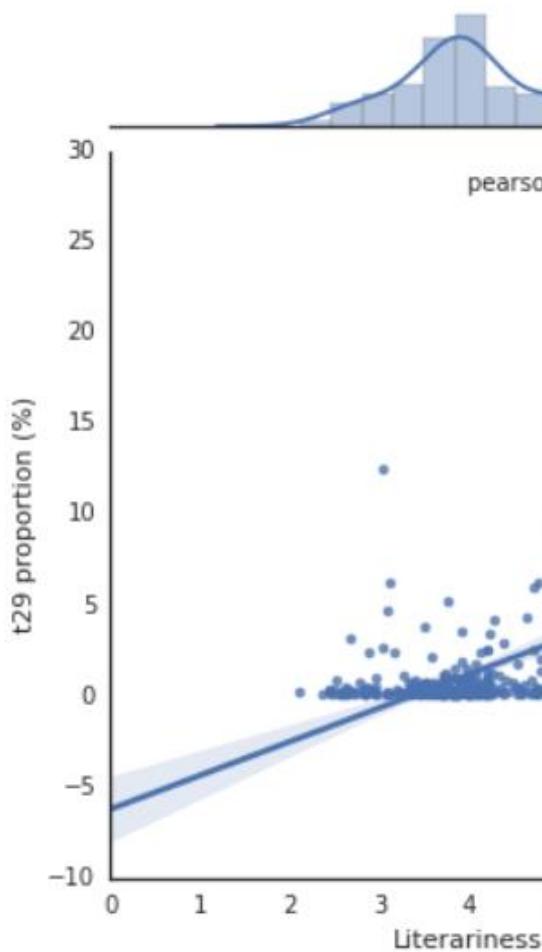
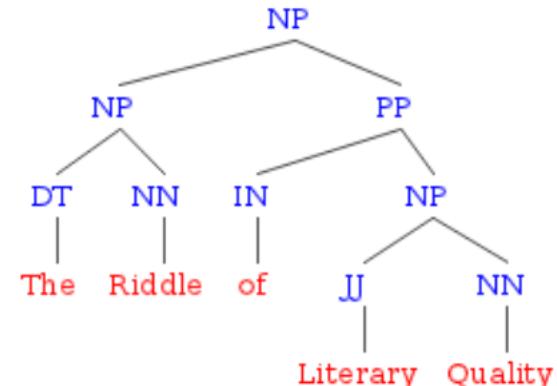


Fig. 4: Correlation between proportion and mean literariness

# Academic homepage of Andreas van Cranenburgh

I am a postdoc at Heinrich Heine Universität Düsseldorf in the [Beyond CFG](#) project. I was previously a PhD candidate in the project [The Riddle of Literary Quality](#). My primary interests are statistical parsing and syntactic patterns, with particular interest in tree fragments and discontinuous constituents.



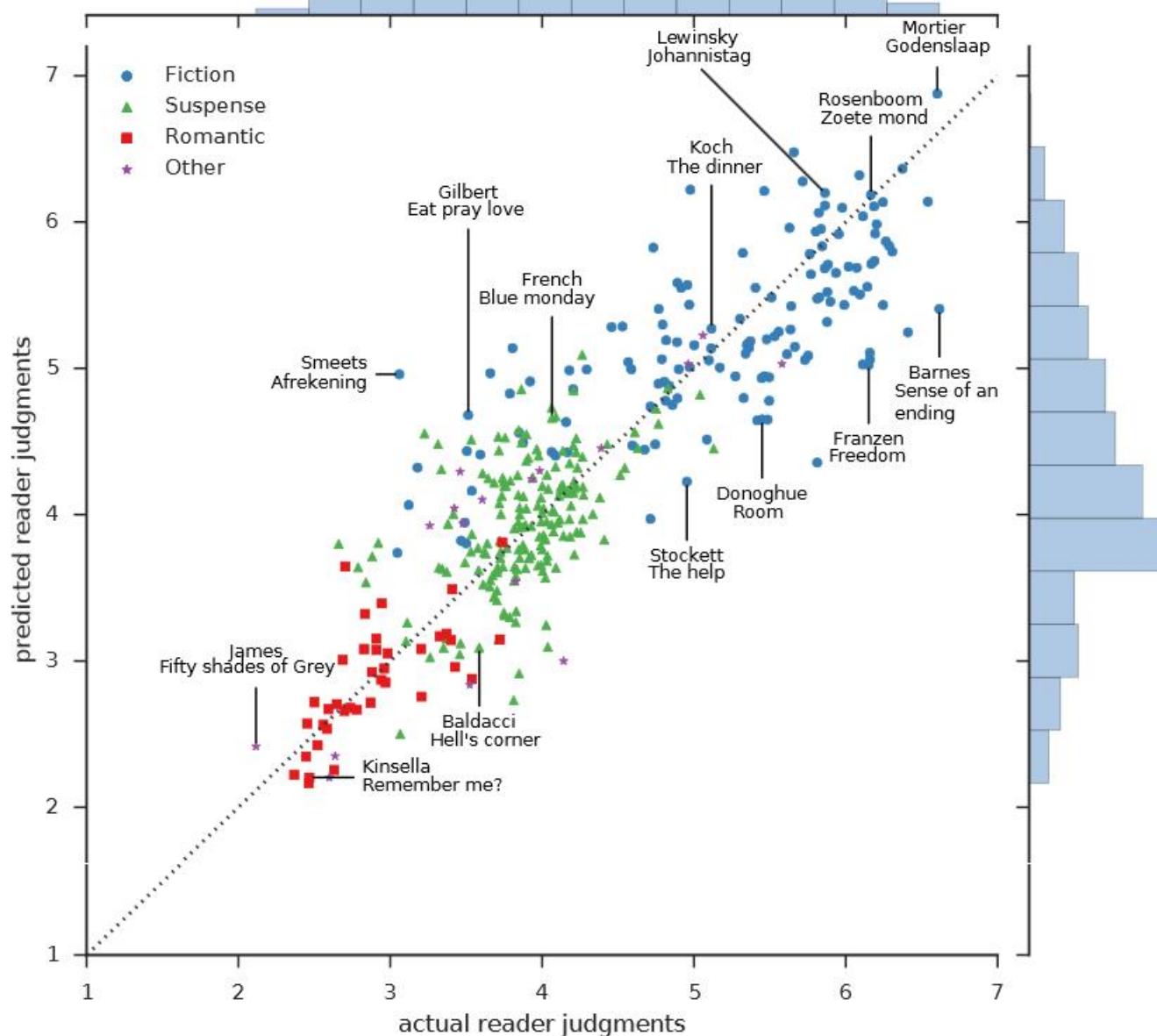
Mail: [cranenburgh@phil.hhu.de](mailto:cranenburgh@phil.hhu.de)

Code: <https://github.com/andreasvc> and <https://gist.github.com/andreasvc/>

Profiles: [Google Scholar](#); [Semantic Scholar](#).

## Education

- PhD in Computational Linguistics (2016), University of Amsterdam. [PhD thesis](#): Rich statistical parsing and literary language ([revised version](#); [errata](#)).
- MSc. in Logic (2011), University of Amsterdam. [MSc. thesis](#): Discontinuous Data-Oriented Parsing through Mild Context-Sensitivity. ([code](#)).
- BSc. in Artificial Intelligence (2009), University of Amsterdam. [BSc. thesis](#): Simulating Language Games in the Two Word Stage.



**Not** The End.



was a dark and stormy night.  
The rain fell in torrents — ex-